CS 4/5789 - Prelim Exam Equation/Definition Sheet

- Infinite Horizon Discounted MDP is $(\mathcal{S}, \mathcal{A}, r, P, \gamma)$, where
  - $\mathcal{S}$: The state space, the set of all states
  - $\mathcal{A}$: Action space, set of all actions.
  - r: Reward function maps $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$. $r(s, a)$ is the reward of taking action a in state s.
  - P: Transition function. Either written as $P(s, a)$ mapping $\mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ (a distribution over states) or mapping $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, where $P(s' \mid s, a)$ is the probability of reaching state s' given that you take action a in state s.
  - γ: Discount factor between 0 and 1.
- Finite Horizon MDP is $(\mathcal{S}, \mathcal{A}, r, P, H)$, where $\mathcal{S}, \mathcal{A}, r, P$ are as above, and
  - H: a positive integer representing the time horizon.
- Optimal Control Problem: Finite Horizon MDP where
  - $\mathcal{S} = \mathbb{R}^{n_s}$ and $\mathcal{A} = \mathbb{R}^{n_a}$.
  - $r(s, a) = -c(s, a)$ where c is a cost function mapping $\mathcal{S} \times A \to \mathbb{R}$.
  - Transitions are deterministic and described by the dynamics function f mapping $\mathcal{S} \times \mathcal{A} \to \mathcal{S}$.
- Policy denoted by π can either be written as $\pi(s)$ a map from state to a distribution over actions $\mathcal{S} \to \Delta(\mathcal{A})$ or as $\pi(a|s)$ a map from a state and action to a probability, $\mathcal{A} \times \mathcal{S} \to [0, 1]$. We write deterministic policies as mapping $\mathcal{S} \to \mathcal{A}$.
- Value function, Q function:
  - **Infinite horizon:**

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s\right], \quad Q^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a\right]$$

  - **Finite horizon:** for $k = 0, \dots H-1$

$$V_k^\pi(s) = \mathbb{E}\left[\sum_{t=k}^{H-1} r(s_t, a_t) \mid s_k = s\right], \quad Q_k^\pi(s, a) = \mathbb{E}\left[\sum_{t=k}^{H-1} r(s_t, a_t) \mid s_k = s, a_k = a\right]$$

- Bellman Expectation Equation:
  - **Infinite horizon:**

$$V^\pi(s) = \mathbb{E}_{a \sim \pi(s)}\left[r(s, a) + \gamma \mathbb{E}_{s' \sim P(s,a)}[V^\pi(s')]\right], \quad Q^\pi(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(s,a)}\left[V^\pi(s')\right]$$

  - **Finite horizon:** for $t = 0, \dots H-1$

$$V_t^\pi(s) = \mathbb{E}_{a \sim \pi_t(s)}\left[r(s, a) + \mathbb{E}_{s' \sim P(s,a)}[V_{t+1}^\pi(s')]\right], \quad Q_t^\pi(s) = r(s, a) + \mathbb{E}_{s' \sim P(s,a)}[V_{t+1}^\pi(s')]$$

- Bellman Optimality Equation:
  - **Infinite horizon:** $\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s, a)$,

$$V^*(s) = \max_{a \in \mathcal{A}}\left[r(s, a) + \gamma \mathbb{E}_{s' \sim P(s,a)}\left[V^*(s')\right]\right], \quad Q^*(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(s,a)}\left[\max_{a' \in \mathcal{A}} Q^*(s', a')\right]$$

- **Finite horizon:** for $t = 0, \ldots H-1$, $\pi_t^*(s) = \arg\max_{a \in \mathcal{A}} Q_t^*(s,a)$

$$V_t^*(s) = \max_{a \in \mathcal{A}} \left[r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \left[V_{t+1}^*(s') \right] \right], \quad Q_t^*(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \left[\max_{a' \in \mathcal{A}} Q_{t+1}^*(s',a') \right]$$

- The Advantage function: $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$
- State-action distributions for initial state distribution μ_0:
  - $\mathbb{P}_{\mu_0}^\pi(s_0, a_0, s_1, a_1, \ldots, s_t, a_t) = \mu_0(s_0)\pi(a_0 \mid s_0) \cdot \prod_{i=1}^{t} P(s_i \mid s_{i-1}, a_{i-1})\pi(a_i \mid s_i)$
  - $\mathbb{P}_t^\pi(s; \mu_0) = \sum_{\substack{s_{0:t-1} \\ a_{0:t}}} \mathbb{P}_{\mu_0}^\pi(s_{0:t-1}, a_{0:t-1}, s_t, a_t \mid s_t = s)$
  - $d_{\mu_0,t}^\pi(s) = \mathbb{P}_t^\pi(s; \mu_0)$ and $d_{\mu_0,t}^\pi = P_\pi^\top d_{\mu_0,t-1}^\pi$ where P_π has the value $\mathbb{E}_{a \sim \pi(s)}[P(s'|s,a)]$ at row s and column s'.
  - $d_{\mu_0}^\pi(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_t^\pi(s; \mu_0)$
- Linear dynamics (transition) function and unrolled trajectory expression:
  - **One step:** $s_{t+1} = As_t + Ba_t$
  - **Multiple step:** $s_t = A^t s_0 + \sum_{k=0}^{t-1} A^k B a_{t-k-1}$
- Stability: for $s_{t+1} = As_t$,
  - $\max_{i \in [n_s]} |\lambda_i(A)| < 1 \rightarrow$ system is stable
  - $\max_{i \in [n_s]} |\lambda_i(A)| = 1 \rightarrow$ system is marginally (un)stable
  - $\max_{i \in [n_s]} |\lambda_i(A)| > 1 \rightarrow$ system is unstable
- Gradient Approximations: to $J(\theta)$
  - For small $\delta > 0$, $\nabla J(\theta) \approx \mathbb{E}_{v \sim \mathcal{N}(0,I)} \left[\frac{1}{2\delta} \left(J(\theta + \delta v) - J(\theta - \delta v) \right) v \right]$
  - If we can write $J(\theta) = \mathbb{E}_{x \sim P_\theta}[h(x)]$ then $\nabla J(\theta) = \mathbb{E}_{x \sim P_\theta}[\nabla_\theta \{\log P_\theta(x)\} h(x)]$
- Entropy of distribution $P \in \Delta(\mathcal{X})$ is defined as $\mathsf{Ent}(P) = \mathbb{E}_{x \sim P}\left[-\log P(x) \right]$
- Constrained optimization: the following are equivalent

$$\max_x \quad f(x) \quad \text{s.t.} \quad g(x) = 0 \iff \max_x \min_w \quad f(x) + w^\top g(x)\,.$$

  Optimal values occur at the critical points of the weighted sum, i.e. where

$$\nabla_x [f(x) + w^\top g(x)] = \nabla_w [f(x) + w^\top g(x)] = 0.$$

- Matrix/vector calculus: $\nabla_x \left\{ a^\top x \right\} = a$, $\nabla_x \left\{ x^\top A x \right\} = (A^\top + A)x$
- Geometric Sum: for any $\gamma \neq 1$, $\sum_{t=0}^{n-1} \gamma^t = \frac{1-\gamma^n}{1-\gamma}$ and for $0 < \gamma < 1$, $\sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma}$.