Update equations cheat sheet

- DP value iteration: $v_{k+1}(s) = \max_a \sum_{s',r} p\left(s', r|s, a\right)\left[r + \gamma v_k\left(s'\right)\right]$
- DP policy evaluation: $v_{k+1}(s) = \sum_a \pi(a|s) \sum_{s',r} p\left(s', r|s, a\right)\left[r + \gamma v_k\left(s'\right)\right]$
- Monte Carlo: $V\left(S_t\right) \leftarrow V\left(S_t\right) + \alpha\left[G_t - V\left(S_t\right)\right]$
- TD(0): $V\left(S_t\right) \leftarrow V\left(S_t\right) + \alpha\left[R_{t+1} + \gamma V\left(S_{t+1}\right) - V\left(S_t\right)\right]$
- SARSA: $Q\left(S_t, A_t\right) \leftarrow Q\left(S_t, A_t\right) + \alpha\left[R_{t+1} + \gamma Q\left(S_{t+1}, A_{t+1}\right) - Q\left(S_t, A_t\right)\right]$
- Expected SARSA: $Q(S_t, A_t) \leftarrow Q\left(S_t, A_t\right) + \alpha\left[R_{t+1} + \gamma \sum_a \pi\left(a|S_{t+1}\right) Q\left(S_{t+1}, a\right) - Q\left(S_t, A_t\right)\right]$
- Q-learning: $Q\left(S_t, A_t\right) \leftarrow Q\left(S_t, A_t\right) + \alpha\left[R_{t+1} + \gamma \max_a Q\left(S_{t+1}, a\right) - Q\left(S_t, A_t\right)\right]$
- Gradient Monte Carlo: $\mathbf{w} \leftarrow \mathbf{w} + \alpha\left[G_t - \hat{v}\left(S_t, \mathbf{w}\right)\right] \nabla \hat{v}\left(S_t, \mathbf{w}\right)$
- Semi-gradient TD: $\mathbf{w} \leftarrow \mathbf{w} + \alpha\left[R + \gamma \hat{v}\left(S', \mathbf{w}\right) - \hat{v}(S, \mathbf{w})\right] \nabla \hat{v}(S, \mathbf{w})$
- LSTD: $\mathbf{w}_t \doteq \widehat{\mathbf{A}}_t^{-1} \widehat{\mathbf{b}}_t$, where $\widehat{\mathbf{A}}_t \doteq \sum_{k=0}^{t-1} \mathbf{x}_k\left(\mathbf{x}_k - \gamma \mathbf{x}_{k+1}\right)^\top + \varepsilon \mathbf{I}$ and $\widehat{\mathbf{b}}_t \doteq \sum_{k=0}^{t-1} R_{k+1} \mathbf{x}_k$
- GTD2: $\mathbf{v}_{t+1} \doteq \mathbf{v}_t + \beta \rho_t\left(\delta_t - \mathbf{v}_t^\top \mathbf{x}_t\right) \mathbf{x}_t, w_{t+1} = \mathbf{w}_t + \alpha \rho_t\left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}\right) \mathbf{x}_t^\top \mathbf{v}_t$ (note: in the GTD2 equation, v is not the value function but a variable storing an intermediate result)

Note: the following methods are given assuming the discount factor $\gamma = 1$.

- Finite difference gradients: $\theta_{k+1} = \theta_k + \alpha \frac{J(\theta_k - \epsilon) - J(\theta_k + \epsilon)}{2\epsilon}$
- original REINFORCE: $\theta_{k+1} = \theta_k + \alpha G(\tau) \sum_{t=0}^T \nabla_\theta \log \pi_\theta(a_t|s_t)$
- REINFORCE v2: $\theta_{k+1} = \theta_k + \alpha \sum_{t=0}^T G_t \nabla_\theta \log \pi_\theta(a_t|s_t)$
- PGT Actor-Critic: $\theta_{t+1} = \theta_t + \alpha \hat{q}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t)$
- Deterministic policy gradient (DPG): $\theta_{t+1} = \theta_t + \alpha \nabla_\theta \pi_\theta(a_t|s_t) \nabla_a q(s_t, a_t)|_{a=\pi_\theta(s)}$
- Natural policy gradient: $\theta_{t+1} = \theta_t + \alpha F^{-1}(\theta) \widehat{\nabla_\theta J(\theta)}$, with $\widehat{\nabla_\theta J(\theta)}$ estimating 'vanilla' policy gradient.