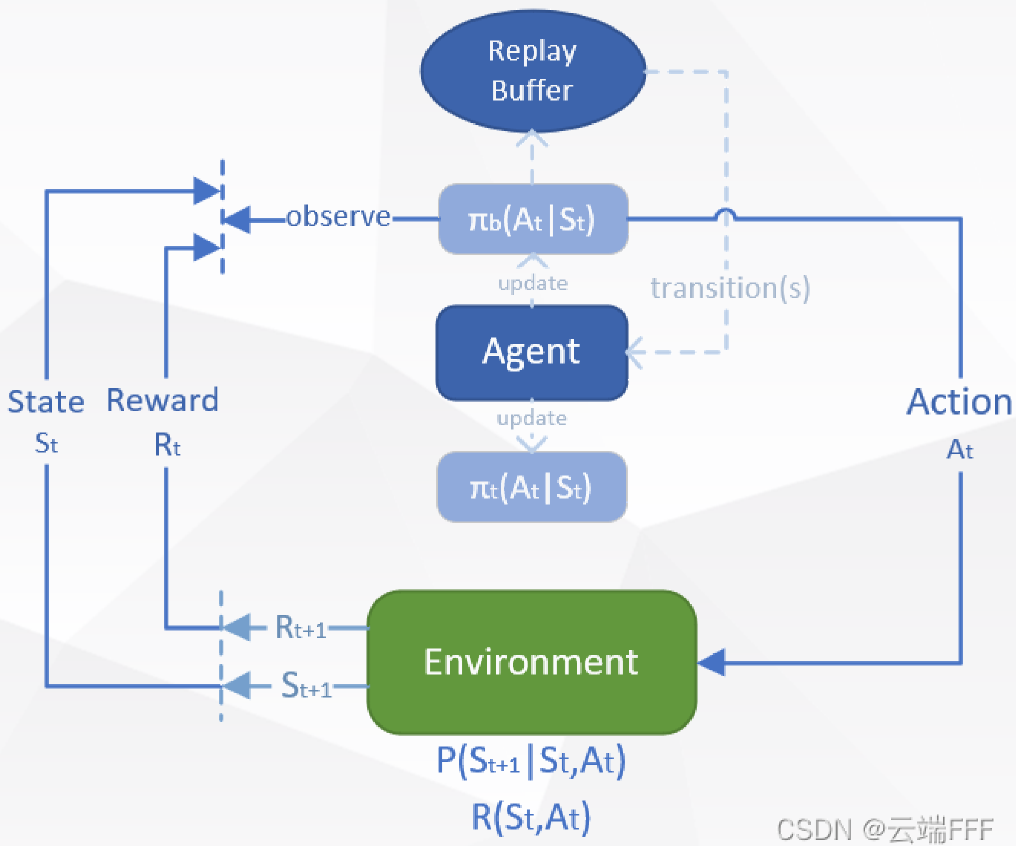
DisCor: Corrective Feedback in Reinforcement Learning via Distribution Correction

Aviral Kumar, Abhishek Gupta, Sergey Levine
Electrical Engineering and Computer Sciences, UC Berkeley
aviralk@berkeley.edu

NIPS 2020

Reinforcement Learning

- Balance between exploration and exploitation
- Agent's action affect the subsequent data it received (actions affects the environment)
- Delayed reward
- Time matters (sequential data, not i.i.d)

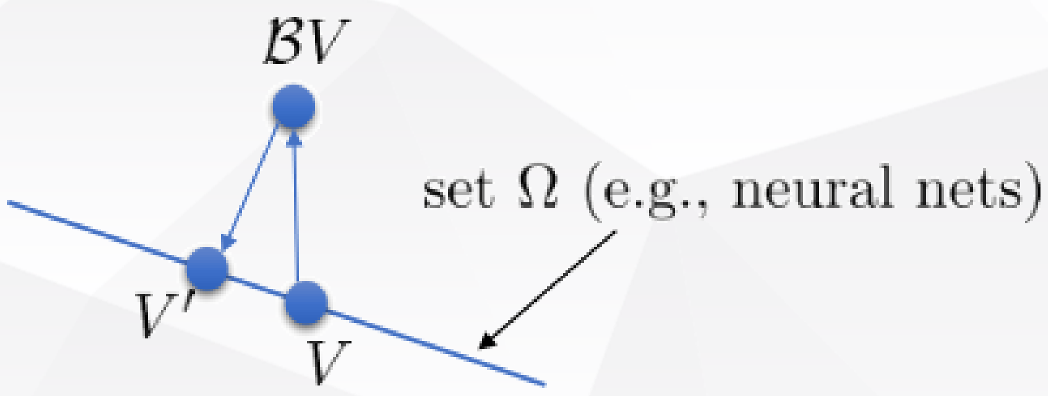


ADP Methods (for prediction)

DP $$(\mathcal{B}^* Q)(s,a) = r(s,a) + \gamma \mathbb{E}_{s'|s,a}\left[\max_{a'} \bar{Q}(s',a')\right]$$

ADP $$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right]$$

$$\theta \leftarrow \arg\min_\theta \mathbb{E}_{s,a\sim\mathcal{D}}\left[(Q_\theta(s,a) - (r(s,a) + \gamma \mathbb{E}_{s'|s,a}[\max_{a'} \bar{Q}(s',a')]))^2\right]$$



Note: TD is a kind of MC

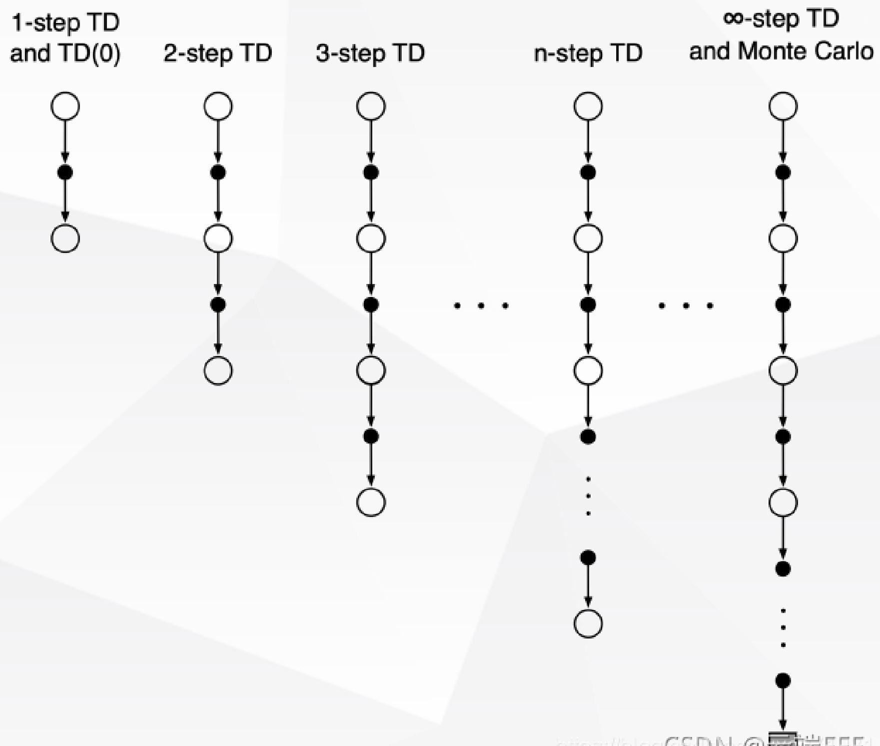
MC

$$v_\pi(s) \leftarrow v_\pi(s) + \frac{1}{N(s)}(g_t - v_\pi(s)) \quad \rightarrow \quad v_\pi(s) = \mathbb{E}_\pi[g_t]$$

$$v_\pi(s) \leftarrow v_\pi(s) + \alpha(g_t - v_\pi(s))$$

TD

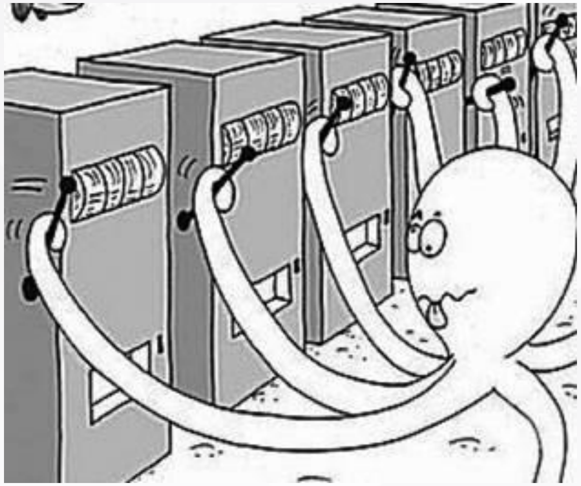
$$v_\pi(s) \leftarrow v_\pi(s) + \alpha[r_{t+1} + \gamma v_\pi(s') - v_\pi(s)] \quad \rightarrow \quad v_\pi(s_t) = \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|s_t]$$



Corrective Feedback

$$\mathcal{L}(Q) = \mathbb{E}_{s \sim \beta(s), a \sim \pi_k(a|s)}[|Q_k(s,a) - Q^*(s,a)|].$$

1. some state value over-estimated
2. policy chooses action correspond to it
3. observes the corresponding r(s,a), or Q*(s,a)
4. minimize $\mathcal{L}(Q)$, which correcte the Q-values precisely



constructive interaction between data collection and error correction

Corrective Feedback is Absent

$$\mathcal{L}(Q) = \mathbb{E}_{s\sim\beta(s),a\sim\pi_k(a|s)}[|Q_k(s,a) - Q^*(s,a)|].$$

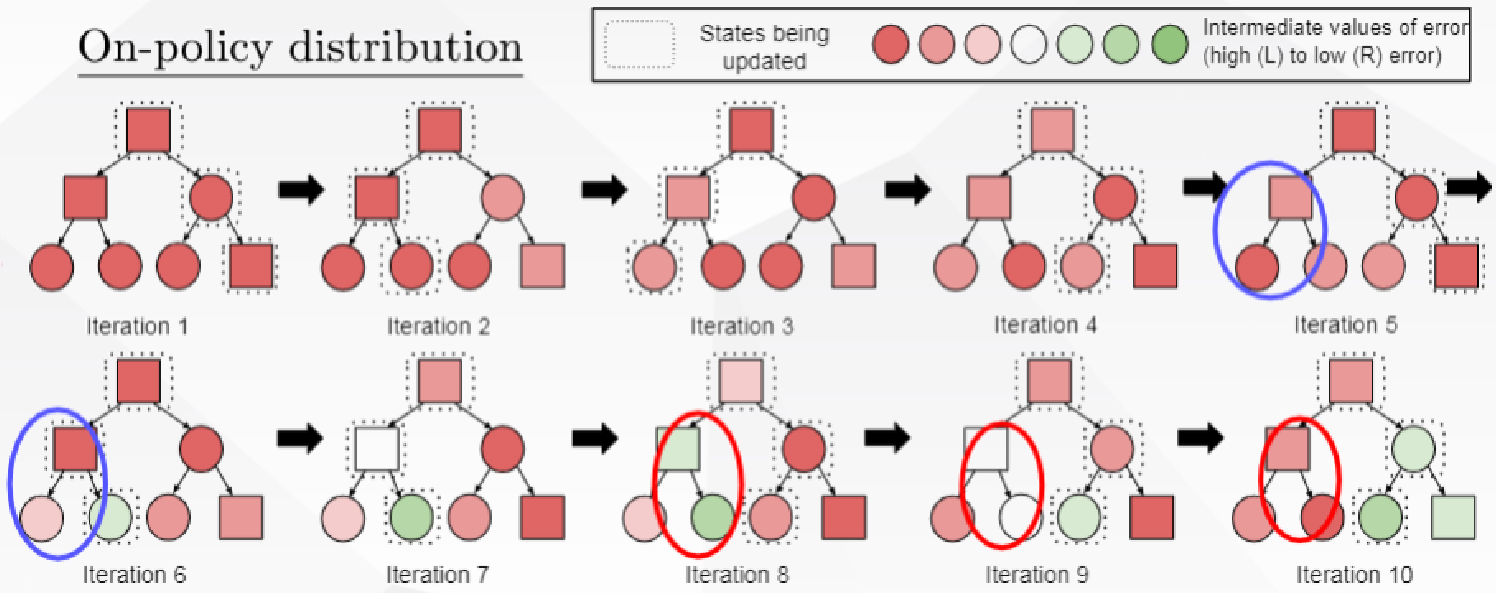
$$\mathcal{L}(Q) = \mathbb{E}_{s\sim\beta(s),a\sim\pi_k(a|s)}[|Q_k(s,a) - \mathcal{B}^*Q_k(s,a)|]$$

can be a wrong target

precise target

function approximator make things worse

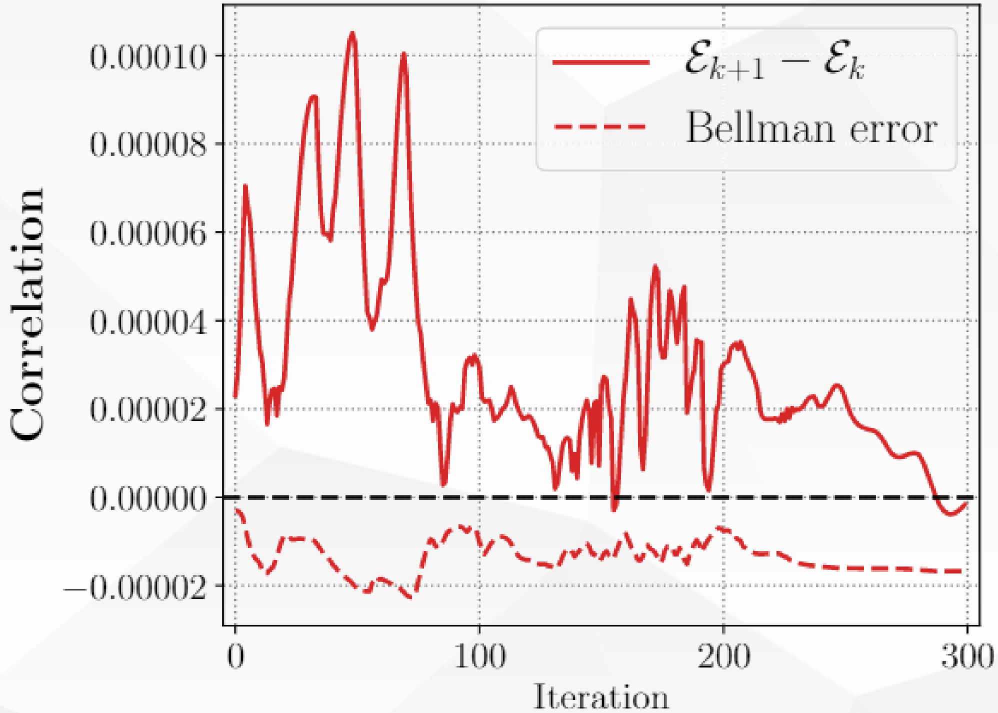
Corrective Feedback is Absent



- leaf state: rarely visit, provide incorrect TD target
- root state: frequently visit, fit to incorrect target
- state with similar features affect each orther

Analyze computationally

Gridworld MDP, training on all transitions to eliminate sampling error



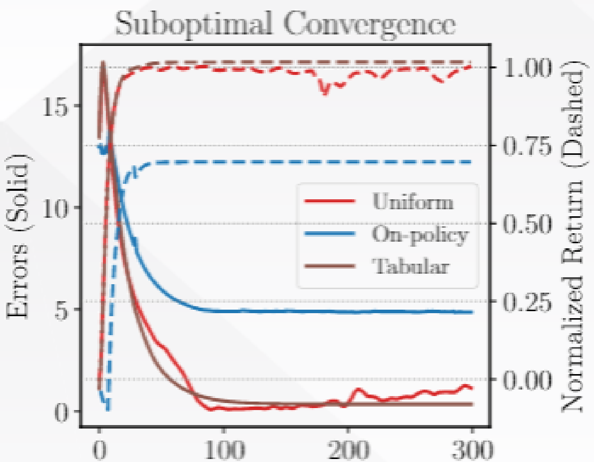
$\mathcal{E}_k = \mathbb{E}_{d^{\pi_k}}\left[|Q_k - Q^*|\right]$

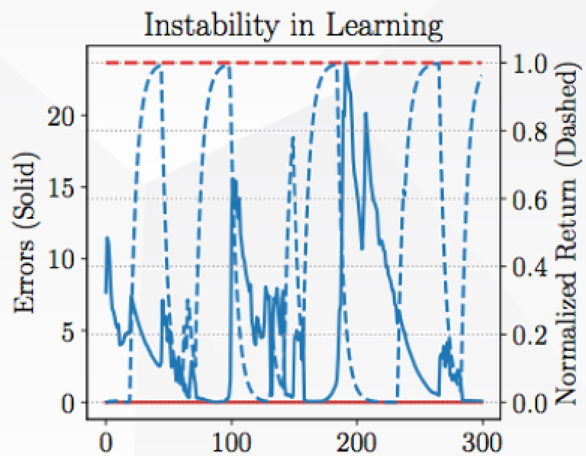
$|Q_{k+1} - \mathcal{B}^* Q_k|(s,a)$

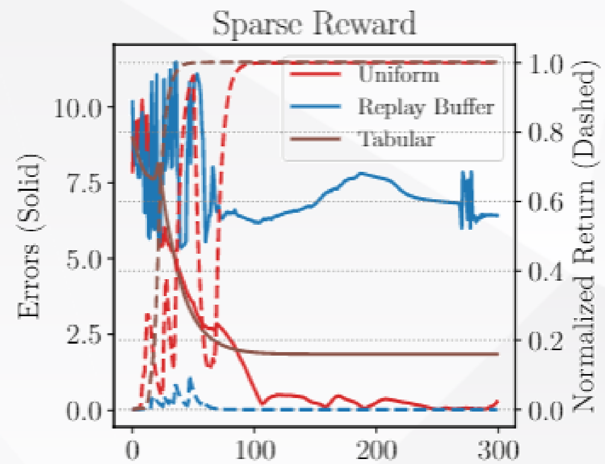
$d^\pi(s) = \sum_{t=0}^{\infty} \gamma^t p(S_t = s|\pi)$

$d^\pi(s,a) = d^\pi(s)\pi(a|s)$

Consequences







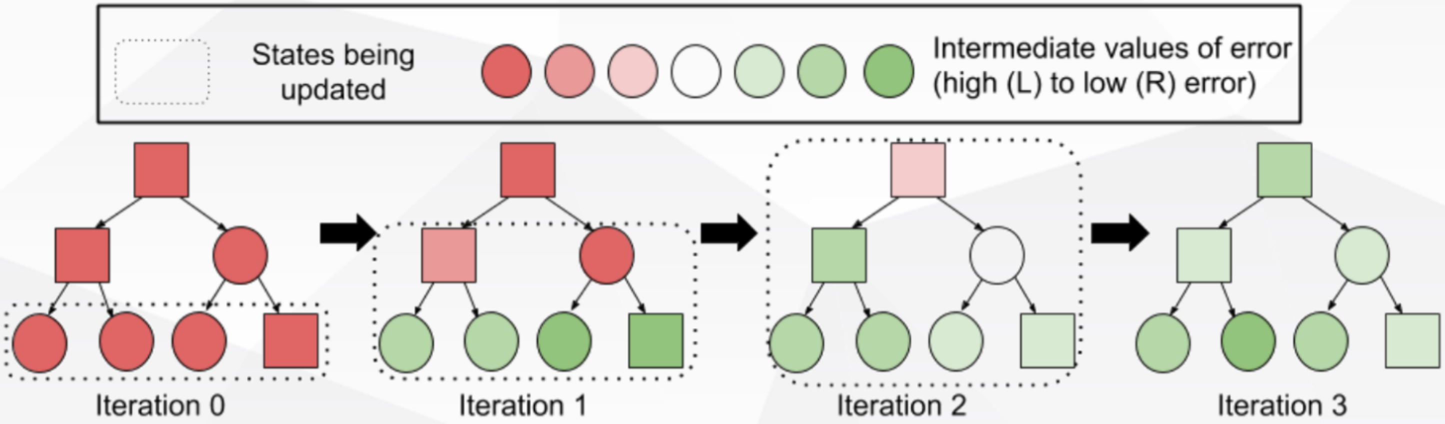
- Convergence to suboptimal Q-functions
- Instability in the learning process
- Inability to learn with low signal-to-noise ratio

Idea

$$\theta \leftarrow \arg\min_{\theta} \mathbb{E}_{s,a\sim\mathcal{D}}\left[\left(Q_\theta(s,a) - (r(s,a) + \gamma \mathbb{E}_{s'|s,a}[\max_{a'} \bar{Q}(s',a')])\right)^2\right]$$

- Computing an "optimal" data distribution that provides maximal corrective feedback, and train Q-functions using this distribution

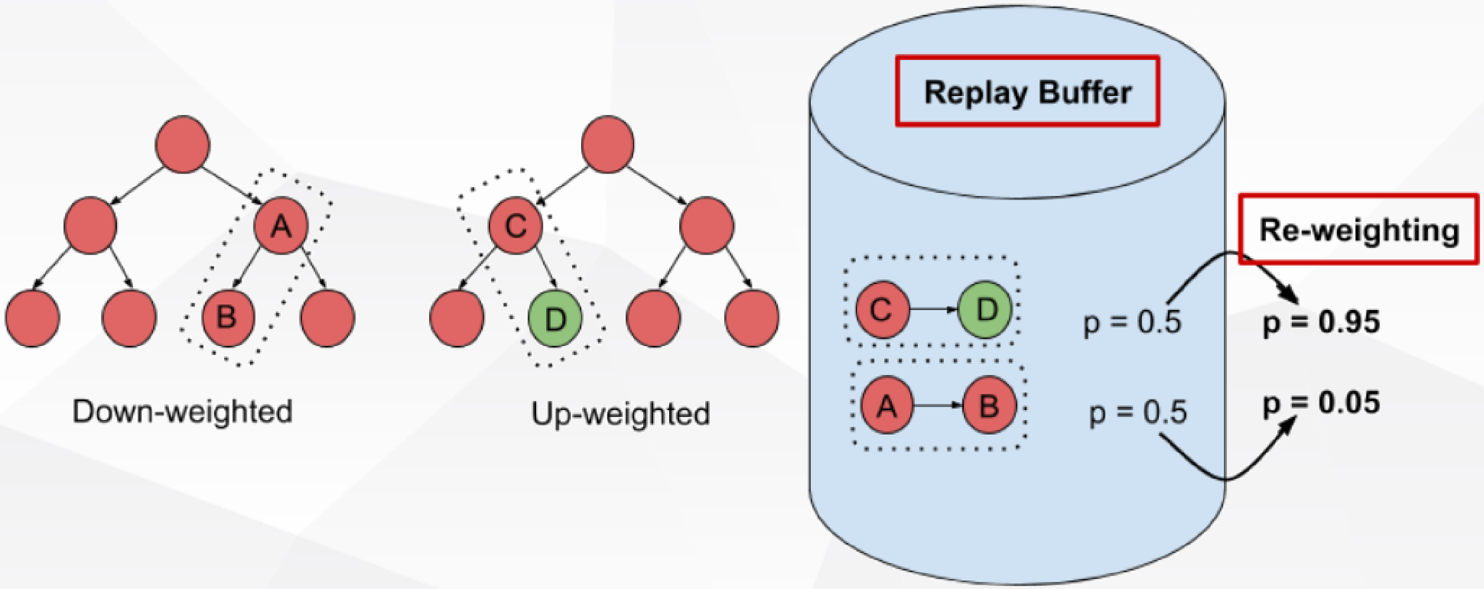
- Once get this optimal distribution, we can then perform a weighted Bellman update that re-weights the data distribution in the replay buffer to this optimal distribution



Idea

$$\theta \leftarrow \arg\min_{\theta} \mathbb{E}_{s,a\sim\mathcal{D}}\left[\left(Q_\theta(s,a) - (r(s,a) + \gamma \mathbb{E}_{s'|s,a}[\max_{a'} \bar{Q}(s',a')])\right)^2\right]$$

$$Q_k \leftarrow \arg\min_{Q} \frac{1}{N}\sum_{i=1}^{N} w_i(s,a)\cdot\left(Q(s,a) - [r(s,a) + \gamma Q_{k-1}(s',a')]\right)^2$$



Formalize the problem

$$\min_{p_k} \mathbb{E}_{d^{\pi_k}} \left[|Q_k - Q^*| \right]$$

$$\text{s.t. } Q_k = \arg\min_Q \mathbb{E}_{p_k} \left[(Q - \mathcal{B}^* Q_{k-1})^2 \right], \quad \sum_{s,a} p_k(s,a) = 1, \quad \forall s,a \;\; p_k(s,a) \geq 0$$

$$p_k(s,a) \propto \exp(-|Q_k - Q^*|(s,a)) \frac{|Q_k - \mathcal{B}^* Q_{k-1}|(s,a)}{\lambda^*}$$

$$\Delta_k(s,a) + \sum_{i=1}^{k} \gamma^{k-i} \alpha_i \geq |Q_k - Q^*|(s,a),$$

$$\alpha_i = \frac{2R_{\max}}{1-\gamma} \mathrm{D}_{\mathrm{TV}}(\pi_i(\cdot \mid s), \pi^*(\cdot \mid s))$$

$$\Delta_k = \sum_{i=1}^{k} \gamma^{k-i} \left(\prod_{j=i}^{k-1} P^{\pi_j} \right) |Q_i - (\mathcal{B}^* Q_{i-1})|, \quad \text{(vector-matrix form)}$$

$$\Longrightarrow \Delta_k(s,a) = |Q_k(s,a) - (\mathcal{B}^* Q_{k-1})(s,a)| + \gamma (P^{\pi_{k-1}} \Delta_{k-1})(s,a).$$

$\forall s, a$ $\quad c_1 \leq |Q_k - \mathcal{B}^* Q_{k-1}|(s,a) \leq c_2$

where $\quad c_1 = \min_{s,a} |Q_{k-1} - \mathcal{B}^* Q_{k-2}|,$

$$c_2 = \max_{s,a} |Q_{k-1} - \mathcal{B}^* Q_{k-2}|$$

Formalize the problem

$$w_k(s,a) = \frac{p_k(s,a)}{\mu(s,a)}$$ → high variance; densities $\mu(s,a)$ are unknown

$$q_k^* = \arg\min_{q_k} -\mathbb{E}_{q_k}[\log p_k] + (\tau)\mathrm{D}_{\mathrm{KL}}(q_k \| \mu)$$

$$\begin{aligned}
\frac{\partial - q_k \log p_k + \tau q_k \log \frac{q_k}{\mu_k}}{\partial q_k} &= -\log p_k + \tau(\log \frac{q_k}{\mu_k} + \frac{\mu_k}{q_k}\frac{1}{\mu_k} q_k) \\
&= -\log p_k + \tau(\log \frac{q_k}{\mu_k} + 1) \\
&\stackrel{\Delta}{=} 0 \\
\Rightarrow \log \frac{q_k^*}{\mu_k} + 1 &= \frac{\log p_k}{\tau} \\
\Rightarrow e\frac{q_k^*}{\mu_k} &= \exp(\frac{\log p_k}{\tau}) \\
\Rightarrow q_k^* &\propto \mu_k \cdot \exp(\frac{\log p_k}{\tau})
\end{aligned}$$

$$q_k^*(s,a) \propto (\mu_k) \cdot \exp\left(\frac{\log p_k(s,a)}{\tau}\right)$$

$$\therefore \frac{q_k^*}{\mu_k} \propto \exp\left(\frac{-|Q_k - Q^*|(s,a)}{\tau}\right) \frac{|Q_k - \mathcal{B}^* Q_{k-1}|(s,a)}{\lambda^*}$$

Formalize the problem

$$\frac{q_k^*}{\mu_k} \propto \exp\left(\frac{-|Q_k - Q^*|(s,a)}{\tau}\right) \frac{|Q_k - \mathcal{B}^* Q_{k-1}|(s,a)}{\lambda^*}$$

$$\Delta_k(s,a) + \sum_{i=1}^{k} \gamma^{k-i} \alpha_i \geq |Q_k - Q^*|(s,a),$$

$$\Delta_k(s,a) = |Q_k(s,a) - (\mathcal{B}^* Q_{k-1})(s,a)| + \gamma \left(P^{\pi_{k-1}} \Delta_{k-1}\right)(s,a).$$

$\forall s, a$ $\quad c_1 \leq |Q_k - \mathcal{B}^* Q_{k-1}|(s,a) \leq c_2$

where $\quad c_1 = \min_{s,a} |Q_{k-1} - \mathcal{B}^* Q_{k-2}|,$

$\quad c_2 = \max_{s,a} |Q_{k-1} - \mathcal{B}^* Q_{k-2}|$

lower bound

$$w_k \propto \exp\left(\frac{-c_2 - \gamma \left[P^{\pi_{k-1}} \Delta_{k-1}\right](s,a)}{\tau}\right) \frac{c_1}{\lambda^*}$$

$$w_k(s,a) \propto \exp\left(-\frac{\gamma \left[P^{\pi_{k-1}} \Delta_{k-1}\right](s,a)}{\tau}\right).$$

Formalize the problem

$$|Q_k - Q^*| \leq \gamma P^{\pi_{k-1}} \Delta_{k-1} + c_2 + \sum_i \gamma^i \alpha_i \quad (30)$$

Using this bound in the expression for w_k, along with the lower bound, $|Q_k - \mathcal{B}^* Q_{k-1}| \geq c_1$, we obtain the following lower bound on weights w_k:

$$w_k \propto \exp\left(\frac{-c_2 - \gamma \left[P^{\pi_{k-1}} \Delta_{k-1}\right](s,a)}{\tau}\right) \frac{c_1}{\lambda^*} \quad (31)$$

Pseudo Code

Algorithm 1 DisCor (Distribution Correction)

Initialize Q-values $Q_\theta(s,a)$, initial distribution $p_0(s,a)$, a replay buffer μ, and an error model $\Delta_\phi(s,a)$.
for step k in $\{1, \dots, \text{N}\}$ **do**
    Collect M samples using π_k, add them to replay buffer μ, sample $\{(s_i, a_i)\}_{i=1}^{N} \sim \mu$
    Evaluate $Q_\theta(s,a)$ and $\Delta_\phi(s,a)$ on samples (s_i, a_i).
    Compute target values for Q and Δ on samples:
    $y_i = r_i + \gamma \max_{a'} Q_{k-1}(s'_i, a')$
    $\hat{a}_i = \arg\max_a Q_{k-1}(s'_i, a)$
    $\hat{\Delta}_i = |Q_\theta(s,a) - y_i| + \gamma \Delta_{k-1}(s'_i, \hat{a}_i)$
    Compute w_k using Equation 7.
    Minimize Bellman error for Q_θ weighted by w_k.
    $\theta_{k+1} \leftarrow \underset{\theta}{\text{argmin}} \frac{1}{N} \sum_i^N w_k(s_i, a_i)(Q_\theta(s_i, a_i) - y_i)^2$
    Minimize ADP error for training ϕ.
    $\phi_{k+1} \leftarrow \underset{\phi}{\text{argmin}} \frac{1}{N} \sum_{i=1}^N (\Delta_\phi(s_i, a_i) - \hat{\Delta}_i)^2$
end for

network output

$$w_k(s,a) \propto \exp\left(-\frac{\gamma \left[P^{\pi_{k-1}} \Delta_{k-1}\right](s,a)}{\tau}\right).$$

Pseudo Code

Algorithm 3 DisCor: Deep RL Version

Initialize online Q-network $Q_\theta(s,a)$, target Q-network, $Q_{\bar{\theta}}(s,a)$, error network $\Delta_\phi(s,a)$, target error network $\Delta_{\bar{\phi}}$, initial distribution $p_0(s,a)$, a replay buffer β and a policy $\pi_\psi(a|s)$, number of gradient steps G, target network update rate η, initial temperature for computing weights w_k, τ_0.

for step k in $\{1, \dots, \}$ **do**

Collect M samples using $\pi_\psi(a|s)$, add them to replay buffer β, sample $\{(s_i, a_i)\}_{i=1}^N \sim \beta$

Evaluate $Q_\theta(s,a)$ and $\Delta_\phi(s,a)$ on samples (s_i, a_i).

Compute target values for Q and Δ on samples:

$$y_i = r_i + \gamma \mathbb{E}_{a' \sim \pi_\psi(a'|s')}[Q_{\bar{\theta}}(s'_i, a')]$$

$$\hat{\Delta}_i = |Q_\theta(s,a) - y_i| + \gamma \mathbb{E}_{\hat{a}_i \sim \pi(a_i|s')}[\Delta_{\bar{\phi}}(s'_i, \hat{a}_i)]$$

Compute w_k using Equation 7 with temperature τ_k

Take G gradient steps on the Bellman error for training Q_θ weighted by w_k.

$$\theta \leftarrow \theta - \alpha \nabla_\theta \frac{1}{N} \sum_{i=1}^{N} w_k(s_i, a_i) \cdot (Q_\theta(s_i, a_i) - y_i)^2$$

Tale G gradient steps to minimize unweighted (regular) Bellman error for training ϕ.

$$\phi \leftarrow \phi - \alpha \nabla_\phi \frac{1}{N} \sum_{i=1}^{N} (\Delta_\theta(s_i, a_i) - \hat{\Delta}_i)^2$$

Update the policy π_ψ if it is explicitly modeled.

$$\psi \leftarrow \psi + \alpha \nabla_\psi \mathbb{E}_{s \sim \beta, a \sim \pi_\psi(a|s)}[Q_\theta(s,a)]$$

Update target networks using soft updates (SAC), hard updates (DQN)

$$\bar{\theta} \leftarrow (1-\eta)\bar{\theta} + \eta\theta$$

$$\bar{\phi} \leftarrow (1-\eta)\bar{\phi} + \eta\phi$$

Update temperature hyperparameter for DisCor:

$$\tau_{k+1} \leftarrow (1-\eta)\tau_k + \eta \ \text{BATCH-MEAN}(\Delta_\phi(s_i, a_i))$$

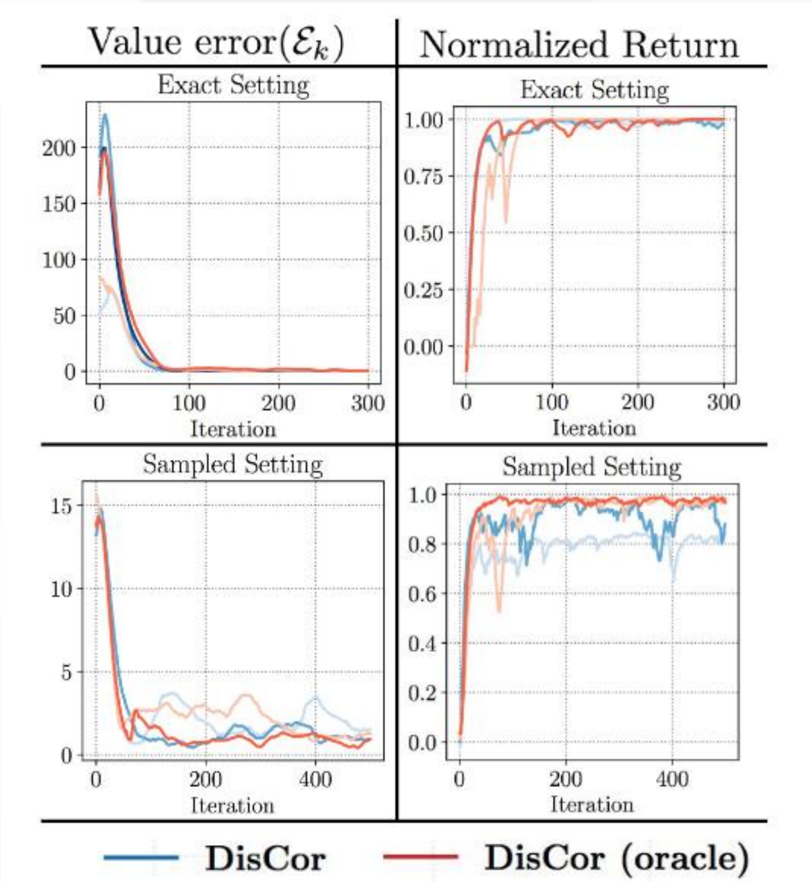
end for

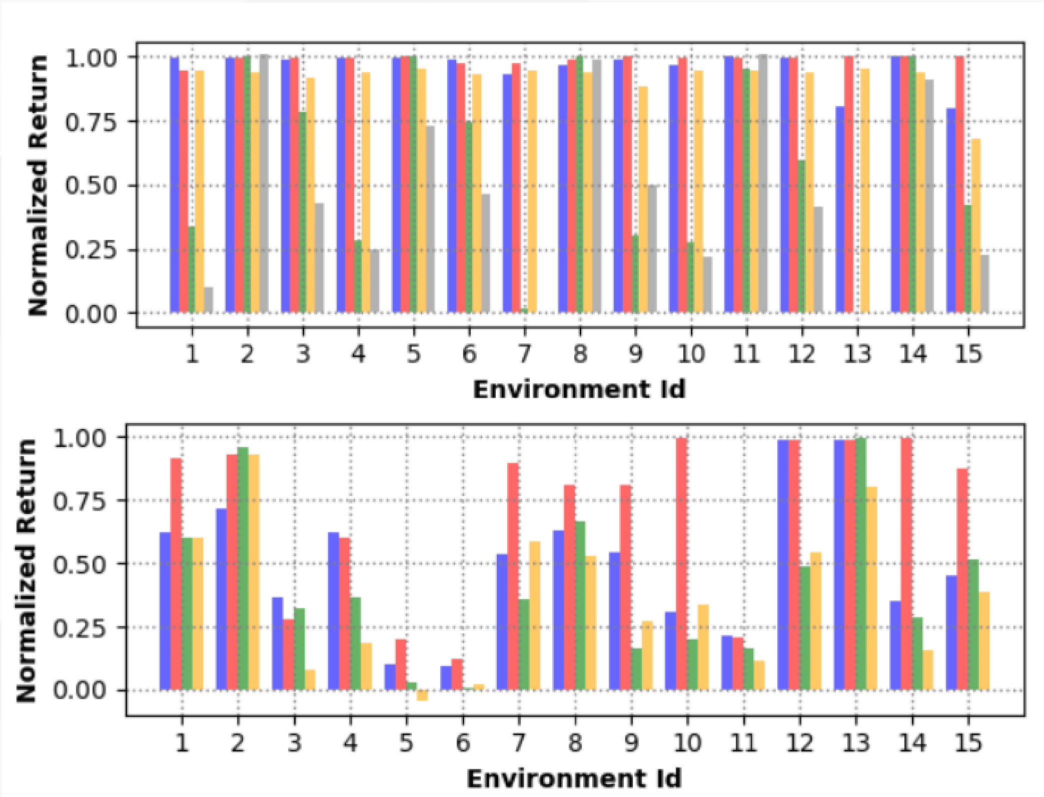
introduce target network

for continuous control domain

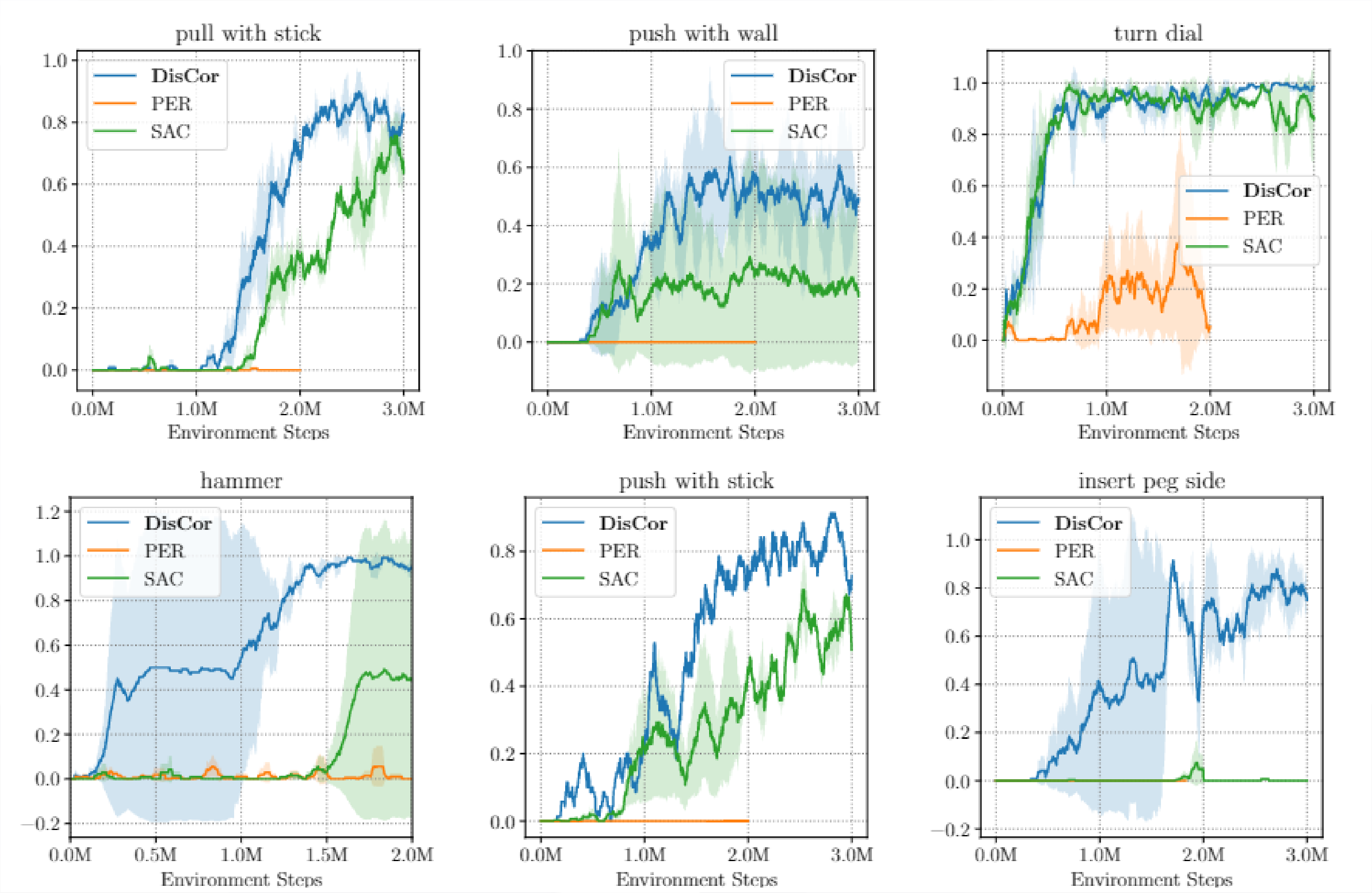
automatically choose temperature

Experiments - grid16

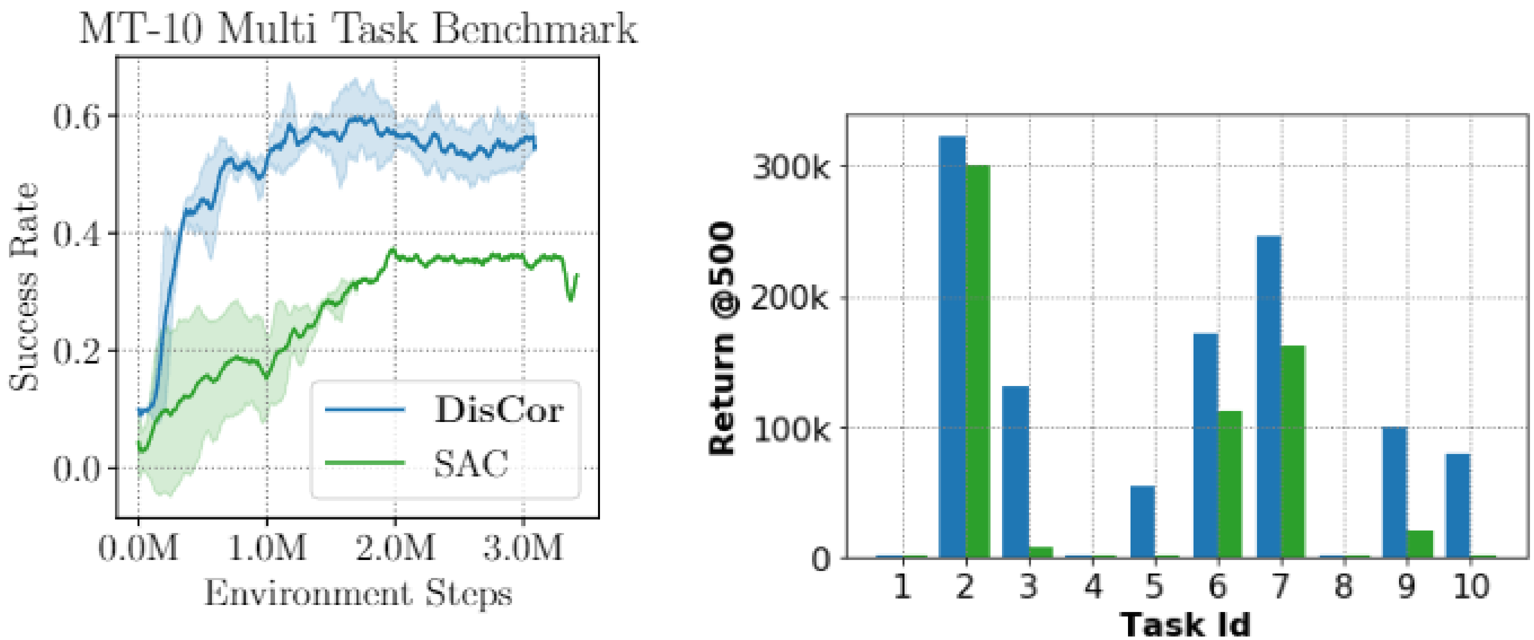




Experiments - MetaWorld

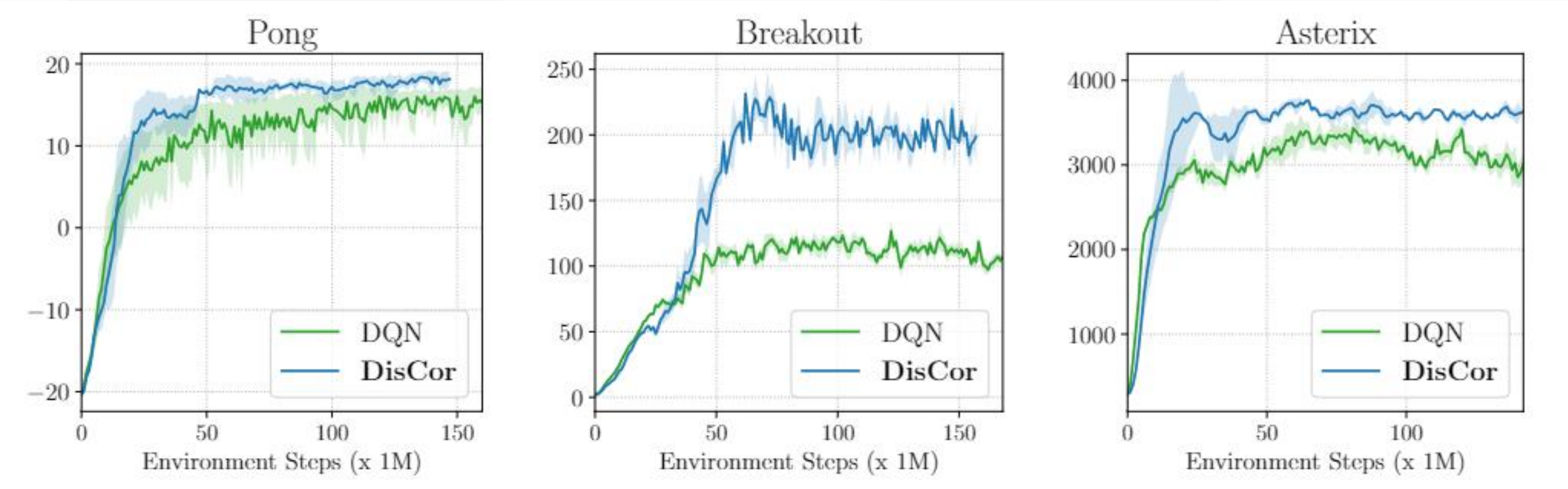


Experiments - MT10



(a) Success rate (b) Per-task return

Experiments - Atari



Proof

$$\min_{p_k} \mathbb{E}_{d^{\pi_k}} \left[|Q_k - Q^*| \right]$$

$$\text{s.t. } Q_k = \arg\min_{Q} \mathbb{E}_{p_k} \left[(Q - \mathcal{B}^* Q_{k-1})^2 \right], \quad \sum_{s,a} p_k(s,a) = 1, \quad \forall s,a \;\; p_k(s,a) \geq 0$$

$$p_k(s,a) \propto \exp\left(-|Q_k - Q^*|(s,a)\right) \frac{|Q_k - \mathcal{B}^* Q_{k-1}|(s,a)}{\lambda^*}$$

1. **引入 Fenchel-Young Inequality**：$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$，对任意凸函数 f 及其 Fenchel 共轭 f^*，有

$$\boldsymbol{x}^\top \boldsymbol{y} \leq f(\boldsymbol{x}) + f^*(\boldsymbol{y})$$

这个不等式是显然的，因为共轭函数的定义就是 $f^*(\boldsymbol{y}) = \sup(\boldsymbol{x}^\top \boldsymbol{y} - f(\boldsymbol{x}))$。注意到优化目标正是 d^{π_k} 和 $|Q_k - Q^*|$ 两个向量内积的形式，所以带入 Fenchel-Young Inequality，得到

$$\mathbb{E}_{d^{\pi_k}} \left[|Q_k - Q^*| \right] \leq f(|Q_k - Q^*|) + f^*(d^{\pi_k}) \tag{9}$$

由于两边都在 $Q_k = Q^*$ 时取得最小值，所以可以用 (9) 中右式的 upper bound 代替 (8) 中的优化目标，求解这个松弛后的优化问题。为了便于处理，f 选择为 $soft-min$ 函数

$$f(x) = -\log\left(\sum_i e^{-x_i}\right), \quad f^*(y) = \mathcal{H}(y) \tag{10}$$

这种选择下 f^* 和香农熵的形式一致，这意味替换 (8) 中优化目标后，我们要同时最小化边际状态动作折扣分布 d^{π_k} 的熵。为了避免优化得到的 p_k 使得 d^{π_k} 的熵大幅下降，作者使用 (s,a) 均匀分布的熵 $\mathcal{H}(\mathcal{U})$ 作为 $\mathcal{H}(y)$ 的 upper bound 代替，这样 f^* 项就变成常数了，最终得到的优化问题为

$$\begin{aligned} &\min_{p_k} -\log\left(\sum_{s,a} \exp(-|Q_k - Q^*|(s,a))\right) \\ &s.t. \;\; Q_k = \arg\min_{Q} \mathbb{E}_{p_k}\left[(Q - \mathcal{B}^* Q_{k-1})^2\right], \quad \sum_{s,a} p_k(s,a) = 1, \quad \forall s,a \;\; p_k(s,a) \geq 0 \end{aligned} \tag{11}$$

Proof

2. **计算拉格朗日函数**：使用拉格朗日乘子法解优化问题 (11)，写出拉格朗日函数

$$\mathcal{L}\left(p_k; \lambda, \mu\right) = -\log\left(\sum_{s,a} \exp(-\left|Q_k - Q^*\right|(s,a))\right) + \lambda\left(\sum_{s,a} p_k(s,a) - 1\right) - \mu^T p_k \tag{12}$$

接下来计算 $\frac{\partial \mathcal{L}}{\partial p_k} = \frac{\partial \mathcal{L}}{\partial Q_k}\frac{\partial Q_k}{\partial p_k}$

3. **使用 implicit function theorem (IFT)**： 考虑如何计算 $\frac{\partial Q_k}{\partial p_k}$，这是两个长 $|\mathcal{S}| \times |\mathcal{A}|$ 向量间求导，最终会得到 $(|\mathcal{S}| \times |\mathcal{A}|) \times (|\mathcal{S}| \times |\mathcal{A}|)$ 的矩阵，注意 Q_k 是一个对应元素相乘的形式，不好用求导公式，根据定义从元素对元素求导的角度出发。这里为了简化运算使用隐函数求导法，先找隐函数，假设 Q_k 满足 $Q_k = \arg\min_Q \mathbb{E}_{p_k}\left[(Q - \mathcal{B}^* Q_{k-1})^2\right]$，则此处对 Q_k 的梯度为零向量（数对向量求导得到等尺寸向量），这就是目标隐函数，即

$$\begin{aligned}
F(p_k, Q_k) &= \left[2p_k(s_0, a_0)[Q_k(s_0, a_0) - \mathcal{B}^* Q_{k-1}(s_0, a_0)] \quad \ldots \quad 2p_k(s_{|\mathcal{S}|}, a_{|\mathcal{A}|})[Q_k(s_{|\mathcal{S}|}, a_{|\mathcal{A}|}) - \mathcal{B}^* Q_{k-1}(s_{|\mathcal{S}|}, a_{|\mathcal{A}|})]\right]^\top \\
&= \mathrm{Diag}(Q_k - \mathcal{B}^* Q_{k-1}) p_k \\
&= \mathrm{Diag}(p_k)(Q_k - \mathcal{B}^* Q_{k-1}) \\
&= \mathbf{0}_{(|\mathcal{S}| \times |\mathcal{A}|) \times 1}
\end{aligned}$$

利用隐函数求导法，有

$$\begin{aligned}
H_Q &= 2\,\mathrm{Diag}(p_k) \quad H_{Q,p_k} = 2\,\mathrm{Diag}(Q_k - \mathcal{B}^* Q_{k-1}) \\
\frac{\partial Q_k}{\partial p_k} &= -[H_Q]^{-1} H_{Q,p_k} = -\mathrm{Diag}\left(\frac{Q_k - \mathcal{B}^* Q_{k-1}}{p_k}\right)
\end{aligned} \tag{14}$$

Proof

4. **计算最优** p_k：令 $\frac{\partial \mathcal{L}(p_k;\lambda,\mu)}{\partial p_k}=\mathbf{0}$ 来求解最优 p_k（本质是对偶问题中的内层极小化问题），这是数对向量求导，还是按定义法从元素对元素求导角度考虑

$$\frac{\partial \mathcal{L}(p_k;\lambda,\mu)}{\partial p_k}=0 \Rightarrow \frac{\operatorname{sgn}(Q_k-Q^*)\exp(-|Q_k-Q^*|(s,a))}{\sum_{s',a'}\exp(-|Q_k-Q^*|(s',a'))}\cdot\frac{\partial Q_k}{\partial p_k}+\lambda-\mu_{s,a}=0 \tag{15}$$

带入上面计算的 $\frac{\partial Q_k}{\partial p_k}$ 得到

$$p_k(s,a)=\frac{\operatorname{sgn}(Q_k-Q^*)\exp(-|Q_k-Q^*|(s,a))}{\sum_{s',a'}\exp(-|Q_k-Q^*|(s',a'))}\cdot\frac{(Q_k-\mathcal{B}^*Q_{k-1})(s,a)}{\mu(s,a)-\lambda}$$

当最优解存在且与原问题一致时，KKT条件成立，有 $\mu^*(s,a)p_k(s,a)=0\ (\forall s,a)$，令所有 (s,a) 都有概率访问到，即 $p_k(s,a)>0$（极值点是约束面的内点），则 $\mu^*(s,a)=0$，且外层最大化解得的 λ^* 要满足 $p_k(s,a)>0$，有

$$p_k(s,a)\propto\exp(-|Q_k-Q^*|(s,a))\frac{|Q_k-\mathcal{B}^*Q_{k-1}|(s,a)}{\lambda^*} \tag{16}$$

Proof

Theorem 4.2. *There exists a $k_0 \in \mathbb{N}$, such that $\forall\, k \geq k_0$ and Δ_k from Equation 5, Δ_k satisfies the following inequality, pointwise, for each s, a, as well as, $\Delta_k \rightarrow |Q_k - Q^*|$ as $\pi_k \rightarrow \pi^*$.*

$$\Delta_k(s,a) + \sum_{i=1}^{k} \gamma^{k-i}\alpha_i \geq |Q_k - Q^*|(s,a), \quad \alpha_i = \frac{2R_{\max}}{1-\gamma} \mathrm{D}_{\mathrm{TV}}(\pi_i(\cdot|s), \pi^*(\cdot|s)).$$

Lemma B.1. *For any $k \in \mathbb{N}$, $|Q_k - Q^*|$ satisfies the following recursive inequality, pointwise for each s, a:*

$$|Q_k - Q^*| \leq |Q_k - \mathcal{B}^* Q_{k-1}| + \gamma P^{\pi_{k-1}} |Q_{k-1} - Q^*| + \frac{2R_{max}}{1-\gamma} \max_s \mathrm{D}_{\mathrm{TV}}(\pi_{k-1}, \pi^*).$$

Proof. Our proof relies on a worst-case expansion of the quantity $|Q_k - Q^*|$. The proof follows the following steps. The first few steps follow common expansions/inequalities operated upon in the work on error propagation in Q-learning [35].

$$\begin{aligned}
|Q_k - Q^*| &\overset{(a)}{=} |Q_k - \mathcal{B}^* Q_{k-1} + \mathcal{B}^* Q_{k-1} - Q^*| \\
&\overset{(b)}{\leq} |Q_k - \mathcal{B}^* Q_{k-1}| + |\mathcal{B}^* Q_{k-1} - \mathcal{B}^* Q^*| \\
&\overset{(c)}{=} |Q_k - \mathcal{B}^* Q_{k-1}| + |R + \gamma P^{\pi_{k-1}} Q_{k-1} - R - \gamma P^{\pi^*} Q^*| \\
&\overset{(d)}{=} |Q_k - \mathcal{B}^* Q_{k-1}| + \gamma |P^{\pi_{k-1}} Q_{k-1} - P^{\pi_{k-1}} Q^* + P^{\pi_{k-1}} Q^* - P^{\pi^*} Q^*| \\
&\overset{(e)}{\leq} |Q_k - \mathcal{B}^* Q_{k-1}| + \gamma P^{\pi_{k-1}} |Q_{k-1} - Q^*| + \gamma |P^{\pi_{k-1}} - P^{\pi^*}||Q^*| \\
&\overset{(f)}{\leq} |Q_k - \mathcal{B}^* Q_{k-1}| + \gamma P^{\pi_{k-1}} |Q_{k-1} - Q^*| + \frac{2R_{\max}}{1-\gamma} \max_s \mathrm{D}_{\mathrm{TV}}(\pi_{k-1}, \pi^*)
\end{aligned}$$

where (a) follows from adding and subtracting $\mathcal{B}^* Q_{k-1}$, (b) follows from an application of triangle inequality, (c) follows from the definition of $\mathcal{B}^*$ applied to two different Q-functions, (d) follows from algebraic manipulation, (e) follows from an application of the triangle inequality, and (f) follows from bounding the maximum difference in transition matrices $|P^{\pi_{k-1}} - P^*|$ by maximum total variation divergence between policy π_{k-1} and π^*, and bounding the maximum possible value of Q^* by $\frac{R_{\max}}{1-\gamma}$.

Proof

Lemma B.2. *For any $k \in \mathbb{N}$, an vector Δ'_k satisfying*

$$\Delta'_k := |Q_k - \mathcal{B}^* Q_{k-1}| + \gamma P^{\pi_{k-1}} \Delta'_{k-1}. \tag{19}$$

with $\alpha_k = \frac{2R_{\max}}{1-\gamma} \max_s \mathrm{D_{TV}}(\pi_k, \pi^*)$, and an initialization $\Delta'_0 := |Q_0 - Q^*|$, pointwise upper bounds $|Q_k - Q^*|$ with an offset depending on α_i, i.e. $\Delta'_k + \sum_i \alpha_i \gamma^{k-i} \geq |Q_k - Q^*|$.

Proof. Let Δ'_k be an estimator satisfying Equation 19. In order to show that $\Delta'_k + \sum_i \gamma^{k-i} \alpha_i \geq |Q_k - Q^*|$, we use the principle of mathematical induction. The base case, $k = 0$ is satisfied, since $\Delta'_0 + \alpha_0 \geq |Q_0 - Q^*|$. Now, let us assume that for a given $k = m$, $\Delta'_m + \sum_i \gamma^{m-i} \alpha_i \geq |Q_m - Q^*|$ pointwise for each (s, a). Now, we need to show that a similar relation holds for $k = m + 1$, and then we can appeal to the principle of mathematical induction to complete the argument. In order to show this, we note that,

$$\begin{aligned}
\Delta'_{m+1} =& |Q_{m+1} - \mathcal{B}^* Q_m| + \gamma P^{\pi_m} \Delta'_m + \sum_i^{m+1} \gamma^{m+1-i} \alpha_i && (20)\\
=& |Q_{m+1} - \mathcal{B}^* Q_m| + \gamma P^{\pi_m} (\Delta'_m + \sum_{i=0}^{m} \gamma^{m-i} \alpha_i) + \alpha_{m+1} && (21)\\
\geq& |Q_{m+1} - \mathcal{B}^* Q_m| + \gamma P^{\pi_m} |Q_m - Q^*| + \alpha_m && (22)\\
\geq& |Q_{m+1} - Q^*| && (23)
\end{aligned}$$

where (20) follows from the definition of Δ'_k, (21) follows by rearranging the recursive sum containing α_i, for $i \leq m$ alongside Δ_m, (22) follows from the inductive hypothesis at $k = m$, and (23) follows from Lemma B.1.

The End

thanks