

باقیه به اندک آپدیت به صورت $2e2y$ و تا حاصل انجام داشته، شکل 5 را داریم؛ پس این

فرآیند به عنوان $importance sampling$ محال می کنند.

$$w_j = \frac{P(\tau) \cdot \exp(r_\psi(\tau_j))}{\pi(\tau_j)} = \frac{P(s_1) \prod_{t=1}^T P(s_{t+1} | s_t, a_t) \cdot \exp(r_\psi(s_t, a_{t+1}))}{P(s_1) \cdot \prod_{t=1}^T P(s_{t+1} | s_t, a_t) \cdot \pi(a_t | s_t)}$$

اقتل این الیمنت در توزیع صحیح

توزیع π که در RL به دست می آید

ψ که در RL به دست می آید

توزیع π که در RL به دست می آید

الف) ادلا در offine RL زحمت اصلاح اشتباه وجود ندارد. اگر فکر کنیم که یک action با بیش از یک پاداش داده و فرقی

اصلاح آن را نداشته باشیم، حواصه آن را انقلب می کنیم و مقدار $Q(s, a)$ اصلاح نمی شود. همین فرایند

آموزشی را باید $Q(s, a) \leftarrow r(s, a) + \max_{a'} Q(s, a')$ باعث می شود که \max گرفتن مشابه $adv attack$

محال کند و به بیش برآورد ما دامن نزنند.

$$\hat{Q}_\pi = \arg \min_Q \max_P \underbrace{\alpha \mathbb{E}_{s \sim D, a \sim p(a|s)} [Q(s, a)]}_{\text{پایه فشرده فواید}} - \underbrace{\alpha \mathbb{E}_{(s, a) \sim D} [Q(s, a)]}_{\text{احتمال بیشتر به فواید موجود در دیتا}} - \underbrace{\mathbb{E}_{s \sim D} [\mathcal{H}(\pi(s))]}_{\text{Regularization}} \quad (ب)$$

$$+ \mathbb{E}_{(s, a, a') \sim D} [Q(s, a) - (r(s, a) + \mathbb{E}[Q(s, a')])^2]$$

عبارت همیشه باید به

مجازات مدل برای رفتن به جهات نادر دیتا نیست!

$$\hat{Q}^{k+1} \leftarrow \arg \min_Q \beta \left(\mathbb{E}_{s, a \sim p(s, a)} [Q(s, a)] - \mathbb{E}_{s, a \sim D} [Q(s, a)] \right)$$

$$+ \frac{1}{2} \mathbb{E}_{s, a, s' \sim d_f} \left[\left(Q(s, a) - \hat{B}^\pi \hat{Q}^k(s, a) \right)^2 \right]. \quad (4)$$

عبارت عمل بهینه سازی

د) بستن به محدودیت تعامل با محیط دارد. در صورت زیاده‌بودن افغان تعامل، کشور به نظم منظم

و رسد. در کل در صورت وجود افغان تعامل محدود همچنان استفاده حفاظت منظم بنظر می‌رسد، چرا که

تعامل محدود نتایج شگفت‌انگیز یا *affine* به طریقی که در صورت *distribution mismatch* زیاد باشد،

با تعامل آنلاین پس از یادگیری آنلاین، به راقی بر می‌گردد.