Discounting-aware Importance Sampling

1 Giới thiệu

Trong học tăng cường ngoài chính sách (off-policy reinforcement learning), ta thường sử dụng **importance sampling** để điều chỉnh sự khác biệt giữa chính sách hành động mà ta muốn đánh giá (π) và chính sách hành động sinh ra dữ liệu (b). Tuy nhiên, phương pháp importance sampling tiêu chuẩn có thể gây ra phương sai cao khi tập dữ liệu gồm các tập trải nghiệm dài (T bước). Bài viết này phân tích cách tiếp cận **Discounting-aware Importance Sampling**, giúp giảm phương sai bằng cách xem xét cấu trúc giảm giá của phần thưởng.

2 Phần 5.8 - Discounting-aware Importance Sampling

2.1 Công thức: $G_0 = R_1$

Ý nghĩa:

- Khi hệ số giảm giá $\gamma = 0$, phần thưởng tổng hợp G_0 chỉ là phần thưởng tức thời tại bước đầu tiên R_1.
- Vì $\gamma = 0$ loại bỏ ảnh hưởng của các phần thưởng tương lai, G_0 không phụ thuộc vào $R_2, R_3, \ldots$.

Triển khai:

- Công thức này đơn giản hóa tính toán phần thưởng tổng hợp trong một kịch bản lý thuyết cực đoan.
- Trong học tăng cường off-policy, điều này giúp xác định các yếu tố quan trọng thực sự ảnh hưởng đến giá trị G_0.

2.2 Công thức: Tỷ lệ Importance Sampling

$$\prod_{t=0}^{99} \frac{\pi(A_t|S_t)}{b(A_t|S_t)} \tag{1}$$

Ý nghĩa:

- Đây là tích của 100 yếu tố, mỗi yếu tố là tỷ lệ xác suất giữa chính sách mục tiêu π và chính sách hành vi b tại mỗi bước từ $t = 0$ đến $t = 99$.
- Khi $\gamma = 0$, $G_0 = R_1$ chỉ phụ thuộc vào A_0 và S_0, nên chỉ cần nhân với hệ số đầu tiên $\frac{\pi(A_0|S_0)}{b(A_0|S_0)}$.

Triển khai:

- Văn bản đề xuất chỉ cần nhân G_0 với $\frac{\pi(A_0|S_0)}{b(A_0|S_0)}$ để giữ nguyên giá trị kỳ vọng nhưng giảm đáng kể phương sai.

3 Flat Partial Returns và Phân tích G_t

3.1 Công thức: Flat Partial Return

$$\bar{G}_{t:h} = \sum_{k=t+1}^{h} R_k, \quad 0 \leq t < h \leq T \tag{2}$$

3.2 Công thức: Conventional Full Return

$$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k \tag{3}$$

3.3 Công thức: Phân tích G_t theo Flat Partial Returns

$$G_t = (1-\gamma) \sum_{h=t+1}^{T} \gamma^{h-t-1} \bar{G}_{t:h} \tag{4}$$

4 Các bộ ước lượng Discounting-aware Importance Sampling

4.1 Công thức: Ordinary Importance-Sampling Estimator

$$V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_{t,h-1} \tilde{G}_{t:h} + \gamma^{T(t)-t-1} \rho_{t,T(t)-1} \tilde{G}_{t:T(t)} \right)}{|\mathcal{T}(s)|} \tag{5}$$

4.2 Công thức: Weighted Importance-Sampling Estimator

$$V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_{t,h-1} \tilde{G}_{t:h} + \gamma^{T(t)-t-1} \rho_{t,T(t)-1} \tilde{G}_{t:T(t)} \right)}{\sum_{t \in \mathcal{T}(s)} \left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_{t,h-1} + \gamma^{T(t)-t-1} \rho_{t,T(t)-1} \right)} \tag{6}$$