Discounting-aware Importance Sampling

1 Giải Thích Các Công Thức Toán Học và Phân Tích Biểu Thức

Các công thức liên quan đến phương pháp Monte Carlo, đặc biệt là importance sampling trong học tăng cường (reinforcement learning). Nội dung được chia thành hai phần chính: phần trên (công thức 5.9 và 5.10) và phần dưới (per-decision importance sampling với công thức 5.11 đến 5.15, cùng bài tập 5.13 và 5.14).

1.1 Công Thức (5.9)

$$V(s) = \frac{\sum_{i \in T(s)} (1-\gamma) \sum_{t=0}^{T(s)-1} \gamma^t R_{t+1} + \gamma^{T(s)} G_{T(s)}}{\sum_{i \in T(s)} (1-\gamma)} \tag{1}$$

- Công thức này tính giá trị mong đợi của trạng thái s dựa trên lợi tức chiết khấu, chuẩn hóa theo tần suất xuất hiện của trạng thái s.

1.2 Công Thức (5.10)

$$V(s) = \frac{\sum_{i \in T(s)} (1-\gamma) \sum_{t=0}^{T(s)-1} \gamma^t \rho_{t,T(s)-1} R_{t+1} + \gamma^{T(s)} \rho_{T(s)-1} G_{T(s)}}{\sum_{i \in T(s)} (1-\gamma)} \tag{2}$$

- Công thức này cải tiến (5.9) bằng cách sử dụng importance sampling với tỷ lệ ρ để điều chỉnh sự khác biệt giữa chính sách mục tiêu và chính sách hành vi.

2 Per-Decision Importance Sampling

2.1 Công Thức (5.15)

$$V(s) = \frac{\sum_{i \in T(s)} \tilde{G}_i}{|T(s)|} \tag{3}$$

- Công thức này ước lượng $V(s)$ bằng trung bình lợi tức không điều chỉnh.

3 Bài Tập 5.14: Sửa Đổi Thuật Toán Monte Carlo Control Off-Policy

3.1 Đề Bài

Sửa đổi thuật toán Monte Carlo control off-policy để sử dụng estimator (5.10), chuyển sang giá trị hành động $Q(s, a)$.

3.2 Giải Pháp

1. **Hiểu Thuật Toán Gốc**: - Học $Q(s, a)$ từ các episode sinh bởi μ, đánh giá π. - Thông thường, cập nhật $Q(s, a)$ bằng cách trung bình lợi tức điều chỉnh bởi $\rho_{t:T-1}$.

2. **Sử Dụng Estimator (5.10) Cho $Q(s, a)$**:

$$Q(s,a) = \frac{\sum_{i \in T(s,a)} (1-\gamma) \sum_{t=0}^{T(s,a)-1} \gamma^t \rho_{t,T(s,a)-1} R_{t+1} + \gamma^{T(s,a)} \rho_{T(s,a)-1} G_{T(s,a)}}{\sum_{i \in T(s,a)} (1-\gamma)} \tag{4}$$

3. **Thuật Toán Sửa Đổi**: - Khởi tạo $Q(s, a) \leftarrow 0$, $N(s, a) \leftarrow 0$. - Với mỗi episode: - Sinh episode bằng μ: $(S_0, A_0, R_1, \ldots, S_T)$. - Tính lợi tức điều chỉnh bằng ρ. - Cập nhật giá trị $Q(s, a)$ dựa trên estimator (5.10).