Xét một quá trình quyết định Markov (MDP) với chân trời vô hạn có chiết khấu (infinite-horizon discounted Markov decision process), được định nghĩa bởi bộ tuple (S,A,P,r,ρ0​,γ).

Trong đó S là tập hợp hữu hạn các trạng thái, A là tập hợp hữu hạn các hành động.

P:S×A×S→R là phân phối xác suất chuyển trạng thái, r:S→R là hàm phần thưởng, ρ0​:S→R là phân phối của trạng thái ban đầu s0​, và γ∈(0,1) là hệ số chiết khấu.

Ký hiệu π:S×A→[0,1] là một chính sách ngẫu nhiên (stochastic policy), và η(π) ký hiệu phần thưởng chiết khấu kỳ vọng của nó:

η(π)=Es0​,a0​,...​[t=0∑∞​γtr(st​)]

trong đó s0​∼ρ0​(s0​), at​∼π(at​∣st​). st+1​∼P(st+1​∣st​,at​).

Chúng ta sẽ sử dụng các định nghĩa chuẩn sau về hàm giá trị trạng thái-hành động Qπ​, hàm giá trị Vπ​, và hàm lợi thế Aπ​:

Qπ​(st​,at​)=Est+1​,at+1​,...​[l=0∑∞​γlr(st+l​)],

Vπ​(st​)=Eat​,st+1​,...​[l=0∑∞​γlr(st+l​)]

Aπ​(s,a)=Qπ​(s,a)−Vπ​(s),

trong đó at​∼π(at​∣st​),st+1​∼P(st+1​∣st​,at​) với t≥0.

Đồng nhất thức hữu ích sau đây biểu diễn phần thưởng kỳ vọng của một chính sách khác π~ theo hàm lợi thế so với π, được tích lũy qua các bước thời gian (xem Kakade & Langford (2002) hoặc Phụ lục A để chứng minh):

η(π~)=η(π)+Es0​,a0​,⋅⋅⋅∼π~​[t=0∑∞​γtAπ​(st​,at​)](1)

trong đó ký hiệu Es0​,a0​,⋅⋅⋅∼π~​[...] chỉ ra rằng các hành động được lấy mẫu at​∼π~(⋅∣st​).

Đặt ρπ​ là tần suất ghé thăm (chưa chuẩn hóa) có chiết khấu:

ρπ​(s)=P(s0​=s)+γP(s1​=s)+γ2P(s2​=s)+...,

trong đó s0​∼ρ0​ và các hành động được chọn theo π.

Chúng ta có thể viết lại Phương trình (1) với tổng theo các trạng thái thay vì các bước thời gian:

η(π~)=η(π)+t=0∑∞​s∑​P(st​=s∣π~)a∑​π~(a∣s)γtAπ​(s,a)=η(π)+s∑​t=0∑∞​γtP(st​=s∣π~)a∑​π~(a∣s)Aπ​(s,a)=η(π)+s∑​ρπ~​(s)a∑​π~(a∣s)Aπ​(s,a)(2)

Phương trình này ngụ ý rằng bất kỳ cập nhật chính sách nào π→π~ mà có lợi thế kỳ vọng không âm tại mọi trạng thái s, tức là ∑a​π~(a∣s)Aπ​(s,a)≥0, được đảm bảo sẽ tăng hiệu suất chính sách η, hoặc giữ nguyên trong trường hợp lợi thế kỳ vọng bằng không ở mọi nơi.

Điều này ngụ ý kết quả cổ điển rằng cập nhật được thực hiện bởi lặp chính sách chính xác (exact policy iteration), sử dụng chính sách tất định (deterministic policy) π~(s)=argmaxa​Aπ​(s,a), sẽ cải thiện chính sách nếu có ít nhất một cặp trạng thái-hành động với giá trị lợi thế dương và xác suất thăm trạng thái khác không, nếu không thì thuật toán đã hội tụ về chính sách tối ưu.

Tuy nhiên, trong bối cảnh xấp xỉ, việc có một số trạng thái s mà lợi thế kỳ vọng âm, tức là ∑a​π~(a∣s)Aπ​(s,a)<0, thường là không thể tránh khỏi do lỗi ước lượng và xấp xỉ.

Sự phụ thuộc phức tạp của ρπ~​(s) vào π~ khiến Phương trình (2) khó tối ưu trực tiếp.

Thay vào đó, chúng ta đưa ra xấp xỉ cục bộ sau đây cho η:

Lπ​(π~)=η(π)+s∑​ρπ​(s)a∑​π~(a∣s)Aπ​(s,a).(3)

Lưu ý rằng Lπ​ sử dụng tần suất ghé thăm ρπ​ thay vì ρπ~​, bỏ qua sự thay đổi mật độ ghé thăm trạng thái do thay đổi chính sách.

Tuy nhiên, nếu chúng ta có một chính sách tham số hóa πθ​, trong đó πθ​(a∣s) là một hàm khả vi của vector tham số θ, thì Lπ​ khớp với η đến bậc nhất (xem Kakade & Langford (2002)).

Tức là, với bất kỳ giá trị tham số θ0​,

Lπθ0​​​(πθ0​​)=η(πθ0​​).

∇θ​Lπθ0​​​(πθ​)∣θ=θ0​​=∇θ​η(πθ​)∣θ=θ0​​(4)

Phương trình (4) ngụ ý rằng một bước đủ nhỏ πθ0​​→π~ mà cải thiện Lπθold​​​ cũng sẽ cải thiện η, nhưng không đưa ra bất kỳ hướng dẫn nào về độ lớn của bước cần thực hiện.

Để giải quyết vấn đề này, Kakade & Langford (2002) đã đề xuất một sơ đồ cập nhật chính sách gọi là conservative policy iteration (lặp chính sách bảo thủ), mà họ có thể cung cấp các cận dưới rõ ràng về sự cải thiện của η.

Để định nghĩa cập nhật lặp chính sách bảo thủ, đặt πold​ là chính sách hiện tại, và đặt π′=argmaxπ′​Lπold​​(π′).

Chính sách mới πnew​ được định nghĩa là hỗn hợp sau:

πnew​(a∣s)=(1−α)πold​(a∣s)+απ′(a∣s).(5)

Kakade và Langford đã đưa ra cận dưới sau đây:

η(πnew​)≥Lπold​​(πnew​)−(1−γ)22ϵγ​α2(6)

trong đó ϵ=maxs​∣Ea∼π′(a∣s)​[Aπ​(s,a)]∣.

(Chúng tôi đã sửa đổi nó để làm cho nó yếu hơn một chút nhưng đơn giản hơn.) Lưu ý, tuy nhiên, cho đến nay ràng buộc này chỉ áp dụng cho các chính sách hỗn hợp được tạo ra bởi Phương trình (5).

Lớp chính sách này cồng kềnh và hạn chế trong thực tế, và mong muốn có một sơ đồ cập nhật chính sách thực tế có thể áp dụng cho tất cả các lớp chính sách ngẫu nhiên chung.