Exercise 2 explanation

2025-03-12

Cambio de Política Óptima con Diferentes Valores de γ

Comparación Detallada de Políticas

Política con γ = 0.95

Características de la Política:

```
+-+-+-+-+
|u|l|d|d|
+-+-+-+-+
|u|x|r|x|
+-+-+-+-+
|l|l|u|x|
+-+-+-+-+
|x|d|l|x|
+-+-+-+-+
```

Política con γ ≈ 1 (0.9999999999)

Características de la Política:

```
+-+-+-+-+
|r|r|l|d|
+-+-+-+-+
|d|x|u|x|
+-+-+-+-+
|l|l|r|x|
+-+-+-+-+
|x|d|d|x|
+-+-+-+-+
```

Análisis Comparativo

Cambios Observados en la Política

1. **Direccionalidad de Movimientos**
    - **γ = 0.95**:
        - Movimientos más directos y localmente optimizados
        - Mayor variedad de direcciones (up, left, down)
        - Parece priorizar rutas más cortas
    - **γ ≈ 1**:
        - Movimientos más consistentes y uniformes
        - Mayor uso de movimientos right (r) y down (d)

        - Sugiere una estrategia más global y cautelosa
2. **Estrategia de Navegación**
    - **Con γ = 0.95**:
        - El agente parece “impaciente”
        - Busca soluciones rápidas
        - Acepta mayor variabilidad en la ruta
        - Prioriza recompensas inmediatas
    - **Con γ ≈ 1**:
        - El agente muestra mayor “previsión”
        - Busca rutas más estables y predecibles
        - Minimiza el riesgo de caer en estados negativos
        - Optimiza la recompensa total a largo plazo

Interpretación Teórica

El cambio de política se explica por la forma en que el factor de descuento (γ) modifica la valoración de recompensas futuras:

- **γ = 0.95**:
    - Descuenta agresivamente recompensas futuras
    - Las penalizaciones por paso tienen un impacto más significativo
    - El agente se comporta más “miope”, buscando soluciones inmediatas
- **γ ≈ 1**:
    - Casi no descuenta recompensas futuras
    - Permite una visión más amplia del espacio de estados
    - El agente se vuelve más “precavido” y estratégico

Considerando la ecuación de Bellman:

$$V^{\pi}(s) = r(s) + \gamma \sum_{s'} P\left(s'|s, \pi(s)\right) \cdot V^{\pi}(s')$$

- Con γ = 0.95, las recompensas futuras se reducen rápidamente
    - Después de 20 pasos: 0.95^20 ≈ 0.36 del valor original
- Con γ ≈ 1, las recompensas futuras mantienen casi todo su valor
    - Después de 20 pasos: Prácticamente sin descuento

Conclusión

El cambio de política ilustra cómo el factor de descuento (γ) determina el horizonte de planificación del agente: - Un γ bajo produce estrategias más cortoplacistas - Un γ cercano a 1 genera estrategias más conservadoras y globales