در یادگیری تقویتی (Reinforcement Learning)، مجموعه‌ای از نمادها و تعاریف پایه‌ای وجود دارد که برای درک الگوریتم‌ها و معادلاتی مانند معادله بلمن باید با آن‌ها آشنا باشیم. در ادامه مهم‌ترین نمادها را همراه با توضیح و مثال‌های ساده به زبان فارسی مرور می‌کنیم.

---

۱. محیط و عامل (Agent & Environment)

- **حالت (State) – s**
  نشان‌دهنده وضعیت فعلیِ محیط است.
  مثال: در بازی شطرنج، یک «حالت» ممکن است آرایش مهره‌ها روی صفحه باشد.
- **اقدام (Action) – a**
  عملی که عامل می‌تواند در حالت s انجام دهد.
  مثال: در همان بازی، حرکت قلعه از خانه e1 به g1 یک اقدام است.
- **سیاست (Policy) – $\pi(a \mid s)$**
  تابعی که احتمال انجام اقدام a را در حالت s مشخص می‌کند.
- **سیاست قطعی (Deterministic):** $\pi(s) = a$
- **سیاست تصادفی (Stochastic):** $\pi(a \mid s) = P[A = a \mid S = s]$

۲. پویایی‌های محیط (Environment Dynamics)

- **تابع گذار احتمال (Transition Probability)** –

$$P(s' \mid s, a)$$

احتمال این که با انجام اقدام a در حالت s، به حالت بعدی s' منتقل شویم.

- گاهی به اختصار $P^a_{ss'}$ یا $P_{ss'}$ هم نوشته می‌شود.

- **تابع پاداش (Reward Function)** –

$$R(s, a, s')$$

پاداشی که محیط پس از انتقال از s به s' به واسطه اقدام a می‌دهد.

- گاهی به اختصار $R^a_{ss'}$ یا $R(s, a)$ نیز می‌آید.

- **پاداش فوری (Immediate Reward)** – r_{t+1}
مقدار عددی که بلافاصله پس از انجام اقدام A_t در حالت S_t و رسیدن به S_{t+1} دریافت می‌شود.

$$R(S_t, A_t, S_{t+1}) = r_{t+1}$$

۳. بازگشت (Return)

- **بازگشت (Return) – G_t**
  مجموع پاداش‌های آینده از گام t به بعد، با اعمال ضریب تنزیل γ:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

که در عمل معمولاً تا یک پایان (Terminal State) یا تعداد محدود گام محاسبه می‌شود.

- **ضریب تنزیل (Discount Factor) – $0 \leq \gamma \leq 1$**
  تعیین می‌کند که پاداش‌های دوردست تا چه حد ارزش دارند.
- اگر $\gamma = 0$، فقط پاداش فوری اهمیت دارد.
- اگر $\gamma \approx 1$، پاداش‌های آینده تقریباً به اندازه پاداش‌های فوری اهمیت دارند.

۴. ارزش‌ها (Value Functions)

- تابع ارزش حالت (State-Value Function) –

$$[s = {}_tS \mid {}_t\mathbb{E}_\pi[G = v_\pi(s)$$

یعنی انتظار بازگشت ${}_tG$ وقتی عامل از سیاست π پیروی کند و در حالت s باشد.

- تابع ارزش جفت حالت-اقدام (Action-Value Function) –

$$[a = {}_ts, A = {}_tS \mid {}_t\mathbb{E}_\pi[G = q_\pi(s, a)$$

یعنی انتظار بازگشت وقتی در s اقدام a اجرا شود و سپس از π پیروی شود.

۵. معادله بلمن (Bellman Equation)

معادله بلمن رابطه بازگشت کوتاه‌مدت و بلندمدت ارزش را برقرار می‌کند.

۵.۱. بلمن برای v_π

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s',r} P(s',r \mid s,a)\left[r + \gamma v_\pi(s')\right]$$

$P(s',r \mid s,a)$ احتمال مشترک رسیدن به حالت s' و دریافت پاداش r.

این معادله می‌گوید: ارزش یک حالت برابر است با میانگین پاداش فوری به اضافه ارزش حالت بعدی (تنزیل‌شده)، که بر اساس سیاست π و دینامیک محیط وزن‌دهی شده‌اند.

۵.۲. بلمن بهینه (Optimal Bellman Equation)

برای سیاست بهینه π^* و ارزش بهینه v_*:

$$v_*(s) = \max_a \sum_{s',r} P(s',r \mid s,a)\left[r + \gamma v_*(s')\right]$$

و به طور مشابه برای q_*:

$$q_*(s,a) = \sum_{s',r} P(s',r \mid s,a)\left[r + \gamma \max_{a'} q_*(s',a')\right]$$

۶. مثال ساده

فرض کنید یک عامل در خانه‌های شماره 1 تا 3 حرکت می‌کند، و اگر به خانه 3 برسد، بازی تمام می‌شود (پایان) با پاداش ۱. هر حرکت • هزینه دارد. ضریب تنزیل $\gamma = 0.9$.

- **حالت‌ها:** $s \ni \{1, 2, 3\}$
- **اقدام‌ها:** $a \ni \{\text{چپ}, \text{راست}\}$ (به ترتیب حرکت به خانه قبلی یا بعدی)
- **دینامیک:**
- از 1 با «چپ» درجا می‌ماند.
- از 2 با «راست» به 3 می‌رود.
- **پاداش:**
- هر حرکت: $r = 0$
- رسیدن به 3: $r = 1$ و سپس پایان.

محاسبه بازگشت و ارزش‌ها

- اگر از 2 با «راست» برویم:
$R_{t+1} = 1$، بازی تمام، پس

$$G_t = 1.$$

- تابع ارزش بهینه (بدون جزئیات):

$$v_*(2) = 0 + 0.9 \times v_*(3) = 1 \times 0.9 = 0.9$$

چون $v_*(3) = 1$ (حالت پایان با ۱ پاداش).

جمع‌بندی نمادها

نماد	معنی
s, s'	حالت فعلی و حالت بعدی
a	اقدام
$\pi(a \mid s)$	سیاست (احتمال یا قاعده انتخاب a در s)
$P(s' \mid s, a)$	احتمال انتقال به s' با اقدام a در s
$R(s, a, s')$	پاداش دریافت‌شده در انتقال $s \to s'$ با a
r_{t+1}	پاداش فوری در گام $t+1$
G_t	بازگشت (مجموع پاداش‌های تنزیل‌شده پس از گام t)
γ	ضریب تنزیل
$v_\pi(s)$	ارزش حالت تحت سیاست π
$q_\pi(s, a)$	ارزش جفت حالت-اقدام تحت سیاست π
$v_*(s), q_*(s, a)$	ارزش‌های بهینه (حداکثرکردن بازگشت)

با این توضیحات و مثال ساده، پایه‌های اصلی یادگیری تقویتی و معادلات بلمن روشن می‌شود.