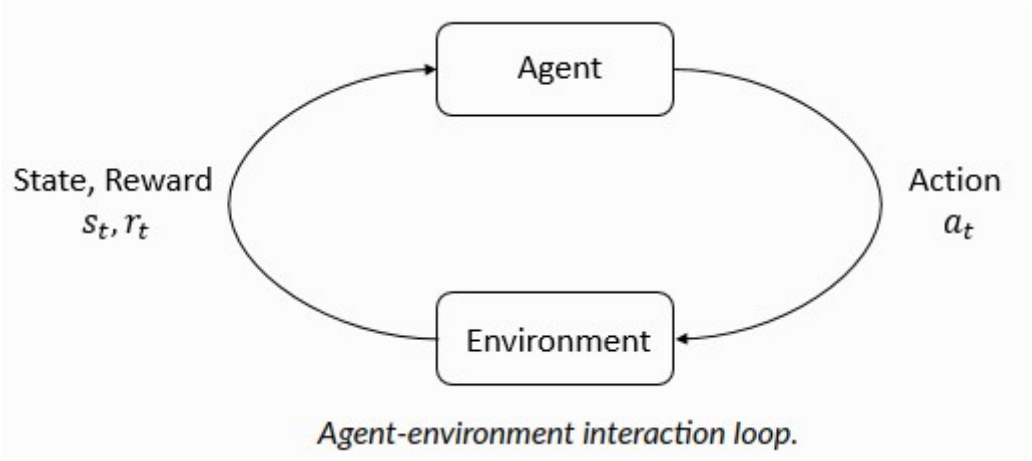
Antes… o que é aprendizado por reforço?

Aprendizado por reforço (*Reinforcement learning* - RL) é o estudo dos agentes e como eles aprendem por tentativa e erro. Formaliza a ideia de que premiar ou punir um agente pelos seu comportamento faz com que seja mais provável que ele repita ou evite o mesmo comportamento no futuro.

Principais conceitos e terminologias:

Os personagens principais do RL são o agente (*agent*) e o ambiente (*environment*). O ambiente é o mundo no qual o agente vive e interage. A cada passo de interação, o agente vê uma observação de um estado (*state*) e então decide qual ação (*action*) tomar. O ambiente pode mudar por conta própria ou reagir a uma ação do agente.



Agent-environment interaction loop.

O agente também recebe uma recompensa (reward) do ambiente, um número que quantifica quão boa ou ruim o estado atual é. O objetivo do agente é maximizar a premiação cumulativa, também chamada de retorno (return). Métodos de RL são as maneiras que o agente pode aprender os comportamentos que o levam a esse objetivo.

Estados e Observações:

Um estado (s) é uma descrição completa do estado do mundo. Não tem nenhuma informação sobre o mundo que se encontra oculta no estado. Um observação (o) é uma descrição parcial do estado, a qual pode omitir alguma informação.

No Deep RL, estado e observações são representados por vetores, matrizes ou tensores. Por exemplo, uma observação visual pode ser representada por uma matriz RGB, um robô, pode ser representado pelo ângulo de suas articulações e velocidades.

Quando o agente é capaz de observar o estado completo do ambiente, dizemos que o ambiente é **completamente observável**. Quando essa obervação não é completa, dizemos que o ambiente é **parcialmente observável**.

> “A notação de formal às vezes coloca o símbolo de estado, s, em locais onde seria tecnicamente mais apropriado escrever o símbolo de observação, o. Especificamente, isso acontece quando

falamos sobre como o agente decide uma ação: muitas vezes sinalizamos em notação que a ação está condicionada ao estado, quando na prática, a ação está condicionada à observação porque o agente não tem acesso ao estado.”

Espaço de Ações:

Diferentes ambientes permitem diferentes tipos de ações. O conjunto de todas as ações válidas é chamado de espaço de ações (*action space*). Alguns ambientes têm espaço de ações **discretos**, nos quais apenas um número finito de movimentos está disponível para o agente. Outros, têm espaço de ações **contínuos**, permitindo infinito número de ações (dentro do conjunto dos números Reais).

Políticas:

A política é uma regra usada por um agente para decidir quais ações tomar. Ela pode ser determinística, nesse caso será denotada por μ:

$$a_t = \mu(s_t),$$

ou ela pode ser estocástica, nesse caso será denotada por π

$$a_t \sim \pi(.|s_t).$$

A polícia é, essencialmente, o cérebro do agente, por isso, é comum utilizar as duas para dizer a mesma coisa. “A política – o agente – está tentando maximizar a recompensa”.

No D-RL, as políticas são parametrizadas, ou seja, são funções que dependem de um conjunto de parâmetros (pesos de uma rede neural) que podem ser atualizados por um algoritmo de otimização. Esses pesos são denotados por θ ou ϕ, portanto, as políticas serão denotadas por: $a_t = \mu_\theta(s_t)$ e $a_t \sim \pi_\theta(.|s_t)$.

Políticas Determinísticas e Estocásticas:

Políticas determinísticas são funções nas quais para um determinado conjunto de entradas, sempre produzirá o mesmo conjunto de saídas. Um exemplo seria uma rede MLP. Politicas estocásticas, por outro lado, para o mesmo conjunto de entradas, diferentes saídas podem ser produzidas em diferentes execuções, obedecendo uma função de probabilidade.

Se o espaço de ação for discreto, a política utilizada será categórica, se o espaço de ação for contínuo, será utilizada a política Gaussiana diagonal (diagonal Gaussian policies – DGP). Duas ações são importantes para política estocástica a amostragem (*sampling*) e o cálculo dás verossimilhanças logarítmicas (*log likelihoods* – ll) de cada ação, $\log \pi_\theta(a|s)$.

A política categórica é construída como um classificado de funções discretas. A rede neural recebe em sua entrada as observações e a camada final resulta nos *logits* de cada ação, seguida por uma *sofmax* para transformar a saída em uma distribuição de probabilidades.

Amostragem: Dada a função de distribuição de probabilidade (pdf) é possível gerar amostras que obedeçam a função geradora.
Verossimilhança: Seja a saída da última camada as probabilidades $P_\theta(s)$ um de tamanho equivalente ao número de ações possíveis, se forma que a ll para uma ação a pode ser obtida por:

$$\log \pi_\theta(a|s) = \log[P_\theta(s)]_a$$

A DGP é construída por uma distribuição Gaussiana multivariável, descrita por um vetor de médias μ e uma matriz de covariância Σ. A Gaussiana diagonal é um caso particular onde a matriz de covariância é uma matriz diagonal e logo pode ser representada por um vetor.
Uma rede neural será responsável por mapear as observações em ações médias $\mu_\theta(s)$. Existem duas formas de se obter a matriz de covariância. Na primeira existe um vetor de logaritmos dos desvios padrões, log σ, que não é função do estado é portanto são parâmetros fixos. No segundo caso existe uma rede neural para estimação dos logaritmos dos desvios padrão log $\sigma(s)$, que serão dependentes das observações.

Amostragem: Dado um vetor de ações médias $\mu_\theta(s)$, um vetor de desvios padrões $\sigma_\theta(s)$ e um vetor de ruído $z \sim N(0, I)$, uma ação pode ser amostrada como

$$a = \mu_\theta(s) + \sigma_\theta(s) \odot z$$

Verossimilhança: A ll the uma ação a k-dimensional, para uma Gaussiana com média $\mu = \mu_\theta(s)$ e desvio padrão $\sigma = \sigma_\theta(s)$ é dado por:

$$\log \pi_\theta(a|s) = \frac{-1}{2} \sum_{i=1}^{k} \left(\frac{(a_i - \mu_i)^2}{\sigma_i^2} + 2 \log(\sigma_i) + k \log(2\pi) \right).$$

Trajetória:

A Trajetória, ou episódio, τ é uma sequencia de estados e ações $\tau = (s_0, a_0, s_1, a_1, \ldots)$. O primeiro estado s_0 é aleatoriamente amostrado por uma distribuição do estado inicial. Os estados de transição são governados pelas leis do ambientes e dependem somente da ação mais recente, podendo ser determinísticos $s_{t+1} = f(s_t, a_t)$ ou estocásticos $st+1 \sim P(.|s_t, a_t)$.

Recompensa e Retorno:

A função de recompensa (*reward*) R depende do estado atual, da ação tomada e do estado futuro

$$r_t = (s_t, a_t, s_{t+1})$$

em alguns casos, pode haver a dependência apenas do estado atual ou do par estado-ação atual.

O objetivo do agente é maximizar a recompensa cumulativa sobre uma trajetória. Uma forma de retorno é o retorno de horizonte finito não descontado, o que significa apenas que o retorno é a soma de todas as recompensas em uma trajetória

$$R(\tau) = \sum_{t=0}^{T} r_t$$

Outro tipo de retorno é o retorno de horizonte infinito descontado, que é a soma de todas as recompensas obtidas pelo agente, mas com um desconto proporcional ao tempo em que essa ação foi tomada. Sendo o desconto $\gamma \in (0,1)$ a recompensa será

$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t$$

O fator de desconto favorece recompensas imediatas e matematicamente garante a convergência do somatório.

O objetivo do RL:

O objetivo do aprendizado por reforço é obter uma política que maximiza o retorno esperado. Sendo as transições do ambiente e a política estocásticos, a probabilidade te uma trajetória de T passos será:

$$P(\tau|\pi)=\rho_0(s_0)\prod_{t=0}^{T-1} P(s_{t+1}|s_t,a_t)\,\pi(a_t|s_t)$$

O retorno esperado será:

$$J(\pi)=\int_{\tau} P(\tau|\pi)R(\tau)=E\left[R(\tau)\right].$$

O problema de otimização será encontrar a política ótima:

$$\pi'=\arg\max_{\pi} J(\pi).$$

Funções de Valores:

Frequentemente necessita-se saber o valor de um estado, ou de um par estado-ação. Matematicamente, valor é a média do valor esperado se uma estado ou par estado-ação forem tomados como iniciais e então agir somente de acordo com uma política específica para sempre. As funções de valores são:

1. Função de valor na política (*On-Policy Value Function*): É definida como o retorno esperado se começar no estado s e somente agir de acordo com a política π

$$V^{\pi}(s)=E_{\tau\sim\pi}[R(\tau)|s_0=s]$$

2. Função de valor-ação na política (*On-Policy Action-Value Function*): É definida como o retorno esperado se iniciar-se no estado s e tomar uma ação arbitrária a (originada ou não da política) e então agir para sempre de acordo com a política π:

$$Q^{\pi}(s,a)=E_{\tau\sim\pi}[R(\tau)|s_0=s,a_0=a]$$

3.Função de valor ótima (*Optimal Value Function*) caso 1 para quando $\pi=\pi'$

$$V^{*}(s)=\max_{\pi} E_{\tau\sim\pi}[R(\tau)|s_0=s]$$

4. Função de ação-valor ótima (*Optimal Action-Value Function*): caso 2 para quando $\pi=\pi'$

$$Q^{*}(s,a)=\max_{\pi} E_{\tau\sim\pi}[R(\tau)|s_0=s,a_0=a]$$

> “Quando falamos sobre funções de valor, se não fizermos referência à dependência do tempo, queremos dizer apenas o retorno esperado descontado no horizonte infinito. Funções de valor para retorno não descontado em horizonte finito precisariam aceitar o tempo como argumento.”

Função Q ótima e Ação ótima:

Existe uma conexão entre função de ação-valor ótima $Q^{*}(s,a)$ e a ação selecionada pela política ótima. Por definição, $Q^{*}(s,a)$ é o retorno esperado por começar em um estado s e tomar uma ação arbitrária a e então agir de acordo com a política ótima para sempre. A política ótima vai selecionar uma ação que maximiza o retorno esperado começando a partir de s, assim, a ação ótima pode ser obtida por:

$$a^{*}=\arg\max_{a} Q^{*}(s,a)$$

Equações de Bellman:

Todas as quatro funções de valor obedecem a equações especiais de autoconsistência chamadas equações de Bellman.

"O valor do seu ponto de partida é a recompensa que você espera receber por estar lá, mais o valor de onde você pousar em seguida."

As equações de Bellman para os valores na política são:

$$V^{\pi} = \underset{\substack{a\sim\pi \\ s'\sim P}}{E}\left[r(s,a) + \gamma V^{\pi}(s')\right],$$

$$Q^{\pi}(s,a) = \underset{s'\sim P}{E}\left[r(s,a) + \gamma \underset{a'\sim\pi}{E}\left[Q^{\pi}(s',a')\right]\right],$$

em que $s' \sim P$ é uma forma abreviada de $s' \sim P(.|s,a)$, indicando que o próximo estado s' é amostrado pelas regras de transição do ambiente. Similarmente, $a \sim \pi = a \sim \pi(.|s)$ e $a' \sim pi = a' \sim \pi(.|s')$.

As equações de Bellman para as funções valores ótimas são:

$$V^{*} = \max_{a} \underset{s'\sim P}{E}\left[r(s,a) + \gamma V^{\pi}(s')\right],$$

$$Q^{\pi}(s,a) = \underset{s'\sim P}{E}\left[r(s,a) + \gamma \max_{a'}\left[Q^{\pi}(s',a')\right]\right],$$

Funções de Vantagem:

As vezes, não é necessário descrever o quão bom uma ação é, somente o quão melhor ela é, na média, quando comparada às outras, ou seja, precisamos saber qual é a vantagem relativa daquela função. Esse conceito é definido por meio da função de vantagem (*advantage function*).

A função de vantagem $A^{\pi}(s,a)$ correspondente à politica π descreve quão melhor é tomar uma ação específica a em um estado s, sobre uma ação selecionada aleatoriamente de acordo com $\pi(.|s)$, assumindo que a política seja mantida para sempre. Matematicamente:

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$$