La ecuación de Bellman nos permite estimar el valor de un estado bajo una política de acción π.

En otras palabras, buscamos la esperanza matemática de la ganancia total que obtendremos partiendo desde ese estado hasta el final de la simulación.

$$v_\pi(s) \doteq E_\pi[G_t \mid S_t = s]$$

Si tenemos un problema de decisión de Markov bien definido (MDP), podemos expresar esta ecuación en términos de la función de probabilidad p que define un MDP.

El valor de un estado es la suma del valor de las acciones que es posible tomar en ese estado ponderadas por la probabilidad de tomar dicha acción.

$$v_\pi(s) = \underbrace{\sum_a}_{\text{Suma sobre las acciones}} \underbrace{\pi(a \mid s)}_{\text{Probabilidad de elegir la acción}} \underbrace{\sum_{s', r} p(s', r \mid s, a)\,[r + \gamma v_\pi(s')]}_{\text{Valor de la acción en ese estado}}$$

El valor de una acción dado un estado está incluido en la ecuación de Bellman.

Si al elegir una acción a en un estado s obtenemos como resultado una recompensa r y un estado s', entonces el valor esperado de la recompensa que obtendremos a partir de ese punto es $[r + \gamma v_\pi(s')]$, que corresponde a la recompensa obtenida mas el valor esperado descontado de lo que obtendremos a partir del nuevo estado.

En palabras más simples, si vemos los resultados de una acción a en un estado s como una tupla (s', r), entonces el valor de esa tupla es $[r + \gamma v_\pi(s')]$.

El valor de una acción en un estado es la suma del valor de los posibles resultados de esa acción ponderados por la probabilidad de que ocurran dichos resultados.

$$q_\pi(a \mid s) = \underbrace{\sum_{s', r}}_{\text{Suma sobre las tuplas resultado de } (s, a)} \underbrace{p(s', r \mid s, a)}_{\text{Probabilidad de que } (s, a) \text{ lleve a un resultado } (s', r)} \underbrace{[r + \gamma v_\pi(s')]}_{\text{Valor del resultado}}$$