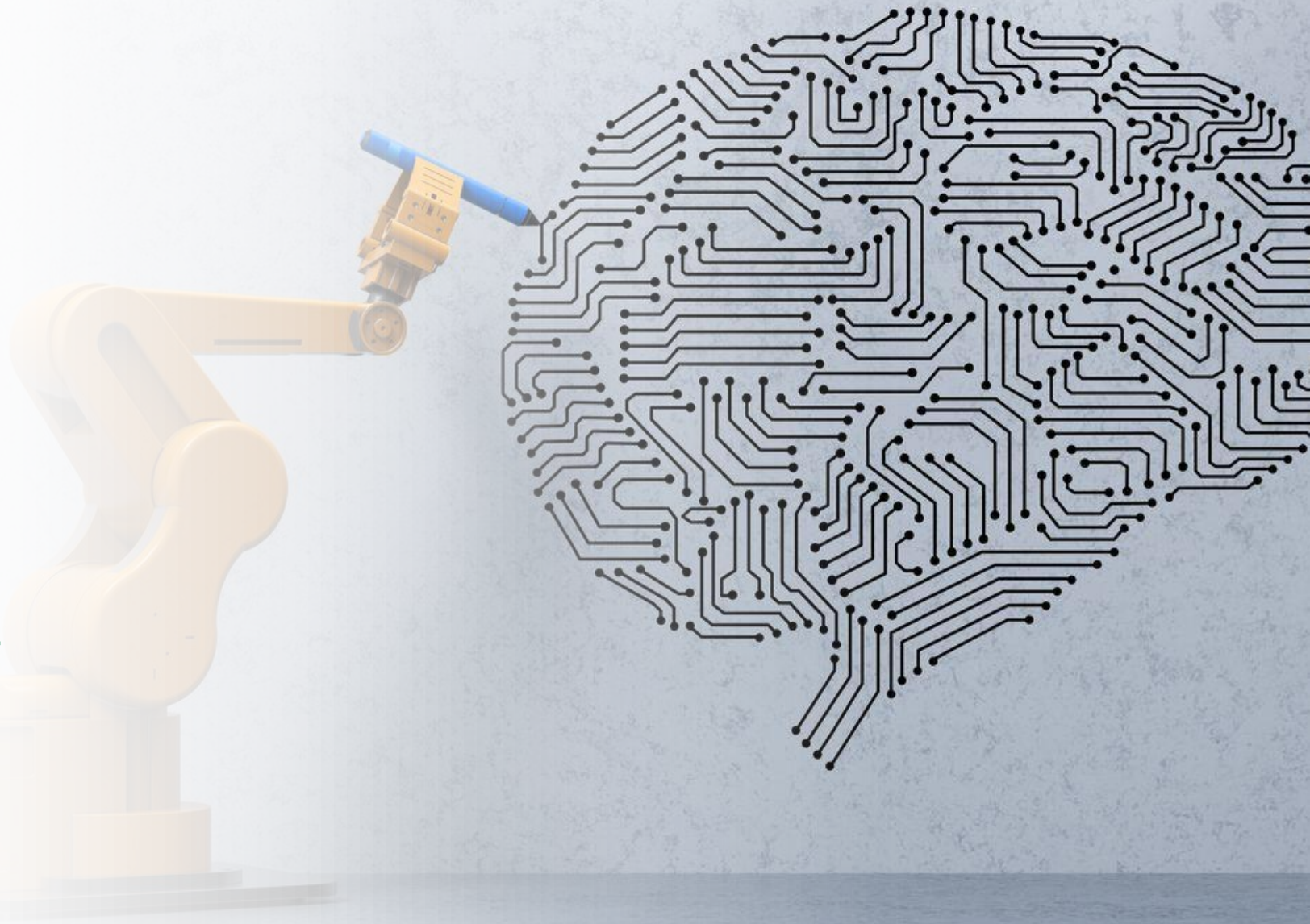
Aprendizaje por refuerzo

Clase 19: Teoría



Antes de empezar...

- Dudas de
  - Tarea 4
  - Proyecto

Para el día de hoy…

- Preliminares
- Operadores de Bellman
- Análisis de
  - Programación dinámica
    Programación dinámica aproximada
  - Política voraz

¿Qué tipo de preguntas nos hacemos?

- ¿Convergen nuestros algoritmos?
- ¿Con que velocidad?
- Si utilizo este algoritmo con N muestras, k iteraciones, que tan bueno será el resultado
- Si utilizo este algoritmo de exploración, ¿cuál es el arrepentimiento?
- Muchas otras

Tipo de suposiciones

- Exploración
  - El desempeño de los métodos se complica con exploración
  - ¿Qué tan probable es encontrar recompensas?
  - Normalmente las garantías teóricas suponen peores escenarios
- Aprendizaje
  - Si podemos abstraer la exploración, ¿cuántas muestras necesitamos para aprender un modelo?

¿Para qué?

- Demostrar que el algoritmo funciona siempre
  - Usualmente no es posible para métodos modernos
- Entender como se afectan los errores por los parámetros
  - ¿Funciona mejor un descuento grande o pequeño?
  - Si queremos reducir el error a la mitad, ¿Cuántas muestras necesitamos?

Algunas definiciones

- Un espacio vectorial normado
  - Espacio vectorial $\mathcal{X}$ + una norma $||\cdot||$ sobre los elementos de $\mathcal{X}$
- Las normas definen un mapeo $\mathcal{X} \rightarrow \mathbb{R}$ tal que
  - $||x|| \geq 0, \forall x \in \mathcal{X}$ y si $||x|| = 0$ entonces $x = 0$
  - $||\alpha x|| = |\alpha|||x||$
  - $||x_1 + x_2|| \leq ||x_1|| + ||x_2||$
- Para hoy:
  - Espacio vectorial: $\mathcal{X} = \mathbb{R}^d$
  - Normas: $||\cdot||_\infty, ||\cdot||_2, ||\cdot||_p$

Mapeo de contracción

- Sea $\mathcal{X}$ un espacio vectorial normado con $||\cdot||$. Un mapeo $\mathcal{T}: \mathcal{X} \to \mathcal{X}$ es una contracción α si para cualquier $x_1, x_2 \in \mathcal{X}$, $\exists a \in [0,1)$ tal que

$$||\mathcal{T}x_1 - \mathcal{T}x_2|| \leq \alpha||x_1 - x_2||$$

- Si $\alpha \in [0,1]$, entonces llamamos $\mathcal{T}$ no expandible
- Cada contracción es también Lipschitz, por lo cual también es continuo
- Si $x_n \to_{||\cdot||} x$ entonces $\mathcal{T}x_n \to_{||\cdot||} \mathcal{T}x$

Puntos fijos

- Punto fijo
  - Un vector $x \in \mathcal{X}$ es un punto fijo de un operador $\mathcal{T}$ si $\mathcal{T}x = x$
- Teorema de punto fijo de Banach
  - Sea $\mathcal{X}$ un espacio vectorial normado con $||\cdot||$ y $\mathcal{T}: \mathcal{X} \to \mathcal{X}$ una contracción γ, entonces
    - $\mathcal{T}$ tiene un punto fijo único $x \in \mathcal{X}: \exists x^* \in \mathcal{X}$ tal que $\mathcal{T}x^* = x^*$
    - $\forall x_0 \in \mathcal{X}$, la secuencia $x_{n+1} = \mathcal{T}x_n$ converge a x^* de forma geométrica
    - $||x_n - x^*|| \leq \gamma^n ||x_0 - x^*||$
    - Entonces $\lim_{n\to\infty} ||x_n - x^*|| \leq \lim_{n\to\infty} (\gamma^n ||x_0 - x^*||) = 0$

Proceso de decisión de Markov

- Un proceso de decisión de Markov es un proceso de recompensa de Markov con decisiones
- Formalmente, un proceso de decisión de Markov es una tupla $\mathrm{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$
  - $\mathcal{S}$ es un conjunto finito de estados
  - $\mathcal{A}$ es un conjunto finito de acciones
  - $\mathcal{P}$ es una matriz de transición, $\mathcal{P}_{ss'}^{a} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$
  - $\mathcal{R}$ es una función de recompensa, $\mathcal{R}_s^a = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$
  - γ es un factor de descuento, $\gamma \in [0, 1]$
- Alternativamente
  - $T(s, a, s')$ define las transiciones o dinámica del modelo
  - $R(s, a, s') | R(s, a) | R(s') | R(s)$ define la función de recompensa

Funciones de valor

- Función de valor, para una política π:

- Función de acción valor, para una política π

- Funciones óptimas

Funciones de valor

- Función de valor, para una política π:

$$v_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | s_0 = s; \pi\right]$$

- Función de acción valor, para una política π

$$q_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | s_0 = s; a_0 = a; \pi\right]$$

- Funciones óptimas

$$q^* = \max_\pi q_\pi$$
$$v^* = \max_\pi v_\pi$$

Ecuaciones de Bellman esperadas

- Dado un MDP $M = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$, para cualquier política π, las función de valor obedecen

$$v_\pi(s) = \sum_a \pi(s,a)[r(s,a) + \gamma \sum_{s'} p(s'|a,s) v_\pi(s')]$$

$$q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|a,s) \sum_{a' \in \mathcal{A}} \pi(a'|s') q_\pi(s',a')$$

Ecuaciones de Bellman óptimas

- Dado un MDP $M = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$, para cualquier política π, las función de valor óptimas obedecen

$$v^*(s) = \max_a \left[r(s,a) + \gamma \sum_{s'} p(s'|a,s) v^*(s') \right]$$

$$q^*(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|a,s) \max_{a' \in \mathrm{A}} q^*(s',a')$$

Operador de Bellman óptimo

- Dado un MDP $M = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$, sea $\mathcal{V} \equiv \mathcal{V}_S$ el espacio acotado de funciones reales sobre $\mathcal{S}$. Definimos, el operador de Bellman $\mathrm{T}_{\mathcal{V}}^{*}: \mathcal{V} \to \mathcal{V}$ como

$$(T_{\mathcal{V}}^{*} f)(s) = \max_{a} \left[r(s,a) + \gamma \sum_{s'} p(s'|a,s) f(s') \right], \forall f \in \mathcal{V}$$

- Se suele usar $T^{*} \equiv T_{\mathcal{V}}^{*}$

Propiedades del operador de Bellman

1. Tiene un único punto fijo v^*

$$T^*v^* = v^*$$

2. T^* es una contracción γ con respecto a $||\cdot||_\infty$

$$||T^*v - T^*u||_\infty \leq \gamma||v - u||_\infty, \forall u, v \in \mathcal{V}$$

3. T^* es monotónico

$\forall u, v \in \mathcal{V}$ tal que $u \leq v$ (por componentes), entonces

$$T^*u \leq T^*v$$

Demostración: T^* es una contracción γ con respecto a $||\cdot||_\infty$

$$
\begin{aligned}
|T^*v(s) - T^*u(s)| &= |\max_a \left[r(s,a) + \gamma \mathbb{E}_{s'|s,a} v(s')\right] - \max_b \left[r(s,b) + \gamma \mathbb{E}_{s''|s,b} u(s'')\right]| && (6)\\
&\leq \max_a |\left[r(s,a) + \gamma \mathbb{E}_{s'|s,a} v(s')\right] - \left[r(s,a) + \gamma \mathbb{E}_{s'|s,a} u(s')\right]| && (7)\\
&= \gamma \max_a |\mathbb{E}_{s'|s,a} \left[v(s') - u(s')\right]| && (8)\\
&\leq \gamma \max_{s'} |\left[v(s') - u(s')\right]| && (9)
\end{aligned}
$$

- Por lo tanto

$$||T^*v - T^*u||_\infty \leq \gamma ||v - u||_\infty, \forall u, v \in \mathcal{V}$$

- Los pasos (6) y (7) usan

$$\max_a f(a) - \max_b g(b) \leq \max_a |f(a) - g(a)|$$

Demostración: T^* es monotónico

- Dado $v(s) \leq u(s), \forall s \rightarrow r(s,a) + \mathbb{E}_{(s'|s,a)} v(s') \leq r(s,a) + \mathbb{E}_{(s'|s,a)} u(s')$

$$
\begin{aligned}
T^*v(s) - T^*u(s) &= \max_a \left[r(s,a) + \gamma \mathbb{E}_{s'|s,a} v(s') \right] - \max_b \left[r(s,b) + \gamma \mathbb{E}_{s''|s,b} u(s'') \right] && (10) \\
&\leq \max_a \left(\left[r(s,a) + \gamma \mathbb{E}_{s'|s,a} v(s') \right] - \left[r(s,a) + \gamma \mathbb{E}_{s'|s,a} u(s') \right] \right) && (11) \\
&\leq 0, \forall s. && (12)
\end{aligned}
$$

- Por lo cual

$$T^*v(s) \leq T^*u(s), \forall s \in \mathcal{S}$$

Iteración de valor desde la perspectiva del operador de Bellman

- Algoritmo
  - Iniciar con v_0
  - Actualizar: $v_{k+1} = T^* v_k$
- Nótese que $k \to \infty, v_k \to_{||\cdot||_\infty} v^*$

$$
\begin{aligned}
\|v_k - v^*\|_\infty &= \|T^* v_{k-1} - v^*\|_\infty \\
&= \|T^* v_{k-1} - T^* v^*\|_\infty \\
&\leq \gamma \|v_{k-1} - v^*\|_\infty \\
&\leq \gamma^k \|v_0 - v^*\|_\infty
\end{aligned}
$$

Operador de Bellman esperado

- Dado un MDP $M = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$, sea $\mathcal{V} \equiv \mathcal{V}_S$ el espacio acotado de funciones reales sobre $\mathcal{S}$. Para cualquier política $\pi: \mathcal{S} \times \mathcal{A} \to [0,1]$, definimos, el operador de Bellman esperado $\mathrm{T}_{\mathcal{V}}^{*}: \mathcal{V} \to \mathcal{V}$ como

$$(T_{\mathcal{V}}^{\pi} f)(s) = \sum_{a} \pi(s,a) \left[r(s,a) + \gamma \sum_{s'} p(s'|a,s) f(s') \right], \forall f \in \mathcal{V}$$

- Se suele usar $T^{\pi} \equiv T_{\mathcal{V}}^{\pi}$

Propiedades del operador de Bellman

1. Tiene un único punto fijo v_π

$$T^\pi v_\pi = v_\pi$$

2. T^π es una contracción γ con respecto a $||\cdot||_\infty$

$$||T^\pi v - T^\pi u||_\infty \leq \gamma ||v - u||_\infty, \forall u, v \in \mathcal{V}$$

3. T^π es monotónico

$\forall u, v \in \mathcal{V}$ tal que $u \leq v$ (por componentes), entonces

$$T^\pi u \leq T^\pi v$$

Demostración: T^* es una contracción γ con respecto a $||\cdot||_\infty$

$$
\begin{aligned}
T^\pi v(s) - T^\pi u(s) &= \sum_a \pi(a|s)\left[r(s,a) + \gamma \mathbb{E}_{s'|s,a} v(s') - r(s,a) - \gamma \mathbb{E}_{s'|s,a} u(s')\right] \\
&= \gamma \sum_a \pi(a|s) \mathbb{E}_{s'|s,a}\left[v(s') - u(s')\right] \qquad (14) \\
\Rightarrow |T^\pi v(s) - T^\pi u(s)| &\leq \gamma \max_{s'} |\left[v(s') - u(s')\right]|
\end{aligned}
$$

- Por lo tanto

$$||T^\pi v - T^\pi u||_\infty \leq \gamma ||v - u||_\infty, \forall u, v \in \mathcal{V}$$

- Por (14) también se demuestra la monotonicidad de T^π

Evaluación de política

- Algoritmo
  - Iniciar con v_0
  - Actualizar: $v_{k+1} = T^{\pi} v_k$
- Nótese que $k \to \infty, v_k \to_{||\cdot||_\infty} v_\pi$ (directa aplicación del teorema de punto fijo de Banach)

En resumen

Value Iteration

- Start with v_0.
- Update values: $v_{k+1} = T^* v_k$.

Policy Iteration

- Start with π_0.
- Iterate:
  - Policy Evaluation: v_{π_i}
    - (E.g. For instance, by iterating T^π: $v_k = T^{\pi_i} v_{k-1} \Rightarrow v_k \to v^{\pi_i}$ as $k \to \infty$)
  - Greedy Improvement: $\pi_{i+1} = \arg\max_a q_{\pi_i}(s, a)$

Programación dinámica aproximada

- Hasta ahora hemos supuesto conocimiento perfecto del MDP así como de las funciones de valor
- De forma realista,
  - No conocemos el MDP
  - No podemos representar la función de valor exactamente
- Objetivo
  - Bajo estas condiciones encontrar una política π óptima (o casi)

Iteración de valor aproximada

Approximate Value Iteration

- Start with v_0.
- Update values: $v_{k+1} = \mathcal{A}T^* v_k$. $(v_{k+1} \approx T^* v_k)$

- ¿$k \to \infty, v_k \to ||\cdot||_\infty v^*$?
- En general, no

Aproximación de la función de valor

- Utilizar una función de aproximación $v_\theta(s)$ con parámetros $\theta \in \mathbb{R}^m$
- La función de valor estimada en la iteración k es $v_k = v_{\theta_k}$
- Utilizar programación dinámica para calcular $v_{\theta_{k+1}}$ desde v_{θ_k}

$$T^* v_k(s) = \max_a \mathbb{E}[R_{t+1} + \gamma v_k(S_{t+1}) | S_t = s]$$

- Ajustar θ_{k+1} tal que $v_{\theta_{k+1}} \approx T^* v_k(s)$. Por ejemplo

$$\theta_{k+1} = \arg\min_{\theta_k+1} \sum_s \left(v_{\theta_{k+1}}(s) - T^* v_k(s) \right)^2$$

Iteración de valor aproximada

Approximate Value Iteration

- Start with v_0.
- Update values: $v_{k+1} = \mathcal{A}T^* v_k$. $(v_{k+1} \approx T^* v_k)$

- $¿k \to \infty, v_k \to ||\cdot||_\infty v^*?$
- En general, no
- ¿Entonces?
  - Las versiones de muestreo convergen bajo ciertas condiciones
  - Para el caso de funciones de aproximación, aunque puede divergir, rara vez sucede en la práctica
  - Existen muchas funciones de valor que inducen la política óptima

Desempeño de una política voraz

- Considere un MDP M. Sea $q: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ una función arbitraria y sea π una política voraz asociada con q entonces
- $||q^* - q^\pi||_\infty \leq \frac{2\gamma}{1-\gamma} ||q^* - q||_\infty$
- Donde q^* es la función de valor óptima asociada con M

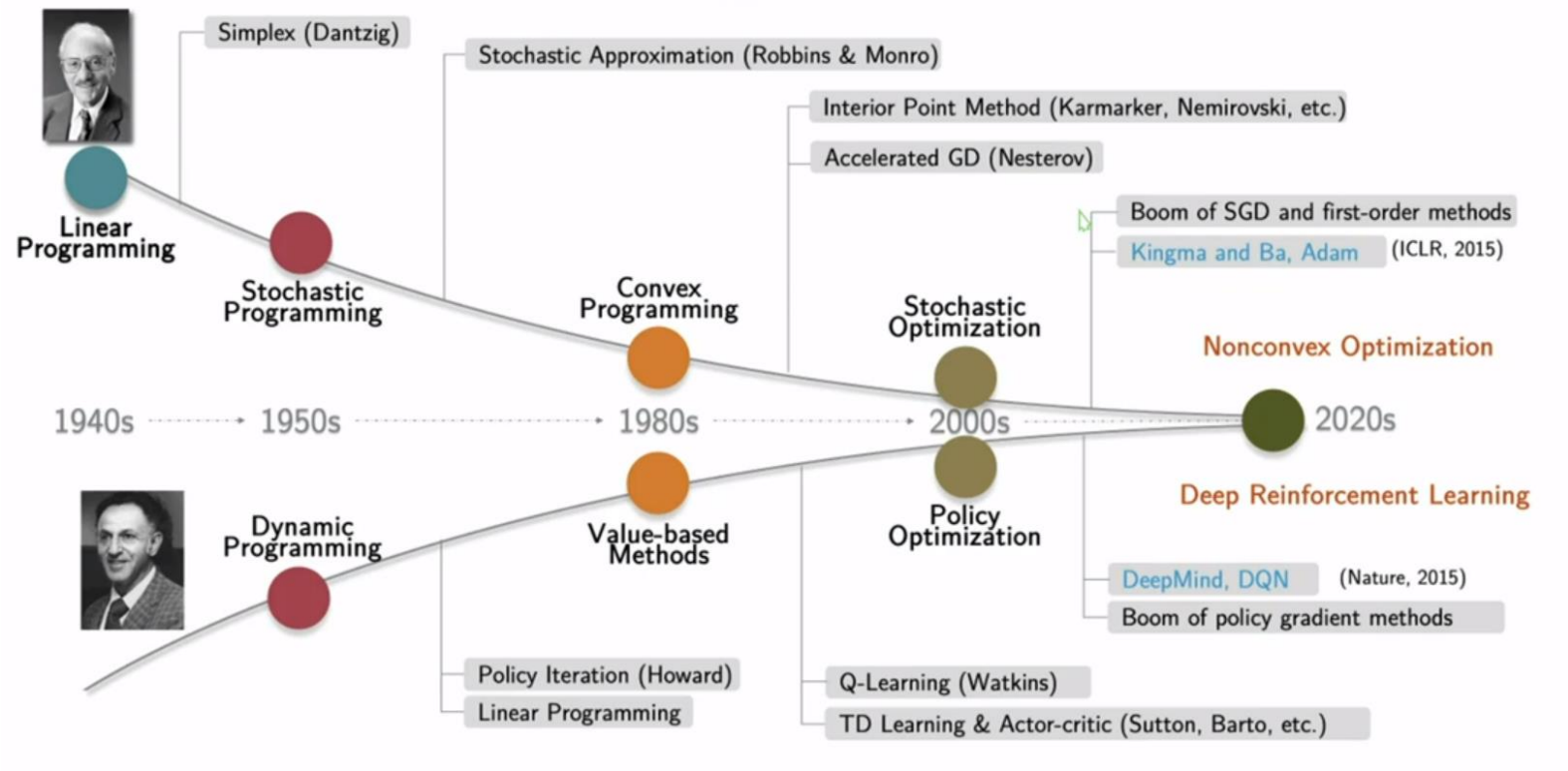
Demostración $||q^* - q^\pi||_\infty \leq \frac{2\gamma}{1-\gamma}||q^* - q||_\infty$

$$
\begin{aligned}
\|q^* - q^\pi\|_\infty &= \|q^* - T^\pi q + T^\pi q - q^\pi\|_\infty && (15)\\
&\leq \|q^* - T^\pi q\|_\infty + \|T^\pi q - q^\pi\|_\infty && (16)\\
&= \|T^* q^* - T^* q\|_\infty + \|T^\pi q - T^\pi q^\pi\|_\infty && (17)\\
&\leq \gamma\|q^* - q\|_\infty + \gamma \underbrace{\|q - q^\pi\|_\infty}_{\leq \|q - q^*\|_\infty + \|q^* - q^\pi\|_\infty} && (18)\\
&\leq 2\gamma\|q^* - q\|_\infty + \gamma\|q^* - q^\pi\|_\infty && (19)
\end{aligned}
$$

- Reacomodando los términos

$$(1-\gamma)||q^* - q^\pi||_\infty \leq 2\gamma||q^* - q||_\infty$$

Una perspectiva



Para la otra vez…

- Algunas aplicaciones