Predicción On-Policy con Aproximación de Funciones

Fernando Lozano

Universidad de los Andes

18 de abril de 2023

Aproximando v_π

Aproximando v_π

- Aproximar v_π con datos generados por π (on policy).

Aproximando v_π

- Aproximar v_π con datos generados por π (on policy).
- Aproximador es función paramétrica $\hat{v}(s, \mathbf{w})$

Aproximando v_π

- Aproximar v_π con datos generados por π (on policy).
- Aproximador es función paramétrica $\hat{v}(s, \mathbf{w})$
  - Función lineal $\mathbf{w}^T\mathbf{s}$.

Aproximando v_π

- Aproximar v_π con datos generados por π (on policy).
- Aproximador es función paramétrica $\hat{v}(s, \mathbf{w})$
  - Función lineal $\mathbf{w}^T\mathbf{s}$.
  - Función lineal de descriptores $\mathbf{z}$, $\mathbf{w}^T\mathbf{z}$.

Aproximando v_π

- Aproximar v_π con datos generados por π (on policy).
- Aproximador es función paramétrica $\hat{v}(s, \mathbf{w})$
  - Función lineal $\mathbf{w}^T\mathbf{s}$.
  - Función lineal de descriptores $\mathbf{z}$, $\mathbf{w}^T\mathbf{z}$.
  - Red neuronal con pesos $\mathbf{w}$.

Aproximando v_π

- Aproximar v_π con datos generados por π (on policy).
- Aproximador es función paramétrica $\hat{v}(s, \mathbf{w})$
  - Función lineal $\mathbf{w}^T\mathbf{s}$.
  - Función lineal de descriptores $\mathbf{z}$, $\mathbf{w}^T\mathbf{z}$.
  - Red neuronal con pesos $\mathbf{w}$.
  - Kernels, árboles de decisión, ...

Aproximando v_π

- Aproximar v_π con datos generados por π (on policy).
- Aproximador es función paramétrica $\hat{v}(s, \mathbf{w})$
  - Función lineal $\mathbf{w}^T\mathbf{s}$.
  - Función lineal de descriptores $\mathbf{z}$, $\mathbf{w}^T\mathbf{z}$.
  - Red neuronal con pesos $\mathbf{w}$.
  - Kernels, árboles de decisión, ...
- Ajustar parámetros $\mathbf{w}$ para que $v_\pi(s) \approx \hat{v}(s, \mathbf{w})$:

Aproximando v_π

- Aproximar v_π con datos generados por π (on policy).
- Aproximador es función paramétrica $\hat{v}(s, \mathbf{w})$
  - Función lineal $\mathbf{w}^T\mathbf{s}$.
  - Función lineal de descriptores $\mathbf{z}$, $\mathbf{w}^T\mathbf{z}$.
  - Red neuronal con pesos $\mathbf{w}$.
  - Kernels, árboles de decisión, ...
- Ajustar parámetros $\mathbf{w}$ para que $v_\pi(s) \approx \hat{v}(s, \mathbf{w})$:
  - Dimensión de $\mathbf{w} \lll |\mathcal{S}|$.

Aproximando v_π

- Aproximar v_π con datos generados por π (on policy).
- Aproximador es función paramétrica $\hat{v}(s, \mathbf{w})$
  - Función lineal $\mathbf{w}^T\mathbf{s}$.
  - Función lineal de descriptores $\mathbf{z}$, $\mathbf{w}^T\mathbf{z}$.
  - Red neuronal con pesos $\mathbf{w}$.
  - Kernels, árboles de decisión, ...
- Ajustar parámetros $\mathbf{w}$ para que $v_\pi(s) \approx \hat{v}(s, \mathbf{w})$:
  - Dimensión de $\mathbf{w} \lll |\mathcal{S}|$.
  - Generalización.

Aproximando v_π

- Aproximar v_π con datos generados por π (on policy).
- Aproximador es función paramétrica $\hat{v}(s, \mathbf{w})$
  - Función lineal $\mathbf{w}^T\mathbf{s}$.
  - Función lineal de descriptores $\mathbf{z}$, $\mathbf{w}^T\mathbf{z}$.
  - Red neuronal con pesos $\mathbf{w}$.
  - Kernels, árboles de decisión, ...
- Ajustar parámetros $\mathbf{w}$ para que $v_\pi(s) \approx \hat{v}(s, \mathbf{w})$:
  - Dimensión de $\mathbf{w} \lll |\mathcal{S}|$.
  - Generalización.
  - Ajustar pesos cambia $\hat{v}(s, \mathbf{w})$ para muchos estados.

Aprendizaje Supervisado vs. aproximación de v_π

Aprendizaje Supervisado vs. aproximación de v_π

- Aprendizaje supervisado:
  - Datos i.i.d. $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ donde se asume que $y_i \approx f(\mathbf{x}_i)$.

Aprendizaje Supervisado vs. aproximación de v_π

- Aprendizaje supervisado:
  - Datos i.i.d. $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ donde se asume que $y_i \approx f(\mathbf{x}_i)$.
  - Aprendizaje ajusta $\mathbf{w}_k$ a $\mathbf{w}_{k+1}$ de manera que $\hat{f}(\mathbf{x}_i, \mathbf{w}_{k+1}) \mapsto y_i$.

Aprendizaje Supervisado vs. aproximación de v_π

- Aprendizaje supervisado:
  - Datos i.i.d. $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ donde se asume que $y_i \approx f(\mathbf{x}_i)$.
  - Aprendizaje ajusta $\mathbf{w}_k$ a $\mathbf{w}_{k+1}$ de manera que $\hat{f}(\mathbf{x}_i, \mathbf{w}_{k+1}) \mapsto y_i$.
- Al aproximar $v_\pi(s)$, queremos ajustar $\mathbf{w}$ para que $\hat{v}(s, \mathbf{w}) \mapsto v_\pi(s)$.

Aprendizaje Supervisado vs. aproximación de v_π

- Aprendizaje supervisado:
  - Datos i.i.d. $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ donde se asume que $y_i \approx f(\mathbf{x}_i)$.
  - Aprendizaje ajusta $\mathbf{w}_k$ a $\mathbf{w}_{k+1}$ de manera que $\hat{f}(\mathbf{x}_i, \mathbf{w}_{k+1}) \mapsto y_i$.
- Al aproximar $v_\pi(s)$, queremos ajustar $\mathbf{w}$ para que $\hat{v}(s, \mathbf{w}) \mapsto v_\pi(s)$.
- No conocemos $v_\pi(s)$, sino valor estimado.

Aprendizaje Supervisado vs. aproximación de v_π

- Aprendizaje supervisado:
  - Datos i.i.d. $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ donde se asume que $y_i \approx f(\mathbf{x}_i)$.
  - Aprendizaje ajusta $\mathbf{w}_k$ a $\mathbf{w}_{k+1}$ de manera que $\hat{f}(\mathbf{x}_i, \mathbf{w}_{k+1}) \mapsto y_i$.
- Al aproximar $v_\pi(s)$, queremos ajustar $\mathbf{w}$ para que $\hat{v}(s, \mathbf{w}) \mapsto v_\pi(s)$.
- No conocemos $v_\pi(s)$, sino valor estimado.
  - DP: $\mathbb{E}_\pi \left[R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}) \mid S_t = s \right]$

Aprendizaje Supervisado vs. aproximación de v_π

- Aprendizaje supervisado:
  - Datos i.i.d. $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ donde se asume que $y_i \approx f(\mathbf{x}_i)$.
  - Aprendizaje ajusta $\mathbf{w}_k$ a $\mathbf{w}_{k+1}$ de manera que $\hat{f}(\mathbf{x}_i, \mathbf{w}_{k+1}) \mapsto y_i$.
- Al aproximar $v_\pi(s)$, queremos ajustar $\mathbf{w}$ para que $\hat{v}(s, \mathbf{w}) \mapsto v_\pi(s)$.
- No conocemos $v_\pi(s)$, sino valor estimado.
  - DP: $\mathbb{E}_\pi \left[R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}) \mid S_t = s \right]$
  - Montecarlo: G_t.

Aprendizaje Supervisado vs. aproximación de v_π

- Aprendizaje supervisado:
  - Datos i.i.d. $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ donde se asume que $y_i \approx f(\mathbf{x}_i)$.
  - Aprendizaje ajusta $\mathbf{w}_k$ a $\mathbf{w}_{k+1}$ de manera que $\hat{f}(\mathbf{x}_i, \mathbf{w}_{k+1}) \mapsto y_i$.
- Al aproximar $v_\pi(s)$, queremos ajustar $\mathbf{w}$ para que $\hat{v}(s, \mathbf{w}) \mapsto v_\pi(s)$.
- No conocemos $v_\pi(s)$, sino valor estimado.
  - DP: $\mathbb{E}_\pi \left[R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}) \mid S_t = s \right]$
  - Montecarlo: G_t.
  - TD(0): $R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w})$

Aprendizaje Supervisado vs. aproximación de v_π

- Aprendizaje supervisado:
  - Datos i.i.d. $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ donde se asume que $y_i \approx f(\mathbf{x}_i)$.
  - Aprendizaje ajusta $\mathbf{w}_k$ a $\mathbf{w}_{k+1}$ de manera que $\hat{f}(\mathbf{x}_i, \mathbf{w}_{k+1}) \mapsto y_i$.
- Al aproximar $v_\pi(s)$, queremos ajustar $\mathbf{w}$ para que $\hat{v}(s, \mathbf{w}) \mapsto v_\pi(s)$.
- No conocemos $v_\pi(s)$, sino valor estimado.
  - DP: $\mathbb{E}_\pi \left[R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}) \mid S_t = s \right]$
  - Montecarlo: G_t.
  - TD(0): $R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w})$
  - TD(n): $G_{t:t+n}$

Aprendizaje Supervisado vs. aproximación de v_π

- Aprendizaje supervisado:
  - Datos i.i.d. $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ donde se asume que $y_i \approx f(\mathbf{x}_i)$.
  - Aprendizaje ajusta $\mathbf{w}_k$ a $\mathbf{w}_{k+1}$ de manera que $\hat{f}(\mathbf{x}_i, \mathbf{w}_{k+1}) \mapsto y_i$.
- Al aproximar $v_\pi(s)$, queremos ajustar $\mathbf{w}$ para que $\hat{v}(s, \mathbf{w}) \mapsto v_\pi(s)$.
- No conocemos $v_\pi(s)$, sino valor estimado.
  - DP: $\mathbb{E}_\pi \left[R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}) \mid S_t = s \right]$
  - Montecarlo: G_t.
  - TD(0): $R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w})$
  - TD(n): $G_{t:t+n}$
- Datos generados incrementalmente

Aprendizaje Supervisado vs. aproximación de v_π

- Aprendizaje supervisado:
  - Datos i.i.d. $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ donde se asume que $y_i \approx f(\mathbf{x}_i)$.
  - Aprendizaje ajusta $\mathbf{w}_k$ a $\mathbf{w}_{k+1}$ de manera que $\hat{f}(\mathbf{x}_i, \mathbf{w}_{k+1}) \mapsto y_i$.
- Al aproximar $v_\pi(s)$, queremos ajustar $\mathbf{w}$ para que $\hat{v}(s, \mathbf{w}) \mapsto v_\pi(s)$.
- No conocemos $v_\pi(s)$, sino valor estimado.
  - DP: $\mathbb{E}_\pi \left[R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}) \mid S_t = s \right]$
  - Montecarlo: G_t.
  - TD(0): $R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w})$
  - TD(n): $G_{t:t+n}$
- Datos generados incrementalmente→ Aprendizaje en línea.

Aprendizaje Supervisado vs. aproximación de v_π

- Aprendizaje supervisado:
  - Datos i.i.d. $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ donde se asume que $y_i \approx f(\mathbf{x}_i)$.
  - Aprendizaje ajusta $\mathbf{w}_k$ a $\mathbf{w}_{k+1}$ de manera que $\hat{f}(\mathbf{x}_i, \mathbf{w}_{k+1}) \mapsto y_i$.
- Al aproximar $v_\pi(s)$, queremos ajustar $\mathbf{w}$ para que $\hat{v}(s, \mathbf{w}) \mapsto v_\pi(s)$.
- No conocemos $v_\pi(s)$, sino valor estimado.
  - DP: $\mathbb{E}_\pi \left[R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}) \mid S_t = s \right]$
  - Montecarlo: G_t.
  - TD(0): $R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w})$
  - TD(n): $G_{t:t+n}$
- Datos generados incrementalmente→ Aprendizaje en línea.
- Aprender función no estacionaria.

Función Objetivo

Función Objetivo

- Error cuadrático medio de valor:

$$\overline{VE}(\mathbf{w}) \doteq \sum_{s \in \mathcal{S}} \mu(s) \left[v_\pi(s) - \hat{v}(s, \mathbf{w}) \right]^2$$

Función Objetivo

- Error cuadrático medio de valor:

$$\overline{VE}(\mathbf{w}) \doteq \sum_{s \in \mathcal{S}} \mu(s) \left[v_\pi(s) - \hat{v}(s, \mathbf{w}) \right]^2$$

  - $\mu(s)$: importancia del estado (p.ej. estados visitados más frecuentemente).

Función Objetivo

- Error cuadrático medio de valor:

$$\overline{VE}(\mathbf{w}) \doteq \sum_{s \in \mathcal{S}} \mu(s) \left[v_\pi(s) - \hat{v}(s, \mathbf{w}) \right]^2$$

  - $\mu(s)$: importancia del estado (p.ej. estados visitados más frecuentemente).
  - $\mu(s)$ es una distribución:

Función Objetivo

- Error cuadrático medio de valor:

$$\overline{VE}(\mathbf{w}) \doteq \sum_{s \in \mathcal{S}} \mu(s) \left[v_\pi(s) - \hat{v}(s, \mathbf{w}) \right]^2$$

  - $\mu(s)$: importancia del estado (p.ej. estados visitados más frecuentemente).
  - $\mu(s)$ es una distribución: $\mu(s) \geq 0$, $\sum_{s \in \mathcal{S}} \mu(s) = 1$

Función Objetivo

- Error cuadrático medio de valor:

$$\overline{VE}(\mathbf{w}) \doteq \sum_{s \in \mathcal{S}} \mu(s) \left[v_\pi(s) - \hat{v}(s, \mathbf{w}) \right]^2$$

  - $\mu(s)$: importancia del estado (p.ej. estados visitados más frecuentemente).
  - $\mu(s)$ es una distribución: $\mu(s) \geq 0$, $\sum_{s \in \mathcal{S}} \mu(s) = 1$
  - En tareas no episódicas $\mu(s)$ es distribución estacionaria.

Función Objetivo

- Error cuadrático medio de valor:

$$\overline{VE}(\mathbf{w}) \doteq \sum_{s \in \mathcal{S}} \mu(s) \left[v_\pi(s) - \hat{v}(s, \mathbf{w}) \right]^2$$

  - $\mu(s)$: importancia del estado (p.ej. estados visitados más frecuentemente).
  - $\mu(s)$ es una distribución: $\mu(s) \geq 0$, $\sum_{s \in \mathcal{S}} \mu(s) = 1$
  - En tareas no episódicas $\mu(s)$ es distribución estacionaria.
- Minimizar $\overline{VE}(\mathbf{w})$:

Función Objetivo

- Error cuadrático medio de valor:

$$\overline{VE}(\mathbf{w}) \doteq \sum_{s \in \mathcal{S}} \mu(s) \left[v_\pi(s) - \hat{v}(s, \mathbf{w}) \right]^2$$

  - $\mu(s)$: importancia del estado (p.ej. estados visitados más frecuentemente).
  - $\mu(s)$ es una distribución: $\mu(s) \geq 0$, $\sum_{s \in \mathcal{S}} \mu(s) = 1$
  - En tareas no episódicas $\mu(s)$ es distribución estacionaria.
- Minimizar $\overline{VE}(\mathbf{w})$:
  - En general no es una función convexa $\rightarrow$ mínimo local.

Función Objetivo

- Error cuadrático medio de valor:

$$\overline{VE}(\mathbf{w}) \doteq \sum_{s \in \mathcal{S}} \mu(s) \left[v_\pi(s) - \hat{v}(s, \mathbf{w}) \right]^2$$

  - $\mu(s)$: importancia del estado (p.ej. estados visitados más frecuentemente).
  - $\mu(s)$ es una distribución: $\mu(s) \geq 0$, $\sum_{s \in \mathcal{S}} \mu(s) = 1$
  - En tareas no episódicas $\mu(s)$ es distribución estacionaria.
- Minimizar $\overline{VE}(\mathbf{w})$:
  - En general no es una función convexa $\rightarrow$ mínimo local.
  - En muchos casos no hay garantía de convergencia.

Descenso de gradiente estocástico

Descenso de gradiente estocástico

- Suponga que se generan estados con probabilidad $\mu(s)$.

Descenso de gradiente estocástico

- Suponga que se generan estados con probabilidad $\mu(s)$.
- SGD:

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t - \frac{1}{2}\alpha\nabla_{\mathbf{w}}\left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t)\right]^2$$

Descenso de gradiente estocástico

- Suponga que se generan estados con probabilidad $\mu(s)$.
- SGD:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t - \frac{1}{2}\alpha\nabla_{\mathbf{w}}\left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t)\right]^2 \\ &= \mathbf{w}_t + \alpha\left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t)\right]\nabla_{\mathbf{w}}\hat{v}(S_t, \mathbf{w}_t)\end{aligned}$$

Descenso de gradiente estocástico

- Suponga que se generan estados con probabilidad $\mu(s)$.
- SGD:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t - \frac{1}{2}\alpha\nabla_{\mathbf{w}}\left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t)\right]^2 \\ &= \mathbf{w}_t + \alpha\left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t)\right]\nabla_{\mathbf{w}}\hat{v}(S_t, \mathbf{w}_t)\end{aligned}$$

- No conocemos $v_\pi(S_t)$, reemplazando por valor estimado U_t:

Descenso de gradiente estocástico

- Suponga que se generan estados con probabilidad $\mu(s)$.
- SGD:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t - \frac{1}{2}\alpha\nabla_{\mathbf{w}}\left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t)\right]^2 \\ &= \mathbf{w}_t + \alpha\left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t)\right]\nabla_{\mathbf{w}}\hat{v}(S_t, \mathbf{w}_t)\end{aligned}$$

- No conocemos $v_\pi(S_t)$, reemplazando por valor estimado U_t:

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\left[U_t - \hat{v}(S_t, \mathbf{w}_t)\right]\nabla_{\mathbf{w}}\hat{v}(S_t, \mathbf{w}_t)$$

Descenso de gradiente estocástico

- Suponga que se generan estados con probabilidad $\mu(s)$.
- SGD:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t - \frac{1}{2}\alpha\nabla_{\mathbf{w}}\left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t)\right]^2 \\ &= \mathbf{w}_t + \alpha\left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t)\right]\nabla_{\mathbf{w}}\hat{v}(S_t, \mathbf{w}_t)\end{aligned}$$

- No conocemos $v_\pi(S_t)$, reemplazando por valor estimado U_t:

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\left[U_t - \hat{v}(S_t, \mathbf{w}_t)\right]\nabla_{\mathbf{w}}\hat{v}(S_t, \mathbf{w}_t)$$

- Si U_t es un estimativo no sesgado de $v_\pi(S_t)$ y $\sum_{t=1}^{\infty}\alpha_t = \infty$, $\sum_{t=1}^{\infty}\alpha_t^2 < \infty$, $\mathbf{w}_t$ converge a un mínimo local de $\overline{VE}(\mathbf{w})$.

Descenso de gradiente estocástico

- Suponga que se generan estados con probabilidad $\mu(s)$.
- SGD:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t - \frac{1}{2}\alpha\nabla_{\mathbf{w}}\left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t)\right]^2 \\ &= \mathbf{w}_t + \alpha\left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t)\right]\nabla_{\mathbf{w}}\hat{v}(S_t, \mathbf{w}_t)\end{aligned}$$

- No conocemos $v_\pi(S_t)$, reemplazando por valor estimado U_t:

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\left[U_t - \hat{v}(S_t, \mathbf{w}_t)\right]\nabla_{\mathbf{w}}\hat{v}(S_t, \mathbf{w}_t)$$

- Si U_t es un estimativo no sesgado de $v_\pi(S_t)$ y $\sum_{t=1}^{\infty}\alpha_t = \infty$, $\sum_{t=1}^{\infty}\alpha_t^2 < \infty$, $\mathbf{w}_t$ converge a un mínimo local de $\overline{VE}(\mathbf{w})$.
- Por ejemplo $\mathbb{E}_\pi\left[G_t \mid S_t = s\right] = v_\pi(s)$

Gradiente Monte Carlo para estimar $v_\pi(s)$

Gradiente Monte Carlo para estimar $v_\pi(s)$

Require: Política π. función diferenciable $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$. $\alpha > 0$

Incialice $\mathbf{w}$

Gradiente Monte Carlo para estimar $v_\pi(s)$

Require: Política π. función diferenciable $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$. $\alpha > 0$

Incialice $\mathbf{w}$

repeat

Gradiente Monte Carlo para estimar $v_\pi(s)$

Require: Política π. función diferenciable $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$. $\alpha > 0$

Incialice $\mathbf{w}$

repeat

Genere episodio $\pi : S_0, A_0, R_1, S_1, A_1, R_2, \ldots S_{T-1}, A_{T-1}, R_T$

Gradiente Monte Carlo para estimar $v_\pi(s)$

Require: Política π. función diferenciable $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$. $\alpha > 0$

Incialice $\mathbf{w}$

repeat

Genere episodio $\pi : S_0, A_0, R_1, S_1, A_1, R_2, \ldots S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Gradiente Monte Carlo para estimar $v_\pi(s)$

Require: Política π. función diferenciable $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$. $\alpha > 0$

Incialice $\mathbf{w}$

repeat

Genere episodio $\pi : S_0, A_0, R_1, S_1, A_1, R_2, \ldots S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

for $t = T-1, T-2, \ldots, 0$ **do**

Gradiente Monte Carlo para estimar $v_\pi(s)$

Require: Política π. función diferenciable $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$. $\alpha > 0$

Incialice $\mathbf{w}$

repeat

Genere episodio $\pi : S_0, A_0, R_1, S_1, A_1, R_2, \ldots S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

for $t = T-1, T-2, \ldots, 0$ **do**

$G \leftarrow G + R_{t+1}$

$\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[G - \hat{v}(S_t, \mathbf{w})\right] \nabla_{\mathbf{w}} \hat{v}(S_t, \mathbf{w})$

Gradiente Monte Carlo para estimar $v_\pi(s)$

Require: Política π. función diferenciable $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$. $\alpha > 0$

Incialice $\mathbf{w}$

repeat

Genere episodio $\pi : S_0, A_0, R_1, S_1, A_1, R_2, \ldots S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

for $t = T-1, T-2, \ldots, 0$ **do**

$G \leftarrow G + R_{t+1}$

$\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[G - \hat{v}(S_t, \mathbf{w})\right] \nabla_{\mathbf{w}} \hat{v}(S_t, \mathbf{w})$

end for

Gradiente Monte Carlo para estimar $v_\pi(s)$

Require: Política π. función diferenciable $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$. $\alpha > 0$

Incialice $\mathbf{w}$

repeat

Genere episodio $\pi : S_0, A_0, R_1, S_1, A_1, R_2, \ldots S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

for $t = T-1, T-2, \ldots, 0$ **do**

$G \leftarrow G + R_{t+1}$

$\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[G - \hat{v}(S_t, \mathbf{w})\right] \nabla_{\mathbf{w}} \hat{v}(S_t, \mathbf{w})$

end for

until ∞

Ejemplo: Random walk con 1000 estados

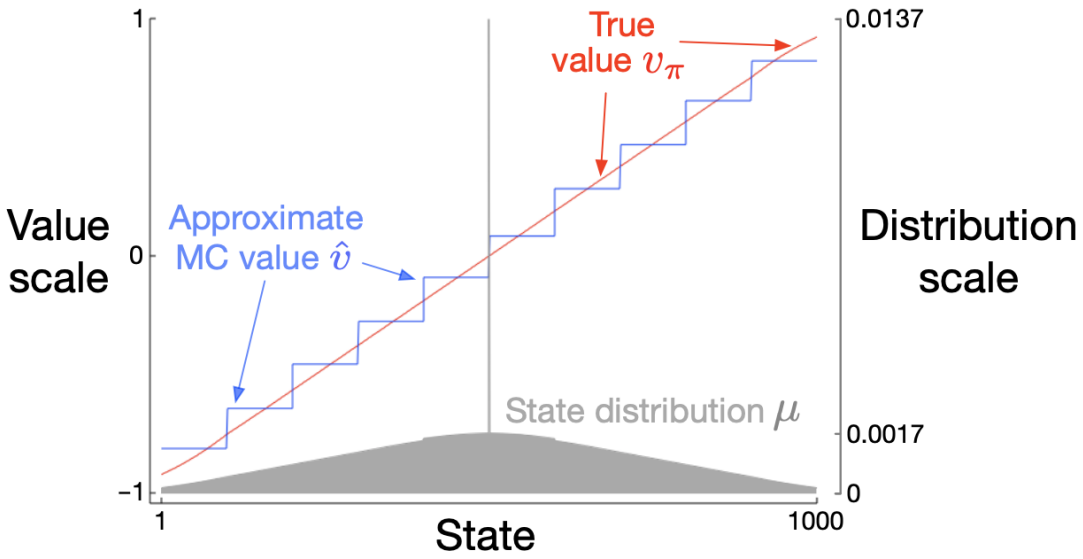


Figure 9.1: Function approximation by state aggregation on the 1000-state random walk task, using the gradient Monte Carlo algorithm (page 202).

- Agregación de a 100 estados.

Ejemplo: Random walk con 1000 estados

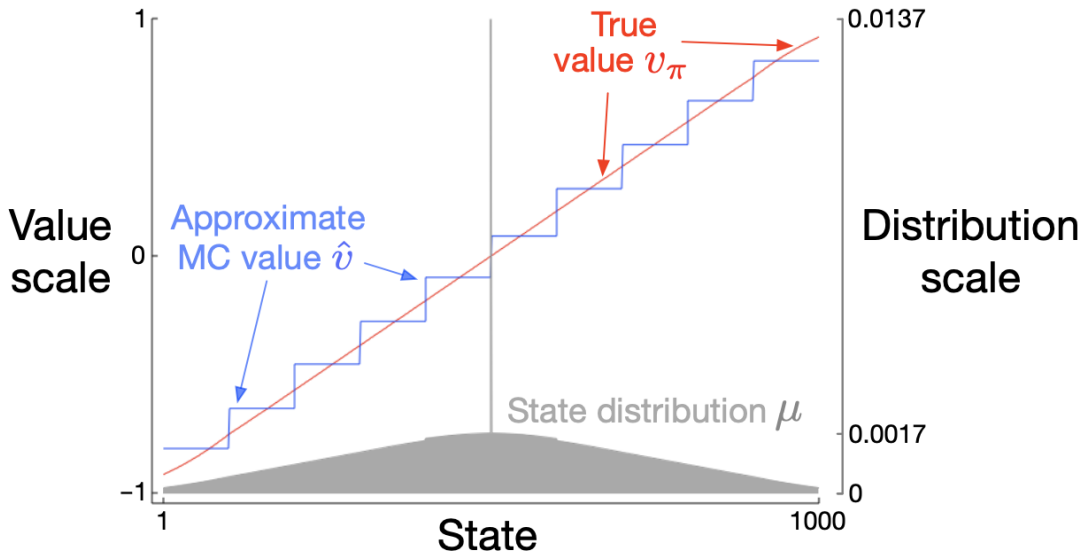


Figure 9.1: Function approximation by state aggregation on the 1000-state random walk task, using the gradient Monte Carlo algorithm (page 202).

- Agregación de a 100 estados.
- $\nabla_{\mathbf{w}} \hat{v}(S_t, \mathbf{w}) = \begin{cases} 1 & \text{Grupo estado visitado} \\ & \text{otros.} \end{cases}$,
- 100000 episodios, $\alpha = 2 \times 10^{-5}$

- Si U_t usa bootstrapping, depende de $\mathbf{w}_t$ y no es un estimativo sin sesgo de $v_\pi(s)$.

- Si U_t usa bootstrapping, depende de $\mathbf{w}_t$ y no es un estimativo sin sesgo de $v_\pi(s)$.
- En la actualización SDG:

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t - \frac{1}{2}\alpha\nabla_{\mathbf{w}}[\underbrace{U_t}_{\text{depende de } \mathbf{w}} - \hat{v}(S_t, \mathbf{w}_t)]^2$$

- Si U_t usa bootstrapping, depende de $\mathbf{w}_t$ y no es un estimativo sin sesgo de $v_\pi(s)$.
- En la actualización SDG:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t - \frac{1}{2}\alpha\nabla_{\mathbf{w}}[\underbrace{U_t}_{\text{depende de } \mathbf{w}} - \hat{v}(S_t, \mathbf{w}_t)]^2 \\ &= \mathbf{w}_t + \alpha\left[U_t - \hat{v}(S_t, \mathbf{w}_t)\right]\nabla_{\mathbf{w}}\left[\hat{v}(S_t, \mathbf{w}_t) - U_t\right]\end{aligned}$$

- Si U_t usa bootstrapping, depende de $\mathbf{w}_t$ y no es un estimativo sin sesgo de $v_\pi(s)$.
- En la actualización SDG:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t - \frac{1}{2}\alpha\nabla_{\mathbf{w}}[\underbrace{U_t}_{\text{depende de } \mathbf{w}} - \hat{v}(S_t, \mathbf{w}_t)]^2 \\ &= \mathbf{w}_t + \alpha\left[U_t - \hat{v}(S_t, \mathbf{w}_t)\right]\nabla_{\mathbf{w}}\left[\hat{v}(S_t, \mathbf{w}_t) - U_t\right]\end{aligned}$$

- Ignorar parte del gradiente: semigradiente.

Métodos de semigradiente

Métodos de semigradiente

- Aprendizaje más rápido.

Métodos de semigradiente

- Aprendizaje más rápido.
- Aprendizaje en línea, sin esperar al final del episodio.

Métodos de semigradiente

- Aprendizaje más rápido.
- Aprendizaje en línea, sin esperar al final del episodio.
- Aplicable en tareas no episódicas.

Métodos de semigradiente

- Aprendizaje más rápido.
- Aprendizaje en línea, sin esperar al final del episodio.
- Aplicable en tareas no episódicas.
- Convergencia menos robusta.

Evaluación de política con TD(0) semigradiente

Require: Política π, $\alpha \in (0, 1]$

Evaluación de política con TD(0) semigradiente

Require: Política π, $\alpha \in (0, 1]$
Require: función diferenciadle $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ con $\hat{v}(\text{terminal}, .) = 0$
Incialice $\mathbf{w}$

Evaluación de política con TD(0) semigradiente

Require: Política π, $\alpha \in (0, 1]$
Require: función diferenciadle $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ con $\hat{v}(\text{terminal}, .) = 0$
  Incialice $\mathbf{w}$
  repeat ▷ para cada episodio

Evaluación de política con TD(0) semigradiente

Require: Política π, $\alpha \in (0, 1]$
Require: función diferenciadle $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ con $\hat{v}(\text{terminal}, .) = 0$
  Incialice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S

Evaluación de política con TD(0) semigradiente

Require: Política π, $\alpha \in (0, 1]$
Require: función diferenciadle $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ con $\hat{v}(\text{terminal}, .) = 0$
  Incialice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    repeat ▷ para cada paso del episodio

Evaluación de política con TD(0) semigradiente

Require: Política π, $\alpha \in (0, 1]$
Require: función diferenciadle $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ con $\hat{v}(\text{terminal}, .) = 0$
  Incialice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    repeat ▷ para cada paso del episodio
      $A \leftarrow$ acción dada por π en S

Evaluación de política con TD(0) semigradiente

Require: Política π, $\alpha \in (0, 1]$
Require: función diferenciadle $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ con $\hat{v}(\text{terminal}, .) = 0$
  Incialice **w**
  repeat ▷ para cada episodio
    Inicialice S
    repeat ▷ para cada paso del episodio
      $A \leftarrow$ acción dada por π en S
      Tome acción A, observe R, y nuevo estado S'

Evaluación de política con TD(0) semigradiente

Require: Política π, $\alpha \in (0, 1]$
Require: función diferenciadle $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ con $\hat{v}(\text{terminal}, .) = 0$
  Incialice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    repeat ▷ para cada paso del episodio
      $A \leftarrow$ acción dada por π en S
      Tome acción A, observe R, y nuevo estado S'
      $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R_t + \gamma \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})\right] \nabla_{\mathbf{w}} \hat{v}(S, \mathbf{w})$

Evaluación de política con TD(0) semigradiente

Require: Política π, $\alpha \in (0, 1]$
Require: función diferenciadle $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ con $\hat{v}(\text{terminal}, .) = 0$

Incialice $\mathbf{w}$
repeat ▷ para cada episodio
  Inicialice S
  repeat ▷ para cada paso del episodio
    $A \leftarrow$ acción dada por π en S
    Tome acción A, observe R, y nuevo estado S'
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R_t + \gamma \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})\right] \nabla_{\mathbf{w}} \hat{v}(S, \mathbf{w})$
    $S \leftarrow S'$

Evaluación de política con TD(0) semigradiente

Require: Política π, $\alpha \in (0, 1]$
Require: función diferenciadle $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ con $\hat{v}(\text{terminal}, .) = 0$

Incialice $\mathbf{w}$
repeat ▷ para cada episodio
  Inicialice S
  repeat ▷ para cada paso del episodio
    $A \leftarrow$ acción dada por π en S
    Tome acción A, observe R, y nuevo estado S'
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R_t + \gamma \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w}) \right] \nabla_{\mathbf{w}} \hat{v}(S, \mathbf{w})$
    $S \leftarrow S'$
  until S es terminal

Evaluación de política con TD(0) semigradiente

Require: Política π, $\alpha \in (0, 1]$
Require: función diferenciadle $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ con $\hat{v}(\text{terminal}, .) = 0$
  Incialice $\mathbf{w}$
  repeat ▷ para cada episodio
    Inicialice S
    repeat ▷ para cada paso del episodio
      $A \leftarrow$ acción dada por π en S
      Tome acción A, observe R, y nuevo estado S'
      $\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R_t + \gamma \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w}) \right] \nabla_{\mathbf{w}} \hat{v}(S, \mathbf{w})$
      $S \leftarrow S'$
    until S es terminal
  until ∞

Funciones lineales (de $\mathbf{w}$)

Funciones lineales (de $\mathbf{w}$)

- Representación del estado en espacio de características:

$$\mathbf{x}(s) = \begin{bmatrix} x_1(s) & x_2(s) & \dots & x_d(s) \end{bmatrix}$$

con $x_i(s) : \mathcal{S} \to \mathbb{R}$

Funciones lineales (de $\mathbf{w}$)

- Representación del estado en espacio de características:

$$\mathbf{x}(s) = \begin{bmatrix} x_1(s) & x_2(s) & \dots & x_d(s) \end{bmatrix}$$

  con $x_i(s) \ : \ \mathcal{S} \to \mathbb{R}$

- Aproximar con función lineal de características:

$$\hat{v}(s, \mathbf{w}) \doteq \mathbf{w}^T \mathbf{x}$$

Funciones lineales (de $\mathbf{w}$)

- Representación del estado en espacio de características:

$$\mathbf{x}(s) = \begin{bmatrix} x_1(s) & x_2(s) & \dots & x_d(s) \end{bmatrix}$$

  con $x_i(s) \ : \ \mathcal{S} \to \mathbb{R}$

- Aproximar con función lineal de características:

$$\hat{v}(s, \mathbf{w}) \doteq \mathbf{w}^T \mathbf{x} \doteq \sum_{i=1}^{d} w_i x_i(s)$$

Funciones lineales (de $\mathbf{w}$)

- Representación del estado en espacio de características:

$$\mathbf{x}(s) = \begin{bmatrix} x_1(s) & x_2(s) & \dots & x_d(s) \end{bmatrix}$$

  con $x_i(s) \ : \ \mathcal{S} \to \mathbb{R}$

- Aproximar con función lineal de características:

$$\hat{v}(s, \mathbf{w}) \doteq \mathbf{w}^T \mathbf{x} \doteq \sum_{i=1}^{d} w_i x_i(s) \longrightarrow \text{funciones base}$$

Funciones lineales (de $\mathbf{w}$)

- Representación del estado en espacio de características:

$$\mathbf{x}(s) = \begin{bmatrix} x_1(s) & x_2(s) & \dots & x_d(s) \end{bmatrix}$$

  con $x_i(s) : \mathcal{S} \to \mathbb{R}$

- Aproximar con función lineal de características:

$$\hat{v}(s, \mathbf{w}) \doteq \mathbf{w}^T \mathbf{x} \doteq \sum_{i=1}^{d} w_i x_i(s) \longrightarrow \text{funciones base}$$

- SGD:

Funciones lineales (de $\mathbf{w}$)

- Representación del estado en espacio de características:

$$\mathbf{x}(s) = \begin{bmatrix} x_1(s) & x_2(s) & \dots & x_d(s) \end{bmatrix}$$

  con $x_i(s) \ : \ \mathcal{S} \to \mathbb{R}$

- Aproximar con función lineal de características:

$$\hat{v}(s, \mathbf{w}) \doteq \mathbf{w}^T \mathbf{x} \doteq \sum_{i=1}^{d} w_i x_i(s) \longrightarrow \text{funciones base}$$

- SGD:

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \left[\hat{v}(S_t, \mathbf{w}_t) - U_t \right] \nabla_{\mathbf{w}} \hat{v}(S_t, \mathbf{w}_t)$$

Funciones lineales (de $\mathbf{w}$)

- Representación del estado en espacio de características:

$$\mathbf{x}(s) = \begin{bmatrix} x_1(s) & x_2(s) & \dots & x_d(s) \end{bmatrix}$$

con $x_i(s) \ : \ \mathcal{S} \to \mathbb{R}$

- Aproximar con función lineal de características:

$$\hat{v}(s, \mathbf{w}) \doteq \mathbf{w}^T \mathbf{x} \doteq \sum_{i=1}^{d} w_i x_i(s) \longrightarrow \text{funciones base}$$

- SGD:

$$\begin{aligned} \mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha \left[\hat{v}(S_t, \mathbf{w}_t) - U_t \right] \nabla_{\mathbf{w}} \hat{v}(S_t, \mathbf{w}_t) \\ &= \mathbf{w}_t + \alpha \left[\hat{v}(S_t, \mathbf{w}_t) - U_t \right] \mathbf{x}(S_t) \end{aligned}$$

- Función $\overline{VE}(\mathbf{w})$ es convexa

- Función $\overline{VE}(\mathbf{w})$ es convexa
  - Mínimo local es mínimo global.

- Función $\overline{VE}(\mathbf{w})$ es convexa
  - Mínimo local es mínimo global.
  - Convergencia a mínimo local $\Rightarrow$ convergencia a mínimo global.

- Función $\overline{VE}(\mathbf{w})$ es convexa
  - Mínimo local es mínimo global.
  - Convergencia a mínimo local $\Rightarrow$ convergencia a mínimo global.
- En Gradiente Montecarlo con $\sum_{t=1}^{\infty} \alpha_t = \infty$, $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$, $\mathbf{w}_t$ converge al mínimo global de $\overline{VE}(\mathbf{w})$.

- Función $\overline{VE}(\mathbf{w})$ es convexa
  - Mínimo local es mínimo global.
  - Convergencia a mínimo local $\Rightarrow$ convergencia a mínimo global.
- En Gradiente Montecarlo con $\sum_{t=1}^{\infty} \alpha_t = \infty$, $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$, $\mathbf{w}_t$ converge al mínimo global de $\overline{VE}(\mathbf{w})$.

- Gradiente TD(0), con $\mathbf{x}_t \doteq \mathbf{x}(S_t)$:

- Gradiente TD(0), con $\mathbf{x}_t \doteq \mathbf{x}(S_t)$:

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^T \mathbf{x}_{t+1} - \mathbf{w}_t^T \mathbf{x}_t \right) \mathbf{x}_t$$

.

- Gradiente TD(0), con $\mathbf{x}_t \doteq \mathbf{x}(S_t)$:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^T \mathbf{x}_{t+1} - \mathbf{w}_t^T \mathbf{x}_t\right) \mathbf{x}_t \\ &= \mathbf{w}_t + \alpha \left(R_{t+1}\mathbf{x}_t - \mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T \mathbf{w}_t\right)\end{aligned}$$

.

- Gradiente TD(0), con $\mathbf{x}_t \doteq \mathbf{x}(S_t)$:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^T \mathbf{x}_{t+1} - \mathbf{w}_t^T \mathbf{x}_t\right) \mathbf{x}_t \\ &= \mathbf{w}_t + \alpha \left(R_{t+1}\mathbf{x}_t - \mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T \mathbf{w}_t\right)\end{aligned}$$

$$\mathbb{E}\left[\mathbf{w}_{t+1} \mid \mathbf{w}_t\right] = \mathbf{w}_t + \alpha \left(\mathbb{E}\left[R_{t+1}\mathbf{x}_t\right] - \mathbb{E}\left[\mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T\right] \mathbf{w}_t\right)$$

- Gradiente TD(0), con $\mathbf{x}_t \doteq \mathbf{x}(S_t)$:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^T \mathbf{x}_{t+1} - \mathbf{w}_t^T \mathbf{x}_t\right) \mathbf{x}_t \\ &= \mathbf{w}_t + \alpha \left(R_{t+1}\mathbf{x}_t - \mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T \mathbf{w}_t\right)\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\mathbf{w}_{t+1} \mid \mathbf{w}_t\right] &= \mathbf{w}_t + \alpha \left(\mathbb{E}\left[R_{t+1}\mathbf{x}_t\right] - \mathbb{E}\left[\mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T\right] \mathbf{w}_t\right) \\ &= \mathbf{w}_t + \alpha \left(\mathbf{b} - \mathbf{A}\mathbf{w}_t\right)\end{aligned}$$

- Gradiente TD(0), con $\mathbf{x}_t \doteq \mathbf{x}(S_t)$:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^T \mathbf{x}_{t+1} - \mathbf{w}_t^T \mathbf{x}_t\right) \mathbf{x}_t \\ &= \mathbf{w}_t + \alpha \left(R_{t+1}\mathbf{x}_t - \mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T \mathbf{w}_t\right)\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\mathbf{w}_{t+1} \mid \mathbf{w}_t\right] &= \mathbf{w}_t + \alpha \left(\mathbb{E}\left[R_{t+1}\mathbf{x}_t\right] - \mathbb{E}\left[\mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T\right] \mathbf{w}_t\right) \\ &= \mathbf{w}_t + \alpha \left(\mathbf{b} - \mathbf{A}\mathbf{w}_t\right) \\ &= (\mathbf{I} - \alpha \mathbf{A})\mathbf{w}_t + \alpha \mathbf{b}\end{aligned}$$

- Gradiente TD(0), con $\mathbf{x}_t \doteq \mathbf{x}(S_t)$:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^T \mathbf{x}_{t+1} - \mathbf{w}_t^T \mathbf{x}_t\right) \mathbf{x}_t \\ &= \mathbf{w}_t + \alpha \left(R_{t+1}\mathbf{x}_t - \mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T \mathbf{w}_t\right)\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\mathbf{w}_{t+1} \mid \mathbf{w}_t\right] &= \mathbf{w}_t + \alpha \left(\mathbb{E}\left[R_{t+1}\mathbf{x}_t\right] - \mathbb{E}\left[\mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T\right] \mathbf{w}_t\right) \\ &= \mathbf{w}_t + \alpha \left(\mathbf{b} - \mathbf{A}\mathbf{w}_t\right) \\ &= (\mathbf{I} - \alpha \mathbf{A})\mathbf{w}_t + \alpha \mathbf{b}\end{aligned}$$

  - Convergencia (punto fijo de TD):

$$\mathbf{A}\mathbf{w} = \mathbf{b} \Rightarrow \mathbf{w}_{\text{TD}} = \mathbf{A}^{-1}\mathbf{b}$$

.

- Gradiente TD(0), con $\mathbf{x}_t \doteq \mathbf{x}(S_t)$:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^T \mathbf{x}_{t+1} - \mathbf{w}_t^T \mathbf{x}_t\right) \mathbf{x}_t \\ &= \mathbf{w}_t + \alpha \left(R_{t+1}\mathbf{x}_t - \mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T \mathbf{w}_t\right)\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\mathbf{w}_{t+1} \mid \mathbf{w}_t\right] &= \mathbf{w}_t + \alpha \left(\mathbb{E}\left[R_{t+1}\mathbf{x}_t\right] - \mathbb{E}\left[\mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T\right] \mathbf{w}_t\right) \\ &= \mathbf{w}_t + \alpha \left(\mathbf{b} - \mathbf{A}\mathbf{w}_t\right) \\ &= (\mathbf{I} - \alpha \mathbf{A})\mathbf{w}_t + \alpha \mathbf{b}\end{aligned}$$

  - Convergencia (punto fijo de TD):

$$\mathbf{A}\mathbf{w} = \mathbf{b} \Rightarrow \mathbf{w}_{\text{TD}} = \mathbf{A}^{-1}\mathbf{b}$$

  - $\mathbf{A}$ depende de $p(s'|s,a)$, π, $\mu(s)$ estacionaria, γ.

- Gradiente TD(0), con $\mathbf{x}_t \doteq \mathbf{x}(S_t)$:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^T \mathbf{x}_{t+1} - \mathbf{w}_t^T \mathbf{x}_t\right) \mathbf{x}_t \\ &= \mathbf{w}_t + \alpha \left(R_{t+1}\mathbf{x}_t - \mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T \mathbf{w}_t\right)\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\mathbf{w}_{t+1} \mid \mathbf{w}_t\right] &= \mathbf{w}_t + \alpha \left(\mathbb{E}\left[R_{t+1}\mathbf{x}_t\right] - \mathbb{E}\left[\mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T\right] \mathbf{w}_t\right) \\ &= \mathbf{w}_t + \alpha \left(\mathbf{b} - \mathbf{A}\mathbf{w}_t\right) \\ &= (\mathbf{I} - \alpha \mathbf{A})\mathbf{w}_t + \alpha \mathbf{b}\end{aligned}$$

  - Convergencia (punto fijo de TD):

$$\mathbf{A}\mathbf{w} = \mathbf{b} \Rightarrow \mathbf{w}_{\text{TD}} = \mathbf{A}^{-1}\mathbf{b}$$

  - $\mathbf{A}$ depende de $p(s'|s,a)$, π, $\mu(s)$ estacionaria, γ.
  - $\mathbf{A}$ es positiva definida

- Gradiente TD(0), con $\mathbf{x}_t \doteq \mathbf{x}(S_t)$:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha\left(R_{t+1} + \gamma \mathbf{w}_t^T \mathbf{x}_{t+1} - \mathbf{w}_t^T \mathbf{x}_t\right)\mathbf{x}_t \\ &= \mathbf{w}_t + \alpha\left(R_{t+1}\mathbf{x}_t - \mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T \mathbf{w}_t\right)\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\mathbf{w}_{t+1} \mid \mathbf{w}_t\right] &= \mathbf{w}_t + \alpha\left(\mathbb{E}\left[R_{t+1}\mathbf{x}_t\right] - \mathbb{E}\left[\mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T\right]\mathbf{w}_t\right) \\ &= \mathbf{w}_t + \alpha\left(\mathbf{b} - \mathbf{A}\mathbf{w}_t\right) \\ &= (\mathbf{I} - \alpha\mathbf{A})\mathbf{w}_t + \alpha\mathbf{b}\end{aligned}$$

  - Convergencia (punto fijo de TD):

$$\mathbf{A}\mathbf{w} = \mathbf{b} \Rightarrow \mathbf{w}_{\text{TD}} = \mathbf{A}^{-1}\mathbf{b}$$

  - $\mathbf{A}$ depende de $p(s'|s,a)$, π, $\mu(s)$ estacionaria, γ.
  - $\mathbf{A}$ es positiva definida $\Rightarrow$si $\alpha \ll \Rightarrow$ valores propios de $(\mathbf{I} - \alpha\mathbf{A})$ son <1

- Gradiente TD(0), con $\mathbf{x}_t \doteq \mathbf{x}(S_t)$:

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha\left(R_{t+1} + \gamma \mathbf{w}_t^T \mathbf{x}_{t+1} - \mathbf{w}_t^T \mathbf{x}_t\right)\mathbf{x}_t \\ &= \mathbf{w}_t + \alpha\left(R_{t+1}\mathbf{x}_t - \mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T \mathbf{w}_t\right)\end{aligned}$$

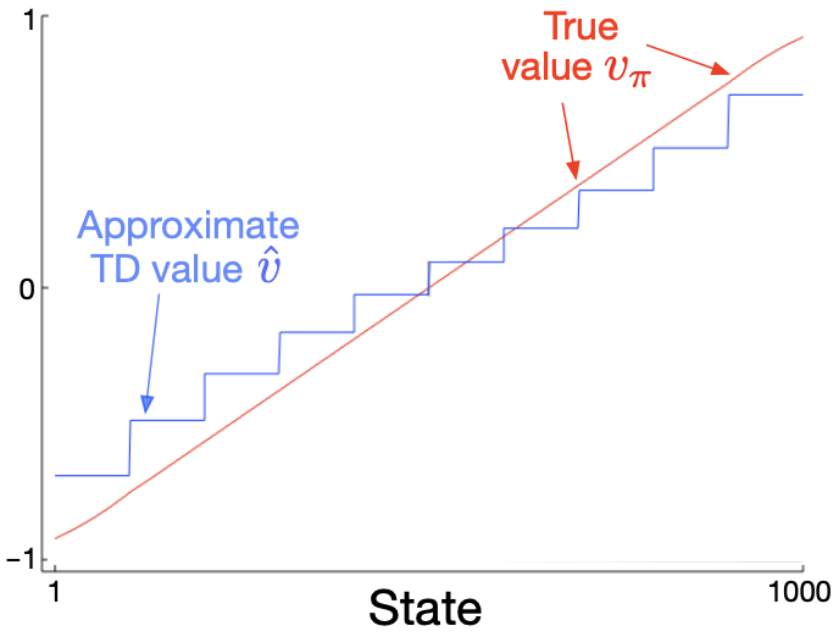
$$\begin{aligned}\mathbb{E}\left[\mathbf{w}_{t+1} \mid \mathbf{w}_t\right] &= \mathbf{w}_t + \alpha\left(\mathbb{E}\left[R_{t+1}\mathbf{x}_t\right] - \mathbb{E}\left[\mathbf{x}_t(\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^T\right]\mathbf{w}_t\right) \\ &= \mathbf{w}_t + \alpha\left(\mathbf{b} - \mathbf{A}\mathbf{w}_t\right) \\ &= (\mathbf{I} - \alpha\mathbf{A})\mathbf{w}_t + \alpha\mathbf{b}\end{aligned}$$

  - Convergencia (punto fijo de TD):

$$\mathbf{A}\mathbf{w} = \mathbf{b} \Rightarrow \mathbf{w}_{\text{TD}} = \mathbf{A}^{-1}\mathbf{b}$$

  - $\mathbf{A}$ depende de $p(s'|s,a)$, π, $\mu(s)$ estacionaria, γ.
  - $\mathbf{A}$ es positiva definida $\Rightarrow$si $\alpha \ll \Rightarrow$ valores propios de $(\mathbf{I} - \alpha\mathbf{A})$ son <1
  -

$$\overline{VE}(\mathbf{w}_{TD}) \leq \frac{1}{1-\gamma} \min_{\mathbf{w}} \overline{VE}(\mathbf{w})$$


1
0
−1
1
1000
State
True
value v_π
Approximate
TD value $\hat{v}$

Polinomios

Polinomios

- Estado $\mathbf{s} = \begin{bmatrix} s_1 & s_2 & \dots & s_d \end{bmatrix}$.

Polinomios

- Estado $\mathbf{s} = \begin{bmatrix} s_1 & s_2 & \dots & s_d \end{bmatrix}$.
- Polinomio de orden 2 (en 2 dimensiones):

$$\mathbf{x}(\mathbf{s}) = \begin{bmatrix} s_1^2 & s_2^2 & s_1 s_2 & 1 \end{bmatrix}$$

Polinomios

- Estado $\mathbf{s} = \begin{bmatrix} s_1 & s_2 & \dots & s_d \end{bmatrix}$.
- Polinomio de orden 2 (en 2 dimensiones):

$$\mathbf{x}(\mathbf{s}) = \begin{bmatrix} s_1^2 & s_2^2 & s_1 s_2 & 1 \end{bmatrix}$$

- Polinomio de orden 2 (en d dimensiones):

$$\mathbf{x}(\mathbf{s}) = \Big[\; s_1^2 \quad s_2^2 \quad \dots \quad s_d^2$$
$$s_1 s_2 \quad s_1 s_3 \quad \dots \quad s_2 s_3 \quad \dots \quad s_{d-1} s_d$$
$$s_1 \quad s_2 \quad \dots \quad s_d \quad 1 \; \Big]$$

Polinomios

- Estado $\mathbf{s} = \begin{bmatrix} s_1 & s_2 & \dots & s_d \end{bmatrix}$.
- Polinomio de orden 2 (en 2 dimensiones):

$$\mathbf{x}(\mathbf{s}) = \begin{bmatrix} s_1^2 & s_2^2 & s_1 s_2 & 1 \end{bmatrix}$$

- Polinomio de orden 2 (en d dimensiones):

$$\mathbf{x}(\mathbf{s}) = \Big[\; s_1^2 \quad s_2^2 \quad \dots \quad s_d^2$$
$$s_1 s_2 \quad s_1 s_3 \quad \dots \quad s_2 s_3 \quad \dots \quad s_{d-1} s_d$$
$$s_1 \quad s_2 \quad \dots \quad s_d \quad 1 \; \Big]$$

$\binom{d+2}{2}$ términos!

Polinomios

- Estado $\mathbf{s} = \begin{bmatrix} s_1 & s_2 & \dots & s_d \end{bmatrix}$.
- Polinomio de orden 2 (en 2 dimensiones):

$$\mathbf{x}(\mathbf{s}) = \begin{bmatrix} s_1^2 & s_2^2 & s_1 s_2 & 1 \end{bmatrix}$$

- Polinomio de orden 2 (en d dimensiones):

$$\mathbf{x}(\mathbf{s}) = \Big[\; s_1^2 \quad s_2^2 \quad \dots \quad s_d^2 \quad s_1 s_2 \quad s_1 s_3 \quad \dots \quad s_2 s_3 \quad \dots \quad s_{d-1} s_d \quad s_1 \quad s_2 \quad \dots \quad s_d \quad 1 \; \Big]$$

  $\binom{d+2}{2}$ términos!
- Polinomios de orden superior

Polinomios

- Estado $\mathbf{s} = \begin{bmatrix} s_1 & s_2 & \dots & s_d \end{bmatrix}$.
- Polinomio de orden 2 (en 2 dimensiones):

$$\mathbf{x}(\mathbf{s}) = \begin{bmatrix} s_1^2 & s_2^2 & s_1 s_2 & 1 \end{bmatrix}$$

- Polinomio de orden 2 (en d dimensiones):

$$\mathbf{x}(\mathbf{s}) = \Big[\; s_1^2 \quad s_2^2 \quad \dots \quad s_d^2 \quad s_1 s_2 \quad s_1 s_3 \quad \dots \quad s_2 s_3 \quad \dots \quad s_{d-1} s_d \quad s_1 \quad s_2 \quad \dots \quad s_d \quad 1 \; \Big]$$

  $\binom{d+2}{2}$ términos!
- Polinomios de orden superior: selección de características.

Serie de Fourier

- Representar como un período de señal periódica de período 1, par.

Serie de Fourier

- Representar como un período de señal periódica de período 1, par.
- Base de Fourier:

$$x_i(s) = \cos(i\pi s)$$

Serie de Fourier

- Representar como un período de señal periódica de período 1, par.
- Base de Fourier:

$$x_i(s) = \cos(i\pi s)$$

- Representación como serie de Fourier truncada:

$$\mathbf{w}^T \mathbf{x}(s) = \sum_{i=0}^{d-1} w_i \cos(i\pi s), \quad s \in [0, 1]$$

Serie de Fourier

- Representar como un período de señal periódica de período 1, par.
- Base de Fourier:

$$x_i(s) = \cos(i\pi s)$$

- Representación como serie de Fourier truncada:

$$\mathbf{w}^T\mathbf{x}(s) = \sum_{i=0}^{d-1} w_i \cos(i\pi s), \quad s \in [0,1]$$

- En más dimensiones:

$$x_i(s) = \cos(i\pi \mathbf{s}^T\mathbf{c}^i), \quad s \in [0,1]$$

donde $c^i_j \in \{0,\ldots,n\}$.

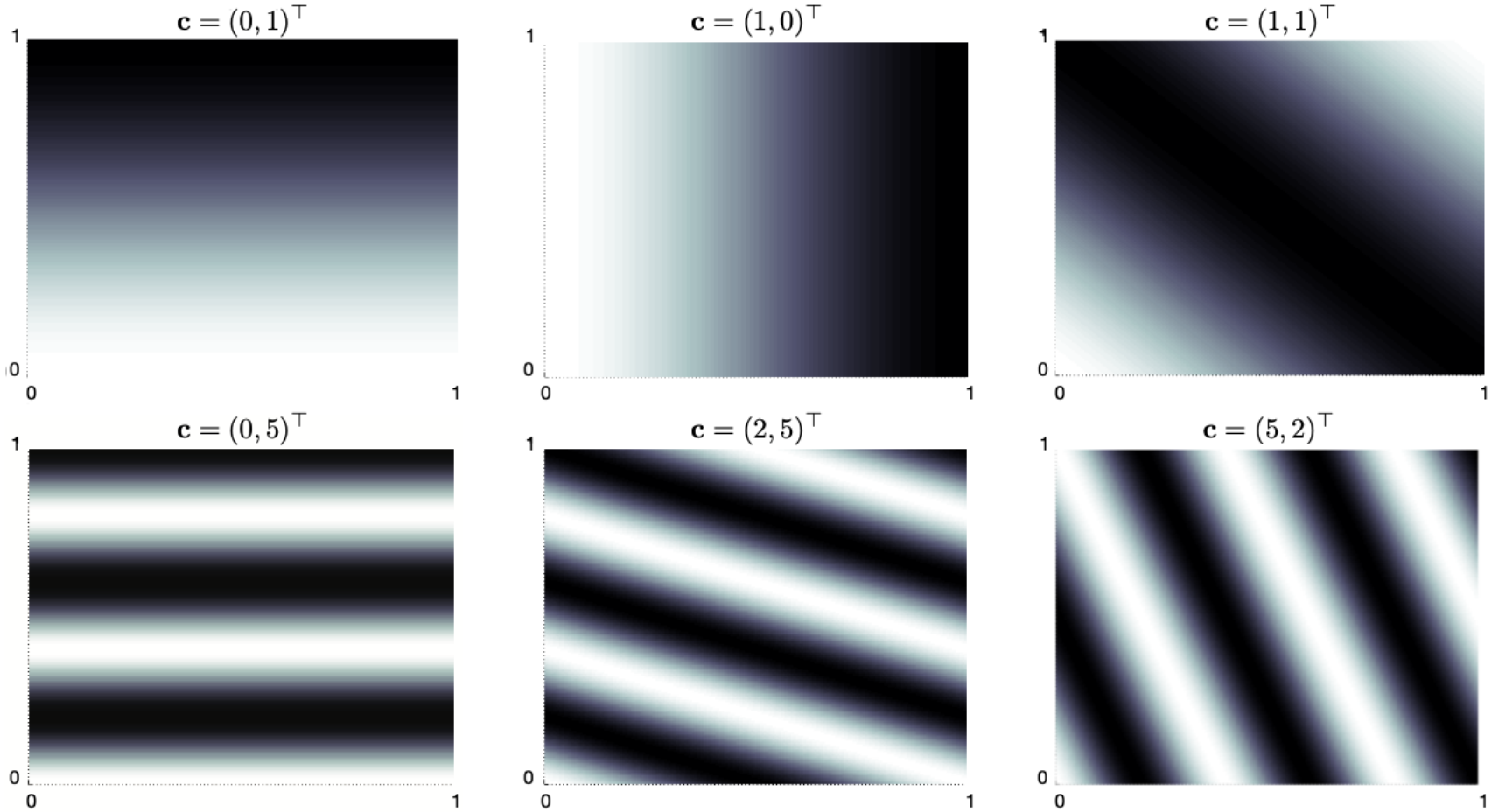
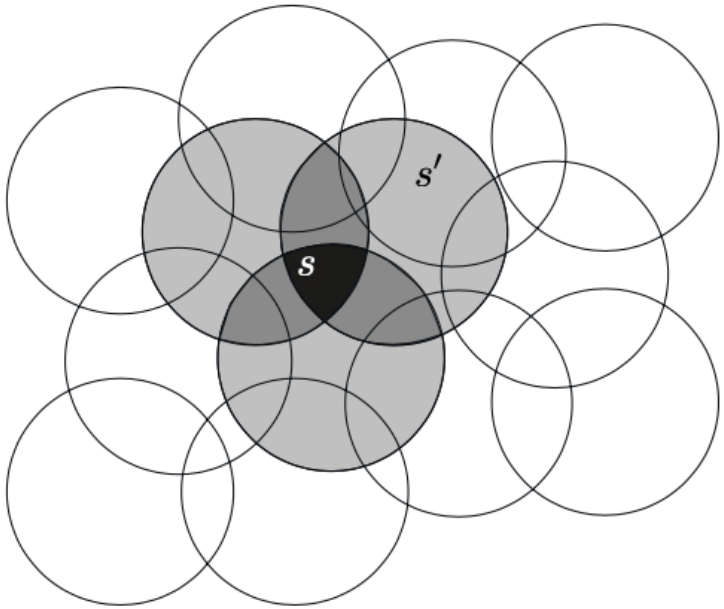
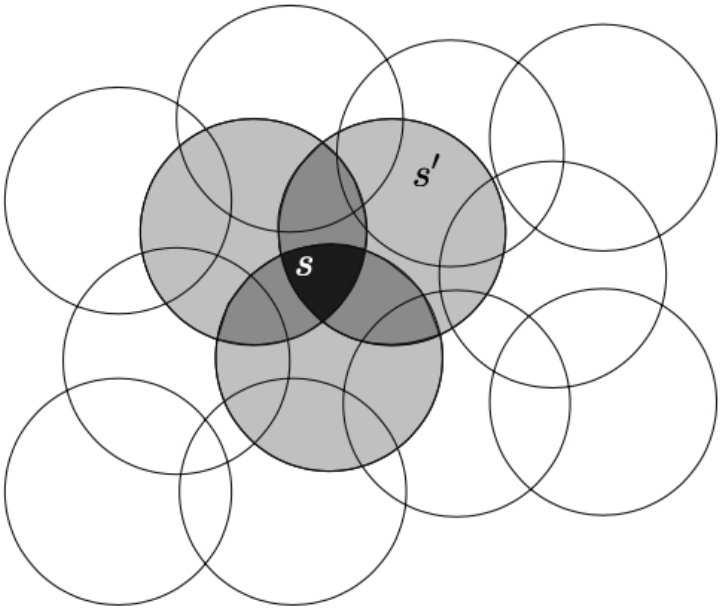


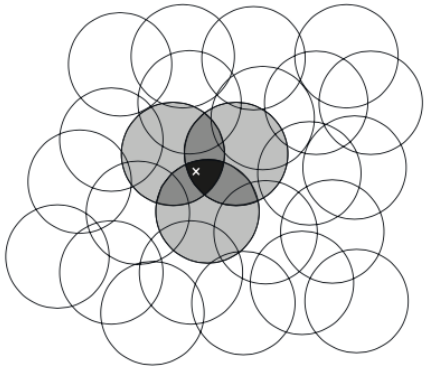
Figure 9.4: A selection of six two-dimensional Fourier cosine features, each labeled by the vector $\mathbf{c}^i$ that defines it (s_1 is the horizontal axis, and $\mathbf{c}^i$ is shown with the index i omitted). After Konidaris et al. (2011).

Codificación gruesa (Coarse coding)

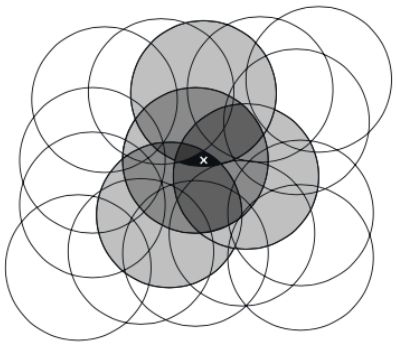


Codificación gruesa (Coarse coding)

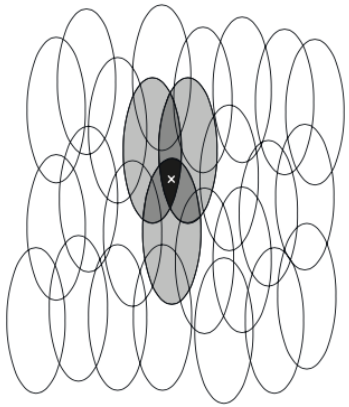




Narrow generalization



Broad generalization



Asymmetric generalization

Codificación con baldosas (Tile Coding)

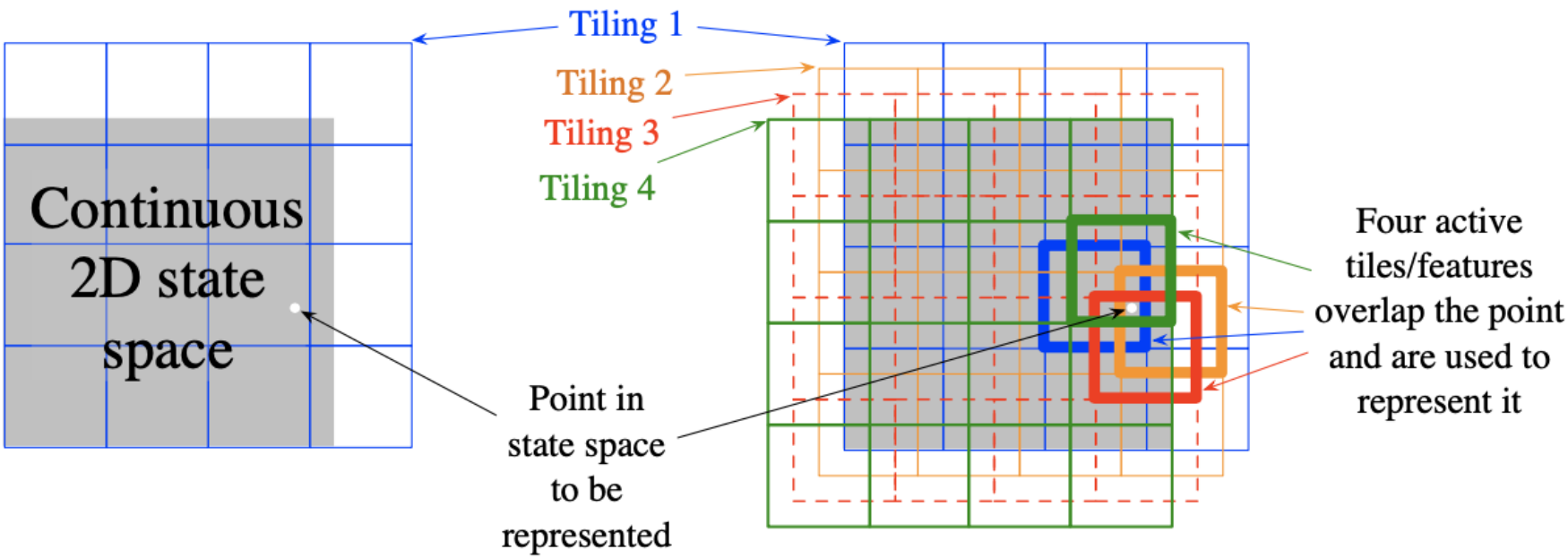


Figure 9.9: Multiple, overlapping grid-tilings on a limited two-dimensional space. These tilings are offset from one another by a uniform amount in each dimension.

Funciones Base Radiales (RBFs)

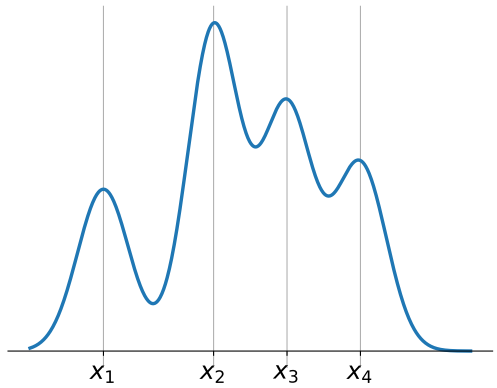
Funciones Base Radiales (RBFs)

- Función decreciente de distancia a un prototipo:

Funciones Base Radiales (RBFs)

- Función decreciente de distancia a un prototipo:

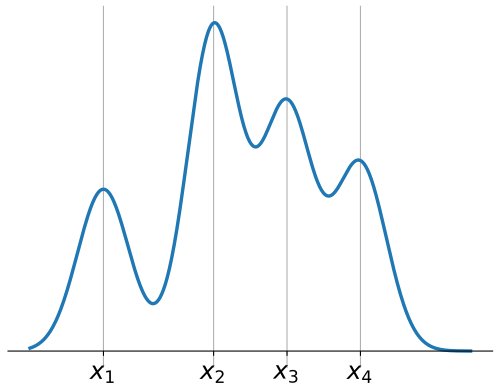
$$x_i(s) = \exp\left(-\frac{\|s - c_i\|^2}{2\sigma_i}\right)$$

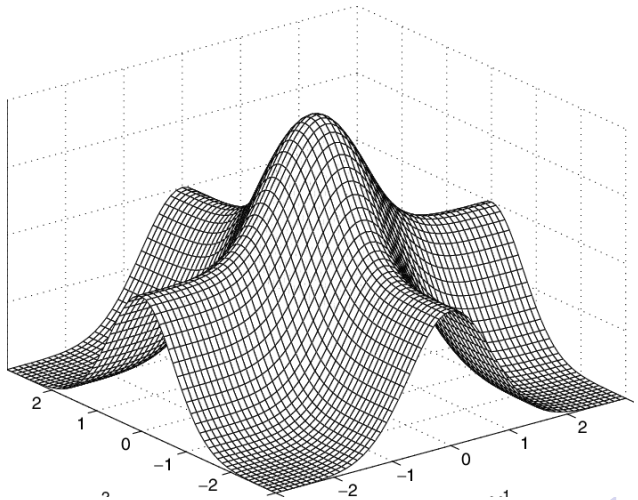


Funciones Base Radiales (RBFs)

- Función decreciente de distancia a un prototipo:

$$x_i(s) = \exp\left(-\frac{\|s - c_i\|^2}{2\sigma_i}\right)$$





Lineal vs. no lineal

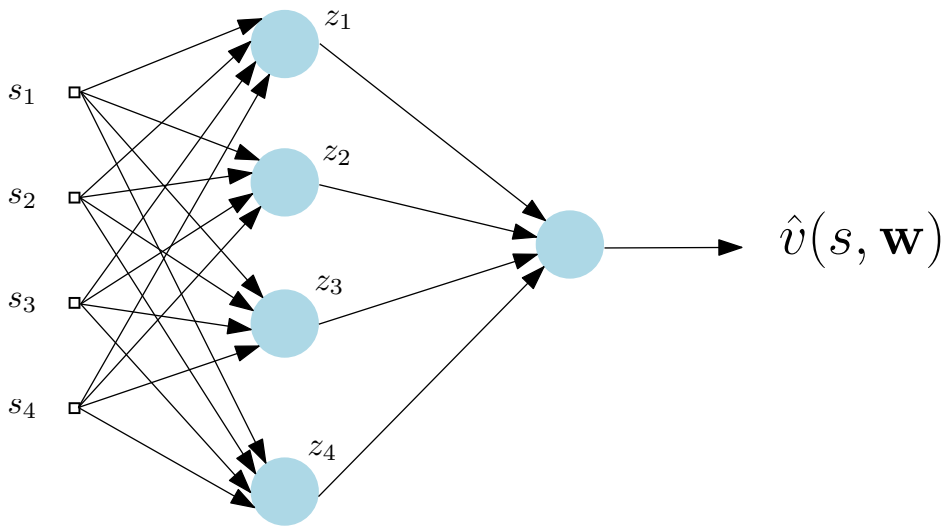
$$\hat{v}(s, \mathbf{w}) = \mathbf{w}^T \mathbf{x}$$

Lineal vs. no lineal

$$\hat{v}(s, \mathbf{w}) = \mathbf{w}^T \mathbf{x} \doteq \sum_{i=1}^{d} w_i x_i(s) \longrightarrow \text{lineal}$$

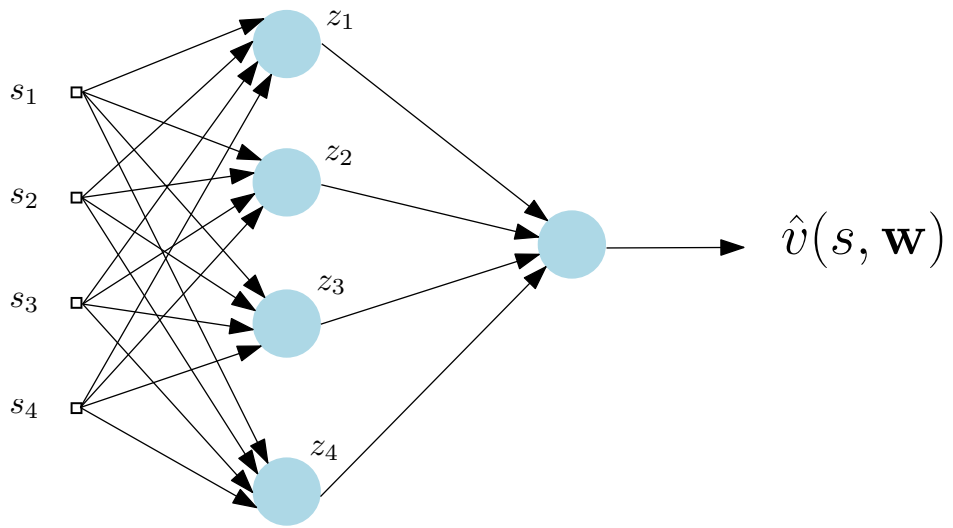
Lineal vs. no lineal

$$\hat{v}(s, \mathbf{w}) = \mathbf{w}^T \mathbf{x} \doteq \sum_{i=1}^{d} w_i x_i(s) \longrightarrow \text{lineal}$$



Lineal vs. no lineal

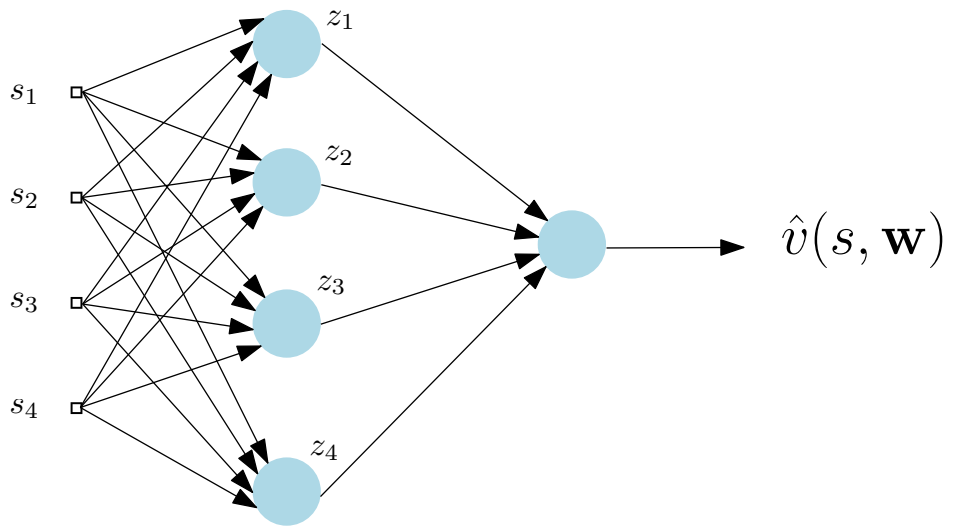
$$\hat{v}(s, \mathbf{w}) = \mathbf{w}^T \mathbf{x} \doteq \sum_{i=1}^{d} w_i x_i(s) \longrightarrow \text{lineal}$$



$$\hat{v}(s, \mathbf{w}) = \sum_{i=1}^{d} a_i f\left(\mathbf{w}^T \mathbf{s}\right) \longrightarrow \text{no lineal}$$

Lineal vs. no lineal

$$\hat{v}(s, \mathbf{w}) = \mathbf{w}^T \mathbf{x} \doteq \sum_{i=1}^{d} w_i x_i(s) \longrightarrow \text{lineal}$$



$$\hat{v}(s, \mathbf{w}) = \sum_{i=1}^{d} a_i f\left(\mathbf{w}^T \mathbf{s}\right) \longrightarrow \text{no lineal}$$

- En aproximación no lineal se requieren exponencialmente menos parámetros para lograr error de aproximación dado.