Learning the value functions

Sébastien Gros

Cybernetic, NTNU
Elec. Eng., Chalmers

TUM lectures on RL

Outline

Recap'

Bellman equations

$$V_\pi(\mathbf{x}) = L(\mathbf{x}, \boldsymbol{\pi}(\mathbf{x})) + \gamma \mathbb{E}\left[V_\pi(\mathbf{x}_+) \,|\, \mathbf{x}, \boldsymbol{\pi}\right] \qquad V_\pi(\mathbf{x}) = Q_\pi(\mathbf{x}, \boldsymbol{\pi}(\mathbf{x}))$$
$$Q_\pi(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[V_\pi(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u}\right]$$

- Policy evaluation $V_{k+1}(\mathbf{x}) \leftarrow L(\mathbf{x}, \boldsymbol{\pi}(\mathbf{x})) + \gamma \mathbb{E}\left[V_k(\mathbf{x}_+) \,|\, \mathbf{x}, \boldsymbol{\pi}\right]$ (sweep over $\mathbf{x}$)
- Policy improvement $\boldsymbol{\pi}'(\mathbf{x}) = \arg\min_{\mathbf{u}} Q_\pi(\mathbf{x}, \mathbf{u})$
- Dynamic programming

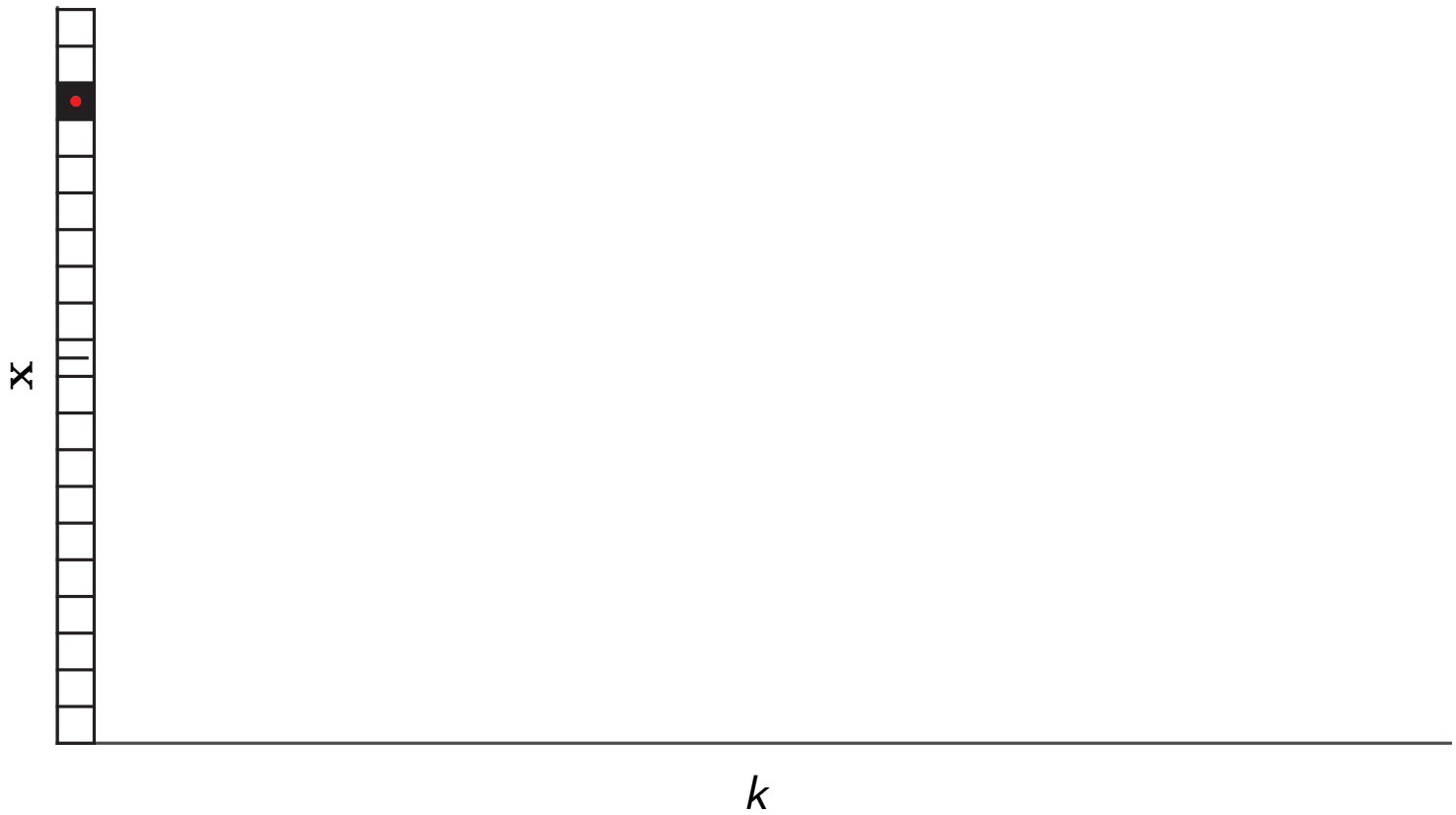
$$Q_+(\mathbf{x}, \mathbf{u}) \leftarrow L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[V(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u}\right]$$
$$V(\mathbf{x}) \leftarrow \min_{\mathbf{u}} Q_+(\mathbf{x}, \mathbf{u})$$

All these techniques require a model of the real system $\mathbb{P}\left[\mathbf{x}_+ \,|\, \mathbf{x}, \mathbf{u}\right]$. What can we do from data?

Let's use the following example (for illustration purposes)

- State $\mathbf{x} \in \mathbb{N}$ and input $\mathbf{u} \in \mathbb{N}$ locked on a grid
- Input $\mathbf{u}$ can only move $\mathbf{x}$ on the grid
- Stage cost: $L(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\mathbf{x}^2 + \frac{1}{2}\mathbf{u}^2$
- Discount $\gamma = 0.9$

Dynamics: $\mathbf{x}_+ = \mathbf{x} + \mathbf{u} + \mathrm{round}(\mathbf{e})$ where $\mathbf{e} \sim \mathcal{N}(0, 1)$

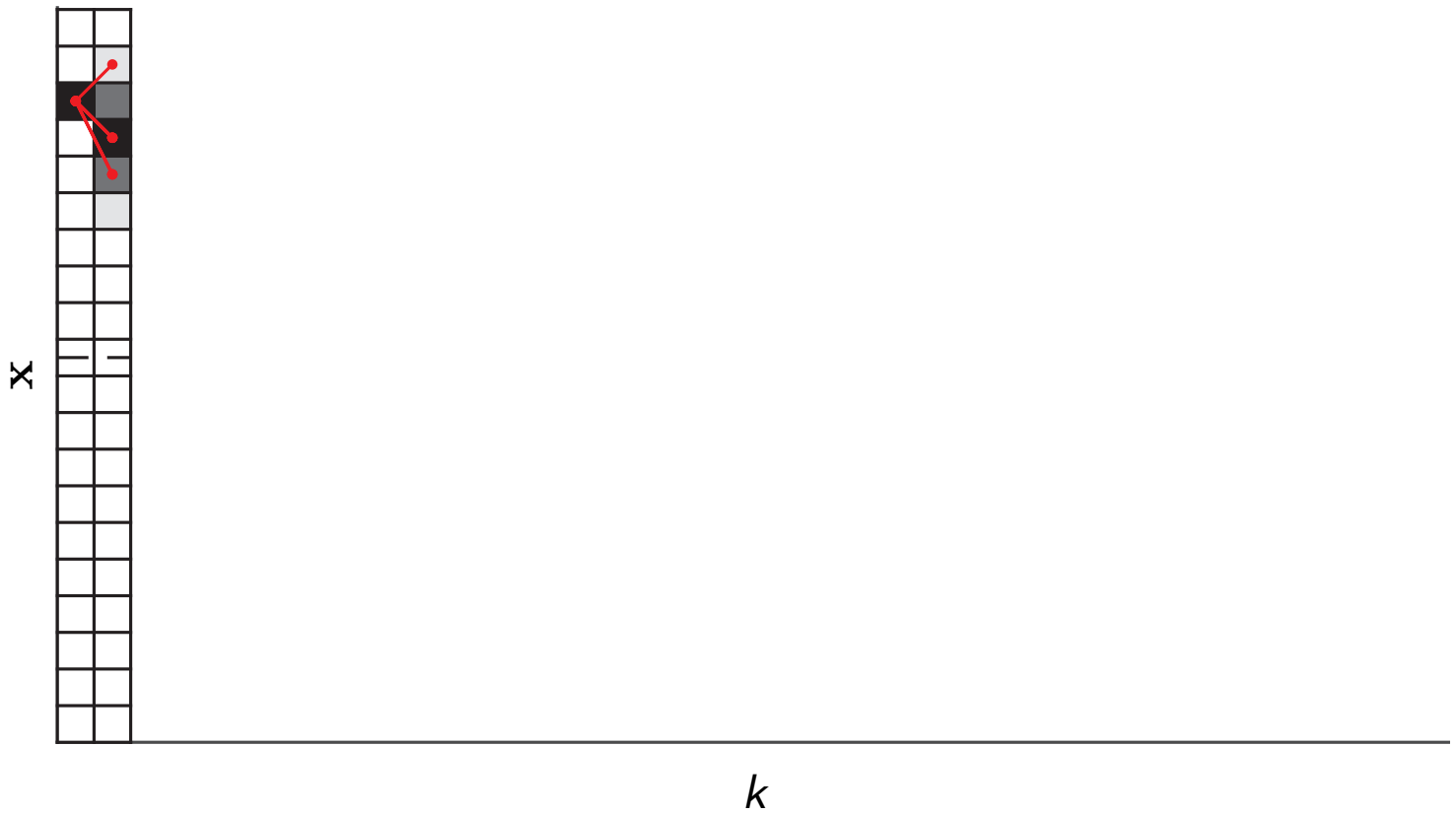


e.g. for
$\mathbf{u} = \mathrm{round}\left(-\frac{\mathbf{x}}{10}\right)$

Let's use the following example (for illustration purposes)

- State $\mathbf{x} \in \mathbb{N}$ and input $\mathbf{u} \in \mathbb{N}$ locked on a grid
- Input $\mathbf{u}$ can only move $\mathbf{x}$ on the grid
- Stage cost: $L(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\mathbf{x}^2 + \frac{1}{2}\mathbf{u}^2$
- Discount $\gamma = 0.9$

Dynamics: $\mathbf{x}_+ = \mathbf{x} + \mathbf{u} + \mathrm{round}(\mathbf{e})$ where $\mathbf{e} \sim \mathcal{N}(0, 1)$

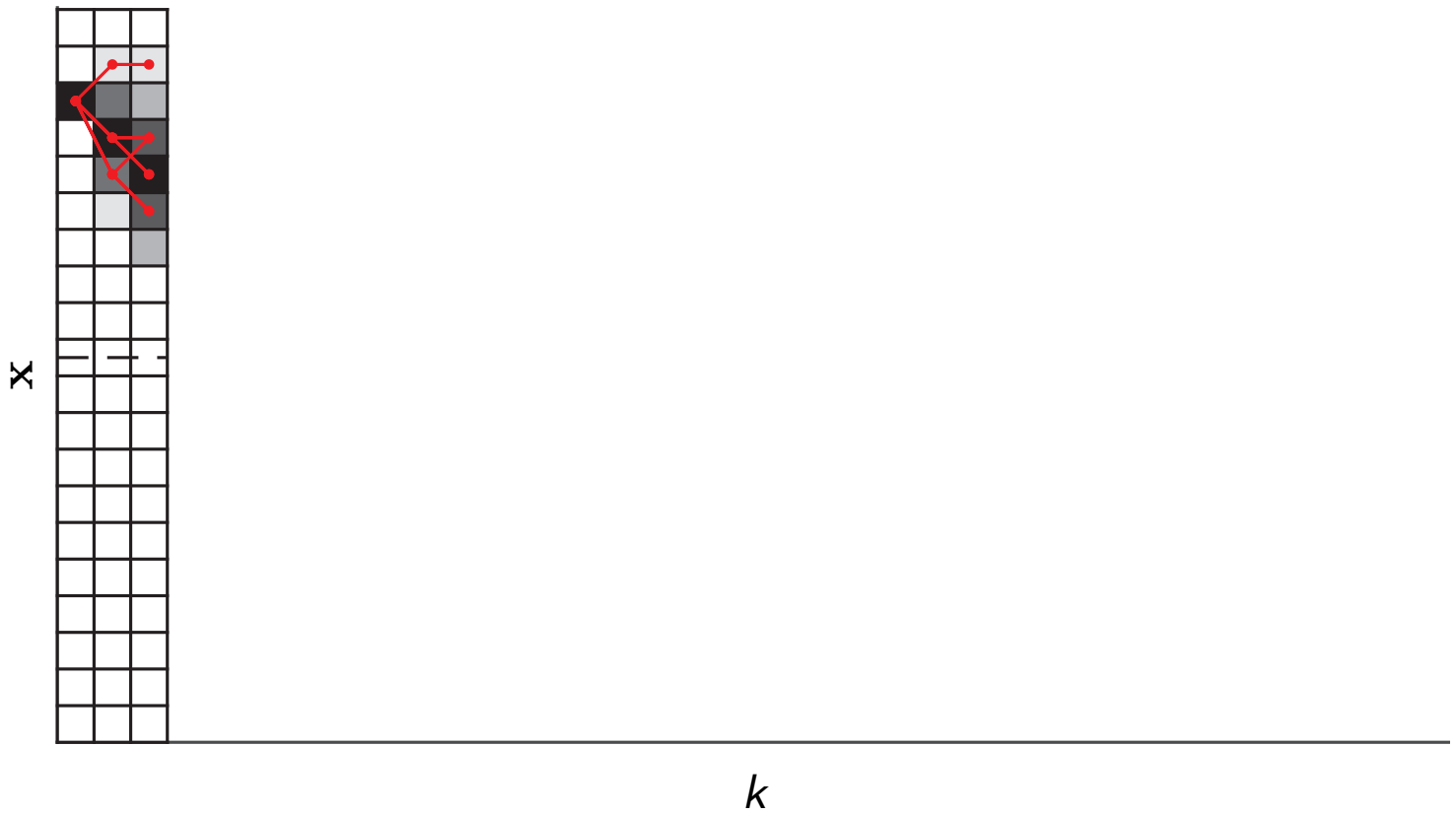


e.g. for
$\mathbf{u} = \mathrm{round}\left(-\frac{\mathbf{x}}{10}\right)$

Let's use the following example (for illustration purposes)

- State $\mathbf{x} \in \mathbb{N}$ and input $\mathbf{u} \in \mathbb{N}$ locked on a grid
- Input $\mathbf{u}$ can only move $\mathbf{x}$ on the grid
- Stage cost: $L(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\mathbf{x}^2 + \frac{1}{2}\mathbf{u}^2$
- Discount $\gamma = 0.9$

Dynamics: $\mathbf{x}_+ = \mathbf{x} + \mathbf{u} + \mathrm{round}(\mathbf{e})$ where $\mathbf{e} \sim \mathcal{N}(0, 1)$

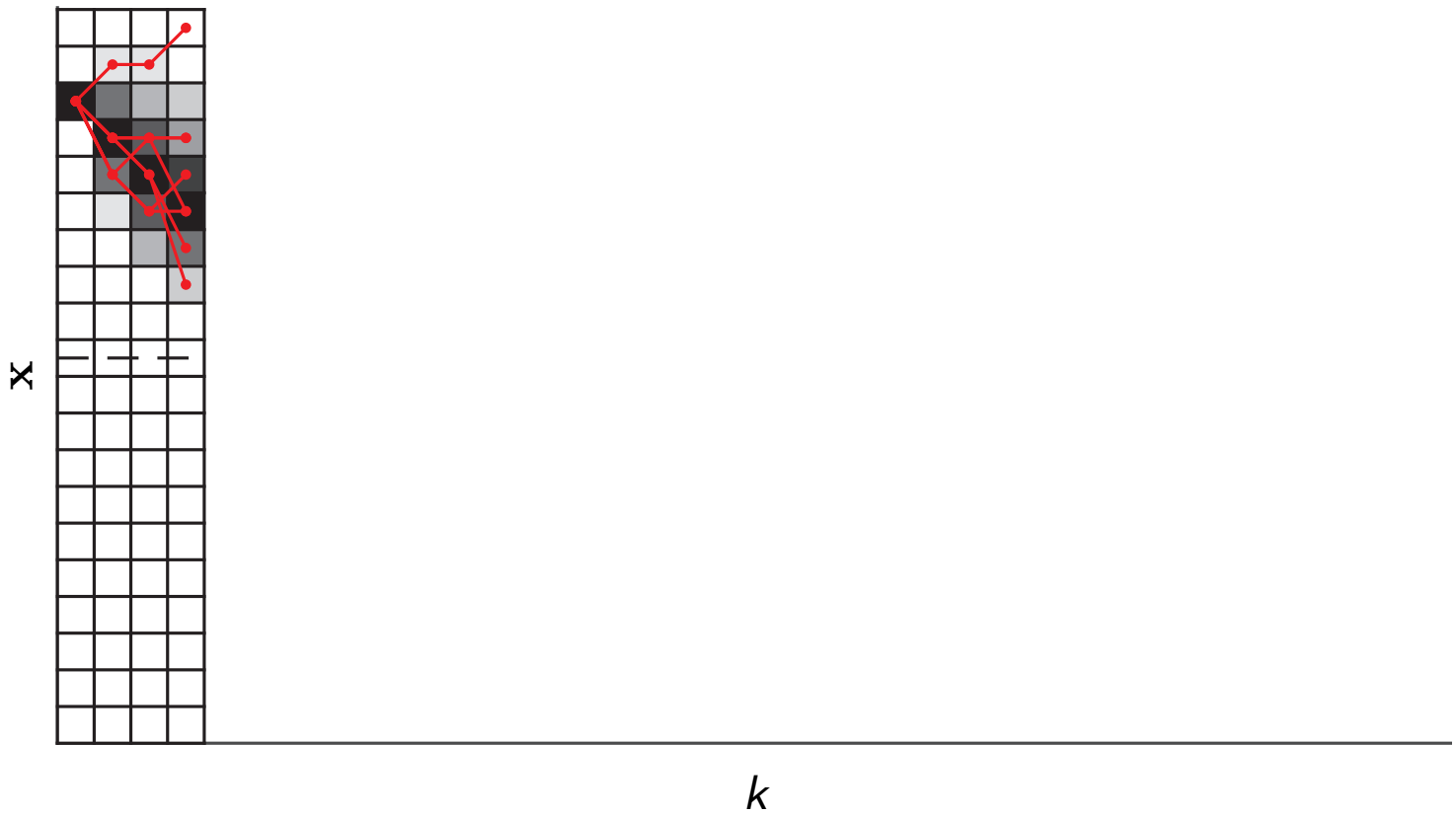


e.g. for
$\mathbf{u} = \mathrm{round}\left(-\frac{\mathbf{x}}{10}\right)$

Let's use the following example (for illustration purposes)

- State $\mathbf{x} \in \mathbb{N}$ and input $\mathbf{u} \in \mathbb{N}$ locked on a grid
- Input $\mathbf{u}$ can only move $\mathbf{x}$ on the grid
- Stage cost: $L(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\mathbf{x}^2 + \frac{1}{2}\mathbf{u}^2$
- Discount $\gamma = 0.9$

Dynamics: $\mathbf{x}_+ = \mathbf{x} + \mathbf{u} + \mathrm{round}(\mathbf{e})$ where $\mathbf{e} \sim \mathcal{N}(0, 1)$

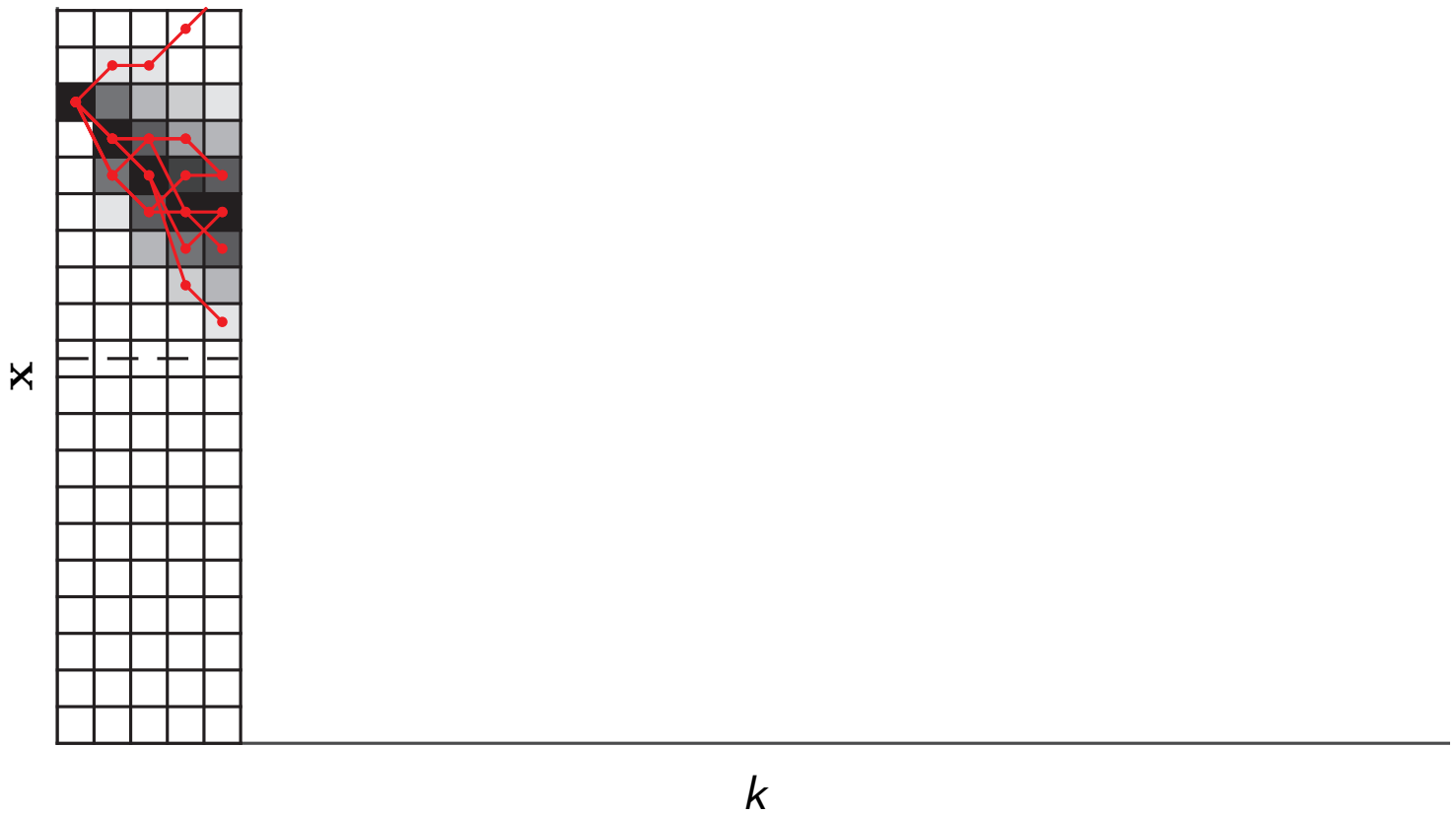


e.g. for
$\mathbf{u} = \mathrm{round}\left(-\frac{\mathbf{x}}{10}\right)$

Let's use the following example (for illustration purposes)

- State $\mathbf{x} \in \mathbb{N}$ and input $\mathbf{u} \in \mathbb{N}$ locked on a grid
- Input $\mathbf{u}$ can only move $\mathbf{x}$ on the grid
- Stage cost: $L(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\mathbf{x}^2 + \frac{1}{2}\mathbf{u}^2$
- Discount $\gamma = 0.9$

Dynamics: $\mathbf{x}_+ = \mathbf{x} + \mathbf{u} + \text{round}(\mathbf{e})$ where $\mathbf{e} \sim \mathcal{N}(0, 1)$

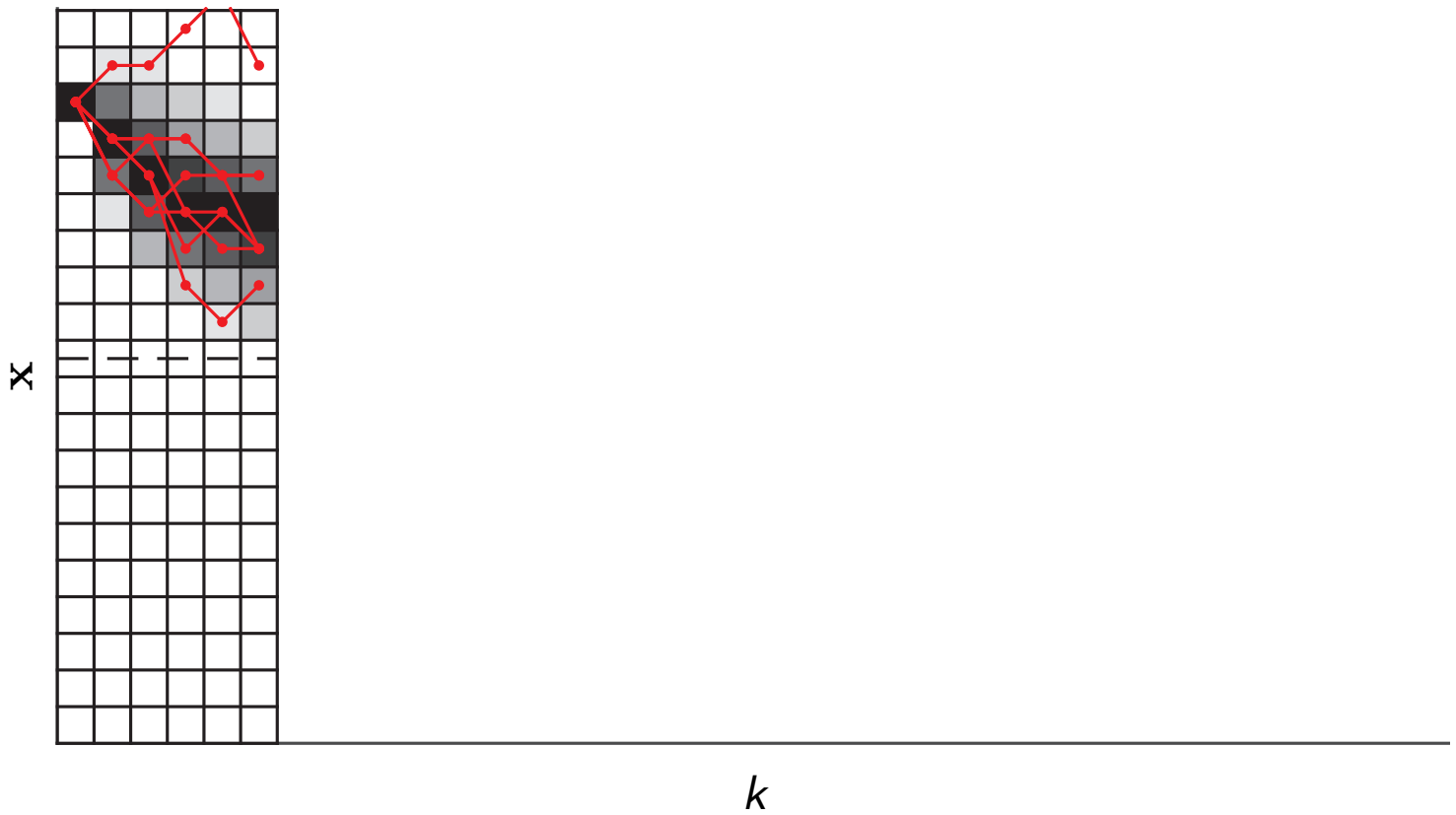


e.g. for
$\mathbf{u} = \text{round}\left(-\frac{\mathbf{x}}{10}\right)$

Let's use the following example (for illustration purposes)

- State $\mathbf{x} \in \mathbb{N}$ and input $\mathbf{u} \in \mathbb{N}$ locked on a grid
- Input $\mathbf{u}$ can only move $\mathbf{x}$ on the grid
- Stage cost: $L(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\mathbf{x}^2 + \frac{1}{2}\mathbf{u}^2$
- Discount $\gamma = 0.9$

Dynamics: $\mathbf{x}_+ = \mathbf{x} + \mathbf{u} + \mathrm{round}(\mathbf{e})$ where $\mathbf{e} \sim \mathcal{N}(0, 1)$

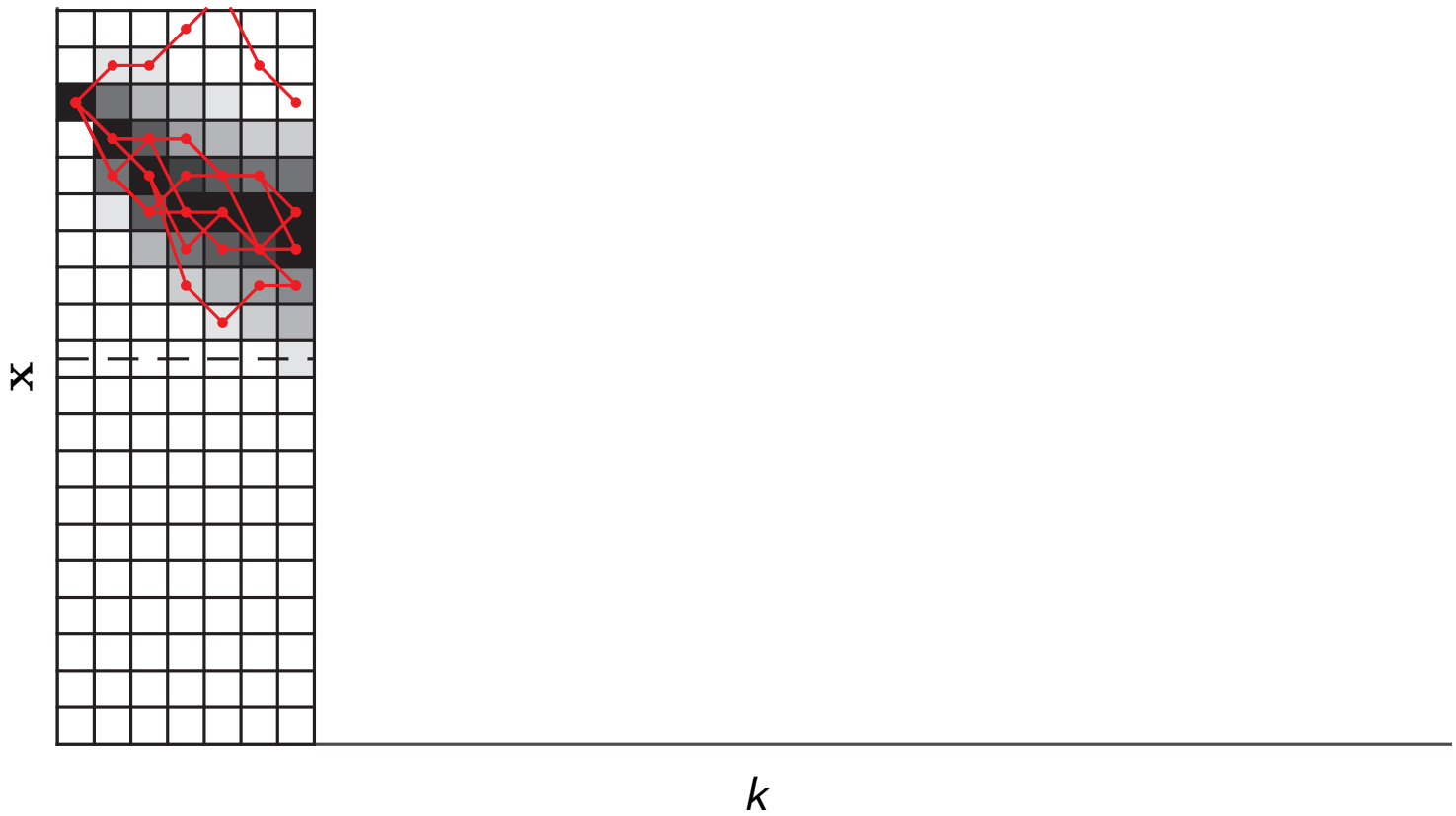


e.g. for
$\mathbf{u} = \mathrm{round}\left(-\frac{\mathbf{x}}{10}\right)$

Let's use the following example (for illustration purposes)

- State $\mathbf{x} \in \mathbb{N}$ and input $\mathbf{u} \in \mathbb{N}$ locked on a grid
- Input $\mathbf{u}$ can only move $\mathbf{x}$ on the grid
- Stage cost: $L(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\mathbf{x}^2 + \frac{1}{2}\mathbf{u}^2$
- Discount $\gamma = 0.9$

Dynamics: $\mathbf{x}_+ = \mathbf{x} + \mathbf{u} + \mathrm{round}(\mathbf{e})$ where $\mathbf{e} \sim \mathcal{N}(0, 1)$

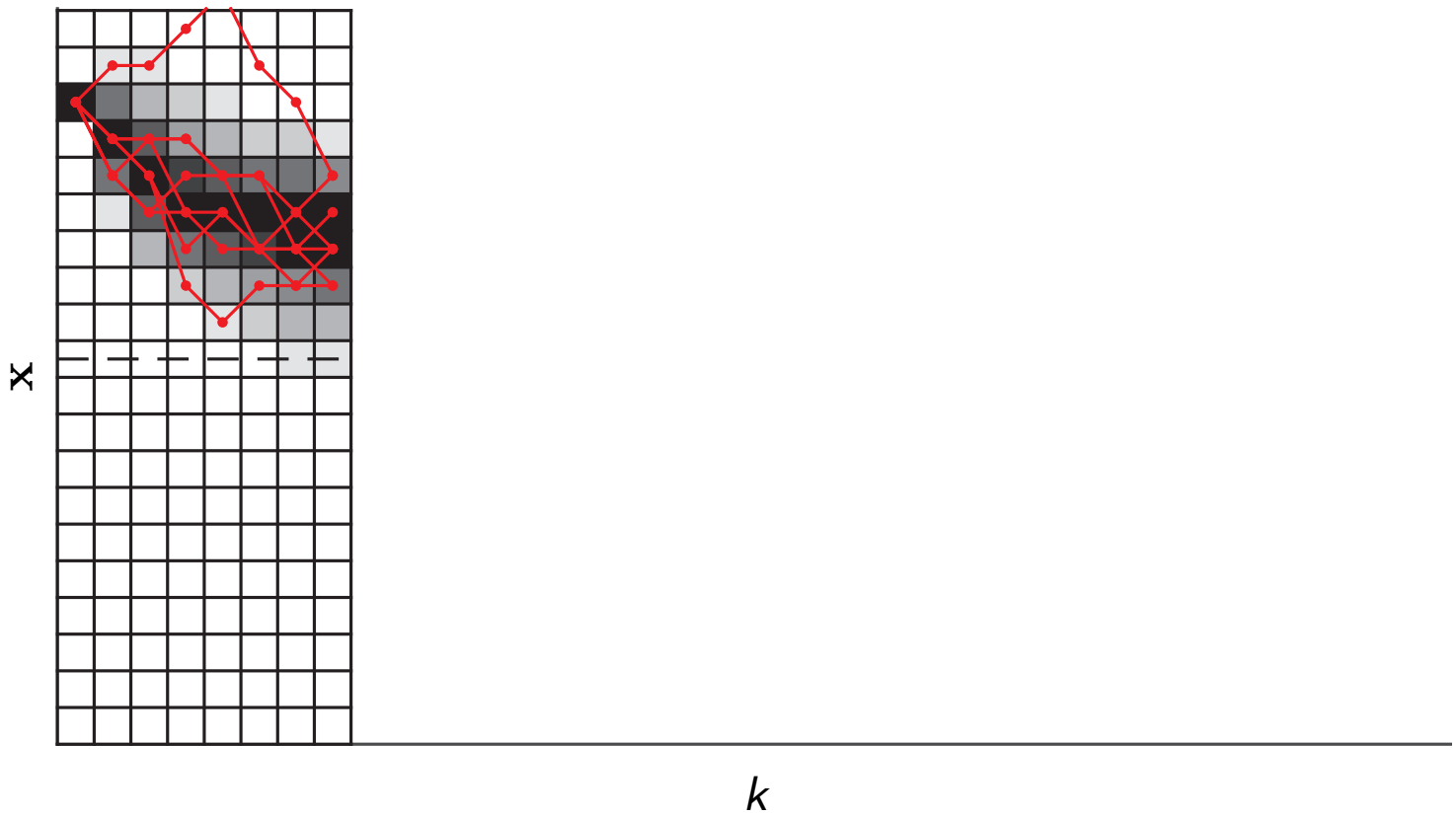


e.g. for
$\mathbf{u} = \mathrm{round}\left(-\frac{\mathbf{x}}{10}\right)$

Let's use the following example (for illustration purposes)

- State $\mathbf{x} \in \mathbb{N}$ and input $\mathbf{u} \in \mathbb{N}$ locked on a grid
- Input $\mathbf{u}$ can only move $\mathbf{x}$ on the grid
- Stage cost: $L(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\mathbf{x}^2 + \frac{1}{2}\mathbf{u}^2$
- Discount $\gamma = 0.9$

Dynamics: $\mathbf{x}_+ = \mathbf{x} + \mathbf{u} + \text{round}(\mathbf{e})$ where $\mathbf{e} \sim \mathcal{N}(0, 1)$

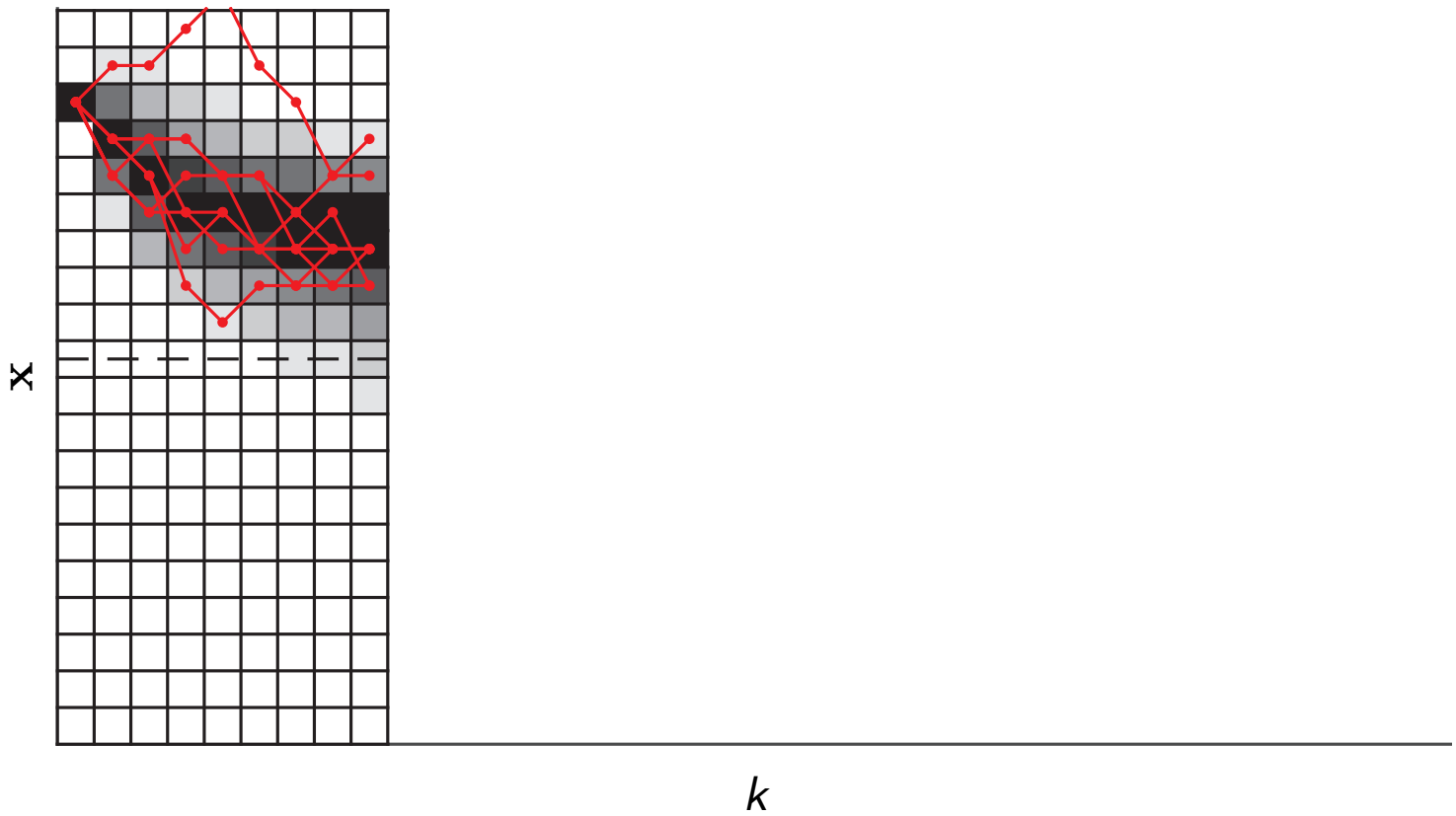


e.g. for
$\mathbf{u} = \text{round}\left(-\frac{\mathbf{x}}{10}\right)$

Let's use the following example (for illustration purposes)

- State $\mathbf{x} \in \mathbb{N}$ and input $\mathbf{u} \in \mathbb{N}$ locked on a grid
- Input $\mathbf{u}$ can only move $\mathbf{x}$ on the grid
- Stage cost: $L(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\mathbf{x}^2 + \frac{1}{2}\mathbf{u}^2$
- Discount $\gamma = 0.9$

Dynamics: $\mathbf{x}_+ = \mathbf{x} + \mathbf{u} + \text{round}(\mathbf{e})$ where $\mathbf{e} \sim \mathcal{N}(0, 1)$

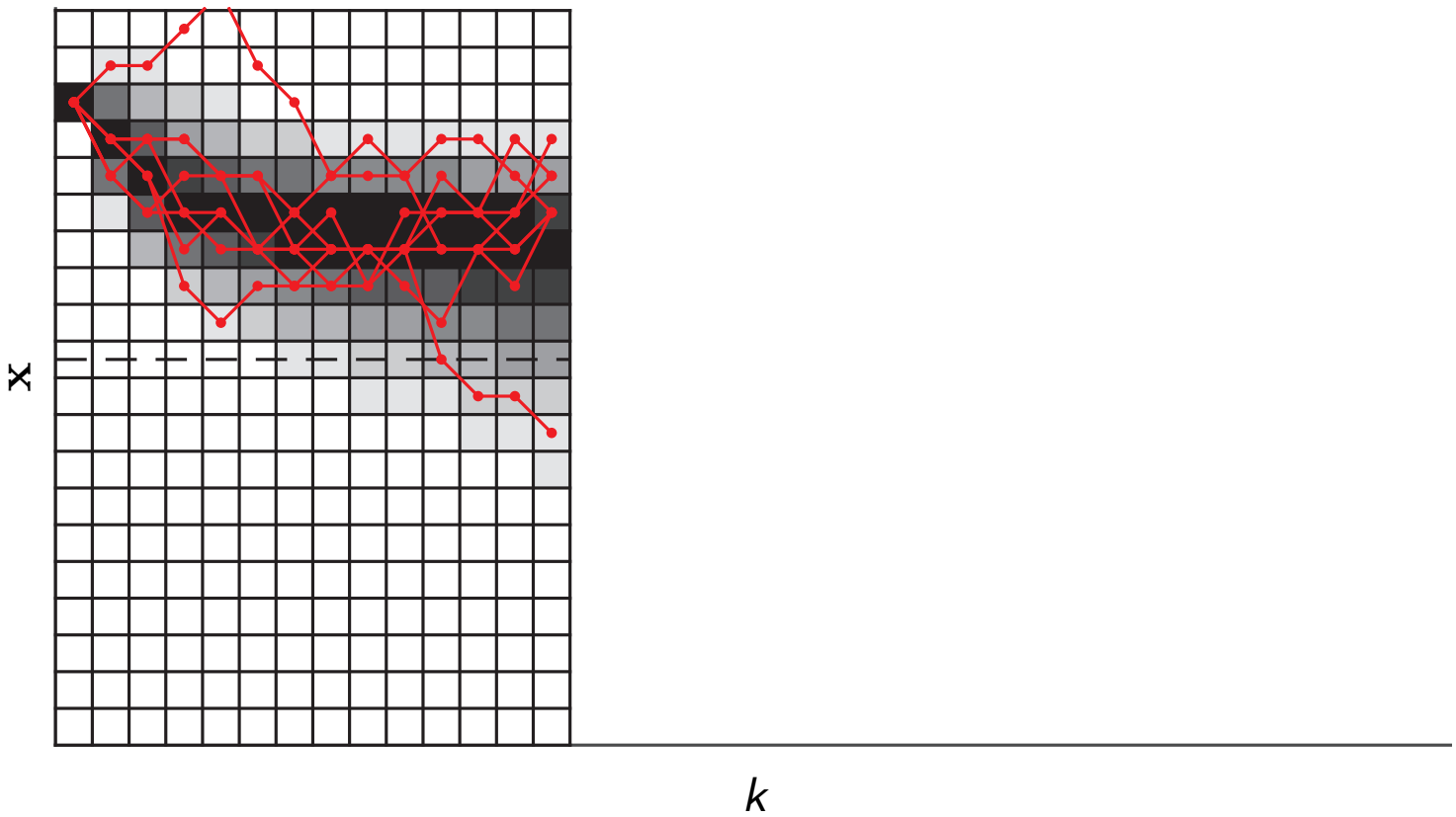


e.g. for
$\mathbf{u} = \text{round}\left(-\frac{\mathbf{x}}{10}\right)$

Let's use the following example (for illustration purposes)

- State $\mathbf{x} \in \mathbb{N}$ and input $\mathbf{u} \in \mathbb{N}$ locked on a grid
- Input $\mathbf{u}$ can only move $\mathbf{x}$ on the grid
- Stage cost: $L(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\mathbf{x}^2 + \frac{1}{2}\mathbf{u}^2$
- Discount $\gamma = 0.9$

Dynamics: $\mathbf{x}_+ = \mathbf{x} + \mathbf{u} + \mathrm{round}(\mathbf{e})$ where $\mathbf{e} \sim \mathcal{N}(0, 1)$

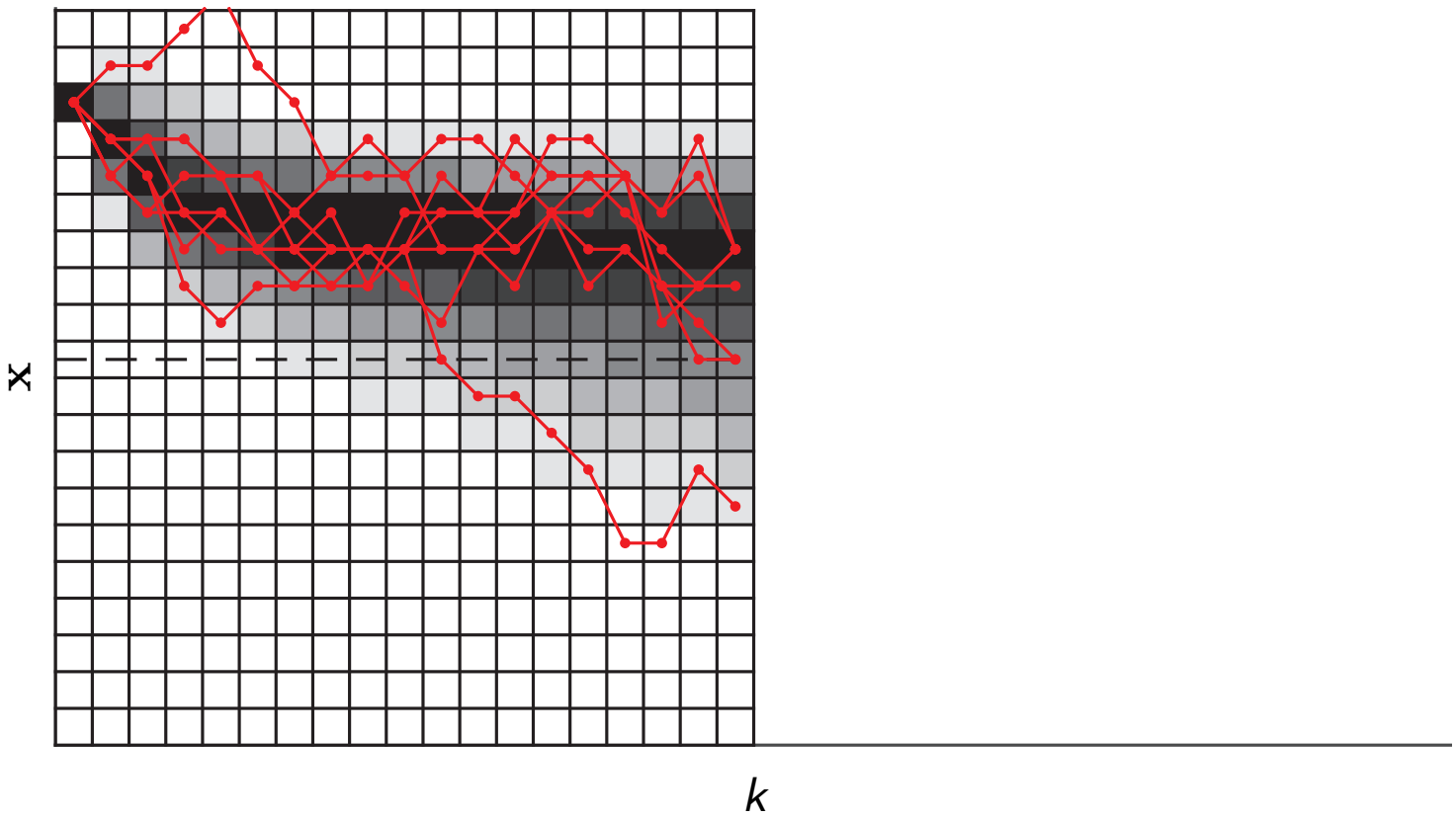


e.g. for
$\mathbf{u} = \mathrm{round}\left(-\frac{\mathbf{x}}{10}\right)$

Let's use the following example (for illustration purposes)

- State $\mathbf{x} \in \mathbb{N}$ and input $\mathbf{u} \in \mathbb{N}$ locked on a grid
- Input $\mathbf{u}$ can only move $\mathbf{x}$ on the grid
- Stage cost: $L(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\mathbf{x}^2 + \frac{1}{2}\mathbf{u}^2$
- Discount $\gamma = 0.9$

Dynamics: $\mathbf{x}_+ = \mathbf{x} + \mathbf{u} + \mathrm{round}(\mathbf{e})$ where $\mathbf{e} \sim \mathcal{N}(0, 1)$

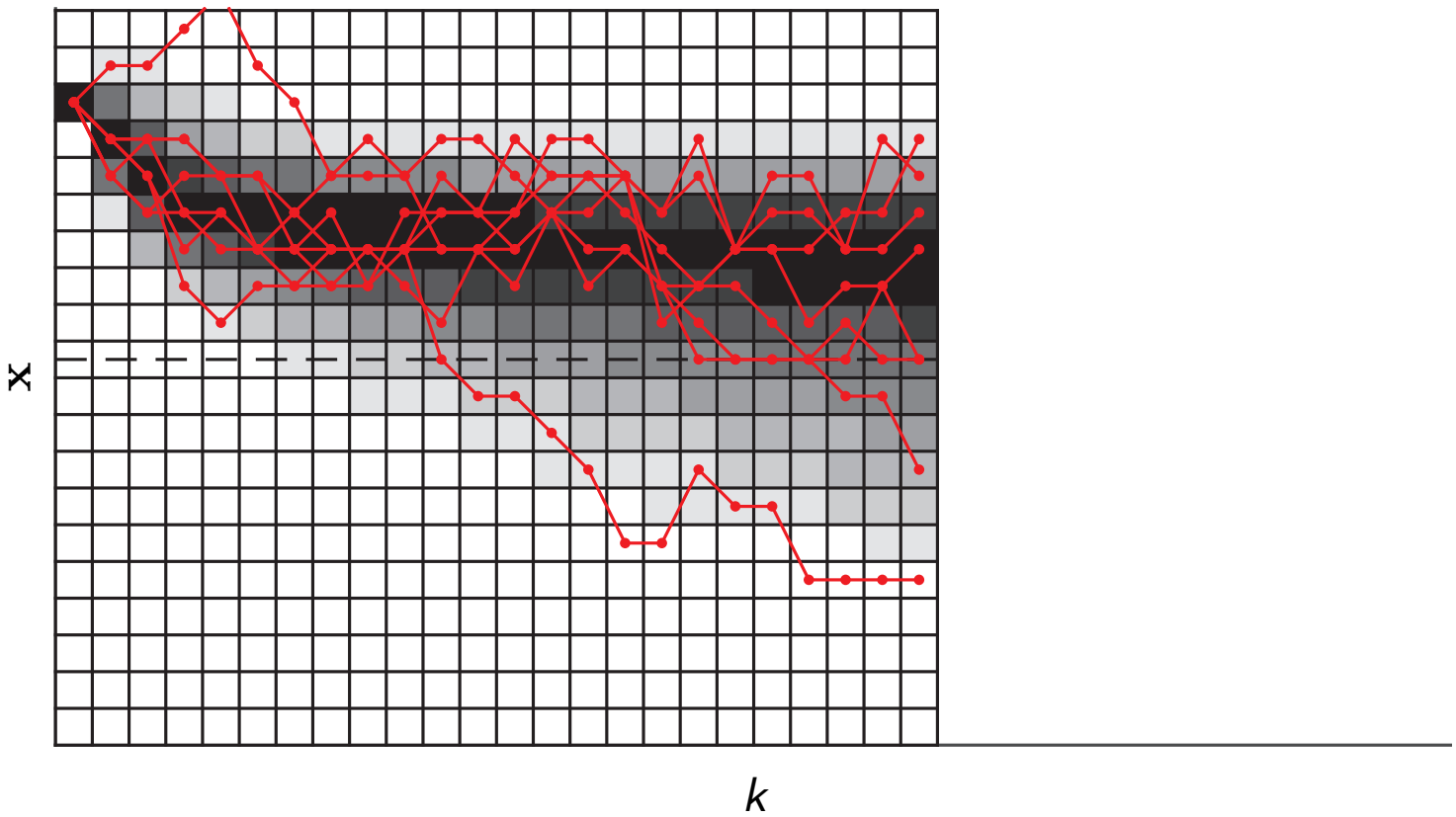


e.g. for
$\mathbf{u} = \mathrm{round}\left(-\frac{\mathbf{x}}{10}\right)$

Let's use the following example (for illustration purposes)

- State $\mathbf{x} \in \mathbb{N}$ and input $\mathbf{u} \in \mathbb{N}$ locked on a grid
- Input $\mathbf{u}$ can only move $\mathbf{x}$ on the grid
- Stage cost: $L(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\mathbf{x}^2 + \frac{1}{2}\mathbf{u}^2$
- Discount $\gamma = 0.9$

Dynamics: $\mathbf{x}_+ = \mathbf{x} + \mathbf{u} + \mathrm{round}(\mathbf{e})$ where $\mathbf{e} \sim \mathcal{N}(0, 1)$

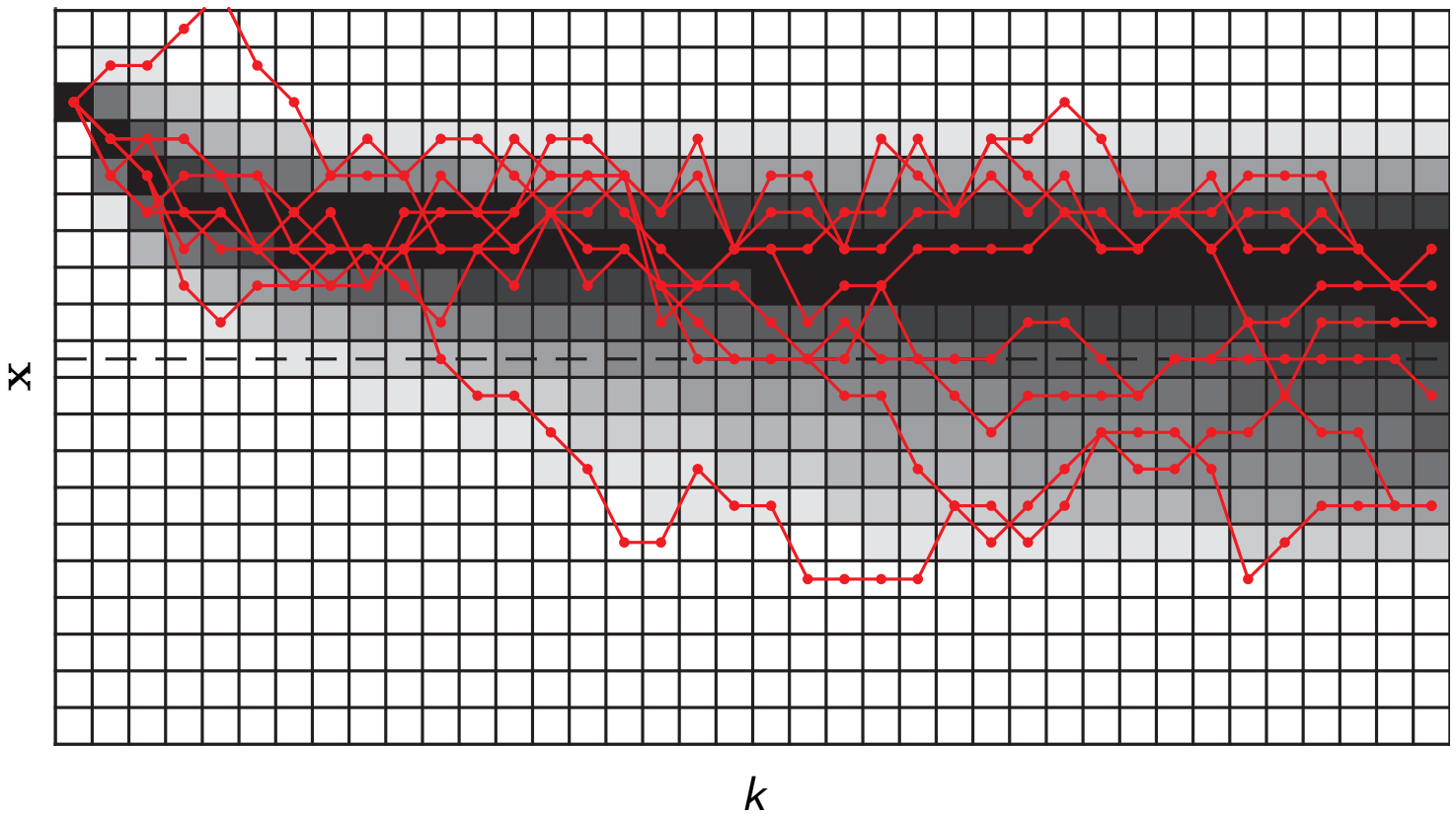


e.g. for
$\mathbf{u} = \mathrm{round}\left(-\frac{\mathbf{x}}{10}\right)$

Let's use the following example (for illustration purposes)

- State $\mathbf{x} \in \mathbb{N}$ and input $\mathbf{u} \in \mathbb{N}$ locked on a grid
- Input $\mathbf{u}$ can only move $\mathbf{x}$ on the grid
- Stage cost: $L(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\mathbf{x}^2 + \frac{1}{2}\mathbf{u}^2$
- Discount $\gamma = 0.9$

Dynamics: $\mathbf{x}_+ = \mathbf{x} + \mathbf{u} + \text{round}(\mathbf{e})$ where $\mathbf{e} \sim \mathcal{N}(0, 1)$



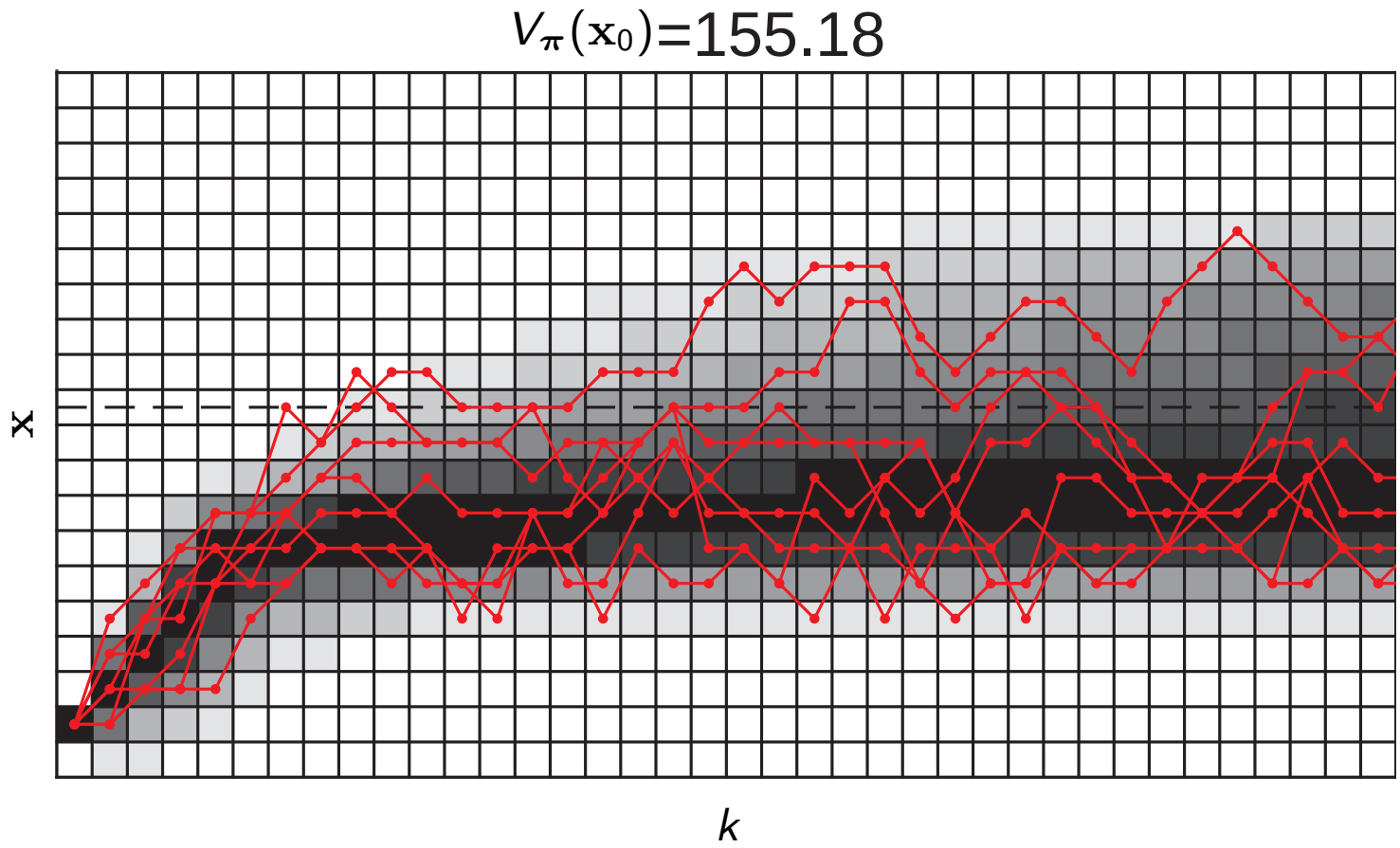
e.g. for
$\mathbf{u} = \text{round}\left(-\frac{\mathbf{x}}{10}\right)$

Outline

Learning the Value function

Value function for policy π:

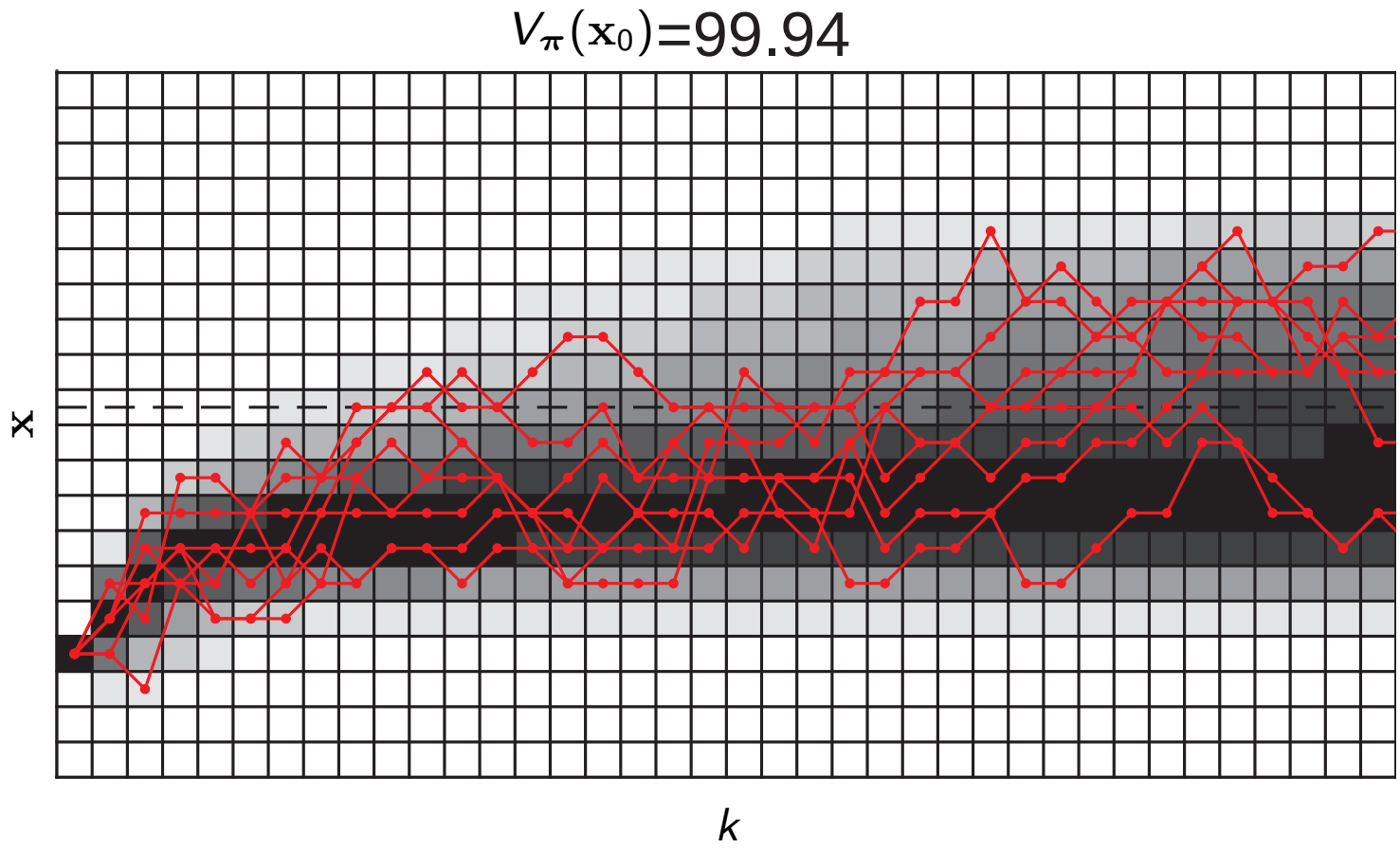
$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$



Learning the Value function

Value function for policy π:

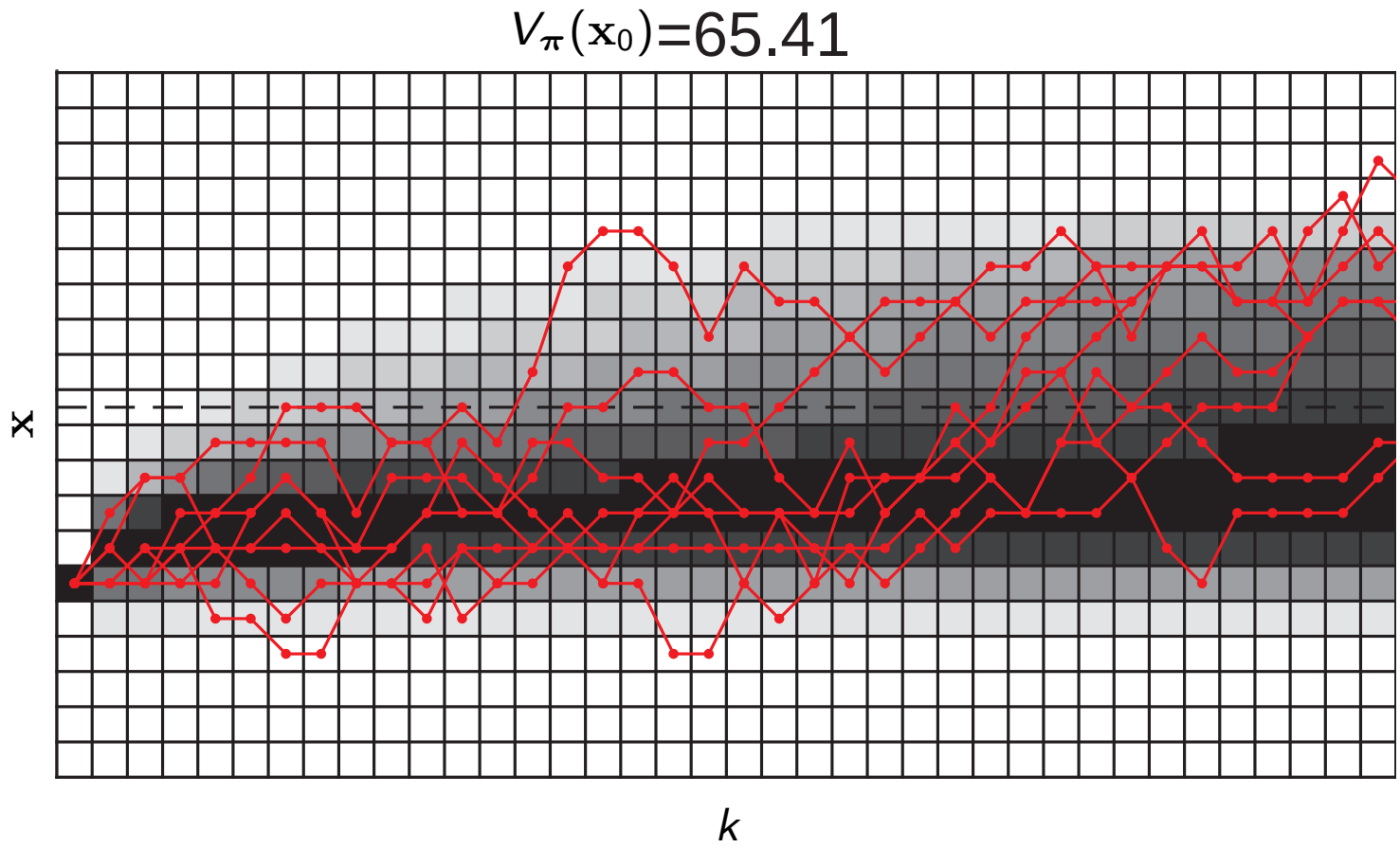
$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$



Learning the Value function

Value function for policy π:

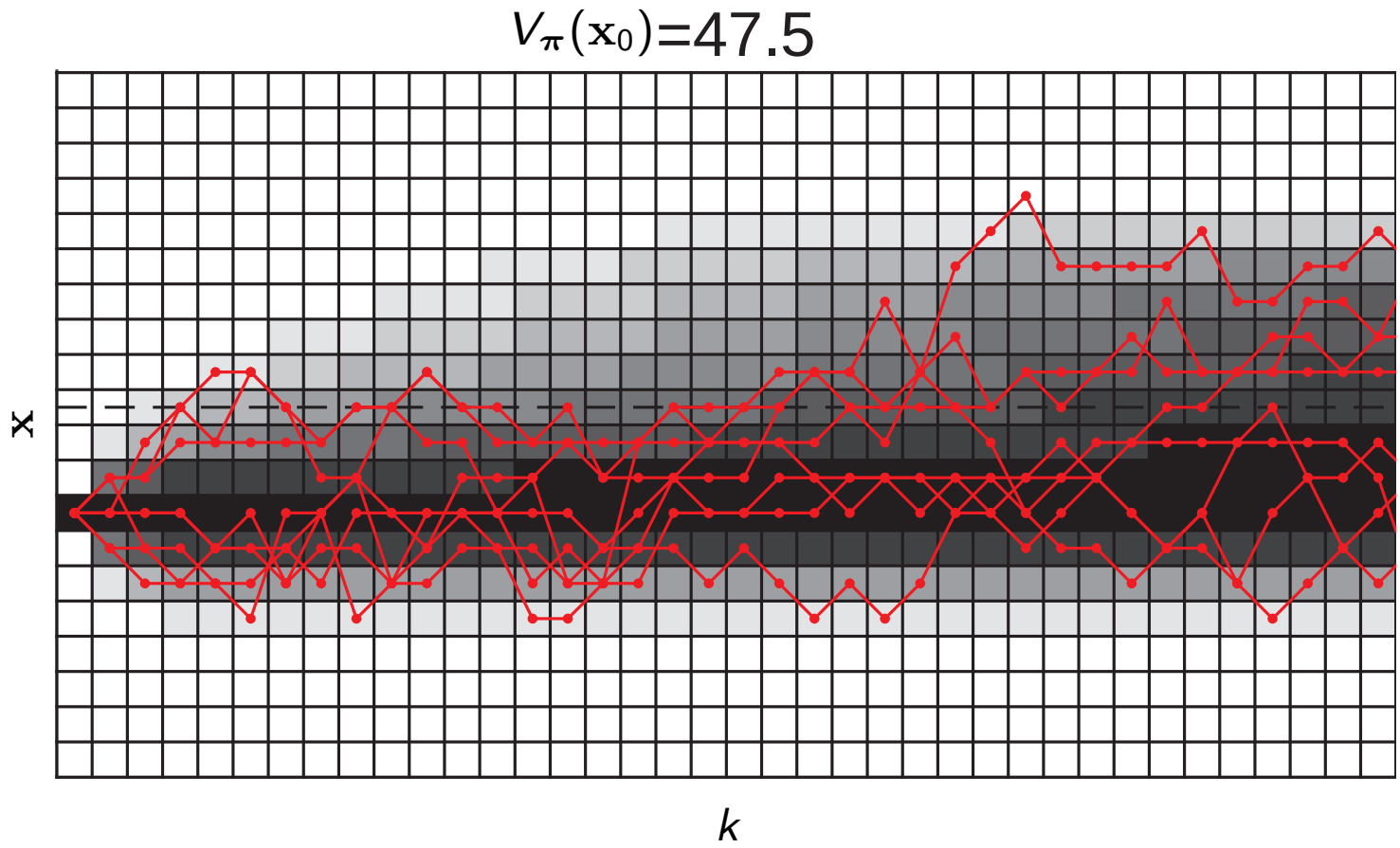
$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$



Learning the Value function

Value function for policy π:

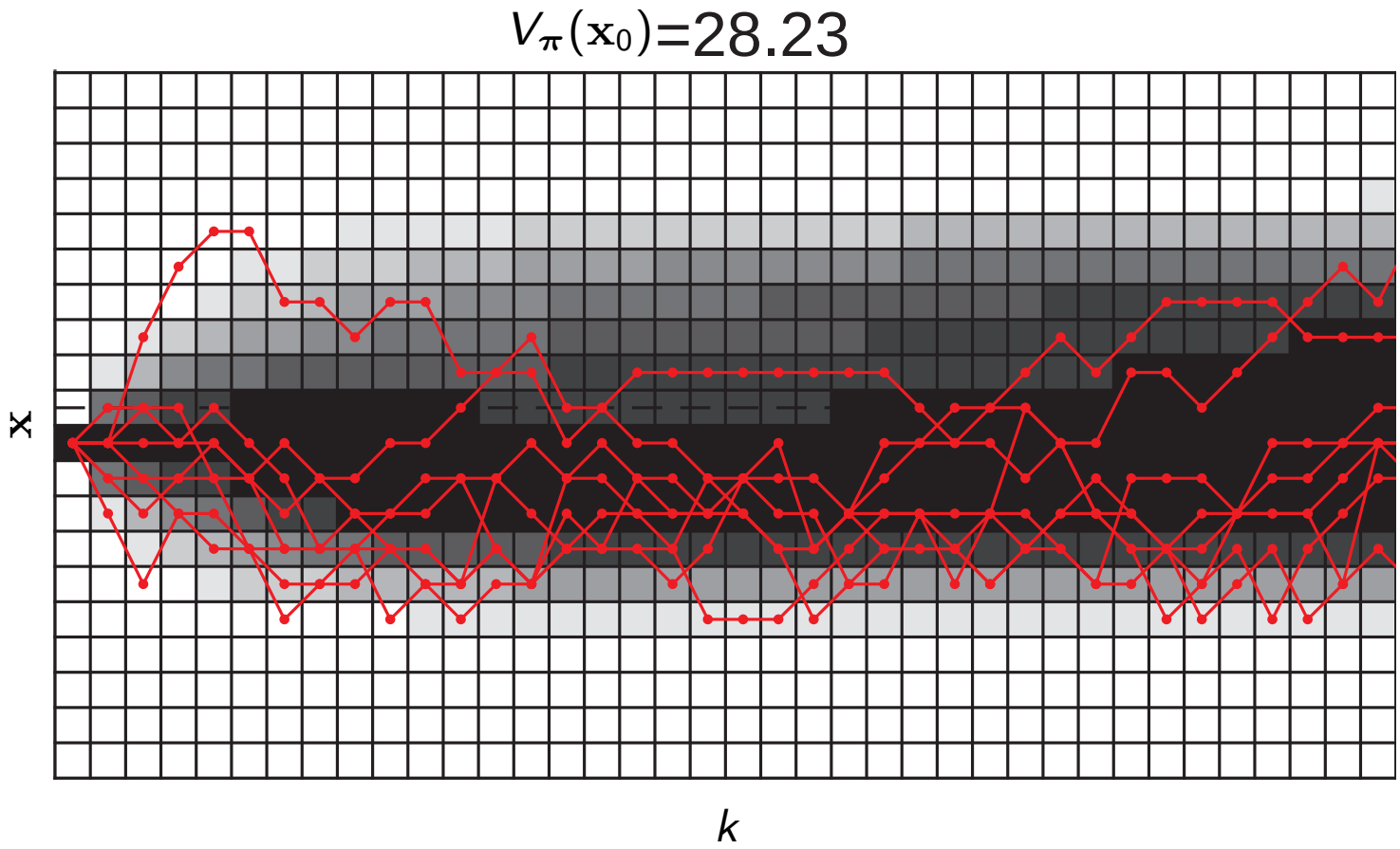
$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$



Learning the Value function

Value function for policy π:

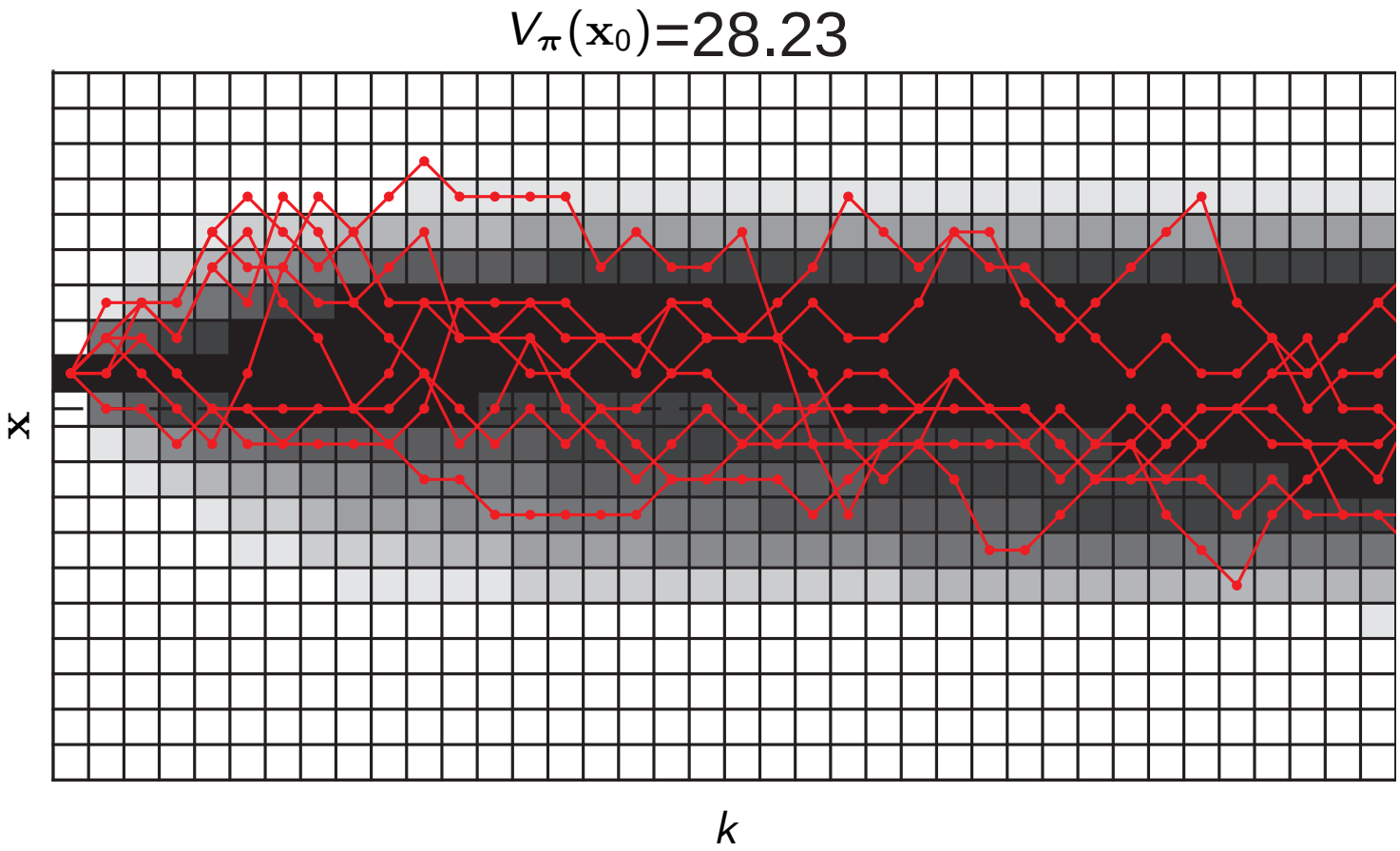
$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$



Learning the Value function

Value function for policy π:

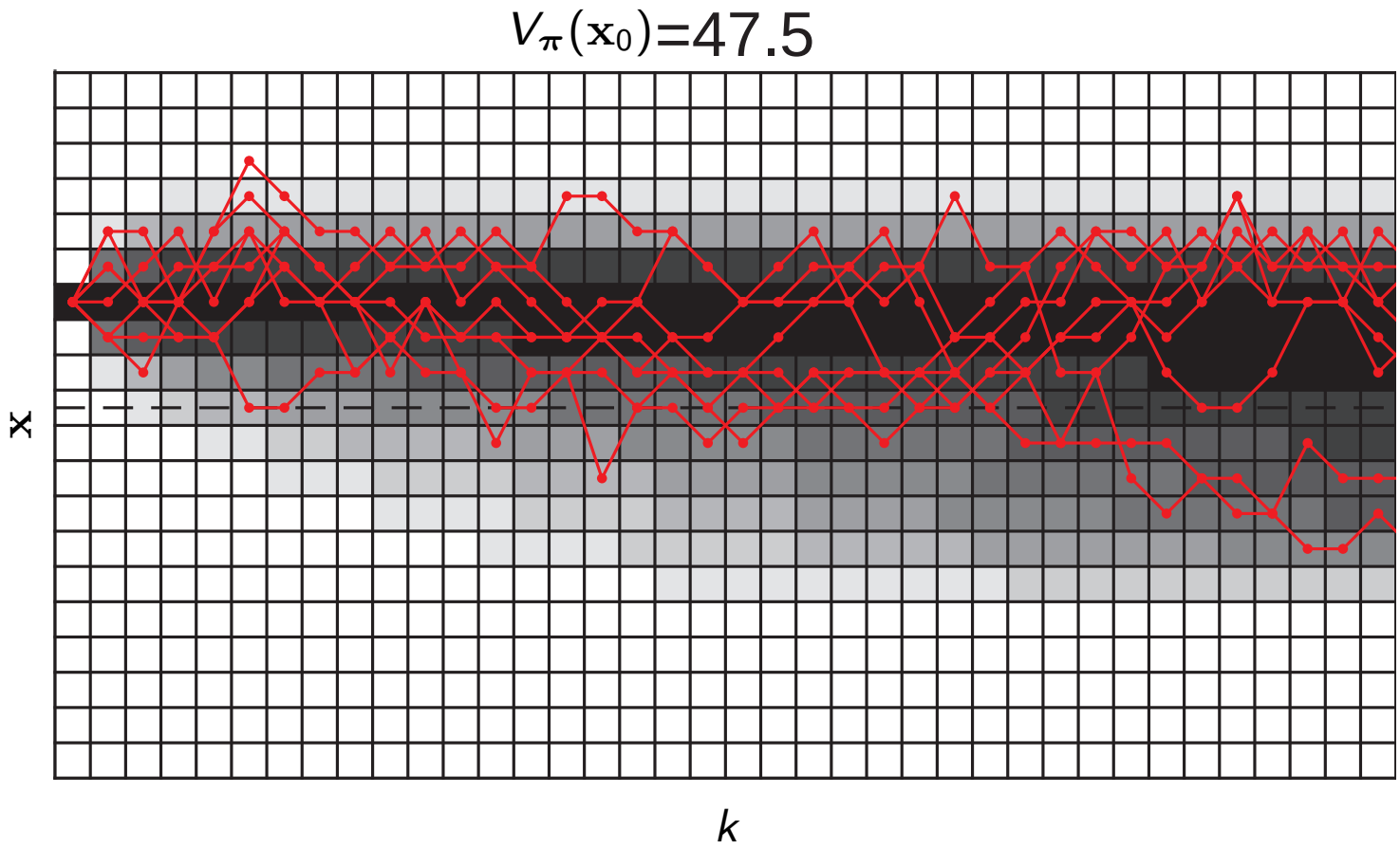
$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$



Learning the Value function

Value function for policy π:

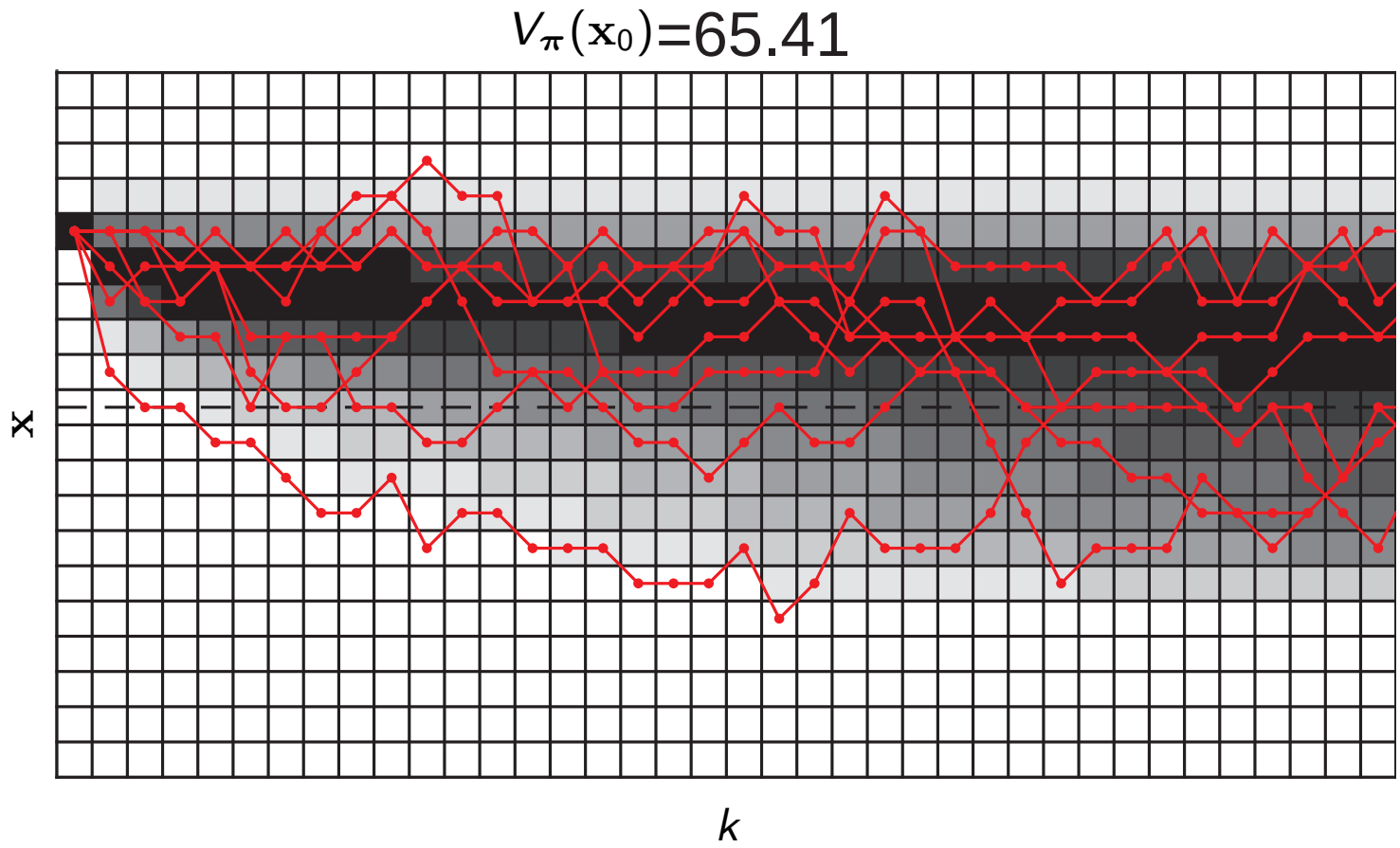
$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$



Learning the Value function

Value function for policy π:

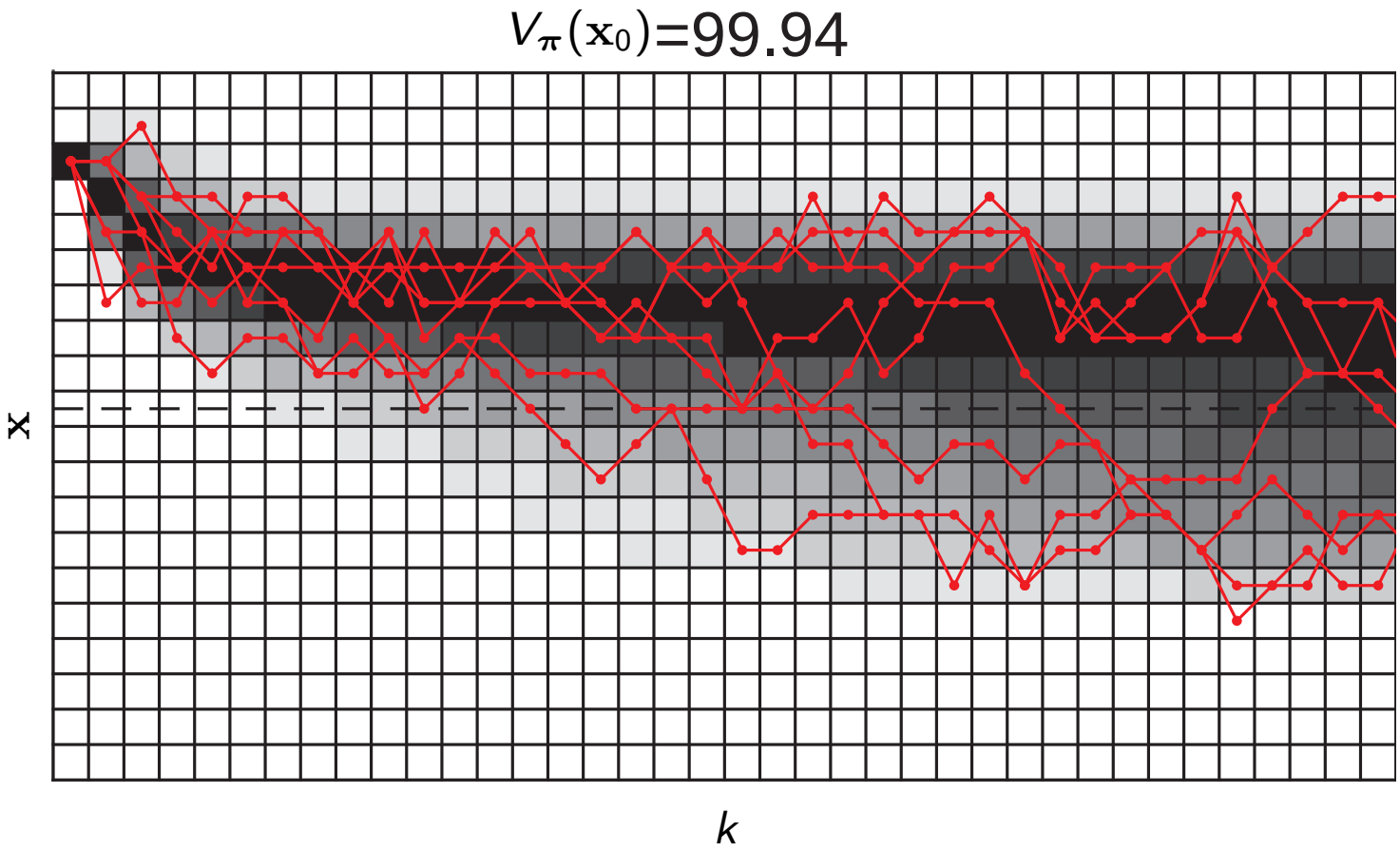
$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

- Reminder **policy evaluation**: for a given policy $\boldsymbol{\pi}$, iterate:

$$V_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow L(\mathbf{x}, \boldsymbol{\pi}(\mathbf{x})) + \gamma \mathbb{E}\left[V_{\boldsymbol{\pi}}(\mathbf{x}_+) \,|\, \mathbf{x}, \boldsymbol{\pi}\right]$$

  starting from any $V(\mathbf{x})$.

Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

- Reminder **policy evaluation**: for a given policy π, iterate:

$$V_{\pi}(\mathbf{x}) \leftarrow L(\mathbf{x}, \boldsymbol{\pi}(\mathbf{x})) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \boldsymbol{\pi}\right]$$

  starting from any $V(\mathbf{x})$.

- Evaluation of $\mathbb{E}\left[V(\mathbf{x}_+) \,|\, \mathbf{x}, \boldsymbol{\pi}\right]$ **requires a model**:

$$\mathbb{E}\left[V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \boldsymbol{\pi}\right] = \int V_{\pi}(\mathbf{x}_+) \underbrace{\mathbb{P}\left[\mathbf{x}_+ \,|\, \mathbf{x}, \boldsymbol{\pi}\right]}_{\text{Model}} \mathrm{d}\mathbf{x}_+$$

Learning the Value function

Value function for policy π:

$$V_\pi(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

- Reminder **policy evaluation**: for a given policy π, iterate:

$$V_\pi(\mathbf{x}) \leftarrow L(\mathbf{x}, \boldsymbol{\pi}(\mathbf{x})) + \gamma \mathbb{E}\left[V_\pi(\mathbf{x}_+) \,|\, \mathbf{x}, \boldsymbol{\pi}\right]$$

  starting from any $V(\mathbf{x})$.
- Evaluation of $\mathbb{E}\left[V(\mathbf{x}_+) \,|\, \mathbf{x}, \boldsymbol{\pi}\right]$ **requires a model**:

$$\mathbb{E}\left[V_\pi(\mathbf{x}_+) \,|\, \mathbf{x}, \boldsymbol{\pi}\right] = \int V_\pi(\mathbf{x}_+) \underbrace{\mathbb{P}\left[\mathbf{x}_+ \,|\, \mathbf{x}, \boldsymbol{\pi}\right]}_{\text{Model}} \mathrm{d}\mathbf{x}_+$$

- How can we **evaluate** $V_\pi(\mathbf{x})$ **from data** rather than from the model?

Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Monte Carlo (MC) evaluation:

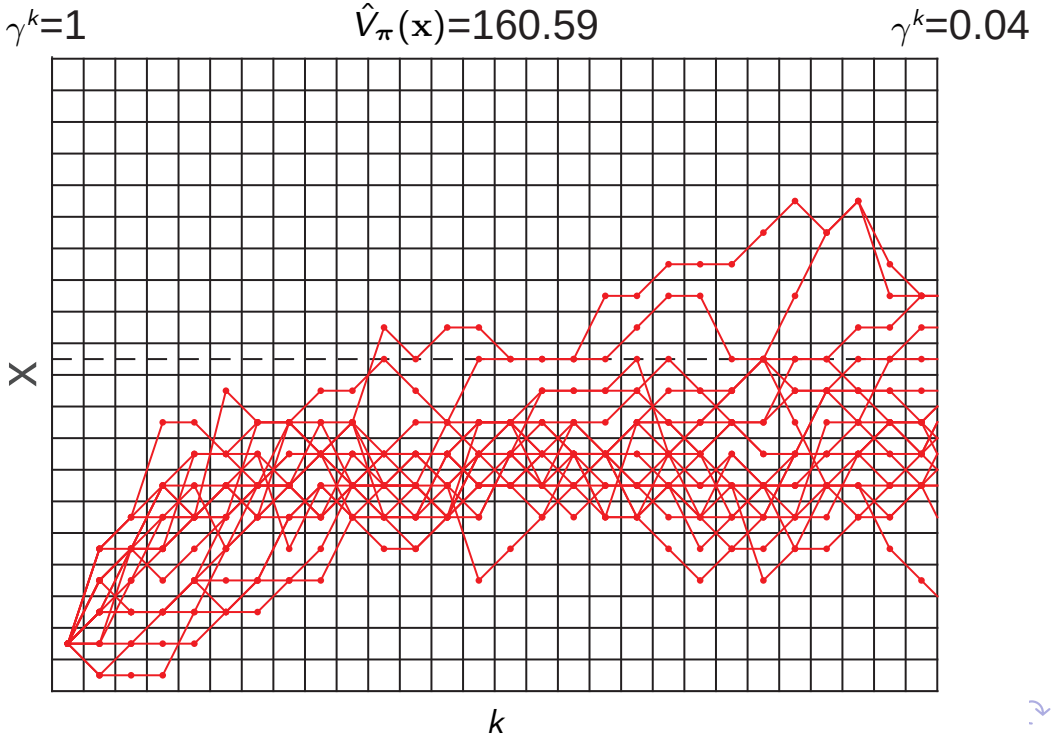
- Start the system in state $\mathbf{x}$
- Unroll N trajectories using $\boldsymbol{\pi}(\mathbf{x})$
- Collect trajectories $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$
- Return:

$$R_i = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k^{(i)}, \boldsymbol{\pi}(\mathbf{x}_k^{(i)}))$$

- Evaluate empirical mean:

$$\hat{V}_{\pi}(\mathbf{x}) \approx \frac{1}{N} \sum_{k=1}^{N} R_i$$

E.g. for $N = 20$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Monte Carlo (MC) evaluation:

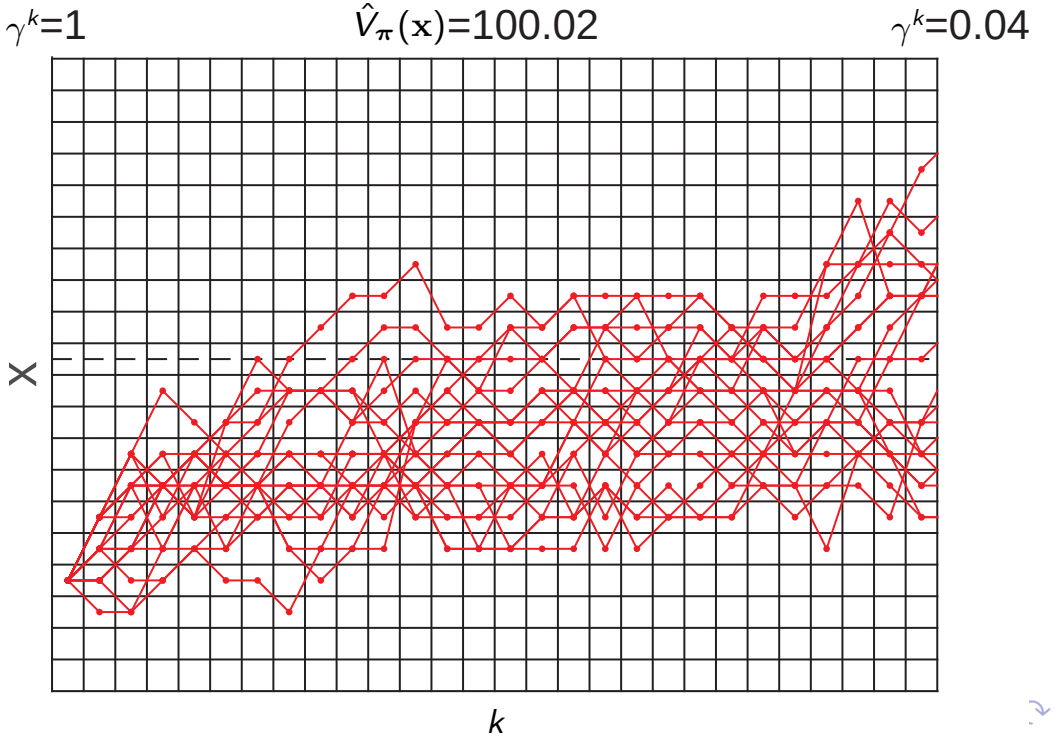
- Start the system in state $\mathbf{x}$
- Unroll N trajectories using $\boldsymbol{\pi}(\mathbf{x})$
- Collect trajectories $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$
- Return:

$$R_i = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k^{(i)}, \boldsymbol{\pi}(\mathbf{x}_k^{(i)}))$$

- Evaluate empirical mean:

$$\hat{V}_{\pi}(\mathbf{x}) \approx \frac{1}{N} \sum_{k=1}^{N} R_i$$

E.g. for $N = 20$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x} \right]$$

Monte Carlo (MC) evaluation:

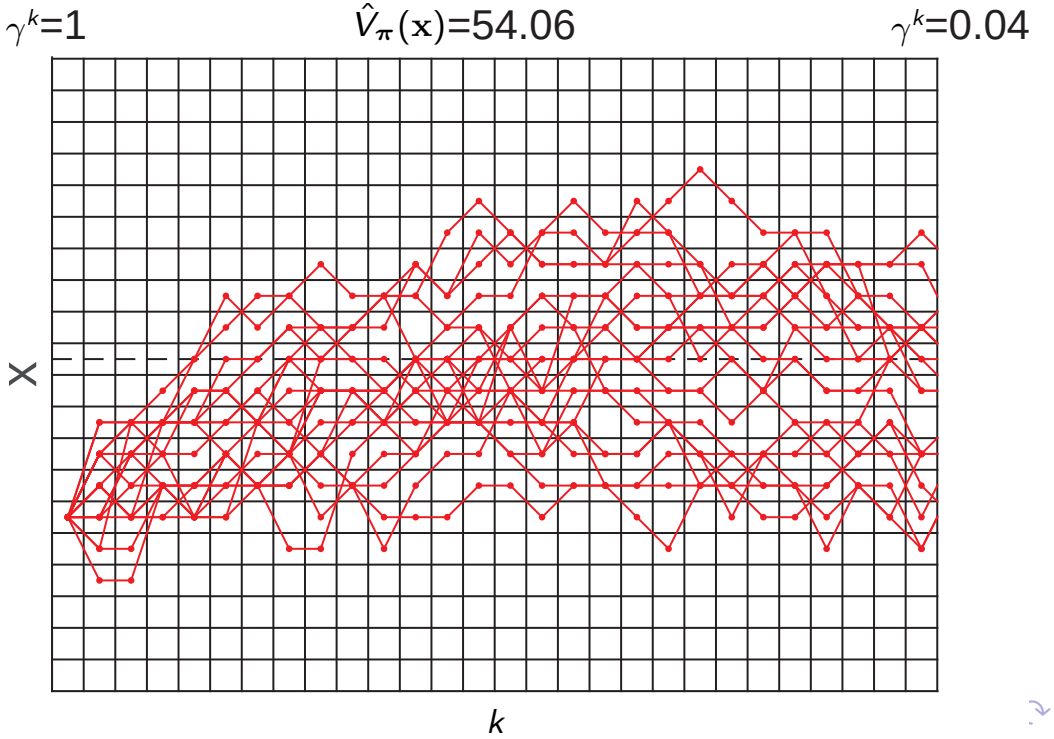
- Start the system in state $\mathbf{x}$
- Unroll N trajectories using $\boldsymbol{\pi}(\mathbf{x})$
- Collect trajectories $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$
- Return:

$$R_i = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k^{(i)}, \boldsymbol{\pi}(\mathbf{x}_k^{(i)}))$$

- Evaluate empirical mean:

$$\hat{V}_{\pi}(\mathbf{x}) \approx \frac{1}{N} \sum_{k=1}^{N} R_i$$

E.g. for $N = 20$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Monte Carlo (MC) evaluation:

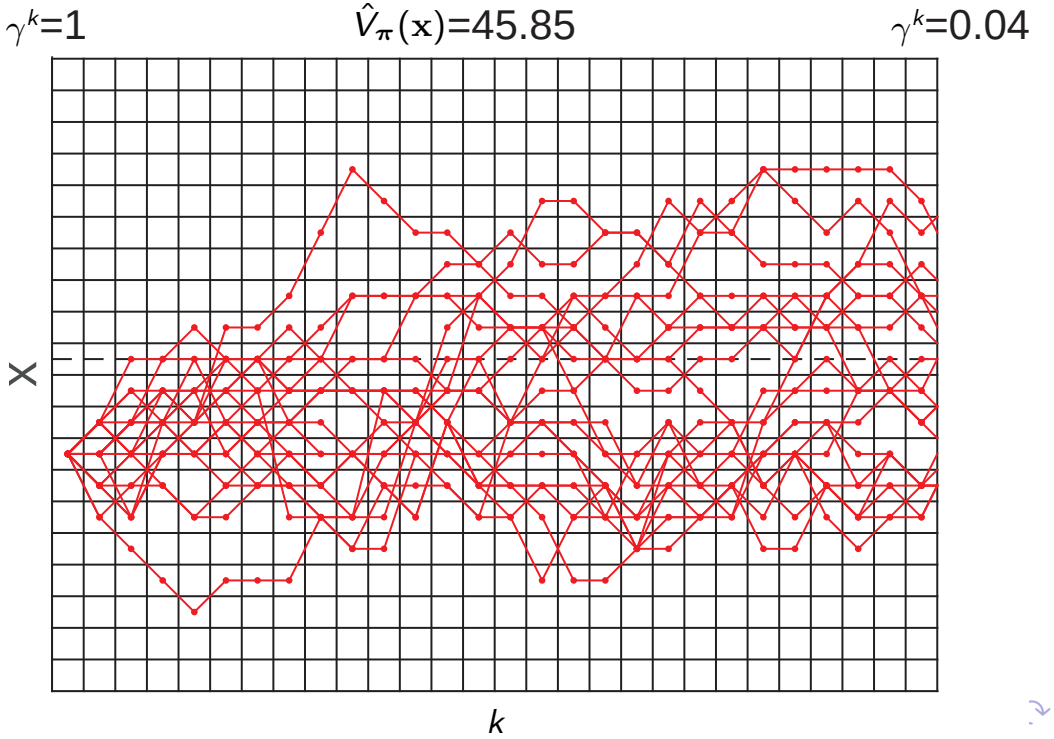
- Start the system in state $\mathbf{x}$
- Unroll N trajectories using $\boldsymbol{\pi}(\mathbf{x})$
- Collect trajectories $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$
- Return:

$$R_i = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k^{(i)}, \boldsymbol{\pi}(\mathbf{x}_k^{(i)}))$$

- Evaluate empirical mean:

$$\hat{V}_{\pi}(\mathbf{x}) \approx \frac{1}{N} \sum_{k=1}^{N} R_i$$

E.g. for $N = 20$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Monte Carlo (MC) evaluation:

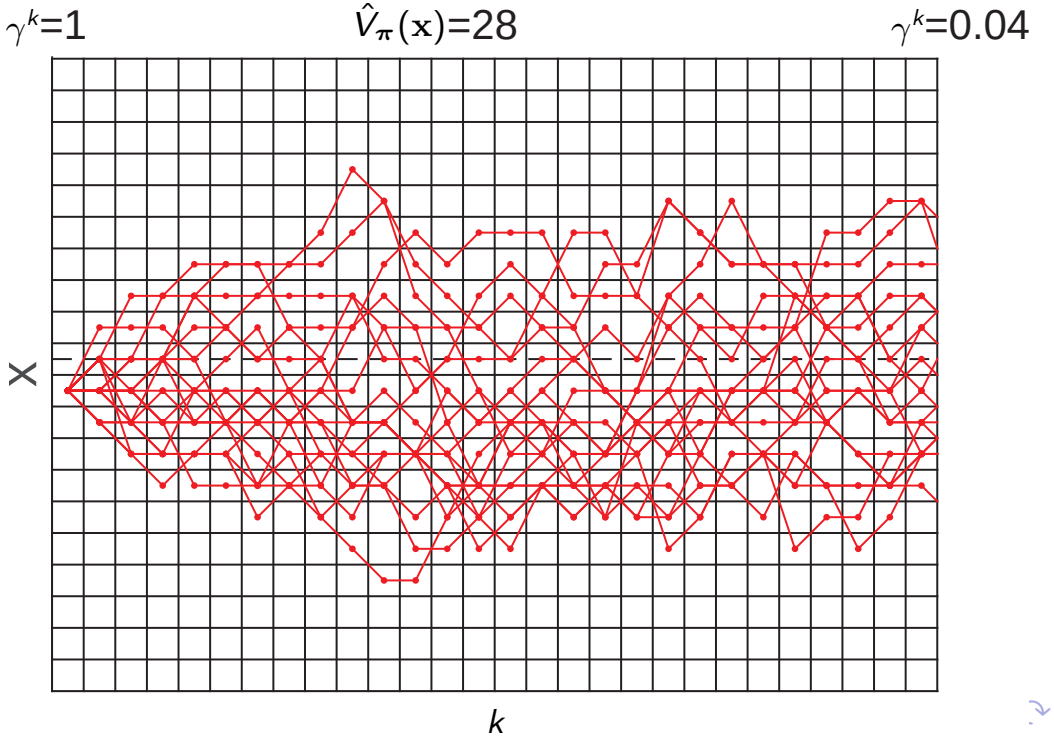
- Start the system in state $\mathbf{x}$
- Unroll N trajectories using $\boldsymbol{\pi}(\mathbf{x})$
- Collect trajectories $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$
- Return:

$$R_i = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k^{(i)}, \boldsymbol{\pi}(\mathbf{x}_k^{(i)}))$$

- Evaluate empirical mean:

$$\hat{V}_{\pi}(\mathbf{x}) \approx \frac{1}{N} \sum_{k=1}^{N} R_i$$

E.g. for $N = 20$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental Monte Carlo (MC)

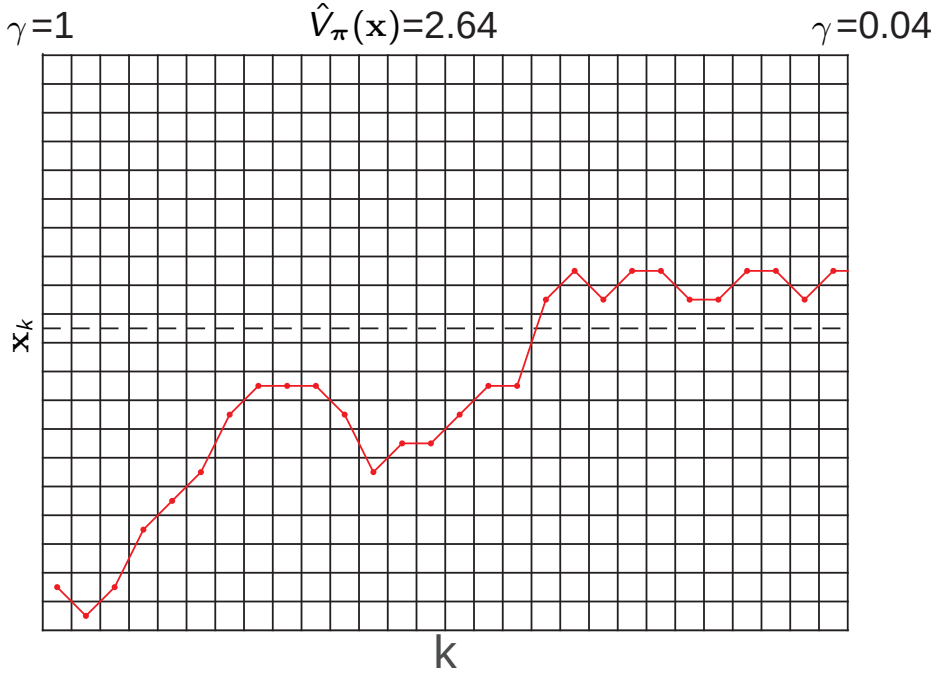
- Start the system in state $\mathbf{x} \equiv \mathbf{x}_0$
- Use $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$
- Update $\hat{V}_{\pi}(\mathbf{x}_0)$ as:

$$\hat{V}_{\pi}(\mathbf{x}_0) \leftarrow \hat{V}_{\pi}(\mathbf{x}_0) + \alpha\left(R - \hat{V}_{\pi}(\mathbf{x}_0)\right)$$

$$R = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Note: R is a noisy estimation of $V_{\pi}(\mathbf{x}_0)$

E.g. $\alpha = 5 \cdot 15^{-3}$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental Monte Carlo (MC)

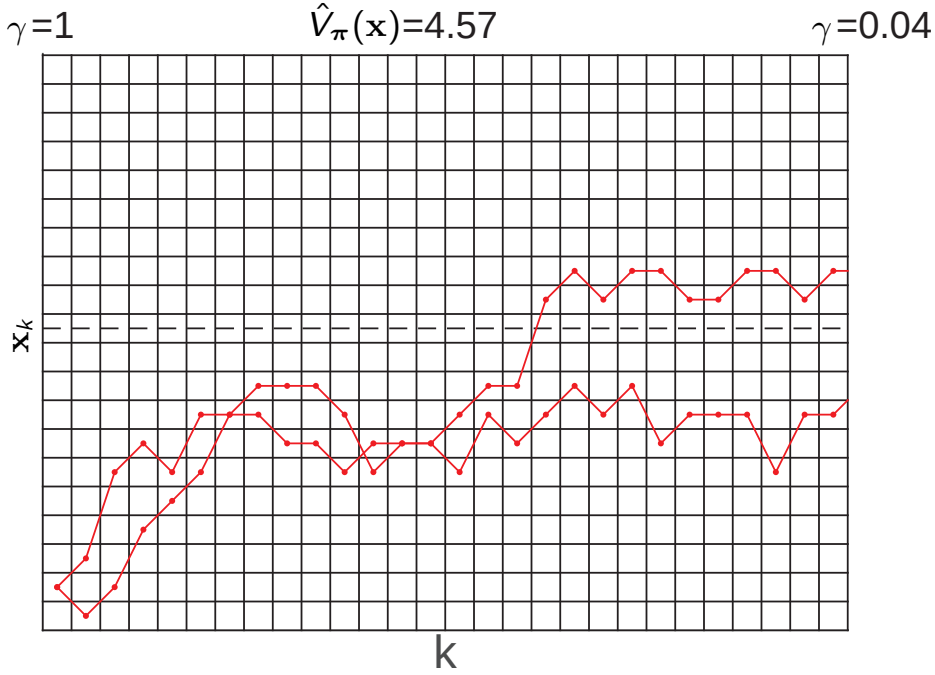
- Start the system in state $\mathbf{x} \equiv \mathbf{x}_0$
- Use $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$
- Update $\hat{V}_{\pi}(\mathbf{x}_0)$ as:

$$\hat{V}_{\pi}(\mathbf{x}_0) \leftarrow \hat{V}_{\pi}(\mathbf{x}_0) + \alpha\left(R - \hat{V}_{\pi}(\mathbf{x}_0)\right)$$

$$R = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Note: R is a noisy estimation of $V_{\pi}(\mathbf{x}_0)$

E.g. $\alpha = 5 \cdot 15^{-3}$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental Monte Carlo (MC)

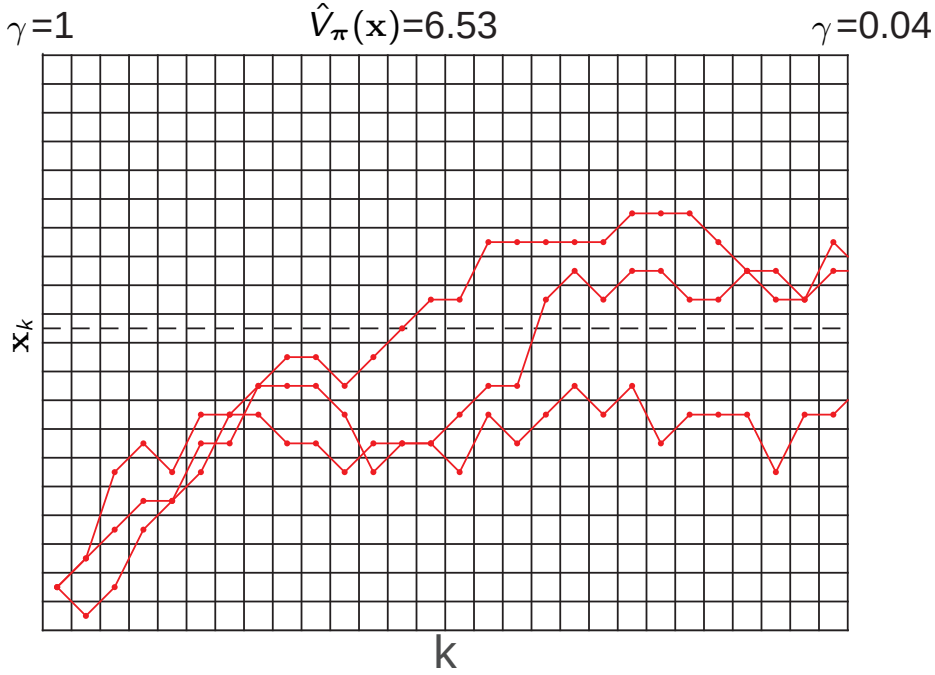
- Start the system in state $\mathbf{x} \equiv \mathbf{x}_0$
- Use $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$
- Update $\hat{V}_{\pi}(\mathbf{x}_0)$ as:

$$\hat{V}_{\pi}(\mathbf{x}_0) \leftarrow \hat{V}_{\pi}(\mathbf{x}_0) + \alpha\left(R - \hat{V}_{\pi}(\mathbf{x}_0)\right)$$

$$R = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Note: R is a noisy estimation of $V_{\pi}(\mathbf{x}_0)$

E.g. $\alpha = 5 \cdot 15^{-3}$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental Monte Carlo (MC)

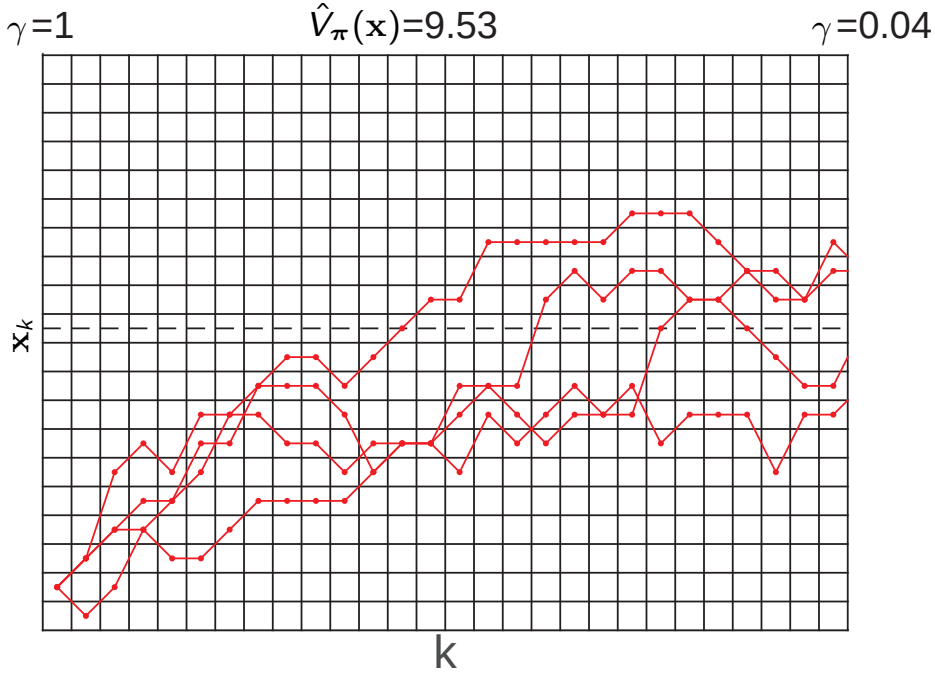
- Start the system in state $\mathbf{x} \equiv \mathbf{x}_0$
- Use $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$
- Update $\hat{V}_{\pi}(\mathbf{x}_0)$ as:

$$\hat{V}_{\pi}(\mathbf{x}_0) \leftarrow \hat{V}_{\pi}(\mathbf{x}_0) + \alpha\left(R - \hat{V}_{\pi}(\mathbf{x}_0)\right)$$

$$R = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Note: R is a noisy estimation of $V_{\pi}(\mathbf{x}_0)$

E.g. $\alpha = 5 \cdot 15^{-3}$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental Monte Carlo (MC)

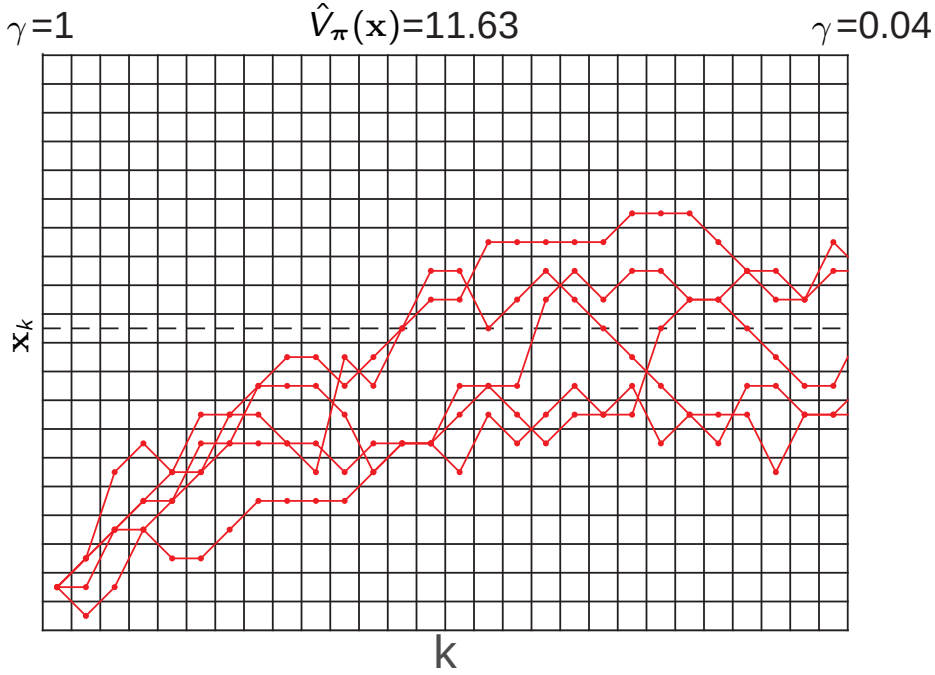
- Start the system in state $\mathbf{x} \equiv \mathbf{x}_0$
- Use $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$
- Update $\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_0)$ as:

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_0) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_0) + \alpha\left(R - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_0)\right)$$

$$R = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Note: R is a noisy estimation of $V_{\pi}(\mathbf{x}_0)$

E.g. $\alpha = 5 \cdot 15^{-3}$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental Monte Carlo (MC)

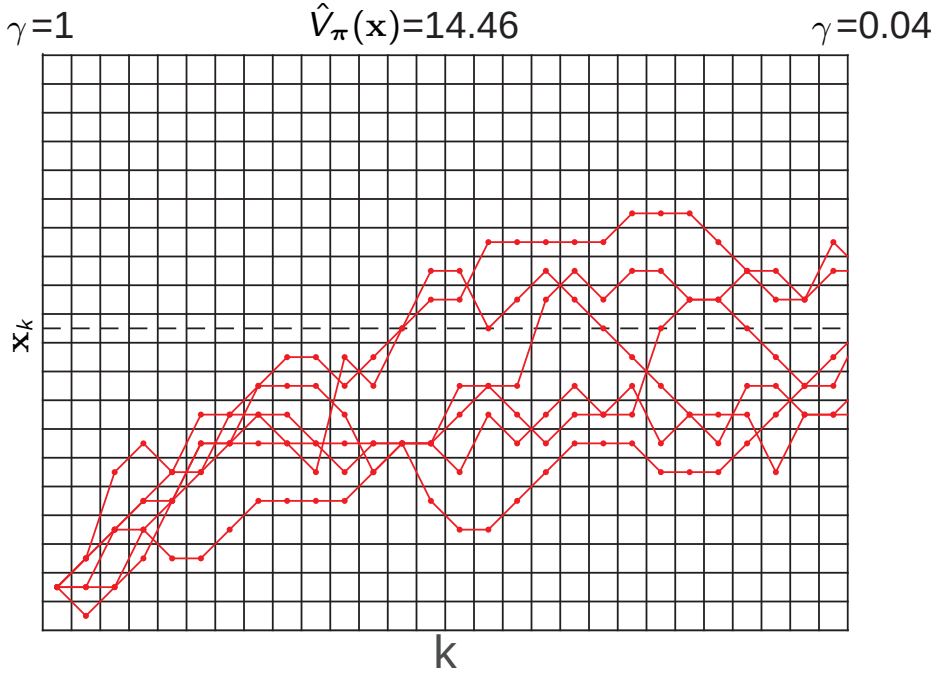
- Start the system in state $\mathbf{x} \equiv \mathbf{x}_0$
- Use $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$
- Update $\hat{V}_{\pi}(\mathbf{x}_0)$ as:

$$\hat{V}_{\pi}(\mathbf{x}_0) \leftarrow \hat{V}_{\pi}(\mathbf{x}_0) + \alpha\left(R - \hat{V}_{\pi}(\mathbf{x}_0)\right)$$

$$R = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Note: R is a noisy estimation of $V_{\pi}(\mathbf{x}_0)$

E.g. $\alpha = 5 \cdot 15^{-3}$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental Monte Carlo (MC)

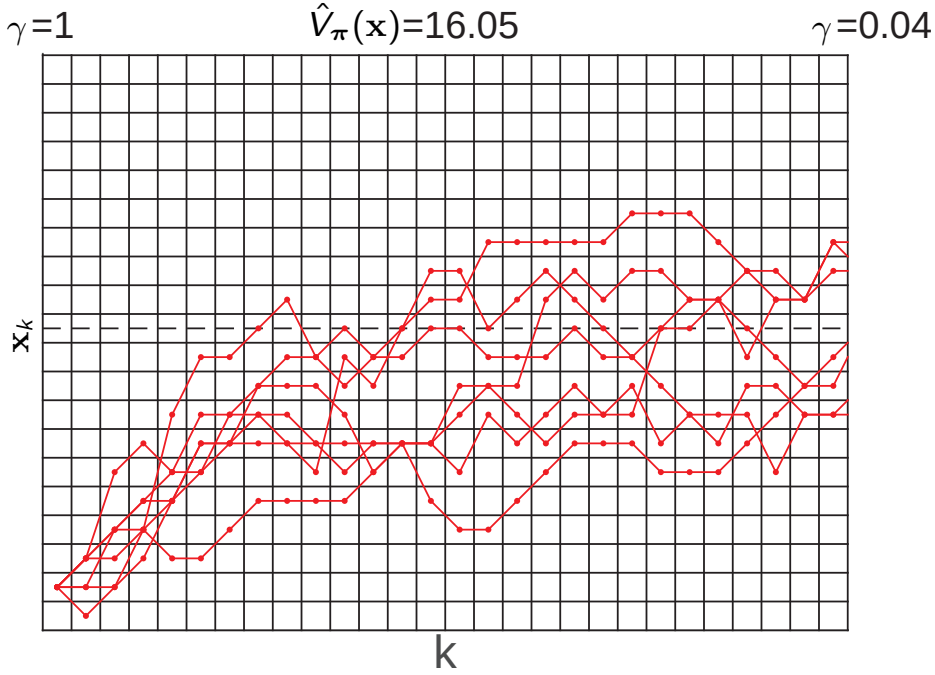
- Start the system in state $\mathbf{x} \equiv \mathbf{x}_0$
- Use $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$
- Update $\hat{V}_{\pi}(\mathbf{x}_0)$ as:

$$\hat{V}_{\pi}(\mathbf{x}_0) \leftarrow \hat{V}_{\pi}(\mathbf{x}_0) + \alpha\left(R - \hat{V}_{\pi}(\mathbf{x}_0)\right)$$

$$R = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Note: R is a noisy estimation of $V_{\pi}(\mathbf{x}_0)$

E.g. $\alpha = 5 \cdot 15^{-3}$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental Monte Carlo (MC)

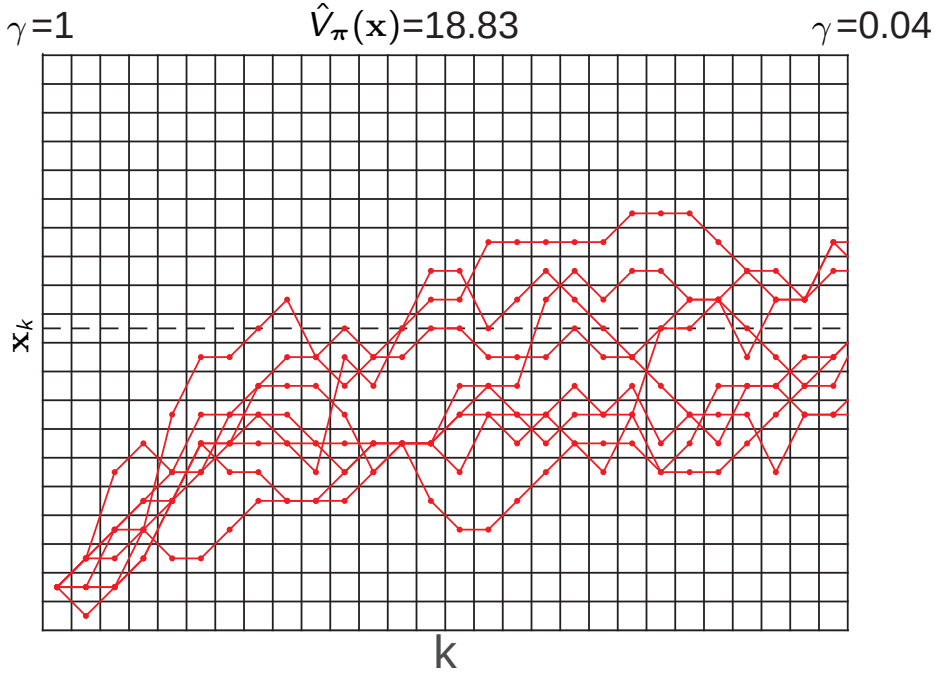
- Start the system in state $\mathbf{x} \equiv \mathbf{x}_0$
- Use $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$
- Update $\hat{V}_{\pi}(\mathbf{x}_0)$ as:

$$\hat{V}_{\pi}(\mathbf{x}_0) \leftarrow \hat{V}_{\pi}(\mathbf{x}_0) + \alpha\left(R - \hat{V}_{\pi}(\mathbf{x}_0)\right)$$

$$R = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Note: R is a noisy estimation of $V_{\pi}(\mathbf{x}_0)$

E.g. $\alpha = 5 \cdot 15^{-3}$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental Monte Carlo (MC)

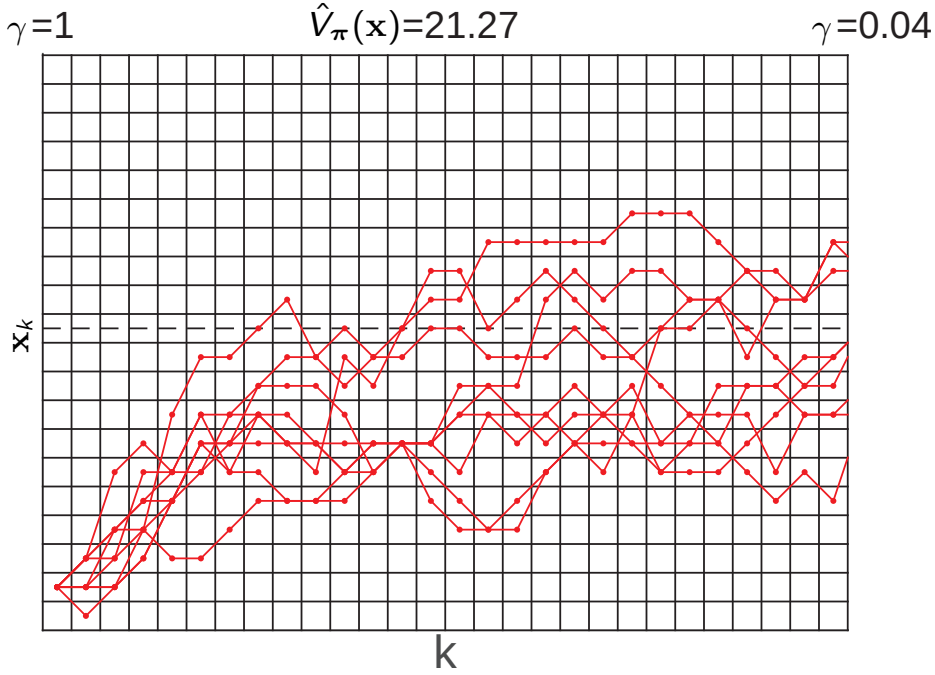
- Start the system in state $\mathbf{x} \equiv \mathbf{x}_0$
- Use $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$
- Update $\hat{V}_{\pi}(\mathbf{x}_0)$ as:

$$\hat{V}_{\pi}(\mathbf{x}_0) \leftarrow \hat{V}_{\pi}(\mathbf{x}_0) + \alpha\left(R - \hat{V}_{\pi}(\mathbf{x}_0)\right)$$

$$R = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Note: R is a noisy estimation of $V_{\pi}(\mathbf{x}_0)$

E.g. $\alpha = 5 \cdot 15^{-3}$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental Monte Carlo (MC)

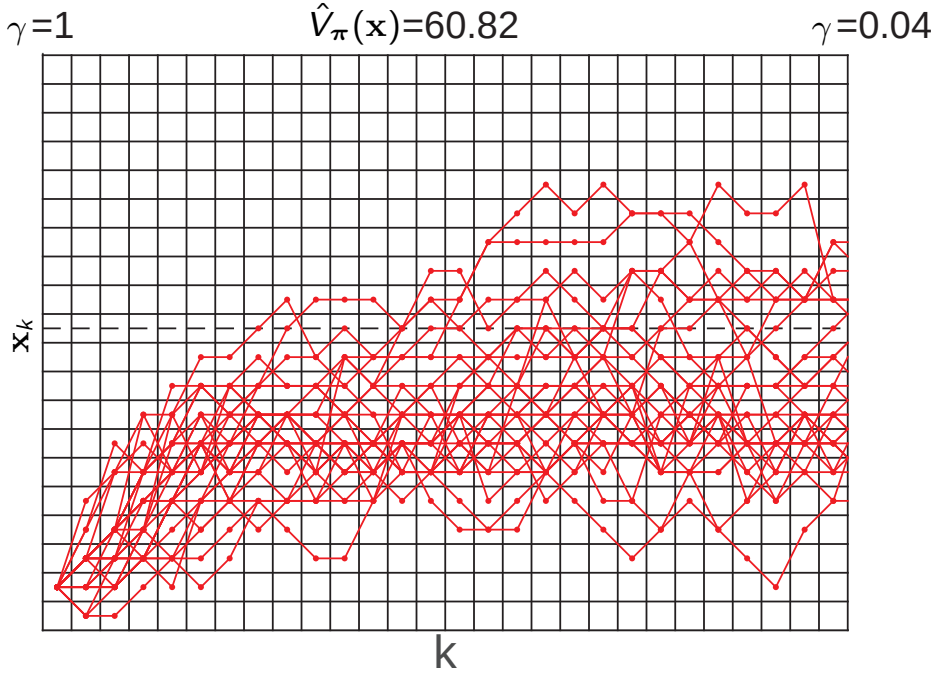
- Start the system in state $\mathbf{x} \equiv \mathbf{x}_0$
- Use $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$
- Update $\hat{V}_{\pi}(\mathbf{x}_0)$ as:

$$\hat{V}_{\pi}(\mathbf{x}_0) \leftarrow \hat{V}_{\pi}(\mathbf{x}_0) + \alpha\left(R - \hat{V}_{\pi}(\mathbf{x}_0)\right)$$

$$R = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Note: R is a noisy estimation of $V_{\pi}(\mathbf{x}_0)$

E.g. $\alpha = 5 \cdot 15^{-3}$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental Monte Carlo (MC)

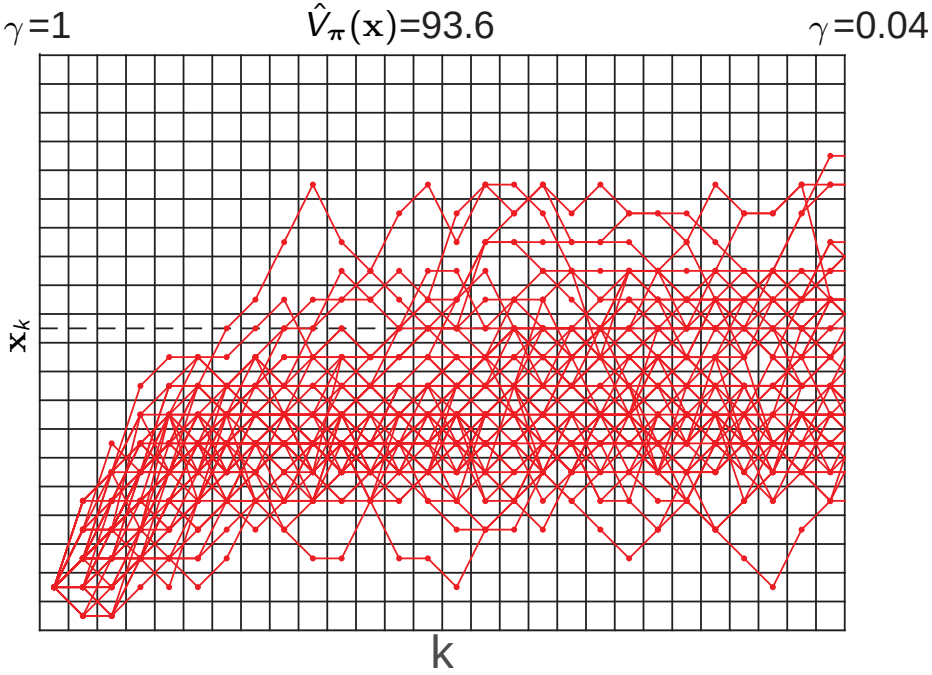
- Start the system in state $\mathbf{x} \equiv \mathbf{x}_0$
- Use $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$
- Update $\hat{V}_{\pi}(\mathbf{x}_0)$ as:

$$\hat{V}_{\pi}(\mathbf{x}_0) \leftarrow \hat{V}_{\pi}(\mathbf{x}_0) + \alpha\left(R - \hat{V}_{\pi}(\mathbf{x}_0)\right)$$

$$R = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Note: R is a noisy estimation of $V_{\pi}(\mathbf{x}_0)$

E.g. $\alpha = 5 \cdot 15^{-3}$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental Monte Carlo (MC)

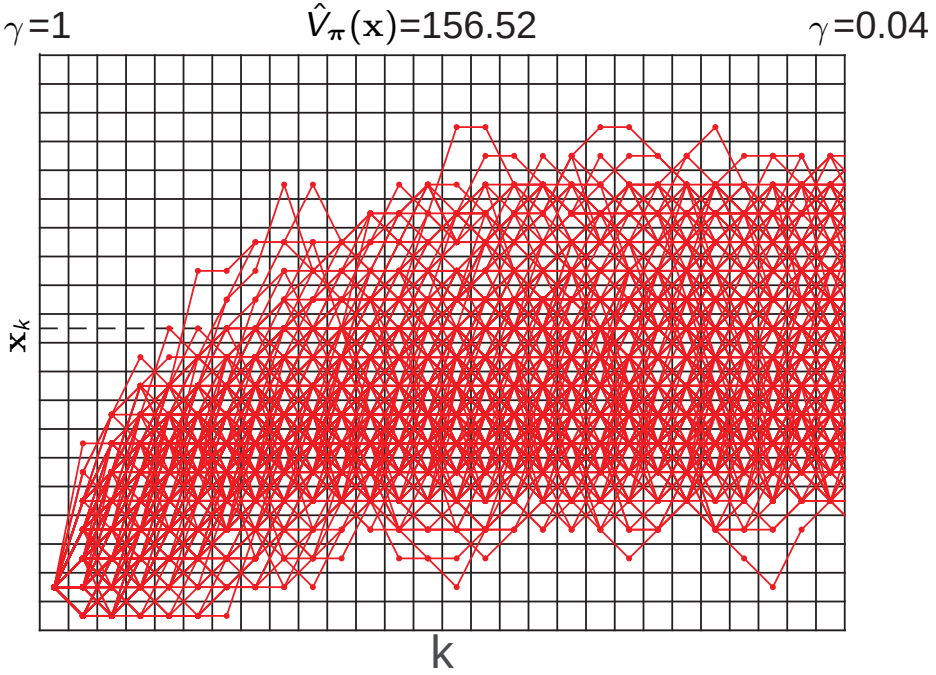
- Start the system in state $\mathbf{x} \equiv \mathbf{x}_0$
- Use $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$
- Update $\hat{V}_{\pi}(\mathbf{x}_0)$ as:

$$\hat{V}_{\pi}(\mathbf{x}_0) \leftarrow \hat{V}_{\pi}(\mathbf{x}_0) + \alpha\left(R - \hat{V}_{\pi}(\mathbf{x}_0)\right)$$

$$R = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Note: R is a noisy estimation of $V_{\pi}(\mathbf{x}_0)$

E.g. $\alpha = 5 \cdot 15^{-3}$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental Monte Carlo (MC)

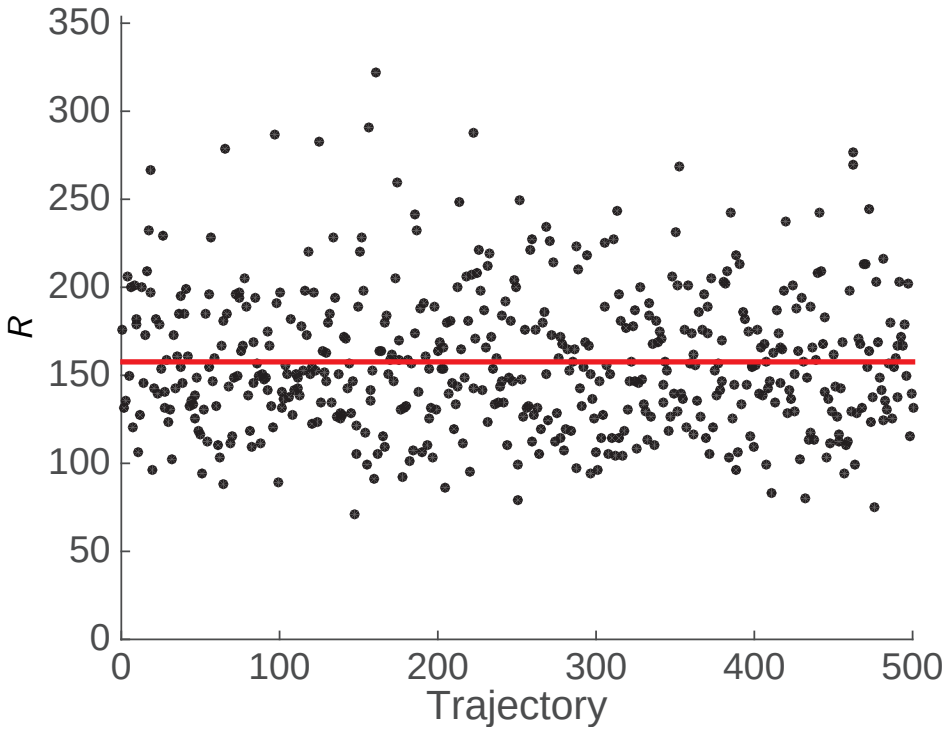
- Start the system in state $\mathbf{x} \equiv \mathbf{x}_0$
- Use $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$
- Update $\hat{V}_{\pi}(\mathbf{x}_0)$ as:

$$\hat{V}_{\pi}(\mathbf{x}_0) \leftarrow \hat{V}_{\pi}(\mathbf{x}_0) + \alpha\left(R - \hat{V}_{\pi}(\mathbf{x}_0)\right)$$

$$R = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Note: R is a noisy estimation of $V_{\pi}(\mathbf{x}_0)$

E.g. $\alpha = 5 \cdot 15^{-3}$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental Monte Carlo (MC)

- Start the system in state $\mathbf{x} \equiv \mathbf{x}_0$
- Use $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$
- Update $\hat{V}_{\pi}(\mathbf{x}_0)$ as:

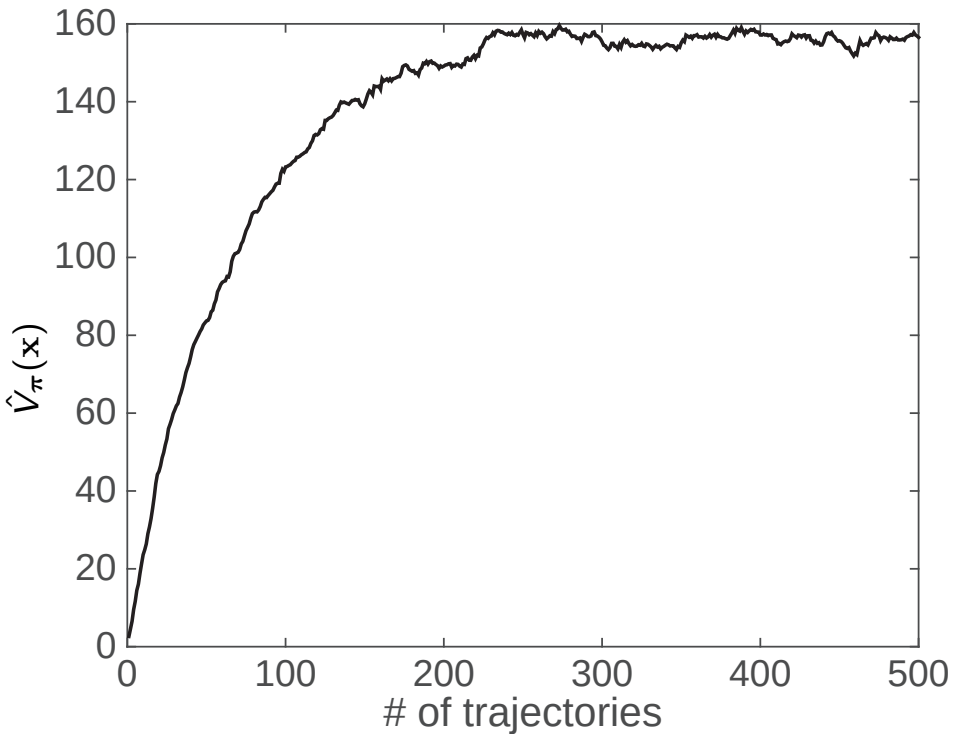
$$\hat{V}_{\pi}(\mathbf{x}_0) \leftarrow \hat{V}_{\pi}(\mathbf{x}_0) + \alpha\left(R - \hat{V}_{\pi}(\mathbf{x}_0)\right)$$

$$R = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Note: R is a noisy estimation of $V_{\pi}(\mathbf{x}_0)$

Can build "on-the-fly" version

E.g. $\alpha = 5 \cdot 15^{-3}$



Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x} \right]$$

Incremental MC

$$\hat{V}_{\pi}(\mathbf{x}) \leftarrow \hat{V}_{\pi}(\mathbf{x}) + \alpha \left(R - \hat{V}_{\pi}(\mathbf{x}) \right)$$

$$R = \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x} \right]$$

Incremental MC

$$\hat{V}_{\pi}(\mathbf{x}) \leftarrow \hat{V}_{\pi}(\mathbf{x}) + \alpha \left(R - \hat{V}_{\pi}(\mathbf{x}) \right)$$

$$R = L(\mathbf{x}, \boldsymbol{\pi}(\mathbf{x})) + \sum_{k=1}^{\infty} \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k))$$

Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x} \right]$$

Incremental MC

$$\hat{V}_{\pi}(\mathbf{x}) \leftarrow \hat{V}_{\pi}(\mathbf{x}) + \alpha \left(R - \hat{V}_{\pi}(\mathbf{x}) \right)$$

$$R = L(\mathbf{x}, \boldsymbol{\pi}(\mathbf{x})) + \gamma \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_{k+1}, \boldsymbol{\pi}(\mathbf{x}_{k+1}))$$

Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\, \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental MC

$$\hat{V}_{\pi}(\mathbf{x}) \leftarrow \hat{V}_{\pi}(\mathbf{x}) + \alpha\left(R - \hat{V}_{\pi}(\mathbf{x})\right)$$

$$R = L(\mathbf{x}, \boldsymbol{\pi}(\mathbf{x})) + \gamma \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_{k+1}, \boldsymbol{\pi}(\mathbf{x}_{k+1}))$$

Note that:

$$\mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_{k+1}, \boldsymbol{\pi}(\mathbf{x}_{k+1})) \,\middle|\, \mathbf{x}_1\right] = V_{\pi}(\mathbf{x}_1)$$

Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Incremental MC

$$\hat{V}_{\pi}(\mathbf{x}) \leftarrow \hat{V}_{\pi}(\mathbf{x}) + \alpha\left(R - \hat{V}_{\pi}(\mathbf{x})\right)$$

$$R = L(\mathbf{x}, \boldsymbol{\pi}(\mathbf{x})) + \gamma \sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_{k+1}, \boldsymbol{\pi}(\mathbf{x}_{k+1}))$$

Note that:

$$\mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_{k+1}, \boldsymbol{\pi}(\mathbf{x}_{k+1})) \,\middle|\, \mathbf{x}_1\right] = V_{\pi}(\mathbf{x}_1)$$

Temporal difference uses

$$R \approx L(\mathbf{x}, \boldsymbol{\pi}(\mathbf{x})) + \gamma \hat{V}_{\pi}(\mathbf{x}_1)$$

Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x} \right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha \delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x} \right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha \delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

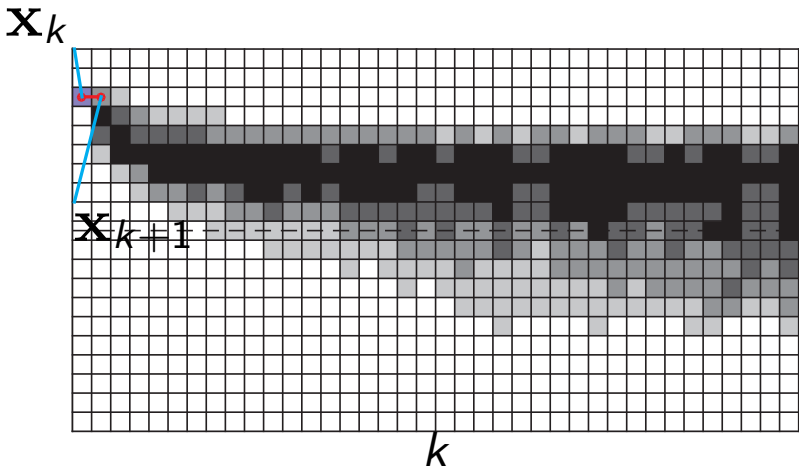
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

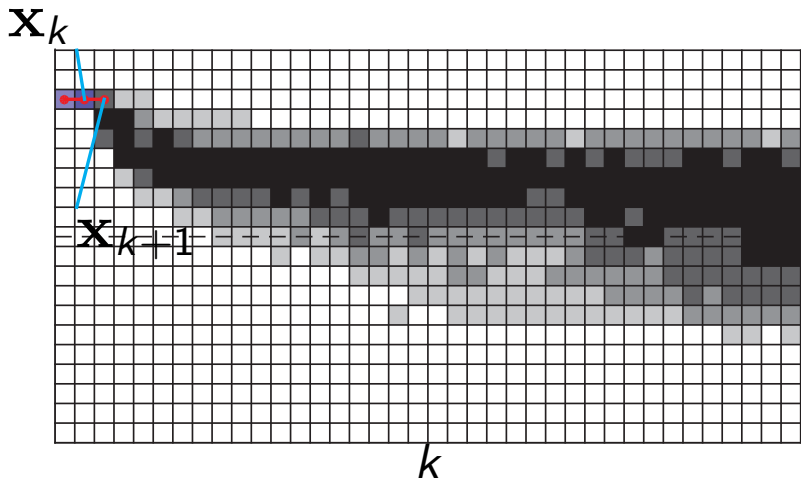
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

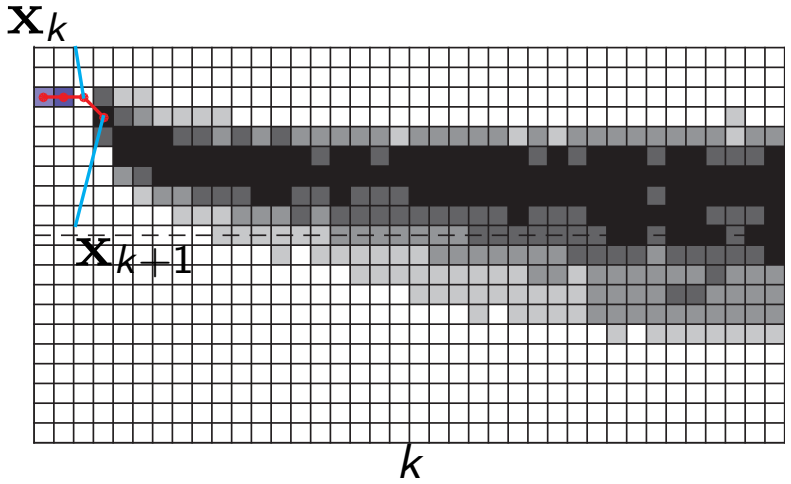
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

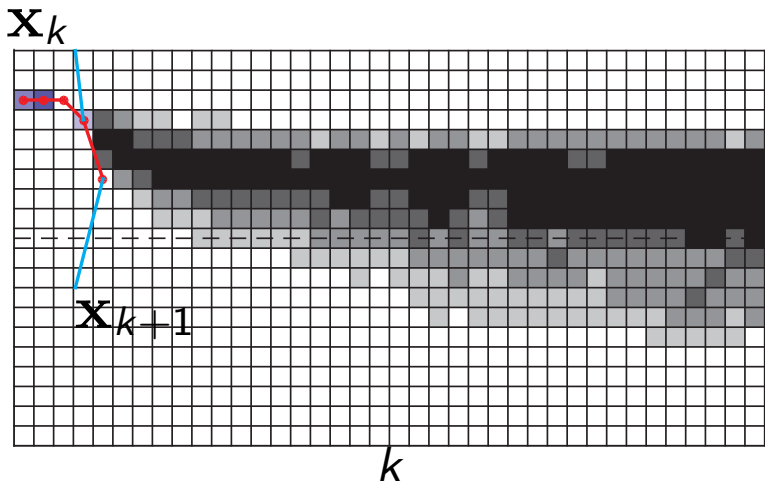
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

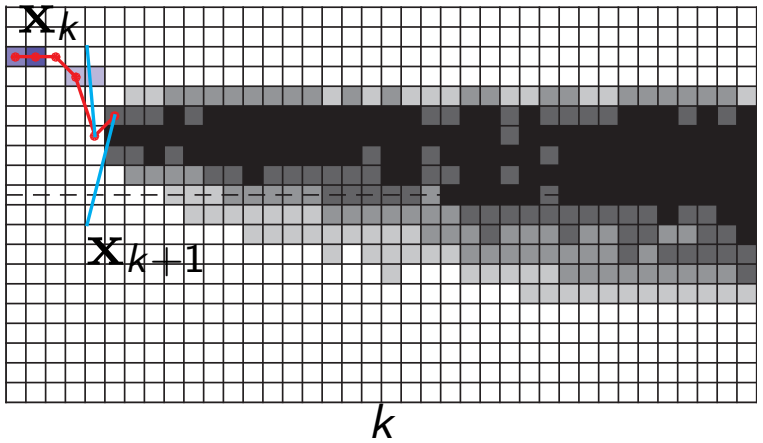
Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$
$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$
$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

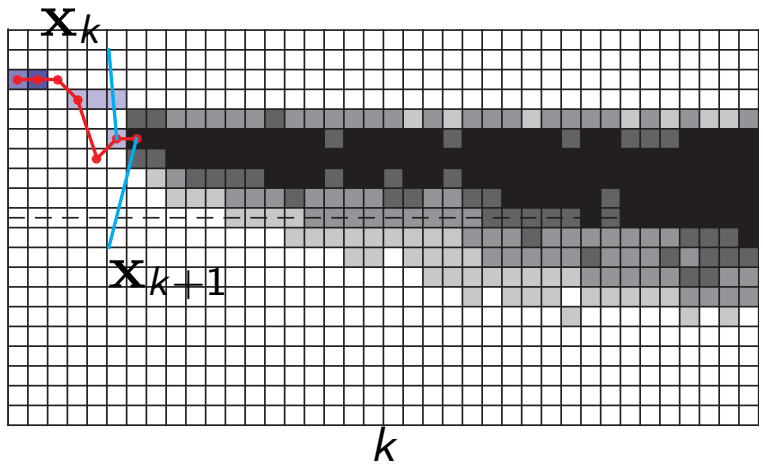
Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$
$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$
$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

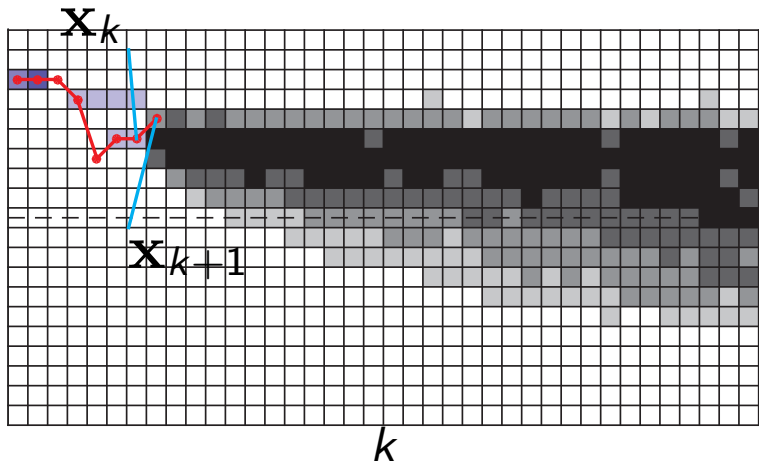
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

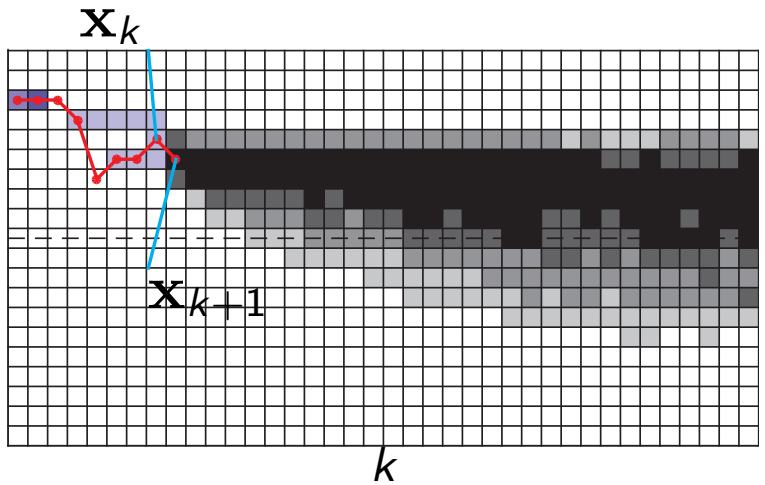
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

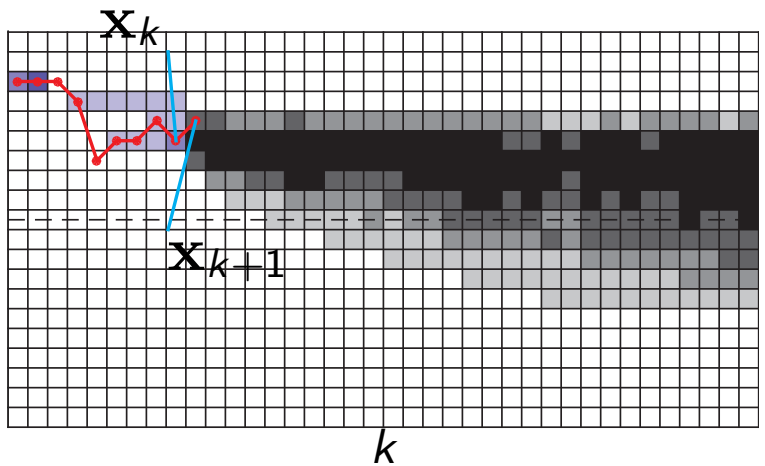
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

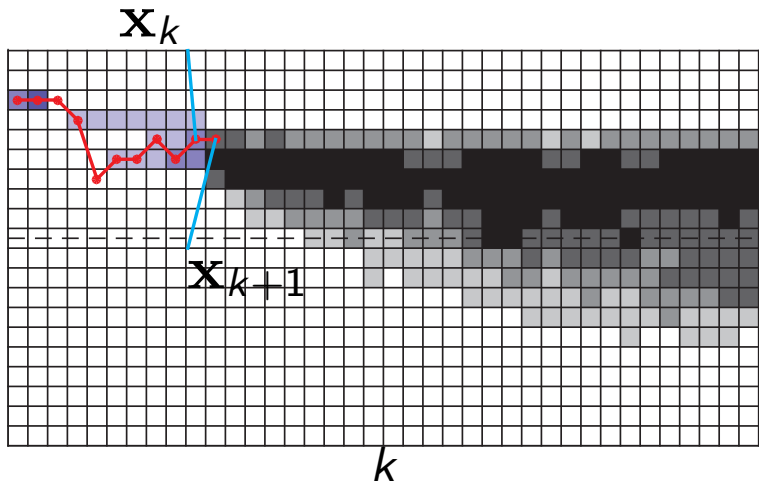
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

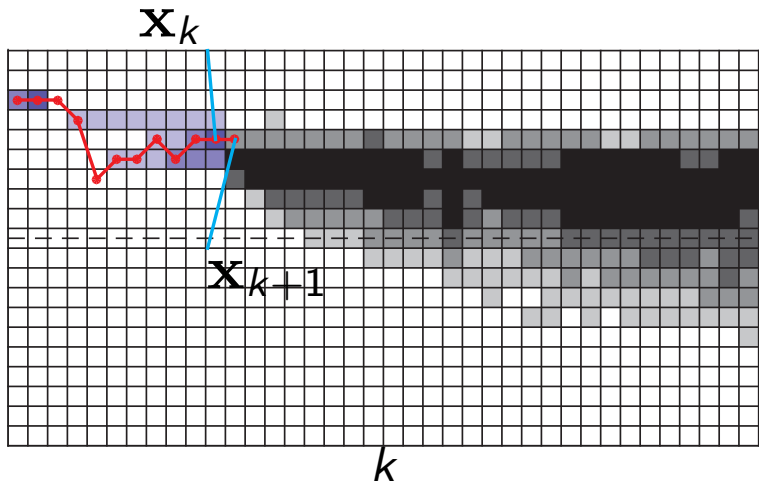
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

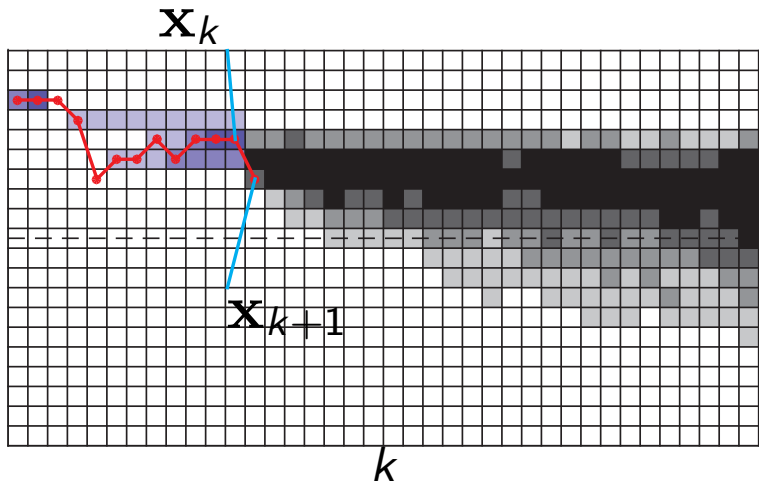
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

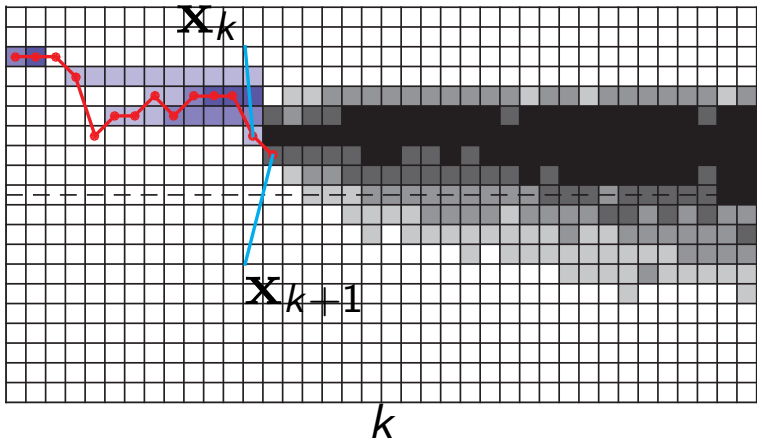
Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$
$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$
$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

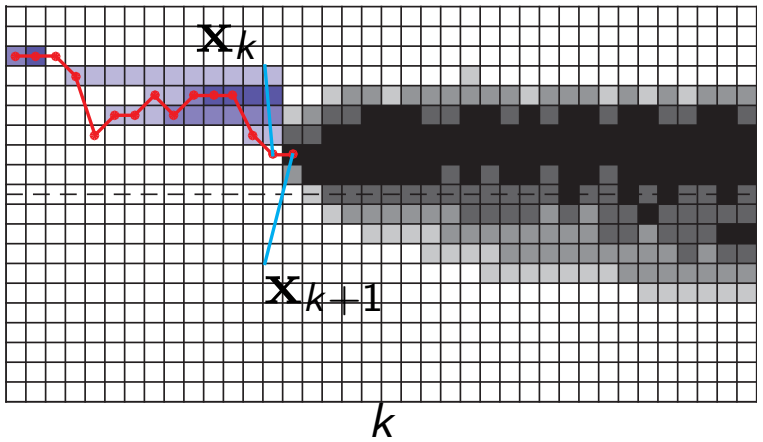
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

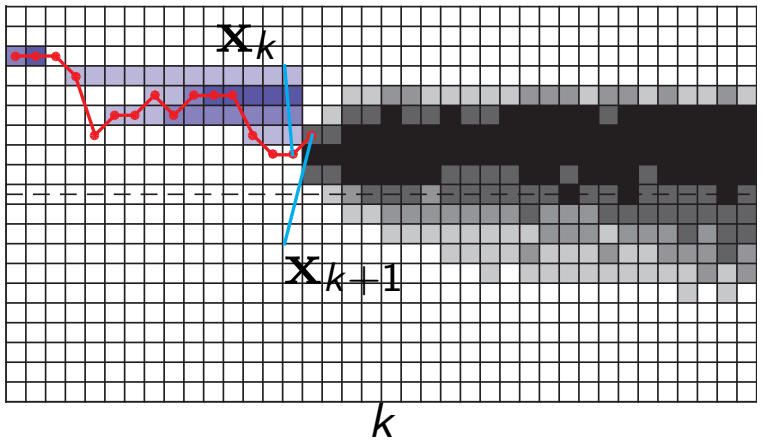
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

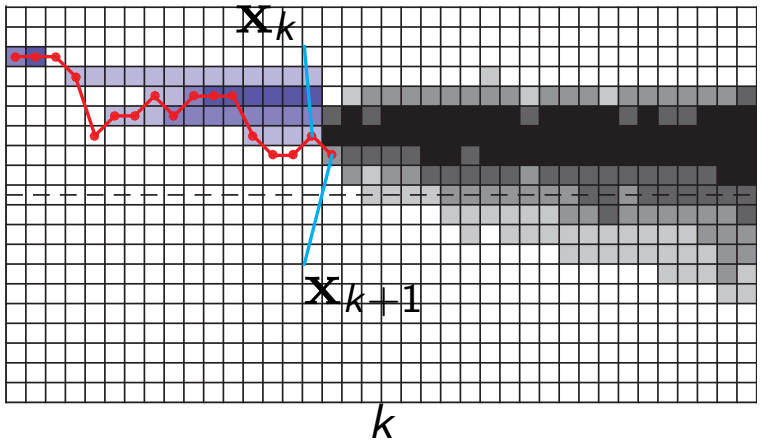
Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$
$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$
$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

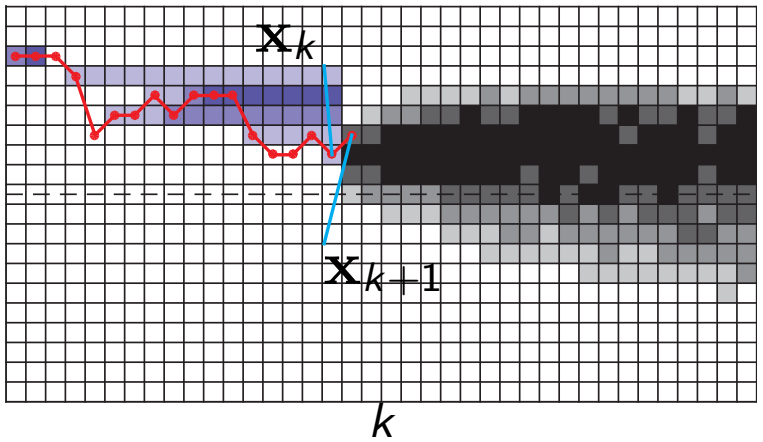
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

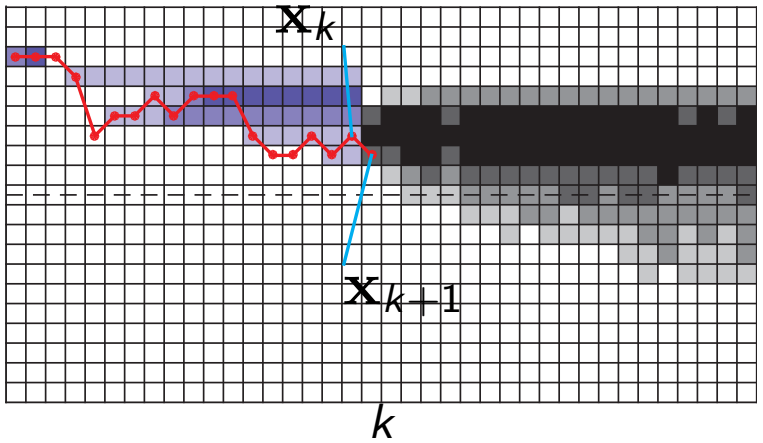
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x} \right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha \delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

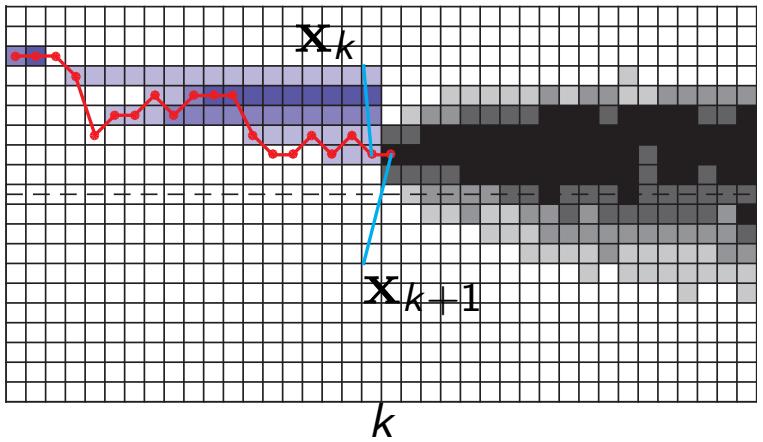
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

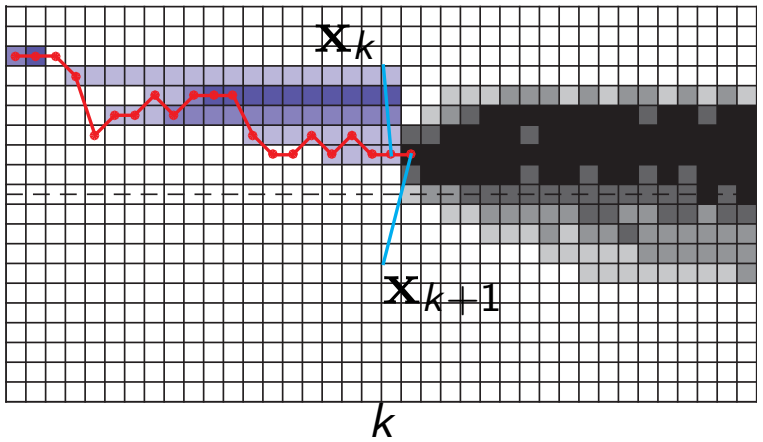
Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$



Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size

Learning the Value function

Value function for policy π:

$$V_\pi(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_\pi(\mathbf{x}_k) \leftarrow \hat{V}_\pi(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_\pi(\mathbf{x}_{k+1}) - \hat{V}_\pi(\mathbf{x}_k)$$

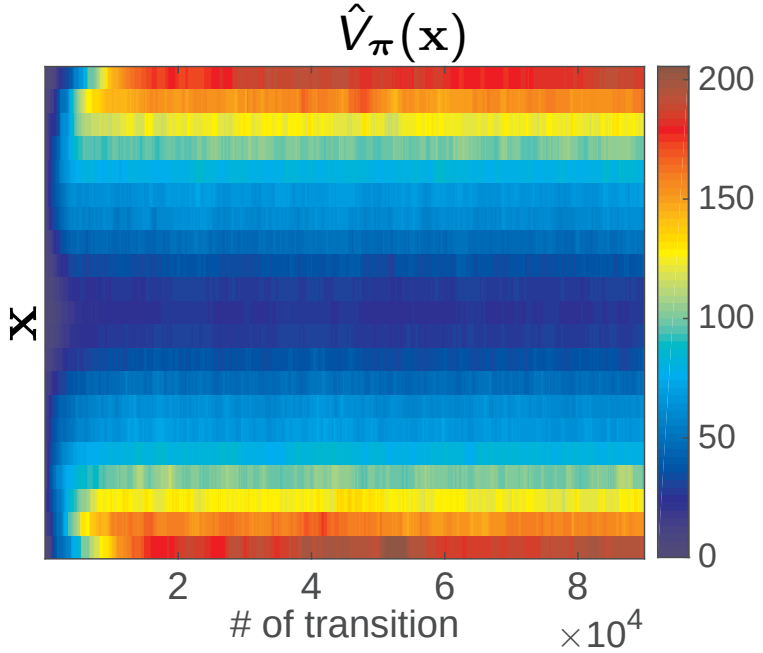
Note: **Policy Evaluation** does

$$V_\pi(\mathbf{x}) \leftarrow V_\pi(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size



10^5 trajectories of length $N = 10$

Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

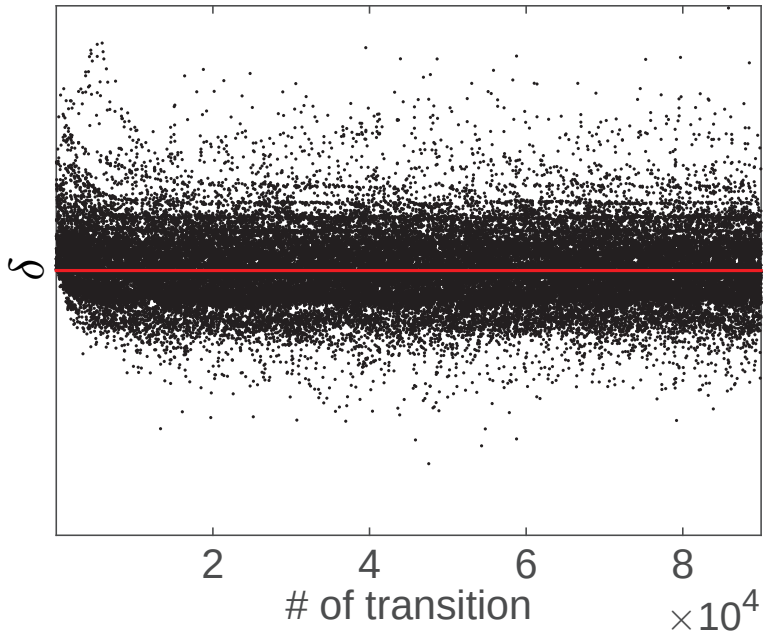
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$
$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$
$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size



10^5 trajectories of length $N = 10$

Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

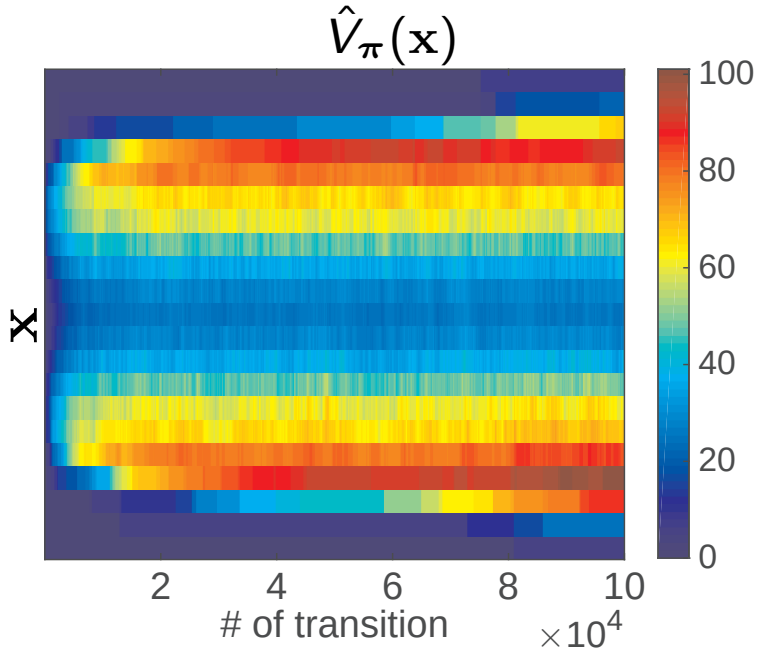
Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size



1 trajectory of length $N = 10^6$

Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

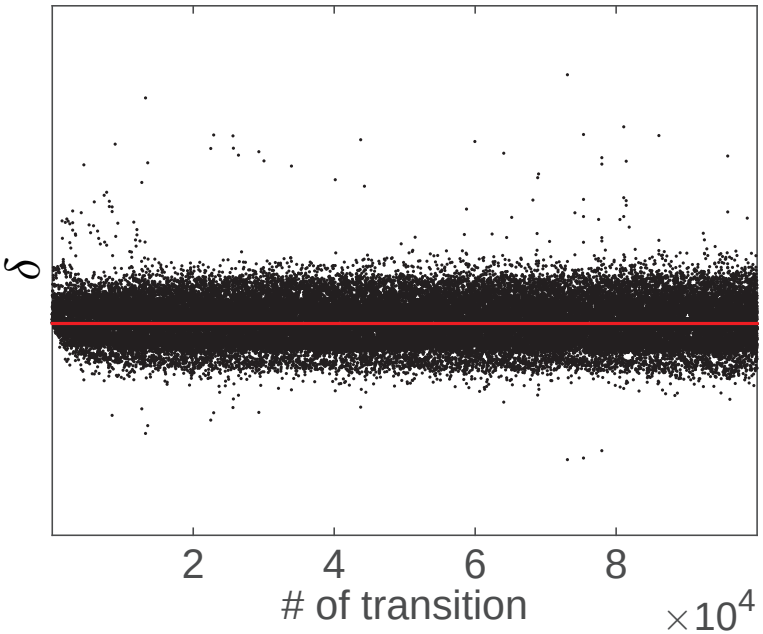
Note: **Policy Evaluation** does

$$V_{\pi}(\mathbf{x}) \leftarrow V_{\pi}(\mathbf{x}) + \bar{\delta}$$

$$\bar{\delta}(\mathbf{x}) = \mathbb{E}\left[\delta(\mathbf{x}, \mathbf{x}_+) \,|\, \mathbf{x}\right], \quad \forall\, \mathbf{x}$$

hence TD uses

- δ as a "noisy" version of $\bar{\delta}$
- $\alpha < 1$ as a "noise filter" & step size



1 trajectory of length $N = 10^6$

Learning the Value function

Value function for policy π:

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k) \,\middle|\, \mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Temporal Difference (TD) uses

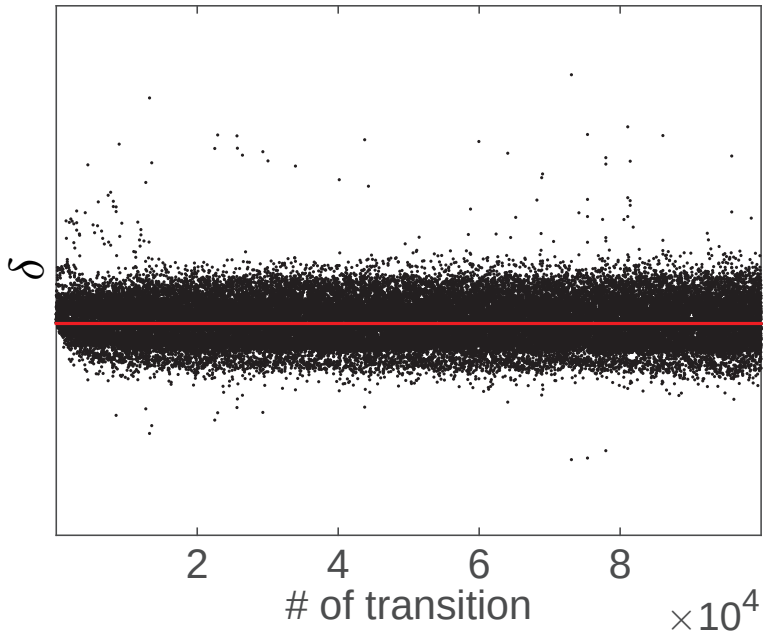
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta(\mathbf{x}_k, \mathbf{x}_{k+1})$$

$$\delta(\mathbf{x}_k, \mathbf{x}_{k+1}) = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

When $\hat{V}_{\pi} = V_{\pi}$

- $\mathbb{E}[\delta \,|\, \mathbf{x}] = 0$ holds
- TD-error δ has zero-mean over trajectories, i.e.

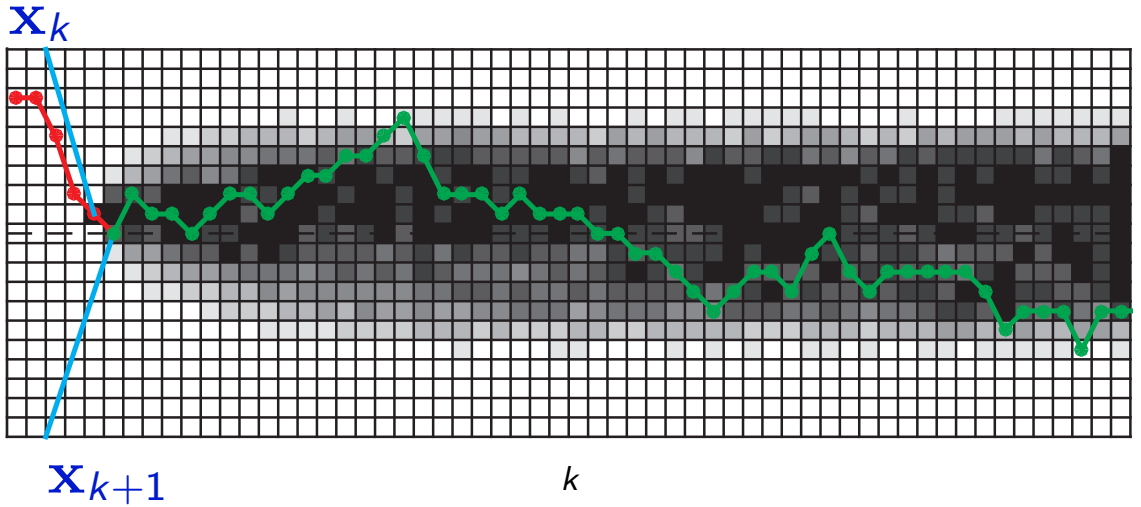
$$\mathbb{E}[\delta(\mathbf{x}_k, \mathbf{x}_{k+1})] = 0$$



Depth of Backup

Backup at time k

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k) + \alpha\delta$$



Incremental Monte Carlo (IMC)

$$\delta = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \sum_{i=1}^{\infty} \gamma^i L(\mathbf{x}_{k+i}, \boldsymbol{\pi}(\mathbf{x}_{k+i})) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

i.e. ∞-deep backup

Temporal Difference (TD)

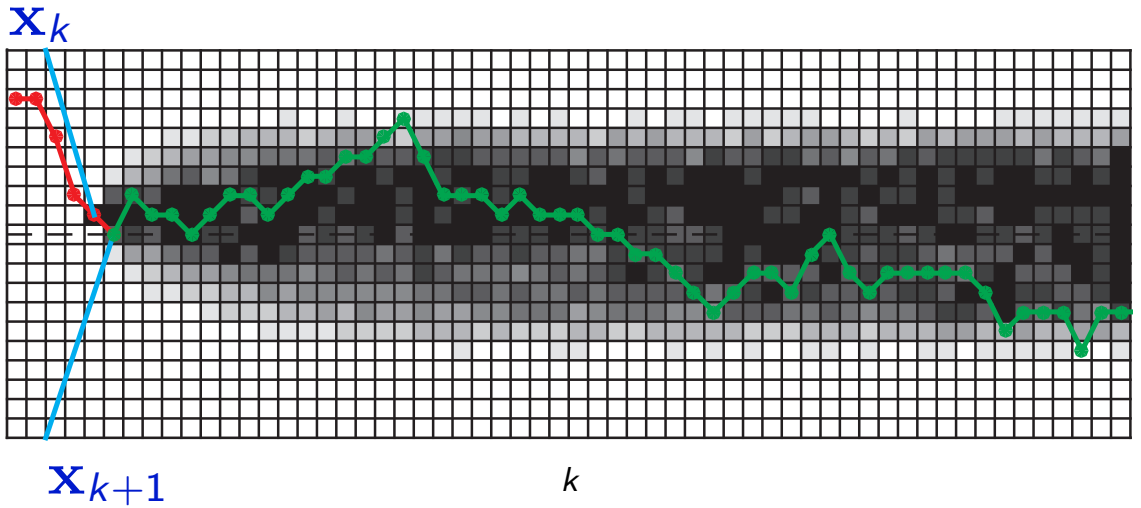
$$\delta = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

i.e. 1-step deep backup

Depth of Backup

Backup at time k

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k) + \alpha\delta$$



Incremental Monte Carlo (IMC)

$$\delta = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \sum_{i=1}^{\infty} \gamma^i L(\mathbf{x}_{k+i}, \boldsymbol{\pi}(\mathbf{x}_{k+i})) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

i.e. ∞-deep backup

Is there something in between 1 and ∞?

Temporal Difference (TD)

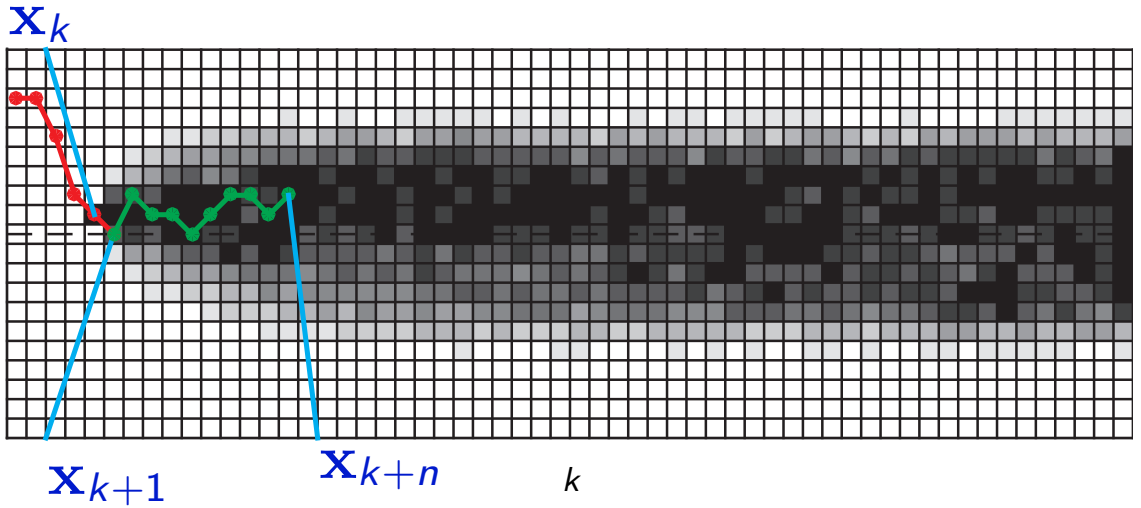
$$\delta = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

i.e. 1-step deep backup

Depth of Backup

Backup at time k

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k) + \alpha\delta$$



Incremental Monte Carlo (IMC)

$$\delta = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \sum_{i=1}^{\infty} \gamma^i L(\mathbf{x}_{k+i}, \boldsymbol{\pi}(\mathbf{x}_{k+i})) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

i.e. ∞-deep backup

Is there something in between 1 and ∞?

Temporal Difference (TD)

$$\delta = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

i.e. 1-step deep backup

n-**step Temporal Difference (TD)** uses

$$\delta = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma^n \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+n}) + \sum_{i=1}^{n-1} \gamma^i L(\mathbf{x}_{k+i}, \boldsymbol{\pi}(\mathbf{x}_{k+i})) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive

TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

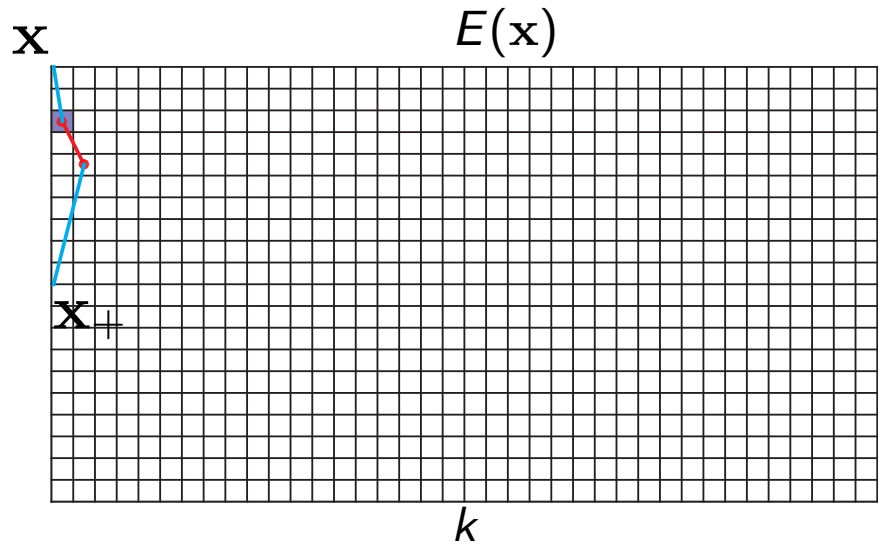
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

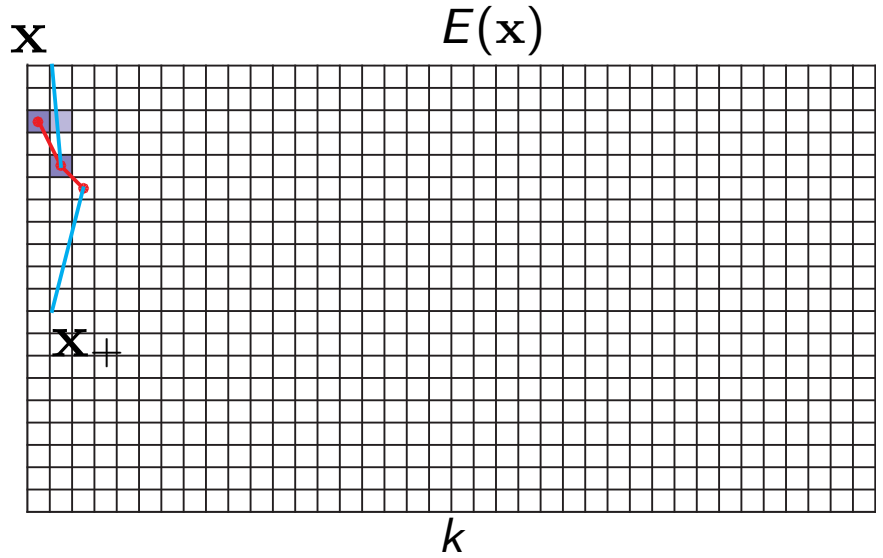
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

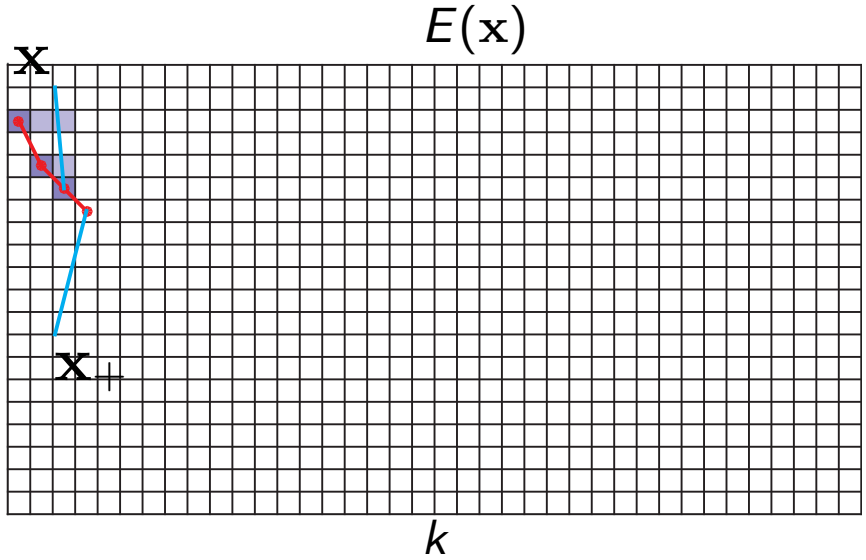
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

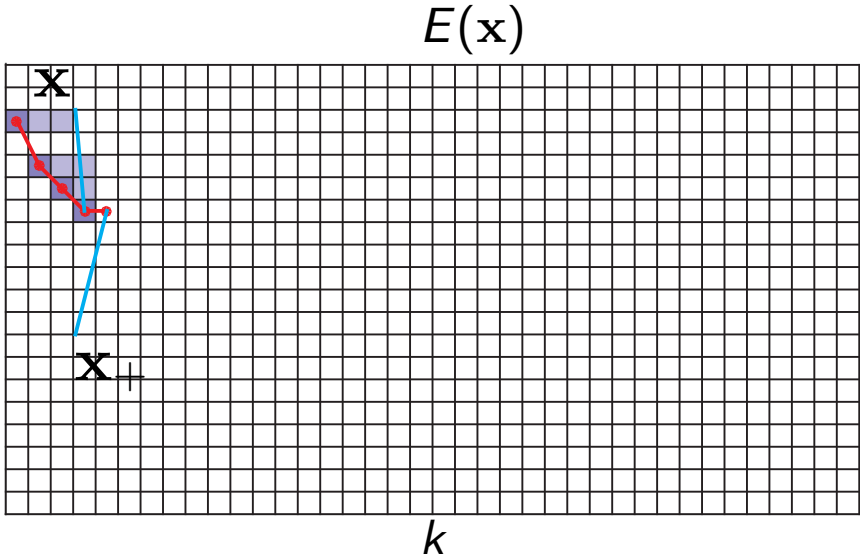
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

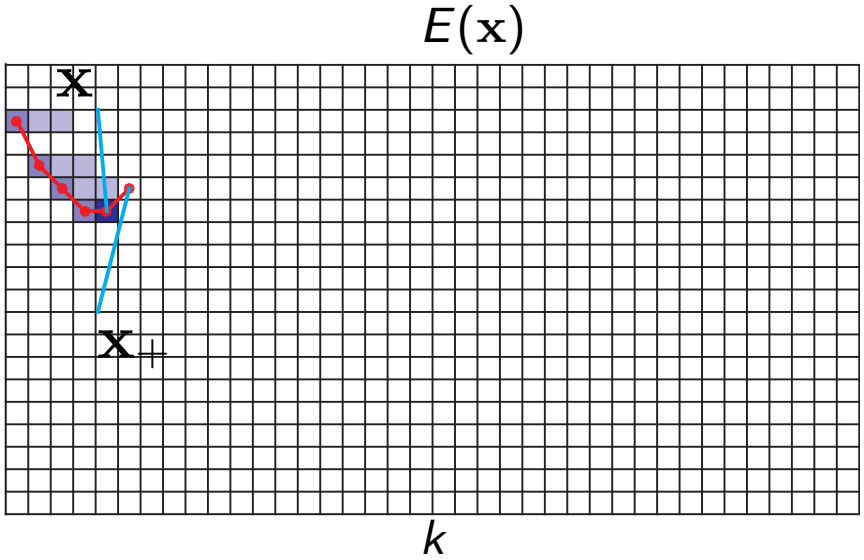
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

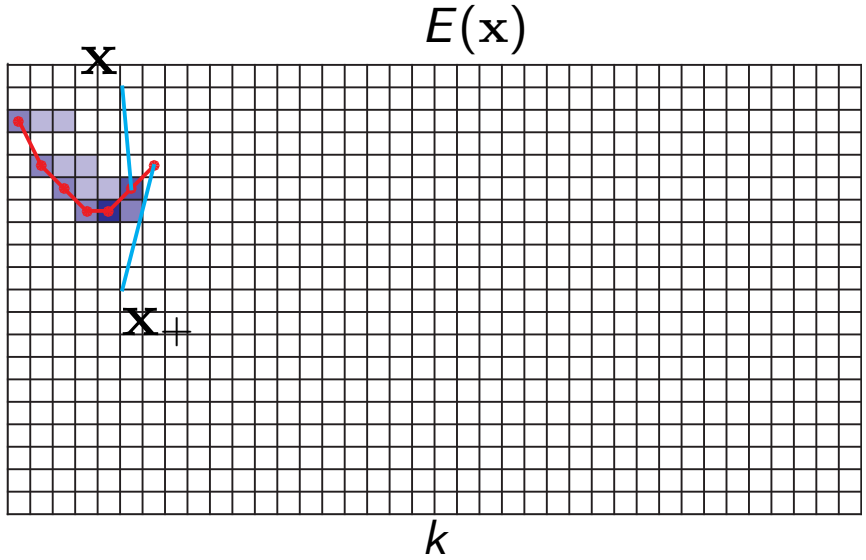
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

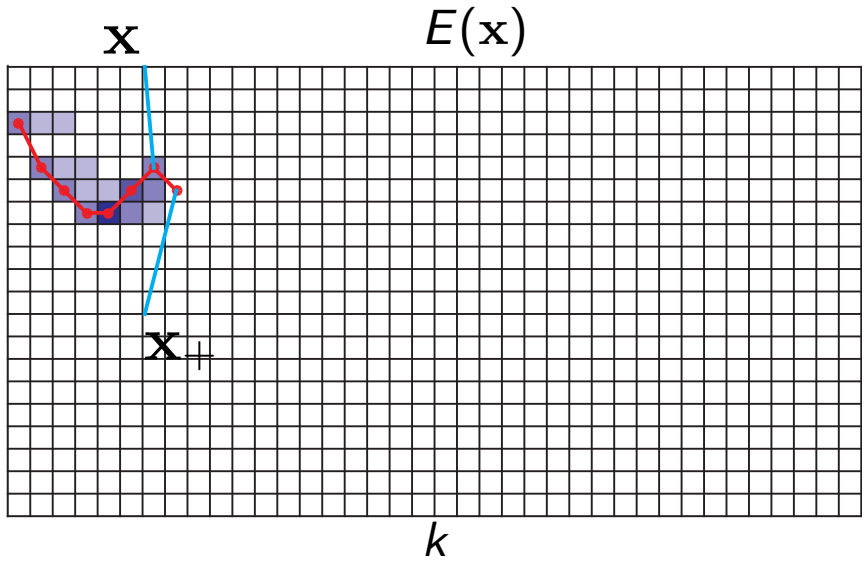
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

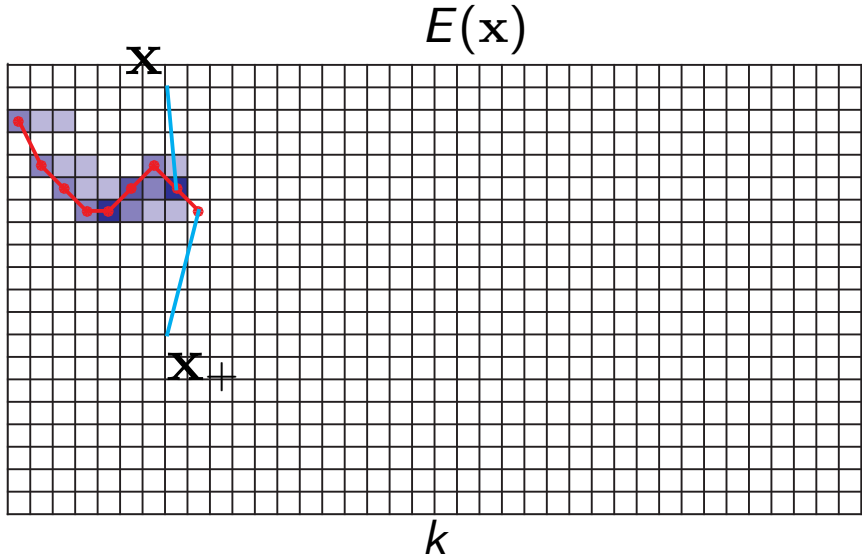
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

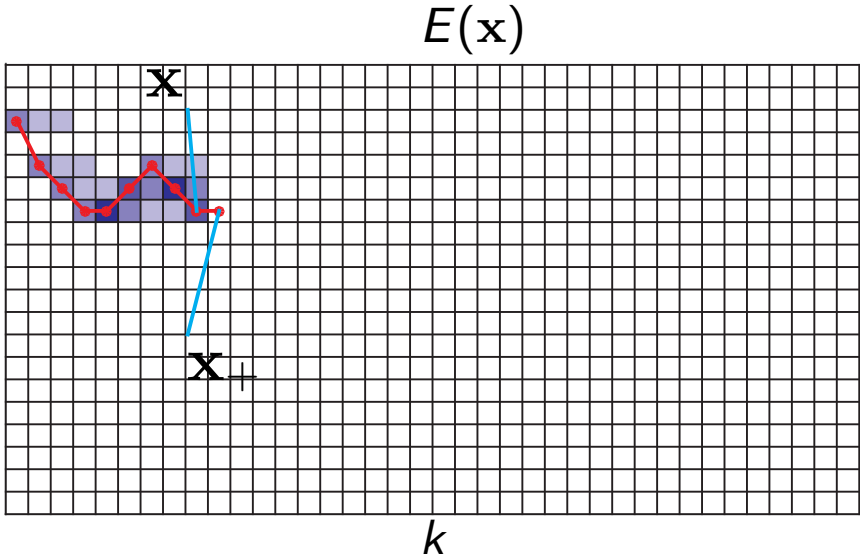
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

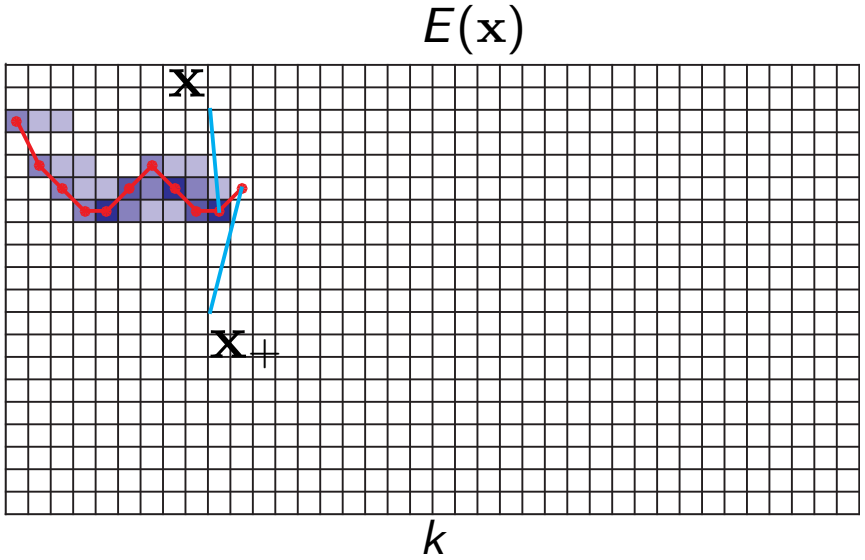
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

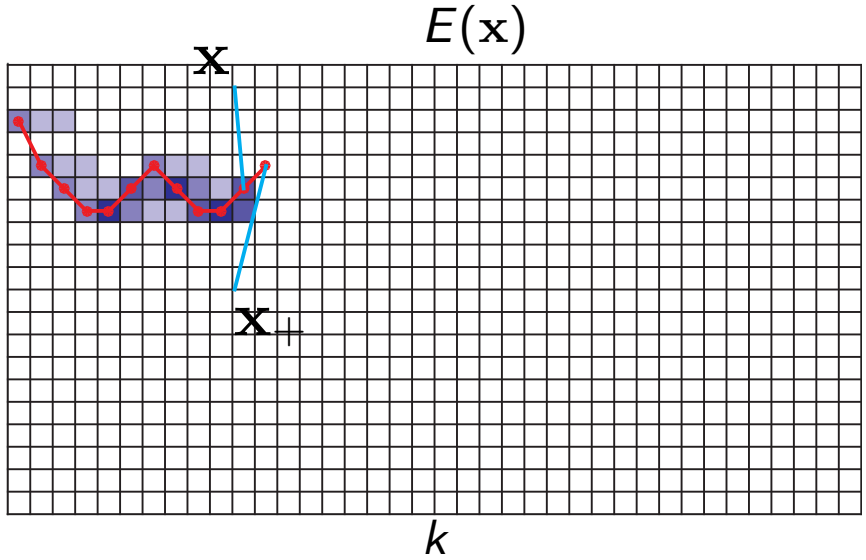
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

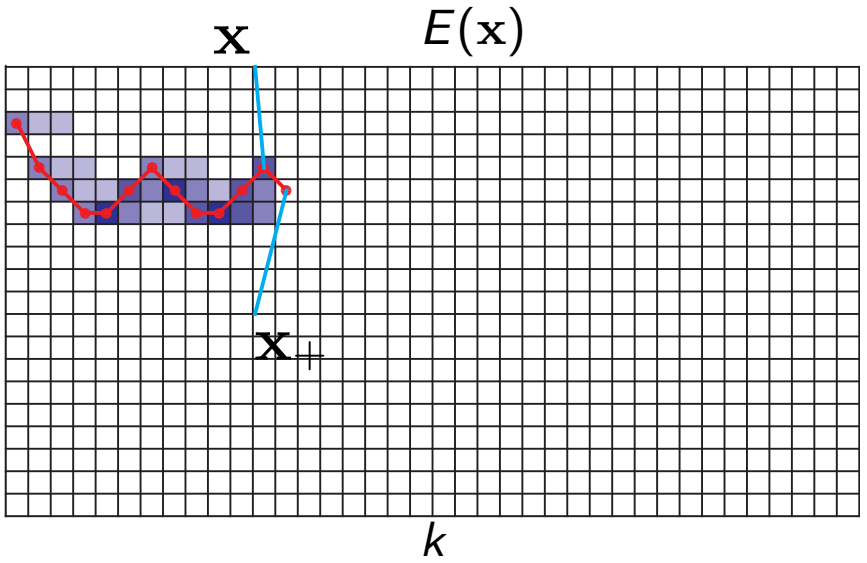
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

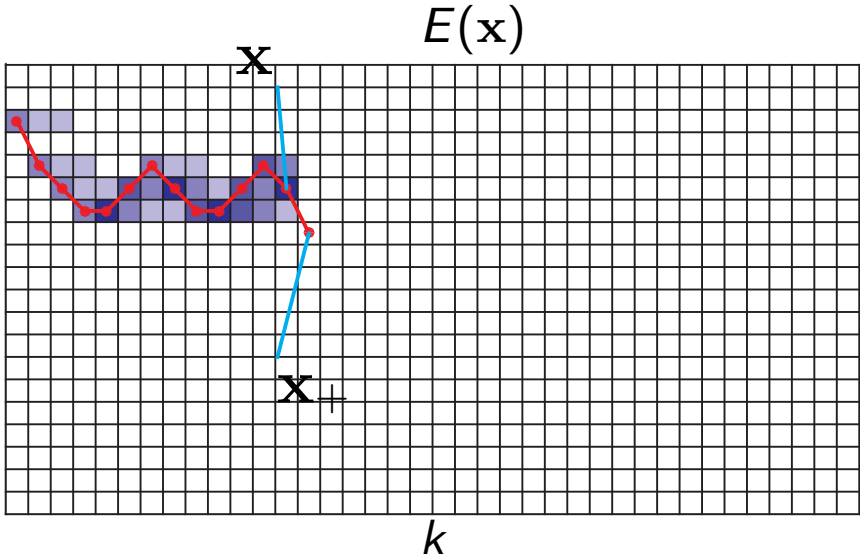
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

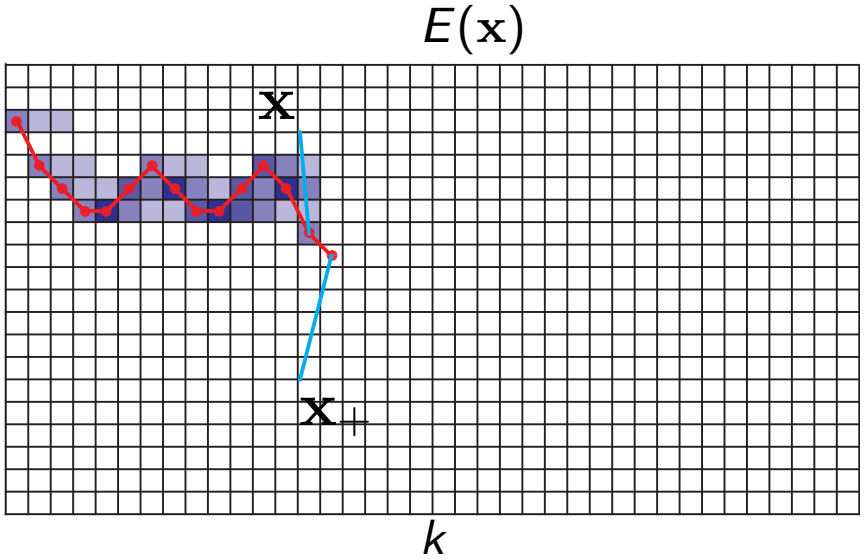
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

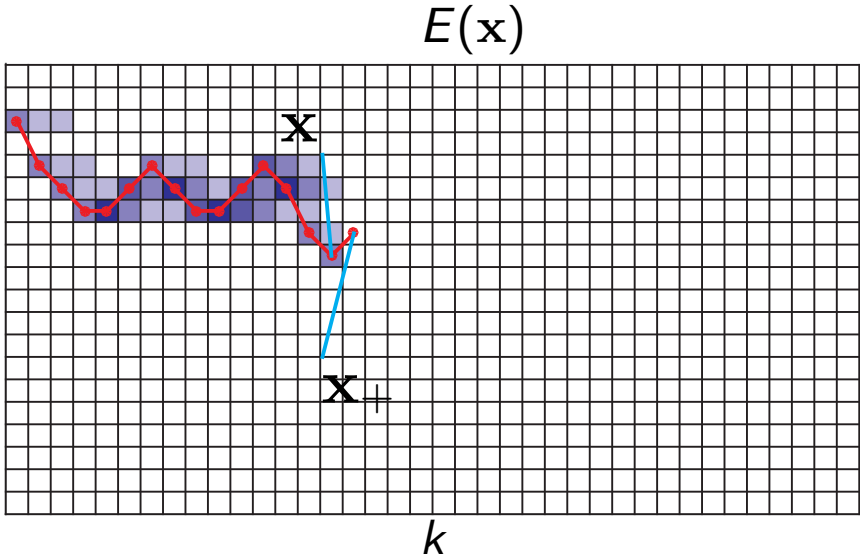
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

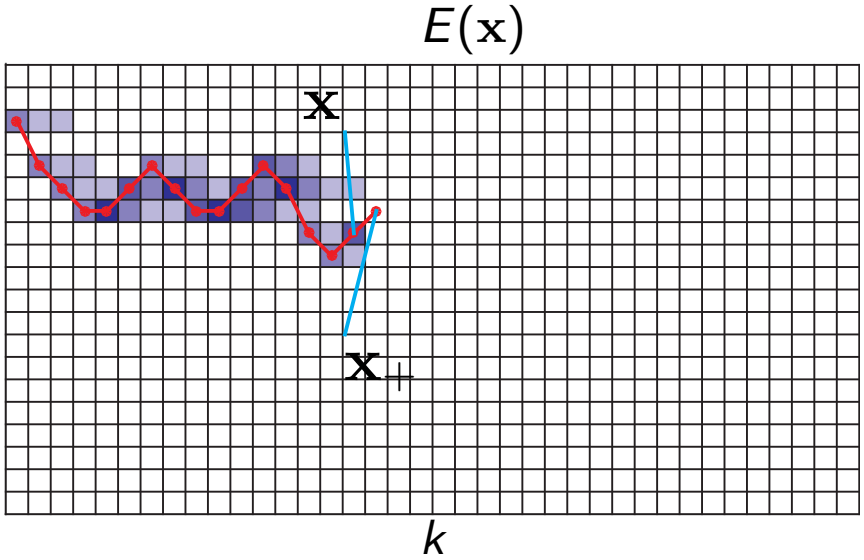
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

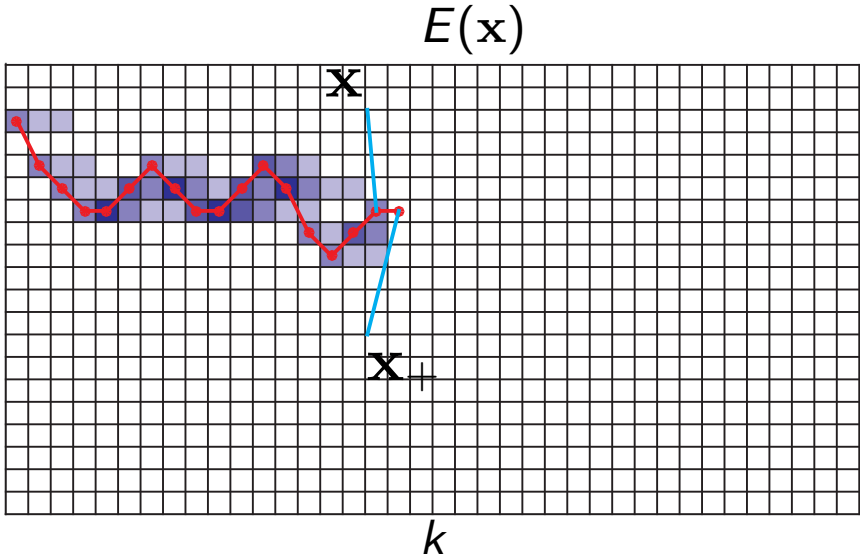
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

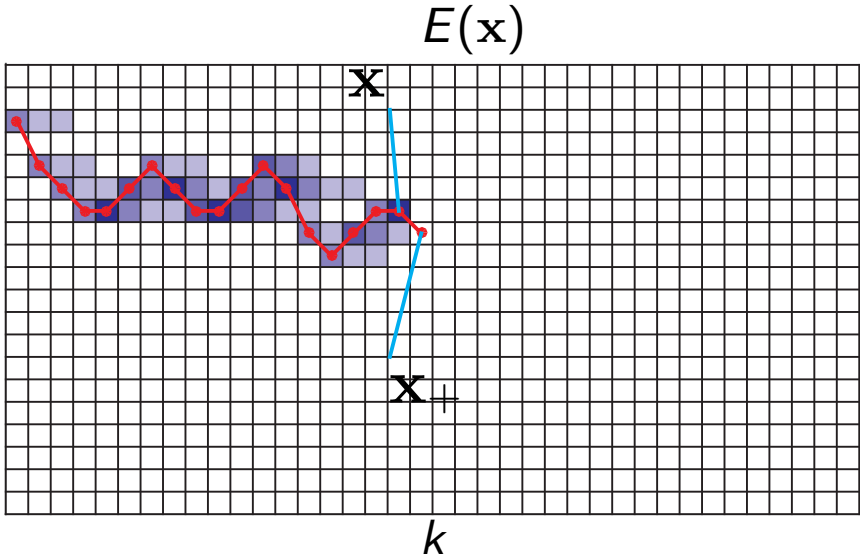
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

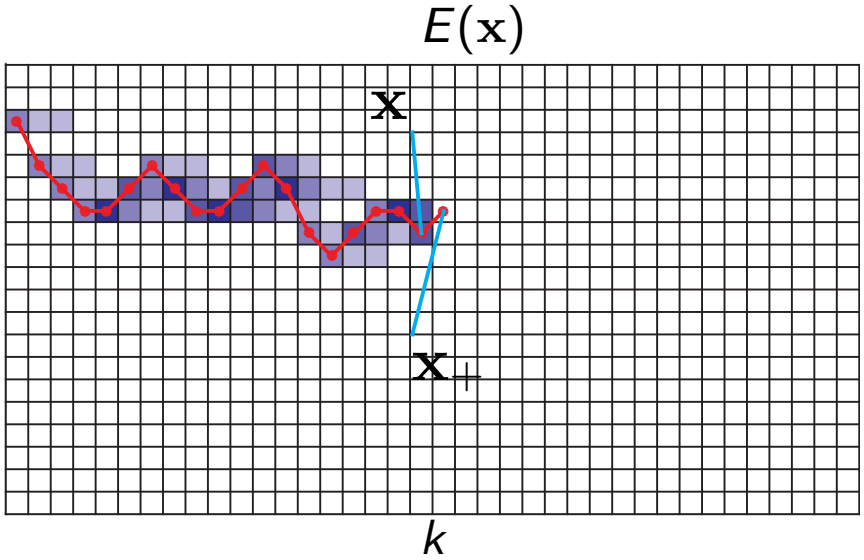
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

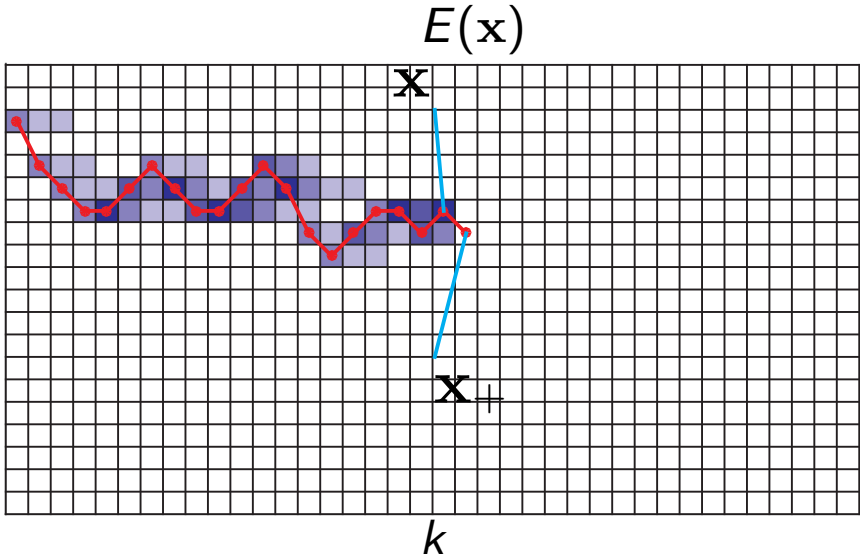
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

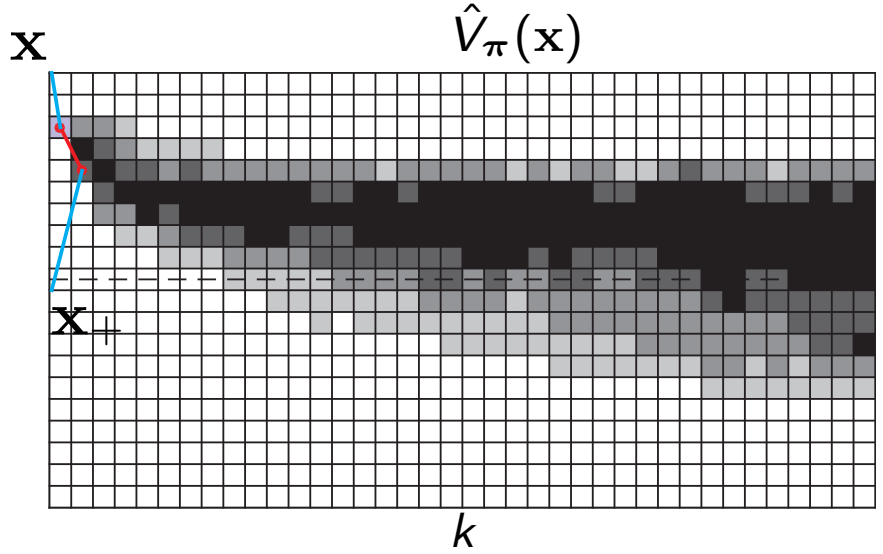
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

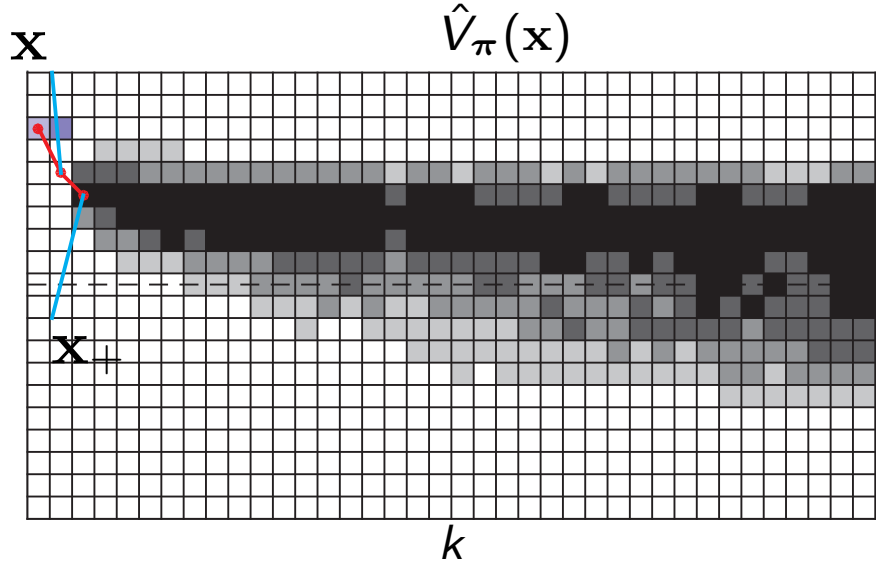
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

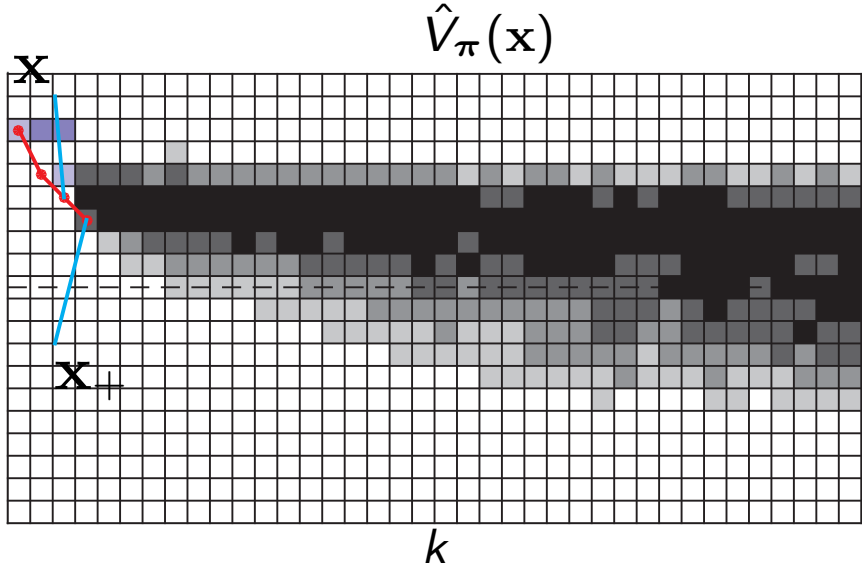
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

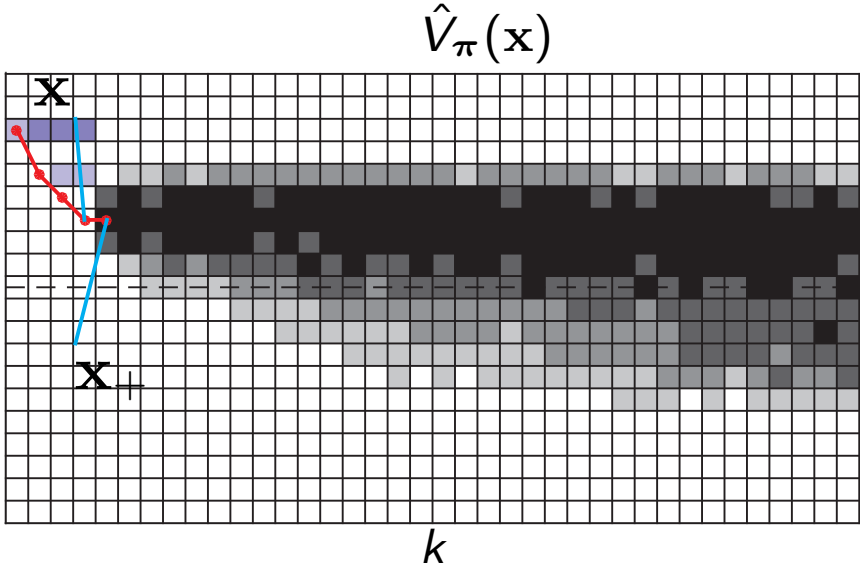
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

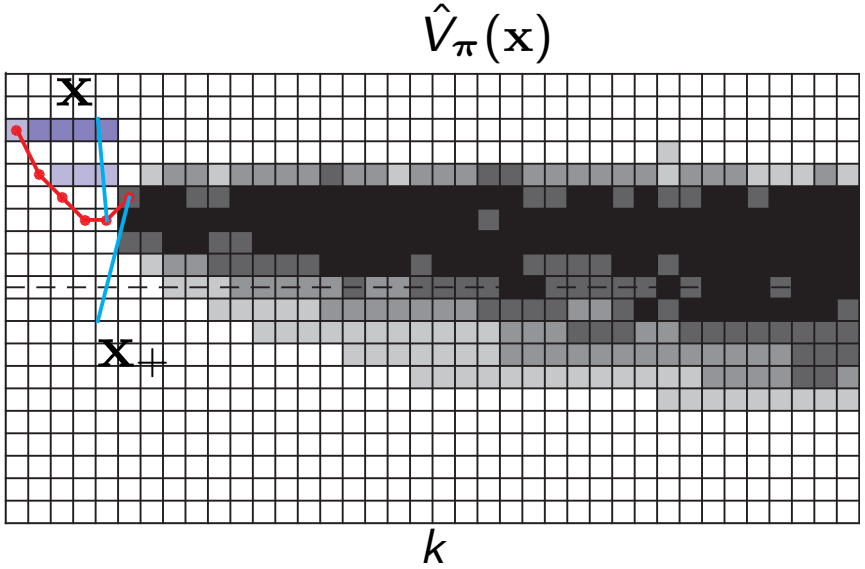
Value function is updated according to

$$\hat{V}_{\pi}(\mathbf{x}) \leftarrow \hat{V}_{\pi}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

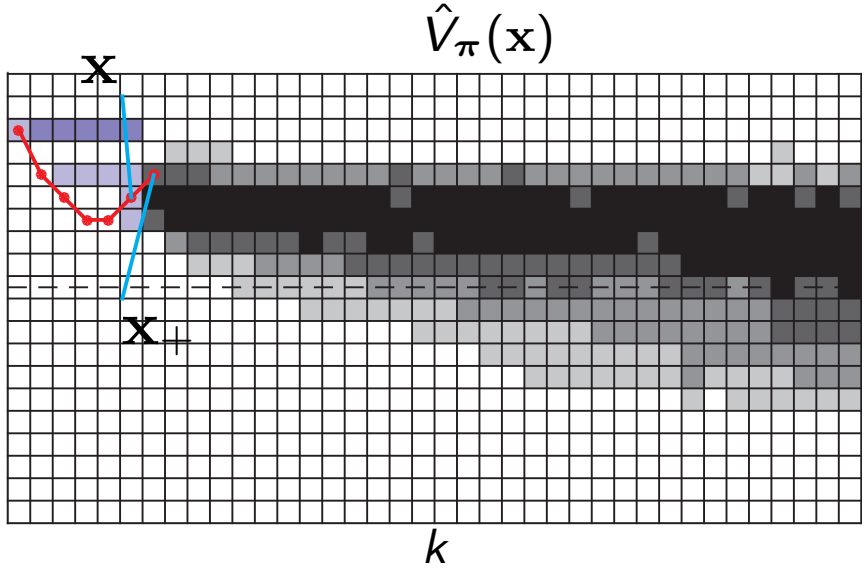
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

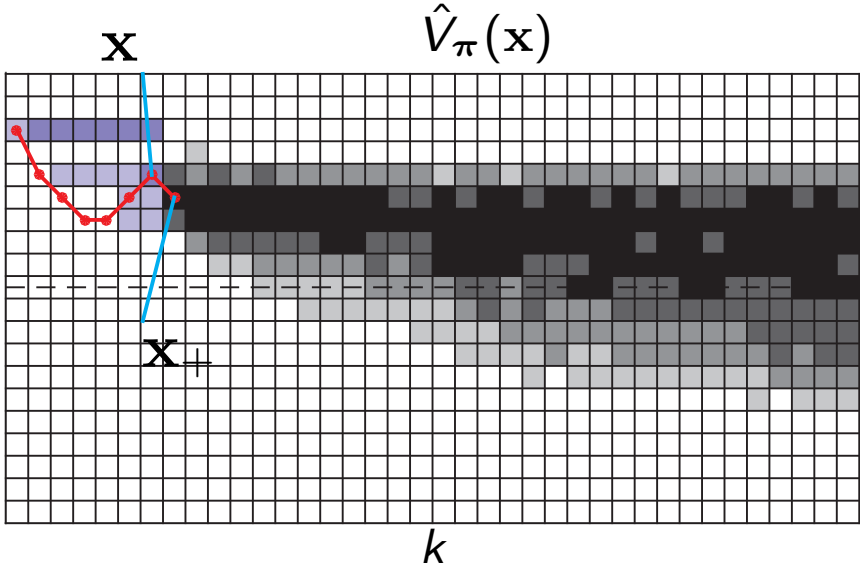
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

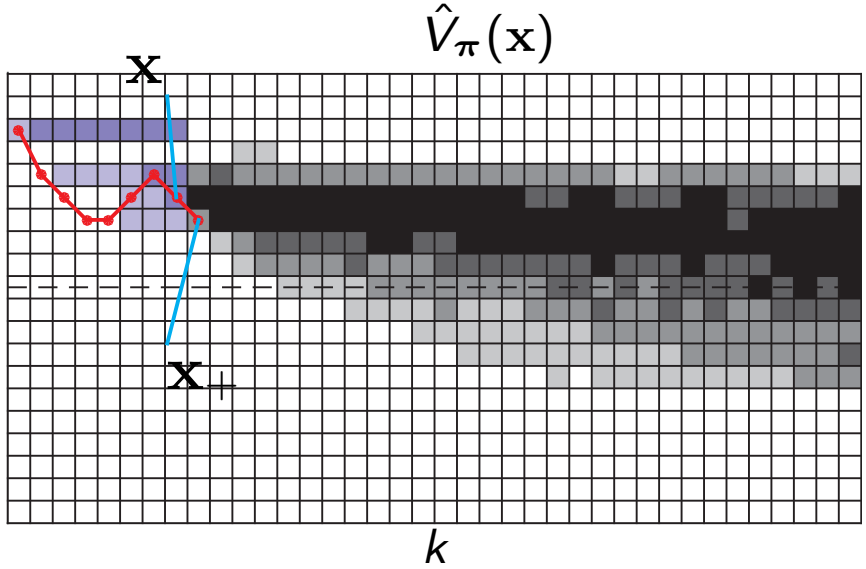
Value function is updated according to

$$\hat{V}_{\pi}(\mathbf{x}) \leftarrow \hat{V}_{\pi}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

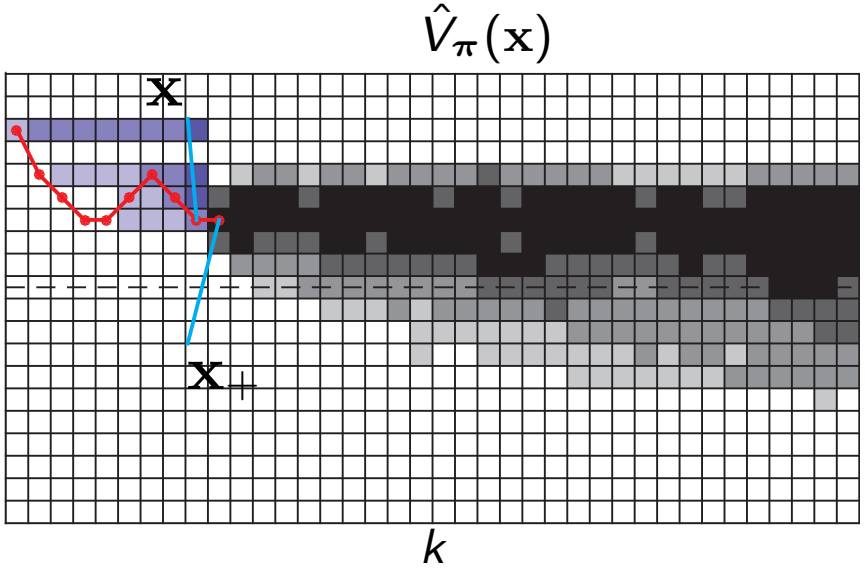
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

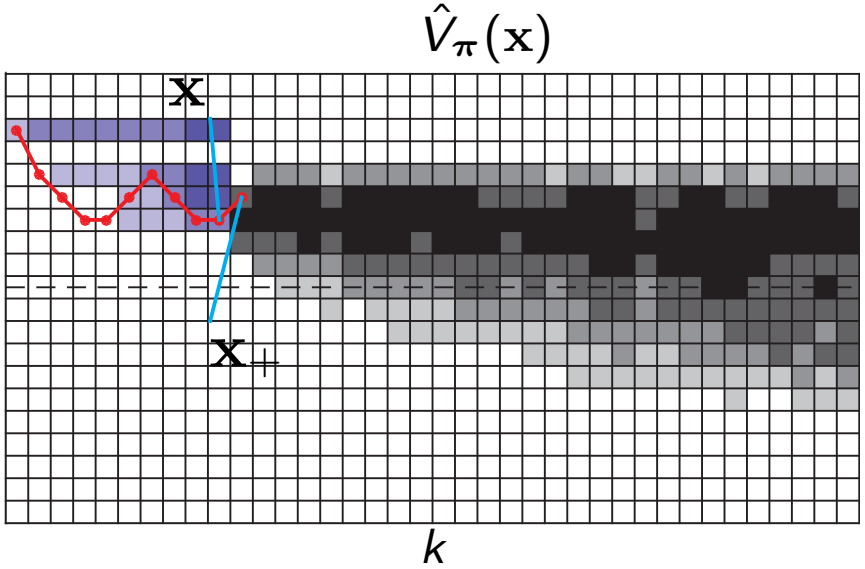
Value function is updated according to

$$\hat{V}_{\pi}(\mathbf{x}) \leftarrow \hat{V}_{\pi}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

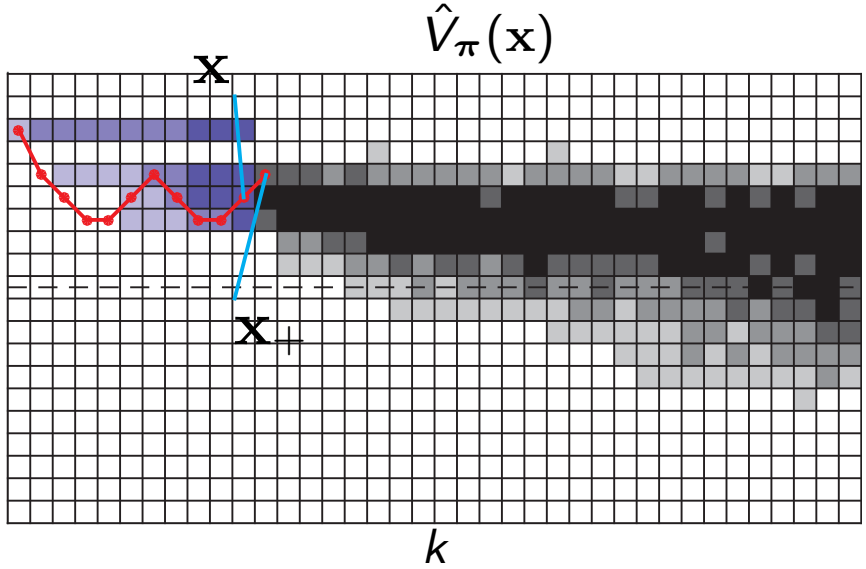
Value function is updated according to

$$\hat{V}_{\pi}(\mathbf{x}) \leftarrow \hat{V}_{\pi}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

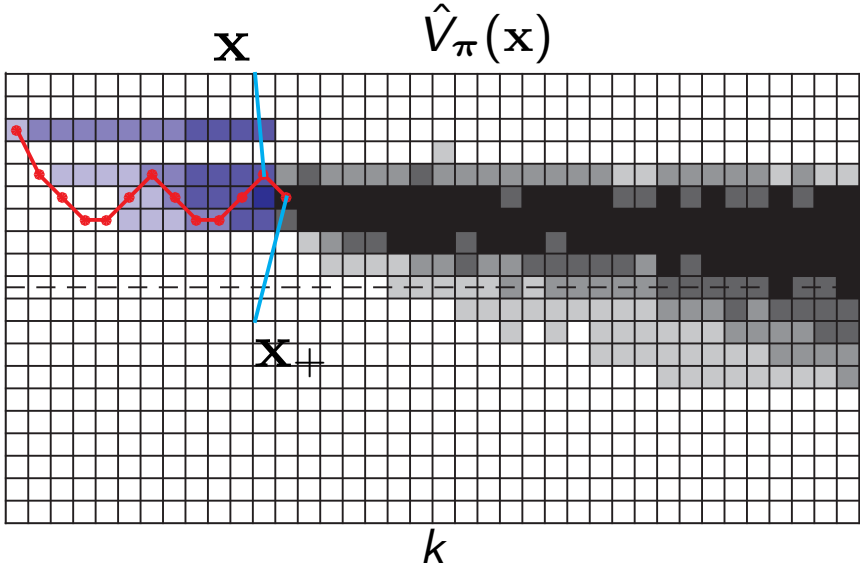
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

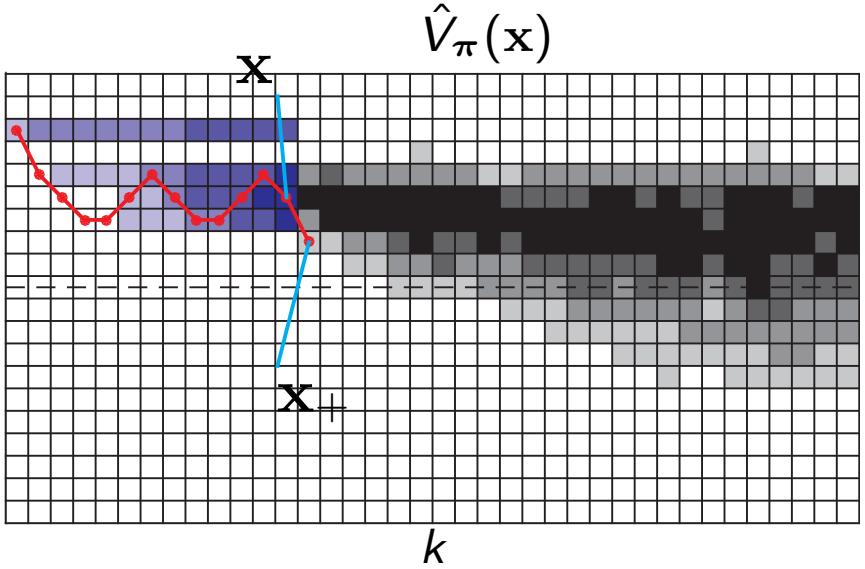
Value function is updated according to

$$\hat{V}_{\pi}(\mathbf{x}) \leftarrow \hat{V}_{\pi}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

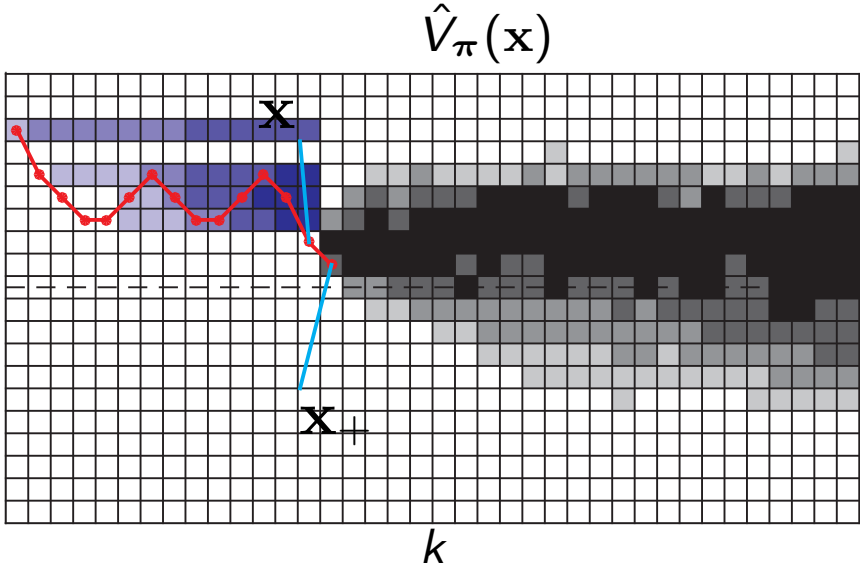
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

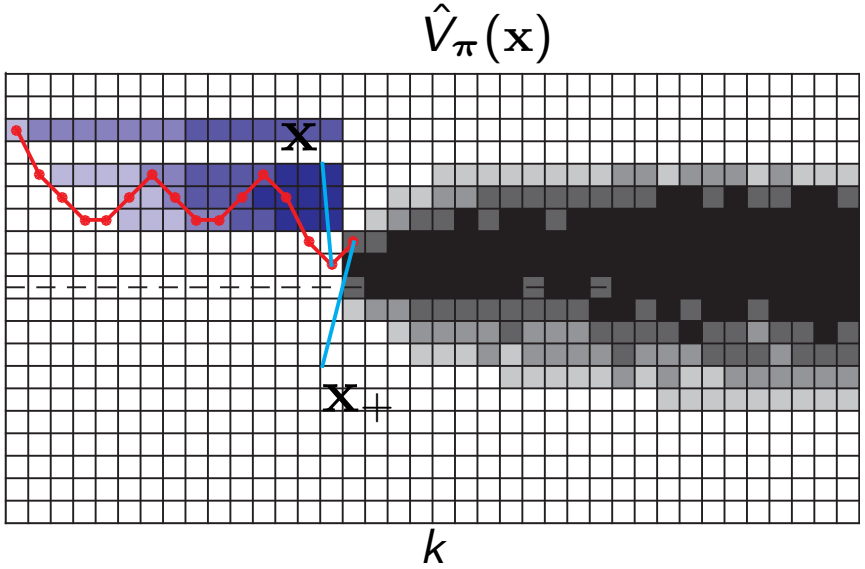
Value function is updated according to

$$\hat{V}_{\pi}(\mathbf{x}) \leftarrow \hat{V}_{\pi}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

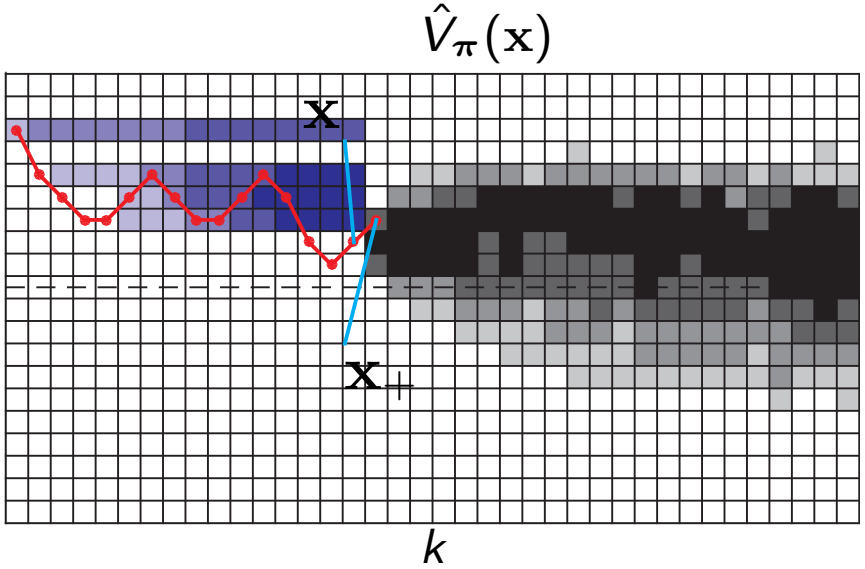
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

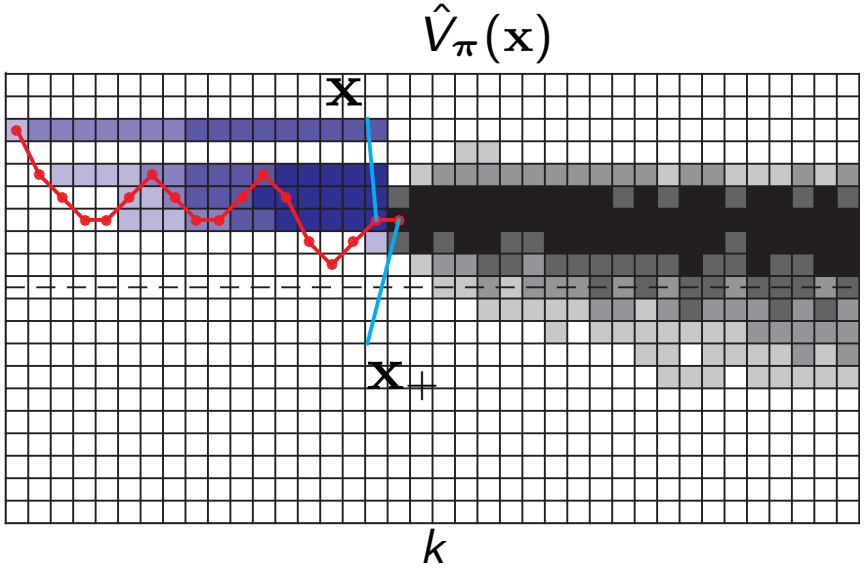
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

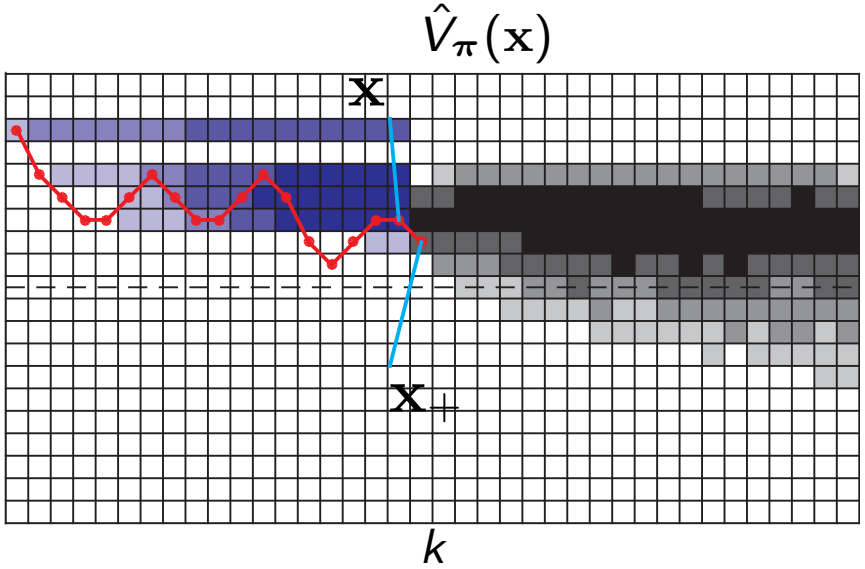
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

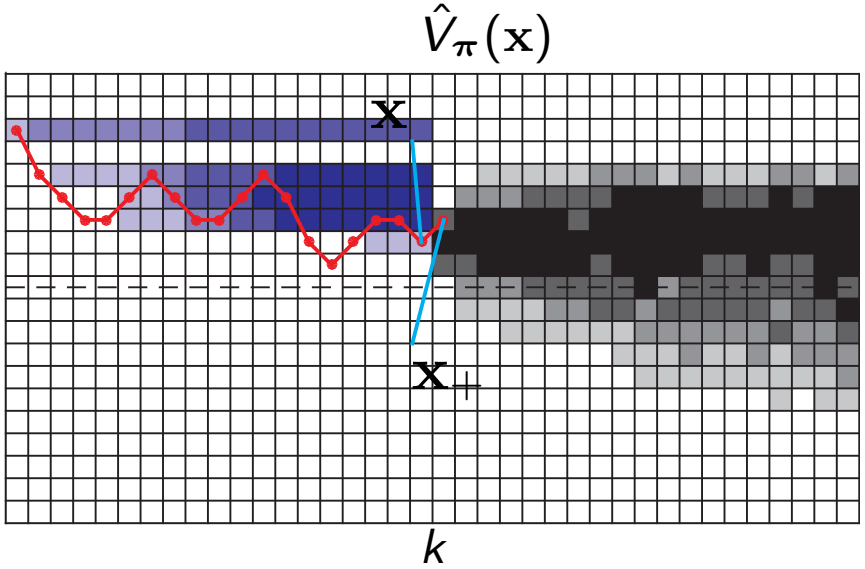
Value function is updated according to

$$\hat{V}_{\pi}(\mathbf{x}) \leftarrow \hat{V}_{\pi}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

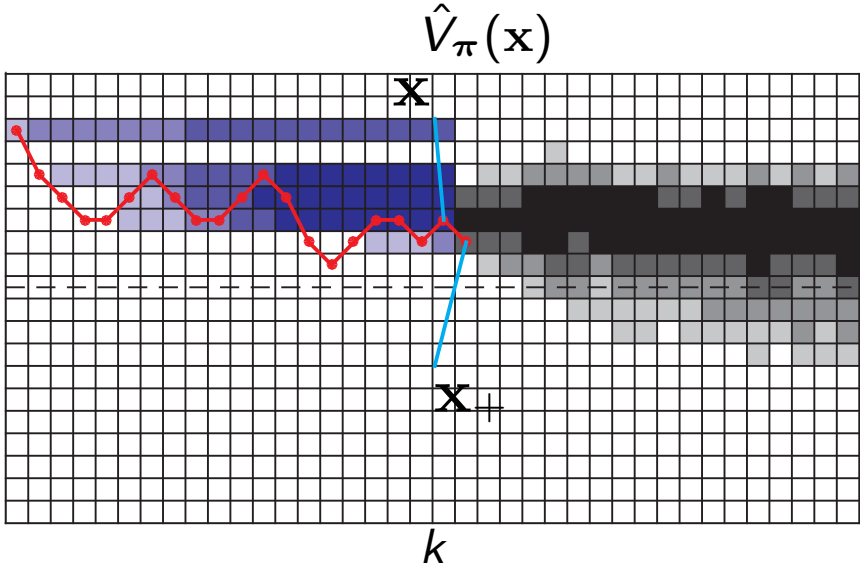
Value function is updated according to

$$\hat{V}_{\pi}(\mathbf{x}) \leftarrow \hat{V}_{\pi}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

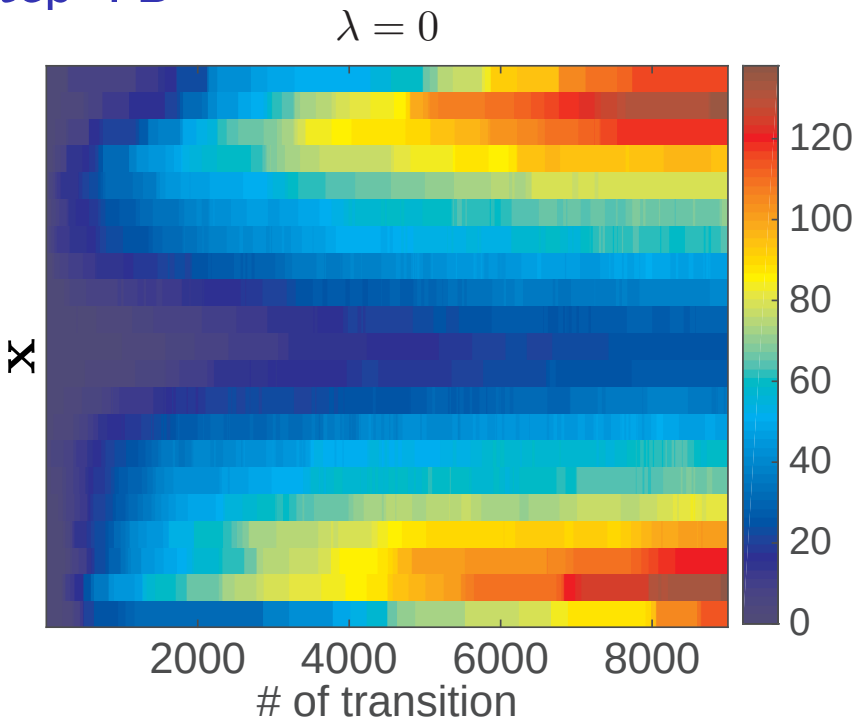
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

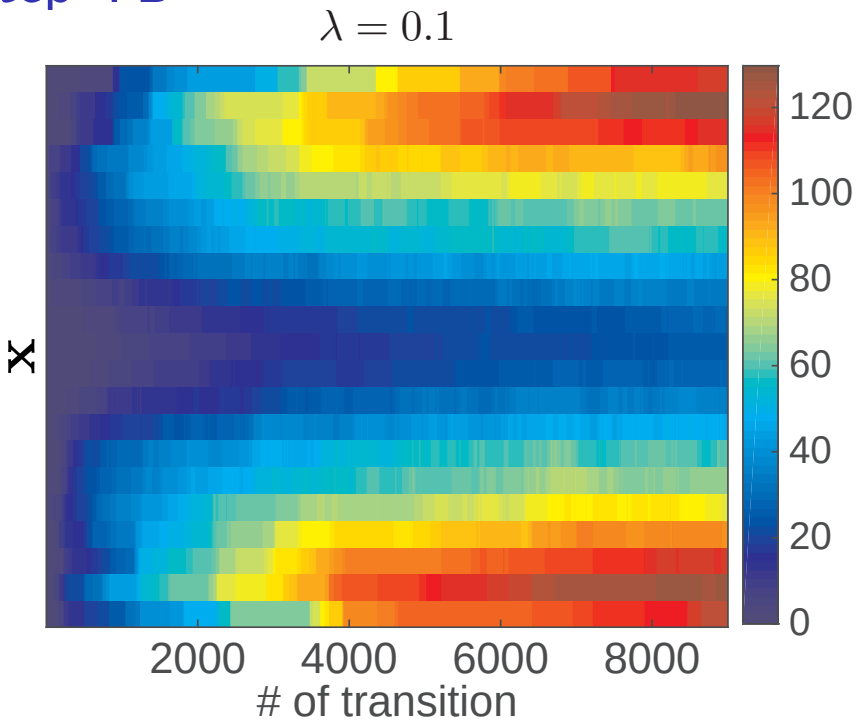
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

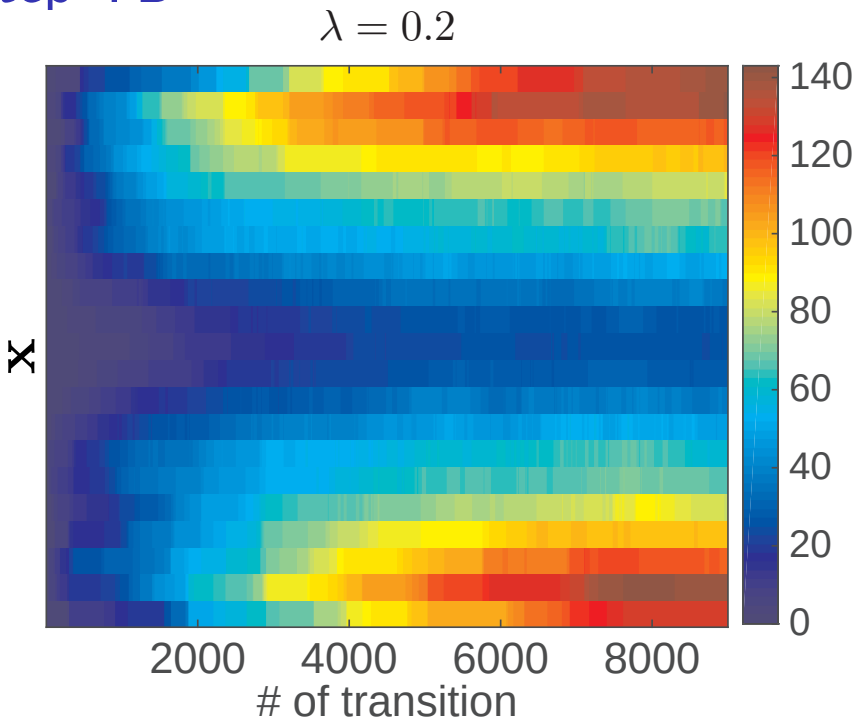
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

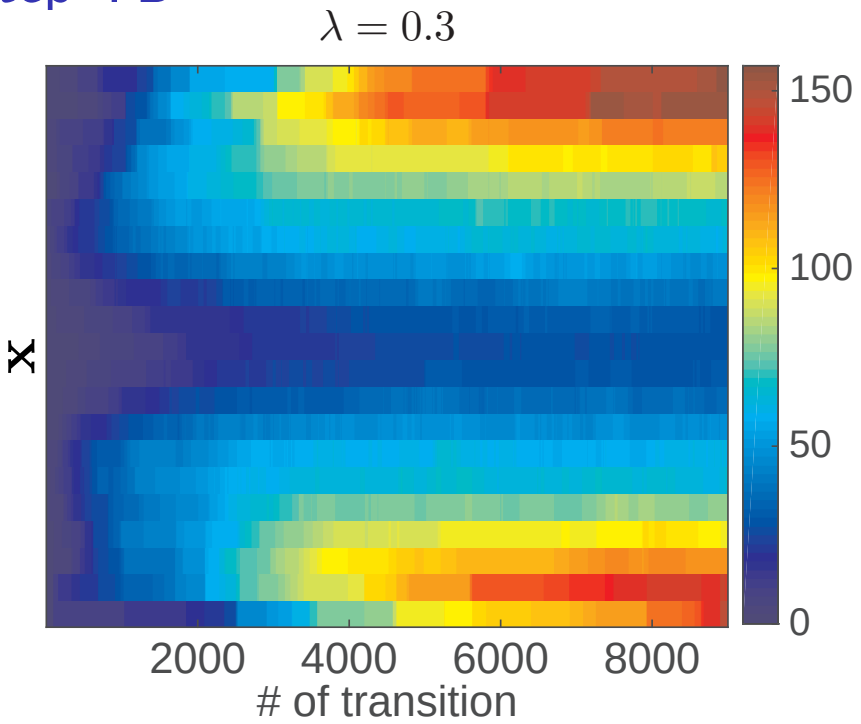
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

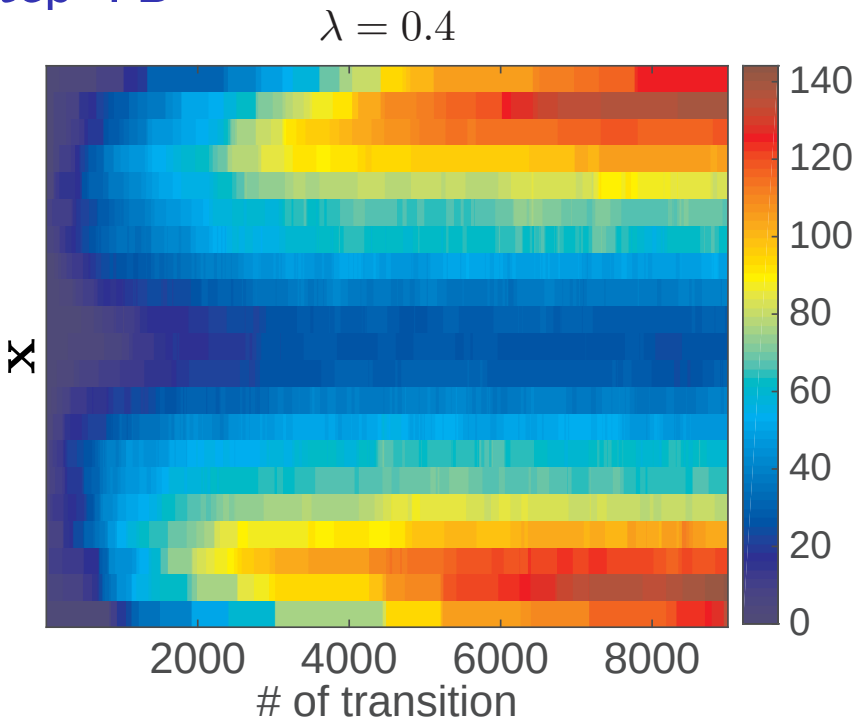
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

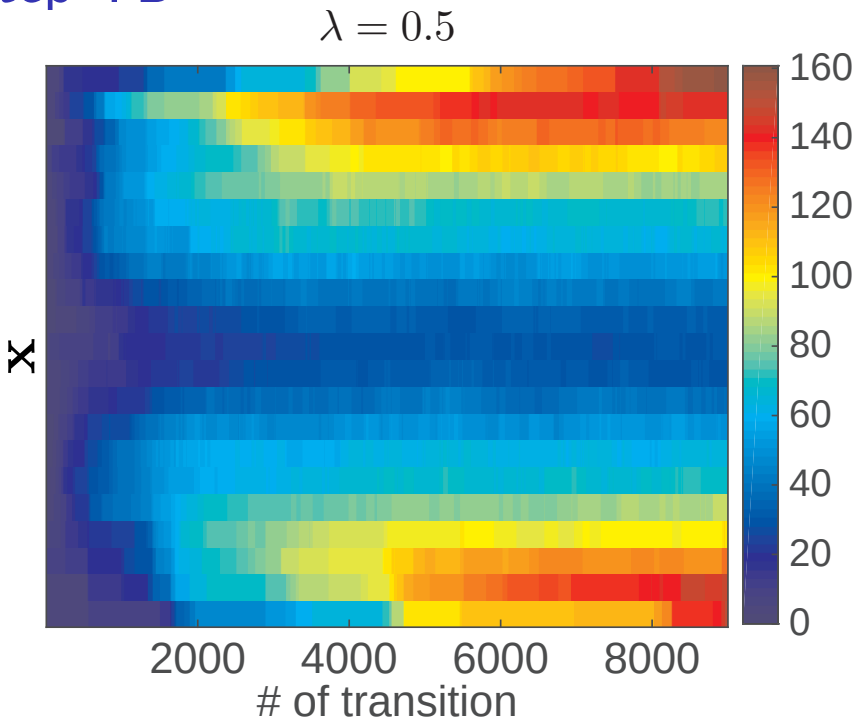
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

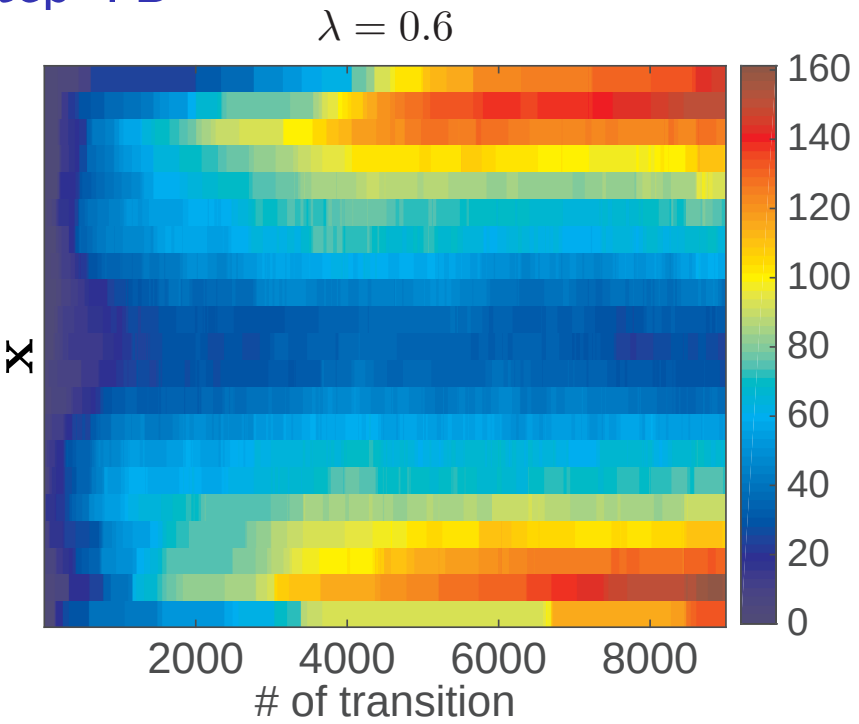
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

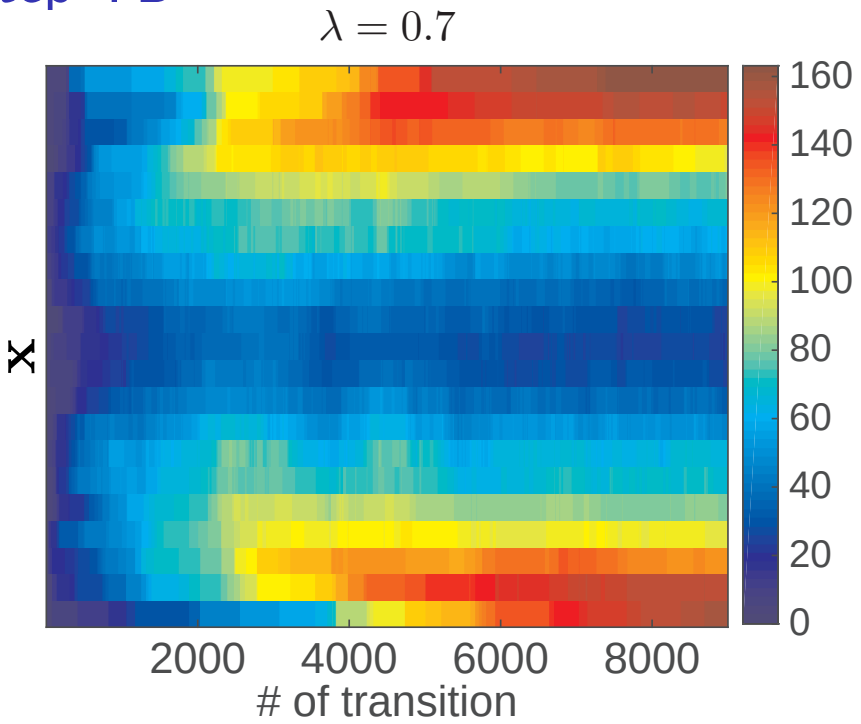
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

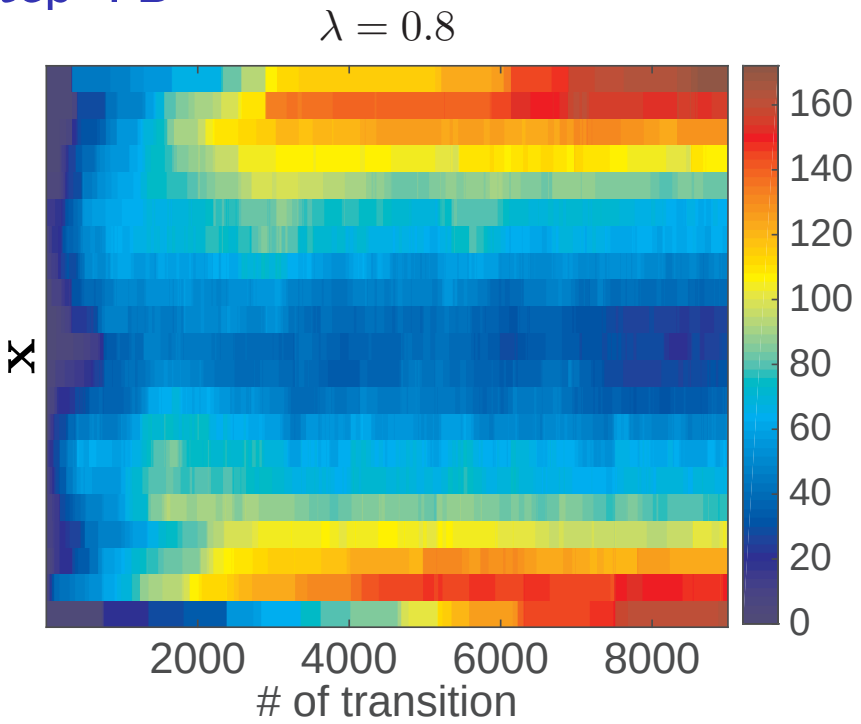
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

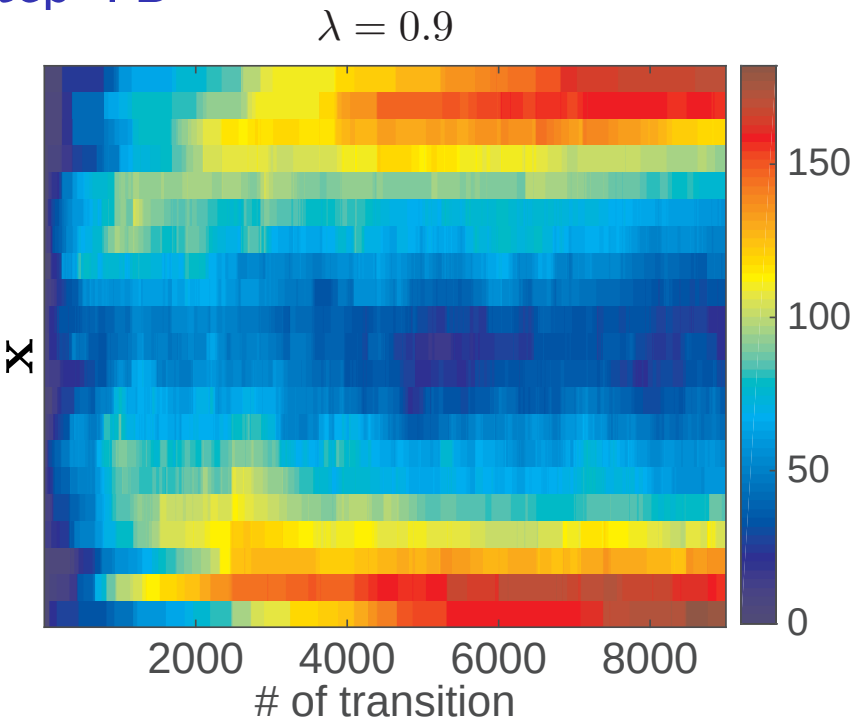
Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

TD(λ) method - Implementing n-step TD

- MC or n-steps TD is hard to implement
- Requires keeping traces of the state trajectories, memory intensive



TD(λ): trace of state visits $\rightarrow$ eligibility trace:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}, \quad \forall \mathbf{x}$$

Value function is updated according to

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

$$\delta_k = L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) + \gamma \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}) - \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k)$$

Outline

Learning the action-value function - Off policy

Action-value function for policy π:

$$Q_{\pi}\left(\mathbf{x},\mathbf{u}\right)=\mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^{k}L(\mathbf{x}_{k},\mathbf{u}_{k})\,\middle|\,\mathbf{u}_{0}=\mathbf{u},\,\mathbf{u}_{k>0}=\boldsymbol{\pi}\left(\mathbf{x}_{k}\right),\,\mathbf{x}_{0}=\mathbf{x}\right]$$

Learning the action-value function - Off policy

Action-value function for policy π:

$$Q_{\pi}\left(\mathbf{x},\mathbf{u}\right)=\mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^{k}L(\mathbf{x}_{k},\mathbf{u}_{k})\,\middle|\,\mathbf{u}_{0}=\mathbf{u},\,\mathbf{u}_{k>0}=\boldsymbol{\pi}\left(\mathbf{x}_{k}\right),\,\mathbf{x}_{0}=\mathbf{x}\right]$$

Policy Evalution does

$$\hat{Q}_{\boldsymbol{\pi}}\left(\mathbf{x},\mathbf{u}\right)\leftarrow L(\mathbf{x},\mathbf{u})+\gamma\mathbb{E}\left[\hat{Q}_{\boldsymbol{\pi}}\left(\mathbf{x}_{+},\boldsymbol{\pi}(\mathbf{x}_{+})\right)\,\middle|\,\mathbf{x},\mathbf{u}\right]$$

Learning the action-value function - Off policy

Action-value function for policy π:

$$Q_\pi(\mathbf{x},\mathbf{u}) = \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{x}_k,\mathbf{u}_k)\,\middle|\,\mathbf{u}_0=\mathbf{u},\ \mathbf{u}_{k>0}=\boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0=\mathbf{x}\right]$$

Policy Evalution does

$$\hat{Q}_\pi(\mathbf{x},\mathbf{u}) \leftarrow L(\mathbf{x},\mathbf{u}) + \gamma\mathbb{E}\left[\hat{Q}_\pi(\mathbf{x}_+,\boldsymbol{\pi}(\mathbf{x}_+))\,\middle|\,\mathbf{x},\mathbf{u}\right]$$

Temporal Difference (TD)

$$\hat{Q}_\pi(\mathbf{x},\mathbf{u}) \leftarrow (1-\alpha)\,\hat{Q}_\pi(\mathbf{x},\mathbf{u}) + \alpha\left(L(\mathbf{x},\mathbf{u}) + \gamma\hat{Q}_\pi(\mathbf{x}_+,\boldsymbol{\pi}(\mathbf{x}_+))\right)$$

where $\mathbf{u}$ can be arbitrary.

Learning the action-value function - Off policy

Action-value function for policy π:

$$Q_\pi(\mathbf{x},\mathbf{u}) = \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{x}_k,\mathbf{u}_k)\,\middle|\,\mathbf{u}_0=\mathbf{u},\ \mathbf{u}_{k>0}=\boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0=\mathbf{x}\right]$$

Policy Evalution does

$$\hat{Q}_\pi(\mathbf{x},\mathbf{u}) \leftarrow L(\mathbf{x},\mathbf{u}) + \gamma\mathbb{E}\left[\hat{Q}_\pi(\mathbf{x}_+,\boldsymbol{\pi}(\mathbf{x}_+))\,\middle|\,\mathbf{x},\mathbf{u}\right]$$

Temporal Difference (TD)

$$\hat{Q}_\pi(\mathbf{x},\mathbf{u}) \leftarrow (1-\alpha)\,\hat{Q}_\pi(\mathbf{x},\mathbf{u}) + \alpha\left(L(\mathbf{x},\mathbf{u}) + \gamma\hat{Q}_\pi(\mathbf{x}_+,\boldsymbol{\pi}(\mathbf{x}_+))\right)$$

where $\mathbf{u}$ can be arbitrary. Hence **TD for Q-learning** does:

$$\begin{aligned}\hat{Q}_\pi(\mathbf{x}_k,\mathbf{u}_k) &\leftarrow \hat{Q}_\pi(\mathbf{x}_k,\mathbf{u}_k) + \alpha\delta_k^Q\\ \delta_k^Q &= L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\underbrace{\hat{Q}_\pi(\mathbf{x}_{k+1},\boldsymbol{\pi}(\mathbf{x}_{k+1}))}_{:=\hat{V}_\pi(\mathbf{x}_{k+1})} - \hat{Q}_\pi(\mathbf{x}_k,\mathbf{u}_k)\end{aligned}$$

This is labelled an off-policy TD evaluation

Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

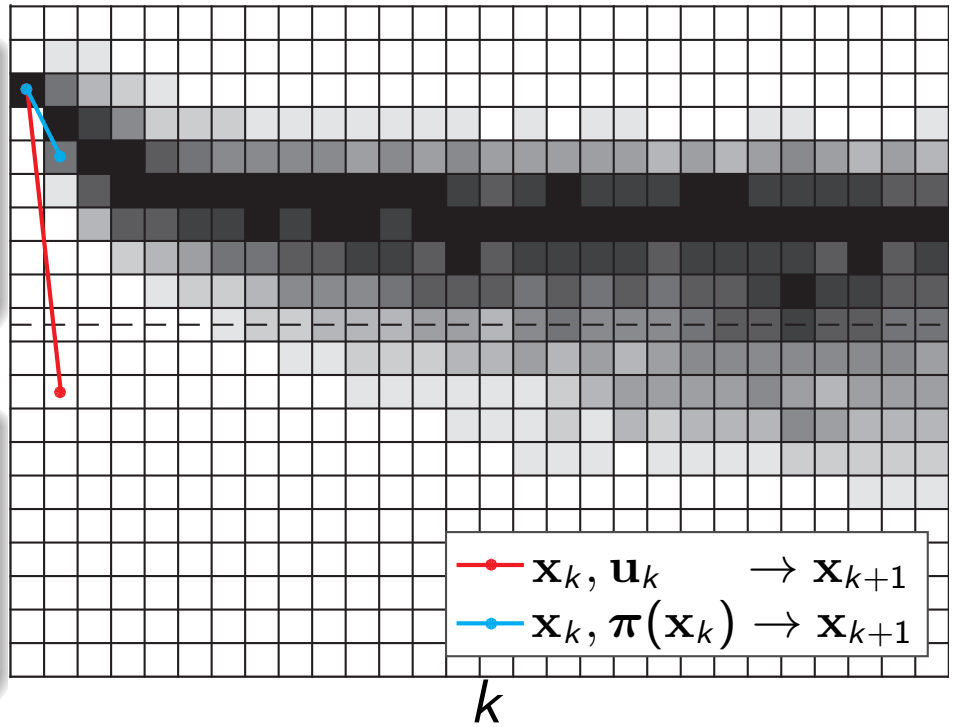
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

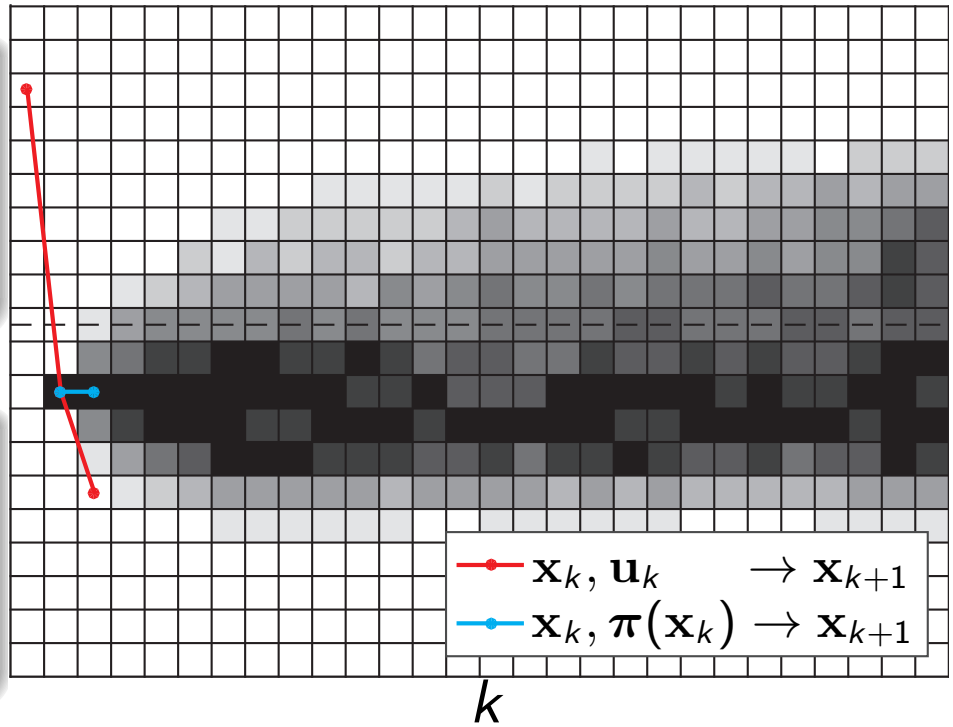
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

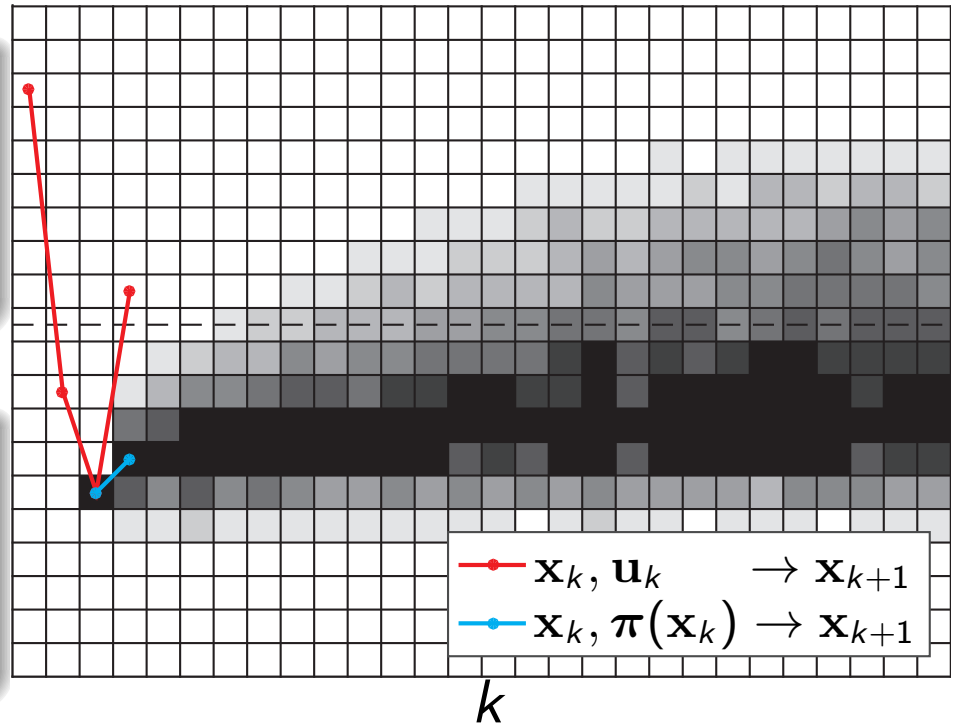
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

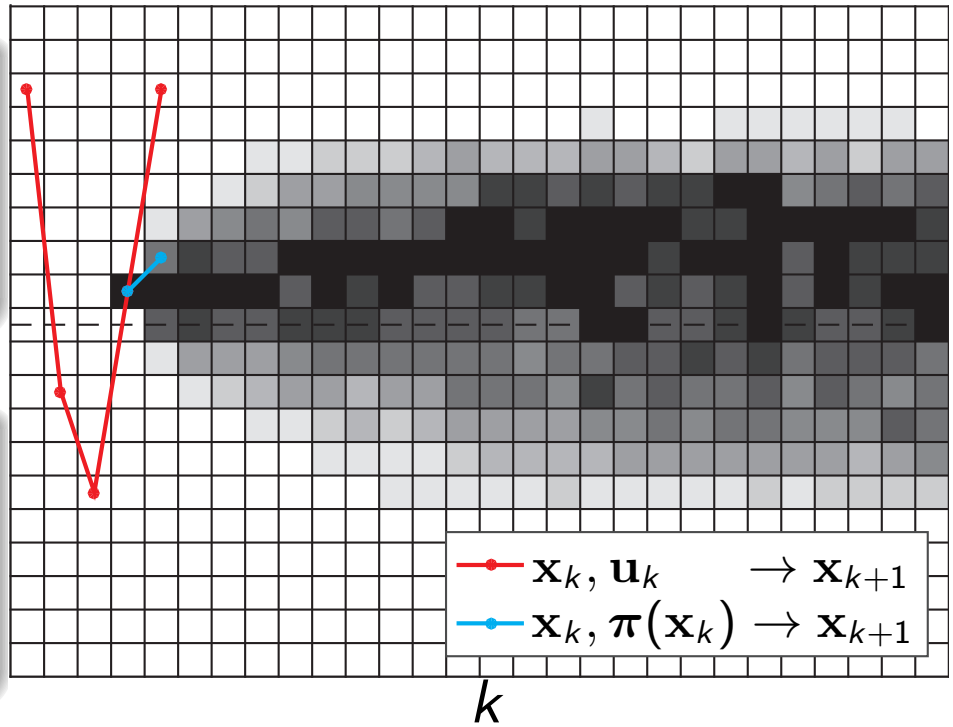
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

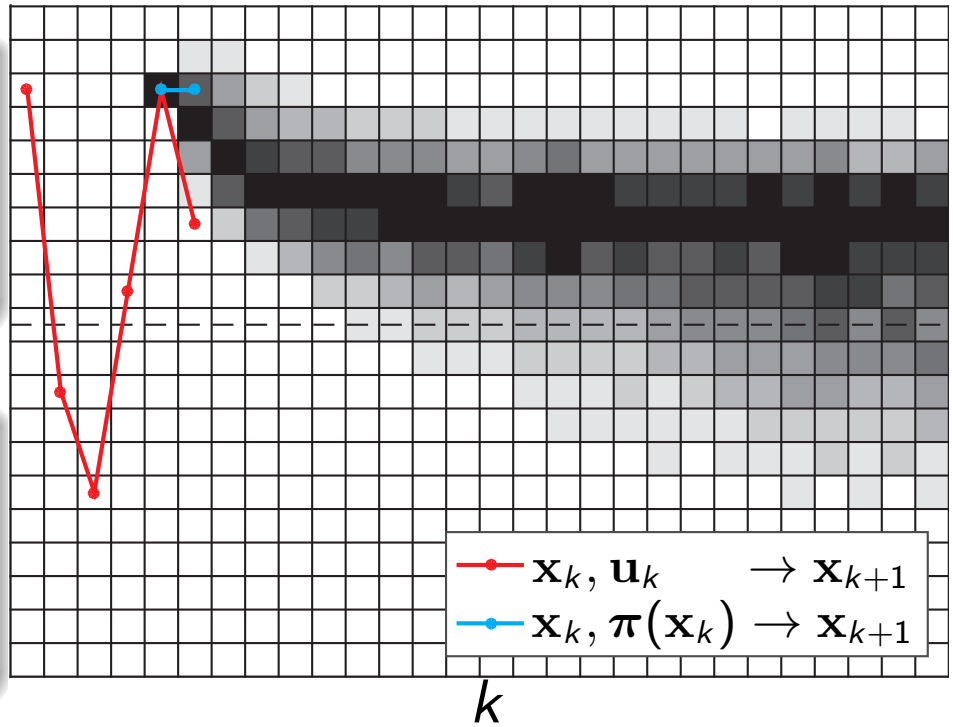
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

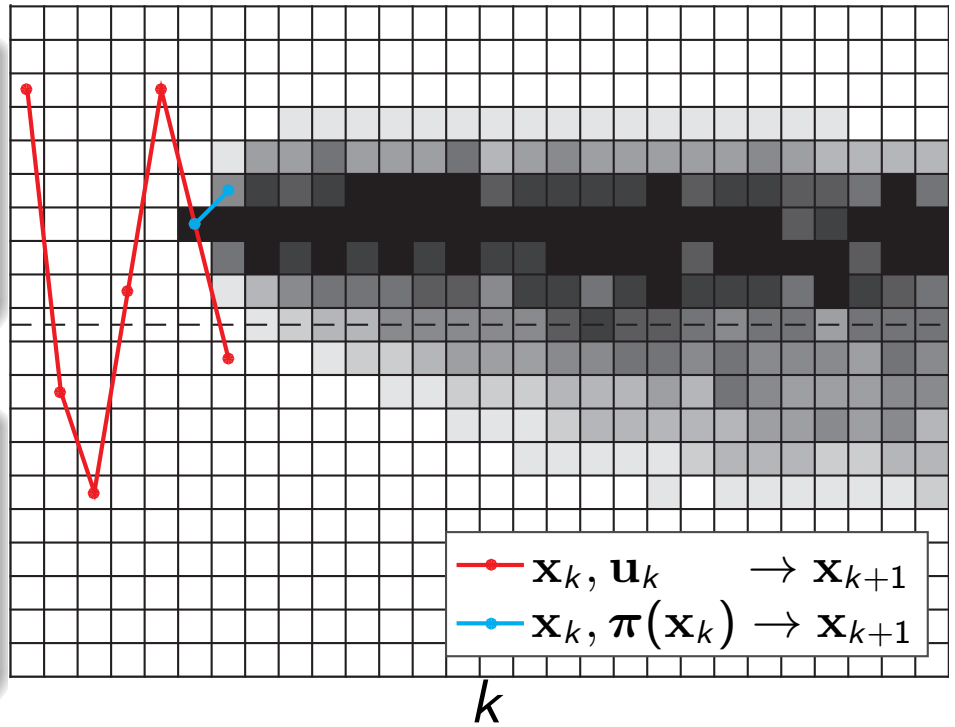
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

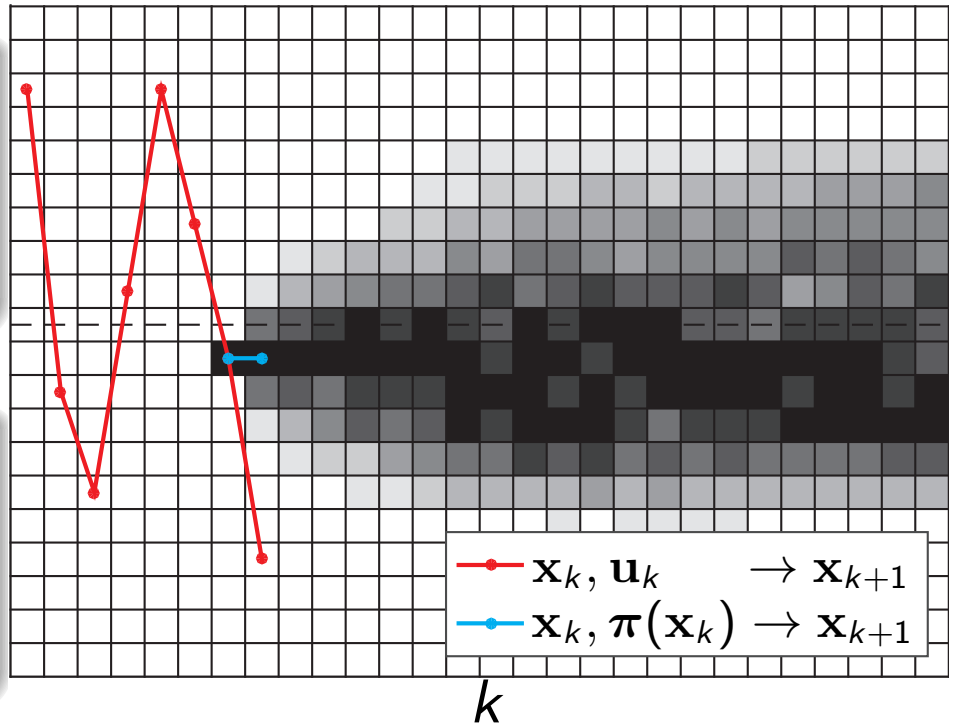
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

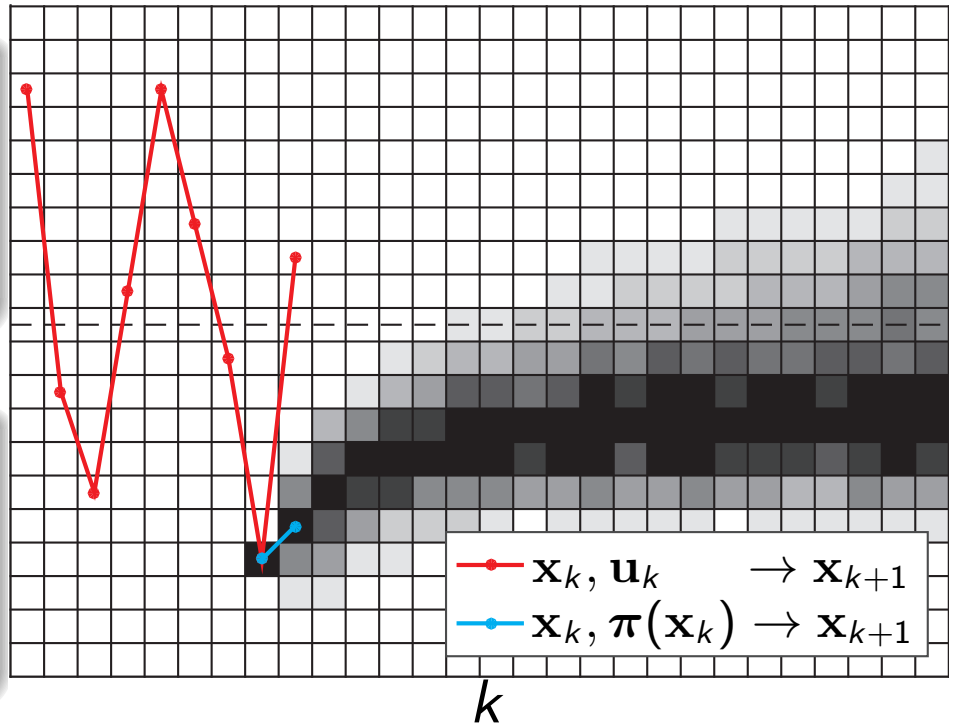
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

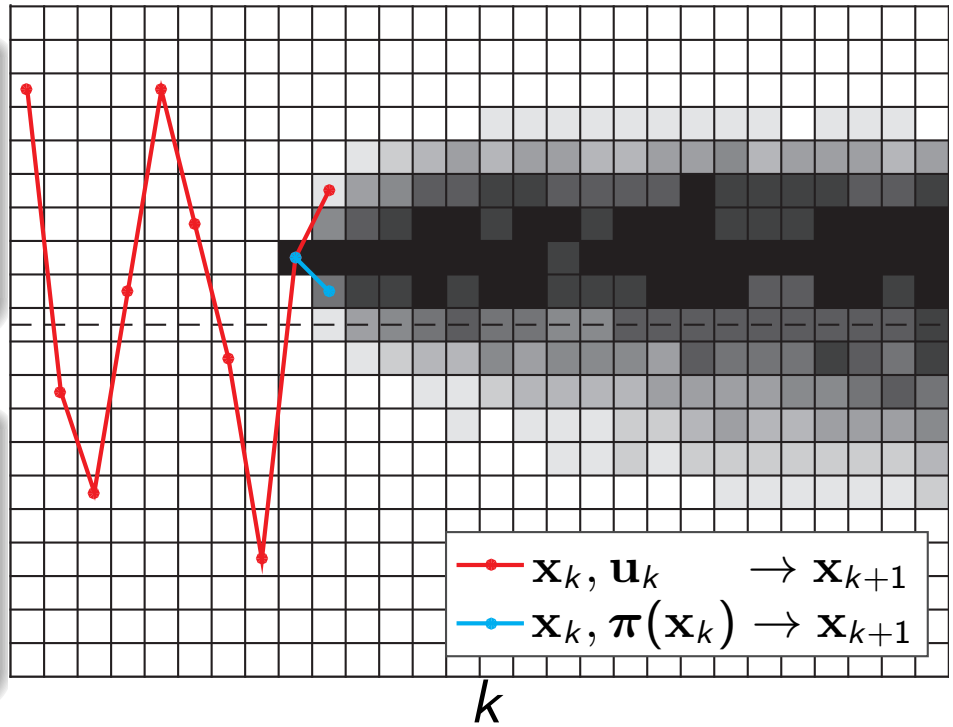
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

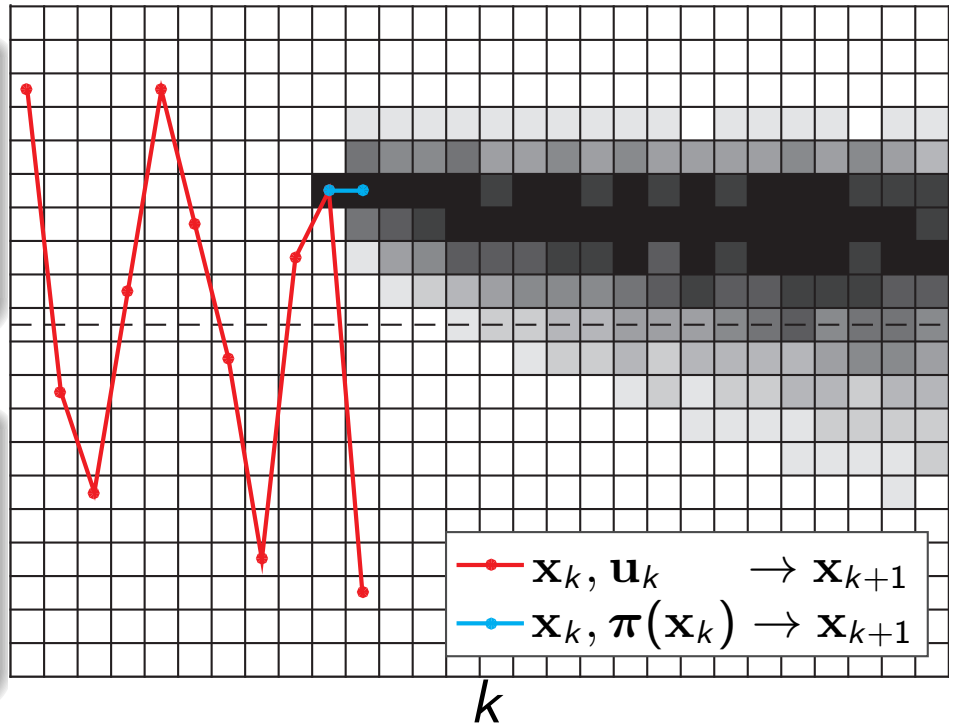
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

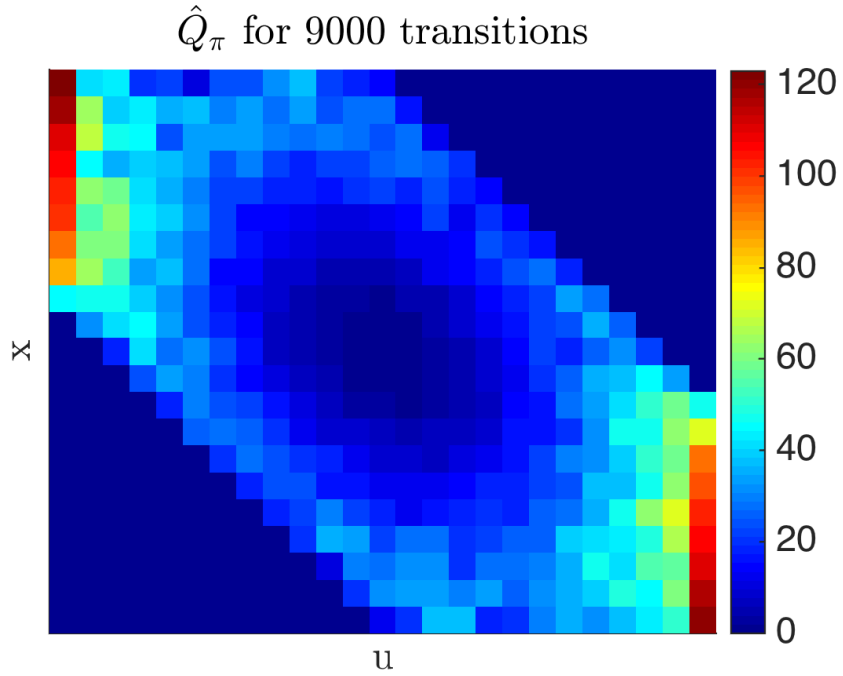
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

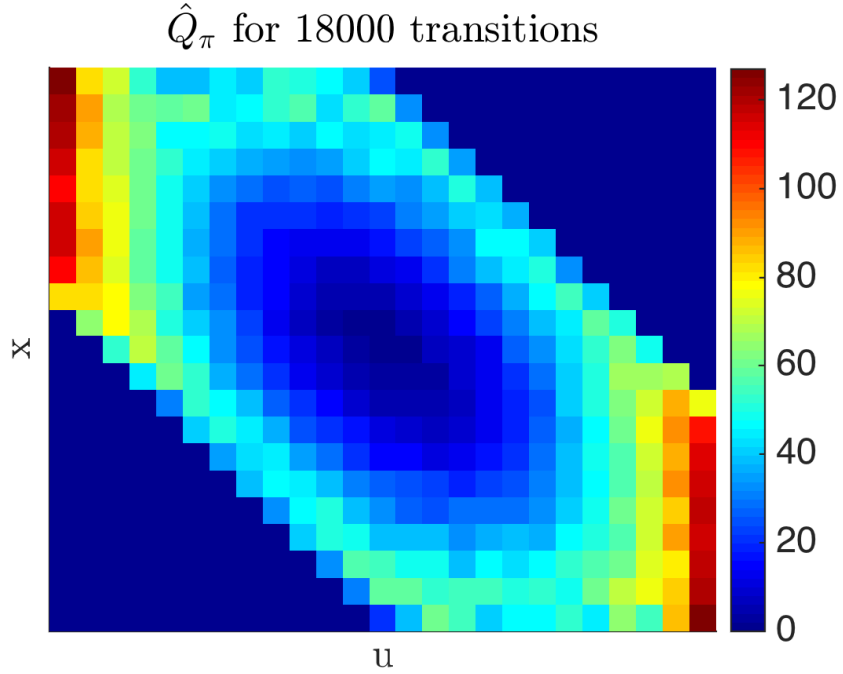
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

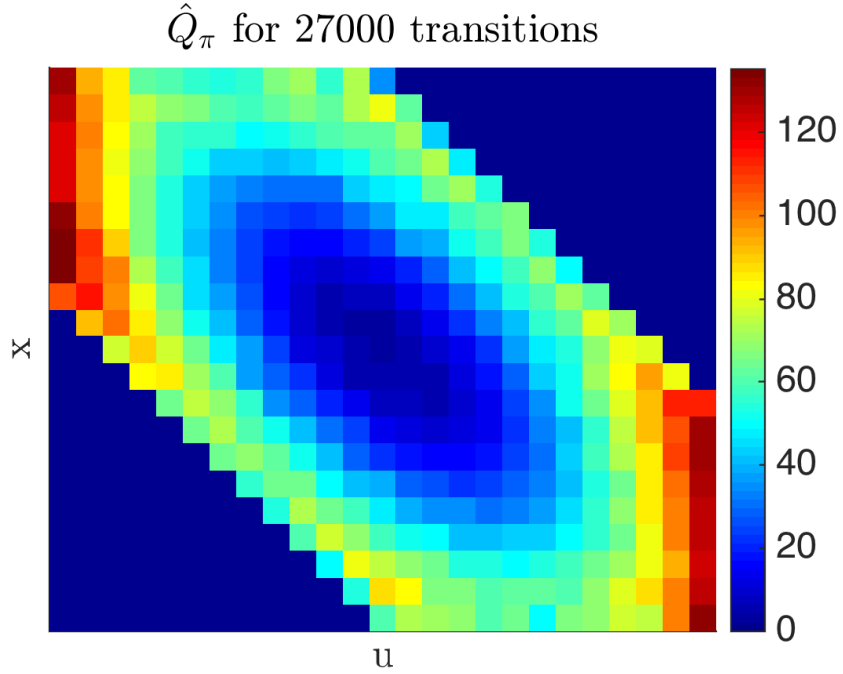
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

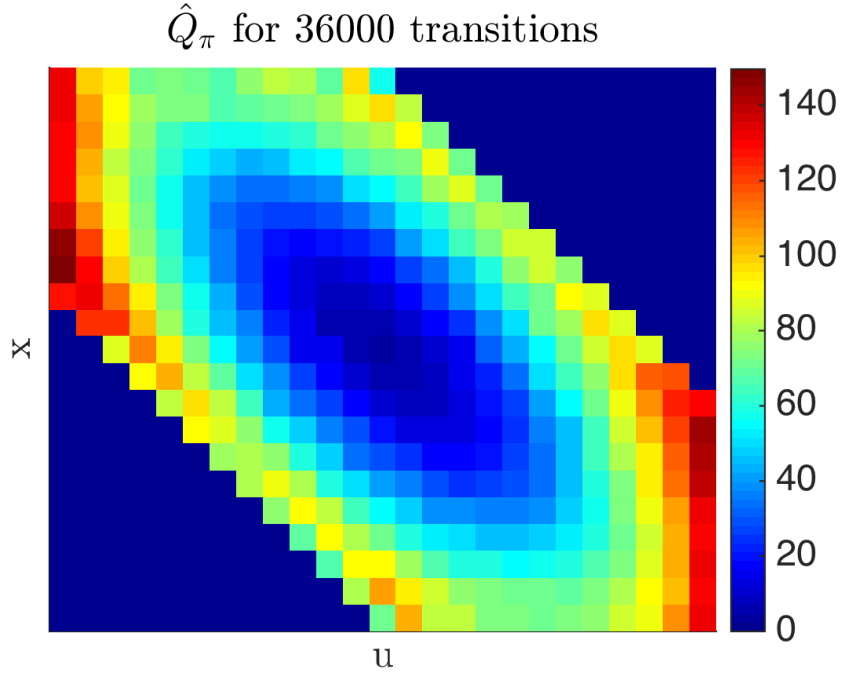
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

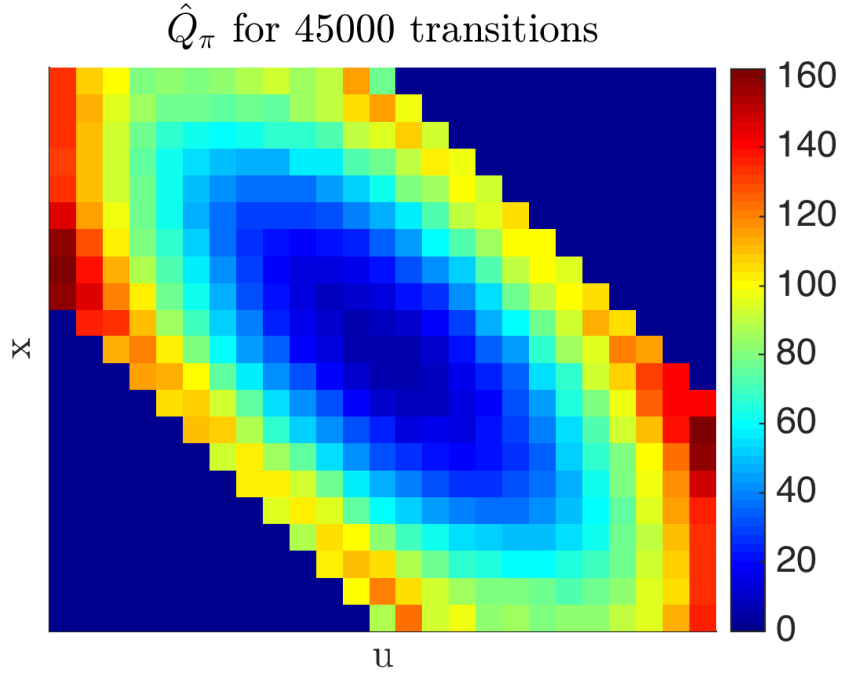
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

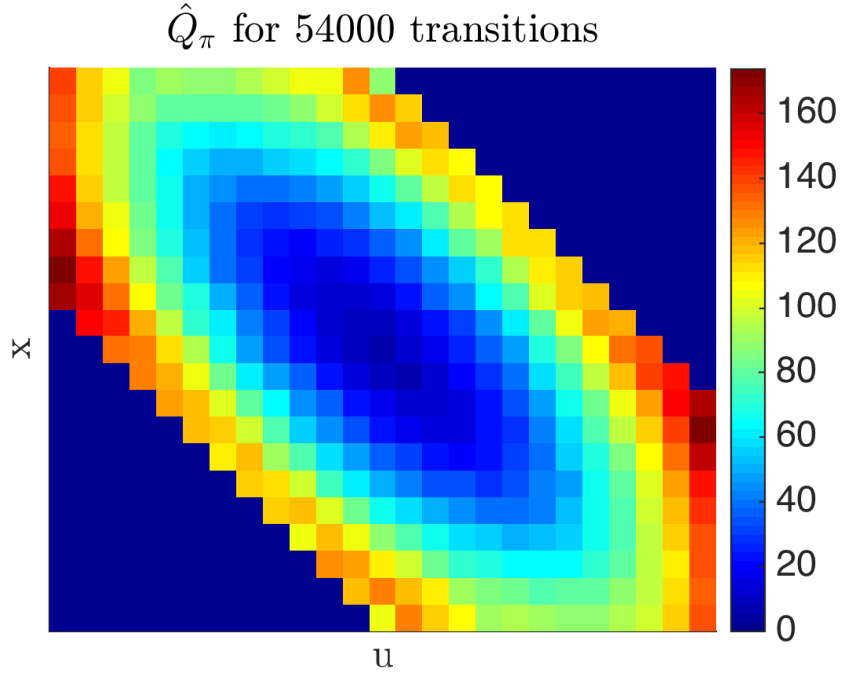
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

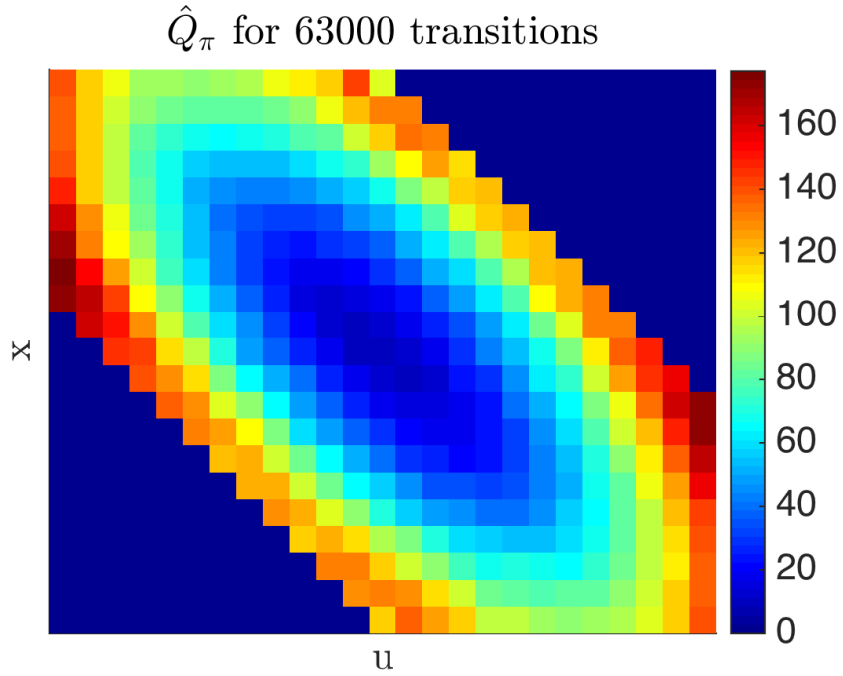
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

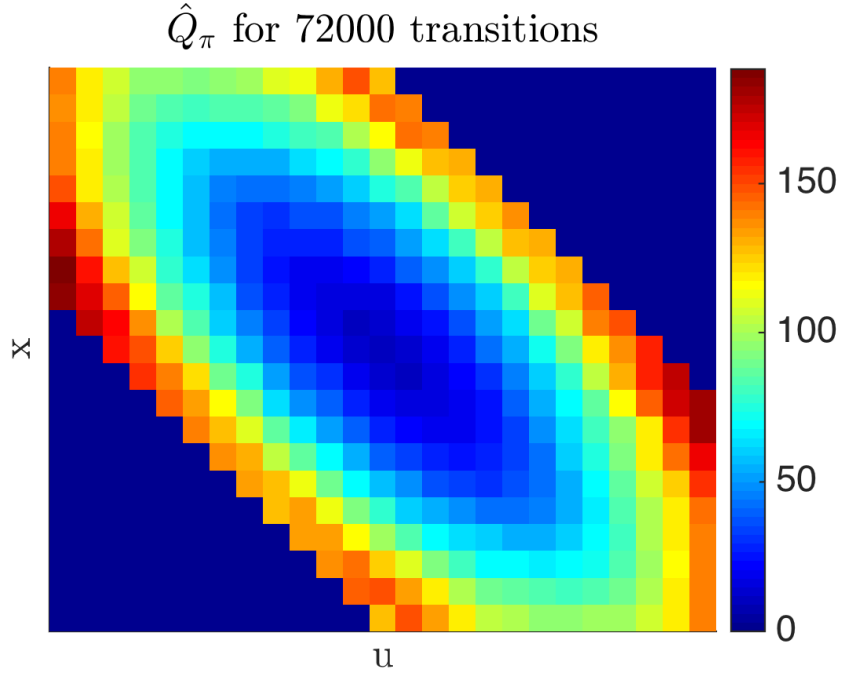
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

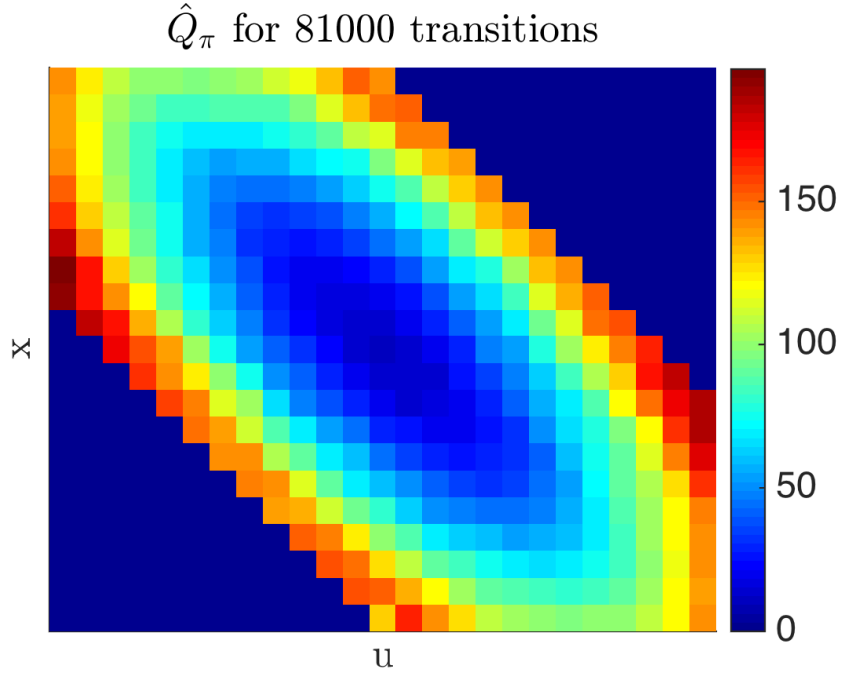
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

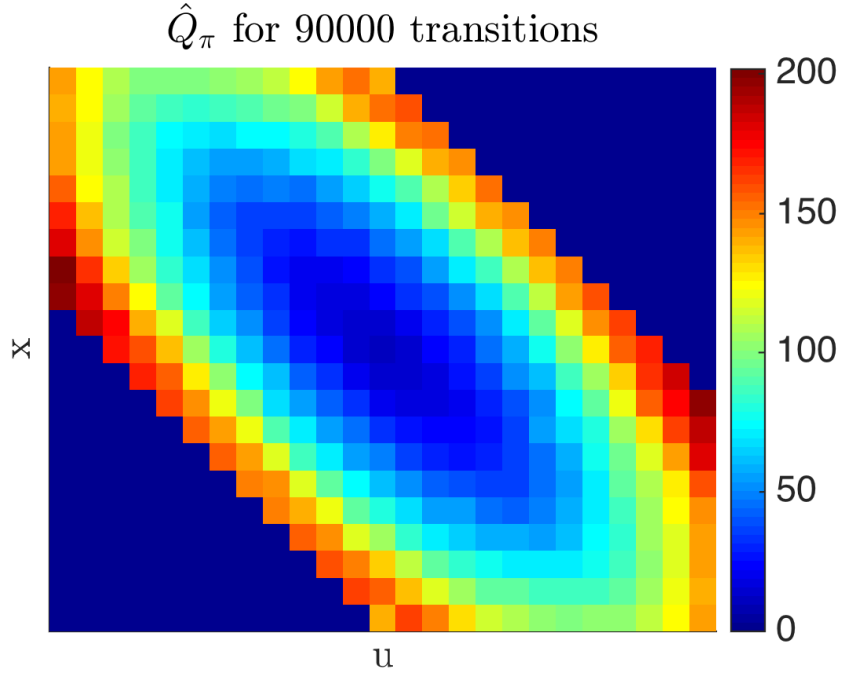
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

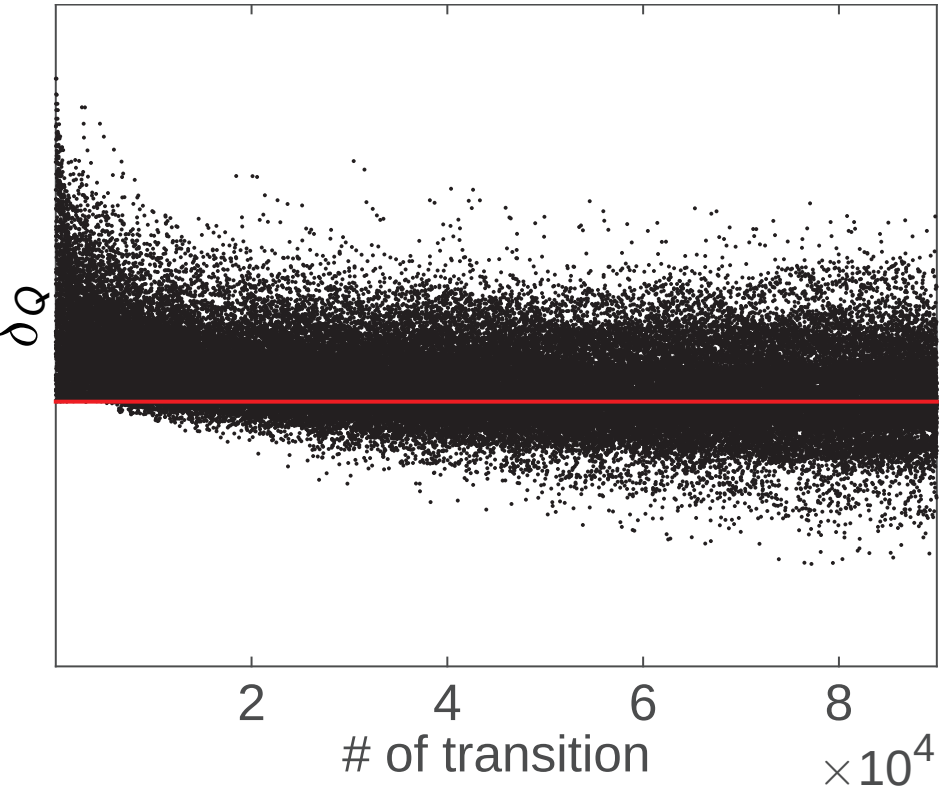
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma\hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Learning the action-value function - Off policy

Temporal Difference (TD) for $\hat{V}_{\pi}$

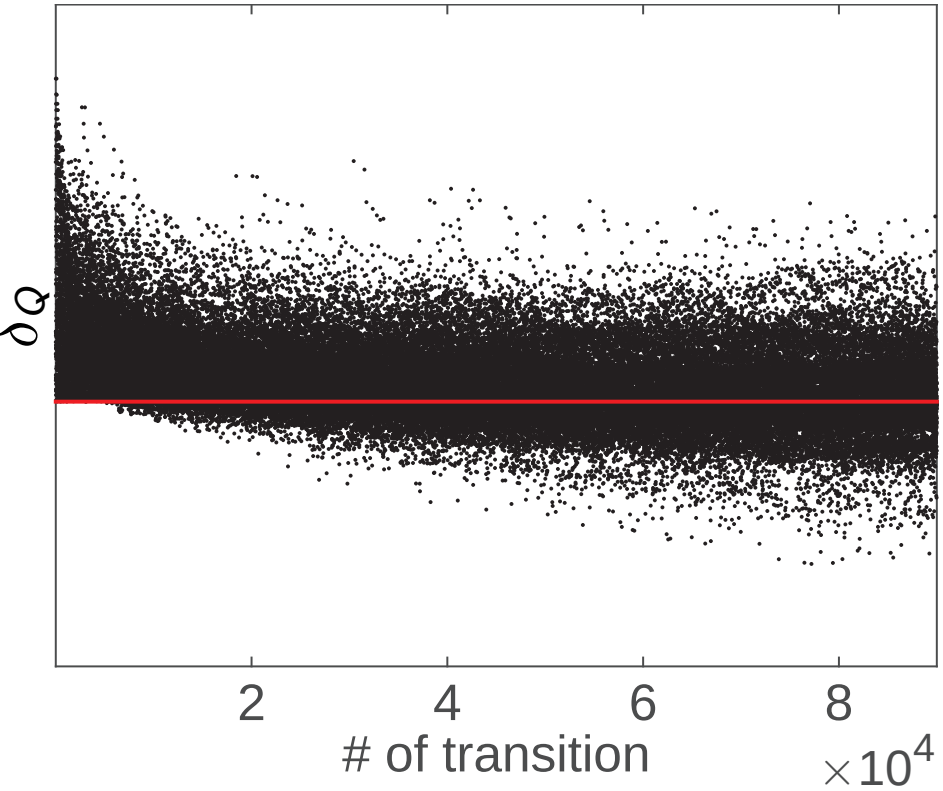
$$\hat{V}_{\pi}(\mathbf{x}_k) \leftarrow \hat{V}_{\pi}(\mathbf{x}_k) + \alpha\delta_k$$

$$\delta_k = L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{V}_{\pi}(\mathbf{x}_k)$$

Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k^Q$$

$$\delta_k^Q = L(\mathbf{x}, \mathbf{u}_k) + \gamma \hat{V}_{\pi}(\mathbf{x}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$



Can we pick $\mathbf{u}_{0,\ldots,\infty}$ arbitrarily? Not really...

- must visit all state-input $\mathbf{x}$, $\mathbf{u}$ pairs infinitely many times to build

$$\hat{Q}_{\pi}(\mathbf{x}, \mathbf{u}) = Q_{\pi}(\mathbf{x}, \mathbf{u})$$

- "partial picture" is built from visiting only subparts of the state-input space

Outline

Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Dynamic Programming (DP)

$$\hat{Q}_\star(\mathbf{x},\mathbf{u}) \leftarrow \mathbb{E}\left[L(\mathbf{x},\mathbf{u}) + \gamma \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x}_+,\mathbf{u}') \,\middle|\, \mathbf{x},\mathbf{u}\right]$$

Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Dynamic Programming (DP)

$$\hat{Q}_\star(\mathbf{x},\mathbf{u}) \leftarrow \mathbb{E}\left[L(\mathbf{x},\mathbf{u}) + \gamma \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x}_+,\mathbf{u}') \,\middle|\, \mathbf{x},\mathbf{u}\right]$$

Temporal Difference (TD)

$$\hat{Q}_\star(\mathbf{x},\mathbf{u}) \leftarrow (1-\alpha)\,\hat{Q}_\star(\mathbf{x},\mathbf{u}) + \alpha\left(L(\mathbf{x},\mathbf{u}) + \gamma \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x}_+,\mathbf{u}')\right)$$

where $\mathbf{u}$ can be arbitrary.

Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

Dynamic Programming (DP)

$$\hat{Q}_\star(\mathbf{x},\mathbf{u}) \leftarrow \mathbb{E}\left[L(\mathbf{x},\mathbf{u}) + \gamma \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x}_+,\mathbf{u}') \,\middle|\, \mathbf{x},\mathbf{u}\right]$$

Temporal Difference (TD)

$$\hat{Q}_\star(\mathbf{x},\mathbf{u}) \leftarrow (1-\alpha)\hat{Q}_\star(\mathbf{x},\mathbf{u}) + \alpha\left(L(\mathbf{x},\mathbf{u}) + \gamma \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x}_+,\mathbf{u}')\right)$$

where $\mathbf{u}$ can be arbitrary.

also reads as **Off-policy TD control**:

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\delta_k$$
$$\delta_k = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x}_{k+1},\mathbf{u}') - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0=\mathbf{u},\ \mathbf{u}_{k>0}=\boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0=\mathbf{x}\right]$$

TD evaluation

$$\hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1},\boldsymbol{\pi}(\mathbf{x}_{k+1})) - \hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}_k,\mathbf{u}_k)$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma \underbrace{\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x}_{k+1},\mathbf{u}')}_{:=\hat{V}_\star(\mathbf{x}_{k+1})} - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0=\mathbf{u},\ \mathbf{u}_{k>0}=\boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0=\mathbf{x}\right]$$

TD evaluation

$$\hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1},\boldsymbol{\pi}(\mathbf{x}_{k+1})) - \hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}_k,\mathbf{u}_k)$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma \underbrace{\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x}_{k+1},\mathbf{u}')}_{:=\hat{V}_\star(\mathbf{x}_{k+1})} - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

- Selection of $\mathbf{u}_{0,\ldots,\infty}$ must yield ∞-many visits of all state-input pairs to get

$$\hat{Q}_\star(\mathbf{x},\mathbf{u}) \to Q_\star(\mathbf{x},\mathbf{u})$$

Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD evaluation

$$\hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1},\boldsymbol{\pi}(\mathbf{x}_{k+1})) - \hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}_k,\mathbf{u}_k)$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma \underbrace{\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x}_{k+1},\mathbf{u}')}_{:=\hat{V}_\star(\mathbf{x}_{k+1})} - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

- Selection of $\mathbf{u}_{0,\ldots,\infty}$ must yield ∞-many visits of all state-input pairs to get

$$\hat{Q}_\star(\mathbf{x},\mathbf{u}) \to Q_\star(\mathbf{x},\mathbf{u})$$

- Optimal policy is $\hat{\pi}_\star(\mathbf{x}) = \arg\min_{\mathbf{u}} \hat{Q}_\star(\mathbf{x},\mathbf{u})$

Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

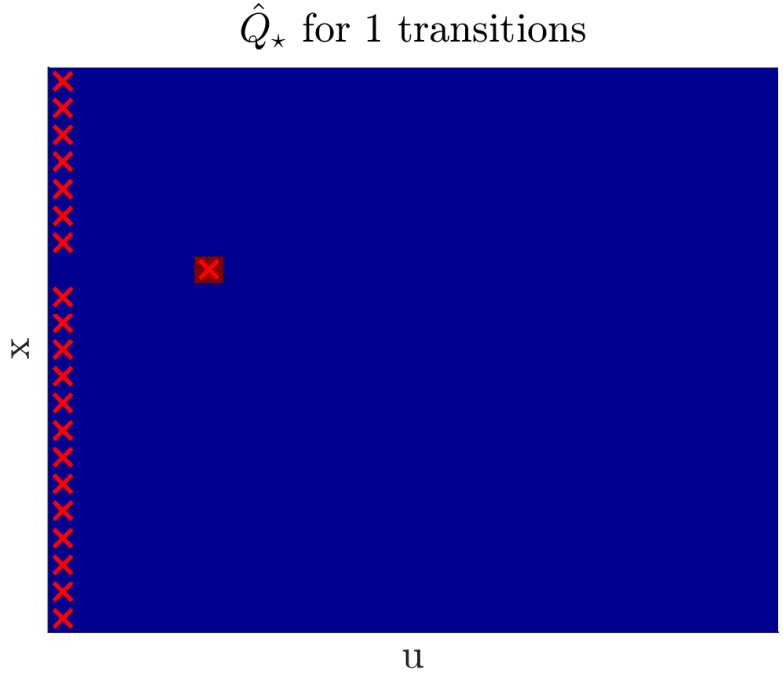
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

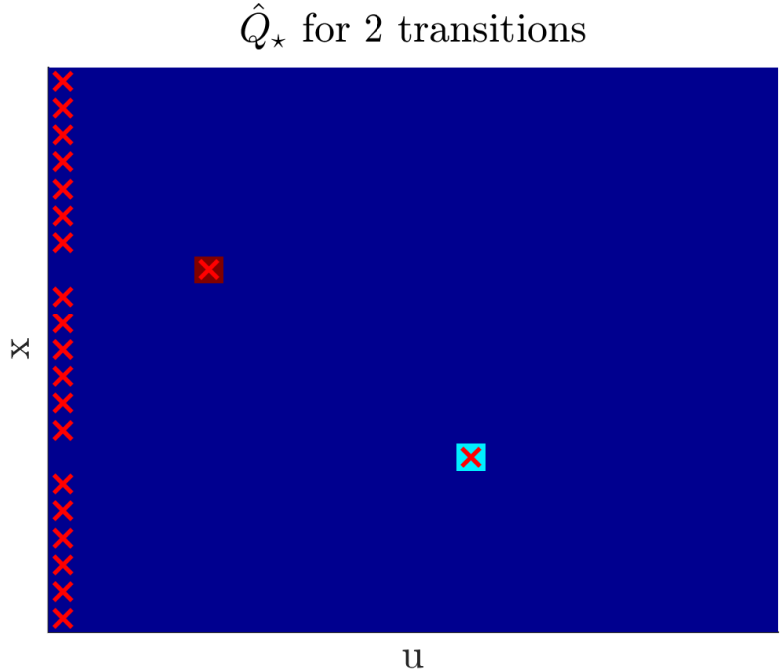
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

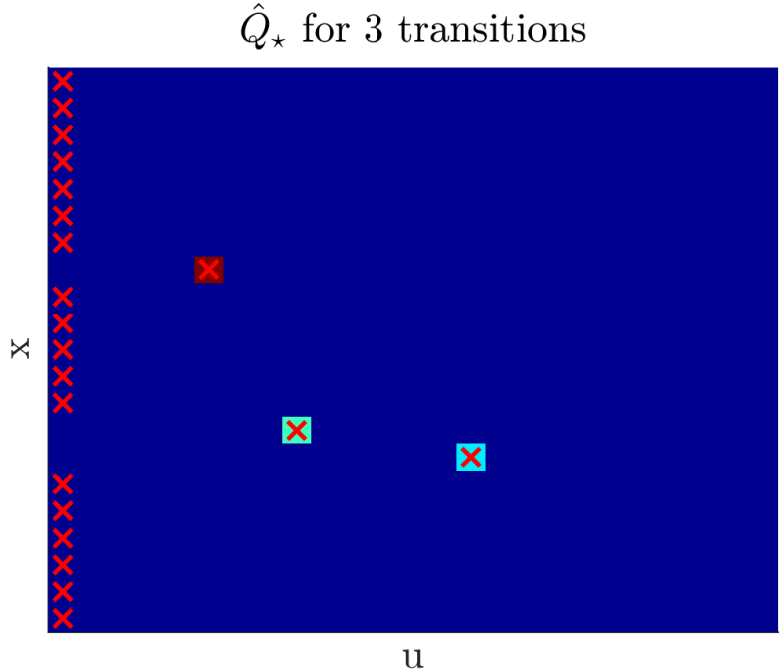
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{x}_k,\mathbf{u}_k)\,\middle|\,\mathbf{u}_0=\mathbf{u},\ \mathbf{u}_{k>0}=\boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0=\mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

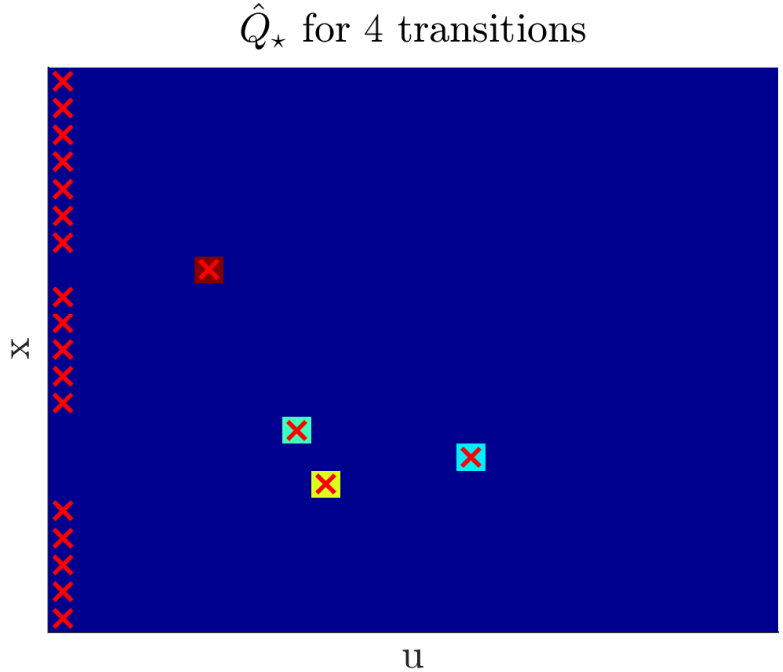
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'}\hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'}\hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

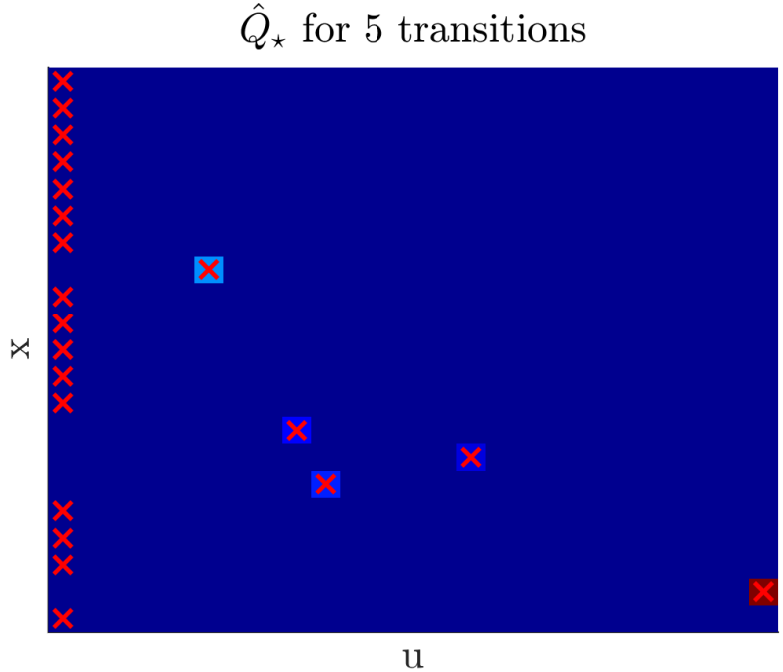
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

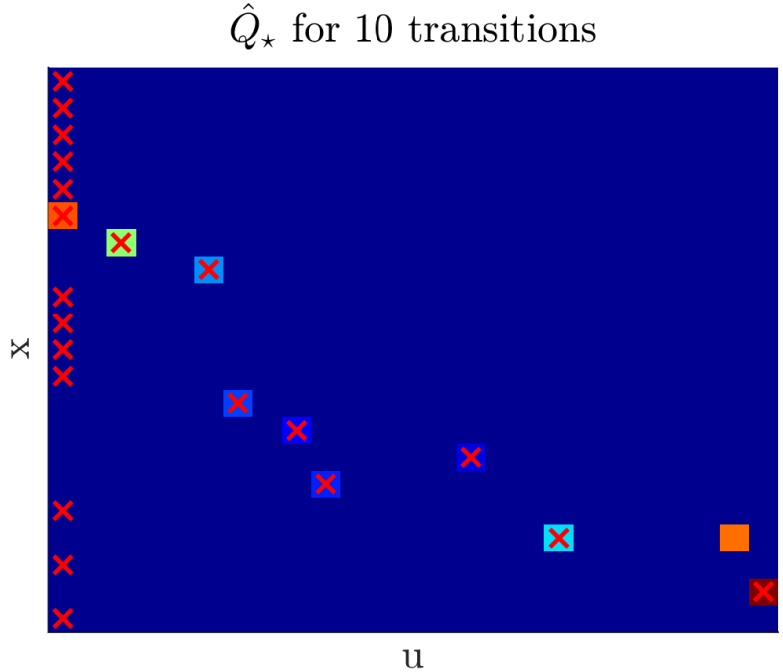
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{x}_k,\mathbf{u}_k)\,\middle|\,\mathbf{u}_0=\mathbf{u},\ \mathbf{u}_{k>0}=\boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0=\mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

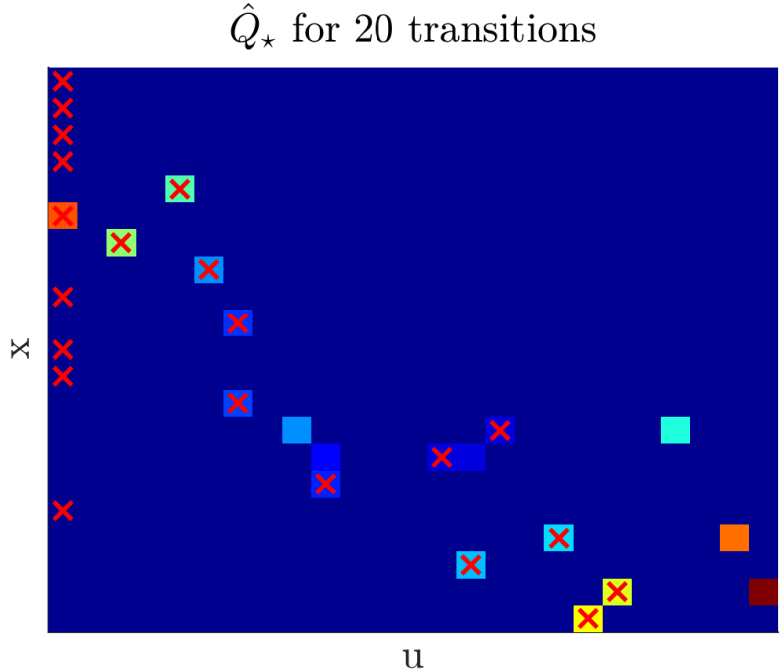
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'}\hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'}\hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

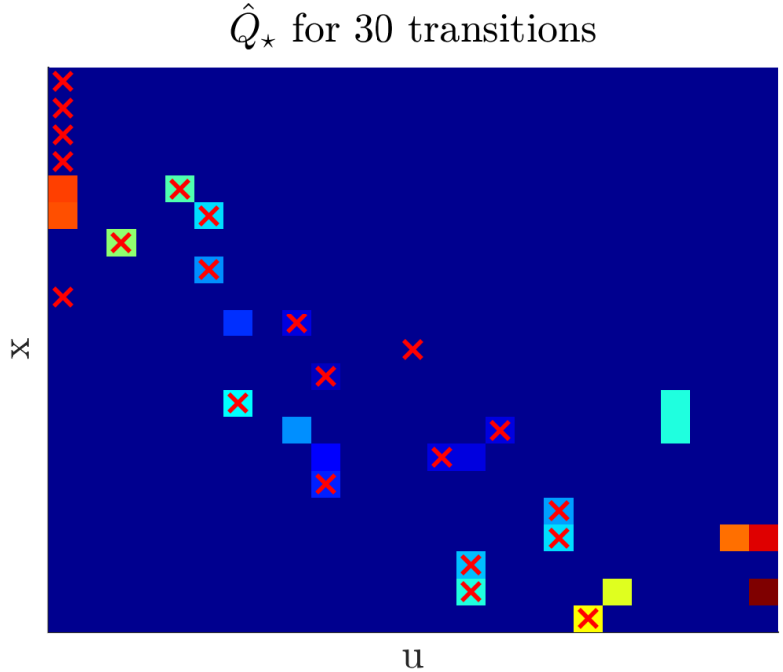
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

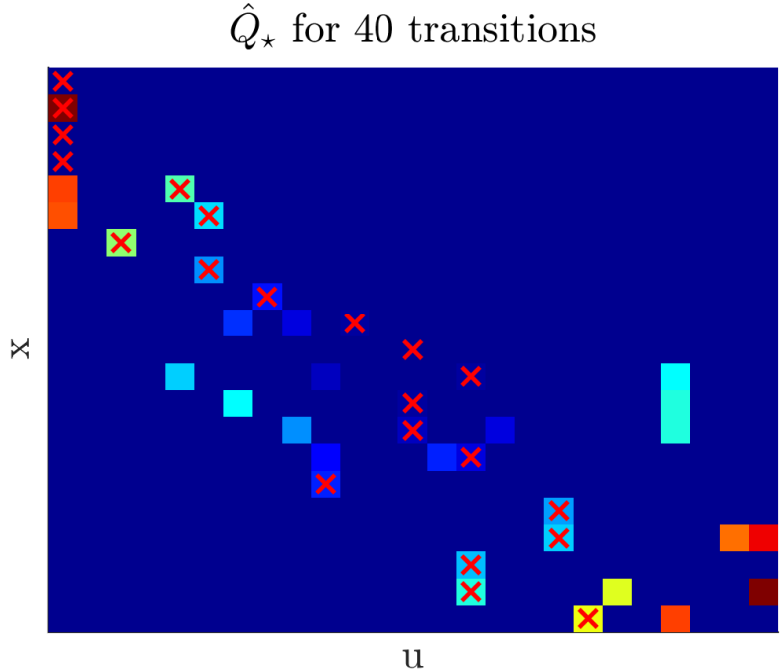
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{x}_k,\mathbf{u}_k)\,\middle|\,\mathbf{u}_0=\mathbf{u},\ \mathbf{u}_{k>0}=\boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0=\mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

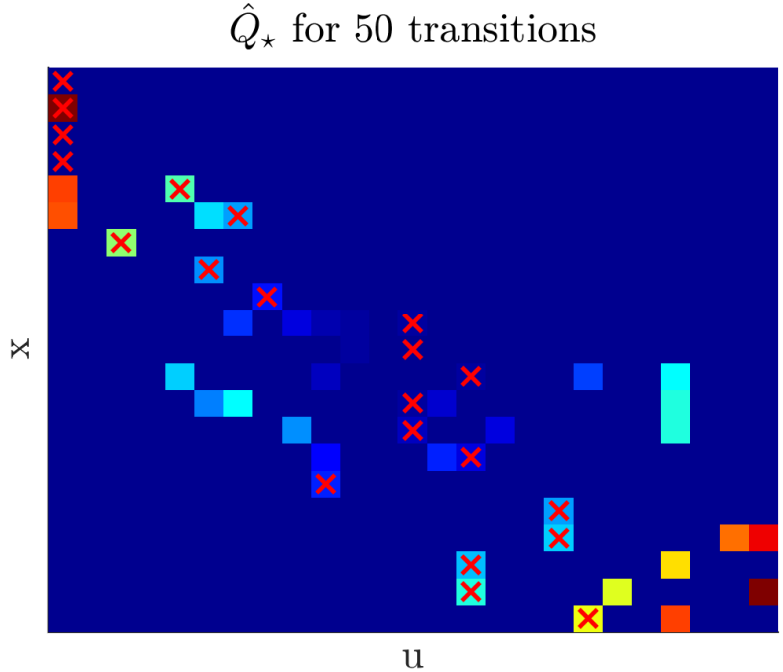
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'}\hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'}\hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{x}_k,\mathbf{u}_k)\,\middle|\,\mathbf{u}_0=\mathbf{u},\ \mathbf{u}_{k>0}=\boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0=\mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

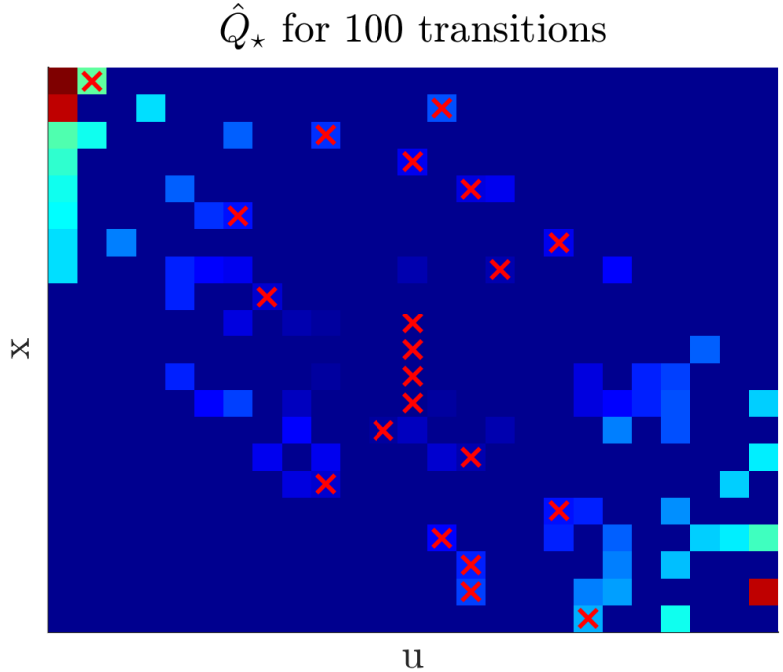
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'}\hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'}\hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

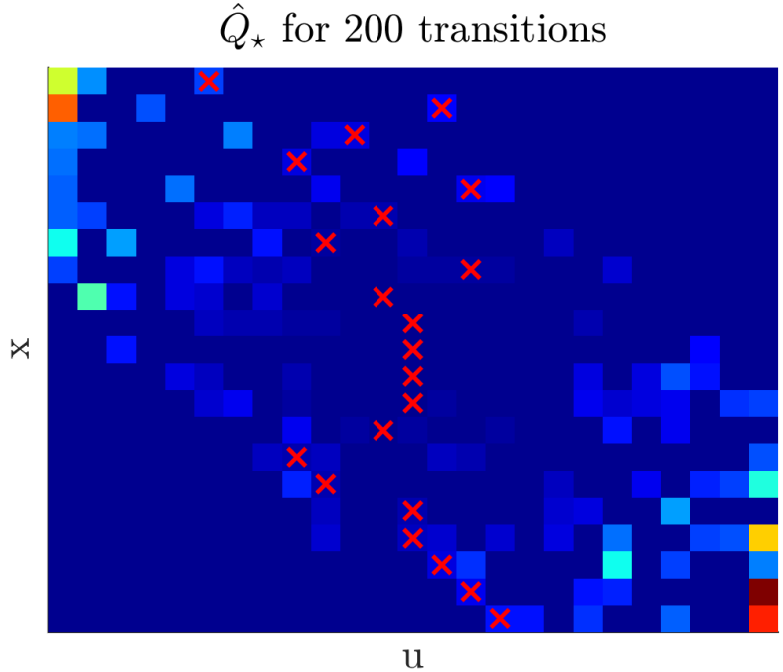
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

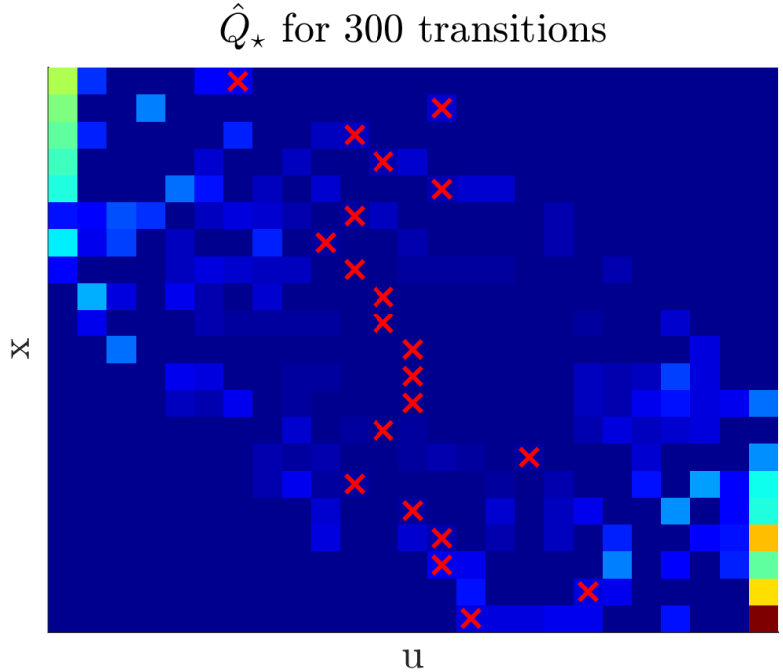
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

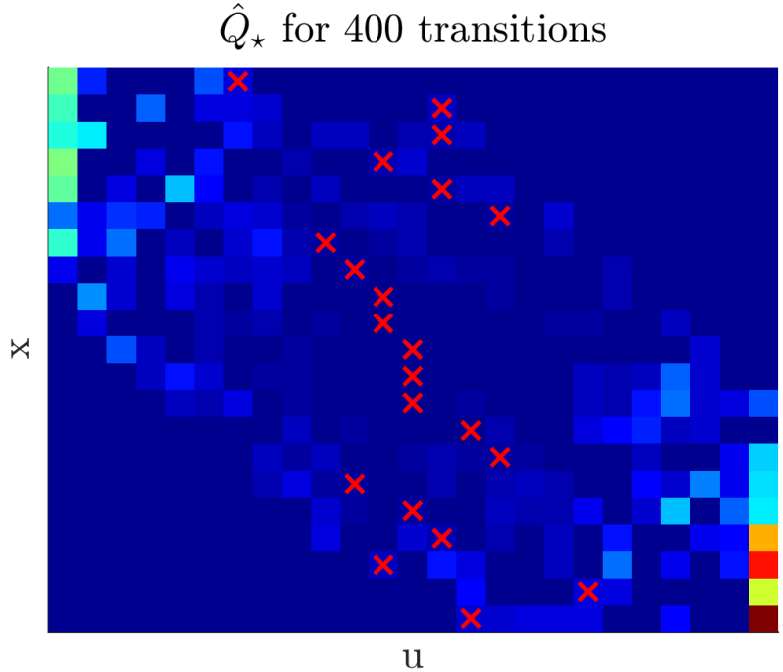
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

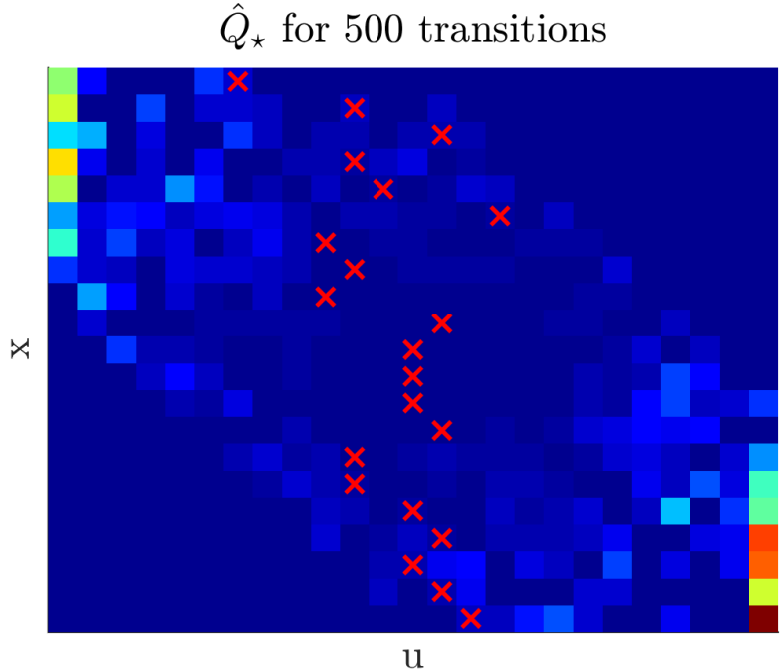
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

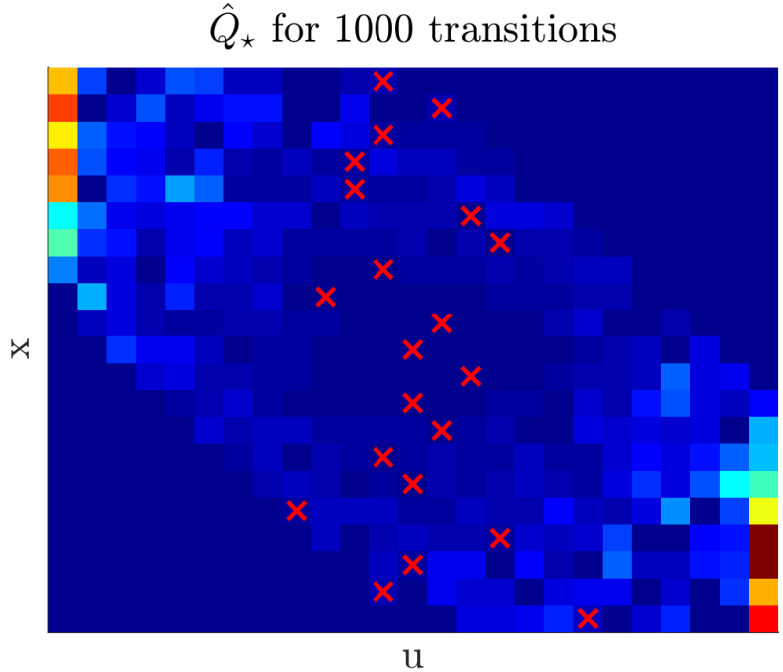
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

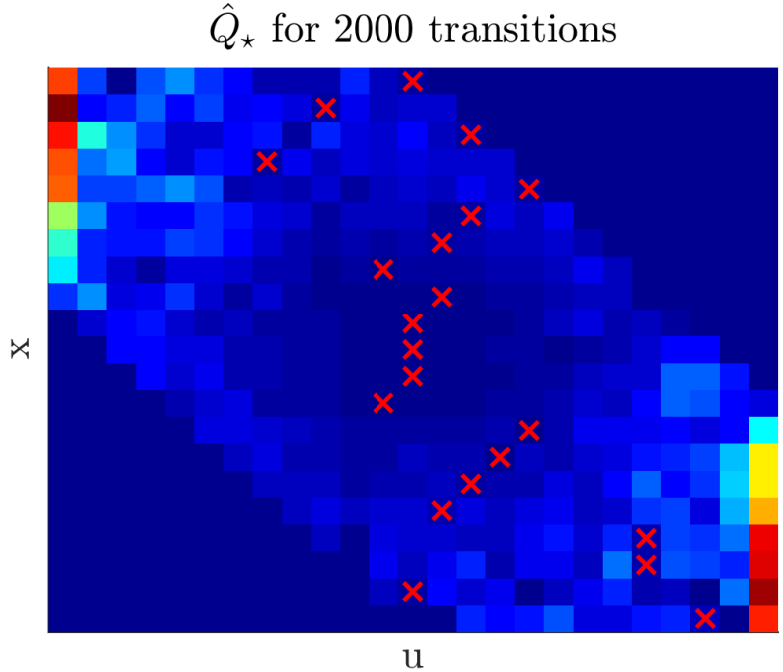
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{x}_k,\mathbf{u}_k)\,\middle|\,\mathbf{u}_0=\mathbf{u},\ \mathbf{u}_{k>0}=\boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0=\mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

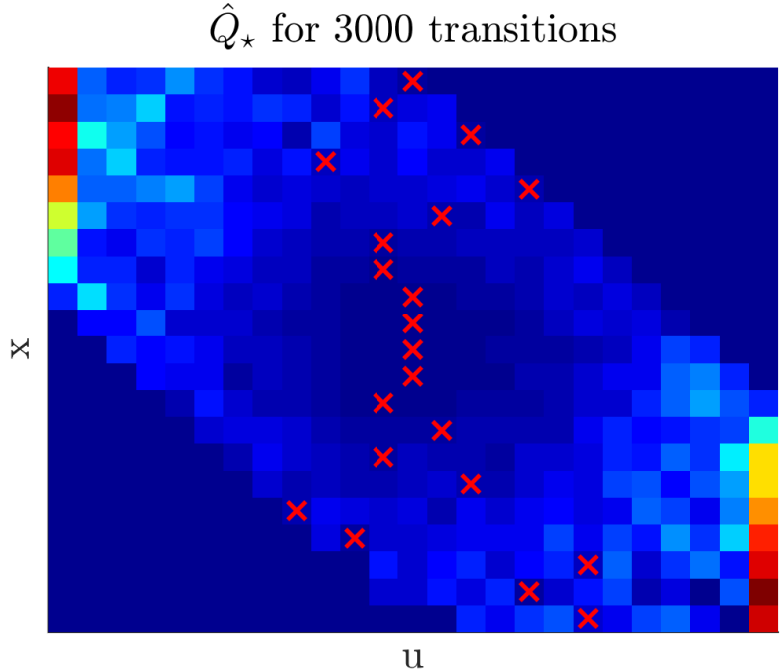
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'}\hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'}\hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

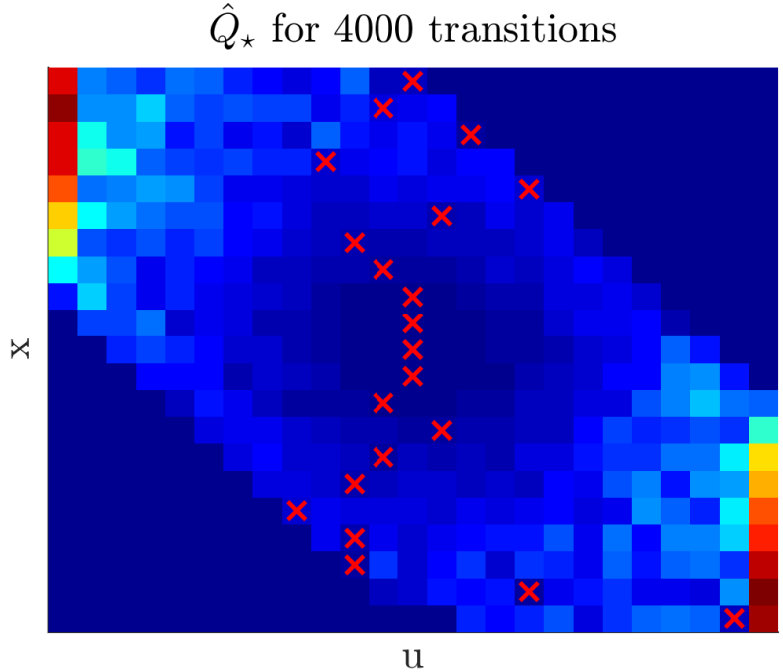
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

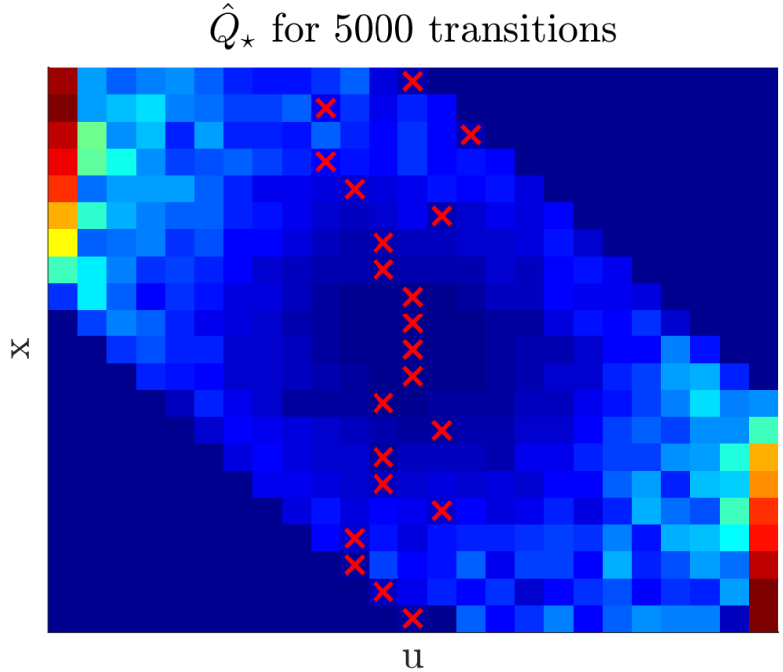
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

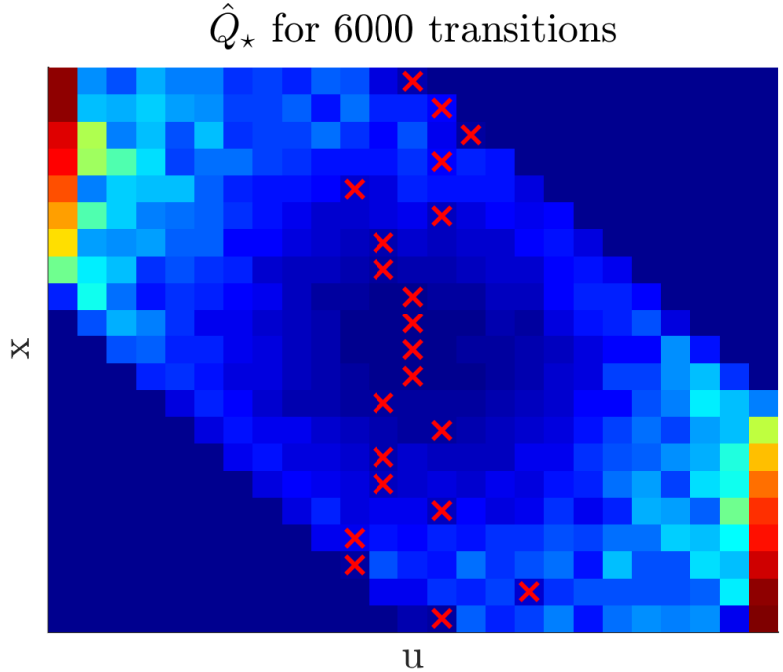
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

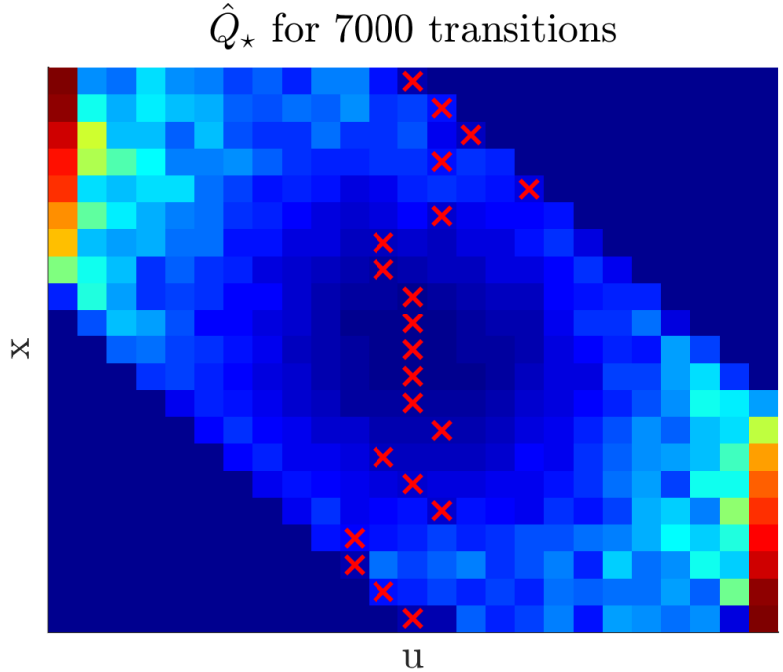
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

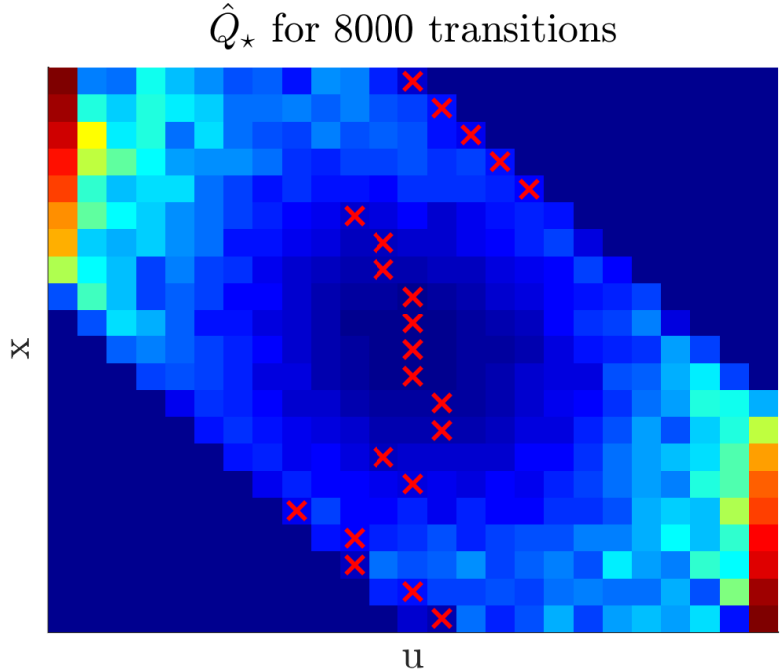
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

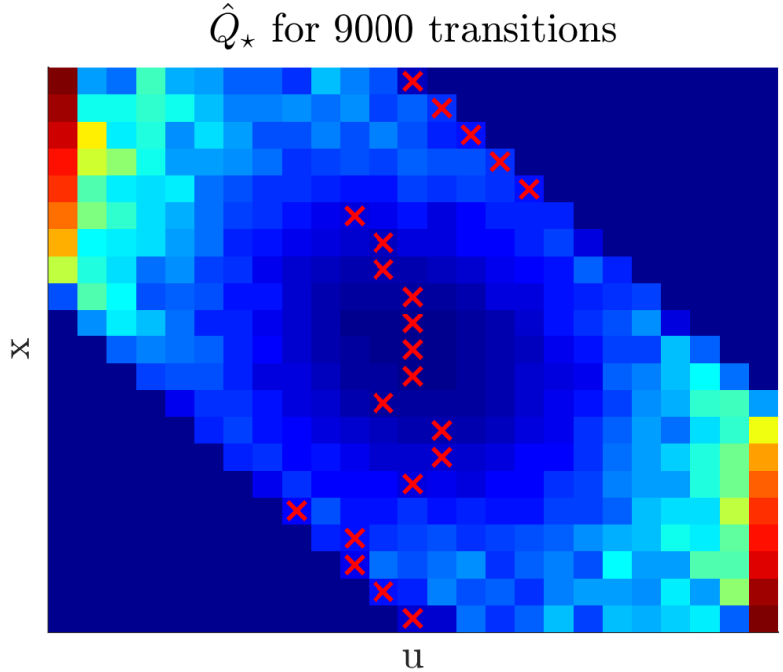
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

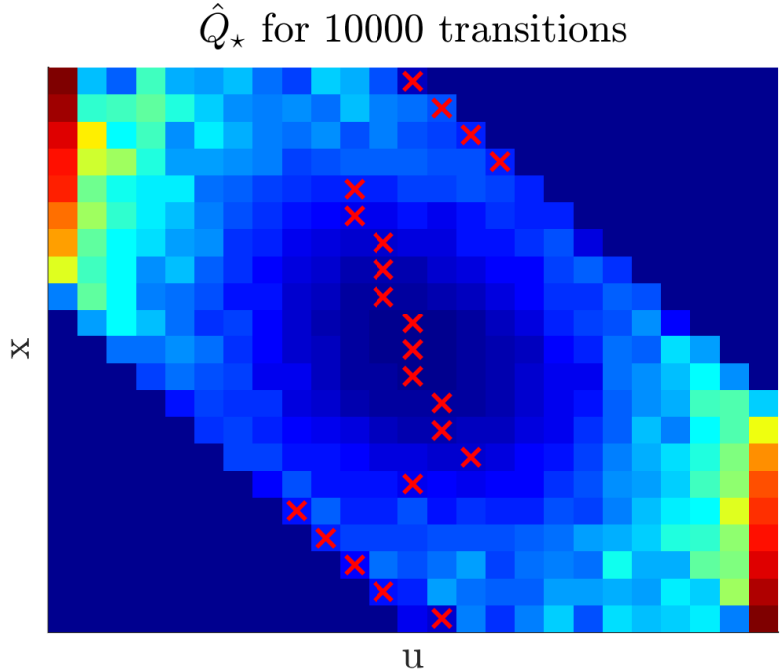
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

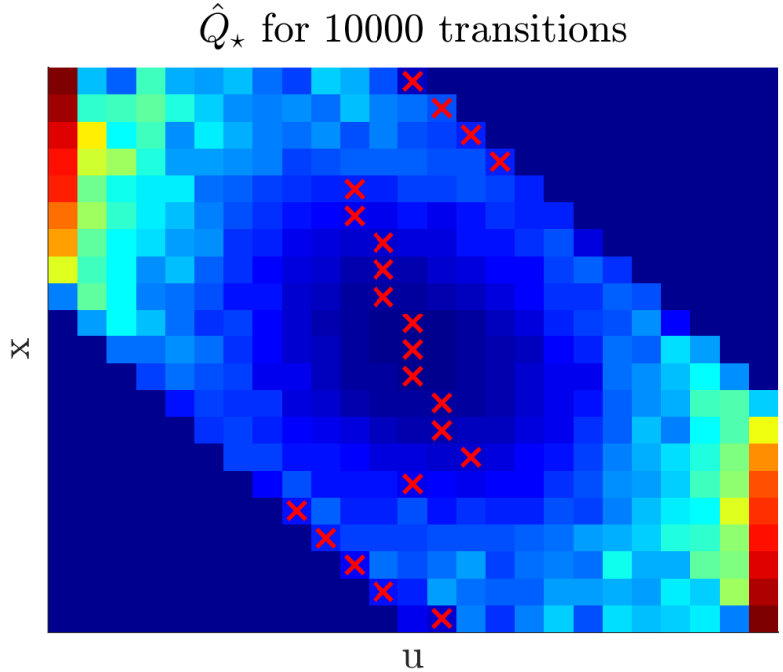
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

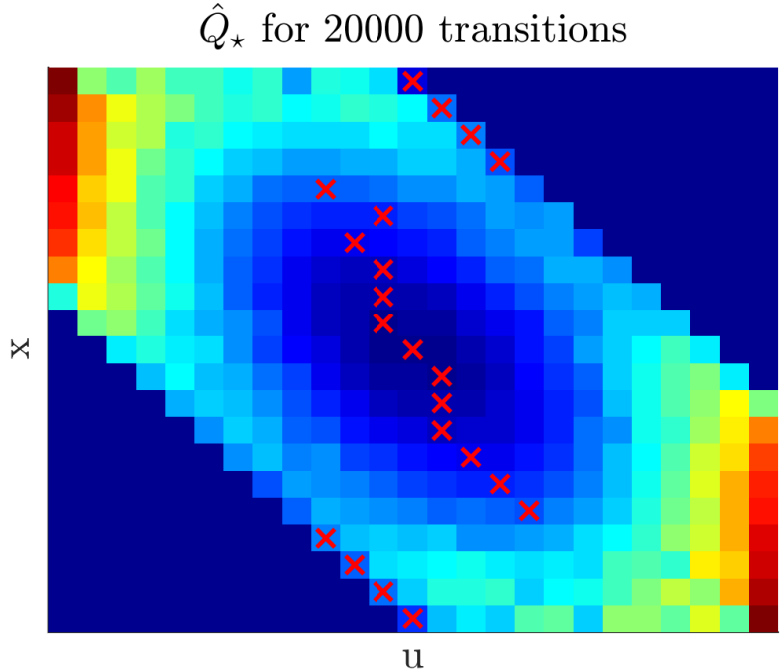
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

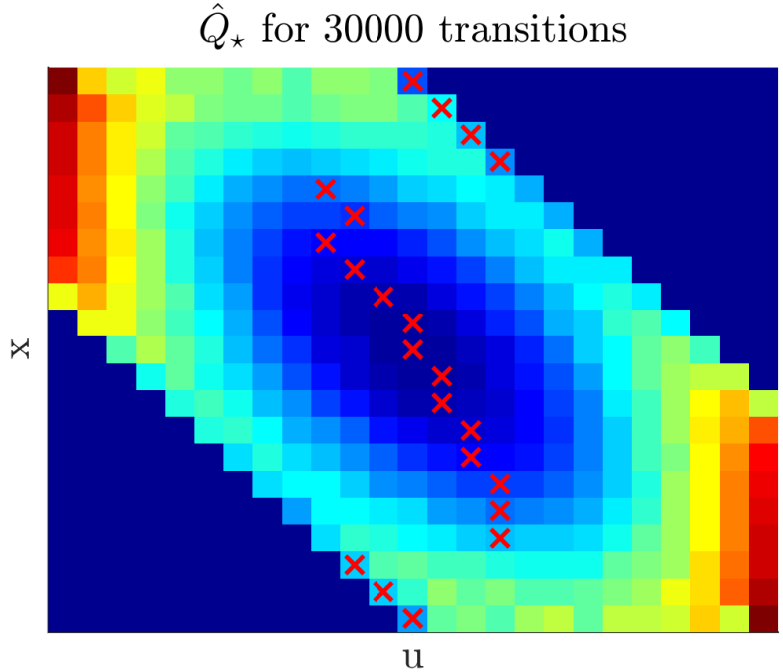
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

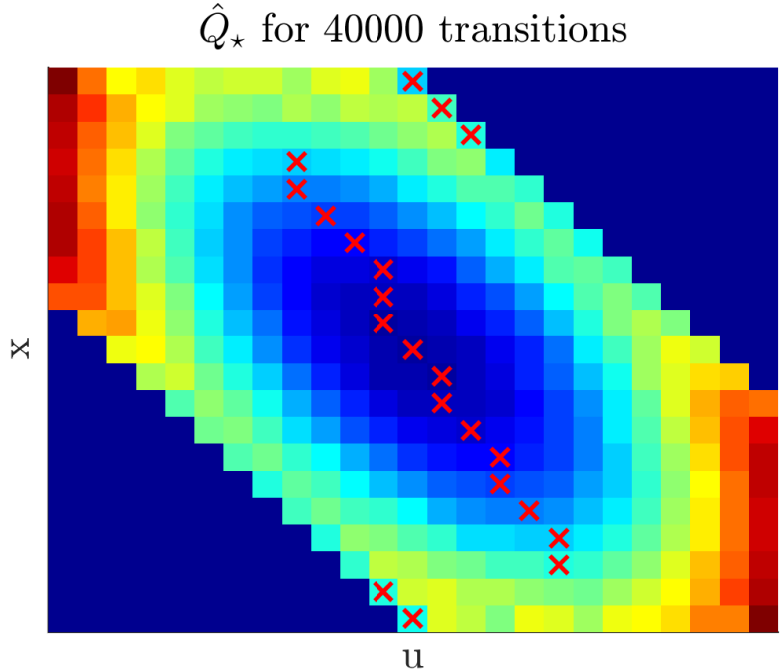
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{x}_k,\mathbf{u}_k)\,\middle|\,\mathbf{u}_0=\mathbf{u},\ \mathbf{u}_{k>0}=\boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0=\mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

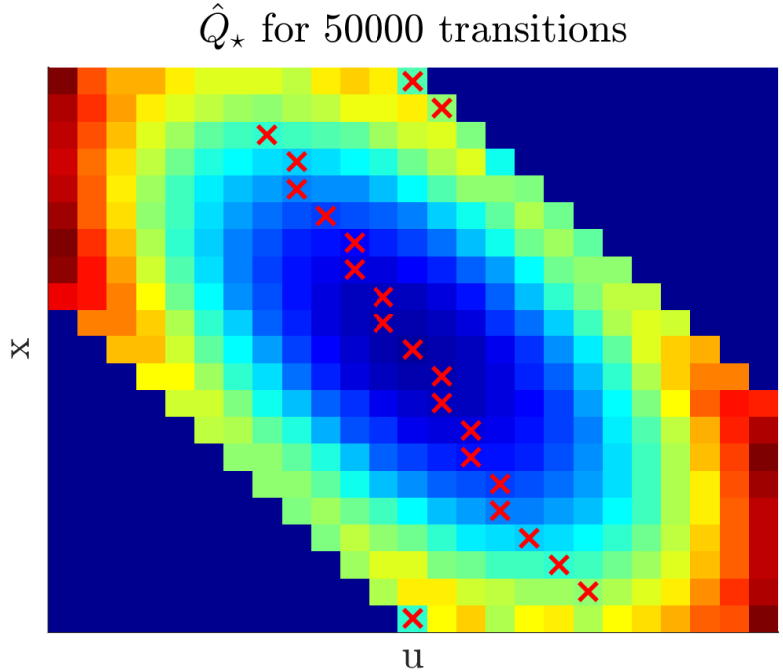
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

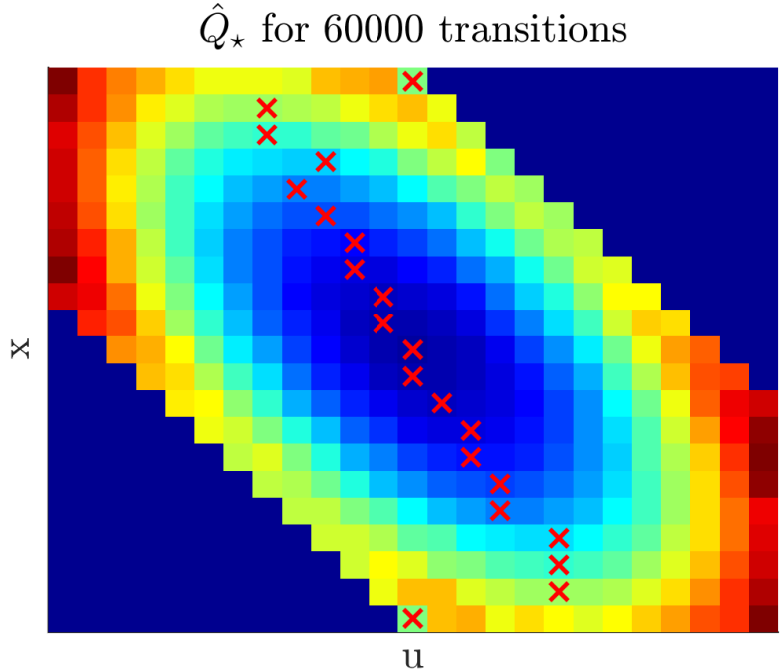
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

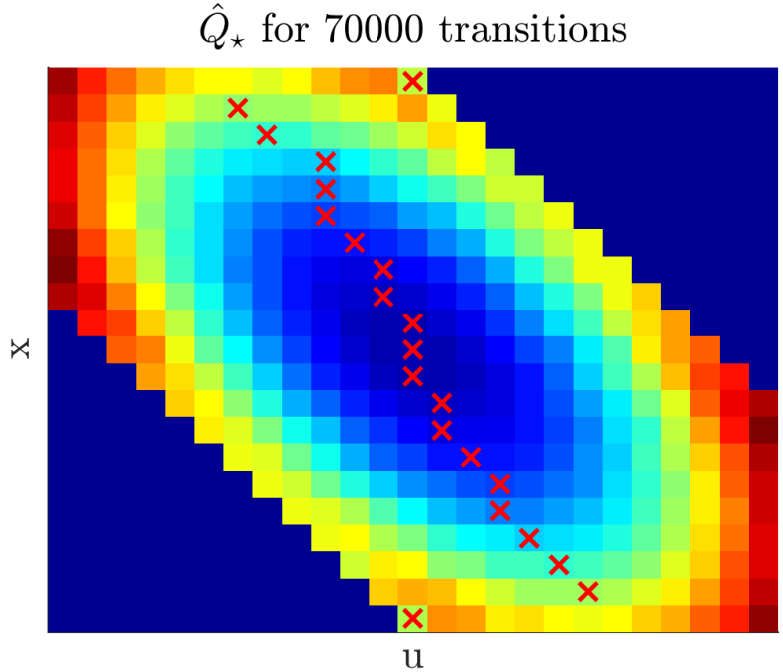
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

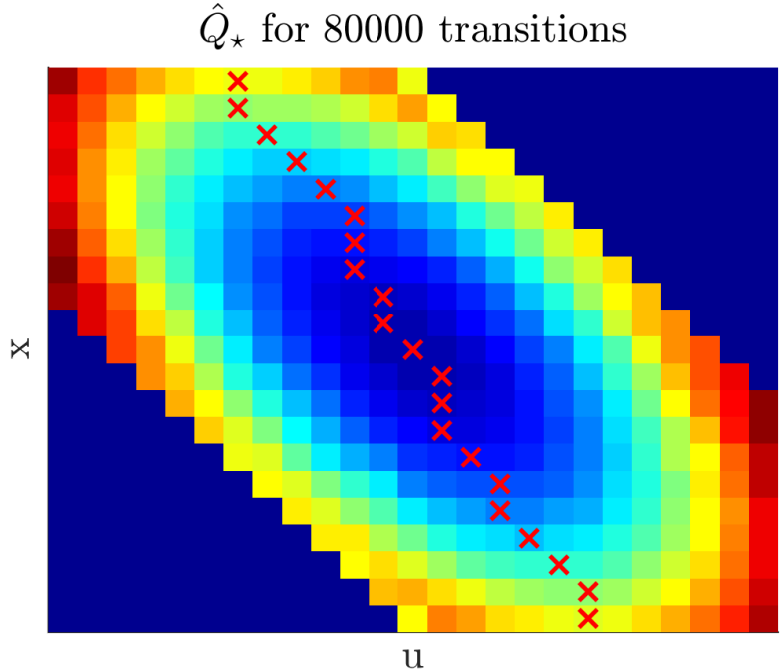
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k,\mathbf{u}_k) \,\middle|\, \mathbf{u}_0 = \mathbf{u},\ \mathbf{u}_{k>0} = \boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0 = \mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

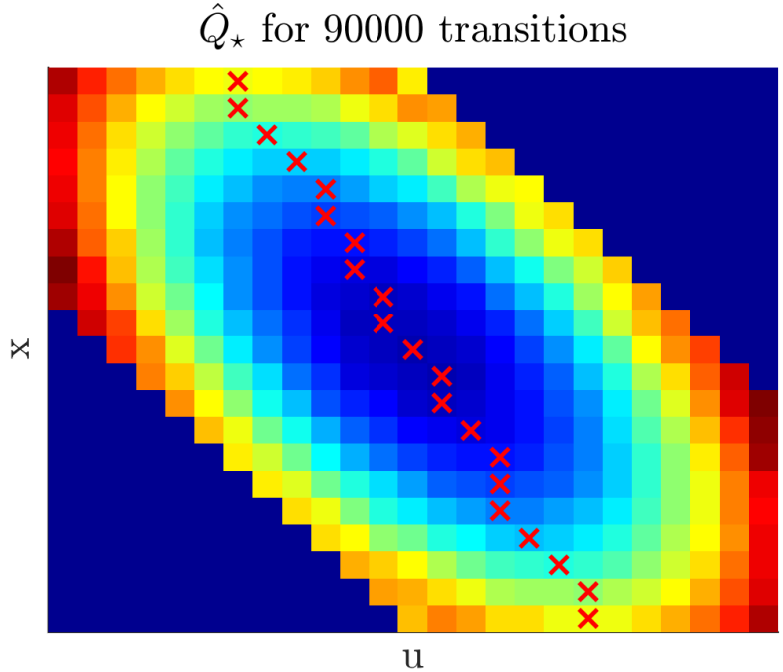
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{x}_k,\mathbf{u}_k)\,\middle|\,\mathbf{u}_0=\mathbf{u},\ \mathbf{u}_{k>0}=\boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0=\mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

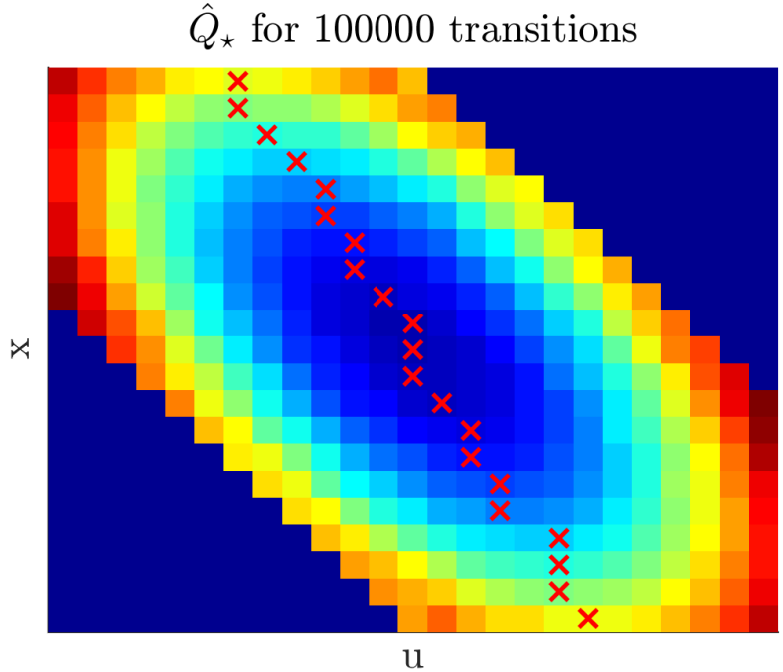
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Learning the optimal action-value function - Off-policy

Optimal action-value function for policy π:

$$Q_\star(\mathbf{x},\mathbf{u}) = \min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{x}_k,\mathbf{u}_k)\,\middle|\,\mathbf{u}_0=\mathbf{u},\ \mathbf{u}_{k>0}=\boldsymbol{\pi}(\mathbf{x}_k),\ \mathbf{x}_0=\mathbf{x}\right]$$

TD control

$$\hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) \leftarrow \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k) + \alpha\,\delta_k^Q$$

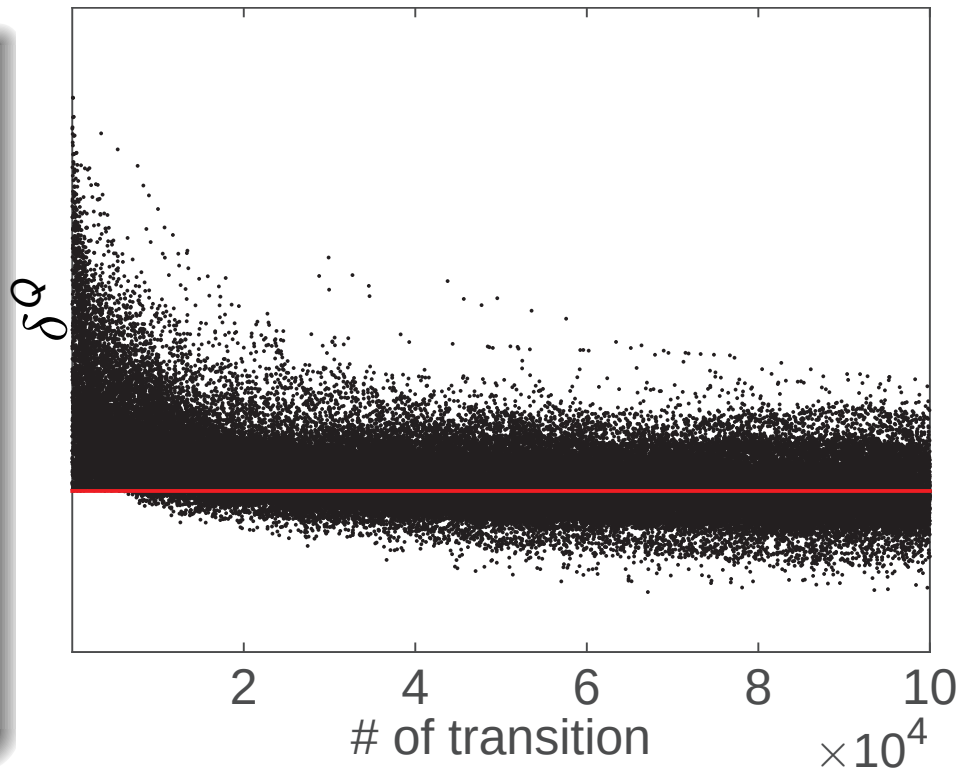
$$\delta_k^Q = L(\mathbf{x}_k,\mathbf{u}_k) + \gamma\,\hat{V}_\star(\mathbf{x}_{k+1}) - \hat{Q}_\star(\mathbf{x}_k,\mathbf{u}_k)$$

where

$$\hat{V}_\star(\mathbf{x}) = \min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$

Policy defined as:

$$\hat{\boldsymbol{\pi}}(\mathbf{x}) = \arg\min_{\mathbf{u}'} \hat{Q}_\star(\mathbf{x},\mathbf{u}')$$



Eligibility trace for learning the optimal action-value function

1-step backup

TD for V_π evaluation:

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}_k) + \alpha\delta_k$$

"∞" backup

TD(λ) for V_π evaluation:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}$$

$$\hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) \leftarrow \hat{V}_{\boldsymbol{\pi}}(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

Eligibility trace for learning the optimal action-value function

1-step backup

TD for V_π evaluation:

$$\hat{V}_\pi(\mathbf{x}_k) \leftarrow \hat{V}_\pi(\mathbf{x}_k) + \alpha\delta_k$$

"∞" backup

TD(λ) for V_π evaluation:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}$$

$$\hat{V}_\pi(\mathbf{x}) \leftarrow \hat{V}_\pi(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

TD for $\hat{Q}_\pi$ - Off-policy

$$\hat{Q}_\pi(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_\pi(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k$$

where $\mathbf{u}$ can be chosen arbitrarily

Eligibility trace for learning the optimal action-value function

1-step backup

TD for V_π evaluation:

$$\hat{V}_\pi(\mathbf{x}_k) \leftarrow \hat{V}_\pi(\mathbf{x}_k) + \alpha\delta_k$$

"∞" backup

TD(λ) for V_π evaluation:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}$$

$$\hat{V}_\pi(\mathbf{x}) \leftarrow \hat{V}_\pi(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

TD for $\hat{Q}_\pi$ - Off-policy

$$\hat{Q}_\pi(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_\pi(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k$$

where $\mathbf{u}$ can be chosen arbitrarily

Can we use E for $\hat{Q}_\pi$???

Eligibility trace for learning the optimal action-value function

1-step backup

TD for V_π evaluation:

$$\hat{V}_\pi(\mathbf{x}_k) \leftarrow \hat{V}_\pi(\mathbf{x}_k) + \alpha\delta_k$$

"∞" backup

TD(λ) for V_π evaluation:

$$E(\mathbf{x}) \leftarrow \gamma\lambda E(\mathbf{x}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k \\ 0 & \text{if} \quad \mathbf{x} \neq \mathbf{x}_k \end{cases}$$

$$\hat{V}_\pi(\mathbf{x}) \leftarrow \hat{V}_\pi(\mathbf{x}) + \alpha E(\mathbf{x})\delta_k, \quad \forall \mathbf{x}$$

TD for $\hat{Q}_\pi$ - Off-policy

$$\hat{Q}_\pi(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_\pi(\mathbf{x}_k, \mathbf{u}_k) + \alpha\delta_k$$

where $\mathbf{u}$ can be chosen arbitrarily

Can we use E for $\hat{Q}_\pi$???

Yes, but only on-policy!

Learning the action-value function - On policy (SARSA)

Off-policy Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha \delta_k^Q$$
$$\delta_k^Q = L(\mathbf{x}_k, \mathbf{u}_k) + \gamma \hat{Q}_{\pi}(\mathbf{x}_{k+1}, \boldsymbol{\pi}(\mathbf{x}_{k+1})) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$

where $\mathbf{u}_k$ can be "arbitrary"

On-policy Temporal Difference (TD) for $\hat{Q}_{\pi}$

$$\hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k) + \alpha \delta_k^Q$$
$$\delta_k^Q = L(\mathbf{x}_k, \mathbf{u}_k) + \gamma \hat{Q}_{\pi}(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) - \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u}_k)$$

where $\mathbf{u}_k$ *must* be chosen from $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k) + \mathrm{d}_k$

Disturbance d_k must decay over the iterations!

Remarks: if $\boldsymbol{\pi}$ is defined as:

$$\boldsymbol{\pi}(\mathbf{x}_k) = \mathrm{a} \min_{\mathbf{u}} \hat{Q}_{\pi}(\mathbf{x}_k, \mathbf{u})$$

throughout the iterations, then

$$\hat{Q}_{\pi} \rightarrow Q_{\star} \text{ and } \boldsymbol{\pi} \rightarrow \boldsymbol{\pi}_{\star}$$

under some conditions...

SARSA(λ)

SARSA with eligibility trace - "Emulates" n-step Monte Carlo:

$$E(\mathbf{x}, \mathbf{u}) \leftarrow \gamma\lambda E(\mathbf{x}, \mathbf{u}) + \begin{cases} 1 & \text{if} \quad \mathbf{x} = \mathbf{x}_k, \mathbf{u} = \mathbf{u} \\ 0 & \text{otherwise} \end{cases}$$

$$\delta^Q = L(\mathbf{x}_k, \mathbf{u}_k) + \gamma \hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) - \hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}_k, \mathbf{u}_k)$$

$$\hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}, \mathbf{u}) \leftarrow \hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}, \mathbf{u}) + \alpha \delta^Q E(\mathbf{x}, \mathbf{u})$$

Remarks

- All $\mathbf{u}_k$ *must* be chosen using $\boldsymbol{\pi}(\mathbf{x}_k) + \mathrm{d}_k$ (c.f. ϵ-greedy)
- δ^Q computed based on latest sarsa transition: $\mathbf{x}_k, \mathbf{u}_k, L, \mathbf{x}_{k+1}, \mathbf{u}_{k+1}$
- Updates of E and $\hat{Q}_{\boldsymbol{\pi}}$ sweep through *all* state-action pairs $\mathbf{x}, \mathbf{u}$
- E keeps a trace of state-input pairs recently visited
- If $\boldsymbol{\pi}$ is updated as

$$\boldsymbol{\pi}(\mathbf{x}_k) = \mathrm{a}\min_{\mathbf{u}} \hat{Q}_{\boldsymbol{\pi}}(\mathbf{x}_k, \mathbf{u})$$

throughout the iterations, then

$$\hat{Q}_{\boldsymbol{\pi}} \to Q_\star \text{ and } \boldsymbol{\pi} \to \boldsymbol{\pi}_\star$$

under some conditions...

Dynamic Programming vs. Temporal-Difference Learning

Full Backup

Iterative π Evaluation

$$V_{\pi}(\mathbf{x}) \leftarrow L(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_{+}) \,|\, \mathbf{x}, \pi(\mathbf{x})\right]$$

sweep on all $\mathbf{x}$

Sampled Backup

TD learning

$$V_{\pi}(\mathbf{x}_k) \leftarrow L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma V_{\pi}(\mathbf{x}_{k+1})$$

update on observed $\mathbf{x}_k \rightarrow \mathbf{x}_{k+1}$

Dynamic Programming vs. Temporal-Difference Learning

Full Backup

Iterative π Evaluation

$$V_{\pi}(\mathbf{x}) \leftarrow L(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_{+}) \,|\, \mathbf{x}, \pi(\mathbf{x})\right]$$

sweep on all $\mathbf{x}$

Q-Evaluation

$$Q_{\pi}(\mathbf{x}, \mathbf{u}) \leftarrow L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[Q_{\pi}(\mathbf{x}_{+}, \pi(\mathbf{x}_{+}) \,|\, \mathbf{x}, \mathbf{u}\right]$$

sweep on all $\mathbf{x}, \mathbf{u}$

Sampled Backup

TD learning

$$V_{\pi}(\mathbf{x}_k) \leftarrow L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma V_{\pi}(\mathbf{x}_{k+1})$$

update on observed $\mathbf{x}_k \rightarrow \mathbf{x}_{k+1}$

TD on Q

$$Q_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow L(\mathbf{x}_k, \mathbf{u}_k) + \gamma Q_{\pi}(\mathbf{x}_{k+1}, \pi(\mathbf{x}_{k+1}))$$

update on observed $\mathbf{x}_k, \mathbf{u}_k \rightarrow \mathbf{x}_{k+1}, \mathbf{u}_{k+1}$

Dynamic Programming vs. Temporal-Difference Learning

Full Backup

Iterative π Evaluation

$$V_{\pi}(\mathbf{x}) \leftarrow L(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_{+}) \,|\, \mathbf{x}, \pi(\mathbf{x})\right]$$

sweep on all $\mathbf{x}$

Q-Evaluation

$$Q_{\pi}(\mathbf{x}, \mathbf{u}) \leftarrow L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[Q_{\pi}(\mathbf{x}_{+}, \pi(\mathbf{x}_{+}) \,|\, \mathbf{x}, \mathbf{u}\right]$$

sweep on all $\mathbf{x}, \mathbf{u}$

Sampled Backup

TD learning

$$V_{\pi}(\mathbf{x}_k) \leftarrow L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma V_{\pi}(\mathbf{x}_{k+1})$$

update on observed $\mathbf{x}_k \rightarrow \mathbf{x}_{k+1}$

SARSA

$$Q_{\pi}(\mathbf{x}_k, \mathbf{u}_k) \leftarrow L(\mathbf{x}_k, \mathbf{u}_k) + \gamma Q_{\pi}(\mathbf{x}_{k+1}, \mathbf{u}_{k+1})$$

update on observed $\mathbf{x}_k, \mathbf{u}_k \rightarrow \mathbf{x}_{k+1}, \mathbf{u}_{k+1}$

Dynamic Programming vs. Temporal-Difference Learning

Full Backup

Iterative π Evaluation

$$V_\pi(\mathbf{x}) \leftarrow L(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E}\left[V_\pi(\mathbf{x}_+) \,|\, \mathbf{x}, \pi(\mathbf{x})\right]$$

sweep on all $\mathbf{x}$

Q-Evaluation

$$Q_\pi(\mathbf{x}, \mathbf{u}) \leftarrow L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[Q_\pi(\mathbf{x}_+, \pi(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u}\right]$$

sweep on all $\mathbf{x}, \mathbf{u}$

Q-Iteration

$$Q_\pi(\mathbf{x}, \mathbf{u}) \leftarrow L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[\min_{\mathbf{u}_+} Q_\pi(\mathbf{x}_+, \mathbf{u}_+) \,|\, \mathbf{x}, \mathbf{u}\right]$$

sweep on all $\mathbf{x}, \mathbf{u}$

Sampled Backup

TD learning

$$V_\pi(\mathbf{x}_k) \leftarrow L(\mathbf{x}_k, \pi(\mathbf{x}_k)) + \gamma V_\pi(\mathbf{x}_{k+1})$$

update on observed $\mathbf{x}_k \rightarrow \mathbf{x}_{k+1}$

SARSA

$$Q_\pi(\mathbf{x}_k, \mathbf{u}_k) \leftarrow L(\mathbf{x}_k, \mathbf{u}_k) + \gamma Q_\pi(\mathbf{x}_{k+1}, \mathbf{u}_{k+1})$$

update on observed $\mathbf{x}_k, \mathbf{u}_k \rightarrow \mathbf{x}_{k+1}, \mathbf{u}_{k+1}$

Q-learning

$$Q_\pi(\mathbf{x}_k, \mathbf{u}_k) \leftarrow L(\mathbf{x}_k, \mathbf{u}_k) + \gamma \min_{\mathbf{u}_+} Q_\pi(\mathbf{x}_{k+1}, \mathbf{u}_+)$$

update on observed $\mathbf{x}_k, \mathbf{u}_k \rightarrow \mathbf{x}_{k+1}$