Continuous Space MDPs

Last Time

- Neural Network Function Approximation

Guiding Questions

Guiding Questions

- What tools do we have to solve MDPs with continuous S and A?

Current Tool-Belt

Today: Four Tools

Notation: Continuous Random Variables

<table>
<tr><th>Term</th><th>Definition</th><th>$\text{Bernoulli}(0.5)$
Coinflip Example</th><th>Uniform Example</th></tr>
<tr><td>support(X)
$x \in X$</td><td>All the values that X can take</td><td>$\{h, t\}$ or $\{0, 1\}$</td><td></td></tr>
<tr><td>Distribution
- Discrete: PMF
- Continuous: PDF</td><td>Maps each value in the support to a real number indicating its probability</td><td>$P(X = 1) = 0.5$
$P(X = 0) = 0.5$
$P(X)$ is a table
<table>
<tr><th>X</th><th>P(X)</th></tr>
<tr><td>0</td><td>0.5</td></tr>
<tr><td>1</td><td>0.5</td></tr>
</table>
</td><td></td></tr>
<tr><td>Expectation
$E[X]$</td><td>First moment of the random variable, "mean"</td><td>$E[X] = \sum_{x \in X} xP(x)$
$= 0.5$</td><td></td></tr>
</table>

Notation: Continuous Random Variables

Term | **Definition** | $\text{Bernoulli}(0.5)$ **Coinflip Example** | $\mathcal{U}([0,1])$ **Uniform Example**

support(X)

$x \in X$

All the values that X can take

$\{h, t\}$ or $\{0, 1\}$

Distribution

- Discrete: PMF
- Continuous: PDF

Maps each value in the support to a real number indicating its probability

$$P(X = 1) = 0.5$$

$$P(X = 0) = 0.5$$

$P(X)$ is a table

X	**P(X)**
0	0.5
1	0.5

Expectation

$E[X]$

First moment of the random variable, "mean"

$$E[X] = \sum_{x \in X} x P(x)$$

$$= 0.5$$

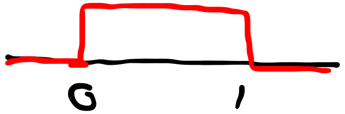
Notation: Continuous Random Variables

<table>
<tr><th>Term</th><th>Definition</th><th>$\text{Bernoulli}(0.5)$
Coinflip Example</th><th>$\mathcal{U}([0,1])$
Uniform Example</th></tr>
<tr><td>support(X)
$x \in X$</td><td>All the values that X can take</td><td>$\{h, t\}$ or $\{0, 1\}$</td><td>$[0, 1]$</td></tr>
<tr><td>Distribution
Discrete: PMF
Continuous: PDF</td><td>Maps each value in the support to a real number indicating its probability</td><td>$P(X = 1) = 0.5$
$P(X = 0) = 0.5$
$P(X)$ is a table<table><tr><th>X</th><th>P(X)</th></tr><tr><td>0</td><td>0.5</td></tr><tr><td>1</td><td>0.5</td></tr></table></td><td></td></tr>
<tr><td>Expectation
$E[X]$</td><td>First moment of the random variable, "mean"</td><td>$E[X] = \sum_{x \in X} xP(x)$
$= 0.5$</td><td></td></tr>
</table>

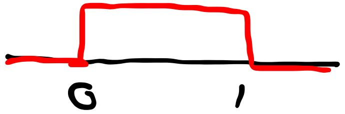
Notation: Continuous Random Variables

<table>
<tr><th>Term</th><th>Definition</th><th>$\text{Bernoulli}(0.5)$
Coinflip Example</th><th>$\mathcal{U}([0,1])$
Uniform Example</th></tr>
<tr><td>support(X)
$x \in X$</td><td>All the values that X can take</td><td>$\{h, t\}$ or $\{0, 1\}$</td><td>$[0, 1]$</td></tr>
<tr><td>Distribution
Discrete: PMF
Continuous: PDF</td><td>Maps each value in the support to a real number indicating its probability</td><td>$P(X = 1) = 0.5$
$P(X = 0) = 0.5$
$P(X)$ is a table<table><tr><th>X</th><th>P(X)</th></tr><tr><td>0</td><td>0.5</td></tr><tr><td>1</td><td>0.5</td></tr></table></td><td>$p(x) = \mathbf{1}_{[0,1]}(x) = \begin{cases} 1 \text{ if } x \in [0,1] \\ 0 \text{ o.w.} \end{cases}$</td></tr>
<tr><td>Expectation
$E[X]$</td><td>First moment of the random variable, "mean"</td><td>$E[X] = \sum_{x \in X} x P(x)$
$= 0.5$</td><td></td></tr>
</table>

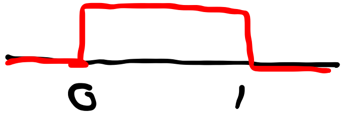
Notation: Continuous Random Variables

<table>
<tr><th>Term</th><th>Definition</th><th>$\text{Bernoulli}(0.5)$
Coinflip Example</th><th>$\mathcal{U}([0,1])$
Uniform Example</th></tr>
<tr><td>support(X)
$x \in X$</td><td>All the values that X can take</td><td>$\{h, t\}$ or $\{0, 1\}$</td><td>$[0, 1]$</td></tr>
<tr><td>Distribution
- Discrete: PMF
- Continuous: PDF</td><td>Maps each value in the support to a real number indicating its probability</td><td>$P(X = 1) = 0.5$
$P(X = 0) = 0.5$
$P(X)$ is a table
<table>
<tr><th>X</th><th>P(X)</th></tr>
<tr><td>0</td><td>0.5</td></tr>
<tr><td>1</td><td>0.5</td></tr>
</table>
</td><td>$p(x) = \mathbf{1}_{[0,1]}(x) = \begin{cases} 1 \text{ if } x \in [0,1] \\ 0 \text{ o.w.} \end{cases}$
 </td></tr>
<tr><td>Expectation
$E[X]$</td><td>First moment of the random variable, "mean"</td><td>$E[X] = \sum_{x \in X} x P(x)$
$= 0.5$</td><td></td></tr>
</table>

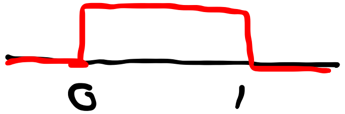
Notation: Continuous Random Variables

<table>
<tr><th>Term</th><th>Definition</th><th>$\text{Bernoulli}(0.5)$
Coinflip Example</th><th>$\mathcal{U}([0,1])$
Uniform Example</th></tr>
<tr><td>support(X)
$x \in X$</td><td>All the values that X can take</td><td>$\{h, t\}$ or $\{0, 1\}$</td><td>$[0, 1]$</td></tr>
<tr><td>Distribution
Discrete: PMF
Continuous: PDF</td><td>Maps each value in the support to a real number indicating its probability</td><td>$P(X=1)=0.5$
$P(X=0)=0.5$
$P(X)$ is a table<table><tr><th>X</th><th>P(X)</th></tr><tr><td>0</td><td>0.5</td></tr><tr><td>1</td><td>0.5</td></tr></table></td><td>$p(x) = \mathbf{1}_{[0,1]}(x) = \begin{cases} 1 \text{ if } x \in [0,1] \\ 0 \text{ o.w.} \end{cases}$

$P(X \in [a,b]) = \int_a^b p(x)dx$</td></tr>
<tr><td>Expectation
$E[X]$</td><td>First moment of the random variable, "mean"</td><td>$$E[X] = \sum_{x \in X} xP(x)$$ $$= 0.5$$</td><td></td></tr>
</table>

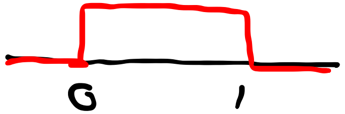
Notation: Continuous Random Variables

<table>
<tr><th>Term</th><th>Definition</th><th>$\text{Bernoulli}(0.5)$
Coinflip Example</th><th>$\mathcal{U}([0,1])$
Uniform Example</th></tr>
<tr><td>support(X)
$x \in X$</td><td>All the values that X can take</td><td>$\{h, t\}$ or $\{0, 1\}$</td><td>$[0, 1]$</td></tr>
<tr><td>Distribution
Discrete: PMF
Continuous: PDF</td><td>Maps each value in the support to a real number indicating its probability</td><td>$P(X = 1) = 0.5$
$P(X = 0) = 0.5$
$P(X)$ is a table<table><tr><th>X</th><th>P(X)</th></tr><tr><td>0</td><td>0.5</td></tr><tr><td>1</td><td>0.5</td></tr></table></td><td>$p(x) = \mathbf{1}_{[0,1]}(x) = \begin{cases} 1 \text{ if } x \in [0,1] \\ 0 \text{ o.w.} \end{cases}$

$P(X \in [a, b]) = \int_a^b p(x)dx$
$P(X = 0.5) =$</td></tr>
<tr><td>Expectation
$E[X]$</td><td>First moment of the random variable, "mean"</td><td>$E[X] = \sum_{x \in X} xP(x)$
$= 0.5$</td><td></td></tr>
</table>

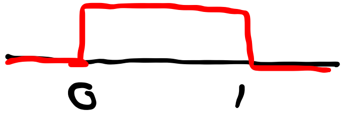
Notation: Continuous Random Variables

<table>
<tr><th></th><th></th><th>$\text{Bernoulli}(0.5)$</th><th>$\mathcal{U}([0,1])$</th></tr>
<tr><th>Term</th><th>Definition</th><th>Coinflip Example</th><th>Uniform Example</th></tr>
<tr><td>support(X)
$x \in X$</td><td>All the values that X can take</td><td>$\{h, t\}$ or $\{0, 1\}$</td><td>$[0, 1]$</td></tr>
<tr><td>Distribution
Discrete: PMF
Continuous: PDF</td><td>Maps each value in the support to a real number indicating its probability</td><td>$P(X = 1) = 0.5$
$P(X = 0) = 0.5$
$P(X)$ is a table<table><tr><th>X</th><th>P(X)</th></tr><tr><td>0</td><td>0.5</td></tr><tr><td>1</td><td>0.5</td></tr></table></td><td>$p(x) = \mathbf{1}_{[0,1]}(x) = \begin{cases} 1 \text{ if } x \in [0,1] \\ 0 \text{ o.w.} \end{cases}$

$P(X \in [a,b]) = \int_a^b p(x)dx$
$P(X = 0.5) = 0$</td></tr>
<tr><td>Expectation
$E[X]$</td><td>First moment of the random variable, "mean"</td><td>$E[X] = \sum_{x \in X} xP(x)$
$= 0.5$</td><td></td></tr>
</table>

Notation: Continuous Random Variables

<table>
<tr><th>Term</th><th>Definition</th><th>$\text{Bernoulli}(0.5)$
Coinflip Example</th><th>$\mathcal{U}([0,1])$
Uniform Example</th></tr>
<tr><td>support(X)
$x \in X$</td><td>All the values that X can take</td><td>$\{h, t\}$ or $\{0, 1\}$</td><td>$[0, 1]$</td></tr>
<tr><td>Distribution
- Discrete: PMF
- Continuous: PDF</td><td>Maps each value in the support to a real number indicating its probability</td><td>$P(X = 1) = 0.5$
$P(X = 0) = 0.5$
$P(X)$ is a table
<table><tr><th>X</th><th>P(X)</th></tr><tr><td>0</td><td>0.5</td></tr><tr><td>1</td><td>0.5</td></tr></table></td><td>$p(x) = \mathbf{1}_{[0,1]}(x) = \begin{cases} 1 \text{ if } x \in [0,1] \\ 0 \text{ o.w.} \end{cases}$

$P(X \in [a, b]) = \int_a^b p(x)dx$
$P(X = 0.5) = 0$</td></tr>
<tr><td>Expectation
$E[X]$</td><td>First moment of the random variable, "mean"</td><td>$E[X] = \sum_{x \in X} xP(x)$
$= 0.5$</td><td>$E[X] = \int_{x \in X} x\, p(x)\, dx$</td></tr>
</table>

Notation: Continuous Random Variables

<table>
<tr><th>Term</th><th>Definition</th><th>$\text{Bernoulli}(0.5)$
Coinflip Example</th><th>$\mathcal{U}([0,1])$
Uniform Example</th></tr>
<tr><td>support(X)
$x \in X$</td><td>All the values that X can take</td><td>$\{h, t\}$ or $\{0, 1\}$</td><td>$[0, 1]$</td></tr>
<tr><td>Distribution
Discrete: PMF
Continuous: PDF</td><td>Maps each value in the support to a real number indicating its probability</td><td>$P(X = 1) = 0.5$
$P(X = 0) = 0.5$
$P(X)$ is a table<table><tr><th>X</th><th>P(X)</th></tr><tr><td>0</td><td>0.5</td></tr><tr><td>1</td><td>0.5</td></tr></table></td><td>$p(x) = \mathbf{1}_{[0,1]}(x) = \begin{cases} 1 \text{ if } x \in [0,1] \\ 0 \text{ o.w.} \end{cases}$

$P(X \in [a,b]) = \int_a^b p(x)dx$
$P(X = 0.5) = 0$</td></tr>
<tr><td>Expectation
$E[X]$</td><td>First moment of the random variable, "mean"</td><td>$E[X] = \sum_{x \in X} xP(x)$
$= 0.5$</td><td>$E[X] = \int_{x \in X} x\, p(x)\, dx$
$= 0.5$</td></tr>
</table>

Rules for Continuous RVs

Discrete

1) a) $0 \leq P(X \mid Y) \leq 1$

   b) $\sum_{x \in X} P(x \mid Y) = 1$

2) $P(X) = \sum_{y \in Y} P(X, y)$

3) $P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$

   $P(X, Y) = P(X \mid Y)\, P(Y)$

Continuous

1)

Rules for Continuous RVs

Discrete

1) a) $0 \le P(X \mid Y) \le 1$

   b) $\sum_{x \in X} P(x \mid Y) = 1$

2) $P(X) = \sum_{y \in Y} P(X, y)$

3) $P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$

$P(X, Y) = P(X \mid Y)\, P(Y)$

Continuous

1) $0 \le p(X \mid Y)$

Rules for Continuous RVs

Discrete

1) a) $0 \leq P(X \mid Y) \leq 1$

   b) $\sum_{x \in X} P(x \mid Y) = 1$

2) $P(X) = \sum_{y \in Y} P(X, y)$

3) $P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$

$P(X, Y) = P(X \mid Y)\, P(Y)$

Continuous

1) $0 \leq p(X \mid Y)$

$$\int_X p(x|Y)\, dx = 1$$

Rules for Continuous RVs

Discrete

1) a) $0 \leq P(X \mid Y) \leq 1$

   b) $\sum_{x \in X} P(x \mid Y) = 1$

2) $P(X) = \sum_{y \in Y} P(X, y)$

3) $P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$

$$P(X, Y) = P(X \mid Y)\, P(Y)$$

Continuous

1) $0 \leq p(X \mid Y)$

$$\int_X p(x|Y)\, dx = 1$$

2)

$$p(X) = \int_Y p(X, y) dy$$

Rules for Continuous RVs

Discrete

1) a) $0 \leq P(X \mid Y) \leq 1$

   b) $\sum_{x \in X} P(x \mid Y) = 1$

2) $P(X) = \sum_{y \in Y} P(X, y)$

3) $P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$

$$P(X, Y) = P(X \mid Y)\, P(Y)$$

Continuous

1) $0 \leq p(X \mid Y)$

$$\int_X p(x|Y)\, dx = 1$$

2)

$$p(X) = \int_Y p(X, y) dy$$

3) $p(X \mid Y) = \frac{p(X,Y)}{p(Y)}$

$$p(X, Y) = p(X \mid Y)\, p(Y)$$

Multivariate Gaussian Distribution

Joint Distribution

Conditional Distribution

Marginal Distribution

Continuous S and A

Continuous S and A

e.g. $S \subseteq \mathbb{R}^n$, $A \subseteq \mathbb{R}^m$

Continuous S and A

e.g. $S \subseteq \mathbb{R}^n$, $A \subseteq \mathbb{R}^m$

The old rules still work!

1. Linear Dynamics, Quadratic Reward

1. Linear Dynamics, Quadratic Reward

$S = \mathbb{R}^n$, $A = \mathbb{R}^m$

1. Linear Dynamics, Quadratic Reward

$S = \mathbb{R}^n$, $A = \mathbb{R}^m$

$T(s' \mid s, a) = \mathcal{N}(T_s s + T_a a, \Sigma)$

1. Linear Dynamics, Quadratic Reward

$S = \mathbb{R}^n$, $A = \mathbb{R}^m$ $\qquad T(s' \mid s, a) = \mathcal{N}(T_s s + T_a a, \Sigma)$

$$s' = T_s s + T_a a + w$$

1. Linear Dynamics, Quadratic Reward

$S = \mathbb{R}^n$, $A = \mathbb{R}^m$ $\qquad T(s' \mid s, a) = \mathcal{N}(T_s s + T_a a, \Sigma)$

$$s' = T_s s + T_a a + w \quad w \sim \mathcal{N}(0, \Sigma)$$

1. Linear Dynamics, Quadratic Reward

$S = \mathbb{R}^n$, $A = \mathbb{R}^m$ $\qquad T(s' \mid s, a) = \mathcal{N}(T_s s + T_a a, \Sigma)$

$s' = T_s s + T_a a + w \quad w \sim \mathcal{N}(0, \Sigma)$ (Also works with other zero-mean w.)

1. Linear Dynamics, Quadratic Reward

$S = \mathbb{R}^n$, $A = \mathbb{R}^m$ $\quad T(s' \mid s, a) = \mathcal{N}(T_s s + T_a a, \Sigma)$ $\quad R(s, a) = s^\top R_s s + a^\top R_a a$

$s' = T_s s + T_a a + w \quad w \sim \mathcal{N}(0, \Sigma)$ (Also works with other zero-mean w.)

1. Linear Dynamics, Quadratic Reward

$S = \mathbb{R}^n$, $A = \mathbb{R}^m$ $\quad T(s' \mid s, a) = \mathcal{N}(T_s s + T_a a, \Sigma)$ $\quad R(s, a) = s^\top R_s s + a^\top R_a a$

$s' = T_s s + T_a a + w \quad w \sim \mathcal{N}(0, \Sigma)$ (Also works with other zero-mean w.)

Finite Horizon:

1. Linear Dynamics, Quadratic Reward

$S = \mathbb{R}^n$, $A = \mathbb{R}^m$ $\quad T(s' \mid s, a) = \mathcal{N}(T_s s + T_a a, \Sigma)$ $\quad R(s, a) = s^\top R_s s + a^\top R_a a$

$s' = T_s s + T_a a + w \quad w \sim \mathcal{N}(0, \Sigma)$ (Also works with other zero-mean w.)

Finite Horizon: $$U_h^*(s) = \max_\pi \left[\sum_{t=0}^{h} R(s_t, a_t)\right]$$

1. Linear Dynamics, Quadratic Reward

$S = \mathbb{R}^n$, $A = \mathbb{R}^m$ $\qquad T(s' \mid s, a) = \mathcal{N}(T_s s + T_a a, \Sigma)$ $\qquad R(s, a) = s^\top R_s s + a^\top R_a a$

$s' = T_s s + T_a a + w \quad w \sim \mathcal{N}(0, \Sigma)$ (Also works with other zero-mean w.)

Finite Horizon: $U_h^*(s) = \max_\pi \left[\sum_{t=0}^{h} R(s_t, a_t) \right]$ $\qquad \pi_h^*$ is "optimal h-step policy"

1. Linear Dynamics, Quadratic Reward

$S = \mathbb{R}^n$, $A = \mathbb{R}^m$ $\quad T(s' \mid s, a) = \mathcal{N}(T_s s + T_a a, \Sigma)$ $\quad R(s, a) = s^\top R_s s + a^\top R_a a$

$s' = T_s s + T_a a + w \quad w \sim \mathcal{N}(0, \Sigma)$ (Also works with other zero-mean w.)

Finite Horizon: $U_h^*(s) = \max_\pi \left[\sum_{t=0}^{h} R(s_t, a_t) \right]$ π_h^* is "optimal h-step policy"

We will show that $U_h^*(s) = s^\top V_h s + q_h$ and $\pi_h^*(s) = -K_h s$

1. Linear Dynamics, Quadratic Reward

1. Linear Dynamics, Quadratic Reward

We will show that $U_h^*(s) = s^\top V_h s + q_h$ and $\pi_h^*(s) = -K_h s$

1. Linear Dynamics, Quadratic Reward

We will show that $U_h^*(s) = s^\top V_h s + q_h$ and $\pi_h^*(s) = -K_h s$ by induction.

1. Linear Dynamics, Quadratic Reward

We will show that $U_h^*(s) = s^\top V_h s + q_h$ and $\pi_h^*(s) = -K_h s$ by induction.

Base: $U_1^*(s) = \max_a \left(s^\top R_s s + a^\top R_a a\right) = s^\top R_s s$

1. Linear Dynamics, Quadratic Reward

We will show that $U_h^*(s) = s^\top V_h s + q_h$ and $\pi_h^*(s) = -K_h s$ by induction.

Base: $U_1^*(s) = \max_a \left(s^\top R_s s + a^\top R_a a\right) = s^\top R_s s$

Inductive step: show that if $U_t^* = s^\top V_t s + q_t$, then $U_{t+1}^* = s^\top V_{t+1} s + q_{t+1}$.

1. Linear Dynamics, Quadratic Reward

We will show that $U_h^*(s) = s^\top V_h s + q_h$ and $\pi_h^*(s) = -K_h s$ by induction.

Base: $U_1^*(s) = \max_a \left(s^\top R_s s + a^\top R_a a\right) = s^\top R_s s$

Inductive step: show that if $U_t^* = s^\top V_t s + q_t$, then $U_{t+1}^* = s^\top V_{t+1} s + q_{t+1}$.

$$U_{t+1}^*(s) = \max_a \left(R(s,a) + \gamma E\left[U_t^*(s)\right]\right)$$

1. Linear Dynamics, Quadratic Reward

We will show that $U_h^*(s) = s^\top V_h s + q_h$ and $\pi_h^*(s) = -K_h s$ by induction.

Base: $U_1^*(s) = \max_a \left(s^\top R_s s + a^\top R_a a\right) = s^\top R_s s$

Inductive step: show that if $U_t^* = s^\top V_t s + q_t$, then $U_{t+1}^* = s^\top V_{t+1} s + q_{t+1}$.

$$\begin{aligned} U_{t+1}^*(s) &= \max_a \left(R(s,a) + \gamma E\left[U_t^*(s)\right]\right) \\ &= \max_a \left(s^\top R_s s + a^\top R_a a + \int p(w) U_t(T_s s + T_a a + w)\, dw\right) \end{aligned}$$

1. Linear Dynamics, Quadratic Reward

We will show that $U_h^*(s) = s^\top V_h s + q_h$ and $\pi_h^*(s) = -K_h s$ by induction.

Base: $U_1^*(s) = \max_a \left(s^\top R_s s + a^\top R_a a\right) = s^\top R_s s$

Inductive step: show that if $U_t^* = s^\top V_t s + q_t$, then $U_{t+1}^* = s^\top V_{t+1} s + q_{t+1}$.

$$\begin{aligned} U_{t+1}^*(s) &= \max_a \left(R(s,a) + \gamma E\left[U_t^*(s)\right]\right) \\ &= \max_a \left(s^\top R_s s + a^\top R_a a + \int p(w) U_t(T_s s + T_a a + w)\, dw\right) \\ &= s^\top R_s s + \max_a \left(a^\top R_a a + \int p(w)(T_s s + T_a a + w)^\top V_t (T_s s + T_a a + w) + q_t\right) dw \end{aligned}$$

1. Linear Dynamics, Quadratic Reward

We will show that $U_h^*(s) = s^\top V_h s + q_h$ and $\pi_h^*(s) = -K_h s$ by induction.

Base: $U_1^*(s) = \max_a \left(s^\top R_s s + a^\top R_a a\right) = s^\top R_s s$

Inductive step: show that if $U_t^* = s^\top V_t s + q_t$, then $U_{t+1}^* = s^\top V_{t+1} s + q_{t+1}$.

$$\begin{aligned}
U_{t+1}^*(s) &= \max_a \left(R(s,a) + \gamma E\left[U_t^*(s)\right]\right) \\
&= \max_a \left(s^\top R_s s + a^\top R_a a + \int p(w) U_t(T_s s + T_a a + w)\, dw\right) \\
&= s^\top R_s s + \max_a \left(a^\top R_a a + \int p(w)(T_s s + T_a a + w)^\top V_t (T_s s + T_a a + w) + q_t\right) dw \\
&= s^\top R_s s + s^\top T_s^\top V_t T_s s + \max_a \left(a^\top R_a a + 2 s^\top T_s^\top V_t T_a a + a^\top T_a^\top V_t T_a a\right) + \int p(w) w^\top V_t w\, dw + q_t
\end{aligned}$$

1. Linear Dynamics, Quadratic Reward

We will show that $U_h^*(s) = s^\top V_h s + q_h$ and $\pi_h^*(s) = -K_h s$ by induction.

Base: $U_1^*(s) = \max_a \left(s^\top R_s s + a^\top R_a a\right) = s^\top R_s s$

Inductive step: show that if $U_t^* = s^\top V_t s + q_t$, then $U_{t+1}^* = s^\top V_{t+1} s + q_{t+1}$.

$$\begin{aligned} U_{t+1}^*(s) &= \max_a \left(R(s,a) + \gamma E\left[U_t^*(s)\right]\right) \\ &= \max_a \left(s^\top R_s s + a^\top R_a a + \int p(w) U_t(T_s s + T_a a + w)\, dw\right) \\ &= s^\top R_s s + \max_a \left(a^\top R_a a + \int p(w)(T_s s + T_a a + w)^\top V_t (T_s s + T_a a + w) + q_t\right) dw \\ &= s^\top R_s s + s^\top T_s^\top V_t T_s s + \max_a \left(a^\top R_a a + 2 s^\top T_s^\top V_t T_a a + a^\top T_a^\top V_t T_a a\right) + \int p(w) w^\top V_t w\, dw + q_t \end{aligned}$$

a^* is where $\nabla_a(\text{max term}) = 0$

$$0 = 2R_a a^* + 2T_a^\top V_t T_s s + 2T_a^\top V_t T_a a^*$$

$$a^* = -\underbrace{(R_a + T_a^\top V_t T_a)^{-1} T_a^\top V_t T_s}_{K_t} s$$

1. Linear Dynamics, Quadratic Reward

We will show that $U_h^*(s) = s^\top V_h s + q_h$ and $\pi_h^*(s) = -K_h s$ by induction.

Base: $U_1^*(s) = \max_a \left(s^\top R_s s + a^\top R_a a\right) = s^\top R_s s$

Inductive step: show that if $U_t^* = s^\top V_t s + q_t$, then $U_{t+1}^* = s^\top V_{t+1} s + q_{t+1}$.

$$\begin{aligned}
U_{t+1}^*(s) &= \max_a \left(R(s,a) + \gamma E\left[U_t^*(s)\right]\right) \\
&= \max_a \left(s^\top R_s s + a^\top R_a a + \textstyle\int p(w) U_t(T_s s + T_a a + w)\, dw\right) \\
&= s^\top R_s s + \max_a \left(a^\top R_a a + \textstyle\int p(w)(T_s s + T_a a + w)^\top V_t (T_s s + T_a a + w) + q_t\right) dw \\
&= s^\top R_s s + s^\top T_s^\top V_t T_s s + \max_a \left(a^\top R_a a + 2 s^\top T_s^\top V_t T_a a + a^\top T_a^\top V_t T_a a\right) + \textstyle\int p(w) w^\top V_t w\, dw + q_t
\end{aligned}$$

a^* is where $\nabla_a(\max \text{term}) = 0$

$$0 = 2R_a a^* + 2T_a^\top V_t T_s s + 2T_a^\top V_t T_a a^*$$

$$a^* = -\underbrace{(R_a + T_a^\top V_t T_a)^{-1} T_a^\top V_t T_s}_{K_t} s$$

$$U_{t+1}^*(s) = s^\top \underbrace{\left(R_s + T_s^\top V_t T_s - (T_a^\top V_t T_s)^\top (R_a + T_a^\top V_t T_a)^{-1} (T_a^\top V_t T_s)\right)}_{V_{t+1}} s + \underbrace{\int p(w) w^\top V_t w\, dw + q_t}_{q_{t+1}}$$

1. Linear Dynamics, Quadratic Reward

We will show that $U_h^*(s) = s^\top V_h s + q_h$ and $\pi_h^*(s) = -K_h s$ by induction.

Base: $U_1^*(s) = \max_a \left(s^\top R_s s + a^\top R_a a\right) = s^\top R_s s$

Inductive step: show that if $U_t^* = s^\top V_t s + q_t$, then $U_{t+1}^* = s^\top V_{t+1} s + q_{t+1}$.

$$\begin{aligned} U_{t+1}^*(s) &= \max_a \left(R(s,a) + \gamma E\left[U_t^*(s)\right]\right) \\ &= \max_a \left(s^\top R_s s + a^\top R_a a + \int p(w) U_t(T_s s + T_a a + w)\, dw\right) \\ &= s^\top R_s s + \max_a \left(a^\top R_a a + \int p(w)(T_s s + T_a a + w)^\top V_t (T_s s + T_a a + w) + q_t\right) dw \\ &= s^\top R_s s + s^\top T_s^\top V_t T_s s + \max_a \left(a^\top R_a a + 2 s^\top T_s^\top V_t T_a a + a^\top T_a^\top V_t T_a a\right) + \int p(w) w^\top V_t w\, dw + q_t \end{aligned}$$

a^* is where $\nabla_a(\text{max term}) = 0$

$$0 = 2R_a a^* + 2T_a^\top V_t T_s s + 2T_a^\top V_t T_a a^*$$

$$a^* = -\underbrace{(R_a + T_a^\top V_t T_a)^{-1} T_a^\top V_t T_s}_{K_t} s$$

$$U_{t+1}^*(s) = s^\top \underbrace{\left(R_s + T_s^\top V_t T_s - (T_a^\top V_t T_s)^\top (R_a + T_a^\top V_t T_a)^{-1} (T_a^\top V_t T_s)\right)}_{V_{t+1}} s + \underbrace{\int p(w) w^\top V_t w\, dw + q_t}_{q_{t+1}}$$

$$U_{t+1}^*(s) = s^\top V_{t+1} s + q_{t+1} \qquad \square$$

1. Linear Dynamics, Quadratic Reward

1. Linear Dynamics, Quadratic Reward

As $h \to \infty$

1. Linear Dynamics, Quadratic Reward

As $h \to \infty$

$$V_\infty = T_s^\top \left(V_\infty - V_\infty T_a \left(T_a^\top V_\infty T_a + R_a \right)^{-1} T_a^\top V_\infty \right) T_s + R_s$$

1. Linear Dynamics, Quadratic Reward

As $h \to \infty$

$$V_\infty = T_s^\top \left(V_\infty - V_\infty T_a \left(T_a^\top V_\infty T_a + R_a \right)^{-1} T_a^\top V_\infty \right) T_s + R_s$$

$$K_\infty = \left(T_a^\top V_\infty T_a + R_a \right)^{-1} T_a^\top V_\infty T_s$$

1. Linear Dynamics, Quadratic Reward

As $h \to \infty$

$$V_\infty = T_s^\top \left(V_\infty - V_\infty T_a \left(T_a^\top V_\infty T_a + R_a \right)^{-1} T_a^\top V_\infty \right) T_s + R_s$$

$$K_\infty = \left(T_a^\top V_\infty T_a + R_a \right)^{-1} T_a^\top V_\infty T_s \qquad \pi^*_\infty(s) = -K_\infty s$$

1. Linear Dynamics, Quadratic Reward

As $h \to \infty$

$$V_\infty = T_s^\top \left(V_\infty - V_\infty T_a \left(T_a^\top V_\infty T_a + R_a \right)^{-1} T_a^\top V_\infty \right) T_s + R_s$$

$$K_\infty = \left(T_a^\top V_\infty T_a + R_a \right)^{-1} T_a^\top V_\infty T_s$$

$$\pi_\infty^*(s) = -K_\infty s$$

(K_∞ has no dependence on Σ)

1. Linear Dynamics, Quadratic Reward

As $h \to \infty$

$$V_\infty = T_s^\top \left(V_\infty - V_\infty T_a \left(T_a^\top V_\infty T_a + R_a\right)^{-1} T_a^\top V_\infty\right) T_s + R_s$$

$$K_\infty = \left(T_a^\top V_\infty T_a + R_a\right)^{-1} T_a^\top V_\infty T_s$$

$$\pi^*_\infty(s) = -K_\infty s$$

(K_∞ has no dependence on Σ)

Certainty-Equivalence Principle: For Linear-Quadratic problems, the optimal policy with noise is the same as the optimal policy without noise!

1. Linear Dynamics, Quadratic Reward

As $h \to \infty$

$$V_\infty = T_s^\top \left(V_\infty - V_\infty T_a \left(T_a^\top V_\infty T_a + R_a\right)^{-1} T_a^\top V_\infty\right) T_s + R_s$$

$$K_\infty = \left(T_a^\top V_\infty T_a + R_a\right)^{-1} T_a^\top V_\infty T_s$$

$$\pi_\infty^*(s) = -K_\infty s$$

(K_∞ has no dependence on Σ)

Certainty-Equivalence Principle: For Linear-Quadratic problems, the optimal policy with noise is the same as the optimal policy without noise!

Practical Implication: If a continuous problem has roughly linear dynamics, a convex cost function, and roughly zero-mean additive noise, you can use *certainty-equivalent control*, i.e. control as if there is no noise.

2. Value Function Approximation

2. Value Function Approximation

$V_\theta(s) = f_\theta(s)$ (e.g. neural network)

2. Value Function Approximation

$V_\theta(s) = f_\theta(s)$ (e.g. neural network)

$V_\theta(s) = \theta^\top \beta(s)$ (linear feature)

2. Value Function Approximation

$V_\theta(s) = f_\theta(s)$ (e.g. neural network)

$V_\theta(s) = \theta^\top \beta(s)$ (linear feature)

Fitted Value Iteration

```
while not converged
    $\theta \leftarrow \theta'$
    $\hat{V}' \leftarrow B_{\text{approx}}[V_\theta]$
    $\theta' \leftarrow \text{fit}(\hat{V}')$
```

2. Value Function Approximation

$V_\theta(s) = f_\theta(s)$ (e.g. neural network)

$V_\theta(s) = \theta^\top \beta(s)$ (linear feature)

Fitted Value Iteration

```
while not converged
    $\theta \leftarrow \theta'$
    $\hat{V}' \leftarrow B_{\text{approx}}[V_\theta]$
    $\theta' \leftarrow \text{fit}(\hat{V}')$
```

$$B_{\text{MC}(N)}[V_\theta](s) = \max_a \left(R(s,a) + \gamma \sum_{i=1}^{N} V_\theta(G(s,a,w_i)) \right)$$

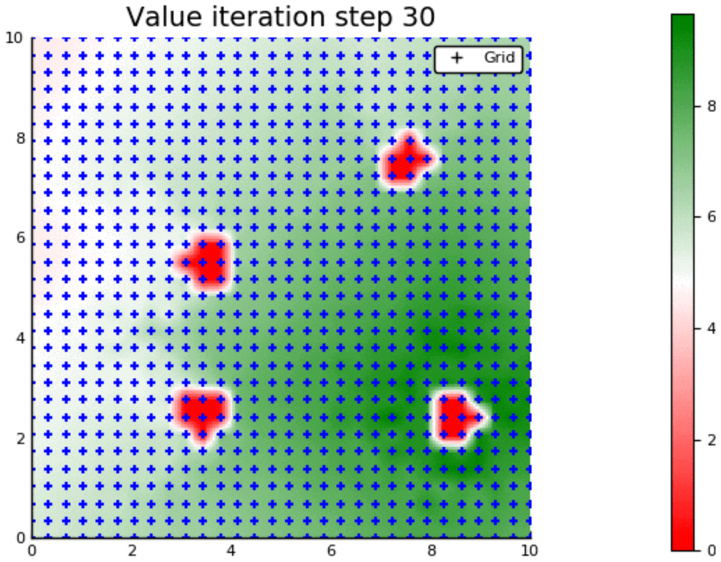
2. Value Function Approximation

$V_\theta(s) = f_\theta(s)$ (e.g. neural network)

$V_\theta(s) = \theta^\top \beta(s)$ (linear feature)

Fitted Value Iteration

```
while not converged
    $\theta \leftarrow \theta'$
    $\hat{V}' \leftarrow B_{\text{approx}}[V_\theta]$
    $\theta' \leftarrow \text{fit}(\hat{V}')$
```


Value iteration step 30

$$B_{\text{MC}(N)}[V_\theta](s) = \max_a \left(R(s,a) + \gamma \sum_{i=1}^{N} V_\theta(G(s,a,w_i)) \right)$$

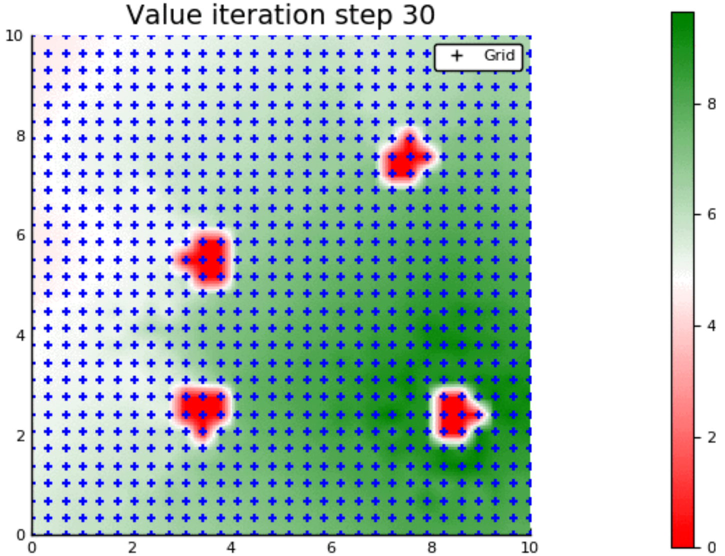
2. Value Function Approximation

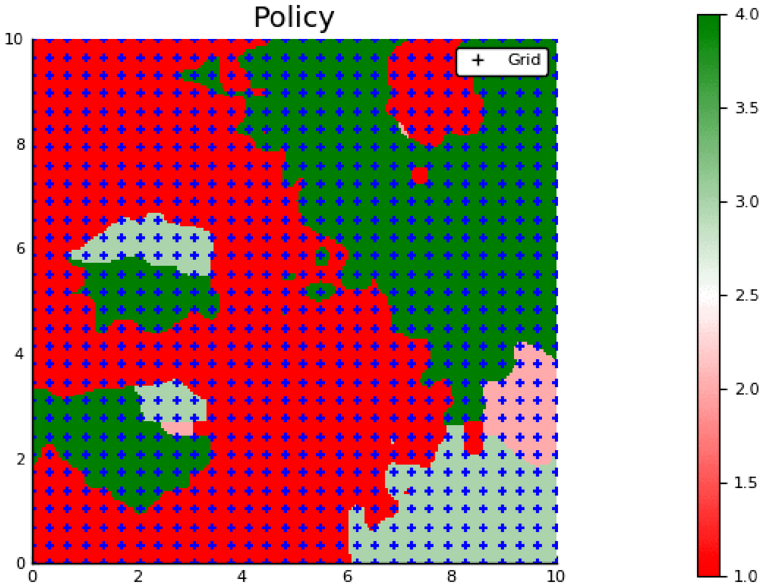
$V_\theta(s) = f_\theta(s)$ (e.g. neural network)

$V_\theta(s) = \theta^\top \beta(s)$ (linear feature)

Fitted Value Iteration

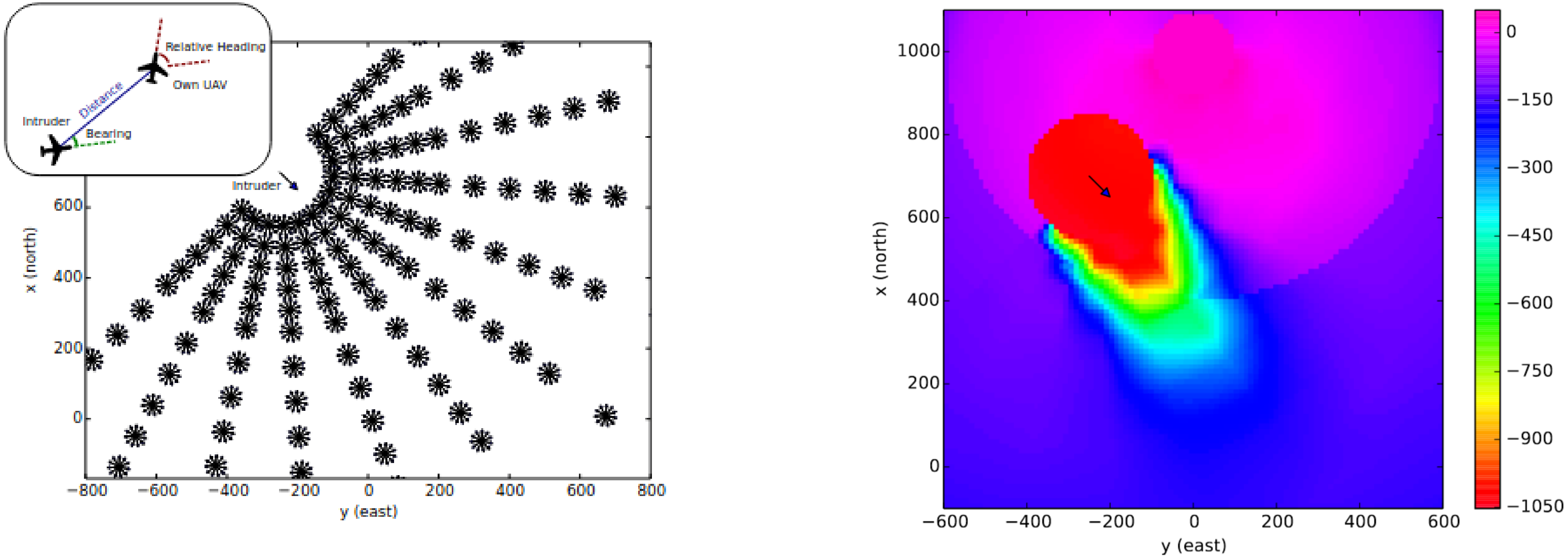
```
while not converged
    $\theta \leftarrow \theta'$
    $\hat{V}' \leftarrow B_{\text{approx}}[V_\theta]$
    $\theta' \leftarrow \text{fit}(\hat{V}')$
```


Value iteration step 30


Policy

$$B_{\text{MC}(N)}[V_\theta](s) = \max_a \left(R(s,a) + \gamma \sum_{i=1}^{N} V_\theta(G(s,a,w_i)) \right)$$

2. Value Function Approximation

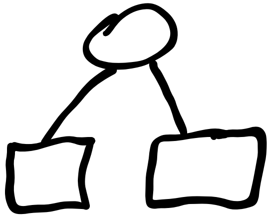


Break

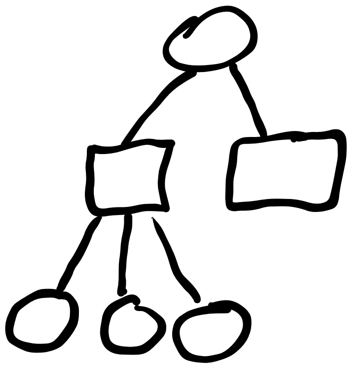
What will a Monte Carlo Tree Search tree look like if run on a problem with continuous spaces?

3. Sparse Tree Search/Progressive Widening

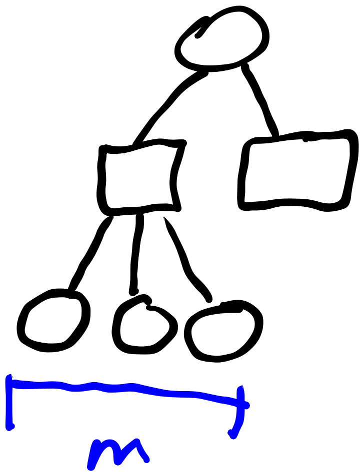
3. Sparse Tree Search/Progressive Widening



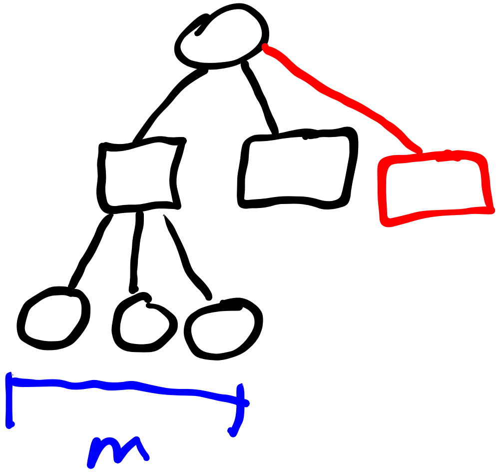
3. Sparse Tree Search/Progressive Widening



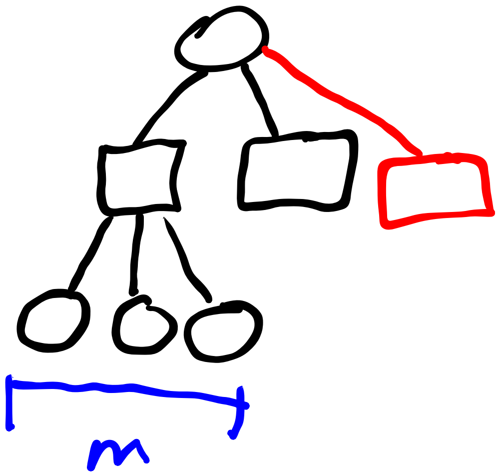
3. Sparse Tree Search/Progressive Widening



3. Sparse Tree Search/Progressive Widening

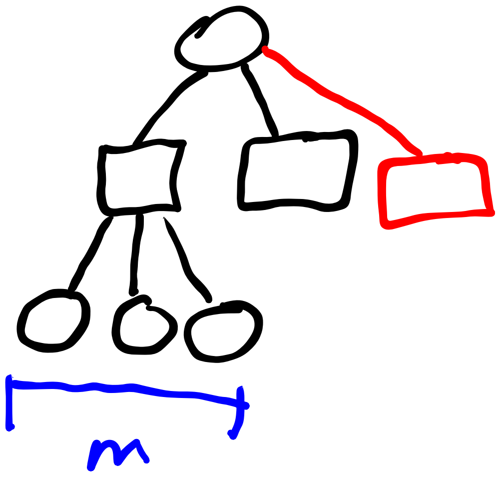


3. Sparse Tree Search/Progressive Widening

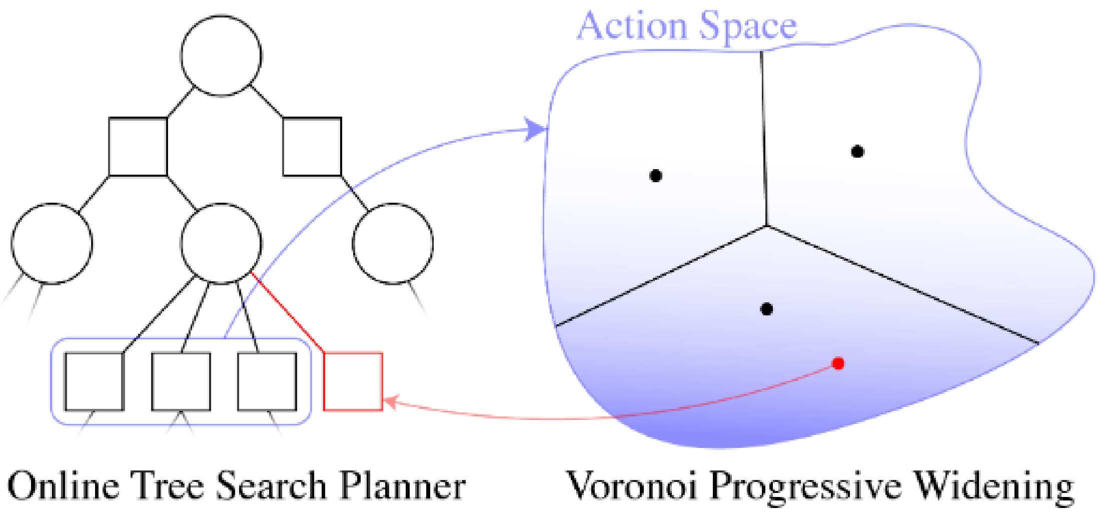


add new branch if $C < kN^\alpha \quad (\alpha < 1)$

3. Sparse Tree Search/Progressive Widening



add new branch if $C < kN^\alpha$ $(\alpha < 1)$



4. Model Predictive Control

(Use off-the-shelf optimization software, e.g. Ipopt)

4. Model Predictive Control

(Use off-the-shelf optimization software, e.g. Ipopt)

Certainty-
Equivalent

$$\begin{aligned} \underset{a_{1:d}, s_{1:d}}{\text{maximize}} \quad & \sum_{t=1}^{d} \gamma^t R(s_t, a_t) \\ \text{subject to} \quad & s_{t+1} = \mathrm{E}_{s' \sim T(s' \mid s, a)}[s'] \quad \forall t \end{aligned}$$

4. Model Predictive Control

(Use off-the-shelf optimization software, e.g. Ipopt)

Certainty-Equivalent

$$\begin{aligned} \underset{a_{1:d}, s_{1:d}}{\text{maximize}} \quad & \sum_{t=1}^{d} \gamma^t R(s_t, a_t) \\ \text{subject to} \quad & s_{t+1} = \mathrm{E}_{s' \sim T(s' \mid s, a)}[s'] \quad \forall t \end{aligned}$$

Open-Loop

$$\begin{aligned} \underset{a_{1:d}, s_{1:d}^{(1:m)}}{\text{maximize}} \quad & \frac{1}{m} \sum_{i=1}^{m} \sum_{t=1}^{d} \gamma^t R(s_t^{(i)}, a_t) \\ \text{subject to} \quad & s_{t+1} = G(s_t^{(i)}, a_t, w_t^{(i)}) \quad \forall t, i \end{aligned}$$

4. Model Predictive Control

(Use off-the-shelf optimization software, e.g. Ipopt)

Certainty-Equivalent

$$\begin{aligned} \underset{a_{1:d}, s_{1:d}}{\text{maximize}} \quad & \sum_{t=1}^{d} \gamma^t R(s_t, a_t) \\ \text{subject to} \quad & s_{t+1} = \mathrm{E}_{s' \sim T(s' \mid s, a)}[s'] \quad \forall t \end{aligned}$$

Open-Loop

$$\begin{aligned} \underset{a_{1:d}, s_{1:d}^{(1:m)}}{\text{maximize}} \quad & \frac{1}{m} \sum_{i=1}^{m} \sum_{t=1}^{d} \gamma^t R(s_t^{(i)}, a_t) \\ \text{subject to} \quad & s_{t+1} = G(s_t^{(i)}, a_t, w_t^{(i)}) \quad \forall t, i \end{aligned}$$

Hindsight Optimization

$$\begin{aligned} \underset{a_{1:d}^{(1:m)}, s_{1:d}^{(1:m)}}{\text{maximize}} \quad & \frac{1}{m} \sum_{i=1}^{m} \sum_{t=1}^{d} \gamma^t R(s_t^{(i)}, a_t^{(i)}) \\ \text{subject to} \quad & s_{t+1} = G(s_t^{(i)}, a_t^{(i)}, w_t^{(i)}) \quad \forall t, i \\ & a_1^{(i)} = a_1^{(j)} \quad \forall i, j \end{aligned}$$

Guiding Questions

- What tools do we have to solve MDPs with continuous S and A?