20th International School on Trustworthy Software
Shanghai, China, July 9 - 12

Trustworthy AI through Neural Certificates, Runtime Monitoring and Multi-Agent Reasoning

Thomas A. Henzinger
Institute of Science and Technology Austria

Djordje Zikelic
Singapore Management University

Lecture 2 and Lecture 3

Neural certificates for
stochastic control with guarantees

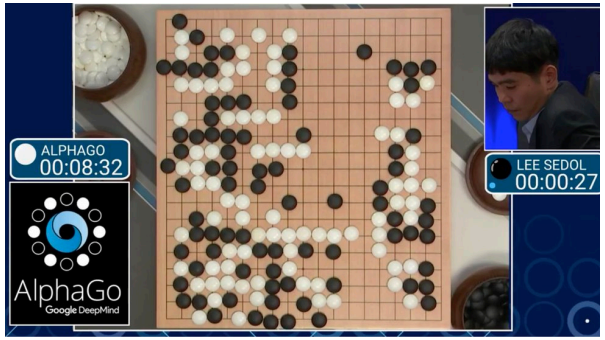
Lecture 2

Theory of certificates
for stochastic control

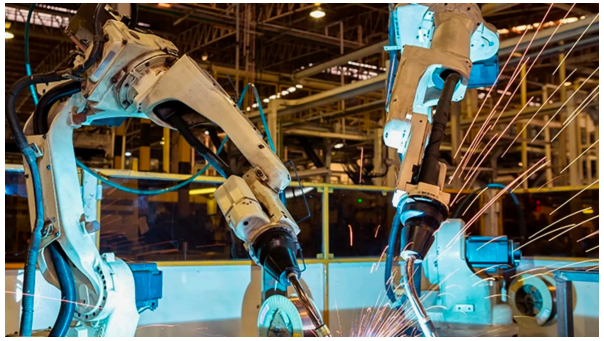
Lecture 3

Learner-verifier
automation

Learning-enabled Controller Synthesis



Not so safety-critical



Safety-critical

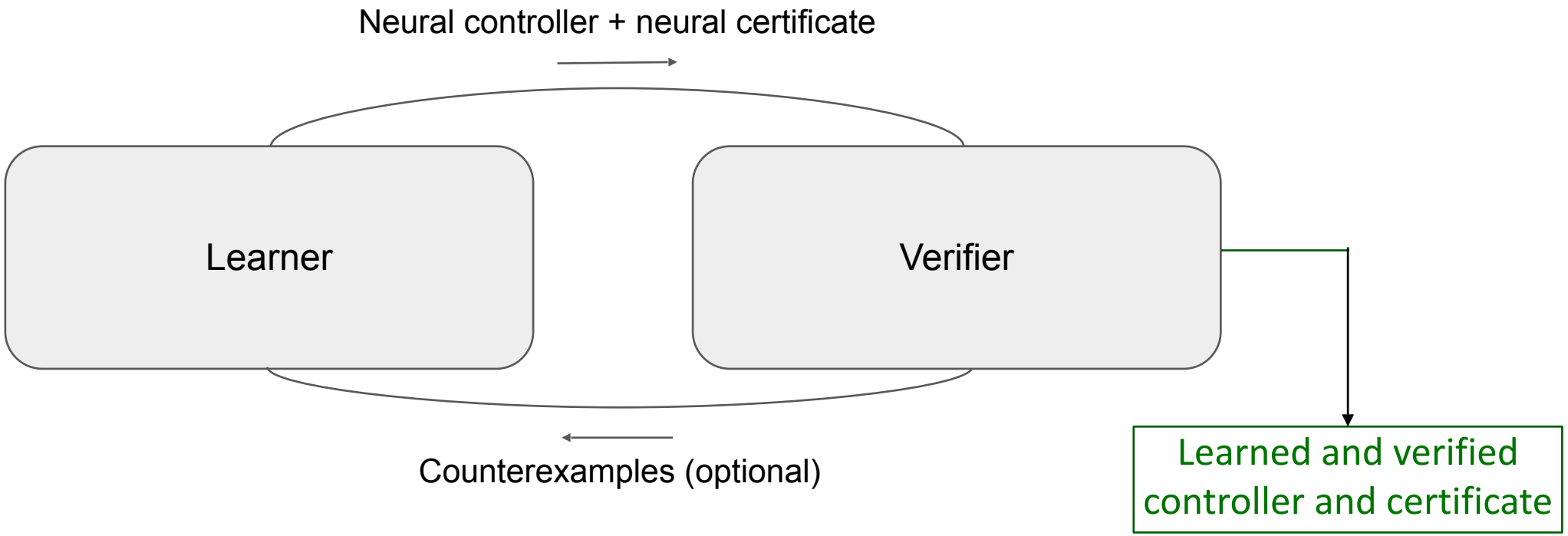
Safety-critical applications require **correctness guarantees**

Machine learning + Formal methods

=

Formally verified learned controllers with <u>certificates</u>

Neural certificates for learning-enabled control



Idea: Learn controllers together with **certificates** of their correctness

Sounds nice, but this is easier said than done…

1. What are the certificates for stochastic control?

2. How to learn neural parametrisations of certificates?

3. How to verify neural parametrisations of certificates?

Sounds nice, but this is easier said than done…

1. What are the certificates for stochastic control? **(Lecture 2)**

2. How to learn neural parametrisations of certificates? **(Lecture 3)**

3. How to verify neural parametrisations of certificates? **(Lecture 3)**

Plan for today

1. Stochastic control problem definition

2. Motivation for martingale-based certificates

3. Crash course on martingale theory

4. Martingale-based certificates

Plan for today

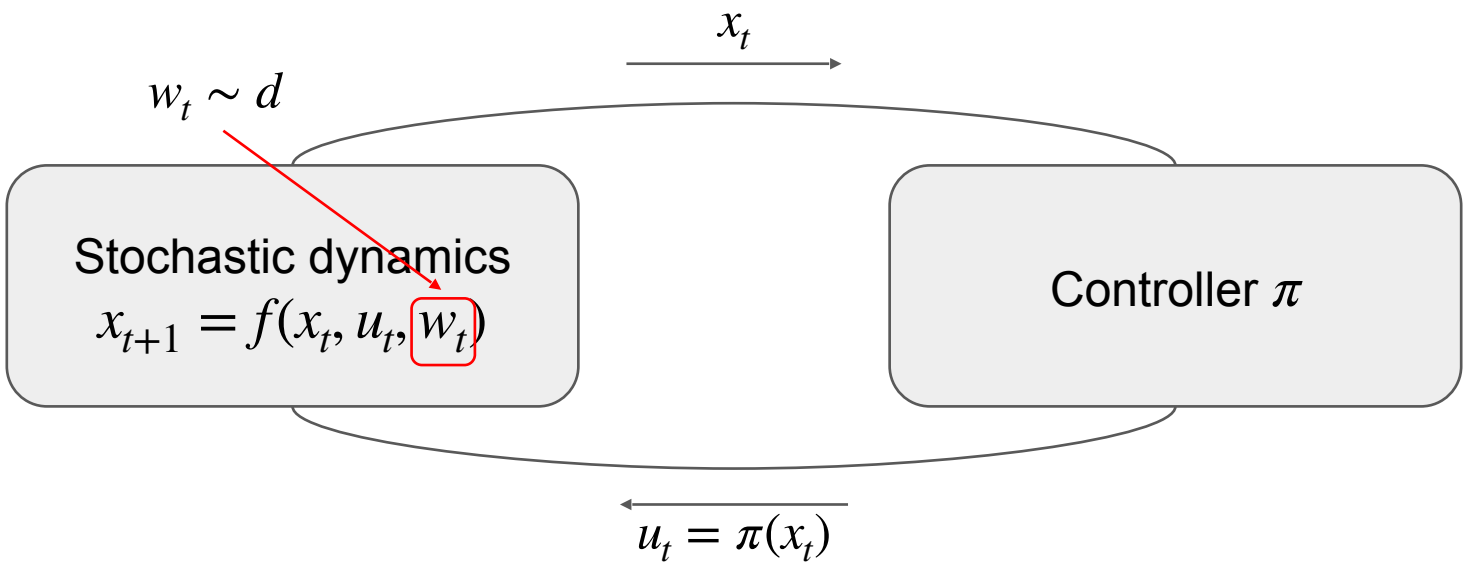
1. **Stochastic control problem definition**

2. Motivation for martingale-based certificates

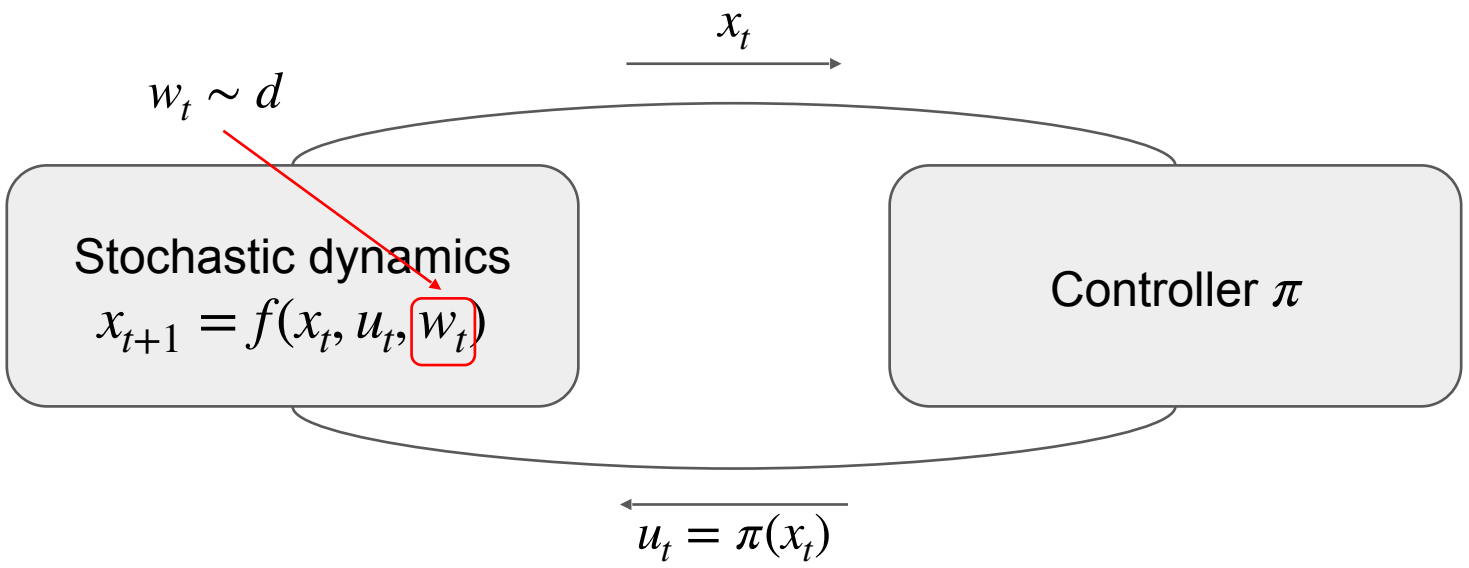
3. Crash course on martingale theory

4. Martingale-based certificates

Model: Stochastic dynamical system (SDS)



Model: Stochastic dynamical system (SDS)

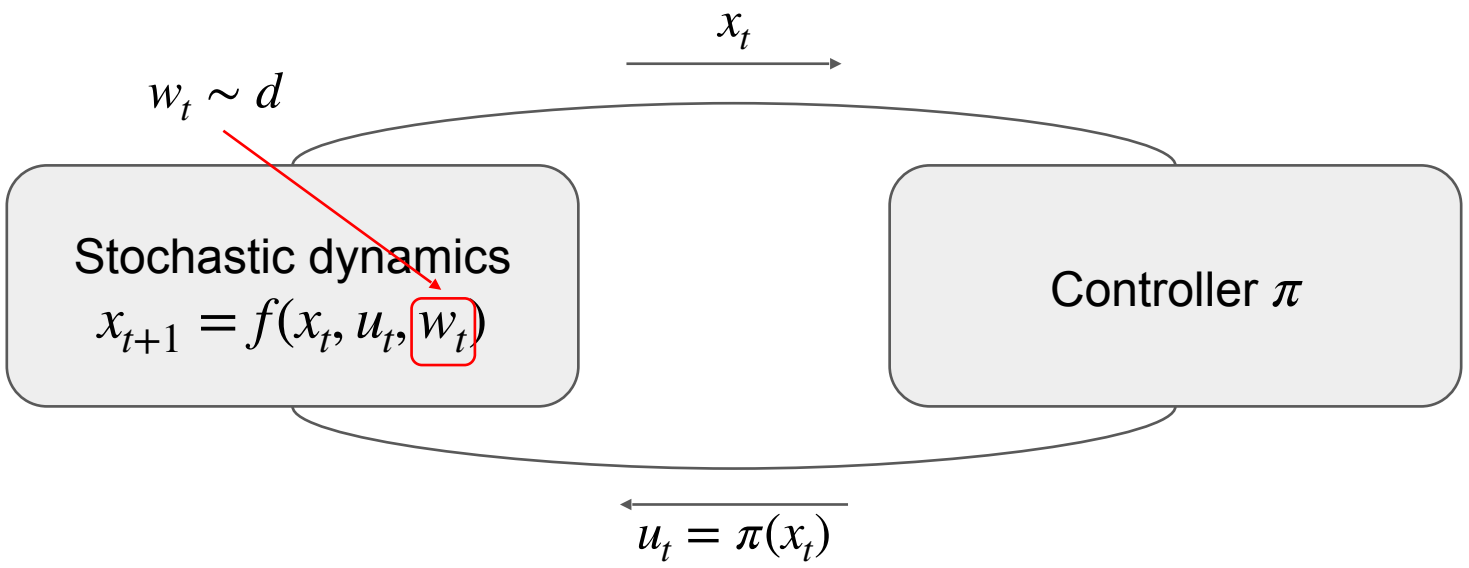


Definition: Stochastic dynamical system

A stochastic dynamical system (SDS) is a tuple (X, U, W, f, d), where

- $X \subseteq \mathbb{R}^n$ is the state space of the system,
- $U \subseteq \mathbb{R}^m$ is the control input space,
- $W \subseteq \mathbb{R}^k$ is the stochastic disturbance space,
- $f : X \times U \times W \to X$ is a continuous dynamics function,
- $d \in \text{dist}(W)$ is a stochastic disturbance distribution.

Model: Stochastic dynamical system (SDS)

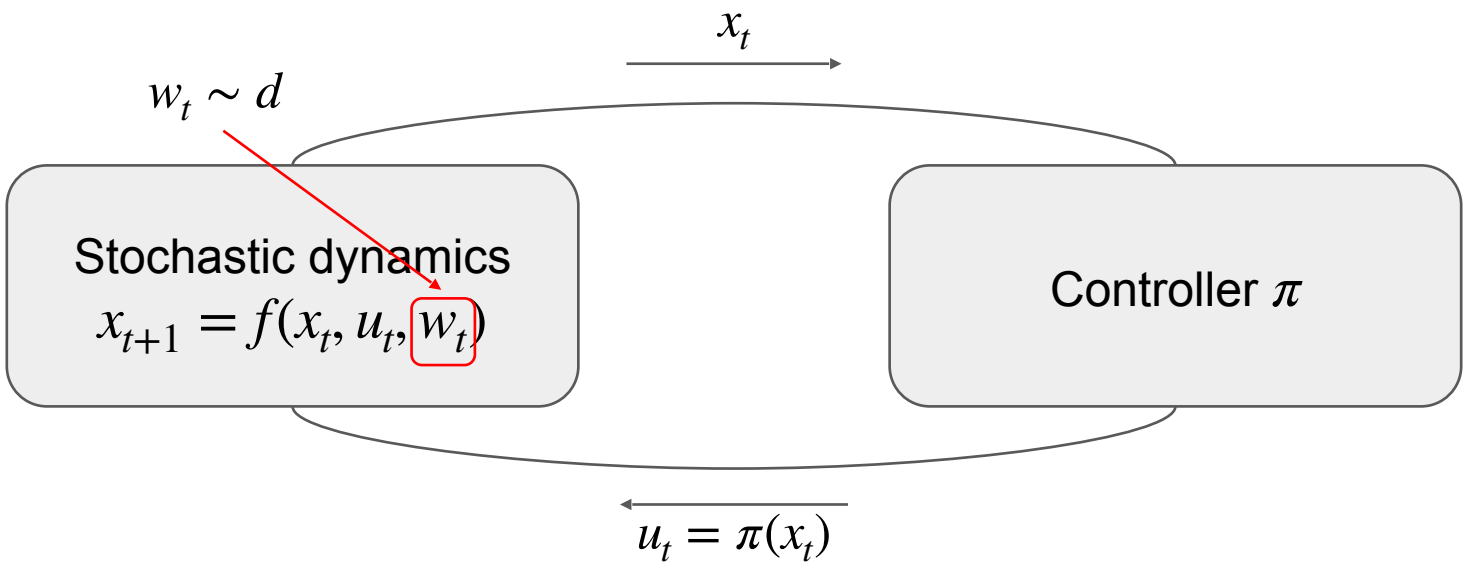


Definition: Controller

Given an SDS (X, U, W, f, d), a controller (or a policy) is a function $\pi : X \to U$ which for each system state $x \in X$ specifies a control input $\pi(x) \in U$.

(<u>Note</u>: We are restricting ourselves to pure memoryless controllers. These are sufficient for all specifications that we will consider.)

Model: Stochastic dynamical system (SDS)

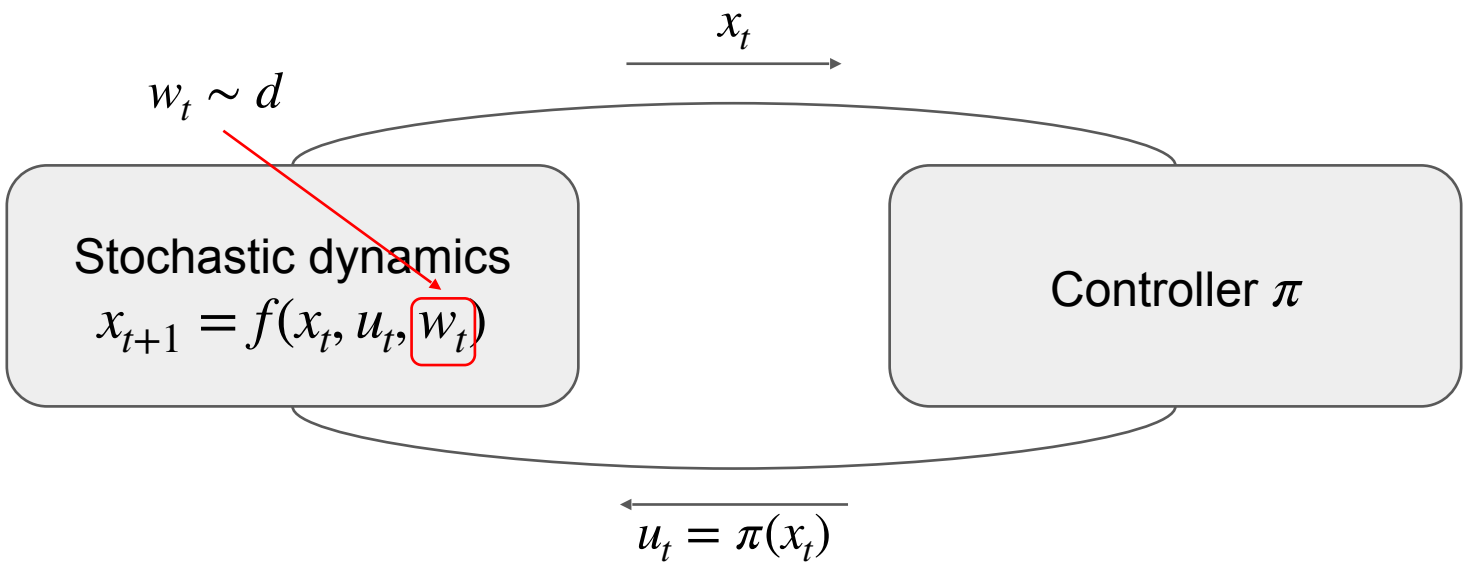


For each initial state $x_0 \in X$, an SDS and a controller together induce a **random system trajectory**

$$(x_0, \pi(x_0), w_0 \sim d) \xrightarrow{f} (x_1, \pi(x_1), w_1 \sim d) \xrightarrow{f} (x_2, \pi(x_2), w_2 \sim d) \xrightarrow{f} (x_3, \pi(x_3), w_3 \sim d) \xrightarrow{f} \ldots$$

We write $\rho = (x_t, \pi(x_t), w_t)_{t=0}^{\infty}$

Model: Stochastic dynamical system (SDS)

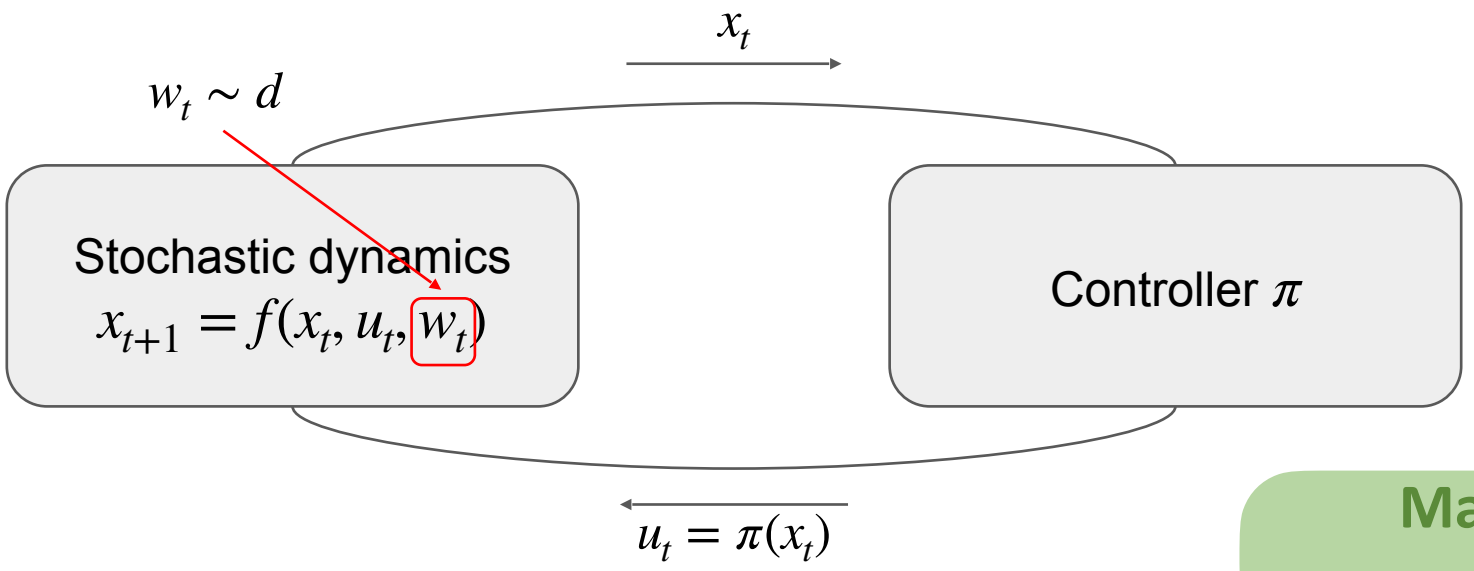


Semantics of SDS

For every initial state $x_0 \in X$, the SDS (X, U, W, f, d) and the controller $\pi : X \to U$ give rise to a **probability distribution** over the set of all random system trajectories with initial system state x_0.

We use $\mathbb{P}_{x_0}$ to denote the **probability measure** of this probability distribution.

Model: Stochastic dynamical system (SDS)



Markov chain semantics

Semantics of SDS

For every initial state $x_0 \in X$, the SDS (X, U, W, f, d) and the controller $\pi : X \to U$ give rise to a **probability distribution** over the set of all random system trajectories with initial system state x_0.

We use $\mathbb{P}_{x_0}$ to denote the **probability measure** of this probability distribution.

Semantics without a controller

Q: Consider an SDS (X, U, W, f, d) in which the controller is not yet fixed. What Markov model can be used to formalise the semantics of SDS?

Semantics without a controller

Q: Consider an SDS (X, U, W, f, d) in which the controller is not yet fixed. What Markov model can be used to formalise the semantics of SDS?

A: Markov decision process (MDP)

Specifications: Reachability, safety, reach-avoidance

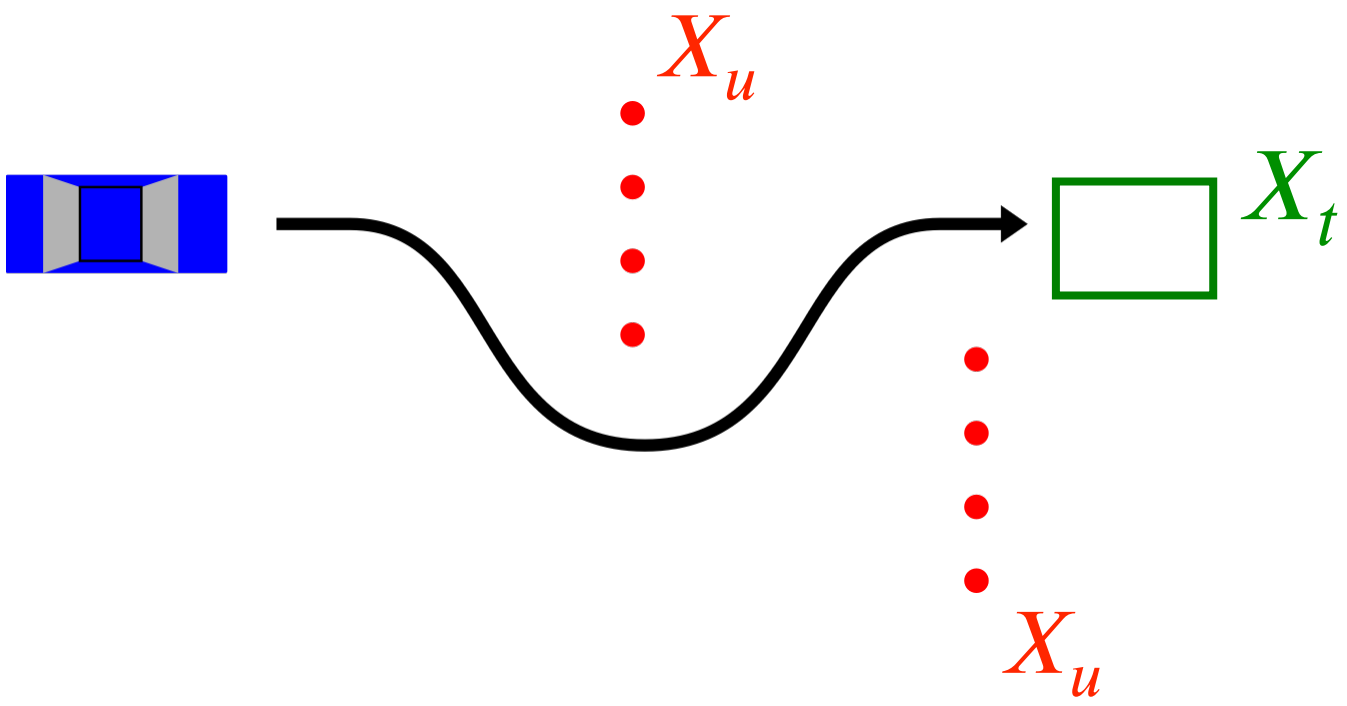
Definition: Reachability, safety, reach-avoidance

Reachability. Given a set $X_T \subseteq X$ of target states in an SDS, a system trajectory $\rho = (x_t, \pi(x_t), w_t)_{t=0}^{\infty}$ satisfies the reachability specification if $x_t \in X_T$ for some $t \geq 0$.

Safety. Given a set $X_U \subseteq X$ of unsafe states in an SDS, a system trajectory $\rho = (x_t, \pi(x_t), w_t)_{t=0}^{\infty}$ satisfies the safety specification if $x_t \notin X_U$ for all $t \geq 0$.

Reach-avoidance. Given a set $X_T \subseteq X$ of target states and a set $X_u \subseteq X$ of unsafe states in an SDS, a system trajectory $\rho = (x_t, \pi(x_t), w_t)_{t=0}^{\infty}$ satisfies the reach-avoidance specification there exists $t \geq 0$ such that $x_t \in X_T$ and $x_{t'} \notin X_U$ for all $t' \leq t - 1$.

Example: Car navigation



Reachability = reach the target set of states

Safety = do not reach the unsafe set of states

Reach-avoidance = reach the target set while avoiding the unsafe set of states

Stochastic control problem

Definition: Reachability, safety, reach-avoidance

Given: An SDS (X, U, W, f, d), initial region $X_0 \subseteq X$, target region $X_T \subseteq X$, unsafe region $X_U \subseteq X$, probability threshold $p \in [0,1]$

Task: Compute a controller $\pi : X \to U$ such that $\mathbb{P}_{x_0}[\text{ReachAvoid}(X_T, X_U)] \geq p$ for every $x_0 \in X_0$.

<u>Notation</u>: ReachAvoid(X_T, X_U) is the set of all trajectories that satisfy the reach-avoidance specification.

Plan for today

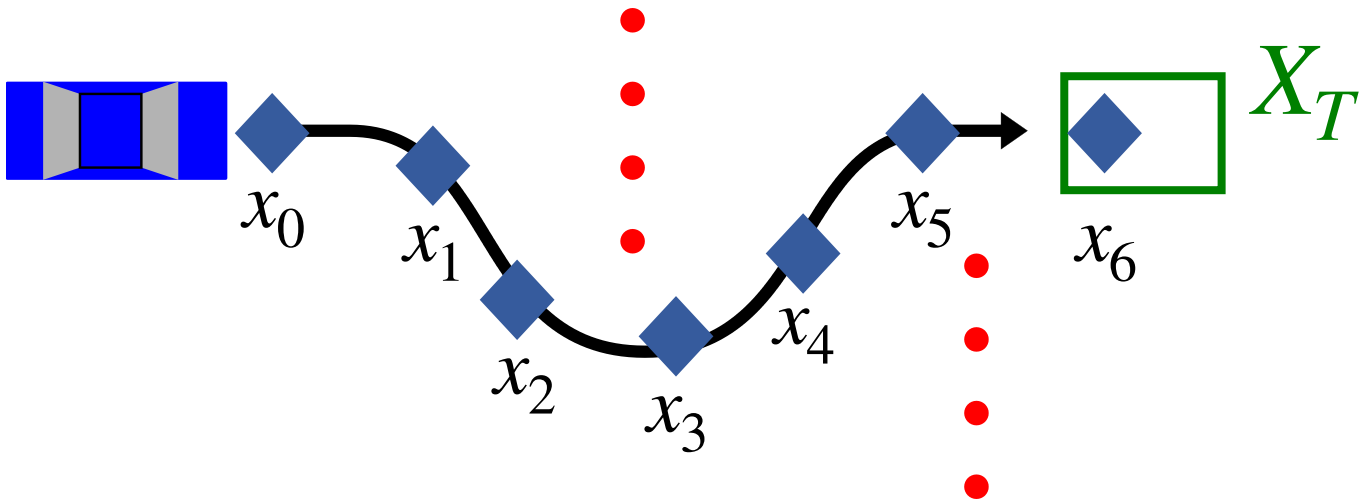
1. Stochastic control problem definition

2. **Motivation for martingale-based certificates**

3. Crash course on martingale theory

4. Martingale-based certificates

Motivation: Deterministic reachability



Q: How do we formally verify that, for every $x_0 \in X_0$, the trajectory eventually reaches X_T?

A: Formally verify that the agent is always making "progress" along each trajectory.

Motivation: Deterministic reachability

Can formally measure progress along trajectories by using **ranking functions**.

Intuitively, in deterministic dynamical systems (so no stochastic disturbances):

Ranking function

A ranking function for the target region $X_T \subseteq X$ under controller π is a **function** $V : X \to \mathbb{R}$ that assigns a real value to each system state, and which is required to be

(1) *Non-negative*. $V(x) \geq 0$ for each $x \in X$.

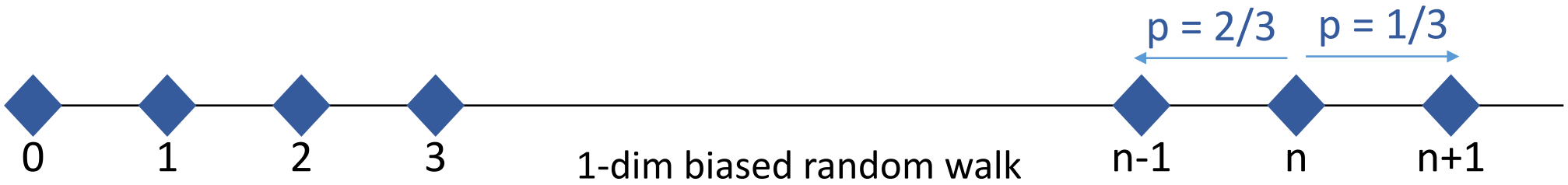
(2) *Strictly decreasing away from target*. There exists $\epsilon > 0$ s.t. $V(x) \geq \epsilon + V(f(x, \pi(x))$ for each $x \in X \backslash X_T$.

Cannot decrease indefinitely while staying non-negative, so target must eventually be reached

What about stochastic reachability?

In SDS, each state has multiple successor states. Can we still use ranking functions?

A1: Require decrease along each trajectory. This works, but is too restrictive! E.g.



1-dim biased random walk terminates **with probability 1**

A2: Require decrease **in expectation**.

Motivation: Ranking supermartingales

Fact: If a ranking function is required to only strictly decrease **in expectation**,
it can be shown that it cannot decrease indefinitely while staying non-negative
(except on a set of trajectories of measure 0).

=>: Target region X_T is reached **with probability 1**.

Stochastic processes that decrease in expectation are called **supermartingales**. That is why we will study martingale theory and use it to design certificates for SDS verification.

Plan for today

1. Stochastic control problem

2. Motivation for certificates

3. **Crash course on martingale theory**

4. Martingale-based certificates

Importance and applications of martingale theory

- Stochastic control
- Verification of probabilistic models and programs
- Design of randomised algorithms
- Financial modelling
- Signal processing
- …

Some definitions

Definition: Probability space

A probability space is a triple $(\Omega, \mathscr{F}, \mathbb{P})$ where:

- Ω is the sample space;
- $\mathscr{F} \subseteq 2^{\Omega}$ is a sigma-algebra (contains the empty set, closed under complementation and countable union);
- $\mathbb{P} : \mathscr{F} \to [0,1]$ is a probability measure that satisfies Kolmogorov axioms.

The elements of $\mathscr{F}$ are called <u>events</u>.

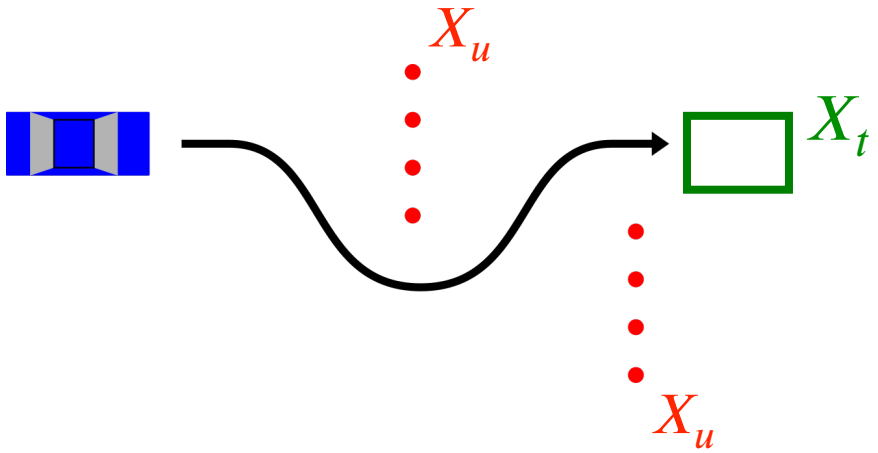
Definition: Random variable

A (real-valued) random variable in a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is a function $X : \Omega \to \mathbb{R}$ that is $\mathscr{F}$-measurable, i.e. for each $a \in \mathbb{R}$ we have that $\{\omega \in \Omega \mid X(\omega) \leq a\} \in \mathscr{F}$.

A sequence of random variables $(X_t)_{t=0}^{\infty}$ is called a <u>stochastic process</u>.

Example: Probability space of all SDS trajectories

- Sample space Ω is the set of all random system trajectories starting in some initial state x_0.
- Sigma-algebra $\mathscr{F}$ is the set generated by all **cylinder sets**, i.e. sets of trajectories with a common prefix $\text{Cylinder}(X_0, X_1, \ldots, X_k) = \{(x_t, \pi(x_t), w_t)_{t=0}^{\infty} \mid x_0 \in X_0, x_1 \in X_1, \ldots, x_k \in X_k\}$
- $\mathbb{P}$ is the probability measure induced by the "natural definition" on cylinder sets*



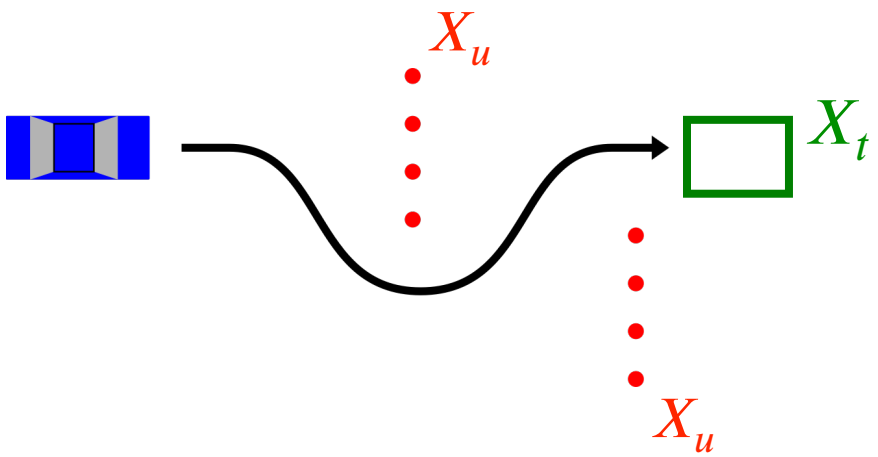
*Formalised via cylinder construction, see e.g. Ash, R., Doléans-Dade, C.: Probability and Measure Theory. Harcourt Academic Press (2000)

Example: Random variable in SDS

Suppose that $V : X \to \mathbb{R}$ is a function that assigns real values to system states.

Random variable $Z_n(\rho) = V(x_n)$ to each $\rho = (x_t, \pi(x_t), w_t)_{t=0}^{\infty}$ assigns the **value of V at its n-th state**.

Stochastic process $Z_0, Z_1, Z_2, \ldots$ tracks the value of V along a random system trajectory.

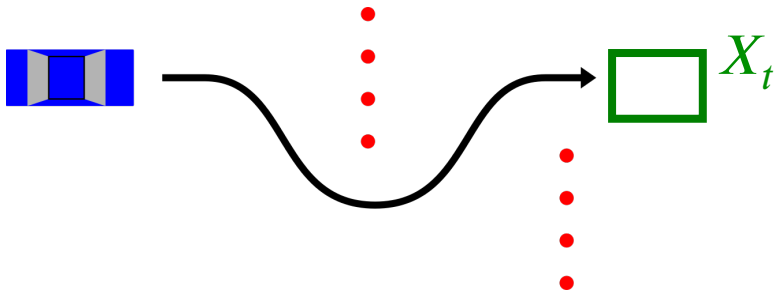


Filtration in a probability space

Definition: Filtration

A filtration in a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is an increasing sequence $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \ldots \subseteq \mathscr{F}$ of sub-sigma-algebras.

Example: Canonical filtration in an SDS



Consider the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ of an SDS.

$\mathscr{F}_0$ = set of all events that can be described in terms of the **first state** along the trajectory

$\mathscr{F}_1$ = set of all events that can be described in terms of the **first two states** along the trajectory

$\mathscr{F}_2$ = set of all events that can be described in terms of the **first three states** along the trajectory

…

$\mathscr{F}_n$ = set of all events that can be described in terms of the **first n states** along the trajectory

…

Conditional expectation

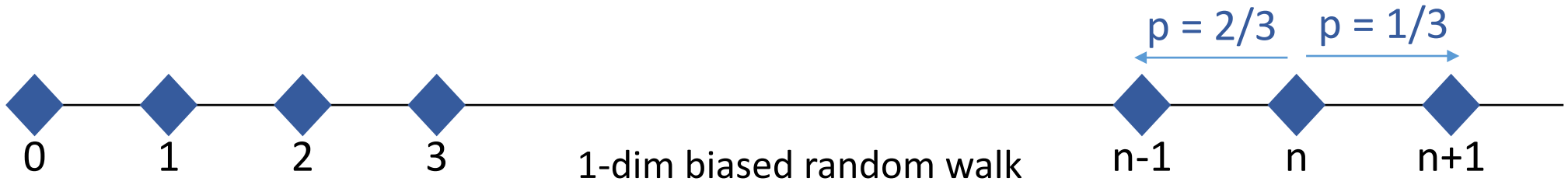
Definition: Conditional expectation

Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$ in it.
Suppose that $\mathscr{F}' \subseteq \mathscr{F}$ is a sub-sigma-algebra. A conditional expectation of X given $\mathscr{F}'$ is an $\mathscr{F}'$-measurable random variable Y, such that $\mathbb{E}[X \cdot \mathbb{I}(A)] = \mathbb{E}[Y \cdot \mathbb{I}(A)]$ for each event $A \in \mathscr{F}'$.

<u>Notation</u>: For event A, the indicator function $\mathbb{I}(A)$ is a random variable defined via $\mathbb{I}(A)(\omega) = 1$ if $\omega \in A$ and $\mathbb{I}(A)(\omega) = 0$ otherwise.

Intuition: Conditional expectation Y behaves exactly like X whenever its expected value is taken over an event in $\mathscr{F}'$.

Example: Conditional expectation on a random walk



1-dim biased random walk terminates **with probability 1**

Stochastic process $(X_n)_{n=0}^{\infty}$ defined by the state of the random walk.

For each n, let $\mathscr{F}_n$ be the sigma-algebra defined by trajectory prefixes of length $\leq n$.

Q: What is the conditional expectation $\mathbb{E}[X_{n+1} \mid \mathscr{F}_n]$?

A: $\mathbb{E}[X_{n+1} \mid \mathscr{F}_n] = X_n - 1/3$

Conditional expectation

Conditional expectation need not necessarily exist. However, if a random variable is **non-negative**, then conditional expectation **exists** and is **almost-surely unique**.

Theorem

Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a sub-sigma-algebra $\mathscr{F}' \subseteq \mathscr{F}$. Let X be an $\mathscr{F}$-measurable random variable which is **non-negative** or **integrable**. Then:

(1) Conditional expectation of X given $\mathscr{F}'$ exists.

(2) Let Y' and Y'' be two conditional expectation of X given $\mathscr{F}'$. Then $\mathbb{P}[Y' = Y''] = 1$.

<u>Notation</u>: Hence, we may fix one conditional expectation of X given $\mathscr{F}'$ and denote it by $\mathbb{E}[X \mid \mathscr{F}']$.

(Super/sub)martingales

Definition: (Super/sub)martingales

Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a filtration $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \ldots \subseteq \mathscr{F}$ in it.

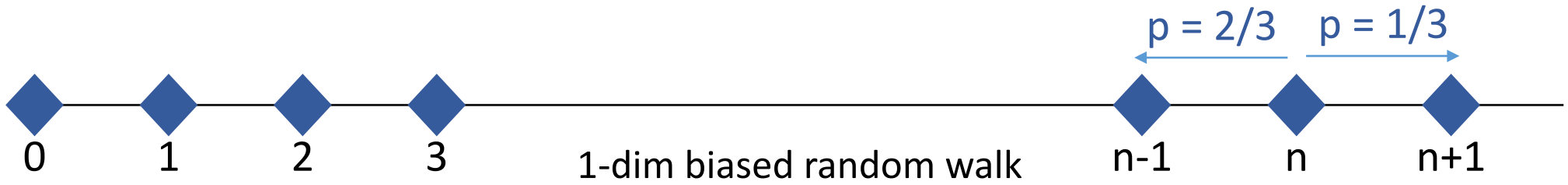
A stochastic process $(X_t)_{t=0}^{\infty}$ is said to be a <u>martingale</u>, if it satisfies the following conditions:

1. Each X_t is $\mathscr{F}_t$ measurable.
2. Each conditional expectation $\mathbb{E}[X_{t+1} \mid \mathscr{F}_t]$ exists.
3. We have $\mathbb{E}[X_{t+1} \mid \mathscr{F}_t] = X_t$ for each t.

A stochastic process $(X_t)_{t=0}^{\infty}$ is a <u>supermartingale</u>, if in condition 3 we only have $\mathbb{E}[X_{t+1} \mid \mathscr{F}_t] \leq X_t$.

A stochastic process $(X_t)_{t=0}^{\infty}$ is a <u>submartingale</u>, if in condition 3 we only have $\mathbb{E}[X_{t+1} \mid \mathscr{F}_t] \geq X_t$.

Example: Supermartingale on a random walk



1-dim biased random walk terminates **with probability 1**

Stochastic process $(X_n)_{n=0}^{\infty}$ defined by the state of the random walk.

For each n, let $\mathscr{F}_n$ be the sigma-algebra defined by trajectory prefixes of length $\leq n$.

Q: What is the conditional expectation $\mathbb{E}[X_{n+1} \mid \mathscr{F}_n]$?

A: $\mathbb{E}[X_{n+1} \mid \mathscr{F}_n] = X_n - 1/3$

$(X_n)_{n=0}^{\infty}$ is a supermartingale

But this is not yet what we want…

We do not want expected decrease at every step, but only **until the target region is reached**.

Also, we want **strict expected decrease**.

Finally, we want the process to have **non-negative values**.

Recall: Ranking function

A ranking function for the target region $X_T \subseteq X$ under controller π is a **function** $V : X \to \mathbb{R}$ that assigns a real value to each system state, and which is required to be

(1) *Non-negative*. $V(x) \geq 0$ for each $x \in X$.

(2) *Strictly decreasing away from target*. There exists $\epsilon > 0$ s.t. $V(x) \geq \epsilon + V(f(x, \pi(x))$ for each $x \in X \backslash X_T$.

Stopping time

Definition: Stopping time

Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a filtration $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \ldots \subseteq \mathscr{F}$ in it.
A stopping time with respect to the filtration is a random variable $T : \Omega \to \mathbb{N}_0 \cup \{\infty\}$, such that for each $t \in \mathbb{N}_0$, we have $\{\omega \in \Omega \mid T(\omega) \leq t\} \in \mathscr{F}_t$.

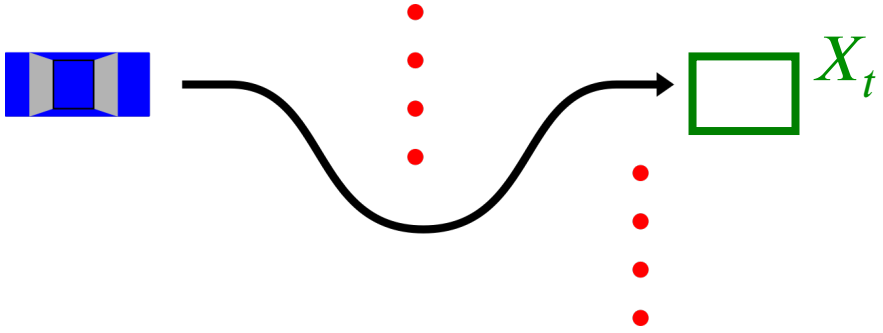
Intuition: The event that the stopping time is less than or equal to t can be observed in the sub-sigma-algebra $\mathscr{F}_t$.

Example: First hitting time in an SDS

Consider the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ of an SDS and the canonical filtration in it.

Define $T : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ to be the first hitting time of the target region X_T, i.e. the index of the first state along the trajectory that is contained in X_T.

Then, for each $t \in \mathbb{N}_0$, we have $\{\omega \in \Omega \mid T(\omega) \leq t\} \in \mathscr{F}_t$. So T is a stopping time.



Ranking supermartingales (RSMs)

Definition: Ranking supermartingale (RSM) [1]

Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a filtration $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \ldots \subseteq \mathscr{F}$ in it.

Let T be a stopping time with respect to the filtration.

A stochastic process $(X_t)_{t=0}^{\infty}$ is said to be a **ranking supermartingale with respect to** T, if it satisfies:

1. Each X_t is $\mathscr{F}_t$-measurable.
2. Each X_t is **non-negative** (hence conditional expectation $\mathbb{E}[X_{t+1} \mid \mathscr{F}_t]$ exists).
3. There exists $\epsilon > 0$, such that $\mathbb{E}[X_{t+1} \mid \mathscr{F}_t] = X_t - \epsilon \cdot \mathbb{I}[t > T]$ for each t.

Intuition: The expected decrease is strict until the stopping time is reached. After that, the expected decrease is non-strict. In addition, the process is always non-negative.

[1] Chakarov, Sankaranarayanan. *Probabilistic Program Analysis with Martingales.* CAV 2013

(Ranking) Supermartingale convergence theorem

Theorem (Supermartingale convergence theorem)

Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a filtration $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \ldots \subseteq \mathscr{F}$ in it. Let $(X_t)_{t=0}^{\infty}$ be a non-negative/integrable supermartingale. Then, there exists a non-negative/integrable random variable X_∞ such that $\mathbb{P}[\lim_{t\to\infty} X_t = X_\infty] = 1$.

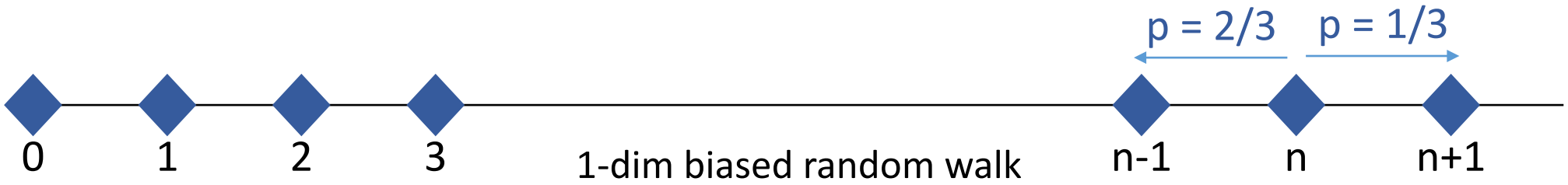
Theorem (RSM convergence theorem) [1]

Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a filtration $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \ldots \subseteq \mathscr{F}$ in it. Let T be a stopping time with respect to the filtration and $(X_t)_{t=0}^{\infty}$ an RSM with respect to T. Then, $\mathbb{P}[T < \infty] = 1$.

Intuition: If there exists a ranking supermartingale with respect to stopping time T, then T is finite with probability 1.

[1] Chakarov, Sankaranarayanan. *Probabilistic Program Analysis with Martingales.* CAV 2013

Example: RW reaches the origin with probability 1



1-dim biased random walk terminates **with probability 1**

Stochastic process $(X_n)_{n=0}^{\infty}$ defined by the state of the random walk.

For each n, let $\mathscr{F}_n$ be the sigma-algebra defined by trajectory prefixes of length $\leq n$.

Let T be the first hitting time of the origin 0.

Then, $(X_n)_{n=0}^{\infty}$ is a **ranking supermartingale** with respect to T.

Hence, the random walk eventually reaches the origin 0 with probability 1.

Another important result on supermartingales

Convergence theorems allow us to reason about probability 1 reachability. The following result will further allow us to reason about quantitative properties (probabilistic reachability and avoidance).

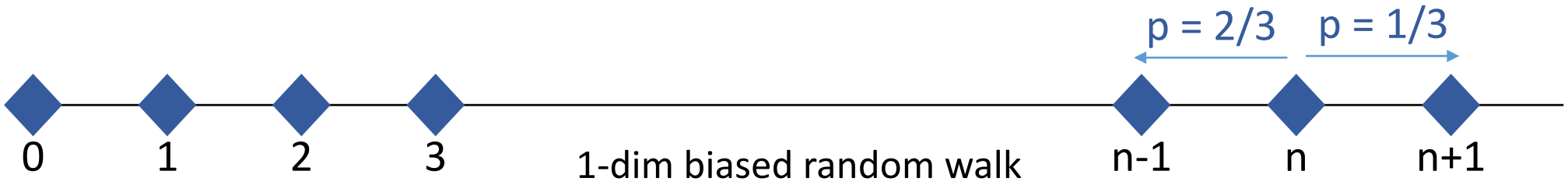
Theorem

Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a filtration $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \ldots \subseteq \mathscr{F}$ in it. Let $(X_t)_{t=0}^{\infty}$ a non-negative supermartingale. Then, for every constant $\lambda > 0$, we have

$$\mathbb{P}[\sup_{t \geq 0} X_t \geq \lambda] \leq \frac{\mathbb{E}[X_0]}{\lambda}.$$

Intuition: Since the process decreases in expectation, we can bound from above the probability that it ever exceeds some value λ in terms of its expected initial value.

Example: Concentration bound for RW



1-dim biased random walk terminates **with probability 1**

Stochastic process $(X_n)_{n=0}^{\infty}$ defined by the state of the random walk.

For each n, let $\mathscr{F}_n$ be the sigma-algebra defined by trajectory prefixes of length $\leq n$.

Let T be the first hitting time of the origin 0.

Then, $(X_n)_{n=0}^{\infty}$ is a **ranking supermartingale** with respect to T.

Hence, if $X_0 = 10$, we have $\mathbb{P}[\sup_{t \geq 0} X_t \geq 1000] \leq 0.01$.

Plan for today

1. Stochastic control problem

2. Motivation for certificates

3. Crash course on martingale theory

4. Martingale-based certificate

Recall: Stochastic control problem

Definition: Reachability, safety, reach-avoidance

Given: An SDS (X, U, W, f, d), initial region $X_0 \subseteq X$, target region $X_T \subseteq X$, unsafe region $X_U \subseteq X$, probability threshold $p \in [0,1]$

Task: Compute a controller $\pi : X \to U$ such that $\mathbb{P}_{x_0}[\text{ReachAvoid}(X_T, X_U)] \geq p$ for every $x_0 \in X_0$.

<u>Notation</u>: ReachAvoid(X_T, X_U) is the set of all trajectories that satisfy the reach-avoidance specification.

Recipe for a reach-avoidance certificate [1,2]

1. Reach the target region X_T **OR** the unsafe region X_U with probability 1

2. Reach the unsafe region X_U with probability $\leq 1 - p$

Reach the target region X_T **AND** avoid the unsafe region X_U with probability $\geq p$

(1) ranking supermartingale convergence theorem
+
(2) bound on the supermartingale supreme value

[1] Zikelic, Lechner, Henzinger, Chatterjee. *Learning Control Policies for Stochastic Systems with Reach-avoid Guarantees.* AAAI 2023
[2] Chatterjee, Novotny, Zikelic. *Stochastic Invariants for Probabilistic Termination.* POPL 2017

Reach-avoid supermartingales

Definition: Reach-avoid supermartingale [1]

Consider an SDS (X, U, W, f, d) with initial region $X_0 \subseteq X$, and let $\pi : X \to U$ be a controller.
A reach-avoid supermartingale with respect to X_T, X_U and p is a (measurable) function $V : X \to \mathbb{R}$, which satisfies the following conditions:

(1) **Non-negativity**. $V(x) \geq 0$ for each state $x \in X$.

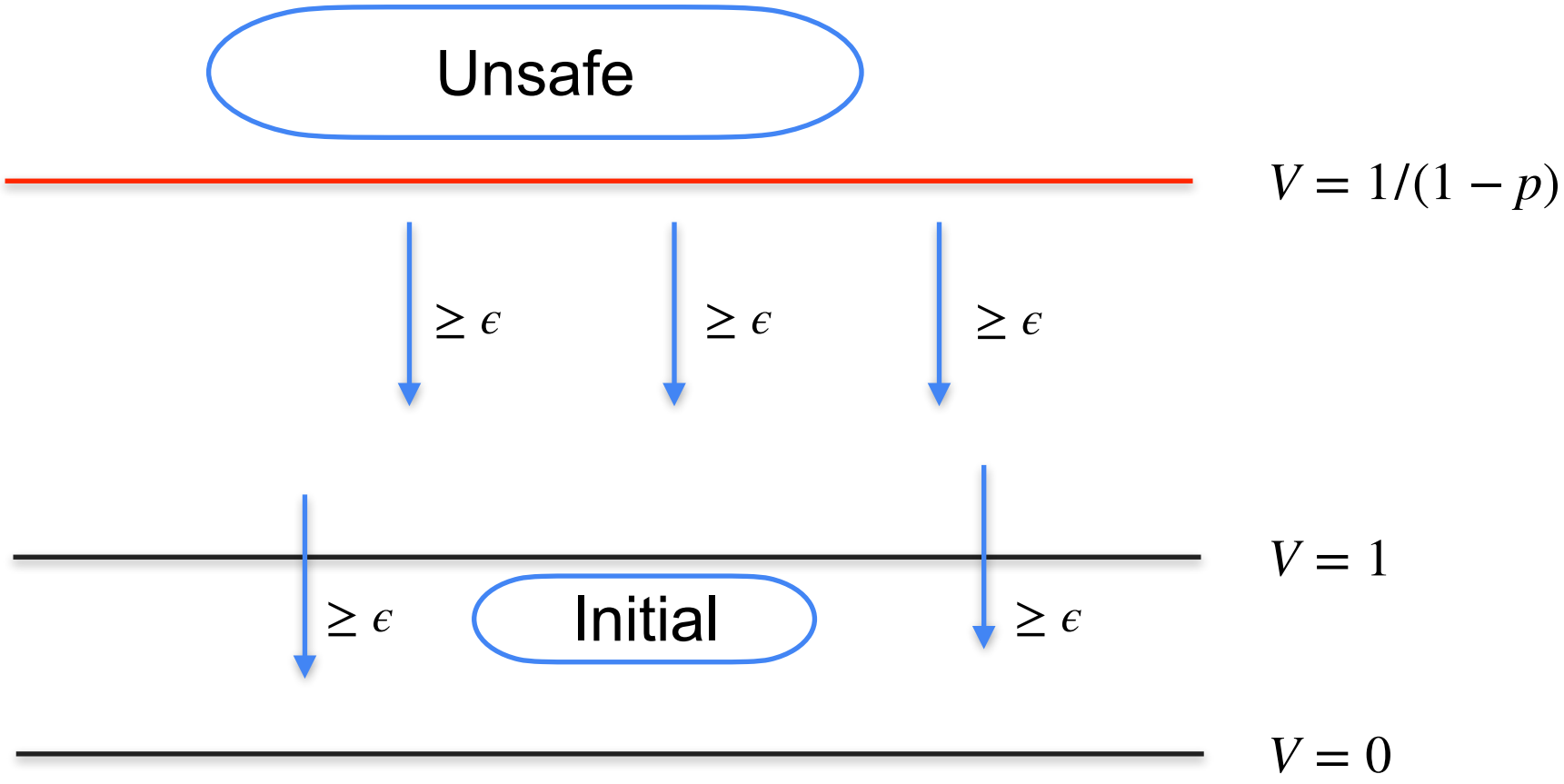
(2) **Strict expected decrease**. There exists $\epsilon > 0$ such that $V(x) \geq \epsilon + \mathbb{E}_{w \sim d}[V(f(x, \pi(x), w)]$ for all $x \in X \backslash X_T$ at which $V(x) \leq 1/(1-p)$

(3) **Initial condition**. $V(x) \leq 1$ for each initial state $x \in X_0$.

(4) **Unsafe condition**. $V(x) \geq 1/(1-p)$ for each unsafe state $x \in X_U$.

[1] Zikelic, Lechner, Henzinger, Chatterjee. *Learning Control Policies for Stochastic Systems with Reach-avoid Guarantees.* AAAI 2023

Intuition behind RASMs



Soundness of RASM certificates

Theorem [1]

Consider an SDS (X, U, W, f, d) with initial region $X_0 \subseteq X$, and let $\pi : X \to U$ be a controller. Suppose that there exists a RASM with respect to X_T, X_U and p. Then, we have that

$$\mathbb{P}_{x_0}[\text{ReachAvoid}(X_T, X_U)] \geq p \text{ for every } x_0 \in X_0.$$

[1] Zikelic, Lechner, Henzinger, Chatterjee. *Learning Control Policies for Stochastic Systems with Reach-avoid Guarantees.* AAAI 2023

Proof sketch of RASM soundness

1. Consider the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ of the SDS and its canonical filtration

2. Consider the stopping time T defined by the first hitting time of $X_T \cup \{x \in X \mid V(x) > 1/(1-p)\}$

3. By the (1) **Non-negativity** and (2) **Strict expected decrease** conditions of RASMs, V defines a ranking supermartingale with respect to the stopping time T. So, by the ranking supermartingale convergence theorem, $X_T \cup \{x \in X \mid V(x) > 1/(1-p)\}$ is reached with probability 1

4. By the (3) **Initial** condition of RASMs and by the supremum bound theorem, $\{x \in X \mid V(x) > 1/(1-p)\}$ is reached with probability at most $1-p$.

5. Hence, X_T is reached while $\{x \in X \mid V(x) > 1/(1-p)\}$ is avoided with probability at least p.

6. By the (4) **Safety** condition of RASMs, we have $X_U \subseteq \{x \in X \mid V(x) > 1/(1-p)\}$. So, we conclude that X_T is reached while X_U is avoided with probability at least p

RASM certificates for reachability and safety

Corollary (Reachability)

Consider an SDS (X, U, W, f, d) with initial region $X_0 \subseteq X$, and let $\pi : X \to U$ be a controller. Suppose that there exists a RASM with respect to X_T, $\varnothing$ and p. Then, we have that

$$\mathbb{P}_{x_0}[\text{Reach}(X_T)] \geq p \text{ for every } x_0 \in X_0.$$

Corollary (Safety)

Consider an SDS (X, U, W, f, d) with initial region $X_0 \subseteq X$, and let $\pi : X \to U$ be a controller. Suppose that there exists a RASM with respect to $\varnothing$, X_U and p. Then, we have that

$$\mathbb{P}_{x_0}[\text{Safe}(X_U)] \geq p \text{ for every } x_0 \in X_0.$$

Improving RASM bounds

Many works on using (super/sub)martingales **for probabilistic reachability and safety**

Exponential decay bounds (under additional assumptions on system dynamics):

- Chatterjee, Novotny, Zikelic. *Stochastic Invariants for Probabilistic Termination.* POPL 2017
- Takisaka, Oyabu, Urabe, Hasuo. *Ranking and Repulsing Supermartingales for Reachability in Randomized Programs*. TOPLAS 2021
- Zikelic, Lechner, Verma, Chatterjee, Henzinger. *Compositional Policy Learning in Stochastic Control Systems with Formal Guarantees.* NeurIPS 2023

Finite-time horizon bounds (for safety):

- Santoyo, Dutreix, Coogan. *A barrier function approach to finite-time stochastic system verification and control.* Automatica 2021
- Zhi, Wang, Liu, Ong, Zhang. *Unifying Qualitative and Quantitative Safety Verification of DNN-Controlled Systems.* CAV 2024

Summary

1. We defined the stochastic control problem under reach-avoidance specifications.

2. We motivated the use of martingale certificates.

3. Crash course on martingale theory.

4. RASMs as stochastic reach-avoidance certificates.

<u>Next time</u>: Neural controller synthesis and verification via RASMs.

Open Positions

SMU
SINGAPORE MANAGEMENT UNIVERSITY

School of
Computing and Information Systems

! Multiple **PhD positions** in computer science (fully funded)

! Multiple **Visiting Research Studentships** (fully funded, 6 months)

Possible topics include (but not limited to):

- Formal methods for probabilistic systems
- Formal methods for trustworthy AI
  (learning-based control, safe RL, NN verification…)
- Program verification (probabilistic and numerical programs)

Apply here (PhD positions): https://computing.smu.edu.sg/phd
Contact me: dzikelic@smu.edu.sg
Personal website: https://djordjezikelic.github.io/

References for today

Probability and martingale theory:

[1] D. Williams. *Probability with martingales*. Cambridge University Press, 1991

Ranking supermartingales:

[2] A. Chakarov, S. Sankaranarayanan. *Probabilistic program analysis with martingales*. CAV, 2013

Reach-avoid supermartingales:

[3] D. Zikelic, M. Lechner, T. A. Henzinger, K. Chatterjee. *Learning control policies for stochastic systems with reach-avoid guarantees*. AAAI, 2023

[4] D. Zikelic, M. Lechner, A. Verma, K. Chatterjee, T. A. Henzinger. *Compositional Policy Learning in Stochastic Control Systems with Formal Guarantees.* NeurIPS 2023