

AN EXPERIMENTALIST'S VENTURE INTO RL THEORY

Two Successes and a Failure

AI Seminar
16 Feb 2024

Abhishek Naik

with thanks to Janey, Yi, and Rich



UNIVERSITY OF
ALBERTA



WHERE IT ALL BEGAN

Learning and Planning in Average-Reward Markov Decision Processes

Yi Wan^{*1} Abhishek Naik^{*1} Richard S. Sutton¹²

Abstract

We introduce learning and planning algorithms for average-reward MDPs, including 1) the first general proven-convergent off-policy model-free control algorithm without reference states, 2) the first proven-convergent off-policy model-free prediction algorithm, and 3) the first off-policy learning algorithm that converges to the actual value function rather than to the value function plus an offset. All of our algorithms are based on us-

with it. For learning and combined methods, both control and prediction problems can be further subdivided into *on-policy* versions, in which data is gathered using the target policy, and *off-policy* versions, in which data is gathered using a second policy, called the *behavior policy*. In general, both policies may be non-stationary. For example, in the control problem, the target policy should converge to a policy that maximizes the reward rate. Useful surveys of average-reward learning are given by Mahadevan (1996) and Dewanto et al. (2020).

WHERE IT ALL BEGAN

Learning and Planning in Average-Reward Markov Decision Processes

Yi Wan^{*1} Abhishek Naik^{*1} Richard S. Sutton^{1 2}

Abstract

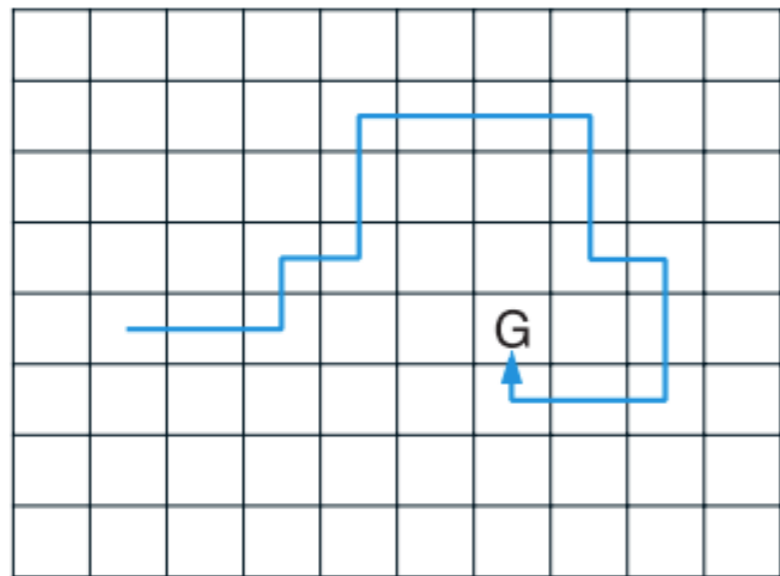
We introduce learning and planning algorithms for average-reward MDPs, including 1) the first general proven-convergent off-policy model-free control algorithm without reference states, 2) the first proven-convergent off-policy model-free prediction algorithm, and 3) the first off-policy learning algorithm that converges to the actual value function rather than to the value function plus an offset. All of our algorithms are based on us-

with it. For learning and combined methods, both control and prediction problems can be further subdivided into *on-policy* versions, in which data is gathered using the target policy, and *off-policy* versions, in which data is gathered using a second policy, called the *behavior policy*. In general, both policies may be non-stationary. For example, in the control problem, the target policy should converge to a policy that maximizes the reward rate. Useful surveys of average-reward learning are given by Mahadevan (1996) and Dewanto et al. (2020).

- ▶ *One-step* tabular average-reward methods

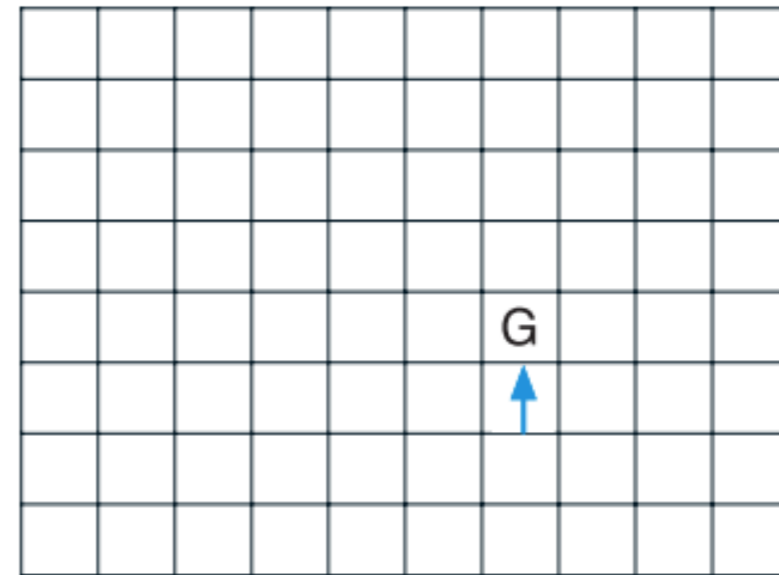
WE WANT MORE EFFICIENT CREDIT ASSIGNMENT

WE WANT MORE EFFICIENT CREDIT ASSIGNMENT



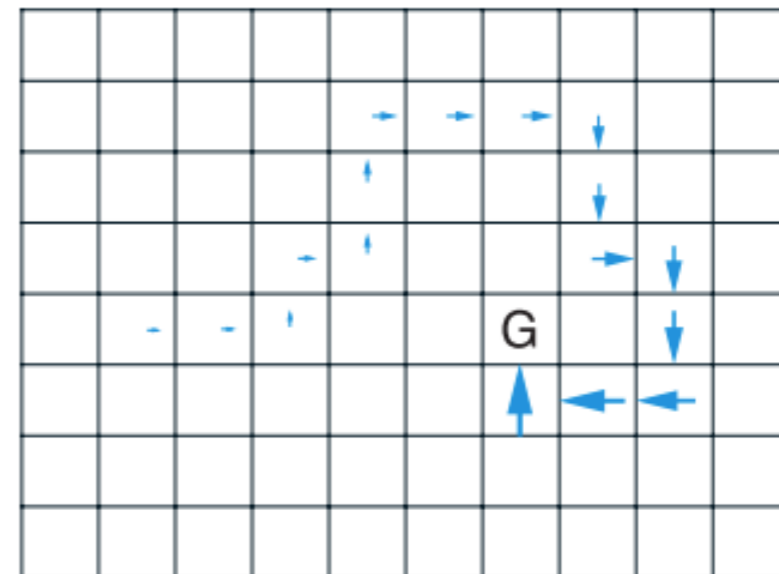
Trajectory

One-step
update

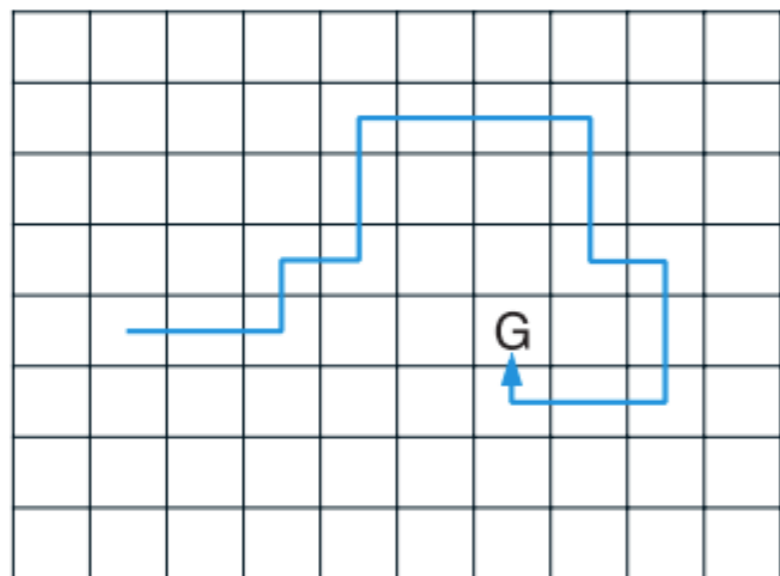


Learned values / policy

Multi-step
update

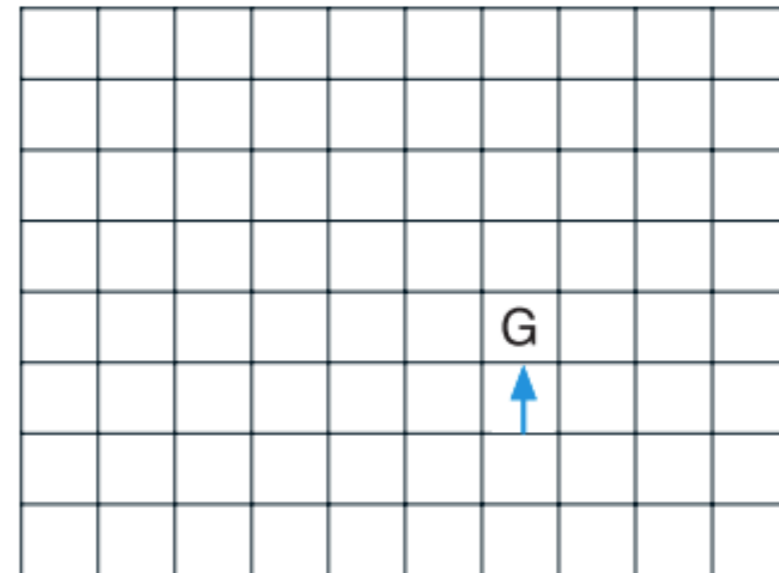


WE WANT MORE EFFICIENT CREDIT ASSIGNMENT



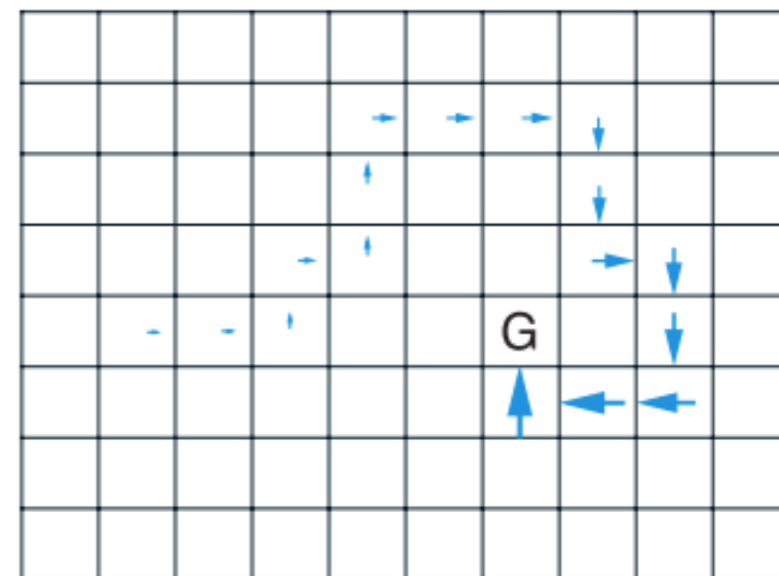
Trajectory

One-step
update



Learned values / policy

Multi-step
update



- ▶ *Multi-step* average-reward methods

PROBLEM SETTING

THE AVERAGE-REWARD FORMULATION

PROBLEM SETTING

THE AVERAGE-REWARD FORMULATION



PROBLEM SETTING

THE AVERAGE-REWARD FORMULATION

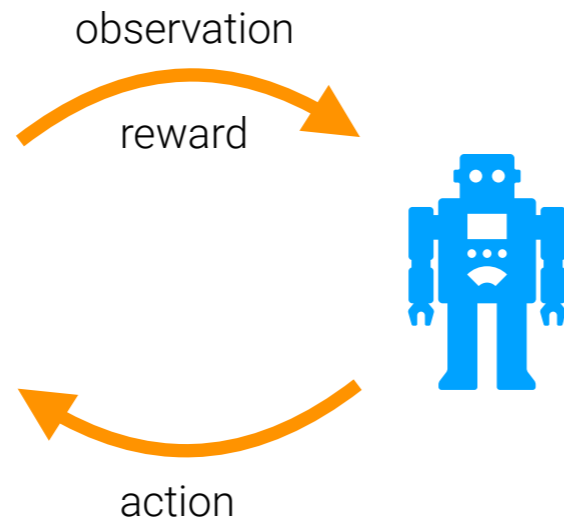
$$S_0 \ A_0 \ R_1 \ S_1 \ A_1, R_2 \dots \ S_t \ A_t \ R_{t+1} \ S_{t+1} \ A_{t+1} \ R_{t+2} \ \dots$$



PROBLEM SETTING

THE AVERAGE-REWARD FORMULATION

$$S_0 \ A_0 \ R_1 \ S_1 \ A_1, R_2 \dots \ S_t \ A_t \ R_{t+1} \ S_{t+1} \ A_{t+1} \ R_{t+2} \ \dots$$



PROBLEM SETTING

THE AVERAGE-REWARD FORMULATION

$S_0 A_0 R_1 S_1 A_1, R_2 \dots S_t A_t R_{t+1} S_{t+1} A_{t+1} R_{t+2} \dots$



$$r(\pi) \doteq \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}_{\pi} \left[\sum_{t=1}^n R_t \right]$$

PROBLEM SETTING

THE AVERAGE-REWARD FORMULATION

$S_0 A_0 R_1 S_1 A_1, R_2 \dots S_t A_t R_{t+1} S_{t+1} A_{t+1} R_{t+2} \dots$



$$r(\pi) \doteq \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}_{\pi} \left[\sum_{t=1}^n R_t \right]$$

PROBLEM SETTING

THE AVERAGE-REWARD FORMULATION

$S_0 A_0 R_1 S_1 A_1, R_2 \dots S_t A_t R_{t+1} S_{t+1} A_{t+1} R_{t+2} \dots$



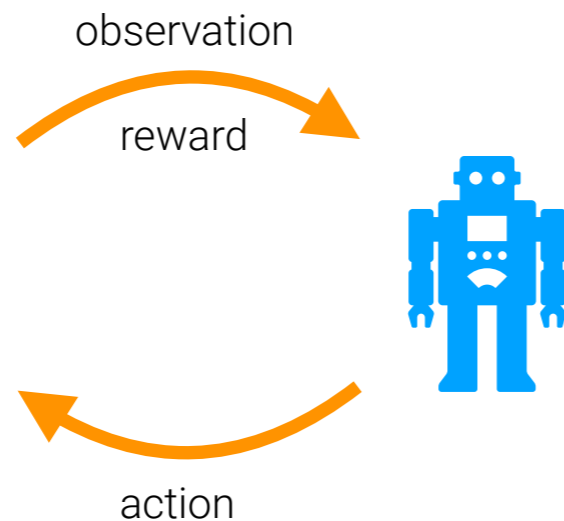
$$r(\pi) \doteq \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}_{\pi} \left[\sum_{t=1}^n R_t \right]$$

$$v_{\pi}(s) \doteq \mathbb{E}_{\pi} [R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + \dots | S_t = s]$$

PROBLEM SETTING

THE AVERAGE-REWARD FORMULATION

$S_0 A_0 R_1 S_1 A_1, R_2 \dots S_t A_t R_{t+1} S_{t+1} A_{t+1} R_{t+2} \dots$



$$r(\pi) \doteq \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}_{\pi} \left[\sum_{t=1}^n R_t \right]$$

$$v_{\pi}(s) \doteq \mathbb{E}_{\pi} [R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + \dots | S_t = s]$$

$$v_{\pi}^{\gamma}(s) \doteq \mathbb{E}_{\pi} [R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots | S_t = s]$$

THE AVERAGE-REWARD FORMULATION

$$S_0 \ A_0 \ R_1 \ S_1 \ A_1, R_2 \dots \ S_t \ A_t \ R_{t+1} \ S_{t+1} \ A_{t+1} \ R_{t+2} \ \dots$$



$$r(\pi) \doteq \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}_{\pi} \left[\sum_{t=1}^n R_t \right]$$

$$v_{\pi}(s) \doteq \mathbb{E}_{\pi} [R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + \dots \mid S_t = s]$$

$$v_{\pi}^{\gamma}(s) \doteq \mathbb{E}_{\pi} [R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s]$$

The Prediction Problem

- ▶ Estimate $r(\pi)$ and v_{π} using data generated by some policy b .

ON-POLICY PREDICTION

ON-POLICY PREDICTION

$$v_{\pi}(s) \approx \mathbf{w}^{\top} \mathbf{x}(s)$$

ON-POLICY PREDICTION

$$v_{\pi}(s) \approx \mathbf{w}^{\top} \mathbf{x}(s)$$

One-step Differential TD

ON-POLICY PREDICTION

$$v_{\pi}(s) \approx \mathbf{w}^{\top} \mathbf{x}(s)$$

One-step Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{x}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t \quad \eta > 0$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^{\top} \mathbf{x}_{t+1} - \mathbf{w}_t^{\top} \mathbf{x}_t$

ON-POLICY PREDICTION

$$v_{\pi}(s) \approx \mathbf{w}^{\top} \mathbf{x}(s)$$

One-step Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{x}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t \quad \eta > 0$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^{\top} \mathbf{x}_{t+1} - \mathbf{w}_t^{\top} \mathbf{x}_t$

Multi-step version

ON-POLICY PREDICTION

$$v_{\pi}(s) \approx \mathbf{w}^{\top} \mathbf{x}(s)$$

One-step Differential TD

$$\begin{aligned} \mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{x}_t \\ \bar{R}_{t+1} &\doteq \bar{R}_t + \eta \alpha_t \delta_t \quad \eta > 0 \end{aligned}$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^{\top} \mathbf{x}_{t+1} - \mathbf{w}_t^{\top} \mathbf{x}_t$

Multi-step version

$$\begin{aligned} \mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t \\ \bar{R}_{t+1} &\doteq \bar{R}_t + \eta \alpha_t \delta_t \end{aligned}$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^{\top} \mathbf{x}_{t+1} - \mathbf{w}_t^{\top} \mathbf{x}_t$
 $\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$

ON-POLICY PREDICTION

$$v_{\pi}(s) \approx \mathbf{w}^{\top} \mathbf{x}(s)$$

One-step Differential TD

$$\begin{aligned} \mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{x}_t \\ \bar{R}_{t+1} &\doteq \bar{R}_t + \eta \alpha_t \delta_t \quad \eta > 0 \end{aligned}$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^{\top} \mathbf{x}_{t+1} - \mathbf{w}_t^{\top} \mathbf{x}_t$

Multi-step version

$$\begin{aligned} \mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t \\ \bar{R}_{t+1} &\doteq \bar{R}_t + \eta \alpha_t \delta_t \end{aligned}$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^{\top} \mathbf{x}_{t+1} - \mathbf{w}_t^{\top} \mathbf{x}_t$
 $\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$

Algorithm 1

ON-POLICY PREDICTION

$$v_{\pi}(s) \approx \mathbf{w}^{\top} \mathbf{x}(s)$$

One-step Differential TD

$$\begin{aligned} \mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{x}_t \\ \bar{R}_{t+1} &\doteq \bar{R}_t + \eta \alpha_t \delta_t \quad \eta > 0 \end{aligned}$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^{\top} \mathbf{x}_{t+1} - \mathbf{w}_t^{\top} \mathbf{x}_t$

Multi-step version

$$\begin{aligned} \mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t \\ \bar{R}_{t+1} &\doteq \bar{R}_t + \eta \alpha_t \delta_t \end{aligned}$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^{\top} \mathbf{x}_{t+1} - \mathbf{w}_t^{\top} \mathbf{x}_t$

$$\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$$

Algorithm 1

Is it guaranteed to converge...?

INTUITIONS ABOUT CONVERGENCE THEORY

SAMPLE-BASED RL ALGORITHMS AS DIFFERENTIAL EQUATIONS

SAMPLE-BASED RL ALGORITHMS AS DIFFERENTIAL EQUATIONS

$$\dot{\mathbf{w}}_{t+1} \doteq \mathbf{w}_t + \alpha_t \mathbf{x}_t [R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t]$$

SAMPLE-BASED RL ALGORITHMS AS DIFFERENTIAL EQUATIONS

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \mathbf{x}_t [R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t]$$

$$V_{t+1}(S_t) \doteq V_t(S_t) + \alpha_t [R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)]$$

$$V_t(S_t) = \mathbf{w}_t^\top \mathbf{x}_t$$

SAMPLE-BASED RL ALGORITHMS AS DIFFERENTIAL EQUATIONS

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \mathbf{x}_t [R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t]$$

$$V_t(S_t) = \mathbf{w}_t^\top \mathbf{x}_t$$

$$V_{t+1}(S_t) \doteq V_t(S_t) + \alpha_t [R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)]$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t [R_{t+1} \mathbf{x}_t + \mathbf{x}_t (\gamma \mathbf{x}_{t+1} - \mathbf{x}_t)^\top \mathbf{w}_t]$$

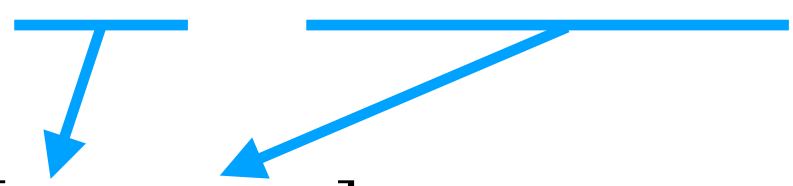
SAMPLE-BASED RL ALGORITHMS AS DIFFERENTIAL EQUATIONS

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \mathbf{x}_t [R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t]$$

$$V_t(S_t) = \mathbf{w}_t^\top \mathbf{x}_t$$

$$V_{t+1}(S_t) \doteq V_t(S_t) + \alpha_t [R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)]$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t [R_{t+1} \mathbf{x}_t + \mathbf{x}_t (\gamma \mathbf{x}_{t+1} - \mathbf{x}_t)^\top \mathbf{w}_t]$$


$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t [\mathbf{b}_t + \mathbf{A}_t \mathbf{w}_t]$$

SAMPLE-BASED RL ALGORITHMS AS DIFFERENTIAL EQUATIONS

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \mathbf{x}_t [R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t]$$

$$V_t(S_t) = \mathbf{w}_t^\top \mathbf{x}_t$$

$$V_{t+1}(S_t) \doteq V_t(S_t) + \alpha_t [R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)]$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t [R_{t+1} \mathbf{x}_t + \mathbf{x}_t (\gamma \mathbf{x}_{t+1} - \mathbf{x}_t)^\top \mathbf{w}_t]$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t [\mathbf{b}_t + \mathbf{A}_t \mathbf{w}_t]$$

$$\mathbf{b}_t \in \mathbb{R}^d$$

$$\mathbf{A}_t \in \mathbb{R}^{d \times d}$$

SAMPLE-BASED RL ALGORITHMS AS DIFFERENTIAL EQUATIONS

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \mathbf{x}_t [R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t]$$

$$V_t(S_t) = \mathbf{w}_t^\top \mathbf{x}_t$$

$$V_{t+1}(S_t) \doteq V_t(S_t) + \alpha_t [R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)]$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t [R_{t+1} \mathbf{x}_t + \mathbf{x}_t (\gamma \mathbf{x}_{t+1} - \mathbf{x}_t)^\top \mathbf{w}_t]$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t [\mathbf{b}_t + \mathbf{A}_t \mathbf{w}_t]$$

$$\mathbf{b}_t \in \mathbb{R}^d$$

$$\mathbf{A}_t \in \mathbb{R}^{d \times d}$$

$$\mathbf{w}_{t+1} - \mathbf{w}_t = \alpha_t [\mathbf{b}_t + \mathbf{A}_t \mathbf{w}_t]$$

SAMPLE-BASED RL ALGORITHMS AS DIFFERENTIAL EQUATIONS

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \mathbf{x}_t [R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t]$$

$$V_t(S_t) = \mathbf{w}_t^\top \mathbf{x}_t$$

$$V_{t+1}(S_t) \doteq V_t(S_t) + \alpha_t [R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)]$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t [R_{t+1} \mathbf{x}_t + \mathbf{x}_t (\gamma \mathbf{x}_{t+1} - \mathbf{x}_t)^\top \mathbf{w}_t]$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t [\mathbf{b}_t + \mathbf{A}_t \mathbf{w}_t]$$

$$\mathbf{b}_t \in \mathbb{R}^d$$

$$\mathbf{A}_t \in \mathbb{R}^{d \times d}$$

$$\frac{\mathbf{w}_{t+1} - \mathbf{w}_t}{(t+1) - t} = \alpha_t [\mathbf{b}_t + \mathbf{A}_t \mathbf{w}_t]$$

SAMPLE-BASED RL ALGORITHMS AS DIFFERENTIAL EQUATIONS

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \mathbf{x}_t [R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t]$$

$$V_t(S_t) = \mathbf{w}_t^\top \mathbf{x}_t$$

$$V_{t+1}(S_t) \doteq V_t(S_t) + \alpha_t [R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)]$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t [R_{t+1} \mathbf{x}_t + \mathbf{x}_t (\gamma \mathbf{x}_{t+1} - \mathbf{x}_t)^\top \mathbf{w}_t]$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t [\mathbf{b}_t + \mathbf{A}_t \mathbf{w}_t]$$

$$\mathbf{b}_t \in \mathbb{R}^d$$

$$\mathbf{A}_t \in \mathbb{R}^{d \times d}$$

$$\frac{\mathbf{w}_{t+1} - \mathbf{w}_t}{(t+1) - t} = \alpha_t [\mathbf{b}_t + \mathbf{A}_t \mathbf{w}_t]$$

$$\frac{d\mathbf{w}_t}{dt} \propto \mathbf{b}_t + \mathbf{A}_t \mathbf{w}_t$$

RECAP: ORDINARY DIFFERENTIAL EQUATIONS

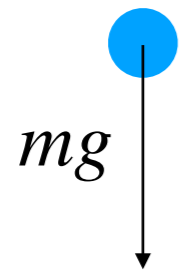
RECAP: ORDINARY DIFFERENTIAL EQUATIONS



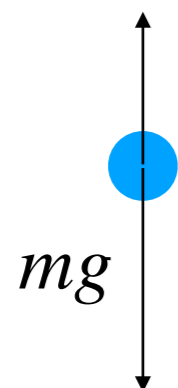
RECAP: ORDINARY DIFFERENTIAL EQUATIONS



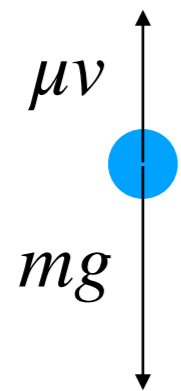
RECAP: ORDINARY DIFFERENTIAL EQUATIONS



RECAP: ORDINARY DIFFERENTIAL EQUATIONS

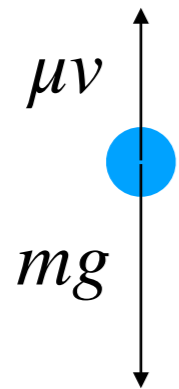


RECAP: ORDINARY DIFFERENTIAL EQUATIONS



RECAP: ORDINARY DIFFERENTIAL EQUATIONS

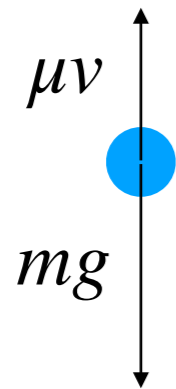
$$F = ma$$



RECAP: ORDINARY DIFFERENTIAL EQUATIONS

$$F = ma$$

$$mg - \mu v_t = ma_t$$

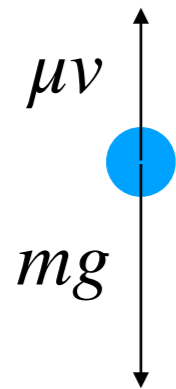


RECAP: ORDINARY DIFFERENTIAL EQUATIONS

$$F = ma$$

$$mg - \mu v_t = ma_t$$

$$mg - \mu v_t = m \frac{dv_t}{dt}$$



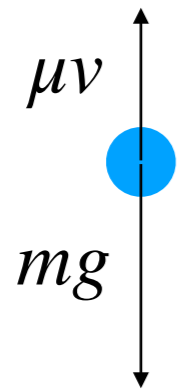
RECAP: ORDINARY DIFFERENTIAL EQUATIONS

$$F = ma$$

$$mg - \mu v_t = ma_t$$

$$mg - \mu v_t = m \frac{dv_t}{dt}$$

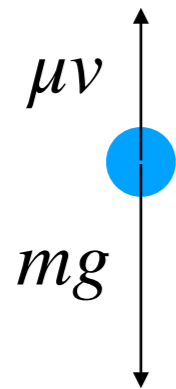
$$\frac{dv_t}{dt} = g - \frac{\mu}{m} v_t$$



RECAP: ORDINARY DIFFERENTIAL EQUATIONS

$$F = ma$$

$$mg - \mu v_t = ma_t$$



$$mg - \mu v_t = m \frac{dv_t}{dt}$$

$$\frac{dv_t}{dt} = g - \frac{\mu}{m} v_t$$

$$\frac{dv_t}{dt} = 10 - 0.2 v_t$$

RECAP: ORDINARY DIFFERENTIAL EQUATIONS

$$F = ma$$

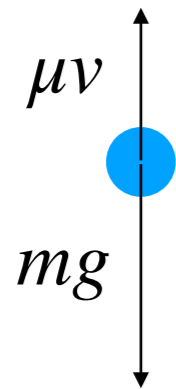
$$mg - \mu v_t = ma_t$$

$$mg - \mu v_t = m \frac{dv_t}{dt}$$

$$\frac{dv_t}{dt} = g - \frac{\mu}{m} v_t$$

$$\frac{dv_t}{dt} = 10 - 0.2 v_t$$

$$\frac{d\mathbf{w}_t}{dt} \propto \mathbf{b}_t + \mathbf{A}_t \mathbf{w}_t$$



RECAP: ORDINARY DIFFERENTIAL EQUATIONS (ODEs)

RECAP: ORDINARY DIFFERENTIAL EQUATIONS (ODEs)

$$\frac{dv_t}{dt} = 10 - 0.2 v_t$$

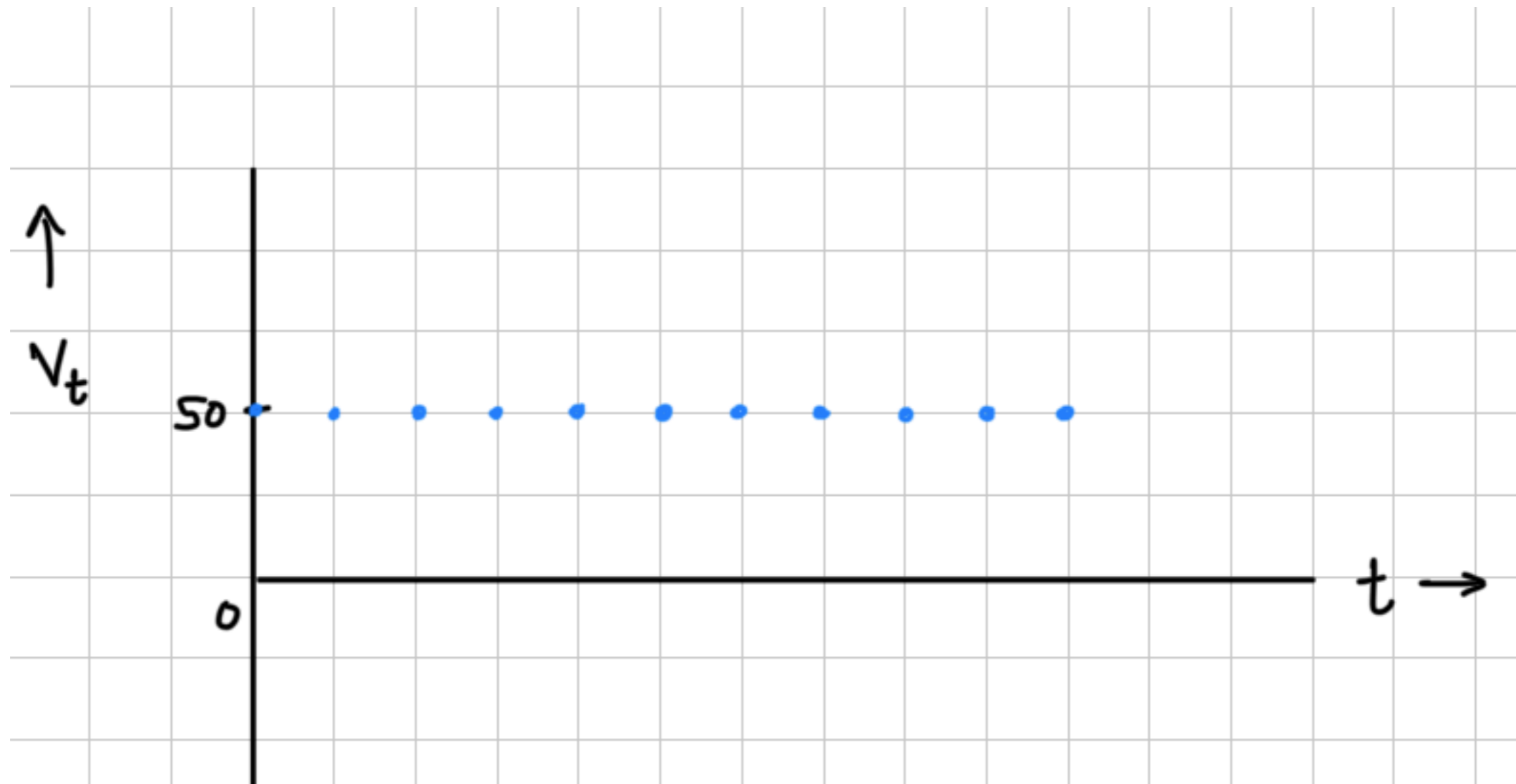
RECAP: ORDINARY DIFFERENTIAL EQUATIONS (ODEs)

$$\frac{dv_t}{dt} = 10 - 0.2v_t$$



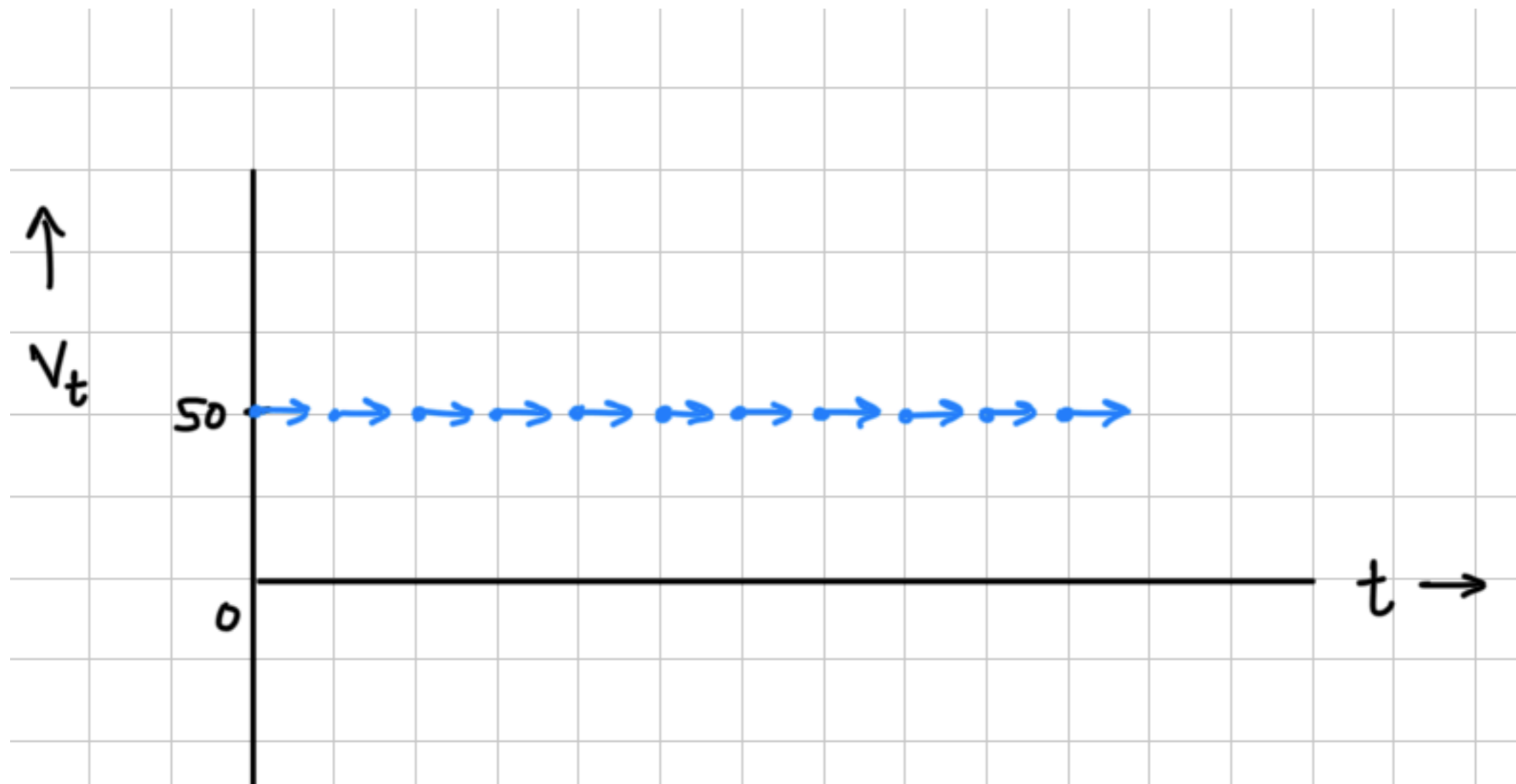
RECAP: ORDINARY DIFFERENTIAL EQUATIONS (ODEs)

$$\frac{dv_t}{dt} = 10 - 0.2 v_t$$



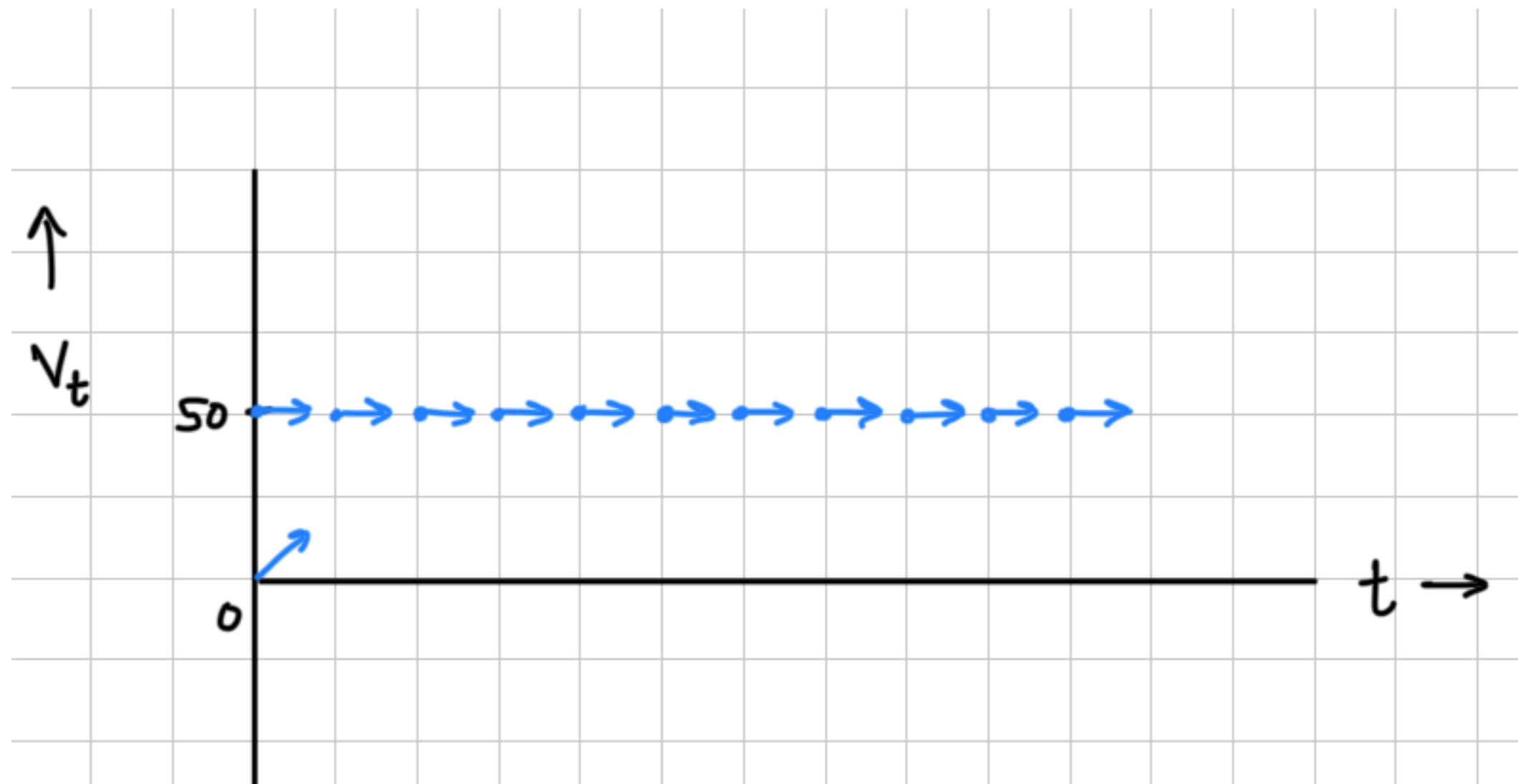
RECAP: ORDINARY DIFFERENTIAL EQUATIONS (ODEs)

$$\frac{dv_t}{dt} = 10 - 0.2v_t$$



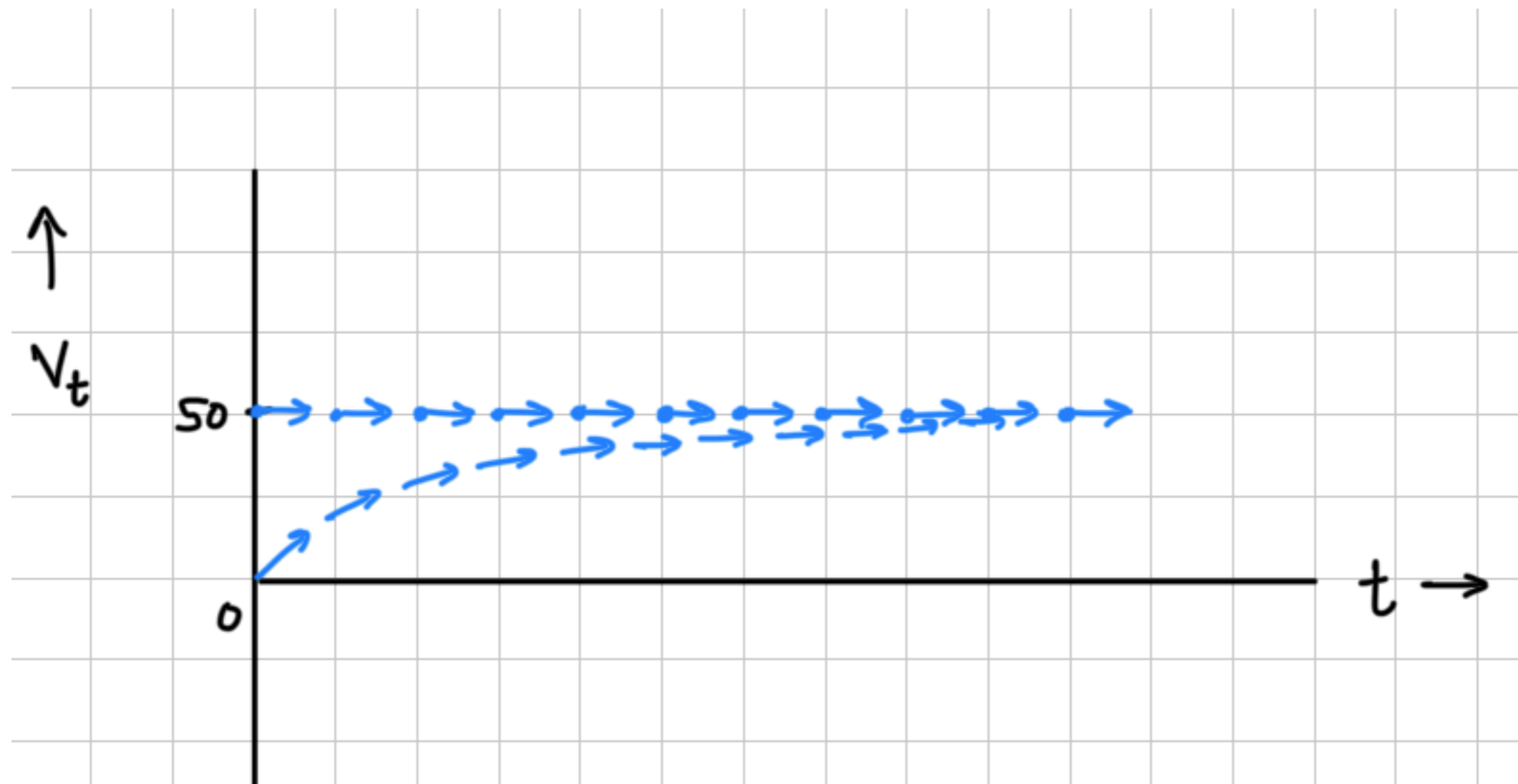
RECAP: ORDINARY DIFFERENTIAL EQUATIONS (ODEs)

$$\frac{dv_t}{dt} = 10 - 0.2v_t$$



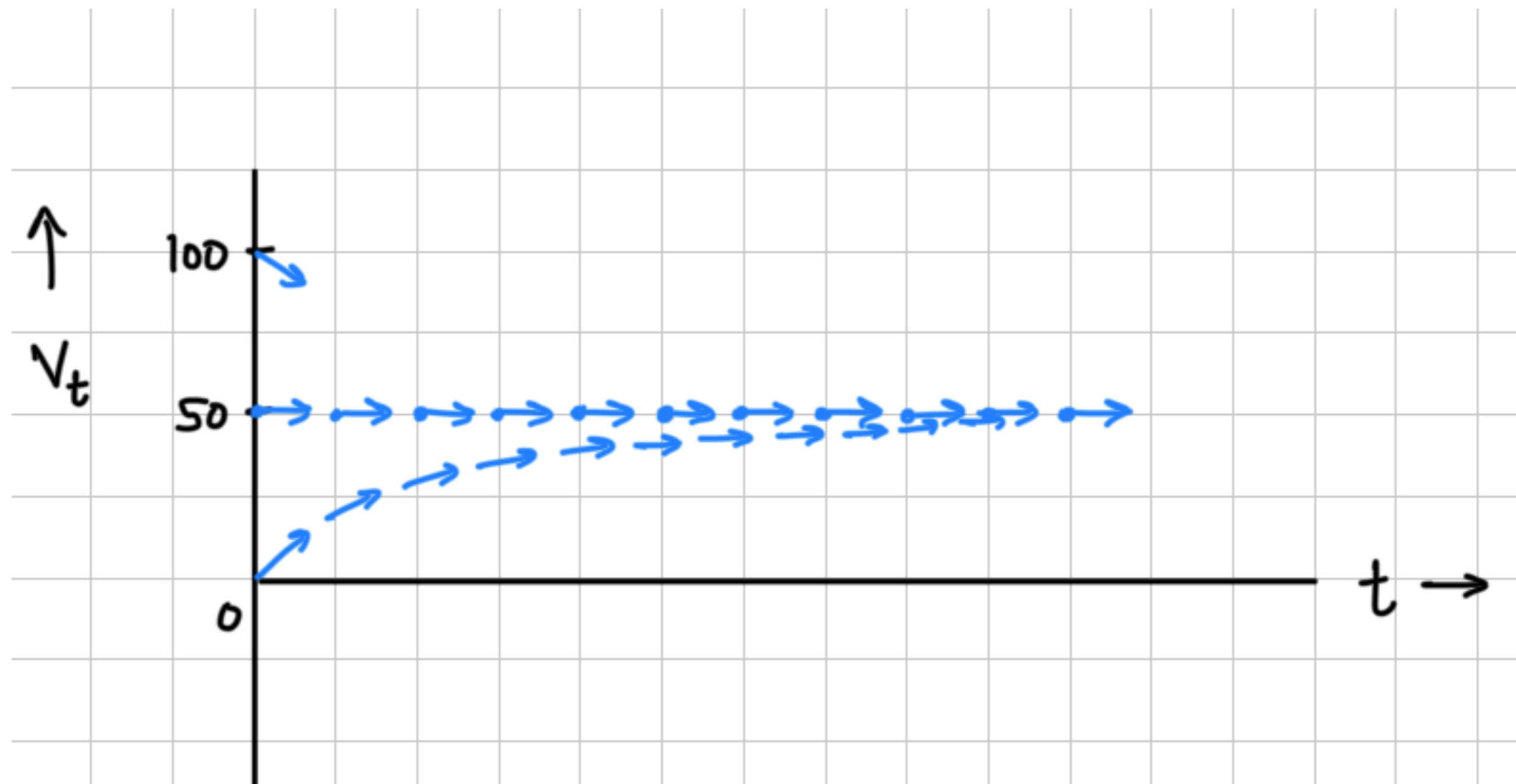
RECAP: ORDINARY DIFFERENTIAL EQUATIONS (ODEs)

$$\frac{dv_t}{dt} = 10 - 0.2v_t$$



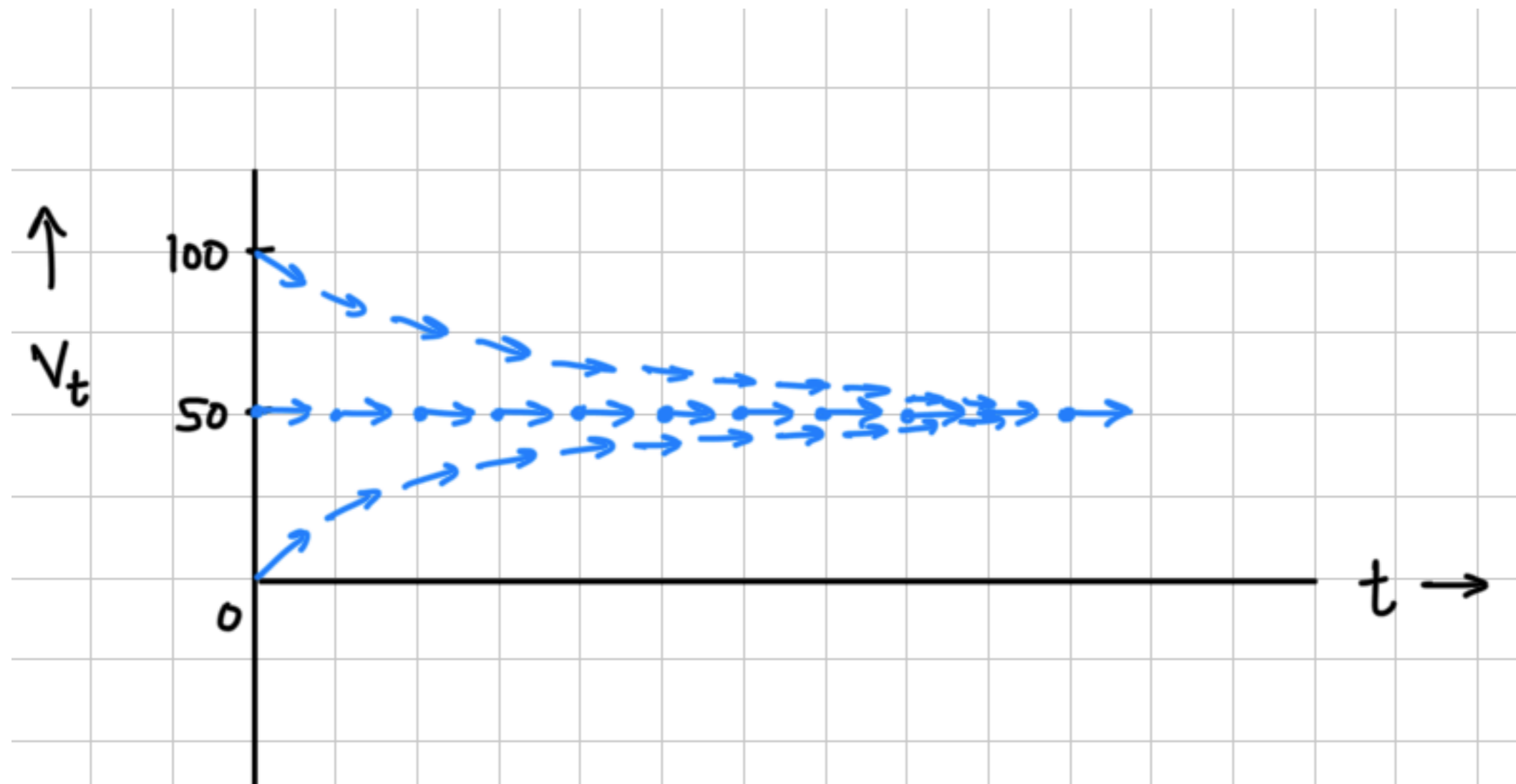
RECAP: ORDINARY DIFFERENTIAL EQUATIONS (ODEs)

$$\frac{dv_t}{dt} = 10 - 0.2 v_t$$



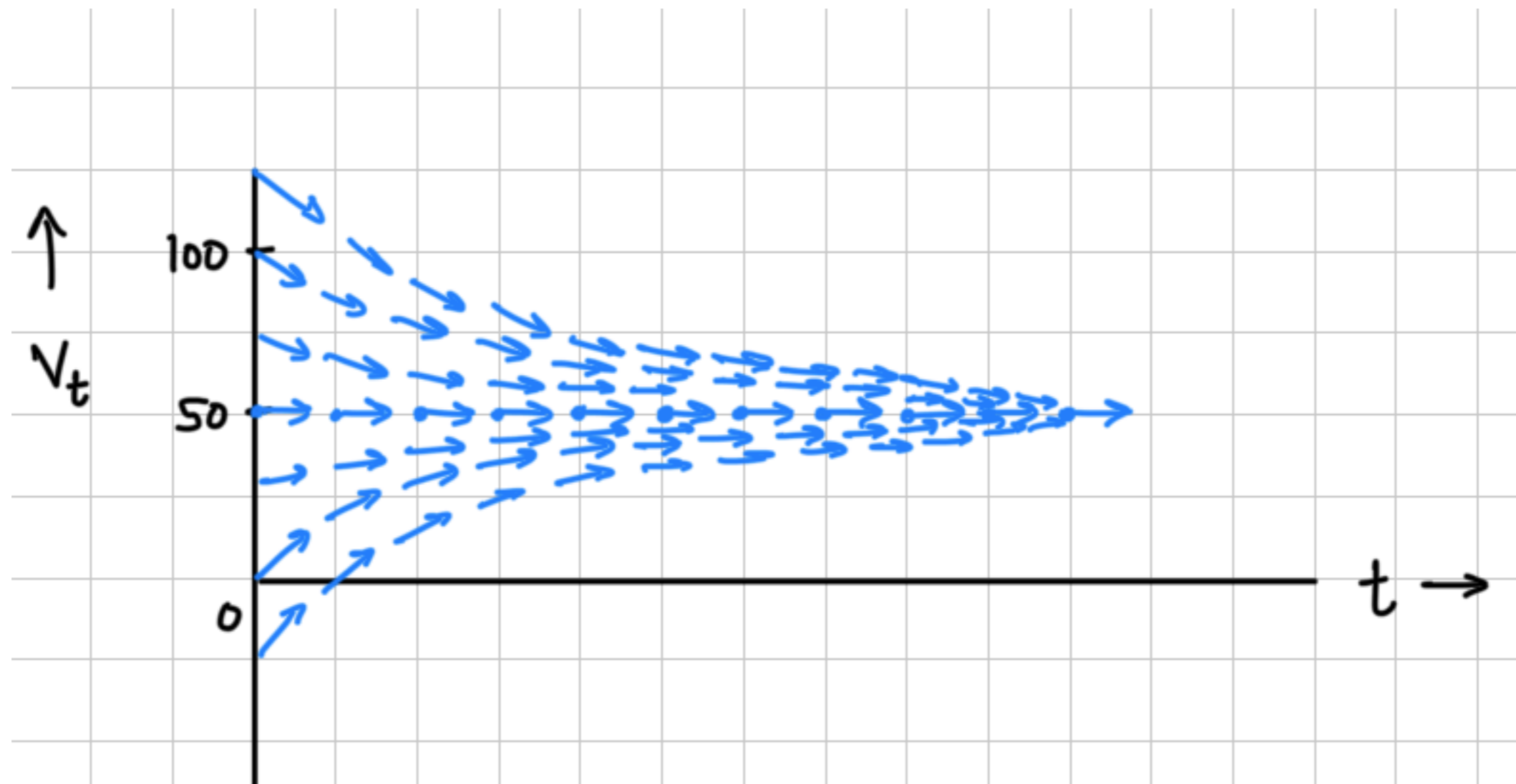
RECAP: ORDINARY DIFFERENTIAL EQUATIONS (ODEs)

$$\frac{dv_t}{dt} = 10 - 0.2 v_t$$



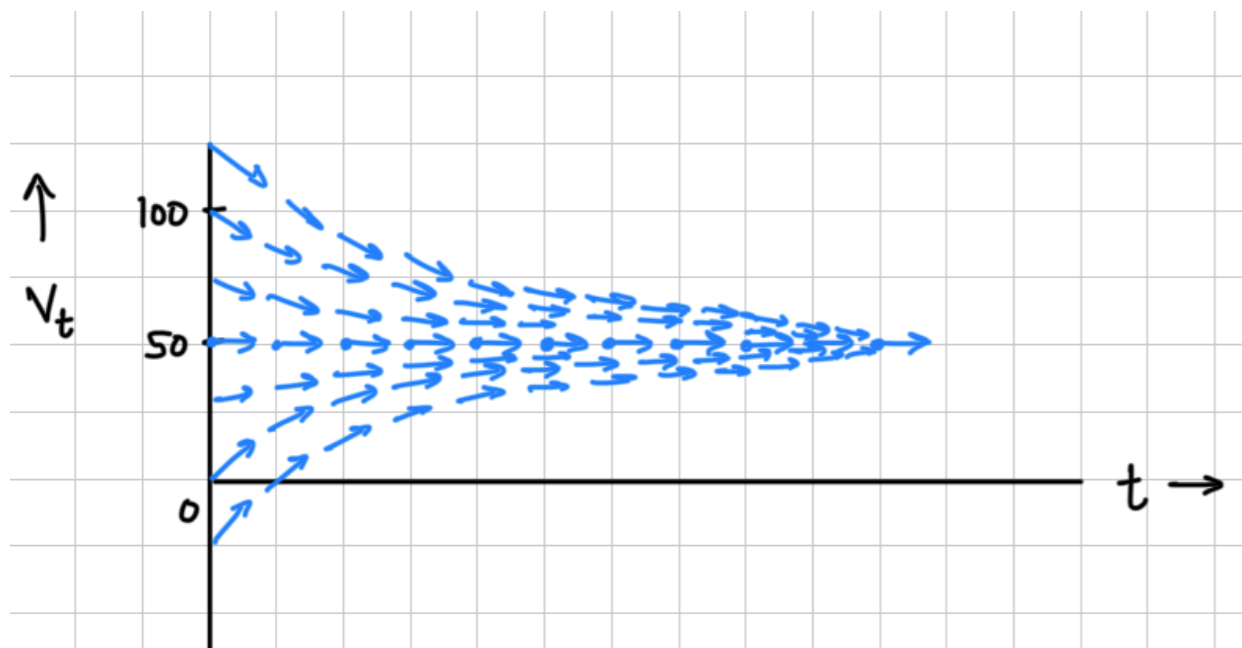
RECAP: ORDINARY DIFFERENTIAL EQUATIONS (ODEs)

$$\frac{dv_t}{dt} = 10 - 0.2v_t$$



RECAP: ORDINARY DIFFERENTIAL EQUATIONS (ODEs)

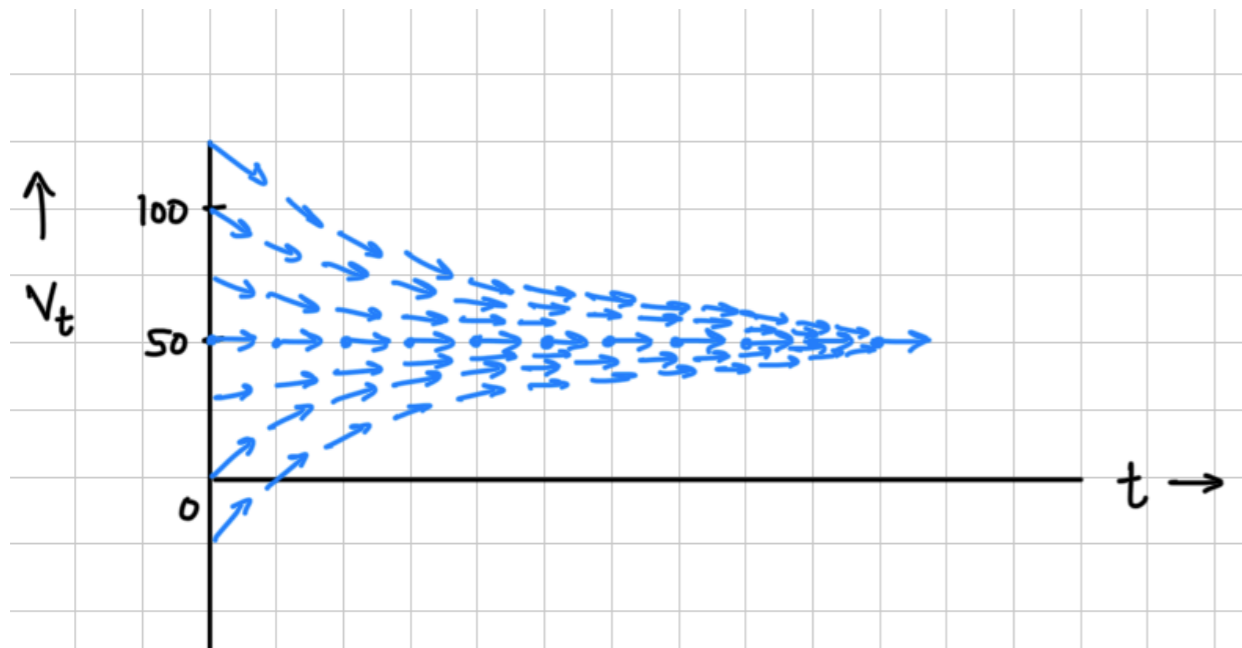
$$\frac{dv_t}{dt} = 10 - 0.2 v_t$$



RECAP: ORDINARY DIFFERENTIAL EQUATIONS (ODEs)

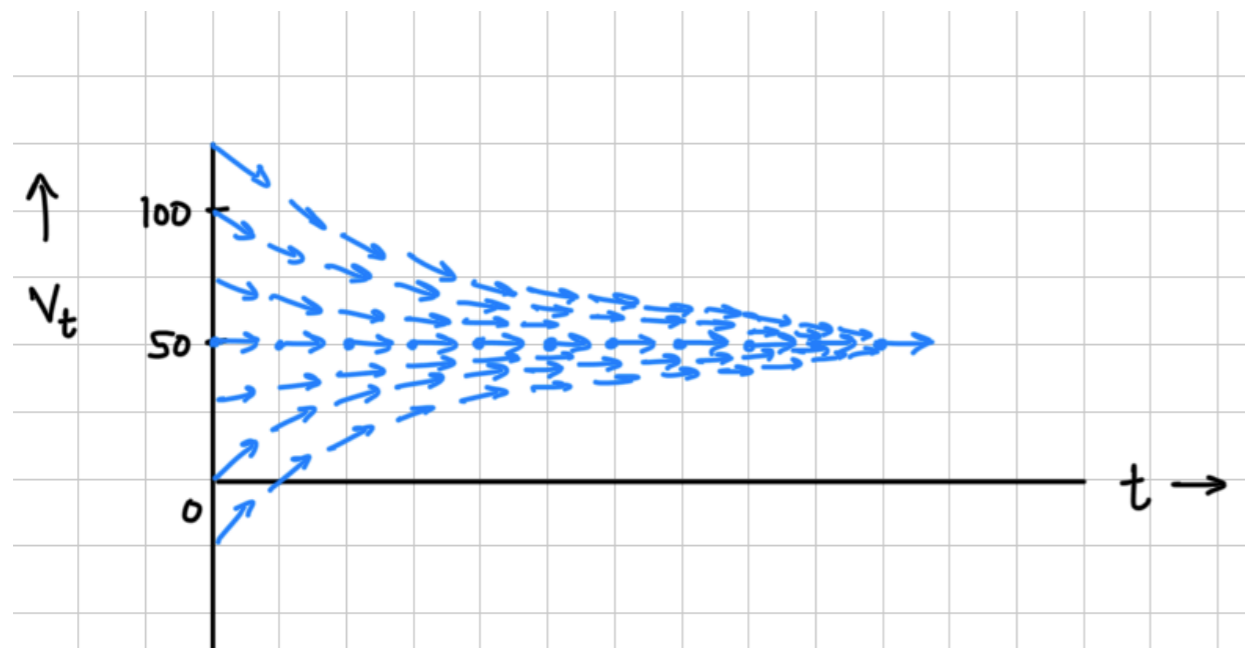
$$\frac{dv_t}{dt} = 10 - 0.2 v_t$$

$$\frac{dv_t}{dt} = 10 + 0.2 v_t$$

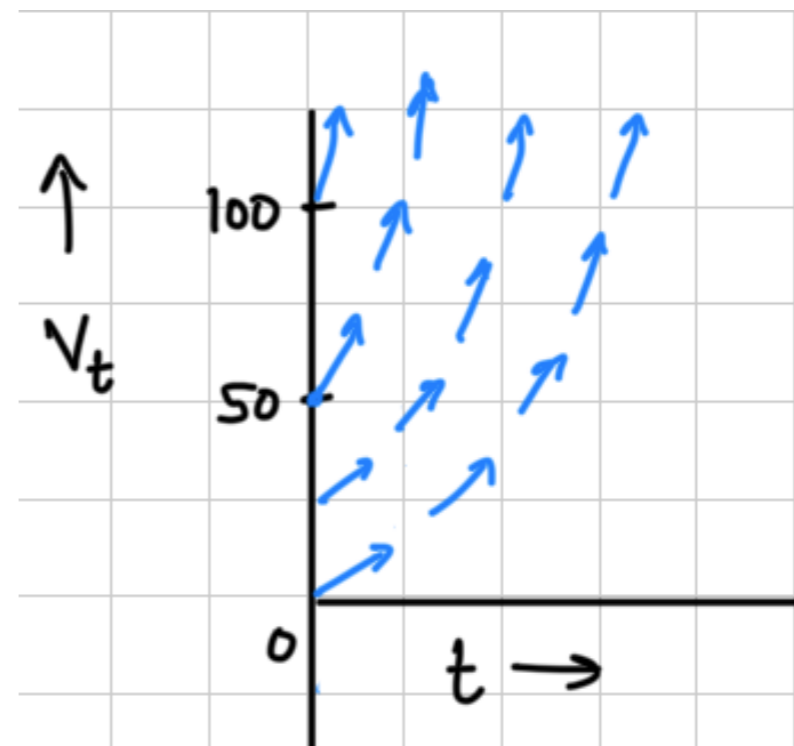


RECAP: ORDINARY DIFFERENTIAL EQUATIONS (ODEs)

$$\frac{dv_t}{dt} = 10 - 0.2 v_t$$

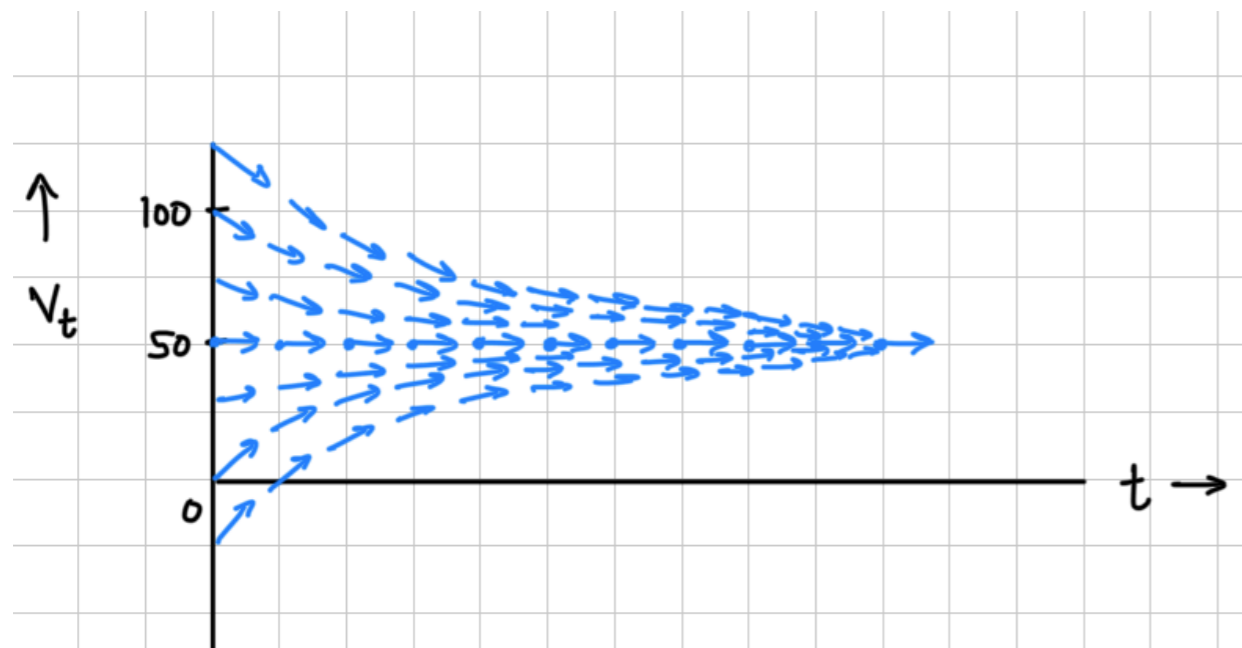


$$\frac{dv_t}{dt} = 10 + 0.2 v_t$$

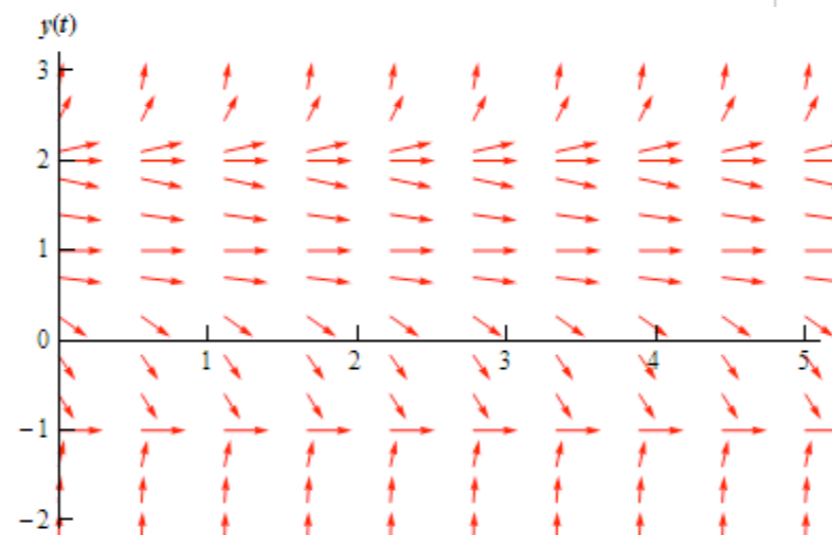
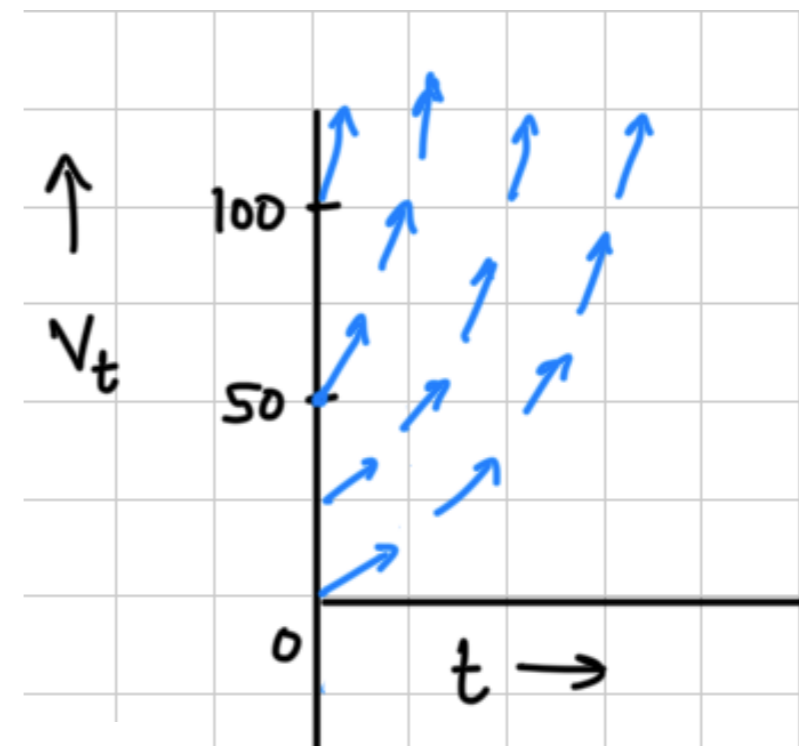


RECAP: ORDINARY DIFFERENTIAL EQUATIONS (ODEs)

$$\frac{dv_t}{dt} = 10 - 0.2 v_t$$



$$\frac{dv_t}{dt} = 10 + 0.2 v_t$$



CONDITION FOR STABILITY OF THE ODE $\frac{d\mathbf{w}_t}{dt} = \mathbf{b} + \mathbf{A}\mathbf{w}_t$

CONDITION FOR STABILITY OF THE ODE $\frac{d\mathbf{w}_t}{dt} = \mathbf{b} + \mathbf{A}\mathbf{w}_t$

All the eigenvalues of \mathbf{A} should have negative real parts.

CONDITION FOR STABILITY OF THE ODE $\frac{d\mathbf{w}_t}{dt} = \mathbf{b} + \mathbf{A}\mathbf{w}_t$

All the eigenvalues of \mathbf{A} should have negative real parts.

That is, \mathbf{A} is a Hurwitz matrix.

CONDITION FOR STABILITY OF THE ODE $\frac{d\mathbf{w}_t}{dt} = \mathbf{b} + \mathbf{A}\mathbf{w}_t$

All the eigenvalues of \mathbf{A} should have negative real parts.

That is, \mathbf{A} is a Hurwitz matrix.

Corollary of Khalil's (1996) Theorem 3.5

The ODE $\frac{d\mathbf{w}_t}{dt} = \mathbf{b} + \mathbf{A}\mathbf{w}_t$ has a globally stable equilibrium point \mathbf{w}^* such that $\mathbf{b} + \mathbf{A}\mathbf{w}^* = \mathbf{0}$ iff \mathbf{A} is Hurwitz.

CONDITION FOR STABILITY OF THE ODE $\frac{d\mathbf{w}_t}{dt} = \mathbf{b} + \mathbf{A}\mathbf{w}_t$

All the eigenvalues of \mathbf{A} should have negative real parts.

That is, \mathbf{A} is a Hurwitz matrix.

Corollary of Khalil's (1996) Theorem 3.5

The ODE $\frac{d\mathbf{w}_t}{dt} = \mathbf{b} + \mathbf{A}\mathbf{w}_t$ has a globally stable equilibrium point \mathbf{w}^* such that $\mathbf{b} + \mathbf{A}\mathbf{w}^* = \mathbf{0}$ iff \mathbf{A} is Hurwitz.

$$\frac{d\mathbf{w}_t}{dt} \propto \mathbf{b}_t + \mathbf{A}_t\mathbf{w}_t$$

CONDITION FOR STABILITY OF THE ODE $\frac{d\mathbf{w}_t}{dt} = \mathbf{b} + \mathbf{A}\mathbf{w}_t$

All the eigenvalues of \mathbf{A} should have negative real parts.

That is, \mathbf{A} is a Hurwitz matrix.

Corollary of Khalil's (1996) Theorem 3.5

The ODE $\frac{d\mathbf{w}_t}{dt} = \mathbf{b} + \mathbf{A}\mathbf{w}_t$ has a globally stable equilibrium point \mathbf{w}^* such that $\mathbf{b} + \mathbf{A}\mathbf{w}^* = \mathbf{0}$ iff \mathbf{A} is Hurwitz.

$$\frac{d\mathbf{w}_t}{dt} \propto \mathbf{b}_t + \mathbf{A}_t\mathbf{w}_t \quad \rightarrow$$

CONDITION FOR STABILITY OF THE ODE $\frac{d\mathbf{w}_t}{dt} = \mathbf{b} + \mathbf{A}\mathbf{w}_t$

All the eigenvalues of \mathbf{A} should have negative real parts.

That is, \mathbf{A} is a Hurwitz matrix.

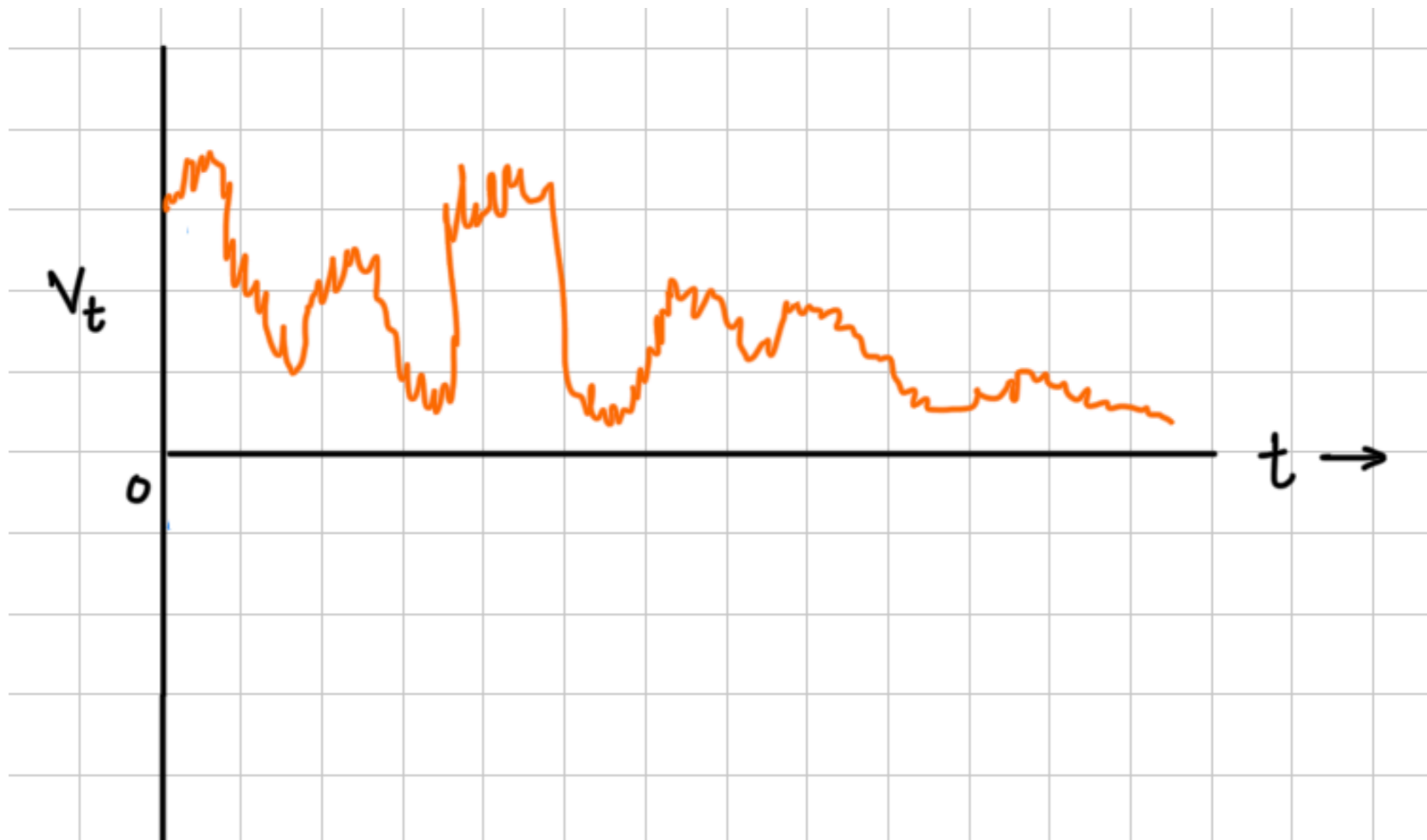
Corollary of Khalil's (1996) Theorem 3.5

The ODE $\frac{d\mathbf{w}_t}{dt} = \mathbf{b} + \mathbf{A}\mathbf{w}_t$ has a globally stable equilibrium point \mathbf{w}^* such that $\mathbf{b} + \mathbf{A}\mathbf{w}^* = \mathbf{0}$ iff \mathbf{A} is Hurwitz.

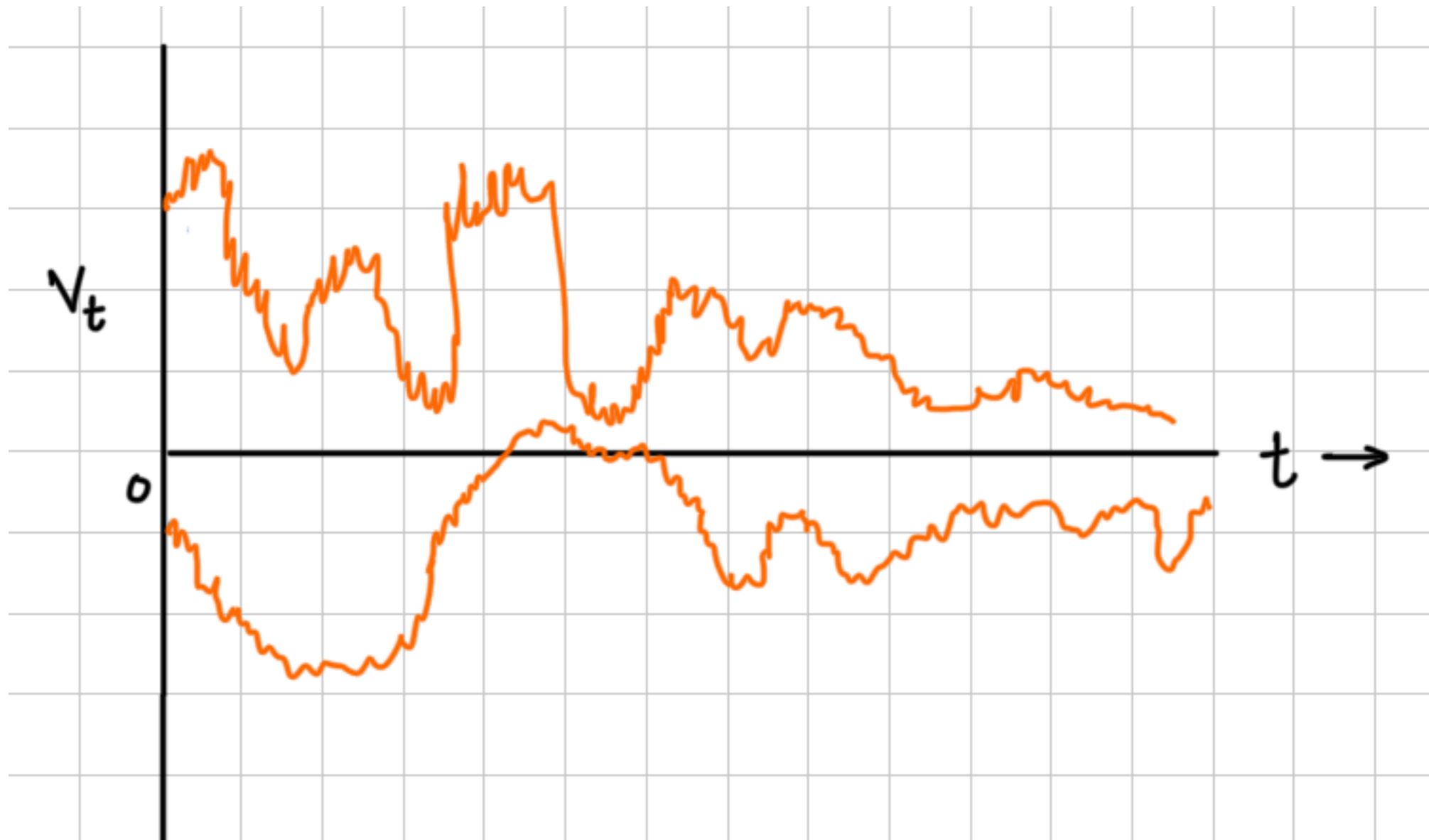
$$\frac{d\mathbf{w}_t}{dt} \propto \mathbf{b}_t + \mathbf{A}_t\mathbf{w}_t \quad \longrightarrow \quad \frac{d\mathbf{w}_t}{dt} = \mathbf{b} + \mathbf{A}\mathbf{w}_t$$

BEHAVIOR OF SAMPLE-BASED ALGORITHMS

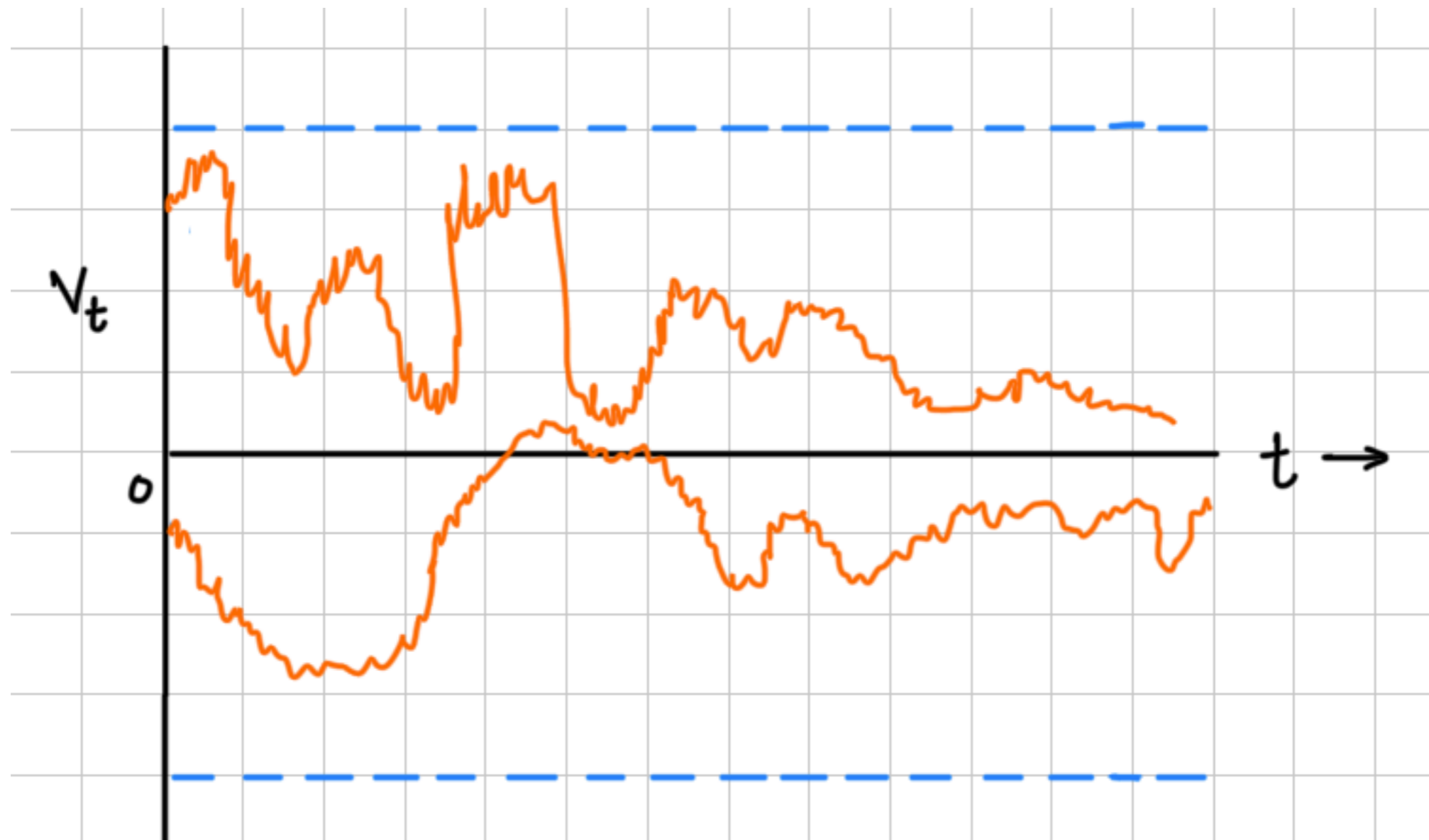
BEHAVIOR OF SAMPLE-BASED ALGORITHMS



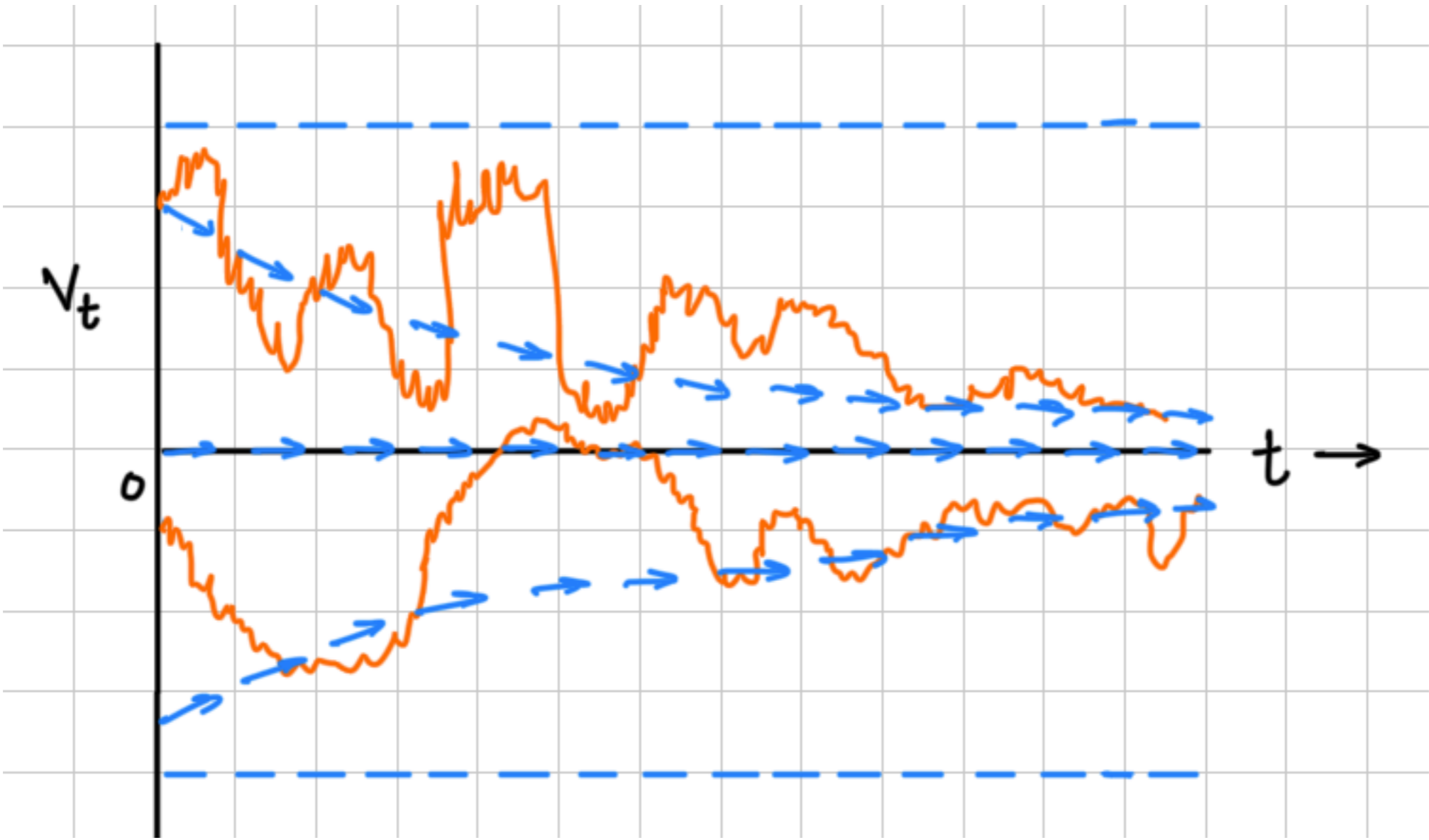
BEHAVIOR OF SAMPLE-BASED ALGORITHMS



BEHAVIOR OF SAMPLE-BASED ALGORITHMS



BEHAVIOR OF SAMPLE-BASED ALGORITHMS



(WHAT I'VE LEARNED ABOUT)

PROVING CONVERGENCE OF SAMPLED-BASED ALGORITHMS

(WHAT I'VE LEARNED ABOUT)

PROVING CONVERGENCE OF SAMPLED-BASED ALGORITHMS

1. Show that the sequence of iterates is bounded and asymptotically converges to the solutions of an ODE.

(WHAT I'VE LEARNED ABOUT)

PROVING CONVERGENCE OF SAMPLED-BASED ALGORITHMS

$\mathbf{w}_0 \ \mathbf{w}_1 \ \dots \ \mathbf{w}_t \ \dots$

1. Show that the sequence of iterates is bounded and asymptotically converges to the solutions of an ODE.

(WHAT I'VE LEARNED ABOUT)

PROVING CONVERGENCE OF SAMPLED-BASED ALGORITHMS

$$\underline{\mathbf{w}_0 \ \mathbf{w}_1 \ \dots \ \mathbf{w}_t \ \dots}$$

1. Show that the sequence of iterates is bounded and asymptotically converges to the solutions of an ODE.
2. Show the ODE has a globally stable equilibrium point.

**APPLYING THESE TECHNIQUES
TO PROVE THE CONVERGENCE OF OUR ALGORITHMS**

ANALYSIS OF ALGORITHM 1

ANALYSIS OF ALGORITHM 1

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} + \mathbf{w}_t^\top \mathbf{x}_t$

$$\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$$

ANALYSIS OF ALGORITHM 1

$$\left. \begin{aligned} \mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t \\ \bar{R}_{t+1} &\doteq \bar{R}_t + \eta \alpha_t \delta_t \end{aligned} \right\} \longrightarrow \mathbf{u}_{t+1} \doteq \mathbf{u}_t + \alpha_t [\mathbf{b}_t + \mathbf{A}_t \mathbf{u}_t]$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} + \mathbf{w}_t^\top \mathbf{x}_t$

$$\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$$

ANALYSIS OF ALGORITHM 1

$$\left. \begin{aligned} \mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t \\ \bar{R}_{t+1} &\doteq \bar{R}_t + \eta \alpha_t \delta_t \end{aligned} \right\} \longrightarrow$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} + \mathbf{w}_t^\top \mathbf{x}_t$

$$\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$$

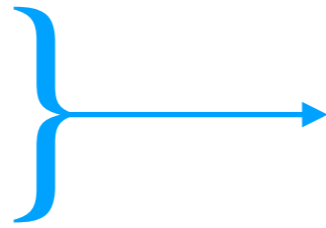
$$\mathbf{u}_{t+1} \doteq \mathbf{u}_t + \alpha_t [\mathbf{b}_t + \mathbf{A}_t \mathbf{u}_t]$$

$$\mathbf{b}_t \doteq \begin{bmatrix} \eta R_{t+1} \\ \mathbf{z}_t R_{t+1} \end{bmatrix}_{(d+1) \times 1}$$

$$\mathbf{A}_t \doteq \begin{bmatrix} -\eta & \eta(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \\ -\mathbf{z}_t & \mathbf{z}_t(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \end{bmatrix}_{(d+1) \times (d+1)}$$

ANALYSIS OF ALGORITHM 1

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t \\ \bar{R}_{t+1} &\doteq \bar{R}_t + \eta \alpha_t \delta_t\end{aligned}$$



$$\mathbf{u}_{t+1} \doteq \mathbf{u}_t + \alpha_t [\mathbf{b}_t + \mathbf{A}_t \mathbf{u}_t]$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} + \mathbf{w}_t^\top \mathbf{x}_t$

$$\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$$

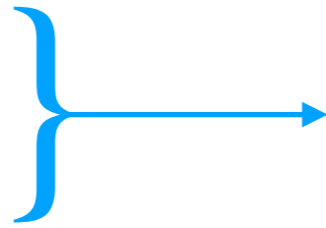
$$\mathbf{b}_t \doteq \begin{bmatrix} \eta R_{t+1} \\ \mathbf{z}_t R_{t+1} \end{bmatrix}_{(d+1) \times 1}$$

$$\mathbf{A}_t \doteq \begin{bmatrix} -\eta & \eta(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \\ -\mathbf{z}_t & \mathbf{z}_t(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \end{bmatrix}_{(d+1) \times (d+1)}$$

$$\mathbf{A} = \mathbb{E}[\mathbf{A}_t] \doteq \begin{bmatrix} -\eta & \mathbf{0}^\top \\ \frac{-1}{1-\lambda} \mathbf{X}^\top \mathbf{D}_\pi \mathbf{1} & \mathbf{X}^\top \mathbf{D}_\pi (\mathbf{P}_\pi^\lambda - \mathbb{I}) \mathbf{X} \end{bmatrix}$$

ANALYSIS OF ALGORITHM 1

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t \\ \bar{R}_{t+1} &\doteq \bar{R}_t + \eta \alpha_t \delta_t\end{aligned}$$



$$\mathbf{u}_{t+1} \doteq \mathbf{u}_t + \alpha_t [\mathbf{b}_t + \mathbf{A}_t \mathbf{u}_t]$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} + \mathbf{w}_t^\top \mathbf{x}_t$

$$\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$$

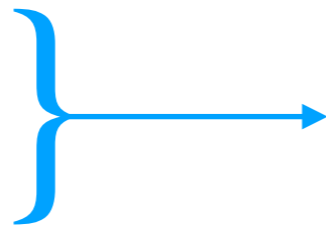
$$\mathbf{b}_t \doteq \begin{bmatrix} \eta R_{t+1} \\ \mathbf{z}_t^\top R_{t+1} \end{bmatrix}_{(d+1) \times 1}$$

$$\mathbf{A}_t \doteq \begin{bmatrix} -\eta & \eta(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \\ -\mathbf{z}_t & \mathbf{z}_t(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \end{bmatrix}_{(d+1) \times (d+1)}$$

$$\mathbf{A} = \mathbb{E}[\mathbf{A}_t] \doteq \begin{bmatrix} -\eta & \mathbf{0}^\top \\ \frac{-1}{1-\lambda} \mathbf{X}^\top \mathbf{D}_\pi \mathbf{1} & \mathbf{X}^\top \mathbf{D}_\pi (\mathbf{P}_\pi^\lambda - \mathbb{I}) \mathbf{X} \end{bmatrix} \text{ is Hurwitz.}$$

ANALYSIS OF ALGORITHM 1

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t \\ \bar{R}_{t+1} &\doteq \bar{R}_t + \eta \alpha_t \delta_t\end{aligned}$$



$$\mathbf{u}_{t+1} \doteq \mathbf{u}_t + \alpha_t [\mathbf{b}_t + \mathbf{A}_t \mathbf{u}_t]$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} + \mathbf{w}_t^\top \mathbf{x}_t$

$$\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$$

$$\mathbf{b}_t \doteq \begin{bmatrix} \eta R_{t+1} \\ \mathbf{z}_t R_{t+1} \end{bmatrix}_{(d+1) \times 1}$$

$$\mathbf{A}_t \doteq \begin{bmatrix} -\eta & \eta(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \\ -\mathbf{z}_t & \mathbf{z}_t(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \end{bmatrix}_{(d+1) \times (d+1)}$$

$$\mathbf{A} = \mathbb{E}[\mathbf{A}_t] \doteq \begin{bmatrix} -\eta & \mathbf{0}^\top \\ \frac{-1}{1-\lambda} \mathbf{X}^\top \mathbf{D}_\pi \mathbf{1} & \mathbf{X}^\top \mathbf{D}_\pi (\mathbf{P}_\pi^\lambda - \mathbb{I}) \mathbf{X} \end{bmatrix} \text{ is Hurwitz.}$$

(Tsitsiklis & Van Roy's (1999) Lemma 7)

ANALYSIS OF ALGORITHM 1

Theorem 1.2. *(Based on Tsitsiklis and Van Roy's (1999) Theorem 2) Consider the iterative algorithm of the form $\mathbf{u}_{t+1} \doteq \mathbf{u}_t + \alpha_t(\mathbf{b}(Y_t) + \mathbf{A}(Y_t)\mathbf{u}_t)$. Suppose the following conditions are satisfied:*

1. *The Markov chain $\{Y_t\}$ evolving in a state space \mathcal{Y} has a unique steady-state distribution. Let $\mathbb{E}_d[\cdot]$ denote the expectation according to this distribution.*
2. *Let $\mathbf{A} \doteq \mathbb{E}_d[\mathbf{A}(Y_t)]$. There exists a diagonal matrix \mathbf{L} with positive diagonal entries such that $\mathbf{L}\mathbf{A}$ is negative definite.*
3. *There exists a constant C such that $\|\mathbf{A}(Y)\| \leq C$ and $\|\mathbf{b}(Y)\| \leq C$ for any $Y \in \mathcal{Y}$.*
4. *There exist scalars C and $\rho \in (0, 1)$ such that $\forall t \geq 0$ and $Y_0 \in \mathcal{Y}$:*

$$\|\mathbb{E}[\mathbf{A}(Y_t) \mid Y_0] - \mathbf{A}\| \leq C\rho^t,$$

$$\|\mathbb{E}[\mathbf{b}(Y_t) \mid Y_0] - \mathbf{b}\| \leq C\rho^t, \quad \text{where, } \mathbf{b} \doteq \mathbb{E}_d[\mathbf{b}(Y_t)].$$

5. *The step sizes α_t are positive, deterministic, and satisfy $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$.*

Then \mathbf{u}_t converges to \mathbf{u}^ with probability one, where \mathbf{u}^* is the unique vector satisfying $\mathbf{A}\mathbf{u}^* + \mathbf{b} = 0$.*

ANALYSIS OF ALGORITHM 1

Theorem 1.2. (Based on Tsitsiklis and Van Roy's (1999) Theorem 2) Consider the iterative algorithm of the form $\mathbf{u}_{t+1} \doteq \mathbf{u}_t + \alpha_t(\mathbf{b}(Y_t) + \mathbf{A}(Y_t)\mathbf{u}_t)$. Suppose the following conditions are satisfied:

1. The Markov chain $\{Y_t\}$ evolving in a state space \mathcal{Y} has a unique steady-state distribution. Let $\mathbb{E}_d[\cdot]$ denote the expectation according to this distribution.
2. Let $\mathbf{A} \doteq \mathbb{E}_d[\mathbf{A}(Y_t)]$. There exists a diagonal matrix \mathbf{L} with positive diagonal entries such that $\mathbf{L}\mathbf{A}$ is negative definite.
3. There exists a constant C such that $\|\mathbf{A}(Y)\| \leq C$ and $\|\mathbf{b}(Y)\| \leq C$ for any $Y \in \mathcal{Y}$.
4. There exist scalars C and $\rho \in (0, 1)$ such that $\forall t \geq 0$ and $Y_0 \in \mathcal{Y}$:

$$\|\mathbb{E}[\mathbf{A}(Y_t) \mid Y_0] - \mathbf{A}\| \leq C\rho^t,$$

$$\|\mathbb{E}[\mathbf{b}(Y_t) \mid Y_0] - \mathbf{b}\| \leq C\rho^t, \quad \text{where, } \mathbf{b} \doteq \mathbb{E}_d[\mathbf{b}(Y_t)].$$

5. The step sizes α_t are positive, deterministic, and satisfy $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$.

Then \mathbf{u}_t converges to \mathbf{u}^* with probability one, where \mathbf{u}^* is the unique vector satisfying $\mathbf{A}\mathbf{u}^* + \mathbf{b} = 0$.

ANALYSIS OF ALGORITHM 1

Theorem 1.1. *Under Assumptions 1.1, 1.2, 1.3, on-policy linear Differential TD(λ)*

(Algorithm 1) converges for all $\lambda \in [0, 1)$ with probability one:

1. \bar{R} converges to the unique reward rate of the target policy $r(\pi)$.
2. \mathbf{w} converges to the unique solution, \mathbf{w}^* , of $\Pi T^\lambda(\mathbf{X}\mathbf{w}) = \mathbf{X}\mathbf{w}$.

The following error bound holds w.r.t. the centered differential value function \mathbf{v}_π :

$$\inf_{c \in \mathbb{R}} \|\mathbf{X}\mathbf{w}^* - (\mathbf{v}_\pi + c\mathbf{1})\|_{\mathbf{d}_\pi} \leq \frac{1}{\sqrt{(1 - \tau_\lambda^2)}} \inf_{c \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d} \|\mathbf{X}\mathbf{w} - (\mathbf{v}_\pi + c\mathbf{1})\|_{\mathbf{d}_\pi},$$

where τ_λ is a function of λ such that $\tau_\lambda \in [0, 1)$ and $\lim_{\lambda \rightarrow 1} \tau_\lambda = 0$;

EXTENSION TO THE OFF-POLICY SETTING

EXTENSION TO THE OFF-POLICY SETTING

One-step off-policy Differential TD

EXTENSION TO THE OFF-POLICY SETTING

One-step off-policy Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \rho_t \delta_t \mathbf{x}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t$

$$\rho_t \doteq \frac{\pi(A_t | S_t)}{b(A_t | S_t)}$$

EXTENSION TO THE OFF-POLICY SETTING

One-step off-policy Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \rho_t \delta_t \mathbf{x}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

Multi-step version?

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t$

$$\rho_t \doteq \frac{\pi(A_t | S_t)}{b(A_t | S_t)}$$

EXTENSION TO THE OFF-POLICY SETTING

One-step off-policy Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \rho_t \delta_t \mathbf{x}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t$

$$\rho_t \doteq \frac{\pi(A_t | S_t)}{b(A_t | S_t)}$$

Algorithm 1

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t$$

where $\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$

Multi-step version?

EXTENSION TO THE OFF-POLICY SETTING

One-step off-policy Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \rho_t \delta_t \mathbf{x}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t$

$$\rho_t \doteq \frac{\pi(A_t | S_t)}{b(A_t | S_t)}$$

Algorithm 1

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t$$

where $\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$

Multi-step version?



EXTENSION TO THE OFF-POLICY SETTING

One-step off-policy Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \rho_t \delta_t \mathbf{x}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t$

$$\rho_t \doteq \frac{\pi(A_t | S_t)}{b(A_t | S_t)}$$

Algorithm 1

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t$$

where $\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$

Multi-step version?



$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

where $\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$

EXTENSION TO THE OFF-POLICY SETTING

One-step off-policy Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \rho_t \delta_t \mathbf{x}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t$

$$\rho_t \doteq \frac{\pi(A_t | S_t)}{b(A_t | S_t)}$$

Algorithm 1

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t$$

where $\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$



Multi-step version?

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

where $\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$

Algorithm 1off

ANALYSIS OF (TABULAR) ALGORITHM 10FF'S "A" MATRIX

ANALYSIS OF (TABULAR) ALGORITHM 10FF'S "A" MATRIX

$$\mathbf{A}^1 \doteq \begin{bmatrix} -\eta & \mathbf{0}^\top \\ \frac{-1}{1-\lambda} \mathbf{D}_\pi \mathbf{1} & \mathbf{D}_\pi (\mathbf{P}_\pi^\lambda - \mathbb{I}) \end{bmatrix}$$

ANALYSIS OF (TABULAR) ALGORITHM 10FF'S "A" MATRIX

$$\mathbf{A}^1 \doteq \begin{bmatrix} -\eta & \mathbf{0}^\top \\ \frac{-1}{1-\lambda} \mathbf{D}_\pi \mathbf{1} & \mathbf{D}_\pi (\mathbf{P}_\pi^\lambda - \mathbb{I}) \end{bmatrix}$$

is Hurwitz.

(Tsitsiklis & Van Roy's
(1999) Lemma 7)

ANALYSIS OF (TABULAR) ALGORITHM 1 OFF'S "A" MATRIX

$$\mathbf{A}^1 \doteq \begin{bmatrix} -\eta & \mathbf{0}^\top \\ \frac{-1}{1-\lambda} \mathbf{D}_\pi \mathbf{1} & \mathbf{D}_\pi (\mathbf{P}_\pi^\lambda - \mathbb{I}) \end{bmatrix}$$

is Hurwitz.

(Tsitsiklis & Van Roy's
(1999) Lemma 7)

$$\mathbf{A}^{1off} \doteq \begin{bmatrix} -\eta & \eta \mathbf{d}_b^\top (\mathbf{P}_\pi - \mathbb{I}) \\ \frac{-1}{1-\lambda} \mathbf{D}_b \mathbf{1} & \mathbf{D}_b (\mathbf{P}_\pi^\lambda - \mathbb{I}) \end{bmatrix}$$

ANALYSIS OF (TABULAR) ALGORITHM 1OFF'S "A" MATRIX

$$\mathbf{A}^1 \doteq \begin{bmatrix} -\eta & \mathbf{0}^\top \\ \frac{-1}{1-\lambda} \mathbf{D}_\pi \mathbf{1} & \mathbf{D}_\pi (\mathbf{P}_\pi^\lambda - \mathbb{I}) \end{bmatrix}$$

is Hurwitz.

(Tsitsiklis & Van Roy's
(1999) Lemma 7)

$$\mathbf{A}^{1off} \doteq \begin{bmatrix} -\eta & \eta \mathbf{d}_b^\top (\mathbf{P}_\pi - \mathbb{I}) \\ \frac{-1}{1-\lambda} \mathbf{D}_b \mathbf{1} & \mathbf{D}_b (\mathbf{P}_\pi^\lambda - \mathbb{I}) \end{bmatrix}$$

is *not* Hurwitz.

(via a simulation analysis)

ANALYSIS OF (TABULAR) ALGORITHM 1OFF'S "A" MATRIX

$$\mathbf{A}^1 \doteq \begin{bmatrix} -\eta & \mathbf{0}^\top \\ \frac{-1}{1-\lambda} \mathbf{D}_\pi \mathbf{1} & \mathbf{D}_\pi (\mathbf{P}_\pi^\lambda - \mathbb{I}) \end{bmatrix} \quad \text{is Hurwitz.}$$

(Tsitsiklis & Van Roy's (1999) Lemma 7)

$$\mathbf{A}^{1off} \doteq \begin{bmatrix} -\eta & \eta \mathbf{d}_b^\top (\mathbf{P}_\pi - \mathbb{I}) \\ \frac{-1}{1-\lambda} \mathbf{D}_b \mathbf{1} & \mathbf{D}_b (\mathbf{P}_\pi^\lambda - \mathbb{I}) \end{bmatrix} \quad \text{is *not* Hurwitz.}$$

(via a simulation analysis)

So Algorithm 1off can diverge... :(

EXTENSION TO THE OFF-POLICY SETTING

EXTENSION TO THE OFF-POLICY SETTING

One-step off-policy Differential TD

EXTENSION TO THE OFF-POLICY SETTING

One-step off-policy Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \rho_t \delta_t \mathbf{x}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t$$

$$\rho_t \doteq \frac{\pi(A_t | S_t)}{b(A_t | S_t)}$$

EXTENSION TO THE OFF-POLICY SETTING

One-step off-policy Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \rho_t \delta_t \mathbf{x}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t$$

$$\rho_t \doteq \frac{\pi(A_t | S_t)}{b(A_t | S_t)}$$

Algorithm 1

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t$$

where $\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$

EXTENSION TO THE OFF-POLICY SETTING

One-step off-policy Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \rho_t \delta_t \mathbf{x}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t$$

$$\rho_t \doteq \frac{\pi(A_t | S_t)}{b(A_t | S_t)}$$

Algorithm 1

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t$$

where $\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

where $\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$

Algorithm 1off

EXTENSION TO THE OFF-POLICY SETTING

One-step off-policy Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \rho_t \delta_t \mathbf{x}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t$$

$$\rho_t \doteq \frac{\pi(A_t | S_t)}{b(A_t | S_t)}$$

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t z_t^{\bar{R}}$$

where $\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$

$$z_t^{\bar{R}} \doteq \rho_t (\lambda z_{t-1}^{\bar{R}} + 1)$$

Algorithm 1

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t$$

where $\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

where $\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$

Algorithm 1off

EXTENSION TO THE OFF-POLICY SETTING

One-step off-policy Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \rho_t \delta_t \mathbf{x}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t$$

$$\rho_t \doteq \frac{\pi(A_t | S_t)}{b(A_t | S_t)}$$

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t z_t^{\bar{R}}$$

where $\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$

$$z_t^{\bar{R}} \doteq \rho_t (\lambda z_{t-1}^{\bar{R}} + 1)$$

Algorithm 2

Algorithm 1

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t$$

where $\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

where $\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$

Algorithm 1off

ANALYSIS OF (TABULAR) ALGORITHM 2

ANALYSIS OF (TABULAR) ALGORITHM 2

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t z_t^{\bar{R}}$$

where $\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$

$$z_t^{\bar{R}} \doteq \rho_t (\lambda z_{t-1}^{\bar{R}} + 1)$$

ANALYSIS OF (TABULAR) ALGORITHM 2

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t z_t^{\bar{R}}$$

where $\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$

$$z_t^{\bar{R}} \doteq \rho_t (\lambda z_{t-1}^{\bar{R}} + 1)$$

$$\bar{R}_t = f(\hat{\mathbf{v}}_t)$$

ANALYSIS OF (TABULAR) ALGORITHM 2

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t z_t^{\bar{R}}$$

where

$$\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$$

$$z_t^{\bar{R}} \doteq \rho_t (\lambda z_{t-1}^{\bar{R}} + 1)$$

$$\bar{R}_t = f(\hat{\mathbf{v}}_t)$$

$$f(\hat{\mathbf{v}}_t) = \eta \mathbf{g}^\top \mathbf{v}_t$$

ANALYSIS OF (TABULAR) ALGORITHM 2

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t z_t^{\bar{R}}$$

where $\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$

$$z_t^{\bar{R}} \doteq \rho_t (\lambda z_{t-1}^{\bar{R}} + 1)$$

$$\bar{R}_t = f(\hat{\mathbf{v}}_t)$$

$$f(\hat{\mathbf{v}}_t) = \eta \mathbf{g}^\top \mathbf{v}_t$$

$$\mathbf{A} = \mathbf{D}_b (\mathbf{P}_\pi^\lambda - \mathbb{I} - \frac{\eta}{1-\lambda} \mathbf{1} \mathbf{g}^\top)$$

ANALYSIS OF (TABULAR) ALGORITHM 2

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t z_t^{\bar{R}}$$

where $\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$

$$z_t^{\bar{R}} \doteq \rho_t (\lambda z_{t-1}^{\bar{R}} + 1)$$

$$\bar{R}_t = f(\hat{\mathbf{v}}_t)$$

$$f(\hat{\mathbf{v}}_t) = \eta \mathbf{g}^\top \mathbf{v}_t \quad \eta > 0$$

\mathbf{g} is a non-negative vector
with at least one positive element

$$\mathbf{A} = \mathbf{D}_b(\mathbf{P}_\pi^\lambda - \mathbb{I} - \frac{\eta}{1-\lambda} \mathbf{1} \mathbf{g}^\top)$$

ANALYSIS OF (TABULAR) ALGORITHM 2

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t z_t^{\bar{R}}$$

where $\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$

$$z_t^{\bar{R}} \doteq \rho_t (\lambda z_{t-1}^{\bar{R}} + 1)$$

$$\bar{R}_t = f(\hat{\mathbf{v}}_t)$$

$$f(\hat{\mathbf{v}}_t) = \eta \mathbf{g}^\top \mathbf{v}_t \quad \eta > 0$$

\mathbf{g} is a non-negative vector
with at least one positive element

$$\mathbf{A} = \mathbf{D}_b(\mathbf{P}_\pi^\lambda - \mathbb{I} - \frac{\eta}{1-\lambda} \mathbf{1g}^\top) \text{ is Hurwitz!}$$

ANALYSIS OF (TABULAR) ALGORITHM 2

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t z_t^{\bar{R}}$$

where

$$\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$$

$$z_t^{\bar{R}} \doteq \rho_t (\lambda z_{t-1}^{\bar{R}} + 1)$$

$$\bar{R}_t = f(\hat{\mathbf{v}}_t)$$

$$f(\hat{\mathbf{v}}_t) = \eta \mathbf{g}^\top \mathbf{v}_t \quad \eta > 0$$

\mathbf{g} is a non-negative vector
with at least one positive element

$$\mathbf{A} = \mathbf{D}_b \left(\mathbf{P}_\pi^\lambda - \mathbb{1} - \frac{\eta}{1 - \lambda} \mathbf{1} \mathbf{g}^\top \right) \text{ is Hurwitz!}$$

For any $\lambda > 0$ there exist $\eta > 0$ such that \mathbf{A} is Hurwitz.

(using the Perron–Frobenius theorem for irreducible non-negative matrices)

ANALYSIS OF ALGORITHM 2

Theorem 1.4 (Based on Borkar's (2009: Chapter 6) Theorem 9 and Corollary 8).

Consider an iterative algorithm of the form: $\mathbf{v}_{n+1} \doteq \mathbf{v}_n + \alpha_n [h(\mathbf{v}_n, Y_n) + \mathbf{m}_{n+1}]$.

Suppose the following conditions are satisfied:

1. The process $\{Y_t\}$ is a weak Feller Markov chain in a compact state space \mathcal{Y} and has a unique invariant probability measure d .
2. The function $h(\mathbf{v}, y)$ is jointly continuous in (\mathbf{v}, y) and is Lipschitz in \mathbf{v} uniformly w.r.t. $y \in \mathcal{Y}$.
3. Define $\tilde{h}(\mathbf{v}) \doteq \mathbb{E}_d[h(\mathbf{v}, Y)]$. The limit $\hat{h}(\mathbf{v}) \doteq \lim_{c \rightarrow \infty} \tilde{h}(c\mathbf{v})/c$ exists uniformly on compact subsets of \mathbf{v} . The ODE $\dot{\mathbf{u}} = \hat{h}(\mathbf{u})$ is well posed and has the origin as the unique globally asymptotically stable solution.
4. The sequence $\{\mathbf{m}_{t+1}\}$ is a martingale difference sequence w.r.t. the increasing σ -fields $\mathcal{F}_t \doteq \sigma(\mathbf{v}_k, Y_k, \mathbf{m}_k, k \leq t), t \geq 0$ (that is, $\mathbb{E}[\|\mathbf{m}_{t+1}\| \mid \mathcal{F}_t] < \infty$ and $\mathbb{E}[\mathbf{m}_{t+1} \mid \mathcal{F}_t] = 0$ almost surely, $\forall t \geq 0$), and $\mathbb{E}[\|\mathbf{m}_{t+1}\|^2 \mid \mathcal{F}_t] < K(1 + \|\mathbf{v}_t\|^2)$ almost surely, $\forall t \geq 0$, for some constant $K > 0$.
5. The step sizes $\{\alpha_t\}$ are positive with $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$.

Then,

- (i) the algorithm is stable, that is, $\sup_t \|\mathbf{v}_t\| < \infty$, almost surely,
- (ii) the algorithm converges almost surely to a compact internally chain transitive invariant set of the ODE $\dot{\mathbf{u}} = \hat{h}(\mathbf{u})$.

ANALYSIS OF ALGORITHM 2

Theorem 1.4 (Based on [Borkar's \(2009: Chapter 6\) Theorem 9 and Corollary 8](#)).

Consider an iterative algorithm of the form: $\mathbf{v}_{n+1} \doteq \mathbf{v}_n + \alpha_n [h(\mathbf{v}_n, Y_n) + \mathbf{m}_{n+1}]$.

Suppose the following conditions are satisfied:

1. The process $\{Y_t\}$ is a weak Feller Markov chain in a compact state space \mathcal{Y} and has a unique invariant probability measure d .
2. The function $h(\mathbf{v}, y)$ is jointly continuous in (\mathbf{v}, y) and is Lipschitz in \mathbf{v} uniformly w.r.t. $y \in \mathcal{Y}$.
3. Define $\tilde{h}(\mathbf{v}) \doteq \mathbb{E}_d[h(\mathbf{v}, Y)]$. The limit $\hat{h}(\mathbf{v}) \doteq \lim_{c \rightarrow \infty} \tilde{h}(c\mathbf{v})/c$ exists uniformly on compact subsets of \mathbf{v} . The ODE $\dot{\mathbf{u}} = \hat{h}(\mathbf{u})$ is well posed and has the origin as the unique globally asymptotically stable solution.
4. The sequence $\{\mathbf{m}_{t+1}\}$ is a martingale difference sequence w.r.t. the increasing σ -fields $\mathcal{F}_t \doteq \sigma(\mathbf{v}_k, Y_k, \mathbf{m}_k, k \leq t), t \geq 0$ (that is, $\mathbb{E}[\|\mathbf{m}_{t+1}\| \mid \mathcal{F}_t] < \infty$ and $\mathbb{E}[\mathbf{m}_{t+1} \mid \mathcal{F}_t] = 0$ almost surely, $\forall t \geq 0$), and $\mathbb{E}[\|\mathbf{m}_{t+1}\|^2 \mid \mathcal{F}_t] < K(1 + \|\mathbf{v}_t\|^2)$ almost surely, $\forall t \geq 0$, for some constant $K > 0$.
5. The step sizes $\{\alpha_t\}$ are positive with $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$.

Then,

- (i) the algorithm is stable, that is, $\sup_t \|\mathbf{v}_t\| < \infty$, almost surely,
- (ii) the algorithm converges almost surely to a compact internally chain transitive invariant set of the ODE $\dot{\mathbf{u}} = \hat{h}(\mathbf{u})$.

ANALYSIS OF ALGORITHM 2

Theorem 1.4 (Based on [Borkar's \(2009: Chapter 6\) Theorem 9 and Corollary 8](#)).

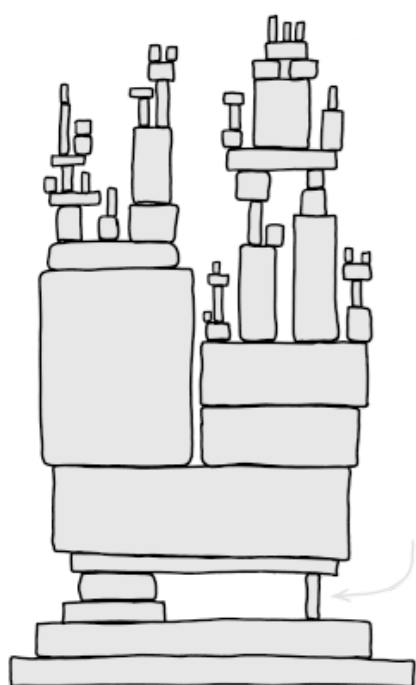
Consider an iterative algorithm of the form: $\mathbf{v}_{n+1} \doteq \mathbf{v}_n + \alpha_n [h(\mathbf{v}_t, Y_t) + \mathbf{m}_{t+1}]$.

Suppose the following conditions are satisfied:

1. The process $\{Y_t\}$ is a weak Feller Markov chain in a compact state space \mathcal{Y} and has a unique invariant probability measure d .
2. The function $h(\mathbf{v}, y)$ is jointly continuous in (\mathbf{v}, y) and is Lipschitz in \mathbf{v} uniformly w.r.t. $y \in \mathcal{Y}$.
3. Define $\tilde{h}(\mathbf{v}) \doteq \mathbb{E}_d[h(\mathbf{v}, Y)]$. The limit $\hat{h}(\mathbf{v}) \doteq \lim_{c \rightarrow \infty} \tilde{h}(c\mathbf{v})/c$ exists uniformly on compact subsets of \mathbf{v} . The ODE $\dot{\mathbf{u}} = \hat{h}(\mathbf{u})$ is well posed and has the origin as the unique globally asymptotically stable solution.
4. The sequence $\{\mathbf{m}_{t+1}\}$ is a martingale difference sequence w.r.t. the increasing σ -fields $\mathcal{F}_t \doteq \sigma(\mathbf{v}_k, Y_k, \mathbf{m}_k, k \leq t), t \geq 0$ (that is, $\mathbb{E}[\|\mathbf{m}_{t+1}\| \mid \mathcal{F}_t] < \infty$ and $\mathbb{E}[\mathbf{m}_{t+1} \mid \mathcal{F}_t] = 0$ almost surely, $\forall t \geq 0$), and $\mathbb{E}[\|\mathbf{m}_{t+1}\|^2 \mid \mathcal{F}_t] < K(1 + \|\mathbf{v}_t\|^2)$ almost surely, $\forall t \geq 0$, for some constant $K > 0$.
5. The step sizes $\{\alpha_t\}$ are positive with $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$.

Then,

- (i) the algorithm is stable, that is, $\sup_t \|\mathbf{v}_t\| < \infty$, almost surely,
- (ii) the algorithm converges almost surely to a compact internally chain transitive invariant set of the ODE $\dot{\mathbf{u}} = \hat{h}(\mathbf{u})$.



xkcd.com/2347

ANALYSIS OF ALGORITHM 2

Theorem 1.4 (Based on [Borkar's \(2009: Chapter 6\) Theorem 9 and Corollary 8](#)).

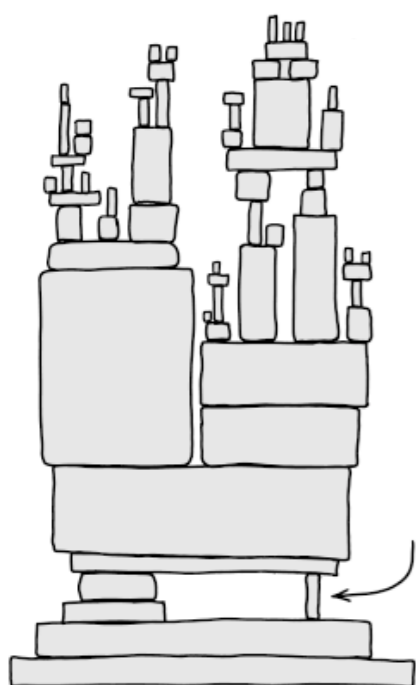
Consider an iterative algorithm of the form: $\mathbf{v}_{n+1} \doteq \mathbf{v}_n + \alpha_n [h(\mathbf{v}_t, Y_t) + \mathbf{m}_{t+1}]$.

Suppose the following conditions are satisfied:

1. The process $\{Y_t\}$ is a weak Feller Markov chain in a compact state space \mathcal{Y} and has a unique invariant probability measure d .
2. The function $h(\mathbf{v}, y)$ is jointly continuous in (\mathbf{v}, y) and is Lipschitz in \mathbf{v} uniformly w.r.t. $y \in \mathcal{Y}$.
3. Define $\tilde{h}(\mathbf{v}) \doteq \mathbb{E}_d[h(\mathbf{v}, Y)]$. The limit $\hat{h}(\mathbf{v}) \doteq \lim_{c \rightarrow \infty} \tilde{h}(c\mathbf{v})/c$ exists uniformly on compact subsets of \mathbf{v} . The ODE $\dot{\mathbf{u}} = \hat{h}(\mathbf{u})$ is well posed and has the origin as the unique globally asymptotically stable solution.
4. The sequence $\{\mathbf{m}_{t+1}\}$ is a martingale difference sequence w.r.t. the increasing σ -fields $\mathcal{F}_t \doteq \sigma(\mathbf{v}_k, Y_k, \mathbf{m}_k, k \leq t), t \geq 0$ (that is, $\mathbb{E}[\|\mathbf{m}_{t+1}\| \mid \mathcal{F}_t] < \infty$ and $\mathbb{E}[\mathbf{m}_{t+1} \mid \mathcal{F}_t] = 0$ almost surely, $\forall t \geq 0$), and $\mathbb{E}[\|\mathbf{m}_{t+1}\|^2 \mid \mathcal{F}_t] < K(1 + \|\mathbf{v}_t\|^2)$ almost surely, $\forall t \geq 0$, for some constant $K > 0$.
5. The step sizes $\{\alpha_t\}$ are positive with $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$.

Then,

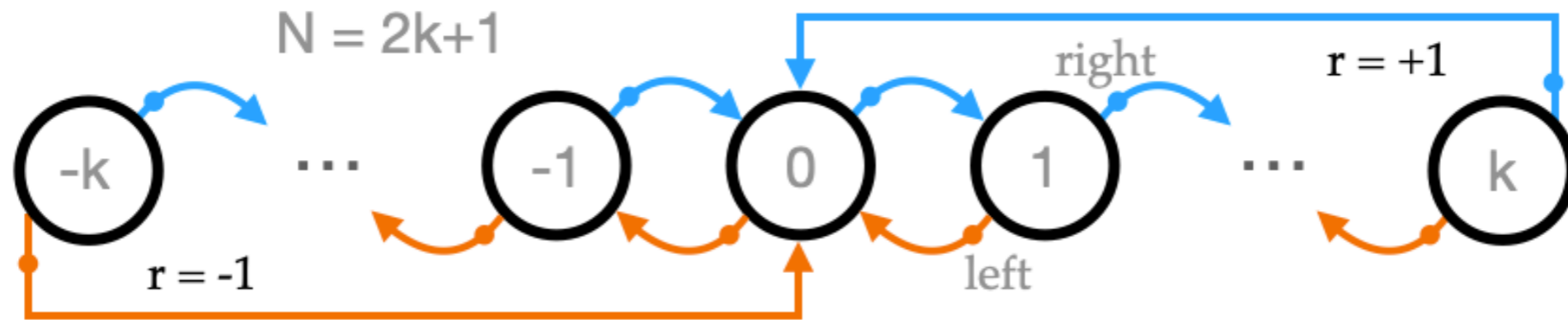
- (i) the algorithm is stable, that is, $\sup_t \|\mathbf{v}_t\| < \infty$, almost surely,
- (ii) the algorithm converges almost surely to a compact internally chain transitive invariant set of the ODE $\dot{\mathbf{u}} = \hat{h}(\mathbf{u})$.



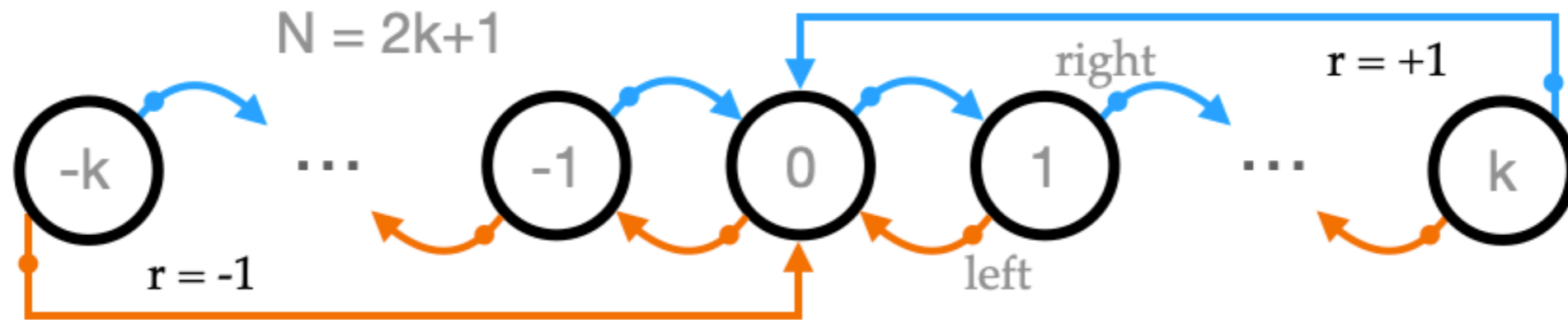
xkcd.com/2347

EXPERIMENTAL ANALYSIS

EXPERIMENTAL ANALYSIS

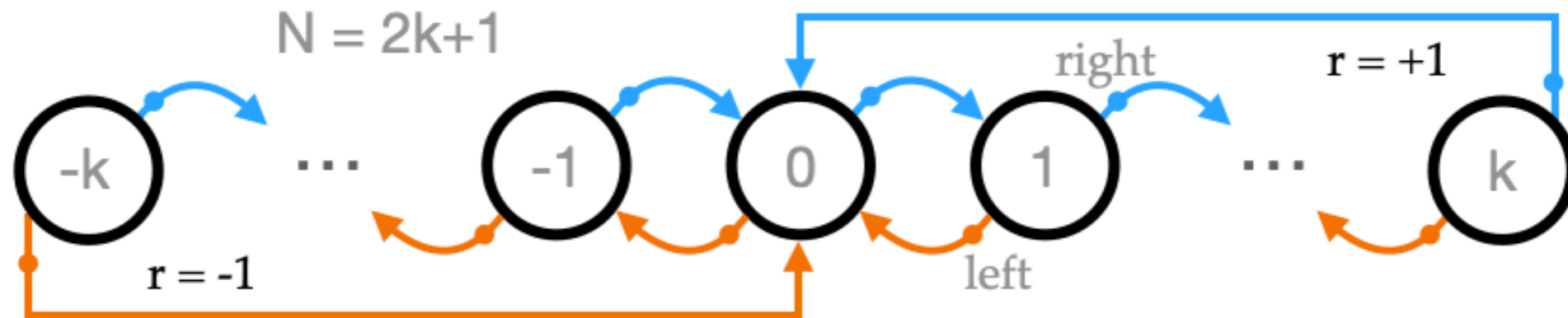


EXPERIMENTAL ANALYSIS



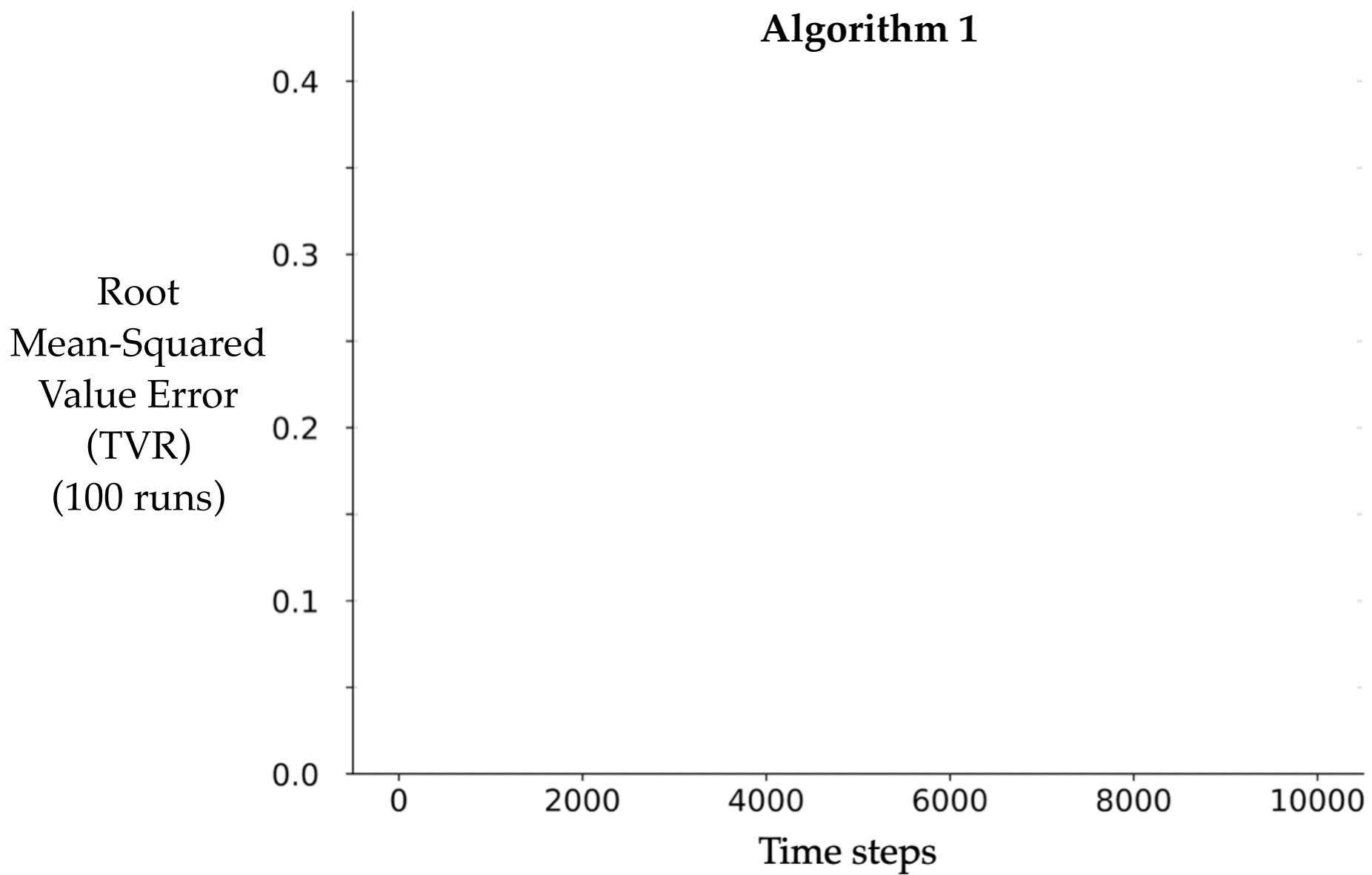
		left	right
Target policy	π	0.5	0.5

EXPERIMENTAL ANALYSIS

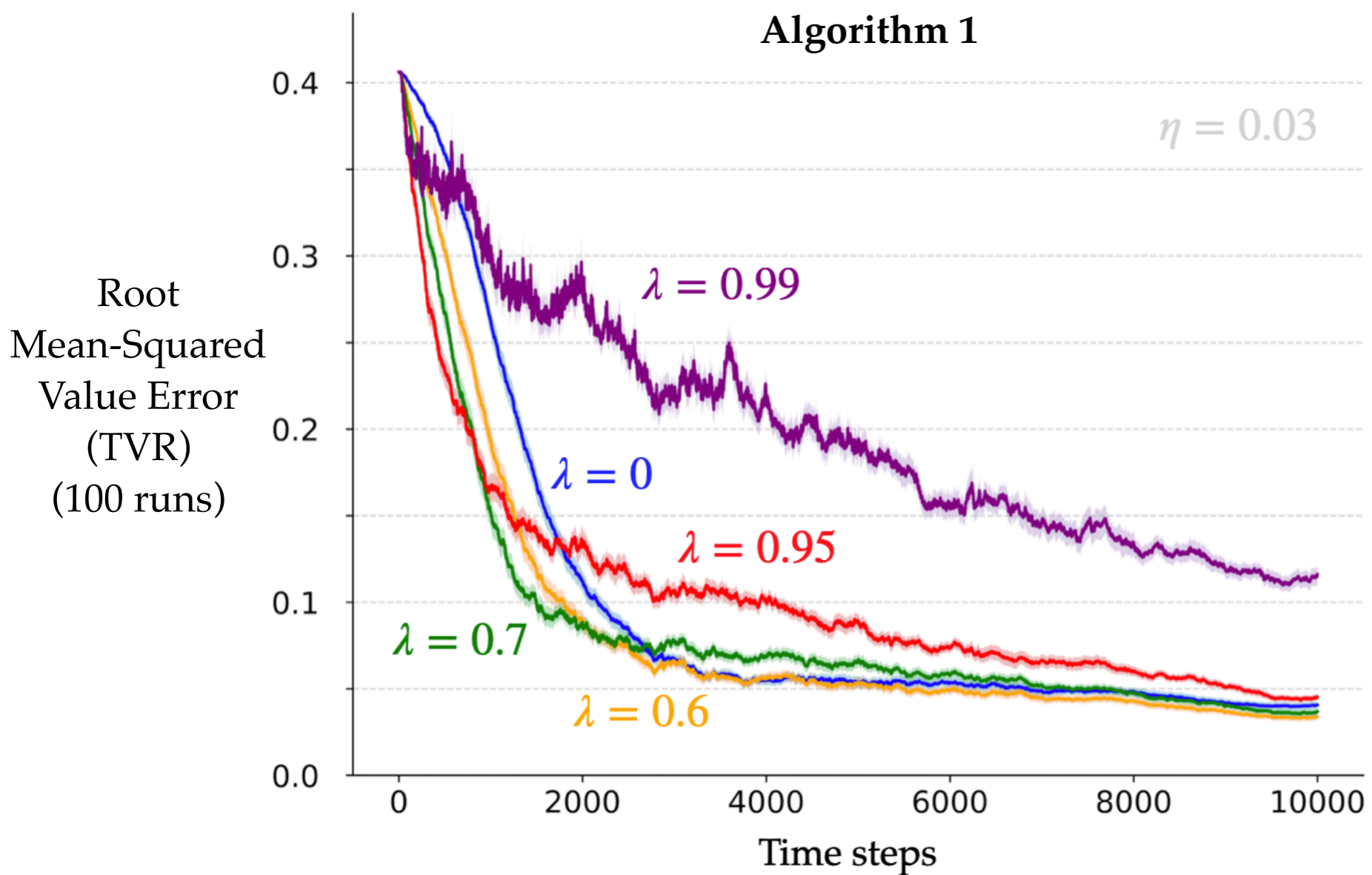


		left	right
Target policy	π	0.5	0.5
Behaviour policies	b	0.5	0.5
		0.55	0.45
		0.6	0.4
		0.65	0.35
		0.7	0.3

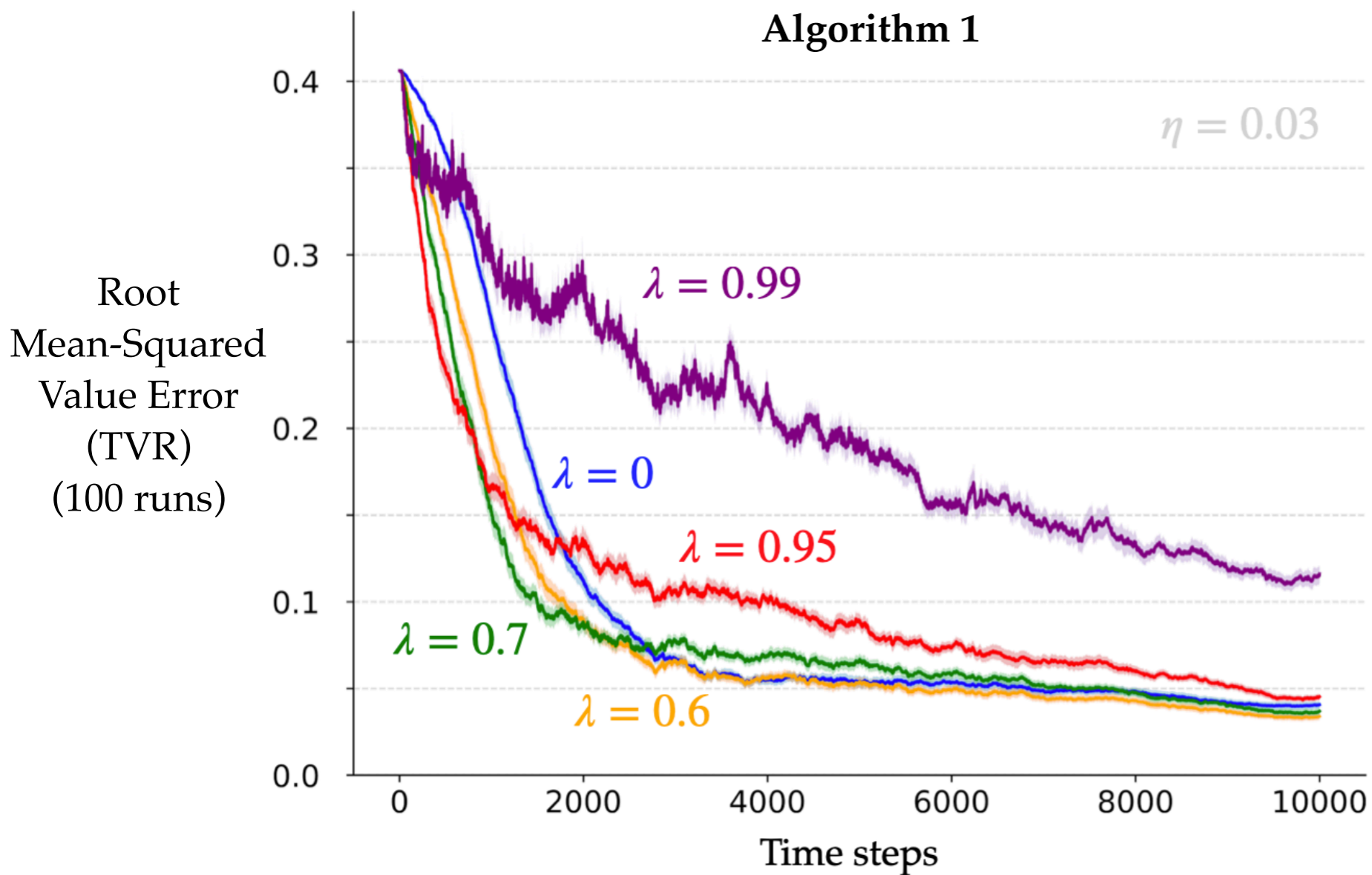
ON-POLICY LEARNING CURVES



ON-POLICY LEARNING CURVES

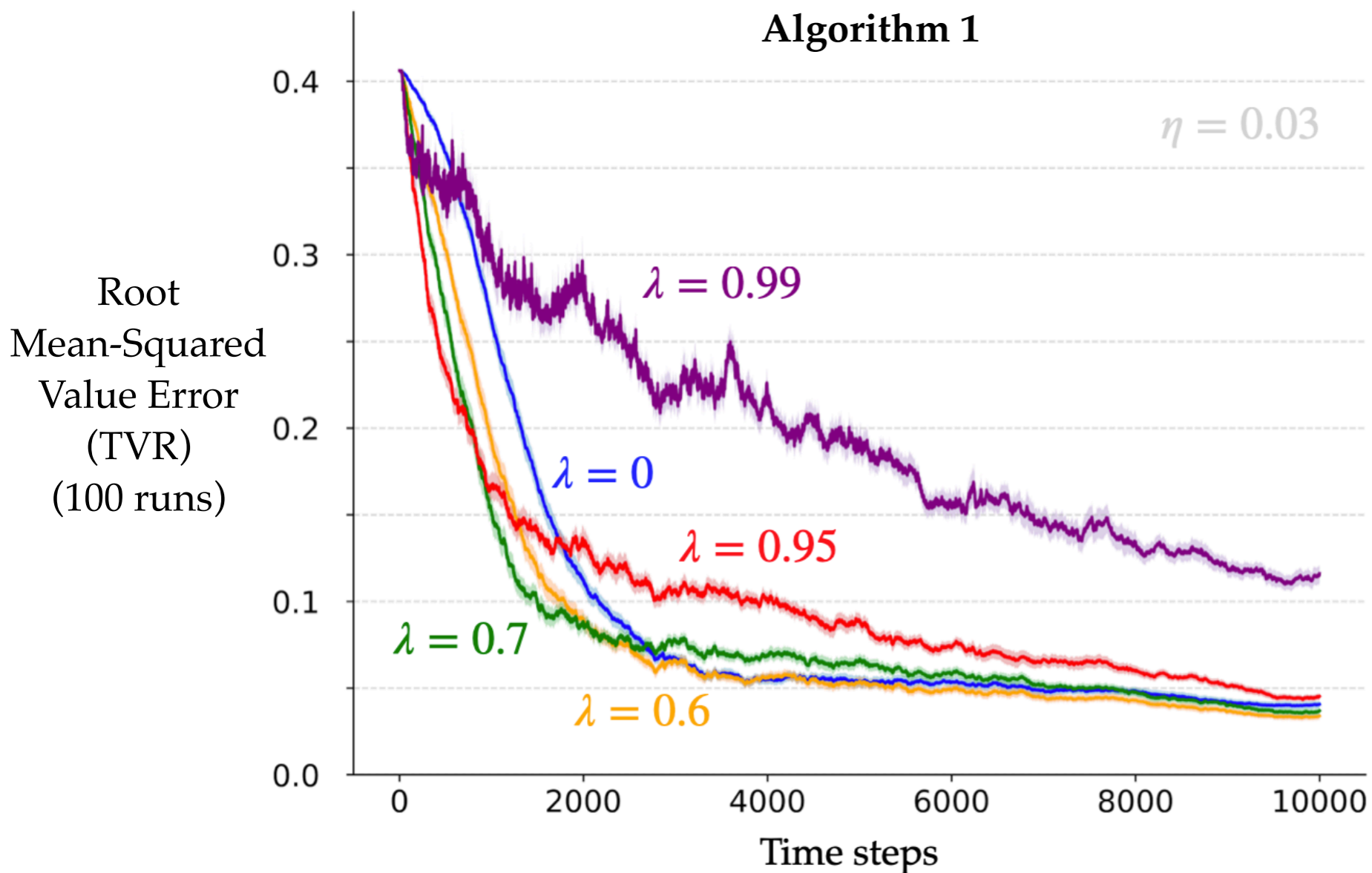


ON-POLICY LEARNING CURVES



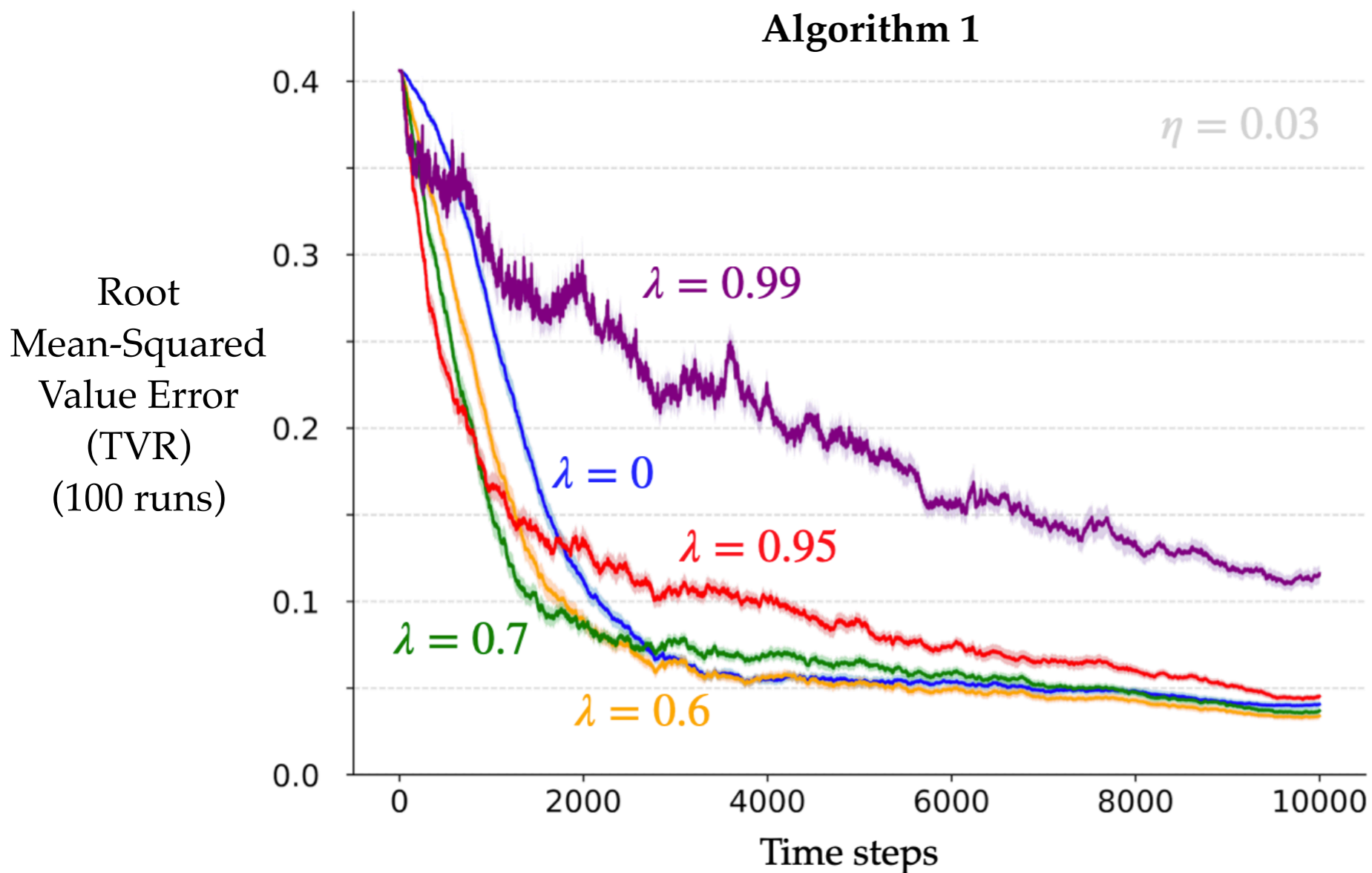
- ▶ Asymptotic convergence for all these values of λ

ON-POLICY LEARNING CURVES



- ▶ Asymptotic convergence for all these values of λ
- ▶ Intermediate value of λ works best

ON-POLICY LEARNING CURVES

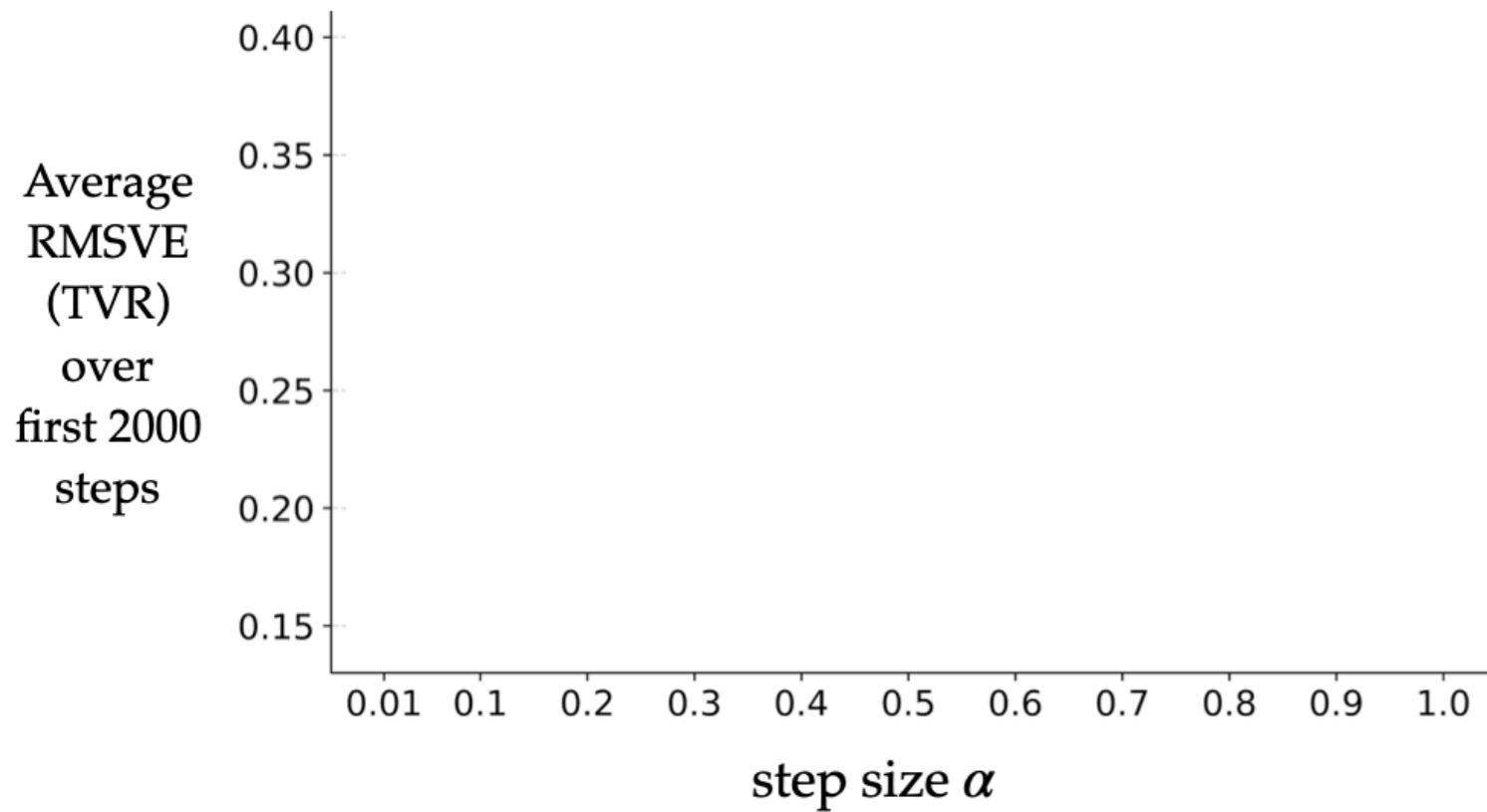


- ▶ Asymptotic convergence for all these values of λ
- ▶ Intermediate value of λ works best
- ▶ Similar trends for other values of η

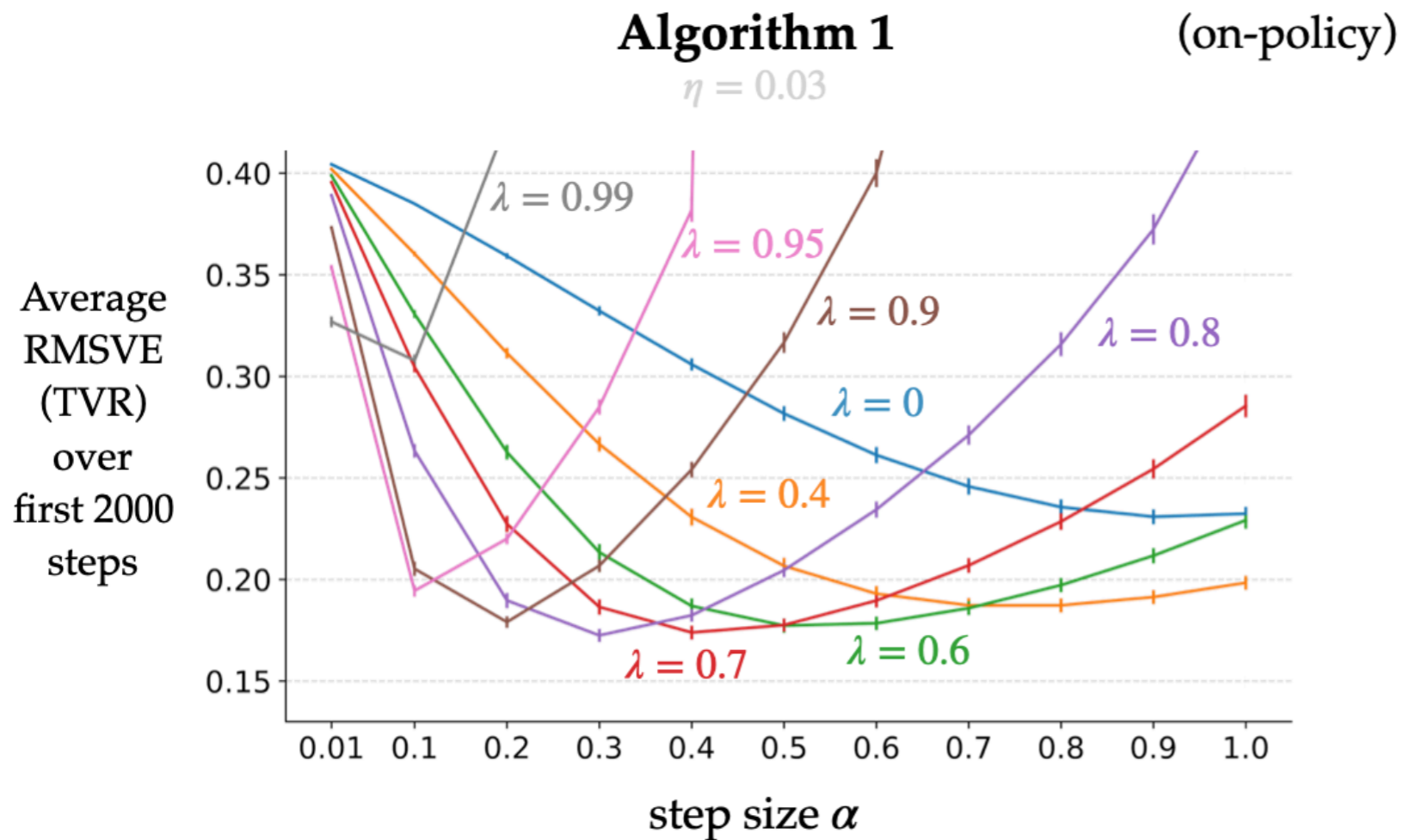
ON-POLICY SENSITIVITY PLOTS

Algorithm 1

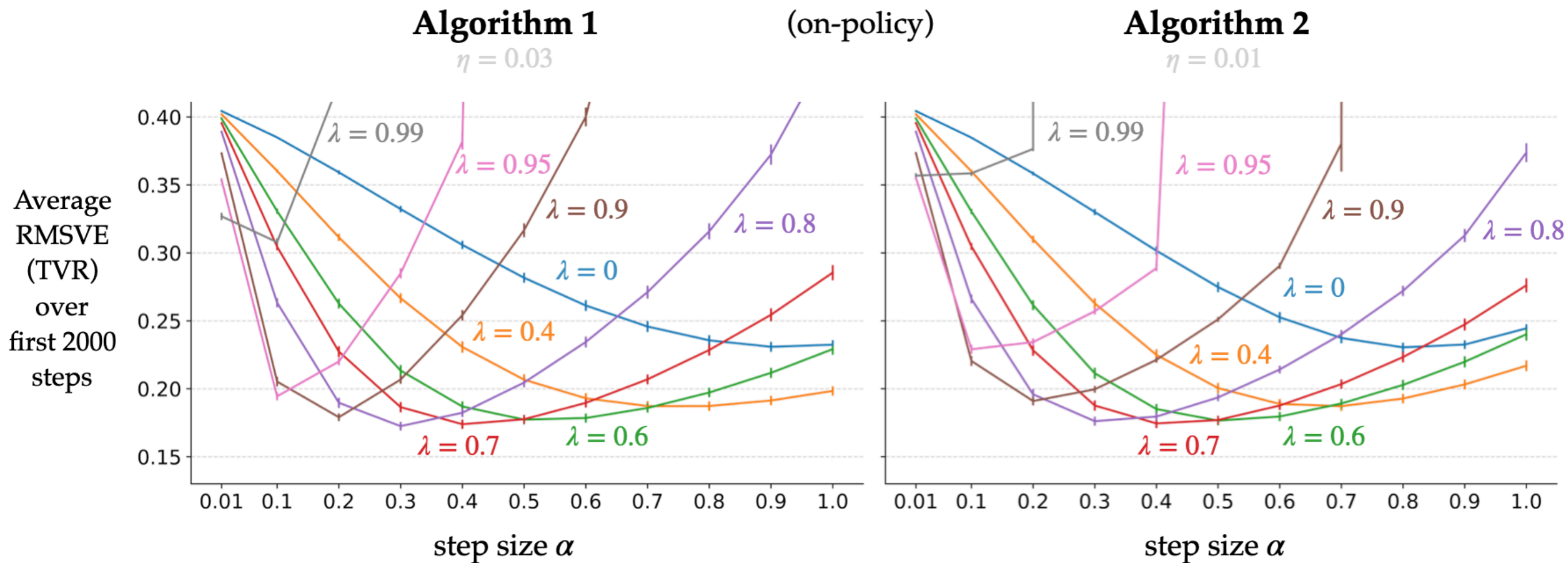
(on-policy)



ON-POLICY SENSITIVITY PLOTS



ON-POLICY SENSITIVITY PLOTS

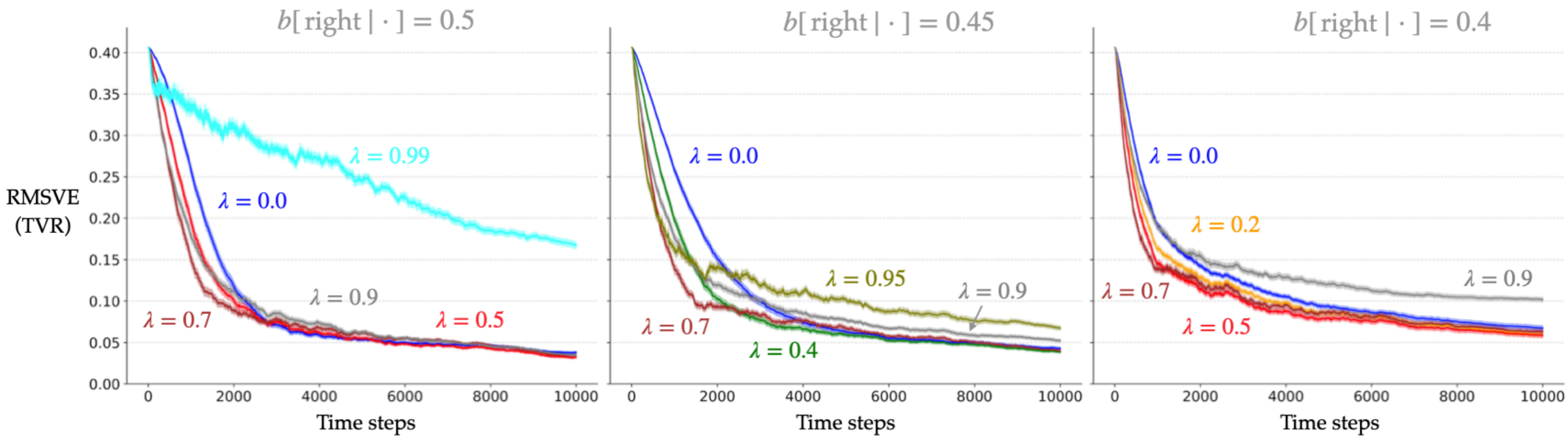


OFF-POLICY LEARNING CURVES

(Algorithm 2)

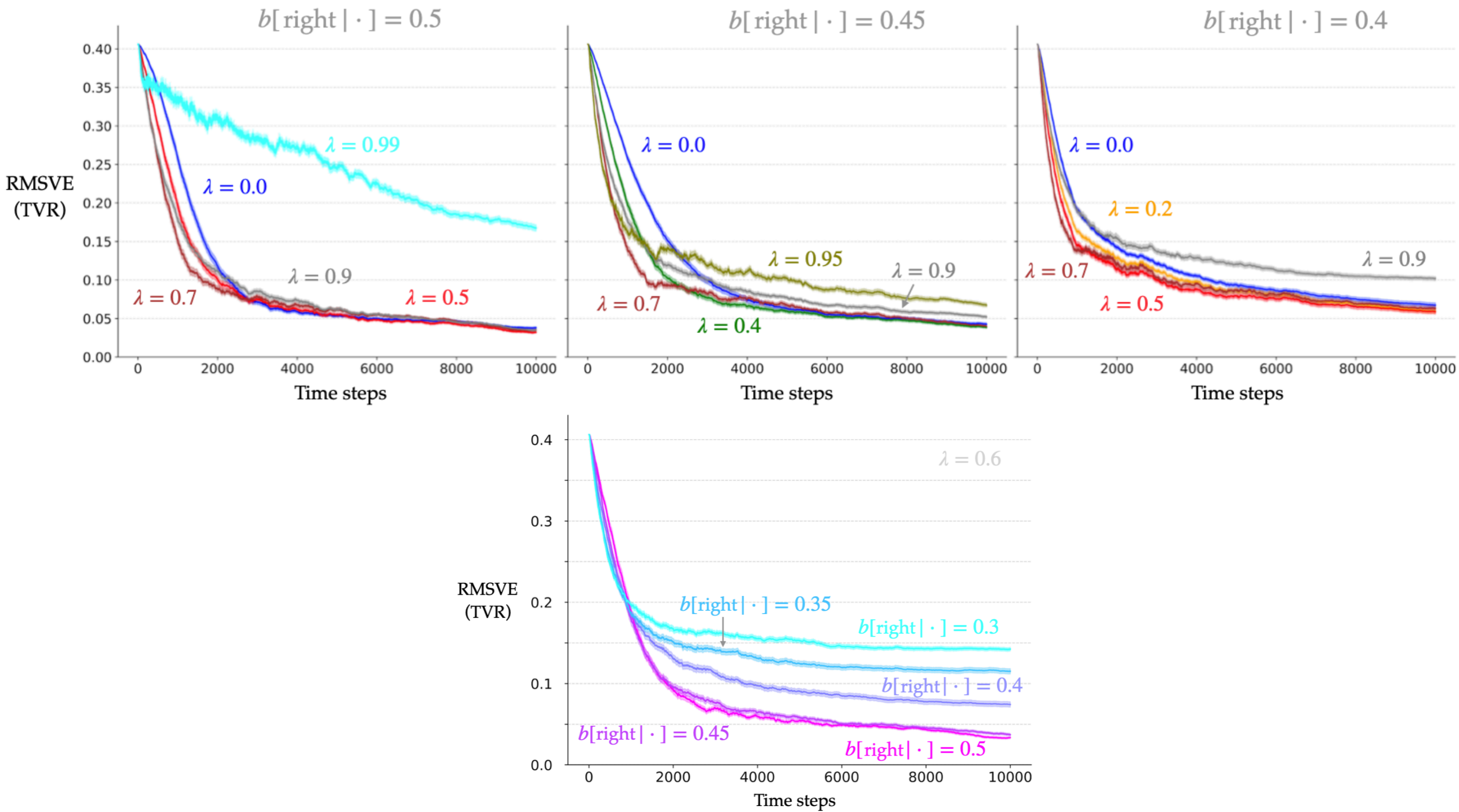
OFF-POLICY LEARNING CURVES

(Algorithm 2)



OFF-POLICY LEARNING CURVES

(Algorithm 2)



THANK YOU

Questions?