Dual Control: Optimization Based Exploration/exploitation

Anders Rantzer

Lund University, Sweden

Exploration/Exploitation

A central problem in reinforcement learning:

Exploitation: To repeat decisions that have worked well so far

Exploration: To try novel decisions, hoping to gain even greater rewards

Studied theoretically in the context of multi-armed bandits.

Heuristic: “Optimism in the face of uncertainty”.

Mostly statistical models, but growing interest in “adversarial learning”, addressing worst case scenarios.

Stochastic Formulations of Dual Control

In the control literature, the term “dual control” was introduced by Feldbaum in 1960 as an optimization approach to adaptive control.

Several attempts to address the problem were made during the 1960-80s, since lack of excitation was a central issue in the study of adaptive control.

In particular, consider minimization of the state variance in

$$y_{t+1} = y_t + bu_t + w_t$$

where w_t is a sequence of zero mean independent random variables.

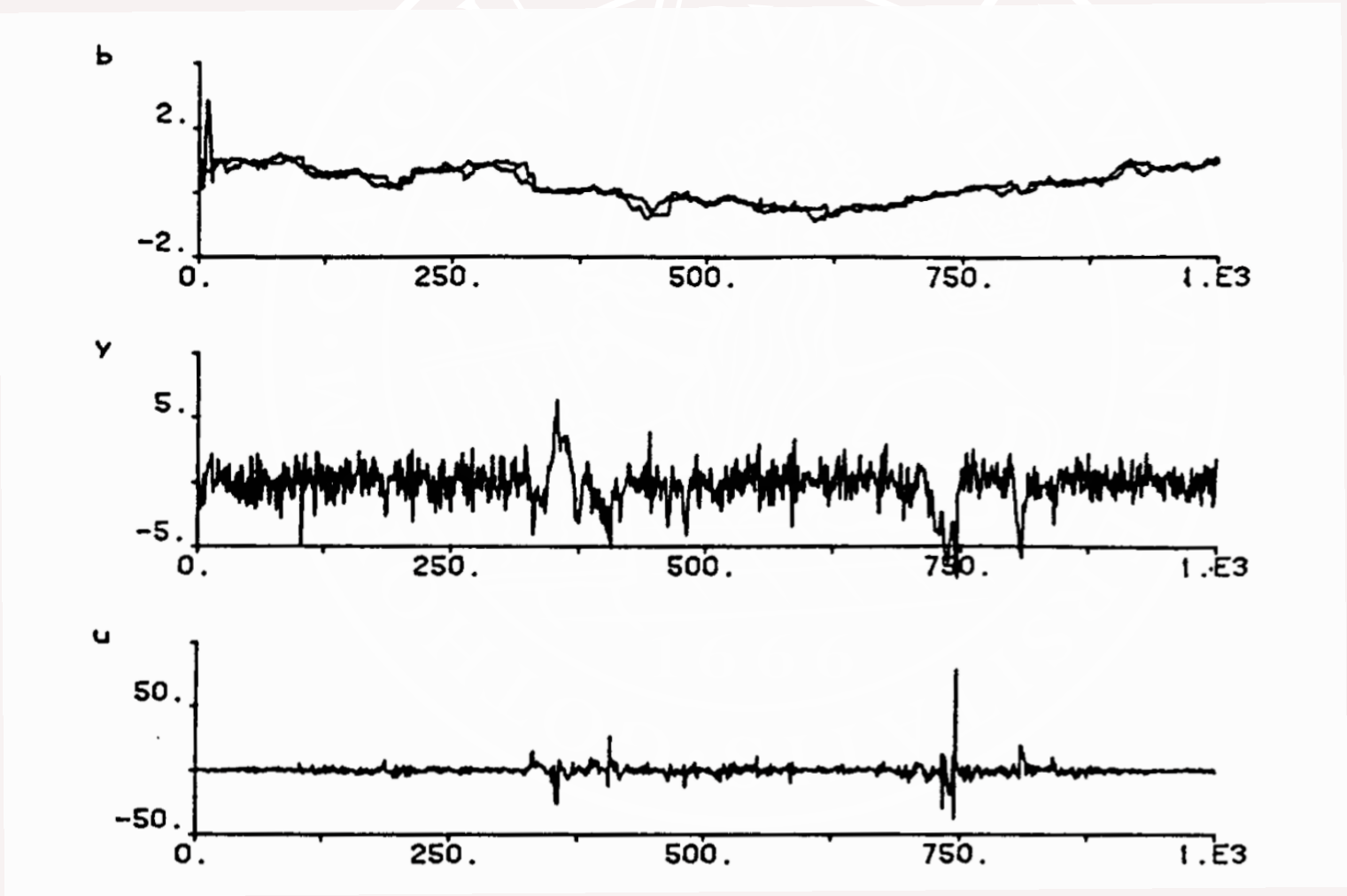
In [Åström/Helmersson, 1986] the constant b was normally distributed.
In [Bernhardsson, 1989] two values of the constant b had given probabilities.

The problems were solved numerically by dynamic programming. In both cases, the optimal controller activates u when the uncertainty in b is big.

[Åström/Helmersson, 1986]

Let the parameter b be moving in a Brownian motion.

Exploration will be activated when there is a sign change:



Limitations of the Stochastic Approach

The stochastic formulations of dual control have severe limitations:

- Prohibitive complexity of the dynamic programming approach
- Poor robustness to structural assumptions (compare H_2 vs. H_∞)

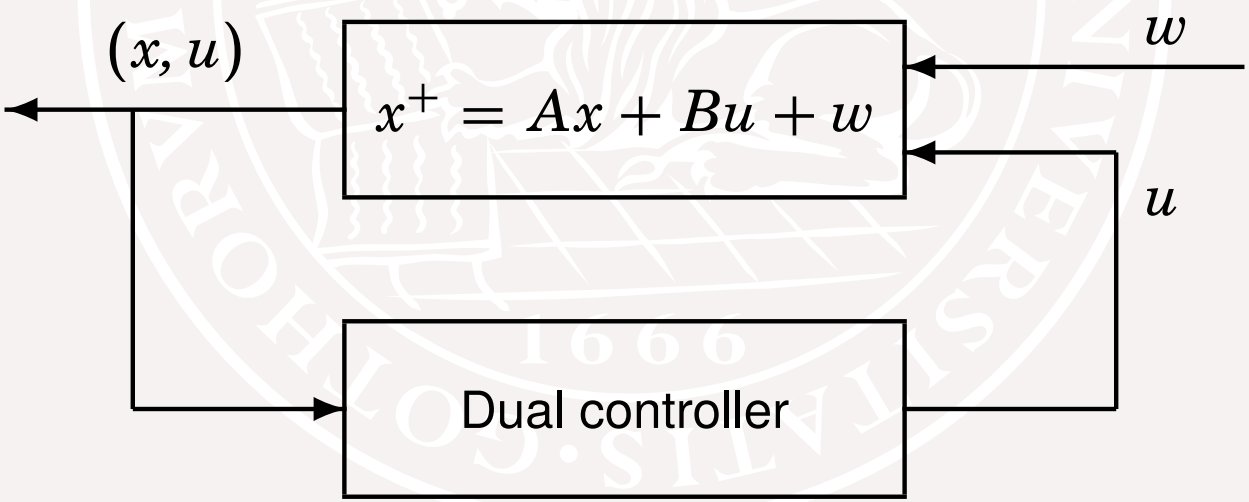
Is there a counter-part to “adversarial learning”?

Limitations of the Stochastic Approach

The stochastic formulations of dual control have severe limitations:

- Prohibitive complexity of the dynamic programming approach
- Poor robustness to structural assumptions (compare H_2 vs. H_∞)

Is there a counter-part to “adversarial learning”?



Classical Linear Quadratic Optimal Control

Let $Q, R \succ 0$ and introduce the notation $|x|_Q^2 = x^\top Q x$.

Consider the problem to find a control law μ that attains the minimum

$$\min_{\mu} \sum_{t=0}^{\infty} \left(|x_t|_Q^2 + |u_t|_R^2 \right)$$

when x_t, u_t are generated according to

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t & t \geq 0 \\ u_t &= \mu_t(x_0, \ldots, x_t, u_0, \ldots, u_{t-1}). \end{aligned}$$

The problem is solved by the state feedback controller $u = -Kx$ defined by the minimizing u in the Riccati equation

$$|x|_P^2 = \min_u \left\{ |x|_Q^2 + |u|_R^2 + |Ax + Bu|_P^2 \right\}$$

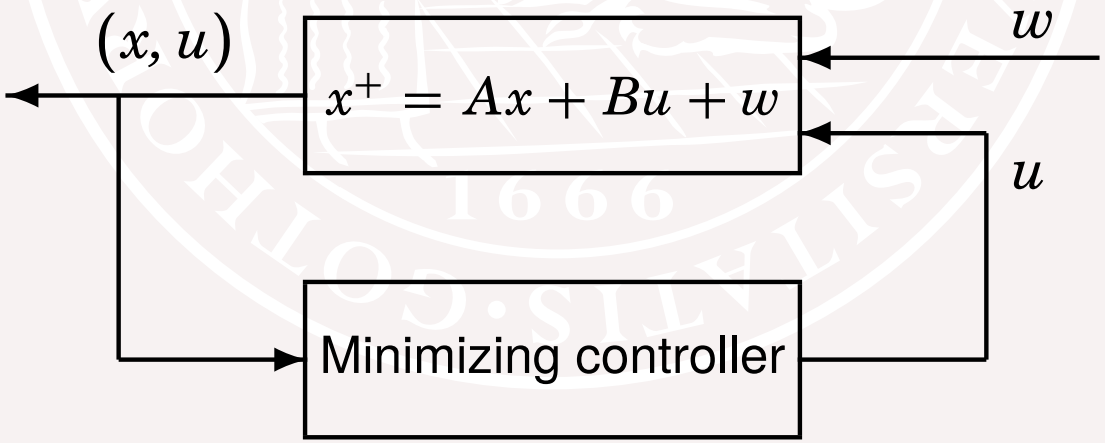
Game Formulation of H_∞ Control

Consider the problem to find a control law μ that attains the minimum

$$\min_{\mu} \max_{w} \sum_{t=0}^{\infty} \left(|x_t|_Q^2 + |u_t|_R^2 - \gamma^2 |w_t|^2 \right)$$

when x_t, u_t are generated according to

$$x_{t+1} = Ax_t + Bu_t + w_t \qquad t \geq 0$$
$$u_t = \mu_t(x_0, \ldots, x_t, u_0, \ldots, u_{t-1}).$$



Game Formulation of H_∞ Control

Consider the problem to find a control law μ that attains the minimum

$$\min_{\mu} \max_{w} \sum_{t=0}^{\infty} \left(|x_t|_Q^2 + |u_t|_R^2 - \gamma^2 |w_t|^2 \right)$$

when x_t, u_t are generated according to

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + w_t && t \geq 0 \\ u_t &= \mu_t(x_0, \ldots, x_t, u_0, \ldots, u_{t-1}). \end{aligned}$$

The problem is solved by the state feedback controller $u = -Kx$ defined by the minimizing u in the Riccati equation

$$|x|_P^2 = \min_{u} \max_{w} \left\{ |x|_Q^2 + |u|_R^2 - \gamma^2 |w|^2 + |Ax + Bu + w|_P^2 \right\}$$

The parameter γ trades “robustness” against “performance”.

Minimax Adaptive Control

Let $Q, R \succ 0$. Given $(A_1, B_1), \ldots, (A_N, B_N)$ and a number $\gamma > 0$, find a control law μ that attains the infimum

$$\inf_{\mu} \sup_{x_0, w, i} \sum_{t=0}^{\infty} \left(|x_t|_Q^2 + |u_t|_R^2 - \gamma^2 |w_t|^2 \right)$$

when $|x_0| = 1$ and supremum is taken over all solutions to

$$\begin{aligned} x_{t+1} &= A_i x_t + B_i u_t + w_t && t \geq 0 \\ u_t &= \mu_t(x_0, \ldots, x_t, u_0, \ldots, u_{t-1}). \end{aligned}$$

- If $N = 1$, then this gives standard H_∞ control.
- γ quantifies robustness to unmodeled dynamics.
- In general, nonlinear feedback with memory is needed.
- Early work by [Didinsky/Basar, CDC 1994]

Minimax Adaptive Control

Let $Q, R \succ 0$. Given $(A_1, B_1), \ldots, (A_N, B_N)$ and a number $\gamma > 0$, find a control law μ that attains the infimum

$$\inf_{\mu} \sup_{x_0, w, i} \sum_{t=0}^{\infty} \left(|x_t|_Q^2 + |u_t|_R^2 - \gamma^2 |w_t|^2 \right)$$

when $|x_0| = 1$ and supremum is taken over all solutions to

$$\begin{aligned} x_{t+1} &= A_i x_t + B_i u_t + w_t && t \geq 0 \\ u_t &= \mu_t(x_0, \ldots, x_t, u_0, \ldots, u_{t-1}). \end{aligned}$$

- If $N = 1$, then this gives standard H_∞ control.
- γ quantifies robustness to unmodeled dynamics.
- In general, nonlinear feedback with memory is needed.
- Early work by [Didinsky/Basar, CDC 1994]

Minimax Adaptive Control

Let $Q, R \succ 0$. Given $(A_1, B_1), \ldots, (A_N, B_N)$ and a number $\gamma > 0$, find a control law μ that attains the infimum

$$\inf_{\mu} \sup_{x_0, w, i} \sum_{t=0}^{\infty} \left(|x_t|_Q^2 + |u_t|_R^2 - \gamma^2 |w_t|^2 \right)$$

when $|x_0| = 1$ and supremum is taken over all solutions to

$$\begin{aligned} x_{t+1} &= A_i x_t + B_i u_t + w_t && t \geq 0 \\ u_t &= \mu_t(x_0, \ldots, x_t, u_0, \ldots, u_{t-1}). \end{aligned}$$

- If $N = 1$, then this gives standard H_∞ control.
- γ quantifies robustness to unmodeled dynamics.
- In general, nonlinear feedback with memory is needed.
- Early work by [Didinsky/Basar, CDC 1994]

Minimax Adaptive Control

Let $Q, R \succ 0$. Given $(A_1, B_1), \ldots, (A_N, B_N)$ and a number $\gamma > 0$, find a control law μ that attains the infimum

$$\inf_{\mu} \sup_{x_0, w, i} \sum_{t=0}^{\infty} \left(|x_t|_Q^2 + |u_t|_R^2 - \gamma^2 |w_t|^2 \right)$$

when $|x_0| = 1$ and supremum is taken over all solutions to

$$\begin{aligned} x_{t+1} &= A_i x_t + B_i u_t + w_t && t \geq 0 \\ u_t &= \mu_t(x_0, \ldots, x_t, u_0, \ldots, u_{t-1}). \end{aligned}$$

- If $N = 1$, then this gives standard H_∞ control.
- γ quantifies robustness to unmodeled dynamics.
- In general, nonlinear feedback with memory is needed.
- Early work by [Didinsky/Basar, CDC 1994]

Minimax Adaptive Control

Let $Q, R \succ 0$. Given $(A_1, B_1), \ldots, (A_N, B_N)$ and a number $\gamma > 0$, find a control law μ that attains the infimum

$$\inf_{\mu} \sup_{x_0, w, i} \sum_{t=0}^{\infty} \left(|x_t|_Q^2 + |u_t|_R^2 - \gamma^2 |w_t|^2 \right)$$

when $|x_0| = 1$ and supremum is taken over all solutions to

$$\begin{aligned} x_{t+1} &= A_i x_t + B_i u_t + w_t && t \geq 0 \\ u_t &= \mu_t(x_0, \ldots, x_t, u_0, \ldots, u_{t-1}). \end{aligned}$$

- If $N = 1$, then this gives standard H_∞ control.
- γ quantifies robustness to unmodeled dynamics.
- In general, nonlinear feedback with memory is needed.
- Early work by [Didinsky/Basar, CDC 1994]

Equivalent Dynamic Game

Given $Q \succ 0$, $R \succ 0$, $\gamma > 0$, find a control law η that attains the infimum

$$\inf_{\eta} \sup_{x_0, v, T} \left[-\gamma^2 \min_i \left\| \begin{bmatrix} I & A_i & B_i \end{bmatrix} \right\|^2_{Z_{T+1}} + \sum_{t=0}^{T} \left(|x_t|^2_Q + |u_t|^2_R \right) \right]$$

when $|x_0| = 1$ and x, u, Z are generated from x_0 and v according to

$$\begin{cases} x_{t+1} = v_t \\ Z_{t+1} = Z_t + \begin{bmatrix} -v_t \\ x_t \\ u_t \end{bmatrix} \begin{bmatrix} -v_t \\ x_t \\ u_t \end{bmatrix}^\top, \qquad Z_0 = 0 \end{cases}$$

and the control law $u_t = \eta(x_t, Z_t)$.

Remark:

Uncertain (A_i, B_i) appears only in terminal cost, not in dynamics

Dynamic Programming Approach

The dynamic game has a finite value if and only if the Bellman equation

$$V_*(x, Z) = \min_u \max_v \left\{ |x|_Q^2 + |u|_R^2 + V_* \left(v, Z + \begin{bmatrix} -v \\ x \\ u \end{bmatrix} \begin{bmatrix} -v \\ x \\ u \end{bmatrix}^\top \right) \right\}$$

has a solution V_* satisfying

$$-\gamma^2 \min_i \| \begin{bmatrix} I & A_i & B_i \end{bmatrix} \|_Z^2 \leq V_*(x, Z) \leq \gamma^2 |x|^2$$

for all $Z \succeq 0$ and $x \in \mathbb{R}^n$.

The value of the game is $V_*(x_0, 0)$.

Example 1: Scalar System with Unknown Input Sign

$$\inf_{\mu} \sup_{w,i} \quad \sum_{t=0}^{\infty} \left(x_t^2 + u_t^2 - \gamma^2 w_t^2 \right)$$

where $\quad x_{t+1} = 1.5 x_t + b u_t + w_t \qquad b \in \{-1, 1\}$

$$u_t = \mu_t(x_0, \ldots, x_t, u_0, \ldots, u_{t-1}) \qquad t \geq 0.$$

Dynamic programming:

$$V_*(x, Z) = \min_u \max_v \left\{ x^2 + u^2 + V_*\left(v, Z + \begin{bmatrix} 1.5x - v \\ u \end{bmatrix} \begin{bmatrix} 1.5x - v \\ u \end{bmatrix}^\top \right) \right\}$$

$$-\gamma^2 \min_{b=\pm 1} \| \begin{bmatrix} 1 & b \end{bmatrix} \|_Z^2 \leq V_*(x, Z) \leq \gamma^2 |x|^2$$

Optimal control law for $\gamma = 6.72$: Use certainty equivalence!

$$u_t = \begin{cases} +0.63 x_t & \text{if } \sum_{k=0}^{t-1} u_k (x_{k+1} - 1.5 x_k) \leq 0 \\ -0.63 x_t & \text{if } \sum_{k=0}^{t-1} u_k (x_{k+1} - 1.5 x_k) > 0 \end{cases}$$

Example 1: Scalar System with Unknown Input Sign

$$\inf_{\mu} \sup_{w,i} \quad \sum_{t=0}^{\infty} \left(x_t^2 + u_t^2 - \gamma^2 w_t^2 \right)$$

$$\text{where} \quad x_{t+1} = 1.5 x_t + b u_t + w_t \qquad b \in \{-1, 1\}$$

$$u_t = \mu_t(x_0, \ldots, x_t, u_0, \ldots, u_{t-1}) \qquad t \geq 0.$$

Dynamic programming:

$$V_*(x, Z) = \min_u \max_v \left\{ x^2 + u^2 + V_* \left(v, Z + \begin{bmatrix} 1.5x - v \\ u \end{bmatrix} \begin{bmatrix} 1.5x - v \\ u \end{bmatrix}^\top \right) \right\}$$

$$- \gamma^2 \min_{b=\pm 1} \| \begin{bmatrix} 1 & b \end{bmatrix} \|_Z^2 \leq V_*(x, Z) \leq \gamma^2 |x|^2$$

Optimal control law for $\gamma = 6.72$: Use certainty equivalence!

$$u_t = \begin{cases} +0.63 x_t & \text{if } \sum_{k=0}^{t-1} u_k (x_{k+1} - 1.5 x_k) \leq 0 \\ -0.63 x_t & \text{if } \sum_{k=0}^{t-1} u_k (x_{k+1} - 1.5 x_k) > 0 \end{cases}$$

Example 1: Scalar System with Unknown Input Sign

$$\inf_{\mu} \sup_{w,i} \quad \sum_{t=0}^{\infty} \left(x_t^2 + u_t^2 - \gamma^2 w_t^2 \right)$$

$$\text{where} \quad x_{t+1} = 1.5 x_t + b u_t + w_t \qquad b \in \{-1, 1\}$$

$$u_t = \mu_t(x_0, \ldots, x_t, u_0, \ldots, u_{t-1}) \qquad t \geq 0.$$

Dynamic programming:

$$V_*(x, Z) = \min_u \max_v \left\{ x^2 + u^2 + V_* \left(v, Z + \begin{bmatrix} 1.5x - v \\ u \end{bmatrix} \begin{bmatrix} 1.5x - v \\ u \end{bmatrix}^\top \right) \right\}$$

$$- \gamma^2 \min_{b=\pm 1} \| \begin{bmatrix} 1 & b \end{bmatrix} \|_Z^2 \leq V_*(x, Z) \leq \gamma^2 |x|^2$$

Optimal control law for $\gamma = 6.72$: Use certainty equivalence!

$$u_t = \begin{cases} +0.63 x_t & \text{if } \sum_{k=0}^{t-1} u_k (x_{k+1} - 1.5 x_k) \leq 0 \\ -0.63 x_t & \text{if } \sum_{k=0}^{t-1} u_k (x_{k+1} - 1.5 x_k) > 0 \end{cases}$$

Example 2: Scalar System with Unknown Input Sign

$$\inf_{\mu} \sup_{w,i} \quad \sum_{t=0}^{\infty} \left(x_t^2 + u_t^2 - \gamma^2 w_t^2 \right)$$

where $x_{t+1} = 1.5x_t + bu_t + w_t \qquad b \in \{-1, 1\}$

$$u_t = \mu_t(x_0, \ldots, x_t, u_0, \ldots, u_{t-1}) \qquad t \geq 0.$$

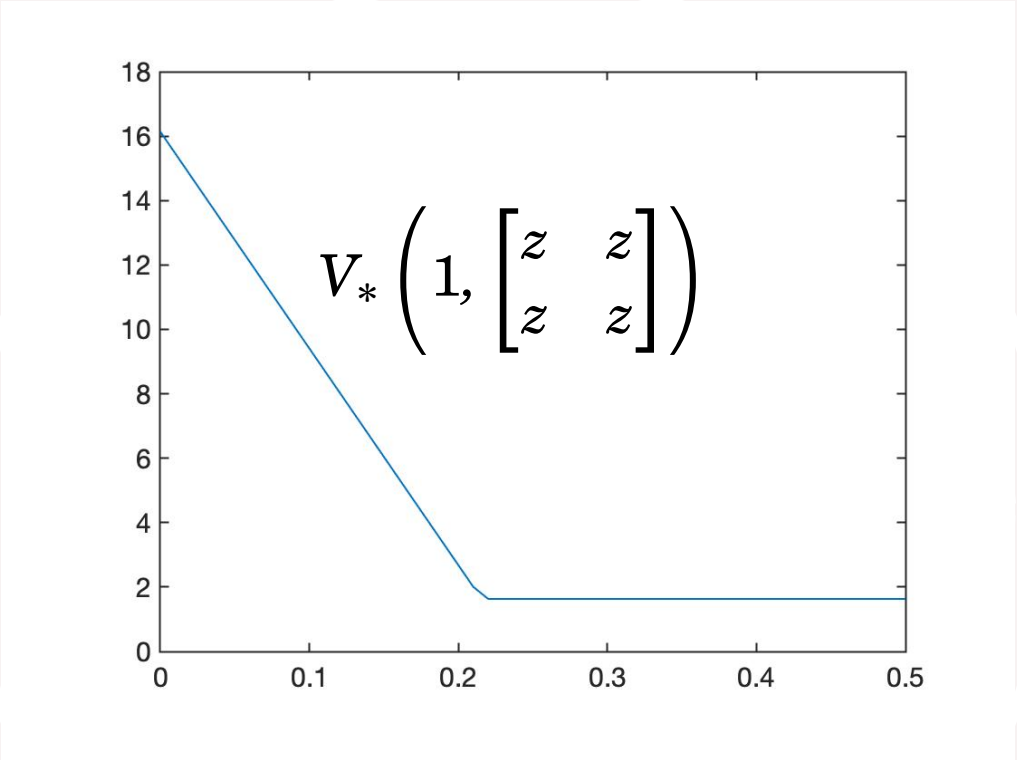
Optimal control law for $\gamma = 6.72$: Use certainty equivalence!

$$u_t = \begin{cases} +0.63x_t & \text{if } \sum_{k=0}^{t-1} u_k(x_{k+1} - 1.5x_k) \leq 0 \\ -0.63x_t & \text{if } \sum_{k=0}^{t-1} u_k(x_{k+1} - 1.5x_k) > 0 \end{cases}$$

Optimal control law for $\gamma = 5.8$: Active exploration when uncertain.

$$u_t = \begin{cases} +0.63x_t & \text{if } \sum_{k=0}^{t-1} u_k(x_{k+1} - 1.5x_k) \in (-\infty, -x_t^2] \\ +1.3x_t & \text{if } \sum_{k=0}^{t-1} u_k(x_{k+1} - 1.5x_k) \in (-x_t^2, 0] \\ -1.3x_t & \text{if } \sum_{k=0}^{t-1} u_k(x_{k+1} - 1.5x_k) \in (0, x_t^2] \\ -0.63x_t & \text{if } \sum_{k=0}^{t-1} u_k(x_{k+1} - 1.5x_k) \in (x_t^2, \infty) \end{cases}$$

The Optimal Cost for Example 2



$$V_*(x, Z) = \max\left\{16.13x^2 - 5.8^2\,\mathrm{tr}(Z)\,,\quad 1.63x^2 - 5.8^2 \min_{b=\pm 1} \left\|\begin{bmatrix}1 & b\end{bmatrix}\right\|_Z^2\right\}$$

Example 3: Unknown Sign of State Dynamics

$$\inf_{\mu} \sup_{w,i} \quad \sum_{t=0}^{\infty} \left(x_t^2 + u_t^2 - \gamma^2 w_t^2 \right)$$

$$\text{where} \quad x_{t+1} = a x_t + u_t + w_t \qquad a \in \{-1, 1\}$$

$$u_t = \mu_t(x_0, \ldots, x_t, u_0, \ldots, u_{t-1}) \qquad t \geq 0.$$

Dynamic programming:

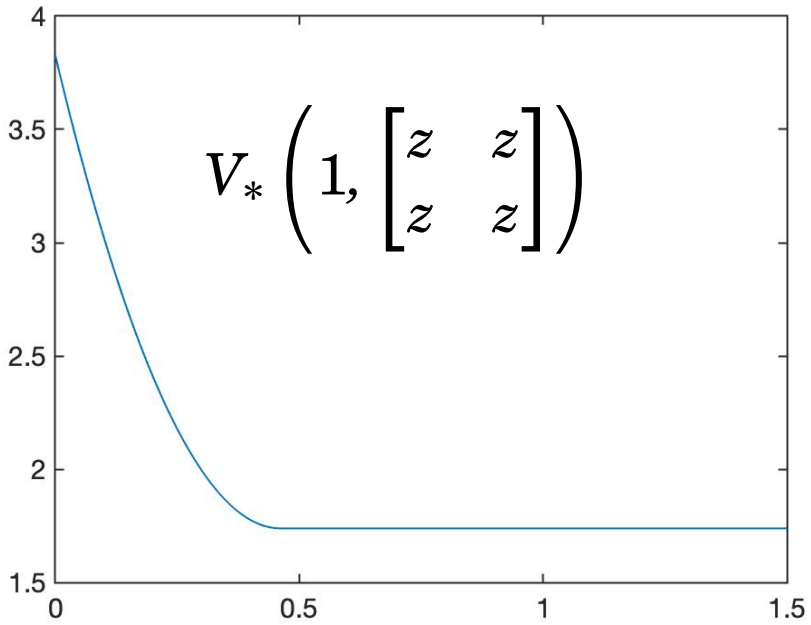
$$V_*(x, Z) = \min_u \max_v \left\{ x^2 + u^2 + V_* \left(v, Z + \begin{bmatrix} u - v \\ x \end{bmatrix} \begin{bmatrix} u - v \\ x \end{bmatrix}^\top \right) \right\}$$

$$-\gamma^2 \min_{a=\pm 1} \| \begin{bmatrix} 1 & a \end{bmatrix} \|_Z^2 \leq V_*(x, Z) \leq \gamma^2 |x|^2$$

Optimal control law for $\gamma = 2.13$:

$$u_t = \text{sat}\left(\frac{\sum_{k=0}^{t-1} (u_k - x_{k+1}) x_k}{x_t^2} \right) \cdot 0.354 x_t$$

The Optimal Cost for Example 3



$$V_*\left(x, \begin{bmatrix} z_{11} & z_{12} \\ z_{12} & z_{22} \end{bmatrix}\right)$$

$$= \begin{cases} 1.74x^2 - 4.52(z_{11} + z_{22} - 2|z_{12}|) & \text{if } z_{12} \geq 0.46 \\ 3.82x^2 - 4.52(z_{11} + z_{22}) + 9.79z_{12}^2 & \text{otherwise} \end{cases}$$

Higher Order Systems

$$\inf_{\mu} \sup_{w,i} \quad \sum_{t=0}^{\infty} \left(|x_t|_Q^2 + |u_t|_R^2 - \gamma^2 |w_t|^2 \right)$$

$$\begin{aligned} \text{where} \quad x_{t+1} &= iAx_t + Bu_t + w_t && i \in \{-1, 1\} \\ u_t &= \mu_t(x_0, \ldots, x_t, u_0, \ldots, u_{t-1}) && t \geq 0. \end{aligned}$$

Dynamic programming:

$$V_*(x, Z) = \min_u \max_v \left\{ |x|_Q^2 + |u|_R^2 + V_* \left(v, Z + \begin{bmatrix} Bu - v \\ x \end{bmatrix} \begin{bmatrix} Bu - v \\ x \end{bmatrix}^\top \right) \right\}$$

$$- \gamma^2 \min_{i=\pm 1} \| \begin{bmatrix} I & iA \end{bmatrix} \|_Z^2 \leq V_*(x, Z) \leq \gamma^2 |x|^2$$

Under general assumption, an **optimal control law** takes the form:

$$u_t = \text{sat} \left(\frac{\sum_{k=0}^{t-1} (Bu_k - x_{k+1})^\top A x_k}{|x_t|_{T-P}^2} \right) \cdot K x_t$$

where K, P, T are obtained from a Riccati equation.

Conclusions

Dual control problems can be stated in terms of stochastic optimal control, but also as zero-sum dynamic games.

The latter has important advantages:

- Robustness guarantees in presence of unmodelled dynamics
- Explicit solutions to the Bellman equation, even for high order systems

First draft available on http://arxiv.org/abs/1912.03550.

European Research Council
Established by the European Commission
VETENSKAPSRÅDET
THE SWEDISH RESEARCH COUNCIL
Stiftelsen för Strategisk Forskning
WASP
WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM