Learning to fly more efficiently in groups

Reinforcement Learning and String Stability

Esteban A.L. Hufstedler[1], James Riehl[1],
Julien M. Hendrickx[2], Philippe Chatelain[1]

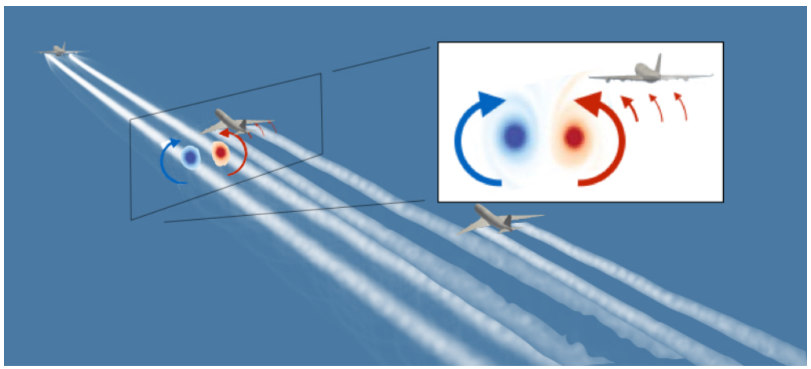
[1]Institute of Mechanics, Materials and Civil Engineering

[2]Institute of Information and Communication Technologies, Electronics and Applied Mathematics

Benelux Meeting on Systems and Control, 19-21 March 2019

Introduction

- Formation flight is great
  - Reduced energy cost
  - Birds save ∼12% [Voelkl et al., 2015]
- Formation flight is hard
  - Find "sweet spot"
  - Compensate for leader's dynamics

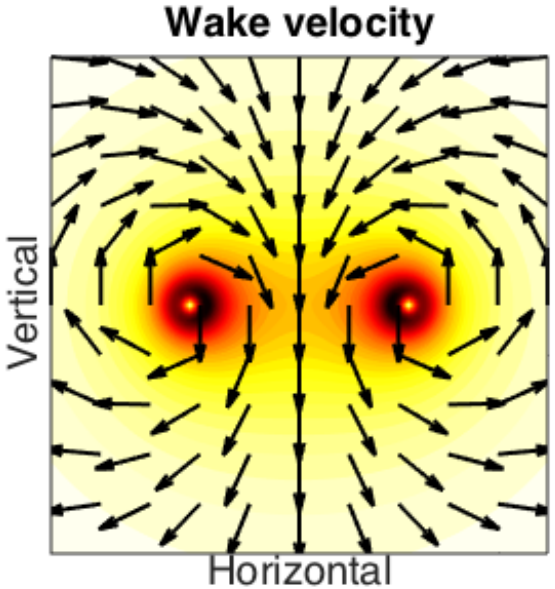


Our goals:

- Stabilize a plane in a wake (LQR)
- Find the point of minimum fuel use (Reinforcement learning)
- Add more planes, stay stable (string instability)

Wakes

- Flying things have wakes
  - Newton's 3rd law
  - Rolls up into vortex pair
- Can 'surf' wakes
  - Upwash is free energy, so reduce thrust
  - Beware downwash
  - Position error of 10% halves benefit [Binetti et al., 2003]



System Model

- Aircraft in wake of another, using published dynamics [Binetti et al., 2003]
- $\dot{x} = Ax + Bu + Fu_{wake}$
- State: velocities, rotations around 3 axes, relative to the leading aircraft
- Motion of plane due to its control (thrust and control surfaces)
- Wake is exogenous (nonlinear) input

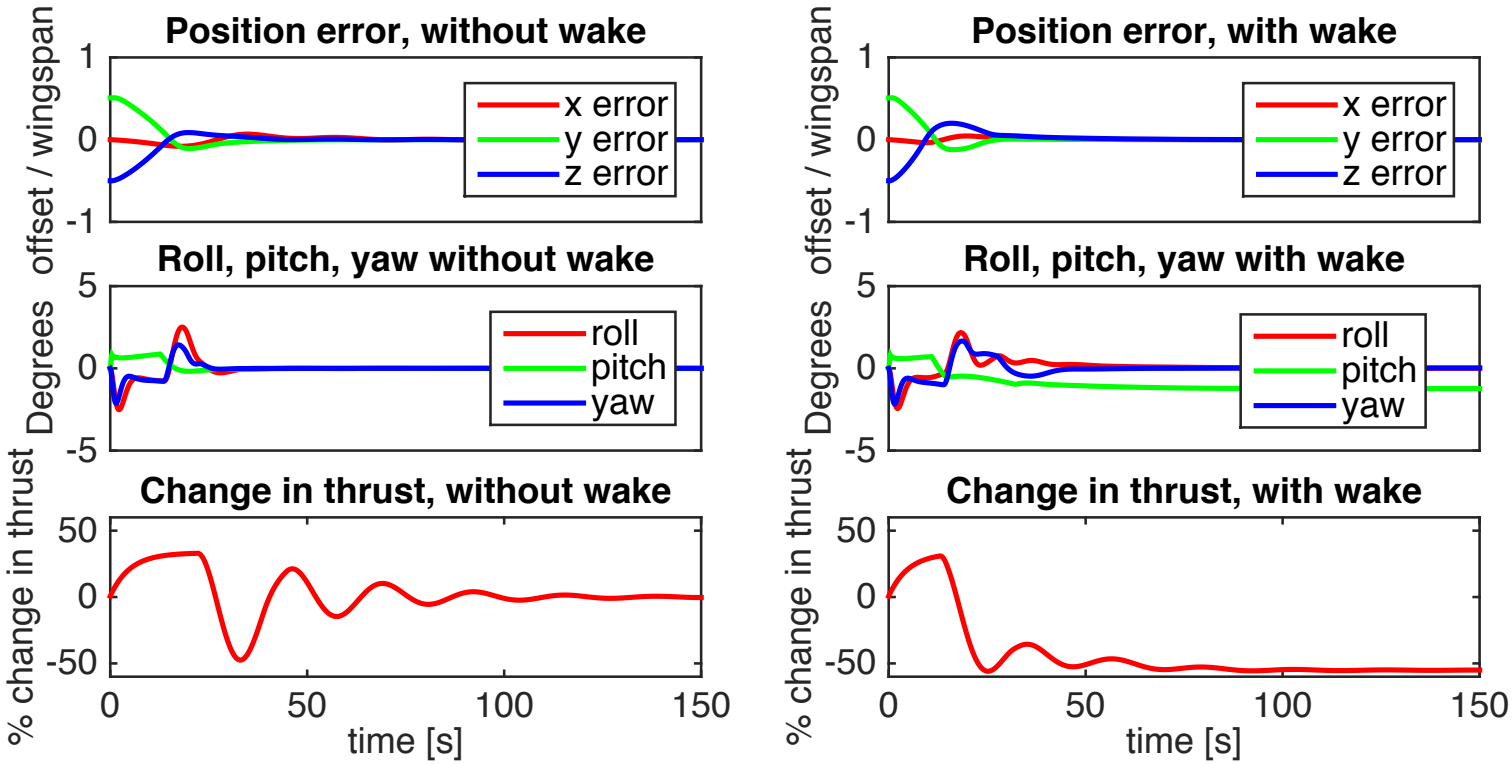
C-5 Galaxy (similar to 747)

- Wingspan: 67.9 m
- Weight: 382,000 kg
- Cruise speed: Mach 0.77 (833 km/h)
- Range: 9165 km

Goal 1: Level flight

- Nearly steady flight conditions (slow maneuvers)
- Basic control: LQR
- Command position relative to leader
- Use integrators for wake effects
- Anti-windup due to control saturation

Goal 1: Level flight, test



- Successfully moved up 0.5 wingspans, and 1.5 wingspans to the left, near sweet spot
- Thrust decreases by 50%!

Goal 2: Peak finding

UCLouvain

- Previous work
  - Extremum-seeking control [Binetti et al., 2003]
  - Wake sensing [Hemati et al., 2014]
- Reinforcement Learning [Sutton et al., 1998]
  - Flexible: variety of wakes, unsteadiness, etc
  - Reward = -Thrust
  - Actions: choose a target position
- Our policy
  - Apply Multi Armed Bandit (MAB) ideas: Upper Confidence Bounds [Auer, 2002]
  - Estimate return with running average
  - More visits, lower uncertainty
  - Choose action to maximize (estimate + uncertainty)

Spatial MABs

Treat peak-finding as a spatial MAB:

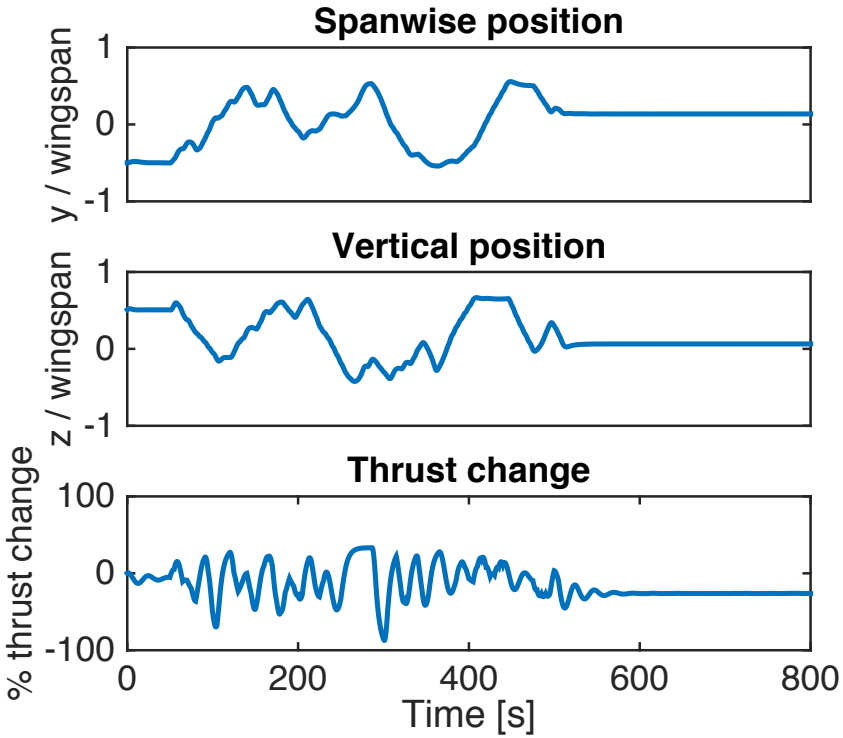
- Assume rewards are smooth over some length scale
- Distribute rewards and visit counts with a finitely-supported gaussian-like function

Details:

Max step size	0.1b
Time between choosing positions	5s
α_{learn}	0.1/s
Max visit count N	5s
Support width	b/3

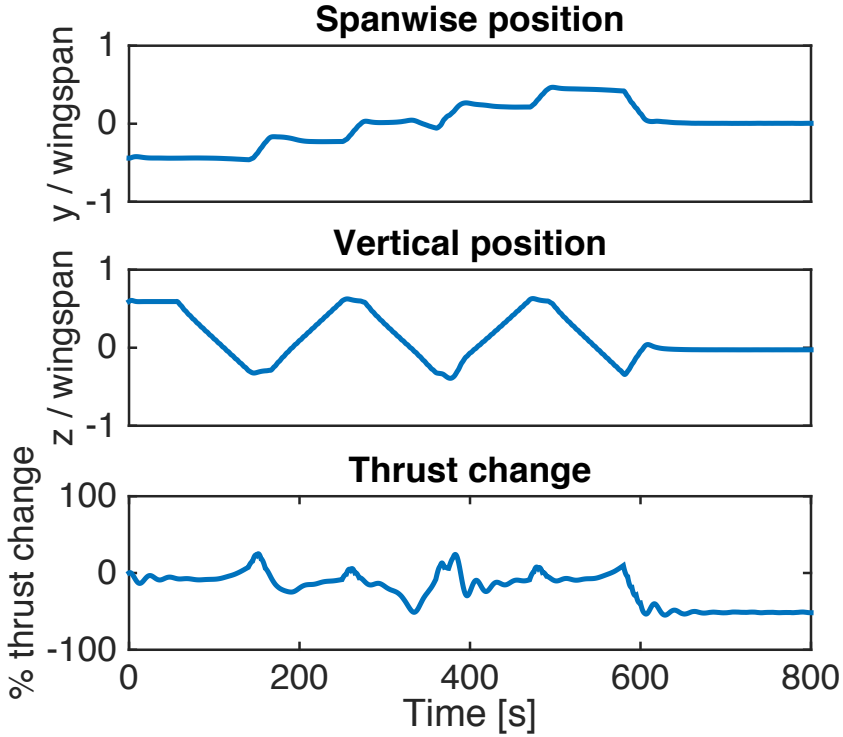
Finding the sweet spot

Searching without a grid



Near the optimum, 26% savings

Searching with a grid

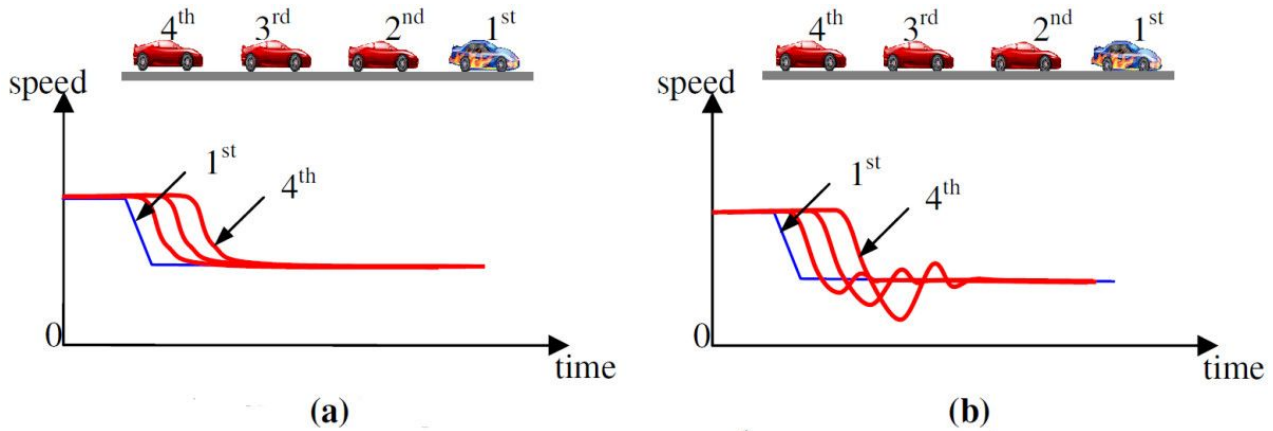


More exhaustive, 51% savings

The MAB approach can find the optimal position.

String Stability

Commonly associated with vehicle platooning

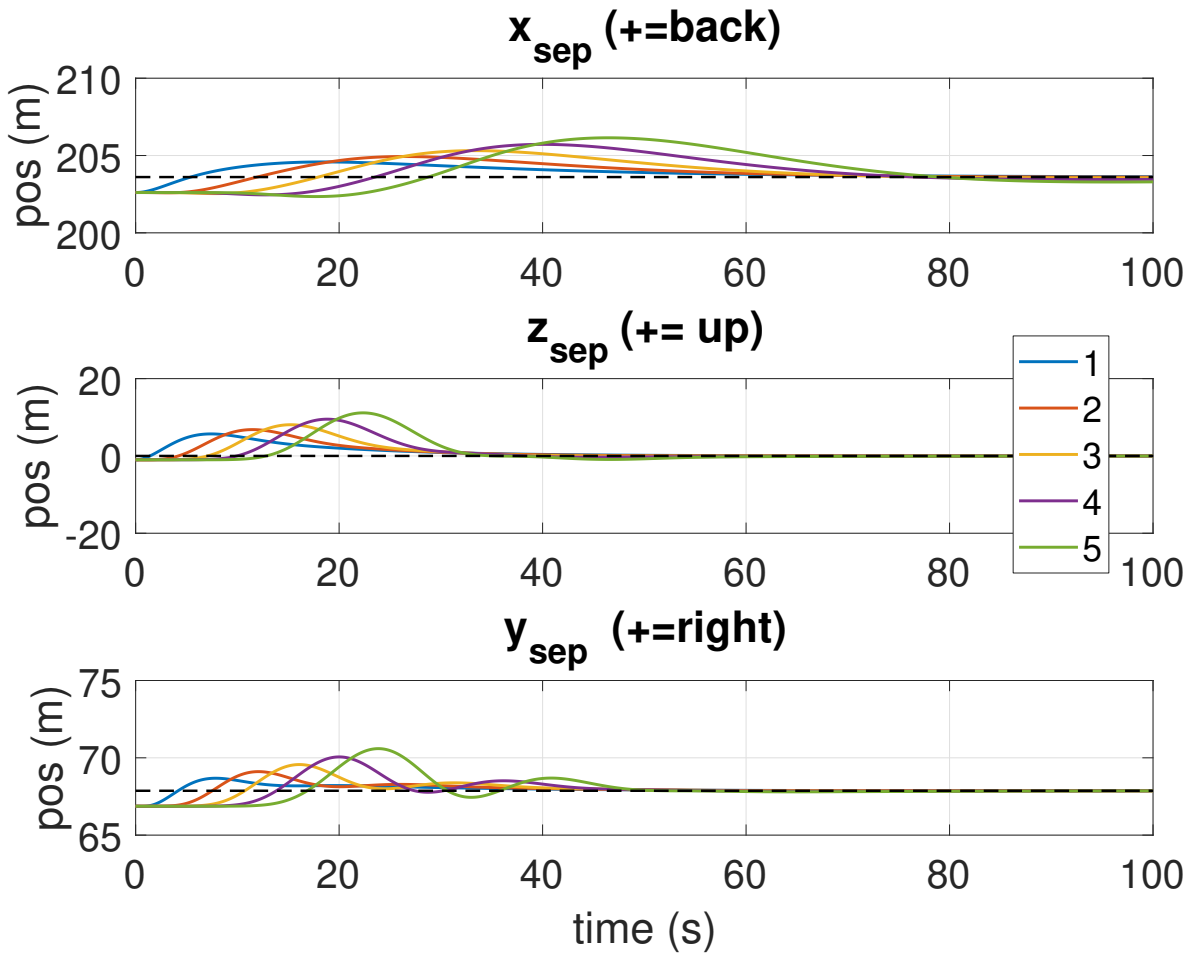


Local (leader-follower) stability is not sufficient for cascaded or "string" systems

Aircraft trying to maintain an energy-saving formation may not be string stable!

String Stability of Aircraft Formation

Using initial LQR design



Effectively limits how long formations can be

String Stability: Definitions

Consider a cascaded system with an equilibrium at the origin:

$$\dot{x}_i = f(x_i, x_{i-1}, \ldots, x_{i-r+1}), \quad i \in \mathcal{N}$$

- **LTI String Stability:** Let $P(s) := X_i(s)/X_{i-1}(s)$. Then we have string stability if $\sup_\omega |P(j\omega)| < 1$.
- $\mathcal{L}_p$ **String Stability**: Defined similar to general $\mathcal{L}_p$ stability [Ploeg et al., 2014].

String Stability: Limitations

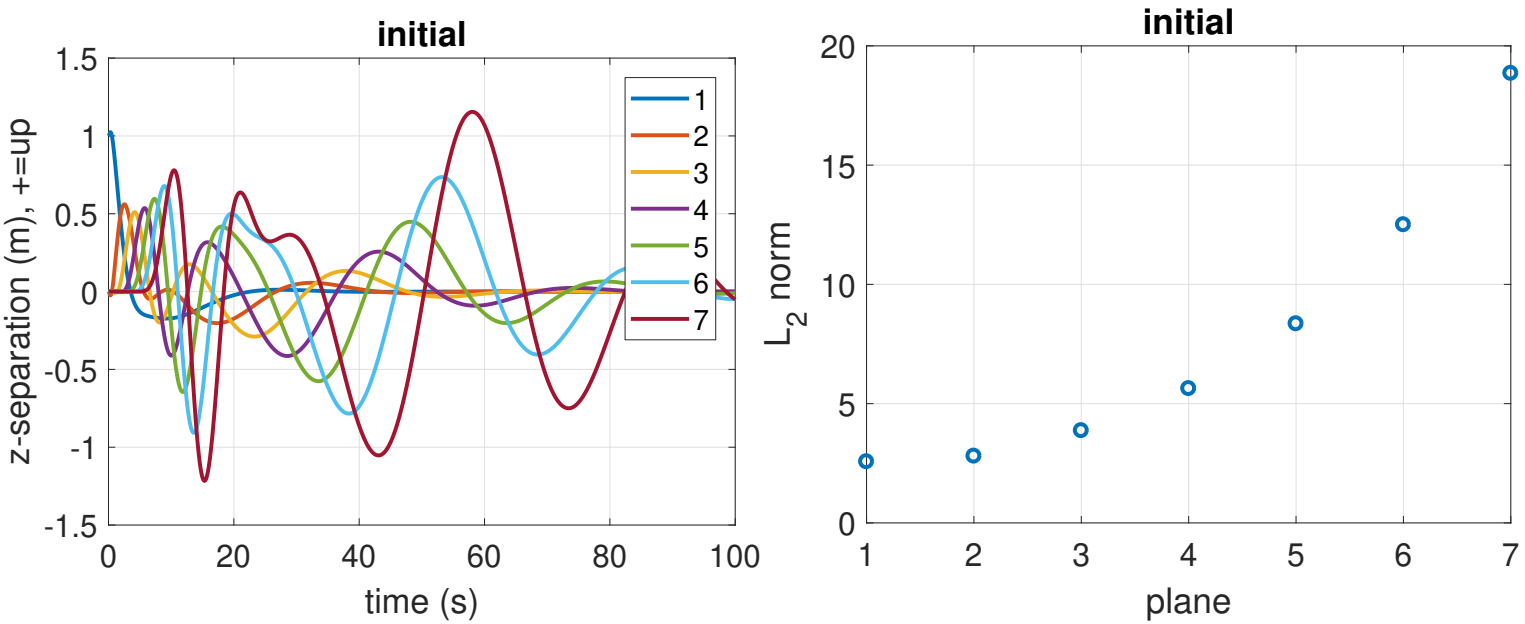
UCLouvain

Hard limit: For LTI systems with two poles at the origin, no linear controller can achieve string stability with a fixed separation distance! [Seiler et al., 2004]

Performance-based limit: Sufficient conditions for exponential growth in disturbance amplification.
[Middleton and Braslavsky, 2010]

String Stability: Disturbance Attenuation

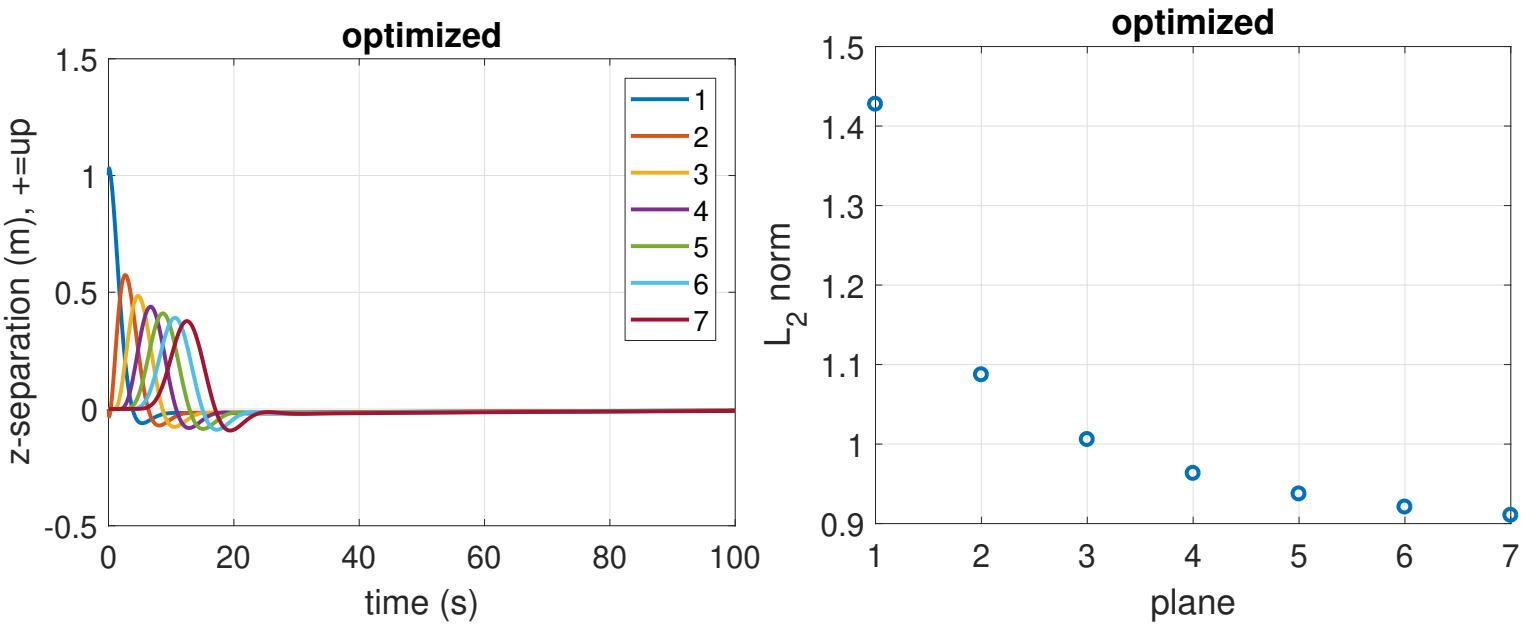
Measure disturbance amplification along string



We want to minimize slope of L_2 norms

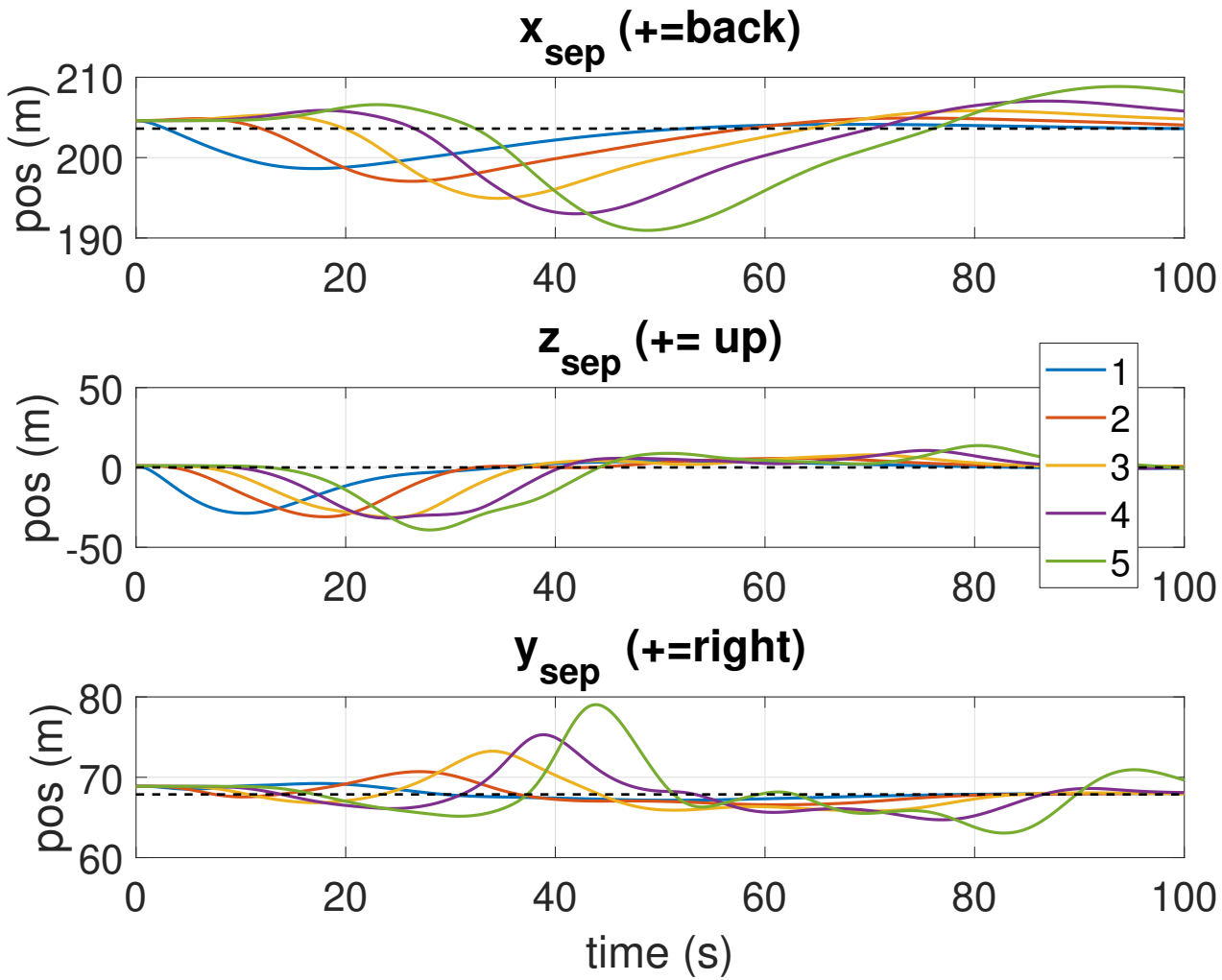
String Stability: Disturbance Attenuation

After optimizing LQR weights (active set method)



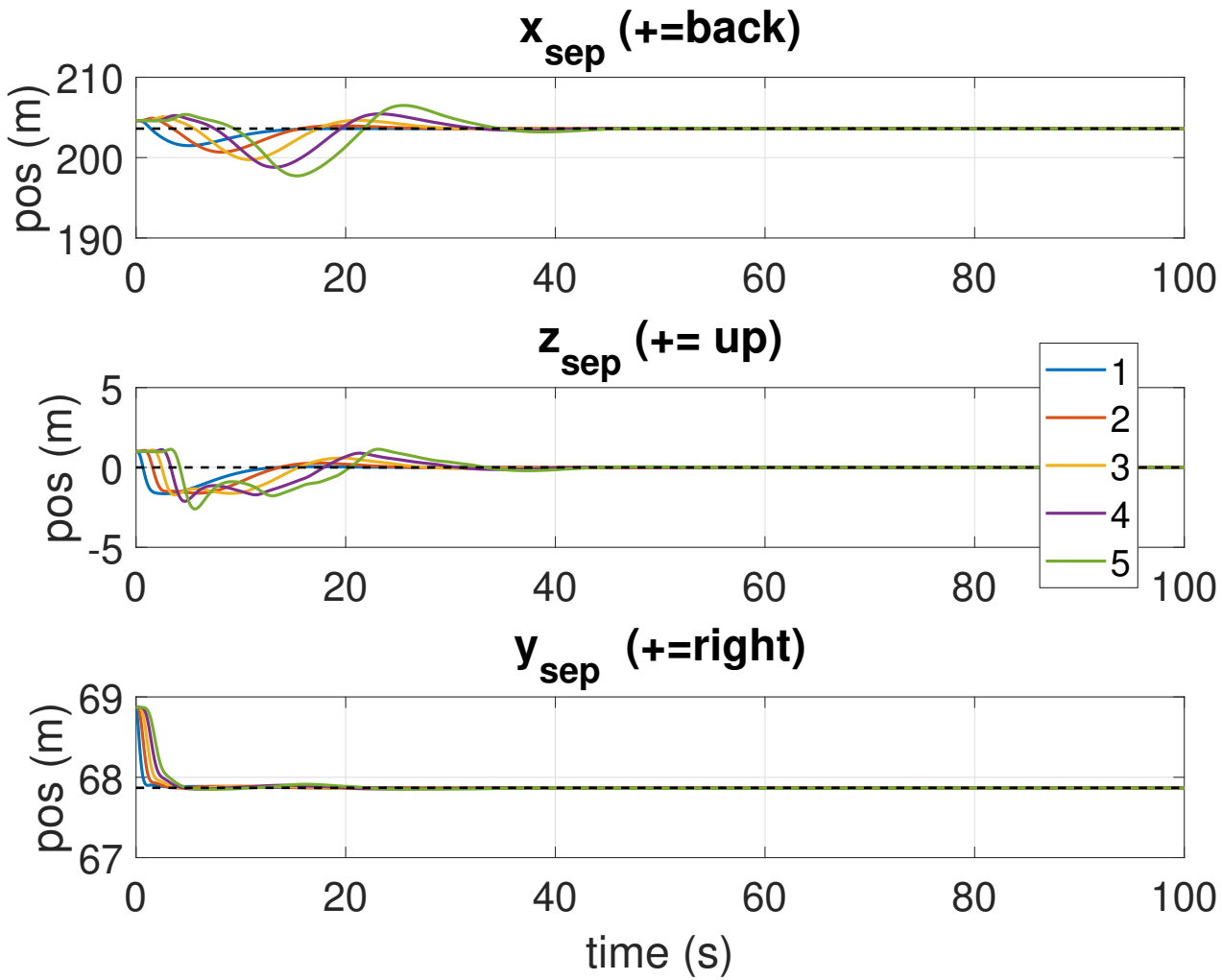
Simulations with Turbulence

Initial controller (Energy savings: -0.1%)



Simulations with Turbulence

Optimized controller (Energy savings: 9.3%)



String Stability: Other Potential Remedies

UCLouvain

1. Use velocity based agent separation distance:

$$e_i(t) = x_{i-1}(t) - x_i(t) + hv_i(t)$$

2. Heterogeneous controller tuning

3. Use state of absolute leader

4. Use control input of agent in front

Conclusions and Future Work

Conclusions:

- Reinforcement learning is a flexible tool which can work for finding the sweet spot
- String stability fixes can stabilize extended formations

Future Work:

- Tune controllers for energy efficiency / string stability trade-off
- Test learning and control on more advanced wake models

References I


Auer, P. (2002).
Using confidence bounds for exploitation-exploration trade-offs.
Journal of Machine Learning Research, 3(Nov):397–422.


Binetti, P., Ariyur, K. B., Krstic, M., and Bernelli, F. (2003).
Formation flight optimization using extremum seeking feedback.
Journal of Guidance, Control, and Dynamics, 26(1):132–142.


Hemati, M. S., Eldredge, J. D., and Speyer, J. L. (2014).
Wake sensing for aircraft formation flight.
Journal of Guidance, Control, and Dynamics, 37(2):513–524.

References II


Middleton, R. H. and Braslavsky, J. H. (2010).
String instability in classes of linear time invariant formation control with limited communication range.
IEEE Transactions on Automatic Control, 55(7):1519–1530.


Ploeg, J., Shukla, D. P., van de Wouw, N., and Nijmeijer, H. (2014).
Controller synthesis for string stability of vehicle platoons.
IEEE Transactions on Intelligent Transportation Systems, 15(2):854–865.


Seiler, P., Pant, A., and Hedrick, K. (2004).
Disturbance propagation in vehicle strings.
IEEE Transactions on automatic control, 49(10):1835–1842.

References III


Sutton, R. S., Barto, A. G., et al. (1998).
Introduction to reinforcement learning, volume 135.
MIT press Cambridge.


Voelkl, B., Portugal, S. J., Unsöld, M., Usherwood, J. R., Wilson, A. M., and Fritz, J. (2015).
Matching times of leading and following suggest cooperation through direct reciprocity during v-formation flight in ibis.
Proceedings of the National Academy of Sciences, 112(7):2115–2120.

Backup slides

UCLouvain

hic sunt dracones

Multi-Armed Bandits

UCLouvain

- Many slot machines
  - How do you find the most rewarding machine?
- Confidence bounds
  - What are the odds $p = \mathbb{P}[|\hat{R} - \bar{R}| < u]$?
  - Black magic yields confidence bound $u(p, N)$
- Policy
  - Pick the bandit that maximizes $\hat{R} + u$, get reward
  - Update your visit counter N for that bandit
  - Update your reward estimate: $\hat{R} \leftarrow \hat{R}(1 - \alpha) + \alpha R$
- Common heuristic
  - Test each bandit once, initial estimate of $\hat{R}$ is R

Upper Confidence Bounds

For simplicity, define $\hat{R}_N$ as the current weighted average: $\hat{R}_N = \sum_{i=1}^{N} c_i R_i$. Also define $S_1 = \sum_{i=1}^{N} c_i$ and $S_2 = \sum_{i=1}^{N} c_i^2$, with i.i.d. rewards R (with zero mean, variance σ)

$$\begin{aligned}\mathbb{E}[\hat{R}_N] &= \mathbb{E}\left[\sum_{i=1}^{N} c_i R_i\right] \\ &= \sum_{i=1}^{N} c_i \mathbb{E}\left[R_i\right] \\ &= 0 S_1 \\ &= 0\end{aligned}$$

$$\begin{aligned}\mathrm{Var}[\hat{R}_N] &= \mathbb{E}\left[\left(\sum_{i=1}^{N} c_i R_i\right)^2\right] \\ &= \sum_{i=1}^{N} c_i^2 \mathbb{E}\left[R_i^2\right] \\ &= \sigma^2 \sum_{i=1}^{N} c_i^2 \\ &= \sigma^2 S_2\end{aligned}$$

Upper Confidence Bounds

UCLouvain

We'll use a Chernoff bound

$$\begin{aligned}\Pr[\hat{R}_N \geq u] &= \Pr\left[\exp(s\hat{R}_N) \geq \exp(su)\right] \\ &\leq e^{-su}\mathbb{E}[\exp(s\hat{R}_N)] \\ &\leq \exp\left(-\frac{2u^2}{(R_{max} - R_{min})^2 S_2}\right)\end{aligned}$$

Where s was chosen to give the tightest bound. If p is the maximum probability, then the uncertainty is:

$$u = |R_{max} - R_{min}|\sqrt{\frac{-\log(p)S_2}{2}}$$

Averages

Running average

$$\hat{R}_N = \left(\frac{N-1}{N}\right)\hat{R}_{N-1} + \left(\frac{1}{N}\right)R_N$$

Exponential Recency Weighted Average

$$\hat{R}_N = (1-\alpha)\hat{R}_{N-1} + (\alpha)R_N$$

$$u_{running} = |b-a|\sqrt{\frac{-\log(p)}{2}}\left(\frac{1}{N}\right)^{1/2}$$

$$u_{discount} = |b-a|\sqrt{\frac{-\log(p)}{2}}\left(\frac{\alpha}{2-\alpha} + \frac{2(1-\alpha)^{2N}}{(1-\alpha)(2-\alpha)}\right)^{1/2}$$

$$u_{running} \leq u_{discount}$$

The running average gives a tighter bound because the discounted average puts too much weight on recent measurements. This is a benefit of discounting, however, as it essentially forgets very old measurements. This is the trade-off for some perpetual uncertainty.