Regret Bounds for the Adaptive Control of Linear Quadratic Systems

Yasin Abbasi-Yadkori  Csaba Szepesvári

University of Alberta
E-mail: szepesva@ualberta.ca

Budapest, July 10, 2011
Outline

1. Problem formulation

2. LQR
   - Main ideas
   - Some details (and the algorithm)

3. Proof sketch (and the result)

4. Conclusions and Open Problems

5. Bibliography

COLT 2011 (Budapest)
Control problems

Agent

Environment

\[
\text{Agent} \quad \text{Environment}
\]

\[
\begin{align*}
\text{Control problems} & \\
\text{Agent} & \quad \text{Environment}
\end{align*}
\]

\[
\sum_{t} c(x_t, u_t) \rightarrow \min
\]

\[
x_{t+1} = f(x_t, u_t, w_t)
\]

COLT 2011 (Budapest)
Control problems

\[
\begin{align*}
\text{Agent} & \quad x_t \\
\text{Environment} & \quad f(x_t, u_t, w_t) \\
\sum_{t} c(x_t, u_t) & \rightarrow \min
\end{align*}
\]

COLT 2011 (Budapest)
Control problems

\[
\begin{align*}
\text{Agent} & \quad \text{Environment} \\
 x_t & \quad u_t \\
 x_{t+1} &= f(x_t, u_t, w_t) \\
\sum_t c(x_t, u_t) &\to \min
\end{align*}
\]

COLT 2011 (Budapest)
Control problems

\[ x_{t+1} = f(x_t, u_t, w_t) \]

\[ \sum_t c(x_t, u_t) \rightarrow \min \]
Control problems

\[ x_{t+1} = f(x_t, u_t, w_t) \]

\[ \sum_t c(x_t, u_t) \rightarrow \min \]
Learning to control

Agent

Environment

$u_t$

$x_t$

$x_{t+1} = f(x_t, u_t, w_t)$

$\sum_t c(x_t, u_t) \rightarrow \min$

$f$ is unknown – yet the goal is to control the environment almost as well as if it was known
Measure of performance of the learner

- Does the average cost converge to the optimal average cost?

\[
\frac{1}{T} \sum_{t=1}^{T} c(x_t, u_t) \rightarrow J^* ?
\]

- How fast is the convergence?
- Compare the total losses ⇒ Regret:

\[
R_T = \sum_{t=1}^{T} c(x_t, u_t) - TJ^* .
\]

- Hannan consistency:

\[
\frac{R_T}{T} \rightarrow 0 \text{ as } T \rightarrow \infty
\]

- Typical result: For some \( \gamma \in (0, 1) \),

\[
R_T = O(T^\gamma) .
\]
Measure of performance of the learner

Does the average cost converge to the optimal average cost?

\[ \frac{1}{T} \sum_{t=1}^{T} c(x_t, u_t) \rightarrow J^* \ ? \]

How fast is the convergence?

Compare the total losses ⇒ Regret:

\[ R_T = \sum_{t=1}^{T} c(x_t, u_t) - TJ^* . \]

Hannan consistency:

\[ \frac{R_T}{T} \rightarrow 0 \text{ as } T \rightarrow \infty \]

Typical result: For some \( \gamma \in (0, 1) \),

\[ R_T = O(T^\gamma) . \]
Measure of performance of the learner

- Does the average cost converge to the optimal average cost?

\[
\frac{1}{T} \sum_{t=1}^{T} c(x_t, u_t) \to J^* \quad ?
\]

- How fast is the convergence?
- Compare the total losses ⇒ Regret:

\[
R_T = \sum_{t=1}^{T} c(x_t, u_t) - TJ^* .
\]

- **Hannan consistency:**

\[
\frac{R_T}{T} \to 0 \text{ as } T \to \infty
\]

- Typical result: For some \( \gamma \in (0, 1) \),

\[
R_T = O(T^\gamma) .
\]
Measure of performance of the learner

- Does the average cost converge to the optimal average cost?
  \[
  \frac{1}{T} \sum_{t=1}^{T} c(x_t, u_t) \to J^* 
  \]

- How fast is the convergence?
- Compare the total losses \( \Rightarrow \) Regret:
  \[
  R_T = \sum_{t=1}^{T} c(x_t, u_t) - TJ^*. 
  \]

- **Hannan consistency:**
  \[
  \frac{R_T}{T} \to 0 \text{ as } T \to \infty 
  \]

- Typical result: For some \( \gamma \in (0, 1) \),
  \[
  R_T = O(T^\gamma). 
  \]
Measure of performance of the learner

- Does the average cost converge to the optimal average cost?
  \[
  \frac{1}{T} \sum_{t=1}^{T} c(x_t, u_t) \to J^*?
  \]

- How fast is the convergence?
- Compare the total losses ⇒ Regret:
  \[
  R_T = \sum_{t=1}^{T} c(x_t, u_t) - TJ^*.
  \]

- Hannan consistency:
  \[
  \frac{R_T}{T} \to 0 \text{ as } T \to \infty
  \]

- Typical result: For some \( \gamma \in (0, 1) \),
  \[
  R_T = O(T^\gamma).
  \]
This talk: Linear Quadratic Regulation

- **Linear dynamics:** $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^d$.
  \[
  f(x_t, u_t, w_{t+1}) = A_* x_t + B_* u_t.
  \]

- **Quadratic cost:** $Q, R \succ 0$
  \[
  c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t.
  \]

- **Noise** $(w_t)_t$: Subgaussian martingale noise, $\mathbb{E} \left[ w_{t+1} w_{t+1}^\top | \mathcal{F}_t \right] = I_n$.

- **LQR problem:** given $A_*, B_*, Q, R$, find an optimal controller

- **LQR learning problem:** given $Q, R$, not knowing $A_*, B_*$, learn to control the system
This talk: Linear Quadratic Regulation

- Linear dynamics: $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^d$.

$$f(x_t, u_t, w_{t+1}) = A_* x_t + B_* u_t.$$  

- Quadratic cost: $Q, R \succ 0$

$$c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t.$$  

- Noise $(w_t)_t$: Subgaussian martingale noise, $\mathbb{E} [w_{t+1} w_{t+1}^\top | \mathcal{F}_t] = I_n$.

- LQR problem: given $A_*, B_*, Q, R$, find an optimal controller

- LQR learning problem: given $Q, R$, not knowing $A_*, B_*$, learn to control the system
This talk: Linear Quadratic Regulation

- **Linear dynamics:** \( x_t \in \mathbb{R}^n, u_t \in \mathbb{R}^d. \)
  \[
f(x_t, u_t, w_{t+1}) = A_* x_t + B_* u_t.
  \]

- **Quadratic cost:** \( Q, R \succ 0 \)
  \[
c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t.
  \]

- **Noise \( (w_t) \):** Subgaussian martingale noise, \( \mathbb{E} [w_{t+1} w_{t+1}^\top | \mathcal{F}_t] = I_n. \)

- **LQR problem:** given \( A_*, B_*, Q, R, \) find an optimal controller
- **LQR learning problem:** given \( Q, R. \) not knowing \( A_*, B_* \), learn to control the system
This talk: Linear Quadratic Regulation

- Linear dynamics: $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^d$.

$$f(x_t, u_t, w_{t+1}) = A_* x_t + B_* u_t.$$ 

- Quadratic cost: $Q, R \succ 0$

$$c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t.$$ 

- Noise $(w_t)_t$: Subgaussian martingale noise, $\mathbb{E} [w_{t+1} w_{t+1}^\top | \mathcal{F}_t] = I_n$.

- LQR problem: given $A_*, B_*, Q, R$, find an optimal controller.

- LQR learning problem: given $Q, R$, not knowing $A_*, B_*$, learn to control the system.
This talk: Linear Quadratic Regulation

- Linear dynamics: \( x_t \in \mathbb{R}^n, u_t \in \mathbb{R}^d. \)

\[
f(x_t, u_t, w_{t+1}) = A_* x_t + B_* u_t.
\]

- Quadratic cost: \( Q, R \succ 0 \)

\[
c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t.
\]

- Noise \( (w_t)_t \): Subgaussian martingale noise, \( \mathbb{E} \left[ w_{t+1} w_{t+1}^\top | \mathcal{F}_t \right] = I_n. \)

- LQR problem: given \( A_*, B_*, Q, R \), find an optimal controller

- LQR learning problem: given \( Q, R \), not knowing \( A_*, B_* \), learn to control the system
The goal and why should we care?

- **Goal**: Design a controller which achieves low regret for a reasonably large class of LQR problems.
  - Simple $\equiv$ beautiful, nice structures!
  - Continuous states and controls!
  - LQR control is actually useful! (even when no learning is involved)
  - Unsolved problem!? 
The goal and why should we care?

- **Goal**: Design a controller which achieves low regret for a reasonably large class of LQR problems.

- **Simple ≡ beautiful, nice structures!**
  - Continuous states and controls!
  - LQR control is actually useful! (even when no learning is involved)
  - Unsolved problem!?
The goal and why should we care?

- **Goal**: Design a controller which achieves low regret for a reasonably large class of LQR problems.

- Simple $\equiv$ beautiful, nice structures!
- Continuous states and controls!
- LQR control is actually useful! (even when no learning is involved)
- Unsolved problem!?
The goal and why should we care?

- **Goal**: Design a controller which achieves low regret for a reasonably large class of LQR problems.
- Simple $\equiv$ beautiful, nice structures!
- Continuous states and controls!
- LQR control is actually useful! (even when no learning is involved)
- Unsolved problem!?
The goal and why should we care?

- **Goal**: Design a controller which achieves low regret for a reasonably large class of LQR problems.
- Simple ⇔ beautiful, nice structures!
- Continuous states and controls!
- LQR control is actually useful! (even when no learning is involved)
- Unsolved problem!? 
Previous works

- Bartlett and Tewari (2009); Auer et al. (2010) – regret analysis of finite MDPs
- Control people!
  - Lai and Wei (1982b, 1987); Chen and Guo (1987); Chen and Zhang (1990); Lai and Ying (2006) – consistency, forced exploration (like ε-greedy)
  - Campi and Kumar (1998); Bittanti and Campi (2006) – consistency, basis of the present work
- Lai and Robbins (1985) – principle in the face of uncertainty for bandits
- Lai and Wei (1982a); Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010) – linear estimation, tail inequalities
Previous works

- Bartlett and Tewari (2009); Auer et al. (2010) – regret analysis of finite MDPs
- Control people!
  - Lai and Wei (1982b, 1987); Chen and Guo (1987); Chen and Zhang (1990); Lai and Ying (2006) – consistency, forced exploration (like $\varepsilon$-greedy)
  - Campi and Kumar (1998); Bittanti and Campi (2006) – consistency, basis of the present work
- Lai and Robbins (1985) – principle in the face of uncertainty for bandits
- Lai and Wei (1982a); Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010) – linear estimation, tail inequalities
Previous works

- Bartlett and Tewari (2009); Auer et al. (2010) – regret analysis of finite MDPs
- Control people!
  - Lai and Wei (1982b, 1987); Chen and Guo (1987); Chen and Zhang (1990); Lai and Ying (2006) – consistency, forced exploration (like $\varepsilon$-greedy)
  - Campi and Kumar (1998); Bittanti and Campi (2006) – consistency, basis of the present work
- Lai and Robbins (1985) – principle in the face of uncertainty for bandits
- Lai and Wei (1982a); Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010) – linear estimation, tail inequalities
Previous works

- Bartlett and Tewari (2009); Auer et al. (2010) – regret analysis of finite MDPs
- Control people!
  - Lai and Wei (1982b, 1987); Chen and Guo (1987); Chen and Zhang (1990); Lai and Ying (2006) – consistency, forced exploration (like $\epsilon$-greedy)
  - Campi and Kumar (1998); Bittanti and Campi (2006) – consistency, basis of the present work
- Lai and Robbins (1985) – principle in the face of uncertainty for bandits
- Lai and Wei (1982a); Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010) – linear estimation, tail inequalities
Previous works

- Bartlett and Tewari (2009); Auer et al. (2010) – regret analysis of finite MDPs
- Control people!
  - Lai and Wei (1982b, 1987); Chen and Guo (1987); Chen and Zhang (1990); Lai and Ying (2006) – consistency, forced exploration (like $\varepsilon$-greedy)
  - Campi and Kumar (1998); Bittanti and Campi (2006) – consistency, basis of the present work
- Lai and Robbins (1985) – principle in the face of uncertainty for bandits
- Lai and Wei (1982a); Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010) – linear estimation, tail inequalities
Previous works

- Bartlett and Tewari (2009); Auer et al. (2010) – regret analysis of finite MDPs
- Control people!
  - Lai and Wei (1982b, 1987); Chen and Guo (1987); Chen and Zhang (1990); Lai and Ying (2006) – consistency, forced exploration (like $\varepsilon$-greedy)
  - Campi and Kumar (1998); Bittanti and Campi (2006) – consistency, basis of the present work
- Lai and Robbins (1985) – principle in the face of uncertainty for bandits
- Lai and Wei (1982a); Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010) – linear estimation, tail inequalities
Previous works

- Bartlett and Tewari (2009); Auer et al. (2010) – regret analysis of finite MDPs
- Control people!
  - Lai and Wei (1982b, 1987); Chen and Guo (1987); Chen and Zhang (1990); Lai and Ying (2006) – consistency, forced exploration (like $\varepsilon$-greedy)
  - Campi and Kumar (1998); Bittanti and Campi (2006) – consistency, basis of the present work
- Lai and Robbins (1985) – principle in the face of uncertainty for bandits
- Lai and Wei (1982a); Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010) – linear estimation, tail inequalities
Previous works

- Bartlett and Tewari (2009); Auer et al. (2010) – regret analysis of finite MDPs
- Control people!
  - Lai and Wei (1982b, 1987); Chen and Guo (1987); Chen and Zhang (1990); Lai and Ying (2006) – consistency, forced exploration (like $\varepsilon$-greedy)
  - Campi and Kumar (1998); Bittanti and Campi (2006) – consistency, basis of the present work
- Lai and Robbins (1985) – principle in the face of uncertainty for bandits
- Lai and Wei (1982a); Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010) – linear estimation, tail inequalities
Outline

1. Problem formulation
2. LQR
   - Main ideas
   - Some details (and the algorithm)
3. Proof sketch (and the result)
4. Conclusions and Open Problems
5. Bibliography
The main ideas of the algorithm

- Estimate the system dynamics
- Be optimistic in selecting the controls
- Avoid frequent changes to the policy
The main ideas of the algorithm

- Estimate the system dynamics
- Be optimistic in selecting the controls
- Avoid frequent changes to the policy
The main ideas of the algorithm

- Estimate the system dynamics
- Be optimistic in selecting the controls
- Avoid frequent changes to the policy
Estimation

\[ x_{t+1} = A_* x_t + B_* u_t + w_{t+1} \]
\[ = \Theta_* \begin{pmatrix} x_t \\ u_t \end{pmatrix} + w_{t+1} \]
\[ = \Theta_* z_t + w_{t+1} \]

- Data: \((z_0, x_1), (z_2, x_2), \ldots, (z_{t-1}, x_t)\)
- \(x_{i+1} = \Theta_* z_i + w_{i+1}\)
- Linear regression with correlated covariates, martingale noise
- \(\Rightarrow\) Use ridge-regression (least-squares, with \(\ell_2\)-penalties)
Estimation

\[ x_{t+1} = A_x x_t + B_u u_t + w_{t+1} \]

\[ = \Theta_z (x_t, u_t) + w_{t+1} \]

\[ = \Theta_z z_t + w_{t+1} \]

- Data: \((z_0, x_1), (z_2, x_2), \ldots, (z_{t-1}, x_t)\)
- \(x_{i+1} = \Theta_z z_i + w_{i+1}\)
- Linear regression with correlated covariates, martingale noise
- \(\Rightarrow\) Use ridge-regression (least-squares, with \(\ell^2\)-penalties)
Estimation

\[ x_{t+1} = A_x x_t + B u_t + w_{t+1} \]
\[ = \Theta_x \begin{pmatrix} x_t \\ u_t \end{pmatrix} + w_{t+1} \]
\[ = \Theta_x z_t + w_{t+1} \]

- **Data:** \((z_0, x_1), (z_2, x_2), \ldots, (z_{t-1}, x_t)\)
- \(x_{i+1} = \Theta_x z_i + w_{i+1}\)
- Linear regression with correlated covariates, martingale noise
- \(\Rightarrow\) Use ridge-regression (least-squares, with \(\ell^2\)-penalties)
Estimation

\[ x_{t+1} = A_x x_t + B_u u_t + w_{t+1} \]

\[ = \Theta_x (x_t) + w_{t+1} \]

\[ = \Theta_x z_t + w_{t+1} \]

- **Data:** \((z_0, x_1), (z_2, x_2), \ldots, (z_{t-1}, x_t)\)
- \(x_{i+1} = \Theta_x z_i + w_{i+1}\)
- Linear regression with correlated covariates, martingale noise
- \(\Rightarrow\) Use ridge-regression (least-squares, with \(\ell^2\)-penalties)
Estimation

\[ x_{t+1} = A_x x_t + B_u u_t + w_{t+1} \]
\[ = \Theta_x \begin{pmatrix} x_t \\ u_t \end{pmatrix} + w_{t+1} \]
\[ = \Theta_z z_t + w_{t+1} \]

- **Data:** \((z_0, x_1), (z_2, x_2), \ldots, (z_{t-1}, x_t)\)
- \(x_{i+1} = \Theta_z z_i + w_{i+1}\)
- Linear regression with correlated covariates, martingale noise
  - Use ridge-regression (least-squares, with \(\ell^2\)-penalties)
Estimation

\[ x_{t+1} = A_x x_t + B_u u_t + w_{t+1} \]

\[ = \Theta_z \begin{pmatrix} x_t \\ u_t \end{pmatrix} + w_{t+1} \]

\[ = \Theta_z z_t + w_{t+1} \]

- **Data:** \((z_0, x_1), (z_2, x_2), \ldots, (z_{t-1}, x_t)\)
- \(x_{i+1} = \Theta_z z_i + w_{i+1}\)
- Linear regression with correlated covariates, martingale noise
- \(\Rightarrow\) Use ridge-regression (least-squares, with \(\ell^2\)-penalties)
Optimism principle

**Optimism Principle**

Let $C_t(\delta)$ be a confidence set for the unknown parameters. Choose the control which gives rise to the best performance.

For given $\Theta$, for the linear system with parameter $\Theta$, let $J(\Theta)$ be the optimal average cost and $\pi_\Theta$ be the corresponding optimal policy. Choose

$$\tilde{\Theta}_t = \arg\min_{\theta \in C_t(\delta)} J(\theta) \quad \text{and} \quad u_t = \pi_{\tilde{\Theta}_t}(x_t).$$

**Caveats**

- $J(\Theta)$ can be ill-defined
- Need restriction on allowed set of parameters
- Finding $\tilde{\Theta}_t$ is a potentially difficult optimization problem
Let $C_t(\delta)$ be a confidence set for the unknown parameters. Choose the control which gives rise to the best performance.

For given $\Theta$, for the linear system with parameter $\Theta$, let $J(\Theta)$ be the optimal average cost and $\pi_\Theta$ be the corresponding optimal policy. Choose

$$\tilde{\Theta}_t = \arg \min_{\theta \in C_t(\delta)} J(\theta) \quad \text{and} \quad u_t = \pi_{\tilde{\Theta}_t}(x_t).$$

Caveats

- $J(\Theta)$, $\pi_\Theta$ can be ill-defined.
- Need restriction on allowed set of parameters.
- Finding $\tilde{\Theta}_t$ is a potentially difficult optimization problem.
Optimism principle

**Optimism Principle**

Let \( C_t(\delta) \) be a confidence set for the unknown parameters. Choose the control which gives rise to the best performance.

For given \( \Theta \), for the linear system with parameter \( \Theta \), let \( J(\Theta) \) be the optimal average cost and \( \pi_\Theta \) be the corresponding optimal policy. Choose

\[
\tilde{\Theta}_t = \arg \min_{\theta \in C_t(\delta)} J(\theta) \quad \text{and} \quad u_t = \pi_{\tilde{\Theta}_t}(x_t).
\]

**Caveats**

- \( J(\Theta), \pi_\Theta \) can be ill-defined
- Need restriction on allowed set of parameters
- Finding \( \tilde{\Theta}_t \) is a potentially difficult optimization problem
Optimism principle

Optimism Principle

Let $C_t(\delta)$ be a confidence set for the unknown parameters. Choose the control which gives rise to the best performance.

For given $\Theta$, for the linear system with parameter $\Theta$, let $J(\Theta)$ be the optimal average cost and $\pi_\Theta$ be the corresponding optimal policy. Choose

$$\tilde{\Theta}_t = \arg \min_{\theta \in C_t(\delta)} J(\theta) \quad \text{and} \quad u_t = \pi_{\tilde{\Theta}_t}(x_t).$$

Caveats

- $J(\Theta)$, $\pi_{\Theta}$ can be ill-defined
  Need restriction on allowed set of parameters
- Finding $\tilde{\Theta}_t$ is a potentially difficult optimization problem
Optimism principle

Optimism Principle

Let $C_t(\delta)$ be a confidence set for the unknown parameters. Choose the control which gives rise to the best performance.

For given $\Theta$, for the linear system with parameter $\Theta$, let $J(\Theta)$ be the optimal average cost and $\pi_\Theta$ be the corresponding optimal policy. Choose

$$\tilde{\Theta}_t = \arg \min_{\theta \in C_t(\delta)} J(\theta) \quad \text{and} \quad u_t = \pi_{\tilde{\Theta}_t}(x_t).$$

Caveats

- $J(\Theta)$, $\pi_\Theta$ can be ill-defined
  - Need restriction on allowed set of parameters
- Finding $\tilde{\Theta}_t$ is a potentially difficult optimization problem
Optimism principle

Optimism Principle

Let $C_t(\delta)$ be a confidence set for the unknown parameters. Choose the control which gives rise to the best performance.

For given $\Theta$, for the linear system with parameter $\Theta$, let $J(\Theta)$ be the optimal average cost and $\pi_\Theta$ be the corresponding optimal policy. Choose

$$\tilde{\Theta}_t = \arg \min_{\theta \in C_t(\delta)} J(\theta)$$

and

$$u_t = \pi_{\tilde{\Theta}_t}(x_t).$$

Caveats

- $J(\Theta), \pi_\Theta$ can be ill-defined
  - Need restriction on allowed set of parameters
- Finding $\tilde{\Theta}_t$ is a potentially difficult optimization problem
Optimism Principle

Let $C_t(\delta)$ be a confidence set for the unknown parameters. Choose the control which gives rise to the best performance.

For given $\Theta$, for the linear system with parameter $\Theta$, let $J(\Theta)$ be the optimal average cost and $\pi_\Theta$ be the corresponding optimal policy. Choose

$$\tilde{\Theta}_t = \arg \min_{\theta \in C_t(\delta)} J(\theta)$$
and
$$u_t = \pi_{\tilde{\Theta}_t}(x_t).$$

Caveats

- $J(\Theta)$, $\pi_\Theta$ can be ill-defined
- Need restriction on allowed set of parameters
- Finding $\tilde{\Theta}_t$ is a potentially difficult optimization problem
Avoiding frequent changes

- Frequent changes are unnecessary
- Saving computation $\implies$ going green!?
- Frequent changes might be a problem (avoiding frequent changes helps with the proof)
Avoiding frequent changes

- Frequent changes are unnecessary
- Saving computation \(\Rightarrow\) going green!?
- Frequent changes might be a problem (avoiding frequent changes helps with the proof)
Avoiding frequent changes

- Frequent changes are unnecessary
- Saving computation \(\Rightarrow\) going green!?
- Frequent changes might be a problem (avoiding frequent changes helps with the proof)
Outline

1. Problem formulation

2. LQR
   - Main ideas
   - Some details (and the algorithm)

3. Proof sketch (and the result)

4. Conclusions and Open Problems

5. Bibliography
How to choose the confidence set?


**Theorem**

Let \( z_t^T = (x_t^T, u_t^T) \in \mathbb{R}^{n+d} \). Let \( \hat{\Theta}_t \) be the ridge-regression parameter estimate with regularization coefficient \( \lambda > 0 \). Let \( V_t = \lambda I + \sum_{i=0}^{t-1} z_i z_i^T \) be the covariance matrix. Then, for any \( 0 < \delta < 1 \), with probability at least \( 1 - \delta \),

\[
\text{trace}((\hat{\Theta}_t - \Theta_*)^T V_t (\hat{\Theta}_t - \Theta_*)) \leq \left(d \sqrt{2 \log \left( \frac{\det(V_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right) + \lambda^{1/2} S^2} \right)^2.
\]
Construction of confidence sets

An ellipsoid centred at $\hat{\Theta}_t$:

$$\text{trace} \left\{ (\Theta - \hat{\Theta}_t)^\top V_t (\Theta - \hat{\Theta}_t) \right\} \leq \beta_t.$$
The algorithm

**Inputs:** $T, S > 0, \delta > 0, Q, L$.
Set $V_0 = I$ and $\hat{\Theta}_0 = 0$, $(\tilde{A}_0, \tilde{B}_0) = \tilde{\Theta}_0 = \text{argmin}_{\Theta \in \mathcal{C}_0(\delta)} J_*(\Theta)$.

for $t := 0, 1, 2, \ldots$ do

Calculate $\hat{\Theta}_t$.

$\hat{\Theta}_t = \text{argmin}_{\Theta \in \mathcal{C}_t(\delta)} J_*(\Theta)$.

Calculate $u_t$ based on the current parameters, $u_t = K(\hat{\Theta}_t)x_t$.

Execute control, observe new state $x_{t+1}$.

$V_{t+1} := V_t + z_t z_t^\top$, where $z_t^\top = (x_t^\top, u_t^\top)$.

end for
Proof sketch

- Fix $T > 0$.
  - With high probability, the state stays $O(\log T)$. ⇒ most of the work is here.
- Decompose the regret
- Analyze each term
Proof sketch

- Fix $T > 0$.
- With high probability, the state stays $O(\log T)$. $\Leftarrow$ most of the work is here..
- Decompose the regret
- Analyze each term
Proof sketch

- Fix $T > 0$.
- With high probability, the state stays $O(\log T)$. ⇐ most of the work is here..
- Decompose the regret
- Analyze each term
Proof sketch

- Fix $T > 0$.
- With high probability, the state stays $O(\log T)$. $\Leftarrow$ most of the work is here..
- Decompose the regret
- Analyze each term
Fix $T > 0$.

With high probability, the state stays $O(\log T)$. $\iff$ most of the work is here..

Decompose the regret

Analyze each term
Regret decomposition

- Dynamic programming equations, \( \mathbb{E}[w_{t+1}|\mathcal{F}_t] = 0 \), Algebra ..

\[
R_1 = \sum_{t=0}^{T} \left\{ x_t^\top P(\tilde{\Theta}_t)x_t - \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_{t+1})x_{t+1} | \mathcal{F}_t \right] \right\}
\]

\[
R_2 = \sum_{t=0}^{T} \mathbb{E} \left[ x_{t+1}^\top \left\{ P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t) \right\} x_{t+1} | \mathcal{F}_t \right]
\]

\[
R_3 = \sum_{t=0}^{T} z_t^\top \left( \Theta_\star^\top P(\tilde{\Theta}_t)\Theta_\star - \tilde{\Theta}_t^\top P(\tilde{\Theta}_t)\tilde{\Theta}_t \right) z_t.
\]

\[
\sum_{t=0}^{T} (x_t^\top Qx_t + u_t^\top Ru_t) = \sum_{t=0}^{T} J(\tilde{\Theta}_t) + R_1 + R_2 + R_3
\]

\[
\leq T J(\Theta_\star) + R_1 + R_2 + R_3.
\]
Regret decomposition

- Dynamic programming equations, $\mathbb{E} [w_{t+1}|\mathcal{F}_t] = 0$, Algebra ..

\[
R_1 = \sum_{t=0}^{T} \left\{ x_t^\top P(\tilde{\Theta}_t) x_t - \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_{t+1}) x_{t+1} | \mathcal{F}_t \right] \right\}
\]

\[
R_2 = \sum_{t=0}^{T} \mathbb{E} \left[ x_{t+1}^\top \left\{ P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t) \right\} x_{t+1} | \mathcal{F}_t \right]
\]

\[
R_3 = \sum_{t=0}^{T} z_t^\top \left( \Theta_*^\top P(\tilde{\Theta}_t) \Theta_* - \tilde{\Theta}_t^\top P(\tilde{\Theta}_t) \tilde{\Theta}_t \right) z_t.
\]

\[
\sum_{t=0}^{T} (x_t^\top Q x_t + u_t^\top R u_t) = \sum_{t=0}^{T} J(\tilde{\Theta}_t) + R_1 + R_2 + R_3
\]

\[
\leq T J(\Theta_*) + R_1 + R_2 + R_3.
\]
Regret decomposition

- Dynamic programming equations, $\mathbb{E} [w_{t+1} | F_t] = 0$, Algebra ..

\[
R_1 = \sum_{t=0}^{T} \left\{ x_t^\top P(\tilde{\Theta}_t)x_t - \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_{t+1})x_{t+1} | F_t \right] \right\}
\]

\[
R_2 = \sum_{t=0}^{T} \mathbb{E} \left[ x_{t+1}^\top \left\{ P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t) \right\} x_{t+1} \left| F_t \right. \right]
\]

\[
R_3 = \sum_{t=0}^{T} z_t^\top \left( \Theta_*^\top P(\tilde{\Theta}_t)\Theta_* - \tilde{\Theta}_t^\top P(\tilde{\Theta}_t)\tilde{\Theta}_t \right) z_t.
\]

\[
\sum_{t=0}^{T} (x_t^\top Qx_t + u_t^\top Ru_t) = \sum_{t=0}^{T} J(\tilde{\Theta}_t) + R_1 + R_2 + R_3
\]

\[
\leq T J(\Theta_*) + R_1 + R_2 + R_3.
\]
Regret decomposition

- **Dynamic programming equations**: $\mathbb{E}[w_{t+1}|\mathcal{F}_t] = 0$, Algebra.

\[
R_1 = \sum_{t=0}^{T} \left\{ x_t^\top P(\tilde{\Theta}_t)x_t - \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_{t+1})x_{t+1} | \mathcal{F}_t \right] \right\}
\]

\[
R_2 = \sum_{t=0}^{T} \mathbb{E} \left[ x_{t+1}^\top \left\{ P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t) \right\} x_{t+1} | \mathcal{F}_t \right]
\]

\[
R_3 = \sum_{t=0}^{T} z_t^\top \left( \Theta_*^\top P(\tilde{\Theta}_t)\Theta_* - \tilde{\Theta}_t^\top P(\tilde{\Theta}_t)\tilde{\Theta}_t \right) z_t.
\]

\[
\sum_{t=0}^{T} (x_t^\top Qx_t + u_t^\top Ru_t) = \sum_{t=0}^{T} J(\tilde{\Theta}_t) + R_1 + R_2 + R_3 \\
\leq T J(\Theta_*) + R_1 + R_2 + R_3.
\]
Regret decomposition

- Dynamic programming equations, $\mathbb{E} [w_{t+1} | \mathcal{F}_t] = 0$, Algebra..

$$R_1 = \sum_{t=0}^{T} \left\{ x_t^\top P(\tilde{\Theta}_t)x_t - \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_{t+1})x_{t+1} \bigg| \mathcal{F}_t \right] \right\}$$

$$R_2 = \sum_{t=0}^{T} \mathbb{E} \left[ x_{t+1}^\top \left\{ P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t) \right\} x_{t+1} \bigg| \mathcal{F}_t \right]$$

$$R_3 = \sum_{t=0}^{T} z_t^\top \left( \Theta^*_\top P(\tilde{\Theta}_t)\Theta_* - \tilde{\Theta}_t^\top P(\tilde{\Theta}_t)\tilde{\Theta}_t \right) z_t.$$
Term $R_1$

$$R_1 = \sum_{t=0}^{T} \left\{ x_t^\top P(\tilde{\Theta}_t)x_t - \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_{t+1})x_{t+1} \mid \mathcal{F}_t \right] \right\}$$

- Regrouping
  - Martingale difference sequence
  - State does not explode
Term $R_1$

\[ R_1 = \sum_{t=0}^{T} \left\{ x_t^\top P(\tilde{\Theta}_t)x_t - \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_{t+1})x_{t+1} \middle| \mathcal{F}_t \right] \right\} \]

- Regrouping
- Martingale difference sequence
- State does not explode
Term $R_1$

$$R_1 = \sum_{t=0}^{T} \left\{ x_t^\top P(\tilde{\Theta}_t)x_t - \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_{t+1})x_{t+1} \mid \mathcal{F}_t \right] \right\}$$

- Regrouping
- Martingale difference sequence
- State does not explode
Term $R_3$

$$R_3 = \sum_{t=0}^{T} z_t^\top \left( \Theta_*^\top P(\tilde{\Theta}_t) \Theta_* - \tilde{\Theta}_t^\top P(\tilde{\Theta}_t) \tilde{\Theta}_t \right) z_t.$$  

- Algebra.. Reduce to

$$O(\sqrt{T}) + \left( \sum_t \|P(\tilde{\Theta}_t)(\tilde{\Theta}_t - \Theta_*)^\top z_t\|^2 \right)^{1/2}$$

- More algebra..
- Choice of confidence set
- State does not explode
Term $R_3$

$$R_3 = \sum_{t=0}^{T} z_t^\top \left( \Theta_*^\top P(\tilde{\Theta}_t) \Theta_* - \tilde{\Theta}_t^\top P(\tilde{\Theta}_t) \tilde{\Theta}_t \right) z_t.$$  

- Algebra.. Reduce to

$$O(\sqrt{T}) + \left( \sum_t \|P(\tilde{\Theta}_t)(\tilde{\Theta}_t - \Theta_*)^\top z_t\|^2 \right)^{1/2}$$

- More algebra..
  - Choice of confidence set
  - State does not explode
Term $R_3$

$$R_3 = \sum_{t=0}^{T} z_t^\top \left( \Theta_*^\top P(\tilde{\Theta}_t) \Theta_* - \tilde{\Theta}_t^\top P(\tilde{\Theta}_t) \tilde{\Theta}_t \right) z_t.$$

- Algebra.. Reduce to

$$O(\sqrt{T}) + \left( \sum_t \|P(\tilde{\Theta}_t)(\tilde{\Theta}_t - \Theta_*)^\top z_t\|^2 \right)^{1/2}$$

- More algebra..
- Choice of confidence set
- State does not explode
Term $R_3$

\[ R_3 = \sum_{t=0}^{T} z_t^\top \left( \Theta_*^\top P(\tilde{\Theta}_t) \Theta_* - \tilde{\Theta}_t^\top P(\tilde{\Theta}_t) \tilde{\Theta}_t \right) z_t. \]

- Algebra.. Reduce to

\[ O(\sqrt{T}) + \left( \sum_{t} \left\| P(\tilde{\Theta}_t)(\tilde{\Theta}_t - \Theta_*)^\top z_t \right\|^2 \right)^{1/2} \]

- More algebra..
- Choice of confidence set
- State does not explode
Term $R_3$

\[ R_3 = \sum_{t=0}^{T} z_t^\top \left( \Theta_\ast^\top P(\tilde{\Theta}_t)\Theta_\ast - \tilde{\Theta}_t^\top P(\tilde{\Theta}_t)\tilde{\Theta}_t \right) z_t. \]

- Algebra.. Reduce to

\[ O(\sqrt{T}) + \left( \sum_t \|P(\tilde{\Theta}_t)(\tilde{\Theta}_t - \Theta_\ast)^\top z_t\|^2 \right)^{1/2} \]

- More algebra..

- Choice of confidence set

- State does not explode
Term $R_2$

$$R_2 = \sum_{t=0}^{T} \mathbb{E} \left[ x_{t+1}^\top \left\{ P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t) \right\} x_{t+1} \bigg| \mathcal{F}_t \right]$$

- Cannot analyze this algorithm!
- What if we change the policies rarely?
Term $R_2$

$$R_2 = \sum_{t=0}^{T} \mathbb{E} \left[ x_{t+1}^\top \left\{ P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t) \right\} x_{t+1} \middle| \mathcal{F}_t \right]$$

- Cannot analyze this algorithm!
- What if we change the policies \textit{rarely}?
Change the policy only when the determinant of confidence ellipsoid doubles.

\[ \tau_s : \text{time of } s\text{th policy change.} \]

\[ O(\log T) \text{ policy changes up to time } T. \]

\[ \sum_{t=0}^{T} \mathbb{E} \left[ x_{t+1}^\top (P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t)) x_{t+1} | \mathcal{F}_t \right] \leq O(\log T). \]
The algorithm

**Inputs:** $T, S > 0, \delta > 0, Q, L$.

Set $V_0 = I$ and $\hat{\Theta}_0 = 0$, $(\tilde{A}_0, \tilde{B}_0) = \tilde{\Theta}_0 = \arg\min_{\Theta \in C_0(\delta)} J_\ast(\Theta)$.

**for** $t := 0, 1, 2, \ldots$ **do**

- **if** $\det(V_t) > 2 \det(V_0)$ **then**
  - Calculate $\hat{\Theta}_t$.
  - $\tilde{\Theta}_t = \arg\min_{\Theta \in C_t(\delta)} J_\ast(\Theta)$.
  - Let $V_0 = V_t$.
- **else**
  - $\tilde{\Theta}_t = \tilde{\Theta}_{t-1}$.

**end if**

- Calculate $u_t$ based on the current parameters, $u_t = K(\tilde{\Theta}_t)x_t$.
- Execute control, observe new state $x_{t+1}$.
- $V_{t+1} := V_t + z_tz_t^\top$, where $z_t^\top = (x_t^\top, u_t^\top)$.

**end for**
Theorem

With probability at least $1 - \delta$, the regret of the algorithm is bounded as follows:

$$R(T) = \tilde{O} \left( \sqrt{T \log(1/\delta)} \right).$$
Conclusions

- First regret result for the problem of linear optimal control
- Algorithm is too expensive!
  Does there exist a cheaper alternative with similar guarantees?
- Relaxing the martingale noise assumption? ($k^{th}$ order Markov noise? ARMA..)
- Extension to linearly parameterized systems?
  \[ x_{t+1} = \theta^\top \varphi(x_t, u_t) + w_{t+1} \]
- Planning? Learning?
- Unrealizable case?
Conclusions

- First regret result for the problem of linear optimal control
- Algorithm is too expensive!

  Does there exist a cheaper alternative with similar guarantees?
- Relaxing the martingale noise assumption? ($k^{th}$ order Markov noise? ARMA..)
- Extension to linearly parameterized systems?

\[
x_{t+1} = \theta^\top \varphi(x_t, u_t) + w_{t+1}
\]
- Planning? Learning?
- Unrealizable case?
Conclusions

- First regret result for the problem of linear optimal control
- Algorithm is too expensive!
  Does there exist a cheaper alternative with similar guarantees?
- Relaxing the martingale noise assumption? ($k^{th}$ order Markov noise? ARMA..)
- Extension to linearly parameterized systems?
  \[ x_{t+1} = \theta^\top \varphi(x_t, u_t) + w_{t+1} \]
- Planning? Learning?
- Unrealizable case?
Conclusions

- First regret result for the problem of linear optimal control
- Algorithm is too expensive!
  - Does there exist a cheaper alternative with similar guarantees?
- Relaxing the martingale noise assumption? ($k^{th}$ order Markov noise? ARMA..)
- Extension to linearly parameterized systems?
  - $x_{t+1} = \theta^\top \varphi(x_t, u_t) + w_{t+1}$
- Planning? Learning?
- Unrealizable case?
Conclusions

- First regret result for the problem of linear optimal control
- Algorithm is too expensive!
  - Does there exist a cheaper alternative with similar guarantees?
- Relaxing the martingale noise assumption? \( k \text{th order Markov noise? ARMA..} \)
- Extension to linearly parameterized systems?
  \[ x_{t+1} = \theta^\top \varphi(x_t, u_t) + w_{t+1} \]
- Planning? Learning?
- Unrealizable case?
Conclusions

- First regret result for the problem of linear optimal control
- Algorithm is too expensive!
  - Does there exist a cheaper alternative with similar guarantees?
- Relaxing the martingale noise assumption? ($k^{th}$ order Markov noise? ARMA..)
- Extension to linearly parameterized systems?
  \[ x_{t+1} = \theta^T \varphi(x_t, u_t) + w_{t+1} \]
- Planning? Learning?
- Unrealizable case?
Conclusions

- First regret result for the problem of linear optimal control
- Algorithm is too expensive!
  - Does there exist a cheaper alternative with similar guarantees?
- Relaxing the martingale noise assumption? ($k^{th}$ order Markov noise? ARMA..)
- Extension to linearly parameterized systems?
  \[ x_{t+1} = \theta^\top \varphi(x_t, u_t) + w_{t+1} \]
- Planning? Learning?
- Unrealizable case?

COLT 2011 (Budapest)
Conclusions

- First regret result for the problem of linear optimal control
- Algorithm is too expensive!
  - Does there exist a cheaper alternative with similar guarantees?
- Relaxing the martingale noise assumption? ($k^{th}$ order Markov noise? ARMA..)
- Extension to linearly parameterized systems?
  \[ x_{t+1} = \theta^\top \varphi(x_t, u_t) + w_{t+1} \]
- Planning? Learning?
- Unrealizable case?
References


