Min Max Generalization for Deterministic Batch Mode Reinforcement Learning: Relaxation Schemes

Raphael Fonteneau

with Damien Ernst, Bernard Boigelot, Quentin Louveaux

Electrical Engineering and Computer Science Department
University of Liège, Belgium

November, 29th, 2013
Maastricht, The Nederlands

Goal

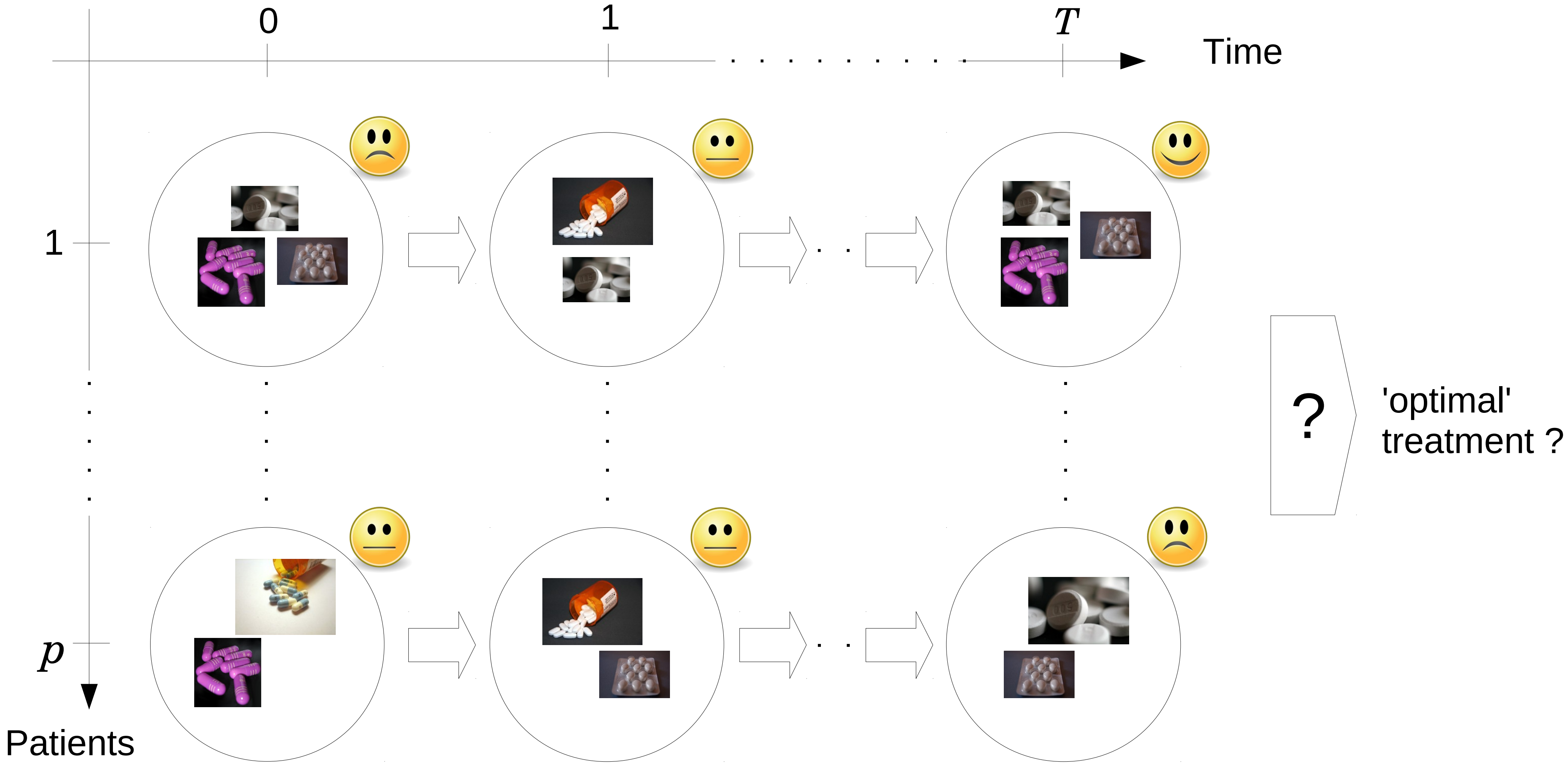
NASA image - public domain – Wikipedia

How to control a system so as to avoid the worst, given the knowledge of:

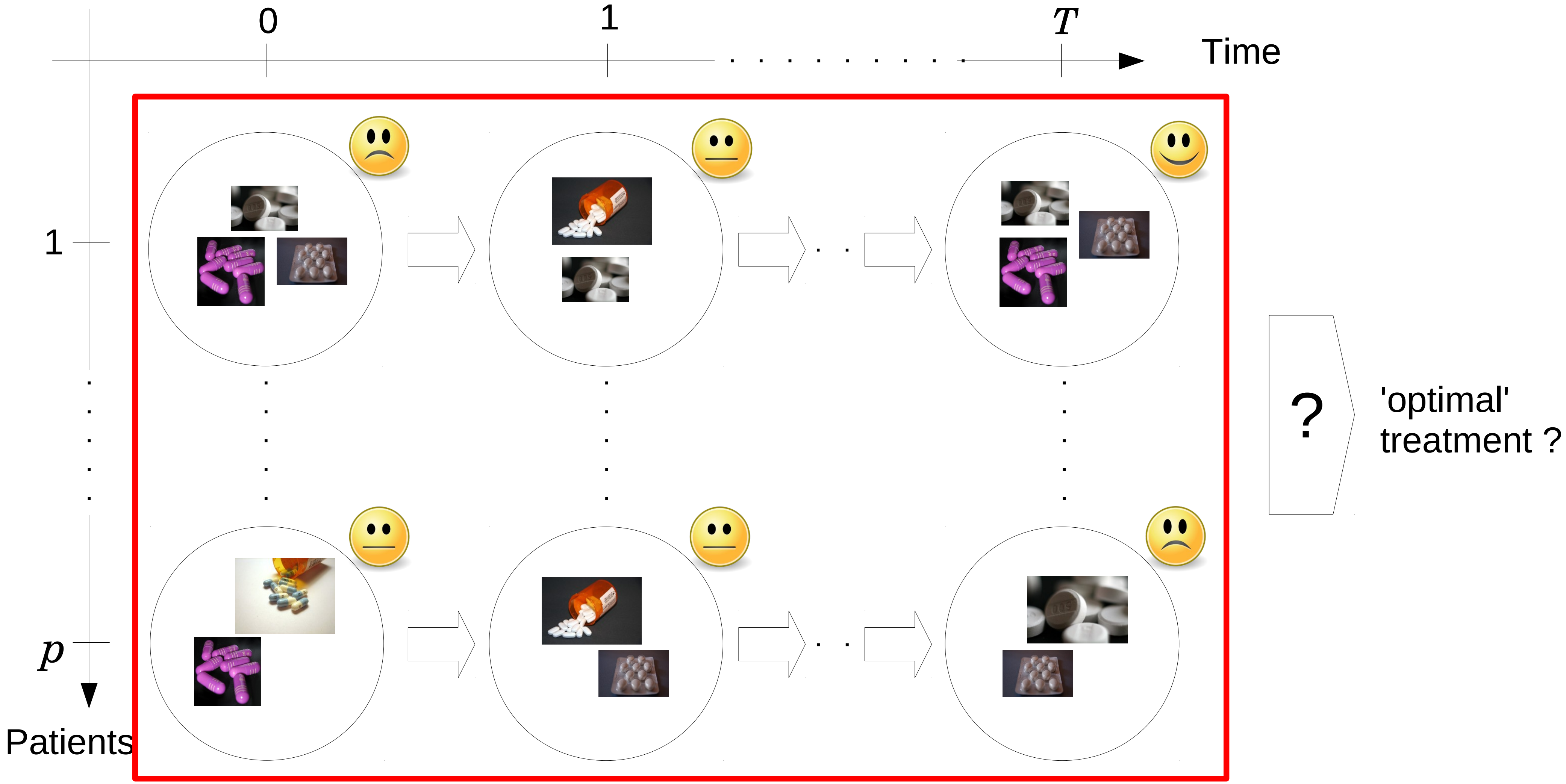
- A batch of (random) trajectories

- Maximal variations of the system, in the form of upper bounds on Lipschitz constants

A motivation: dynamic treatment regimes



A motivation: dynamic treatment regimes



Batch collection of trajectories of patients

Formalization

- Deterministic dynamics: $x_{t+1} = f(x_t, u_t) \quad t = 0, \ldots, T-1$
- Deterministic reward function: $r_t = \rho(x_t, u_t) \in \mathbb{R}$
- Fixed initial state: $x_0 \in \mathcal{X}$
- Continuous sate space, finite action space: $\mathcal{U} = \left\{u^{(1)}, \ldots, u^{(m)}\right\}$

  $\mathcal{X} \subset \mathbb{R}^d$

- Return of a sequence of actions:

$$J(u_0, \ldots, u_{T-1}) \triangleq \sum_{t=0}^{T-1} \rho(x_t, u_t)$$

- Optimal return:

$$J_T^* \triangleq \max_{(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T} J(u_0, \ldots, u_{T-1})$$

The "batch" mode setting

Learning from trajectories

- System dynamics and reward function are **unnkown**
- For every action $u \in \mathcal{U}$, a **set of transitions** is known:

$$\mathcal{F}^{(u)} = \left\{ \left(x^{(u),k}, r^{(u),k}, y^{(u),k} \right) \right\}_{k=1}^{n^{(u)}}$$

$$y^{(u),k} = f\left(x^{(u),k}, u\right) \text{ and } r^{(u),k} = \rho\left(x^{(u),k}, u\right)$$

- Each set of transition is non-empty:

$$\forall u \in \mathcal{U}, \qquad n^{(u)} > 0$$

- Define:

$$\mathcal{F} = \mathcal{F}^{(1)} \cup \ldots \cup \mathcal{F}^{(m)}$$

Lipschitz continuity

Assumption about maximal variations

- We assume that the system dynamics and reward function are Lipschitz continuous:

$$\forall (x, x') \in \mathcal{X}^2, \forall u \in \mathcal{U},$$

$$\begin{aligned} \|f(x,u) - f(x',u)\| &\leq L_f \|x - x'\| \\ |\rho(x,u) - \rho(x',u)| &\leq L_\rho \|x - x'\| \end{aligned}$$

where $\|\cdot\|$ denotes the Euclidean norm over the state space

- We also assume that two constants L_f and L_ρ satisfying the above equations are **known**

Min max generalization

- One can define the sets of Lipschitz continuous functions compatible with the data:

$$\mathcal{L}_{\mathcal{F}}^{f} = \left\{ f' : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{X} \middle| \begin{cases} \forall x', x'' \in \mathcal{X}, \forall u \in \mathcal{U}, \\ \| f'(x', u) - f'(x'', u) \| \leq L_f \| x' - x'' \| , \\ \forall k, f'(x^{(u),k}, u) = f(x^{(u),k}, u) = y^{(u),k} \end{cases} \right\}$$

$$\mathcal{L}_{\mathcal{F}}^{\rho} = \left\{ \rho' : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R} \middle| \begin{cases} \forall x', x'' \in \mathcal{X}, \forall u \in \mathcal{U} , \\ | \rho'(x', u) - \rho'(x'', u) | \leq L_\rho \| x' - x'' \| , \\ \forall k, \rho'(x^{(u),k}, u) = \rho(x^{(u),k}, u) = r^{(u),k} \end{cases} \right\}$$

and the return associated with a couple of fonctions taken in those two ensembles :

$$J_{(f', \rho')}(u_0, \ldots, u_{T-1}) = \sum_{t=0}^{T-1} \rho'(x'_t, u_t)$$

$$x'_{t+1} = f'(x'_t, u_t)$$

Min max generalization

- One can then define:

$$B^*(\mathcal{F}, u_0, \ldots, u_{T-1}) = \min_{(f', \rho') \in \mathcal{L}_{\mathcal{F}}^f \times \mathcal{L}_{\mathcal{F}}^\rho} \left\{ J_{(f', \rho')}(u_0, \ldots, u_{T-1}) \right\}$$

- And the solution of the min max generalization problem can be defined as follows:

$$(u_0, \ldots, u_{T-1}) \in \arg\max_{(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T} B^*(\mathcal{F}, u_0, \ldots, u_{T-1})$$

Reformulation

- According to previous research [1], we know that computing the optimal bound for a given sequence of actions can be reformalized as follows:

$$(\mathcal{P}(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1})):$$

$$\min_{\substack{\hat{\mathbf{r}}_0 \quad \ldots \quad \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_0 \quad \ldots \quad \hat{\mathbf{x}}_{T-1} \in \mathcal{X}}} \sum_{t=0}^{T-1} \hat{\mathbf{r}}_t,$$

subject to

(3.1)
$$\left|\hat{\mathbf{r}}_t - r^{(u_t),k_t}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2 \ \forall (t, k_t) \in \{0, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\},$$

(3.2)
$$\left\|\hat{\mathbf{x}}_{t+1} - y^{(u_t),k_t}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2 \ \forall (t, k_t) \in \{0, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\},$$

(3.3)
$$|\hat{\mathbf{r}}_t - \hat{\mathbf{r}}_{t'}|^2 \leq L_\rho^2 \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'}\|^2 \ \forall t, t' \in \{0, \ldots, T-1 | u_t = u_{t'}\},$$

(3.4)
$$\|\hat{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_{t'+1}\|^2 \leq L_f^2 \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'}\|^2 \ \forall t, t' \in \{0, \ldots, T-2 | u_t = u_{t'}\},$$

(3.5)
$$\hat{\mathbf{x}}_0 = x_0.$$

[1] "**Towards Min Max Generalization in Reinforcement Learning**". R. Fonteneau, S.A. Murphy, L. Wehenkel and D. Ernst. Agents and Artificial Intelligence: International Conference, ICAART 2010, Valencia, Spain, January 2010, Revised Selected Papers. Series: Communications in Computed and Information Science (CCIS), Volume 129, pp. 61-77. Editors: J. Filipe, A. Fred, and B.Sharp. Springer, Heidelberg, 2011.

Reformulation

- According to previous research [1], we know that computing the optimal bound for a given sequence of actions can be reformalized as follows:

$(\mathcal{P}(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1}))$:

$$\min_{\substack{\hat{\mathbf{r}}_0 \ \ldots \ \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_0 \ \ldots \ \hat{\mathbf{x}}_{T-1} \in \mathcal{X}}} \sum_{t=0}^{T-1} \hat{\mathbf{r}}_t,$$

subject to

(3.1)

$$\left|\hat{\mathbf{r}}_t - r^{(u_t),k_t}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2 \ \forall (t, k_t) \in \{0, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\},$$

(3.2)

$$\left\|\hat{\mathbf{x}}_{t+1} - y^{(u_t),k_t}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2 \ \forall (t, k_t) \in \{0, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\},$$

(3.3) $\quad |\hat{\mathbf{r}}_t - \hat{\mathbf{r}}_{t'}|^2 \leq L_\rho^2 \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'}\|^2 \ \forall t, t' \in \{0, \ldots, T-1 | u_t = u_{t'}\},$

(3.4)

$$\|\hat{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_{t'+1}\|^2 \leq L_f^2 \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'}\|^2 \ \forall t, t' \in \{0, \ldots, T-2 | u_t = u_{t'}\},$$

(3.5) $\quad \hat{\mathbf{x}}_0 = x_0.$

[1] proposes a lower bound on the optimal bound (computed independently from this reformulation)

Here, we directly target this problem in order to find a bound tighter than [1]

[1] "**Towards Min Max Generalization in Reinforcement Learning**". R. Fonteneau, S.A. Murphy, L. Wehenkel and D. Ernst. Agents and Artificial Intelligence: International Conference, ICAART 2010, Valencia, Spain, January 2010, Revised Selected Papers. Series: Communications in Computed and Information Science (CCIS), Volume 129, pp. 61-77. Editors: J. Filipe, A. Fred, and B.Sharp. Springer, Heidelberg, 2011.

Small simplification

- One can show that type (3.3) constraints are redundant:

LEMMA 4.1. *Consider* $(\hat{\mathbf{r}}^*, \hat{\mathbf{x}}^*) \in \mathbb{R}^T \times \mathcal{X}^T$ *an optimal solution to* $\bar{\mathcal{P}}(\mathcal{F}, u_0, \ldots, u_{T-1})$. *Then, for all* t, t' *such that* $u_t = u_{t'}$,

$$\left|\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^*\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t^* - \hat{\mathbf{x}}_{t'}^*\right\|^2 .$$

- We can deduce the solution for time t=0 :

LEMMA 4.2. *The solution of the problem* $(\mathcal{P}'(\mathcal{F}, u_0))$ *is*

$$\hat{\mathbf{r}}_0^* = \max_{k_0 \in \left\{1, \ldots, n^{(u_0)}\right\}} r^{(u_0), k_0} - L_\rho \left\| x_0 - x^{(u_0), k_0} \right\| .$$

New problem

$(\mathcal{P}''(\mathcal{F}, u_0, \ldots, u_{T-1})):$

$$\min_{\substack{\hat{\mathbf{r}}_1 \quad \ldots \quad \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_0 \quad \ldots \quad \hat{\mathbf{x}}_{T-1} \in \mathcal{X}}} \sum_{t=1}^{T-1} \hat{\mathbf{r}}_t,$$

subject to

(5.1)

$$\left|\hat{\mathbf{r}}_t - r^{(u_t),k_t}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2 \ \forall (t, k_t) \in \{1, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\},$$

(5.2)

$$\left\|\hat{\mathbf{x}}_{t+1} - y^{(u_t),k_t}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2 \ \forall (t, k_t) \in \{0, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\},$$

(5.3)

$$\|\hat{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_{t'+1}\|^2 \leq L_f^2 \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'}\|^2 \ \forall t, t' \in \{0, \ldots, T-2 | u_t = u_{t'}\},$$

(5.4) $\quad \hat{\mathbf{x}}_0 = x_0 \ .$

Complexity

- One can show that such a problem is **NP-hard**
- We propose relaxation schemes of **polynomial complexity**
- We want those relaxation schemes to preserve the philosophy of the original problem, i.e., to provide lower bounds
- We propose two types of relaxations:
  - The Intertwined Trust-Region (ITR) relaxation scheme
  - The Lagrangian relaxation scheme
- We show that those relaxations are more efficient than previous solution given in [1]

Relaxation schemes

(I) Intertwined trust-region

- First approach: remove constraints until the problem becomes polynomial

$(\mathcal{P}''(\mathcal{F}, u_0, \ldots, u_{T-1}))$:

$$\min_{\substack{\hat{\mathbf{r}}_1 \quad \ldots \quad \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_0 \quad \ldots \quad \hat{\mathbf{x}}_{T-1} \in \mathcal{X}}} \quad \sum_{t=1}^{T-1} \hat{\mathbf{r}}_t,$$

subject to

Only one constraint

(5.1)

$$\left|\hat{\mathbf{r}}_t - r^{(u_t),k_t}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2 \ \forall (t, k_t) \in \{1, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\},$$

(5.2)

$$\left\|\hat{\mathbf{x}}_{t+1} - y^{(u_t),k_t}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2 \ \forall (t, k_t) \in \{0, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\},$$

(5.3)

$$\|\hat{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_{t'+1}\|^2 \leq L_f^2 \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'}\|^2 \ \forall t, t' \in \{0, \ldots, T-2 | u_t = u_{t'}\},$$

(5.4) $$\hat{\mathbf{x}}_0 = x_0 \ .$$

(I) Intertwined trust-region

- We get the ITR problem:

$$\left(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1})\right):$$

$$\min_{\substack{\hat{\mathbf{r}}_1, \ldots, \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_0, \ldots, \hat{\mathbf{x}}_{T-1} \in \mathcal{X}}} \quad \sum_{t=1}^{T-1} \hat{\mathbf{r}}_t$$

subject to

$$(5.5) \qquad \left| \hat{\mathbf{r}}_t - r^{(u_t), \bar{k}_t} \right|^2 \leq L_\rho^2 \left\| \hat{\mathbf{x}}_t - x^{(u_t), \bar{k}_t} \right\|^2, \qquad t \in \{1, \ldots, T-1\},$$

$$(5.6) \qquad \left\| \hat{\mathbf{x}}_t - y^{(u_{t-1}), \bar{k}_{t-1}} \right\|^2 \leq L_f^2 \left\| \hat{\mathbf{x}}_{t-1} - x^{(u_{t-1}), \bar{k}_{t-1}} \right\|^2, \qquad t \in \{1, \ldots, T-1\},$$

$$(5.7) \qquad \hat{\mathbf{x}}_0 = x_0.$$

- A closed-form solution of this problem can be obtained

(I) Intertwined trust-region

THEOREM 5.4. *The solution to* $(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}))$ *is given by*

$$B''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}) = \sum_{t=1}^{T-1} \hat{\mathbf{r}}_t^*,$$

where

$$\begin{aligned}
\hat{\mathbf{r}}_t^* &= r^{(u_t),\bar{k}_t} - L_\rho \left\| \hat{\mathbf{x}}_t^*(\bar{k}_0, \ldots, \bar{k}_t) - x^{(u_t),\bar{k}_t} \right\|, \\
\hat{\mathbf{x}}_t^*(\bar{k}_0, \ldots, \bar{k}_t) &\doteq y^{(u_{t-1}),\bar{k}_{t-1}} \\
&\quad + L_f \frac{\left\| \hat{\mathbf{x}}_{t-1}^*(\bar{k}_0, \ldots, \bar{k}_{t-1}) - x^{(u_{t-1}),\bar{k}_{t-1}} \right\|}{\left\| y^{(u_{t-1}),\bar{k}_{t-1}} - x^{(u_t),\bar{k}_t} \right\|} \left(y^{(u_{t-1}),\bar{k}_{t-1}} - x^{(u_t),\bar{k}_t} \right)
\end{aligned}$$

if $y^{(u_{t-1}),\bar{k}_{t-1}} \neq x^{(u_t),\bar{k}_t}$

and, if $y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t}$, $\hat{\mathbf{x}}_t^*(\bar{k}_0, \ldots, \bar{k}_t)$ *can be any point of the sphere centered in* $y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t}$ *with radius* $L_f \|\hat{\mathbf{x}}_{t-1}^*(\bar{k}_0, \ldots, \bar{k}_{t-1}) - x^{(u_{t-1}),\bar{k}_{t-1}}\|$.

(I) Intertwined trust-region

- The ITR problem can be solved for any selection of constraints
- One can thus define a maximal ITR bound :

DEFINITION 5.5 (ITR bound $B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1})$).

$$B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}) \triangleq \hat{\mathbf{r}}_0^* + \max_{\substack{\bar{k}_{T-1} \in \{1, \ldots, n^{(u_{T-1})}\} \\ \ldots \\ \bar{k}_0 \in \{1, \ldots, n^{(u_0)}\}}} B''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1})$$

(II) Lagrangian relaxation

$(\mathcal{P}''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}))$:

$$\max_{\substack{\nu_{t,t'} \in \mathbb{R} \\ \lambda_{t,k_t} \in \mathbb{R} \\ \mu_{t,k_t} \in \mathbb{R}}} \quad \min_{\substack{\hat{\mathbf{r}}_1, \ldots, \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{T-1} \in \mathcal{X}}}$$

$$\hat{\mathbf{r}}_1 + \cdots + \hat{\mathbf{r}}_{T-1}$$

$$+ \sum_{(t,k_t) \in \{1,\ldots,T-1\} \times \{1,\ldots,n^{(u_t)}\}} \mu_{t,k_t} \left(\left| \hat{\mathbf{r}}_t - r^{(u_t),k_t} \right|^2 - L_\rho^2 \left\| \hat{\mathbf{x}}_t - x^{(u_t),k_t} \right\|^2 \right)$$

$$+ \sum_{(t,k_t) \in \{1,\ldots,T-1\} \times \{1,\ldots,n^{(u_t)}\}} \lambda_{t,k_t} \left(\left\| \hat{\mathbf{x}}_{t+1} - y^{(u_t),k_t} \right\|^2 - L_f^2 \left\| \hat{\mathbf{x}}_t - x^{(u_t),k_t} \right\|^2 \right)$$

$$+ \sum_{t,t' \in \{0,\ldots,T-2 | u_t = u_{t'}\}} \nu_{t,t'} \left(\left\| \hat{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_{t'+1} \right\|^2 - L_f^2 \left\| \hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'} \right\|^2 \right).$$

- Polynomial complexity

Tightness of the bounds

- Comparison with the relaxation proposed in [1] :

DEFINITION 5.8 (CGRL bound $B_{CGRL}(\mathcal{F}, u_0, \ldots, u_{T-1})$).

$$
\begin{aligned}
&B_{CGRL}(\mathcal{F}, u_0, \ldots, u_{T-1}) \\
&\triangleq \max_{\substack{\bar{k}_{T-1} \in \{1, \ldots, n^{(u_{T-1})}\} \\ \ldots \\ \bar{k}_0 \in \{1, \ldots, n^{(u_0)}\}}} r^{(u_0),\bar{k}_0} - L_\rho \left(1 + L_f + L_f^2 + \cdots + L_f^{T-2}\right) \left\| x^{(u_0),\bar{k}_0} - x_0 \right\| \\
&\qquad + \cdots + \\
&\qquad + r^{(u_{T-2}),\bar{k}_{T-2}} - L_\rho \left(1 + L_f\right) \left\| y^{(u_{T-3}),\bar{k}_{T-3}} - x^{(u_{T-2}),\bar{k}_{T-2}} \right\| \\
&\qquad + r^{(u_{T-1}),\bar{k}_{T-1}} - L_\rho \left\| y^{(u_{T-2}),\bar{k}_{T-2}} - x^{(u_{T-1}),\bar{k}_{T-1}} \right\| .
\end{aligned}
$$

Tightness of the bounds

- ITR versus [1] :

THEOREM 5.9.

$$B_{CGRL}(\mathcal{F}, u_0, \ldots, u_{T-1}) \leq B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1})$$

Sketch of proof :

- Compute the ITR relaxation with the constraints used by the CGRL bound

Tightness of the bounds

- Lagrangian relaxation versus ITR :

THEOREM 5.17.

$$B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}) \leq B_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1})$$

Sketch of proof:

  - Strong duality holds for the Lagrangian relaxation of the ITR problem

Tightness of the bounds

- Synthesis:

$$\begin{aligned} B_{CGRL}(\mathcal{F}, u_0, \ldots, u_{T-1}) &\leq B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}) \\ &\leq B_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}) \\ &\leq B^*(\mathcal{F}, u_0, \ldots, u_{T-1}) \\ &\leq J(u_0, \ldots, u_{T-1}) \ . \end{aligned}$$

- All these bounds converge to the actual return of sequences of actions when the dispersion decreases towards zero

Illustration

- Dynamics: $\forall (x,u) \in \mathcal{X} \times \mathcal{U}, \qquad f(x,u) = x + 3.1416 \times u \times 1_d$
- Reward function: $\forall (x,u) \in \mathcal{X} \times \mathcal{U}, \qquad \rho(x,u) = \sum_{i=1}^{d} x(i)$
- Initial state: $x_0 = 0.5772 \times 1_d$
- Decision space: $\mathcal{U} = \{0, 0.1\}$
- Grid :

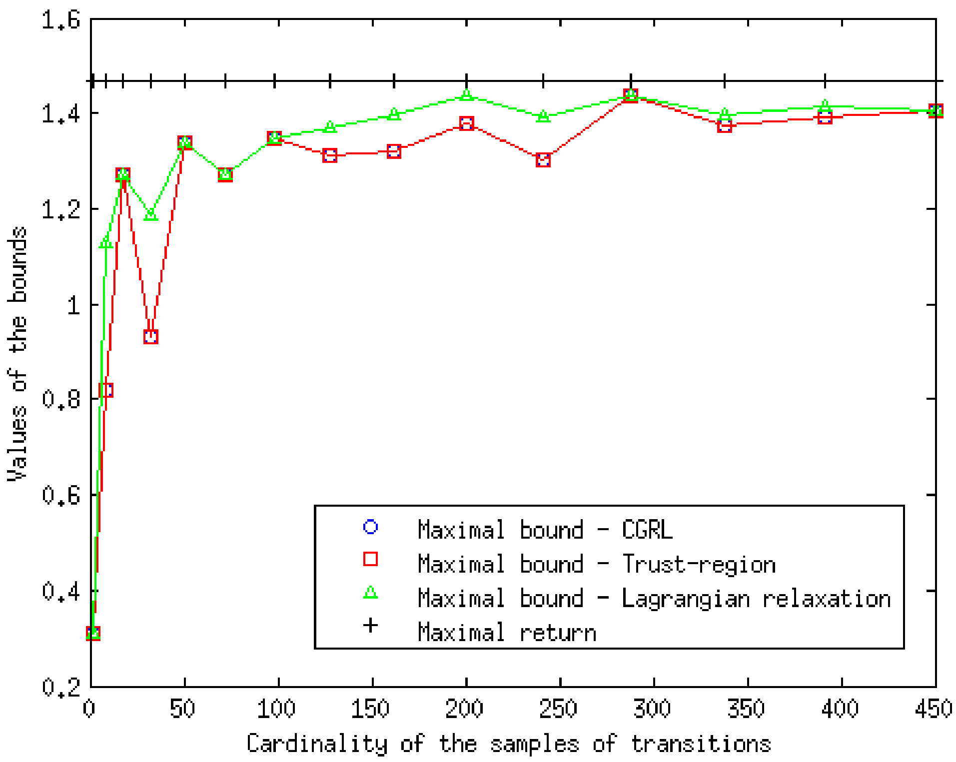
$$\forall u \in \mathcal{U}, \mathcal{F}_{c_i}^{(u)} = \left\{ \left(\left[\frac{i_1}{i}; \frac{i_2}{i} \right], u, \rho\left(\left[\frac{i_1}{i}; \frac{i_2}{i} \right], u \right), f\left(\left[\frac{i_1}{i}; \frac{i_2}{i} \right], u \right) \right) \middle| (i_1, i_2) \in \{1, \ldots, i\}^2 \right\}$$

- 100 samples of transitions drawn uniformly at random

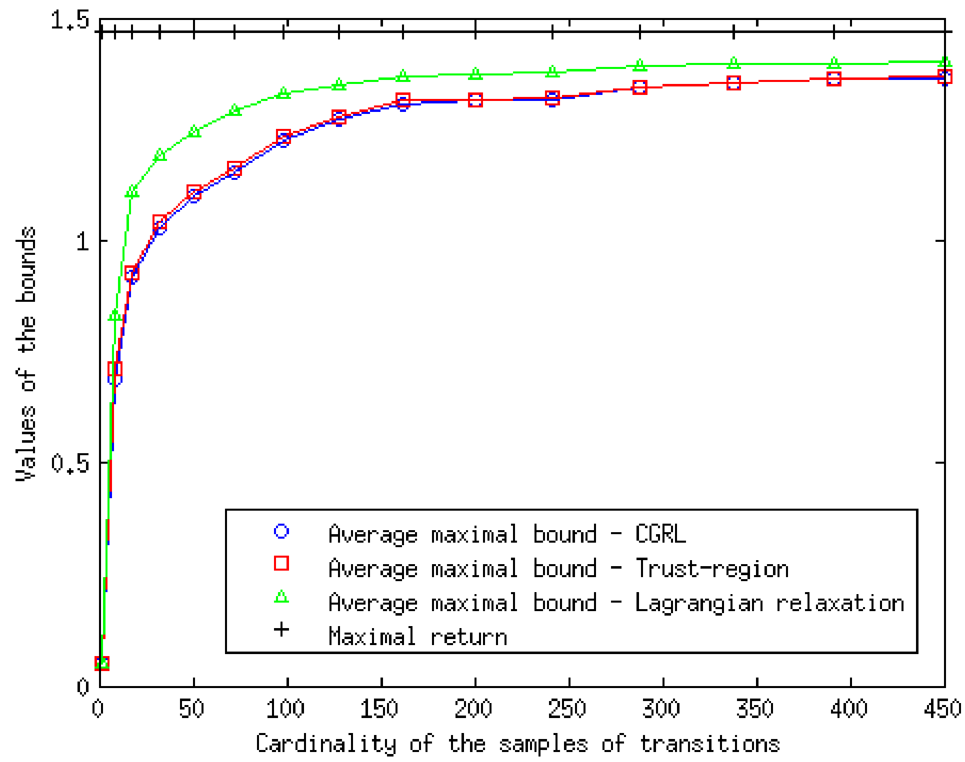
Illustration

Maximal bounds

Grid



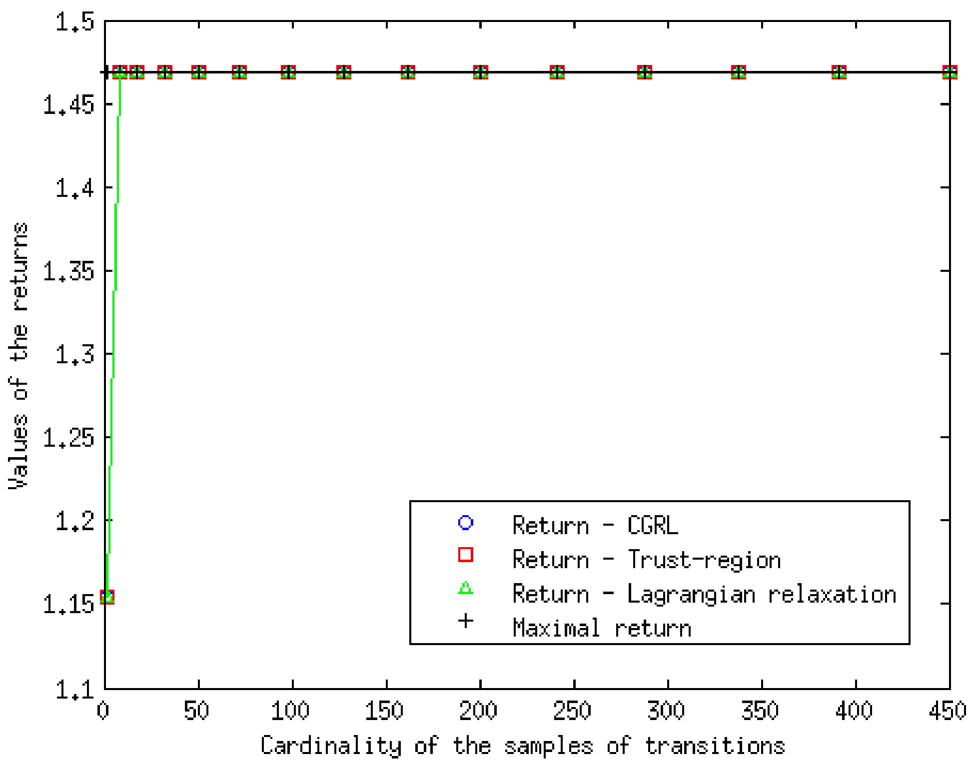
Empirical average over random samples



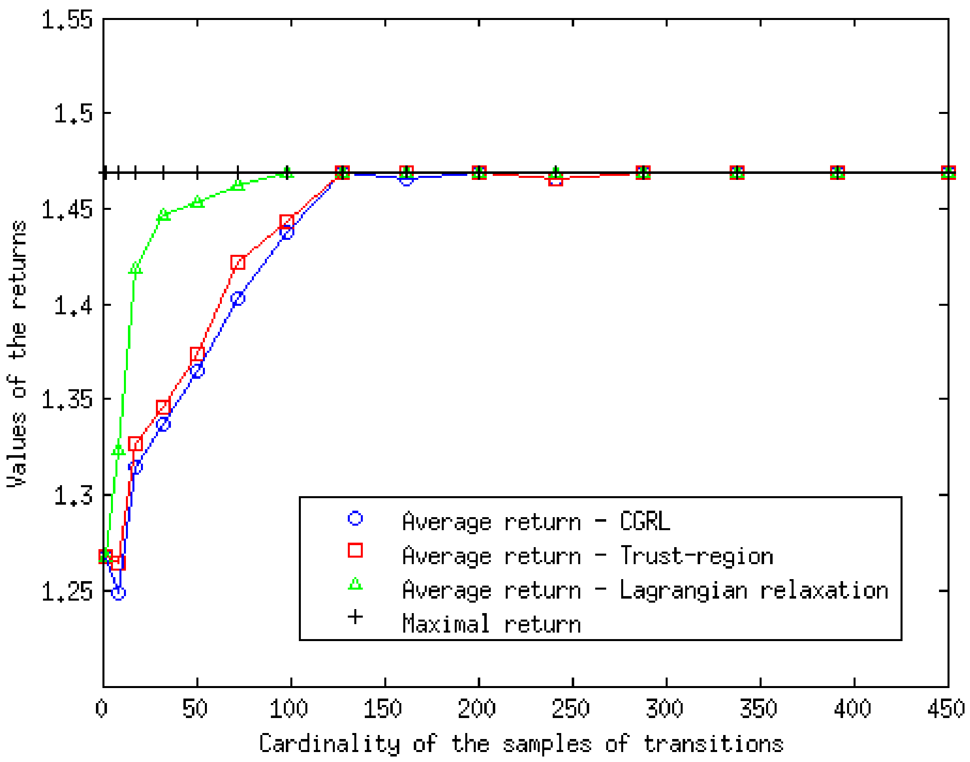
Illustration

Returns of sequences

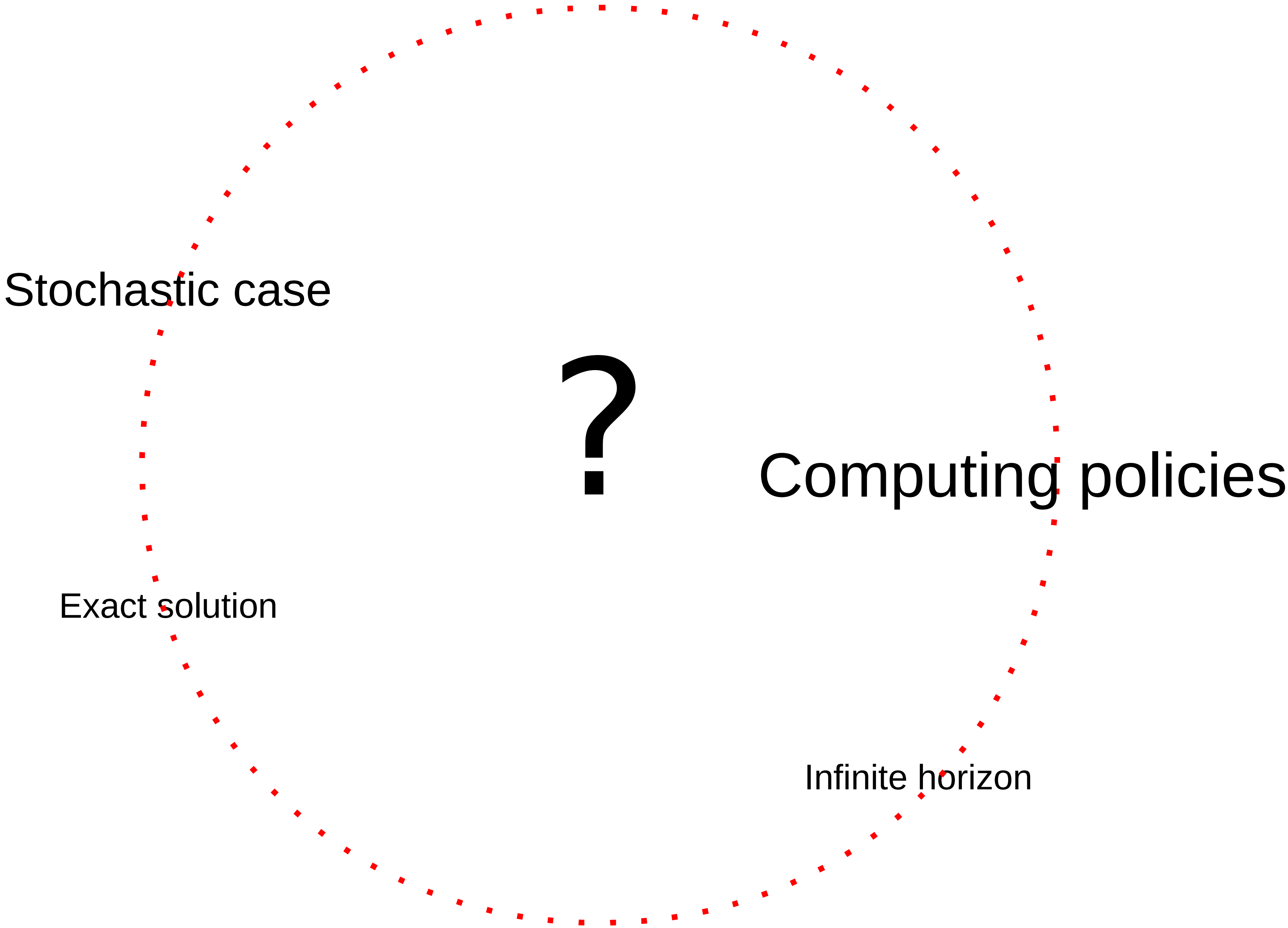
Grid



Empirical average over random samples



Future work



Min Max Generalization for Deterministic Batch Mode Reinforcement Learning: Relaxation Schemes. R. Fonteneau, D. Ernst, B. Boigelot, Q. Louveaux. SIAM Journal on Control and Optimization, Volume 51, Issue 5, pp 3355-3385, 2013.

! Advertisement !
The French Workshop on Planning, Decision Making and Learning 2014
will be in Liège, May 12-13, 2014
Hope to see you there!
http://sites.google.com/site/jfpda14/