Policy Transfer using Value Function as Prior Information

Samy Aittahar, Aivar Sootla and Damien Ernst
University of Liège, Department of Electrical Engineering and Computer Science

Motivations

- Expensive data generation in biological systems;
- Minimization of both objective function and required number of data;
- Data-based distance between models' dynamics.

Dynamics

Main variables

- $S, U \subseteq \mathbb{R}_{\geq 0}$, are respectively state and action spaces;
- $s^i \in S$ is the concentration of gene i;
- $b \in \mathbb{R}_{\geq 0}$, is the power of the light pulse;
- $u \in U$ is the light activation action;
- $p \in \mathbb{R}_{\geq 0}$, is the cost related to the light pulse.

Protein concentration transition for each gene (Simplified version)

$$\begin{aligned} \dot{s}^1 &= \frac{c_1}{1+(s^2/r_1)^{\alpha_1}} - c_2 s^1 + bu \,, \\ \dot{s}^2 &= \frac{c_3}{1+(s^1/r_2)^{\alpha_2}} - c_4 s^2 \,, \end{aligned} \tag{2}$$

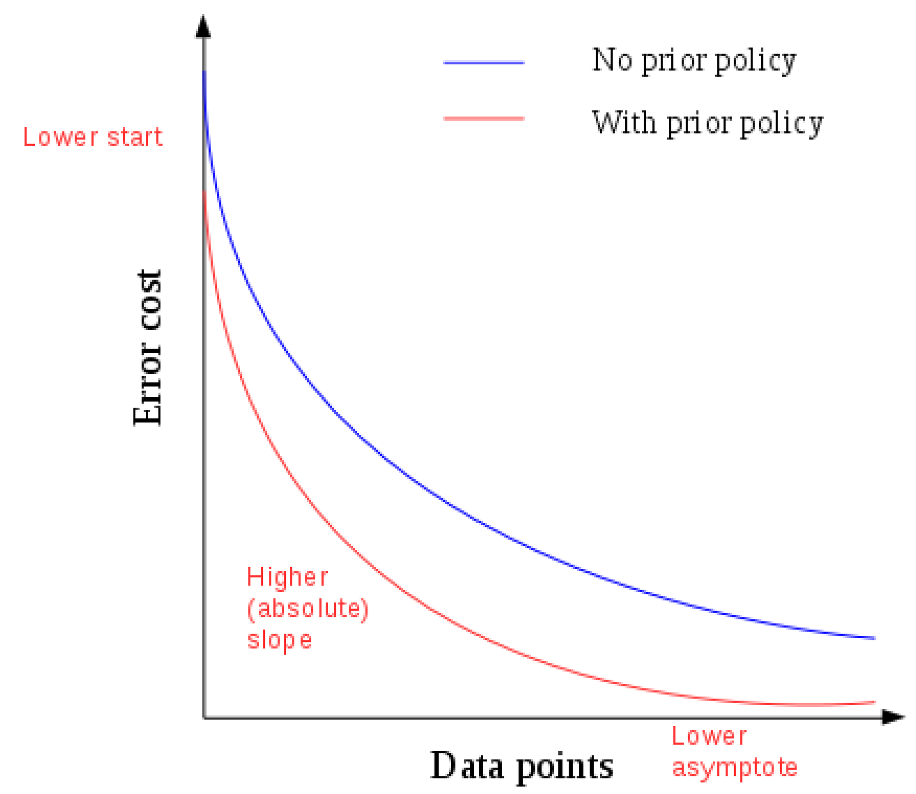
Transition cost

$$c(\langle s^1, s^2 \rangle, u, \langle s^1_+, s^2_+ \rangle) = -s^1_+ + s^2_+ + pu \,, \tag{3}$$

Problem setting

- Known generation model but parameters are hard to identify;
- Small amount of available data;
- Set of near-optimal policies available from similar generation models;
- Deterministic protein concentration transitions.

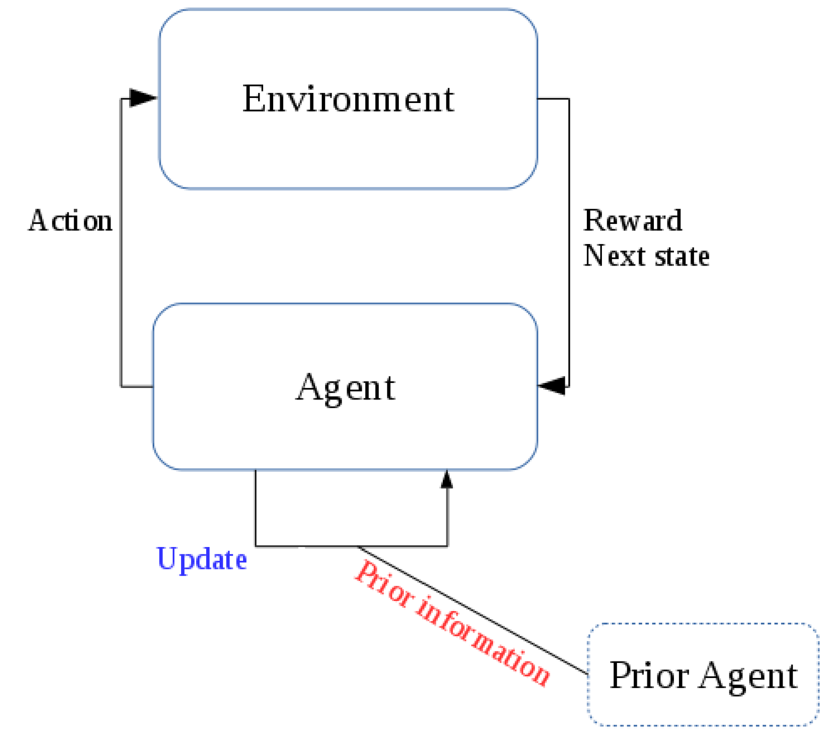
Learning with prior policy



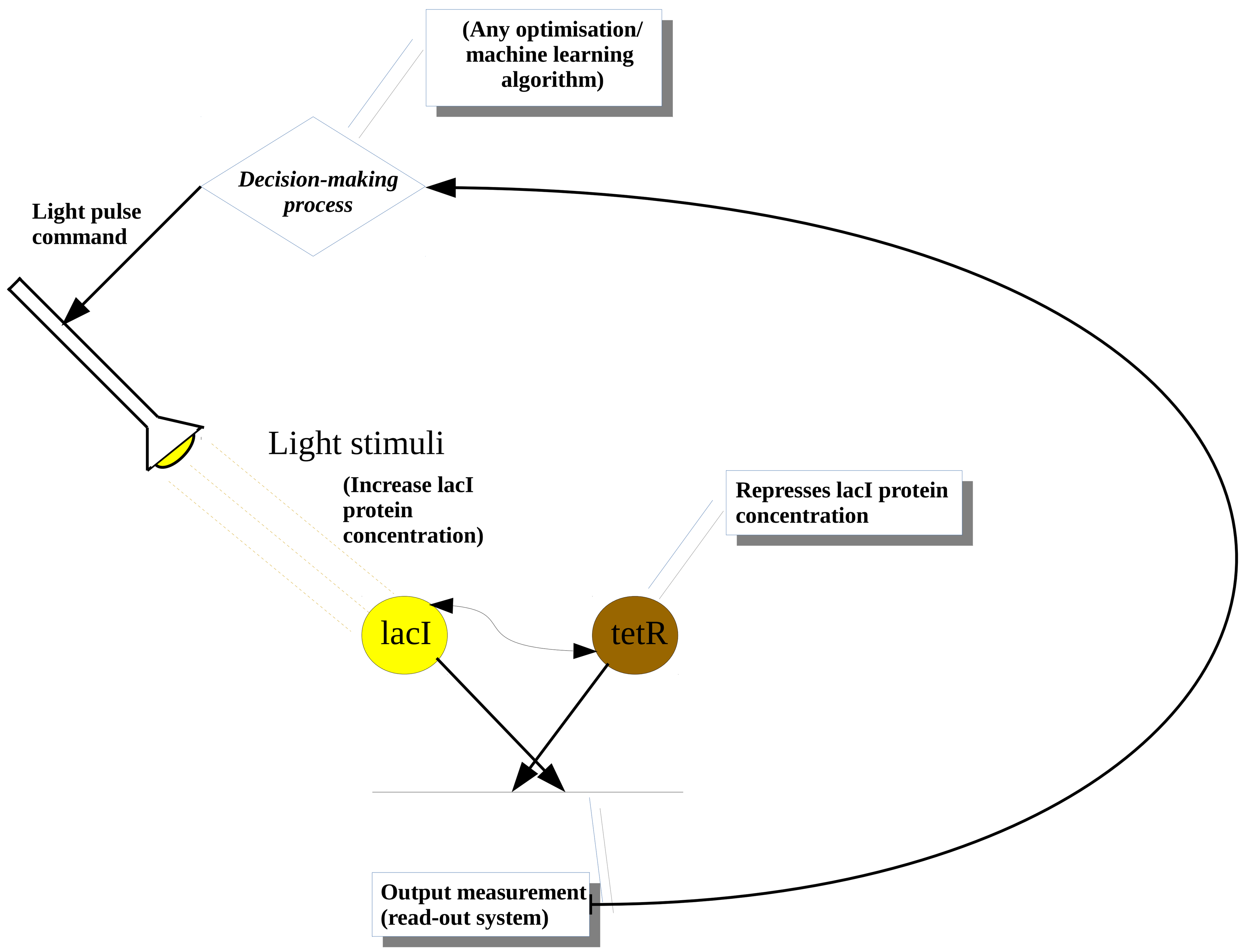
Speed up learning process with less data

Reinforcement learning

Knowledge extraction by exploring environment (possibly with prior information)



Genetic Toggle Switch System



Challenge : Find minimum number light pulses (rationale : avoid bad protein concentration) necessary to grow protein concentration of lacI and overcome the repression gene.

Optimal Value/Q function

Infinite-dimensional optimization problem and Q-function reformulation

$$\begin{aligned} &V(s_t) = \min_{\mu^*(\cdot)} \sum_{i=t}^{+\infty} \gamma^{i-t} c(s_i, \mu^*(s_i), f(s_i, \mu^*(s_i))) \,, \\ &Q_k(s,u) = c(s,u,f(s,u)) + \gamma V_p(f(s,u)) - V_p(s) + \\ &\gamma \min_{v \in U} Q_{k-1}(f(s,v), v) \,. \end{aligned} \tag{1}$$

- $V(s_t)$ is the *value function* (V_p is the existing value function, equals to zero everywhere when no value function is to be used);
- γ is the discount factor;
- $\mu^*(\cdot) : S \to U$ is the optimal state-action mapping (also called an *optimal policy*);
- $f(\cdot)$ is the transition function;
- $Q_k(\cdot)$ is recursive version of the Q-function (computed with *Fitted-Q-Iteration* algorithm in our case) with k the number of processed iterations;
- When V_p is the prior information used in the form of a value function (*reward shaping* technique)

Preliminary conclusions

- Learning process can be speeded up ten times in terms of number of data points;
- Number of data seems not to be correlated to *regret* score (i.e., difference of cost between two policies on the same model).

Future work

- Change problem setting for stochastic dynamics;
- Data selection to avoid negative transfer;
- Aggregation of several prior policies (more specifically, value functions).