Des réseaux de gènes aux lois de mélange et inversement

Elias Ventre

supervisé par Thibault Espinasse, Thomas Lepoutre, Olivier Gandrillon

Context

- We consider a cell in a given **environment**
- Its evolution in the gene expression space depends on its **GRN**
- Due to the **stochastic** nature of the underlying chemical reactions, we observe variations between different cells

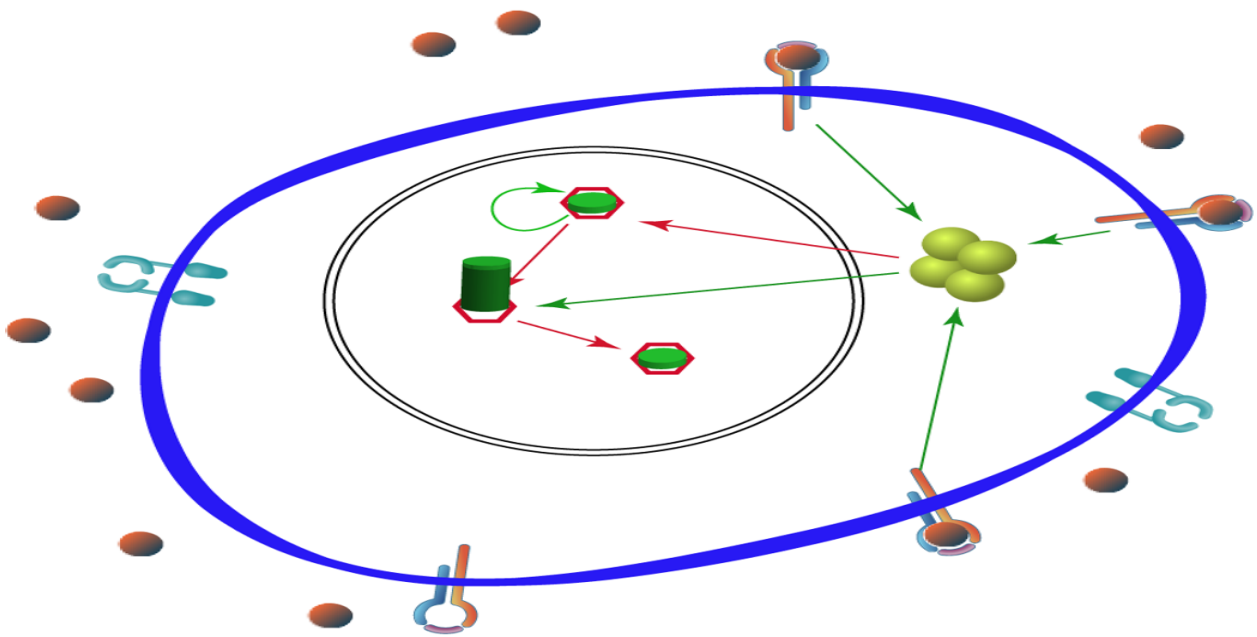


Table of Contents

Stochastic Two States Model

- There is a simple existing stochastic model for the expression of a gene in a single cell :

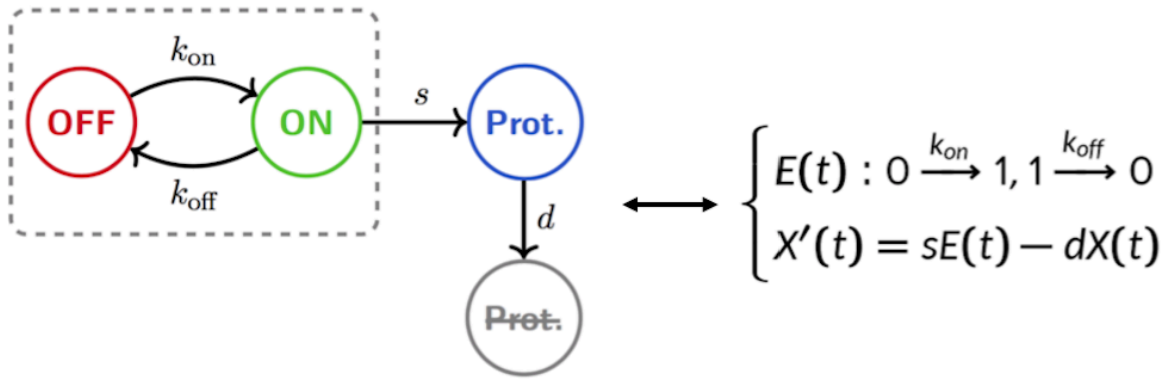
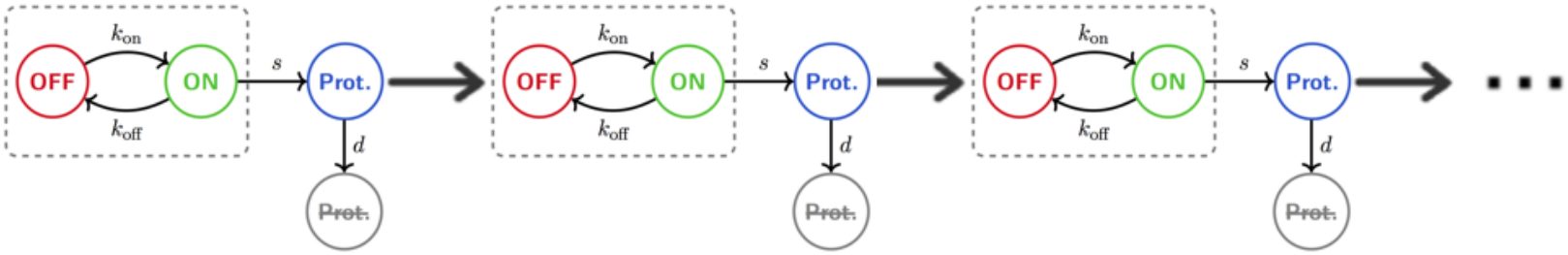


Figure: Two states model. Figure from U. Herbach

- If k_{on} and k_{off} are both constant, and $s = d$, the stationary distribution of such model is a **Beta distribution** of parameters $(\frac{k_{on}}{d}, \frac{k_{off}}{d})$.

Stochastic Two States Model

- We put this model into a network :



- $X = (X_1, \ldots, X_n)$ is now a vector in the **gene expression space**

- k_{on} and k_{off} now depend on **the global protein level**

$$\Longrightarrow k_{on,i}(X) = f_i(X_1, \cdots, X_n)$$

Stochastic two states model

$$\begin{cases} E_i(t) : 0 \xrightarrow{k_{on,i}(X)} 1, 1 \xrightarrow{k_{off,i}} 0 \\ X_i'(t) = d_i(E_i(t) - X_i(t)) \end{cases}$$

- We denote : $\Theta \in M(\mathbb{R}^n)$ a $n \times n$ matrix characterizing the GRN

- The effect of the GRN manifests itself through the function

$k_{on} = k_{on,\Theta}$. Each Θ will generate different cellular behaviours

Deterministic approximation

- We consider that promoters switches are frequent in regard to protein dynamics, and introduce a scaling factor ε :

$$(k_{on}, k_{off}) \longleftrightarrow (\frac{\tilde{k}_{on}}{\varepsilon}, \frac{\tilde{k}_{off}}{\varepsilon})$$

- **scaling factor $\sim$ noise coefficient**

If $\varepsilon \ll 1$, we can derive a deterministic limit :

$$\dot{X}(t) = d\,(E(t) - X(t)) \sim \dot{X}(t) = d\left(\frac{k_{on}}{k_{off} + k_{on}}(X(t)) - X(t)\right)$$

$$\Longrightarrow \dot{X}(t) = F(X(t))$$

Deterministic approximation

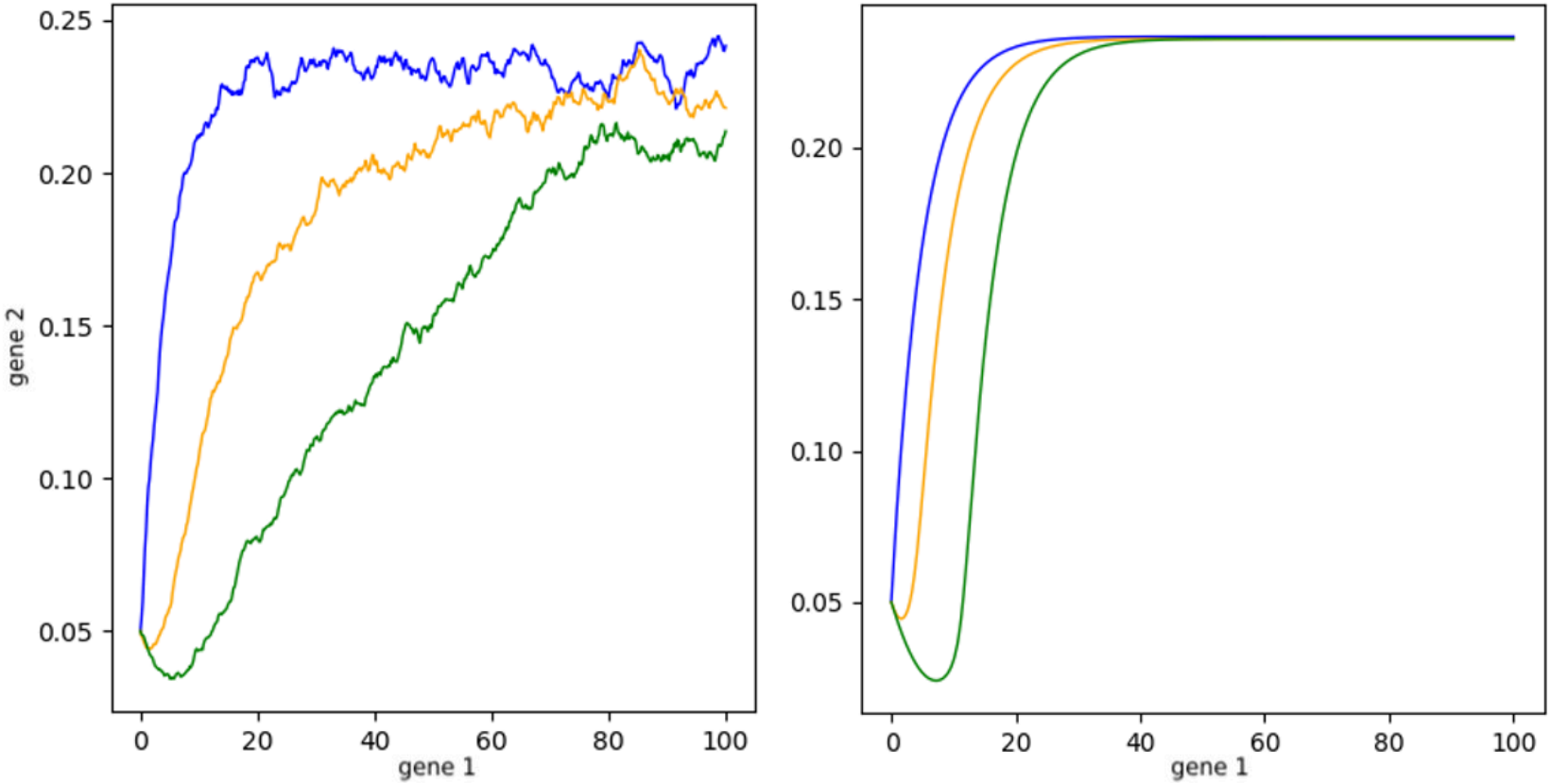


Figure: Comparison between the mean trajectories from the PDMP and the trajectories generated by the deterministic system for a signaling pathway network : 1 ⟶ 2 ⟶ 3

Phase portrait for the toggle-switch

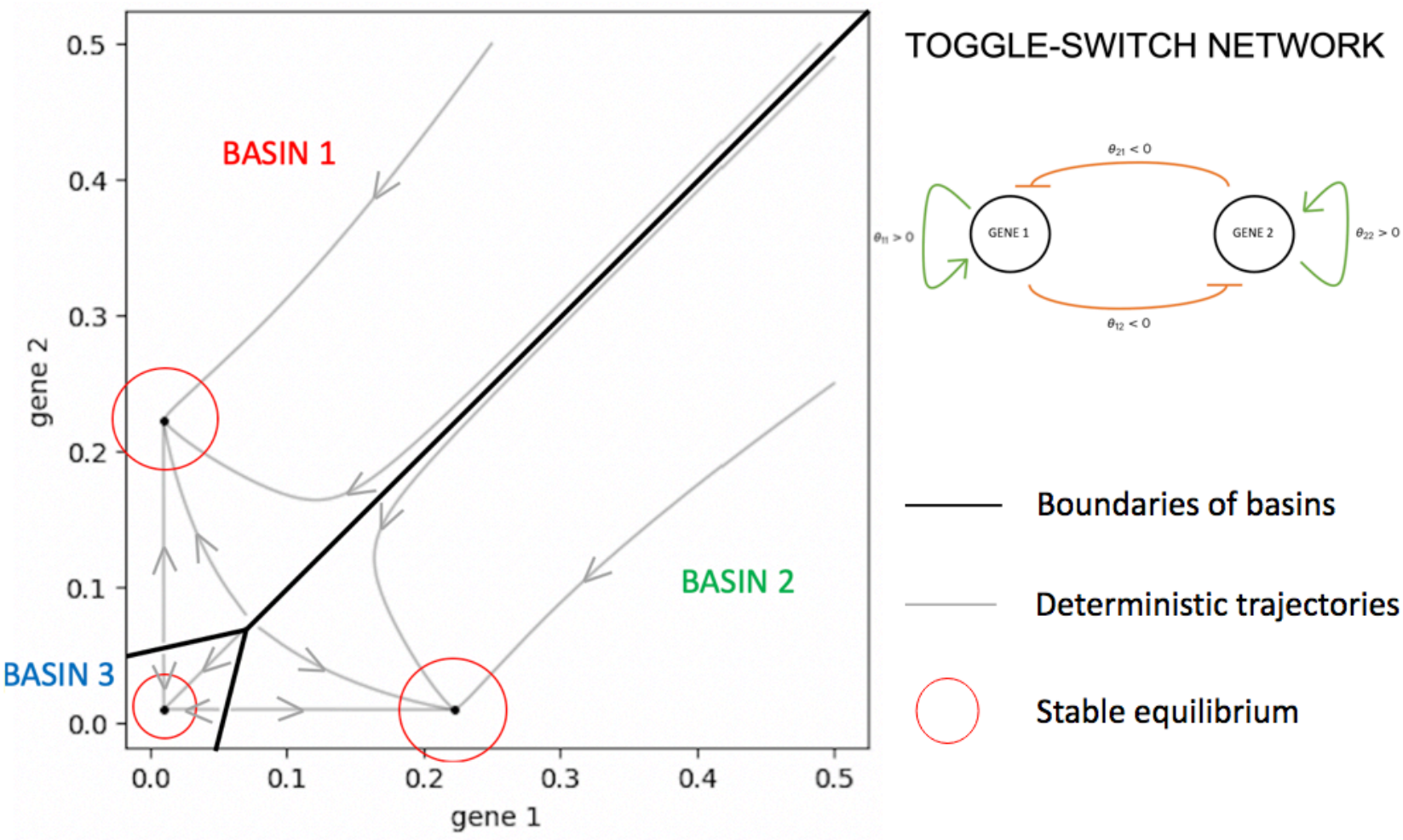


Figure: Phase portrait of the deterministic approximation for a symmetric toggle switch with strong inhibition

Stochastic trajectory

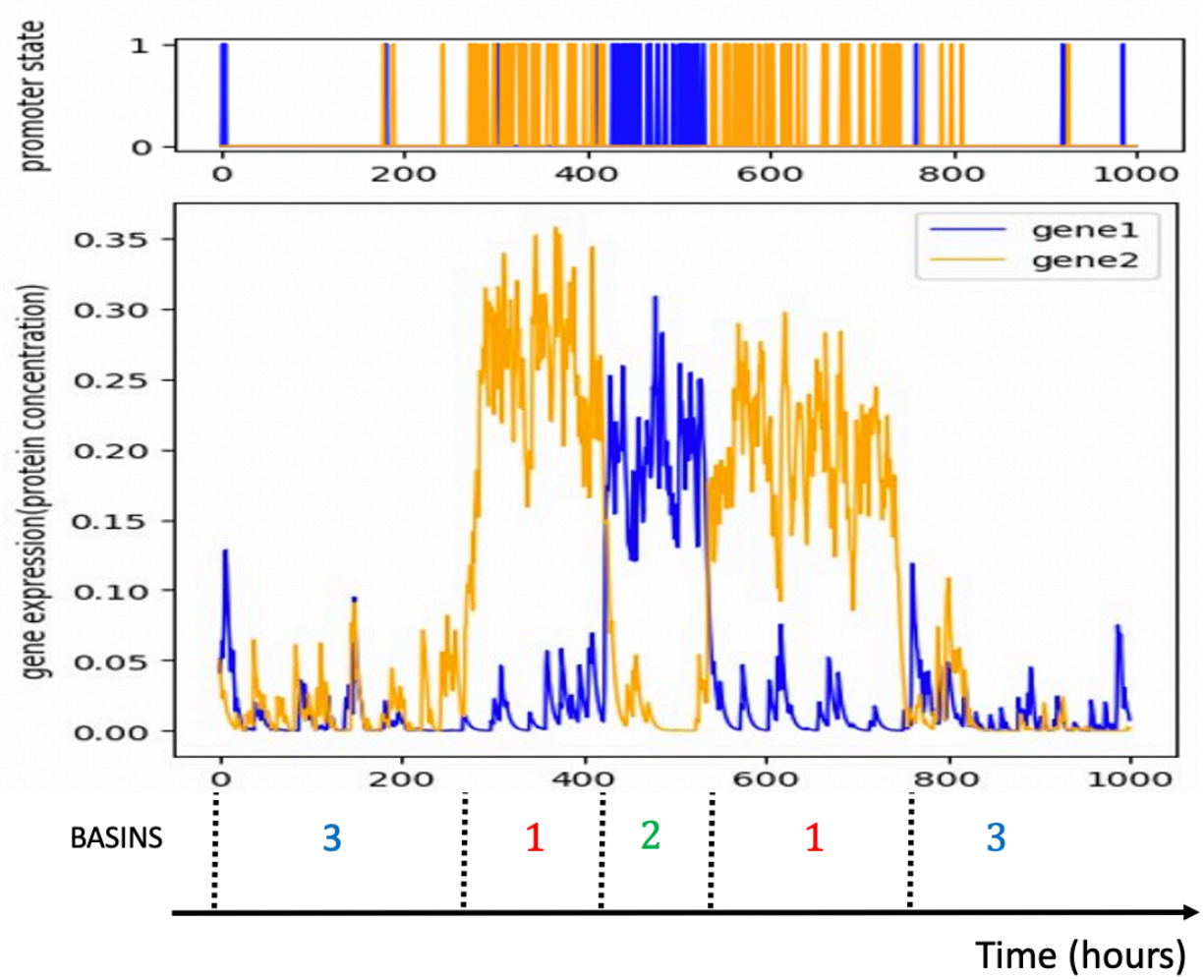


Figure: Example of a stochastic trajectory generated by the toggle switch

Discrete representation

- **Metastability** ~ Cellular type ⟷ basins of attraction

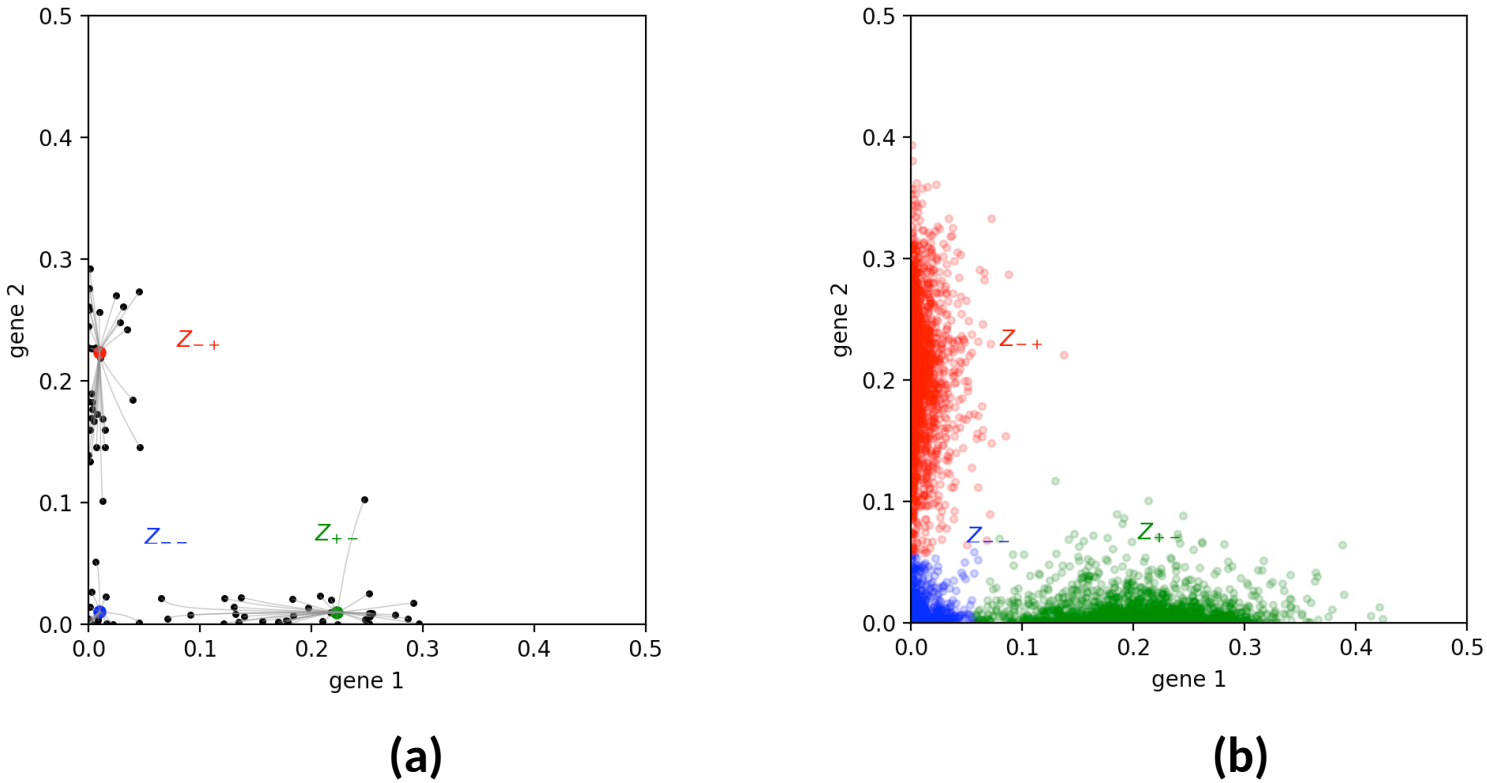


Figure: A cell in the gene expression space can always be associated to one attractive basin (a). Simulating many cells, we can get the proportion of each basin in the process (b)

Transition between basin

- When $\varepsilon \ll 1$, the process spends in a basin a time long enough to **equilibrate inside**:

$\Longrightarrow$ the hitting time of a new basin can be considered as a law without memory

- We build a new **Markovian discrete process**, continuous in time, on the basins

$\Longrightarrow$ the transition probability between two basins Z_i and Z_j can be approximated by an **exponential law**

Exponential fitting

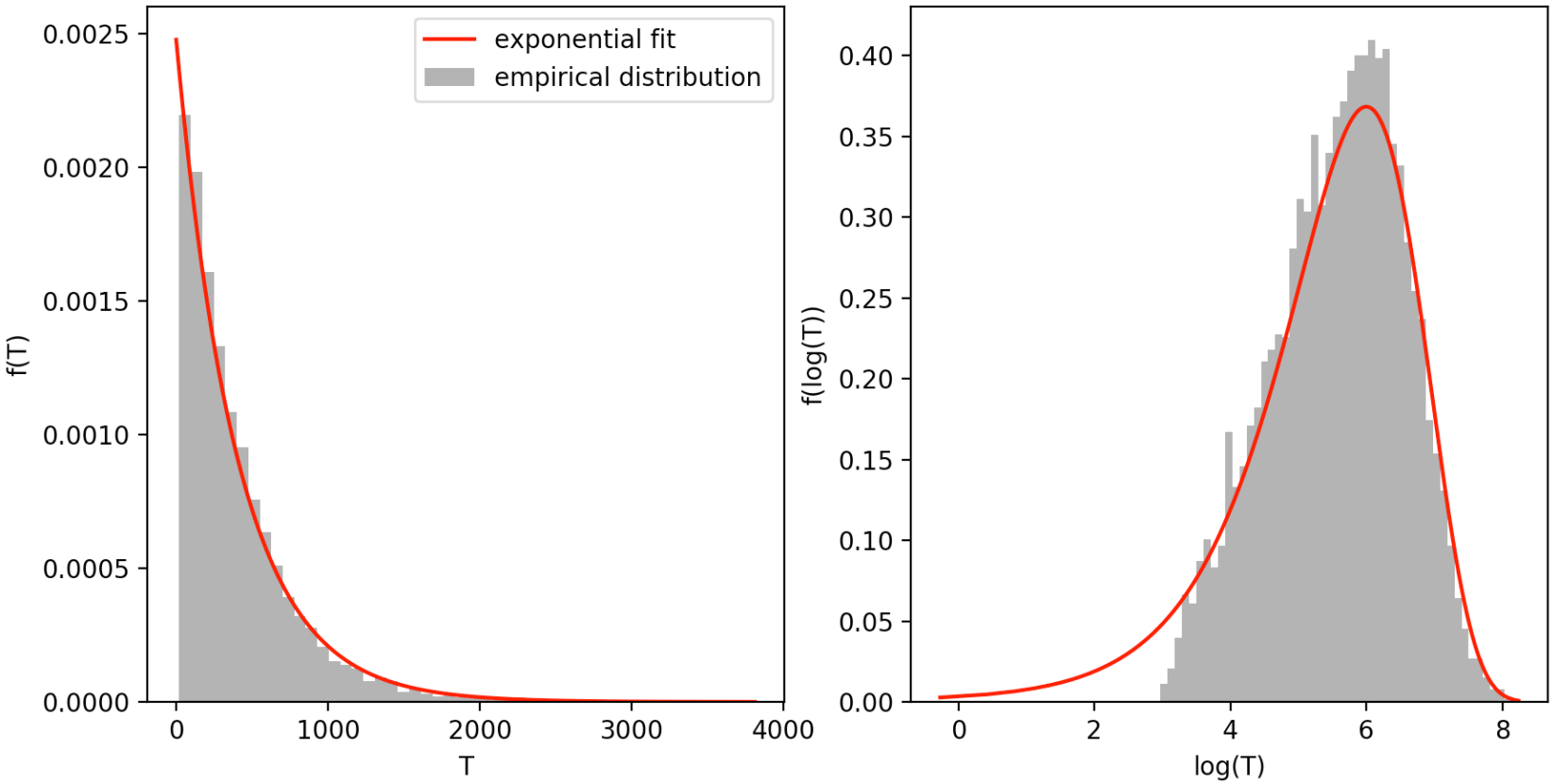


Figure: Empirical distribution of the time passage between two basins in normal and log scale

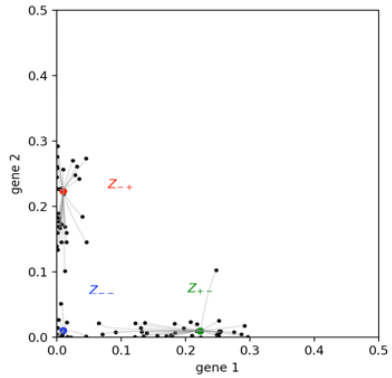
Comparison between stationary distributions

PDMP model

$$\begin{cases} E_i(t) : 0 \xrightarrow{k_{on,i}(X)} 1, 1 \xrightarrow{k_{off,i}} 0 \\ X_i'(t) = d_i(E_i(t) - X_i(t)) \end{cases}$$

Simulation →

Repartition of cells

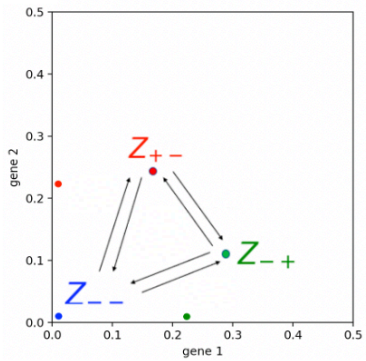


Ratio →

Stationary distribution

μ_b

Coarse grained model



MFPT →

Matricial form

$$\begin{pmatrix} 0 & a_{Z_{--}Z+-} & a_{Z_{--}Z-+} \\ a_{Z_{+-}Z--} & 0 & a_{Z_{+-}Z-+} \\ a_{Z_{-+}Z--} & a_{Z_{-+}Z+-} & 0 \end{pmatrix}$$

Invariant measure →

Stationary distribution

μ_Z

Comparison between stationary distributions

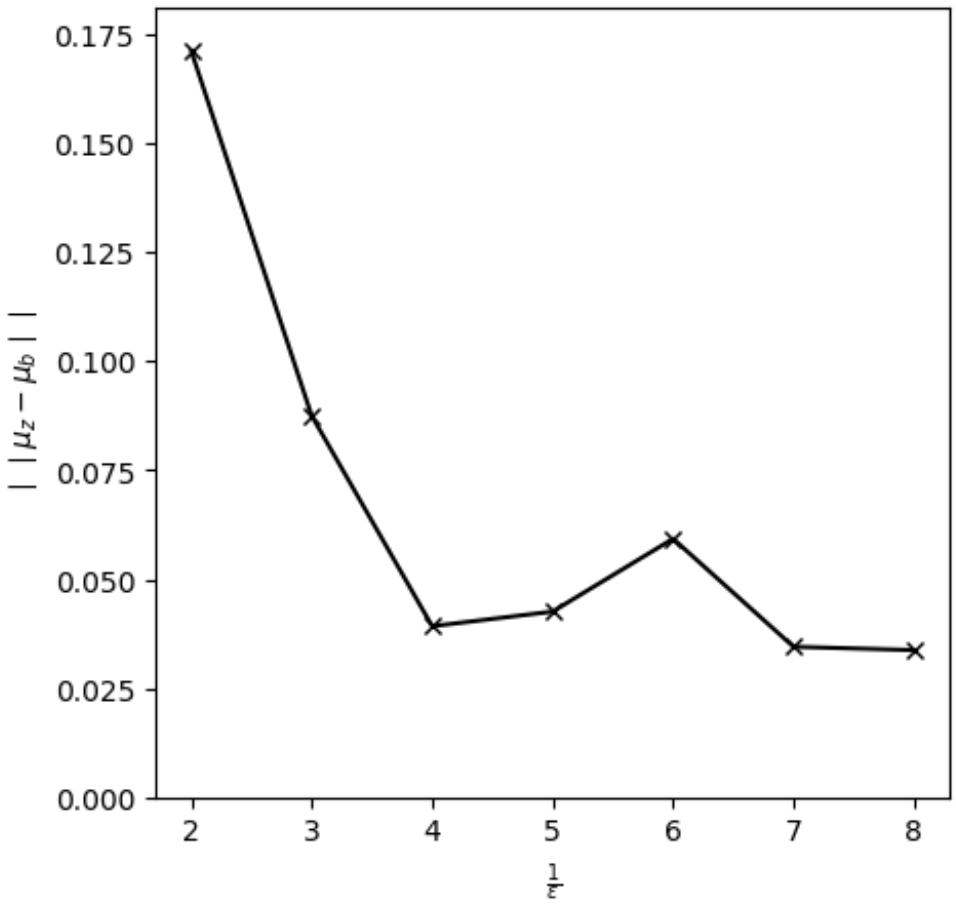
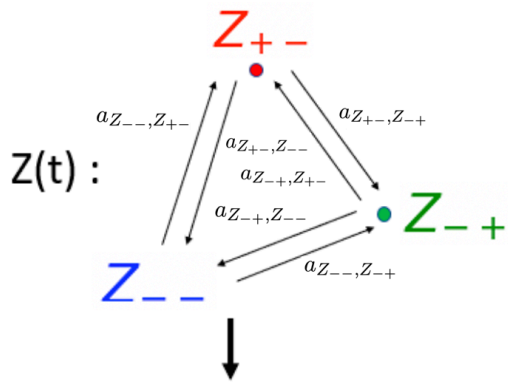


Figure: Comparison between the stationary distribution of the coarse grained model and the one deduced from the PDMP

Phenomenological model



$$\begin{cases} E_i(t) : 0 \xrightarrow{k_{on,i}(X_{eq,Z(t)})} 1,\ 1 \xrightarrow{k_{off,i}} 0 \\ \dot{X}_i(t) = d_i(E_i(t) - X_i(t)) \end{cases}$$

The approximate stationary distribution appears as a Beta mixture :

$$u \sim \sum_{z \in Z} \mu_z \prod_{i=1}^{n} Beta(\frac{k_{z_i}}{d_i}, \frac{k_{off,i}}{i}),$$

where $k_{z_i} = k_{on,\Theta,i}(X_{eq}, Z)$

Importance of the function k_{on}

- For a given network Θ, we denote :

$$\alpha_\Theta = \left(\mu_z, (k_{z_i}, k_{off,i})_{i=1,\cdots,n}\right)_{z\in Z}$$

- We define the function $k_{on,\alpha}$:

$$k_{on,\alpha_\Theta,j}(x) = \frac{\sum_{z\in Z} \mu_z k_{z,j} \prod_{i=1}^{n} Beta(k_{z_i}, k_{off,i})(x)}{\sum_{z\in Z} \mu_z \prod_{i=1}^{n} Beta(k_{z_i}, k_{off,i})(x)} = \mathbb{E}(k_{z_j} \mid X)$$

Theorem

The stationary distribution of the PDMP driven by the function $k_{on,\alpha_\Theta}(x)$ is exactly the Beta mixture of parameters α_Θ

Transition

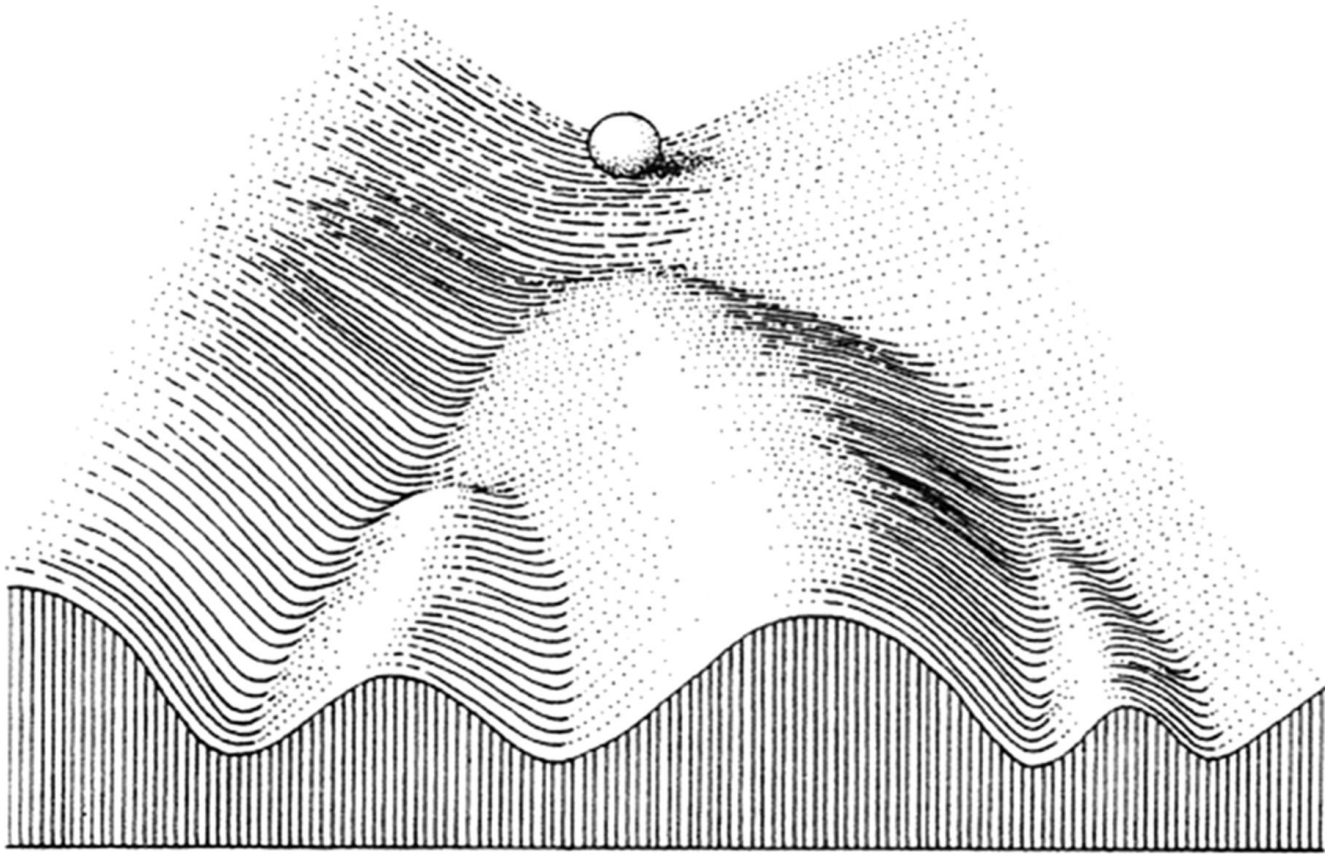


Figure: The Beta mixture is a mathematical representation of the Waddington's epigenetic landscape

Transition

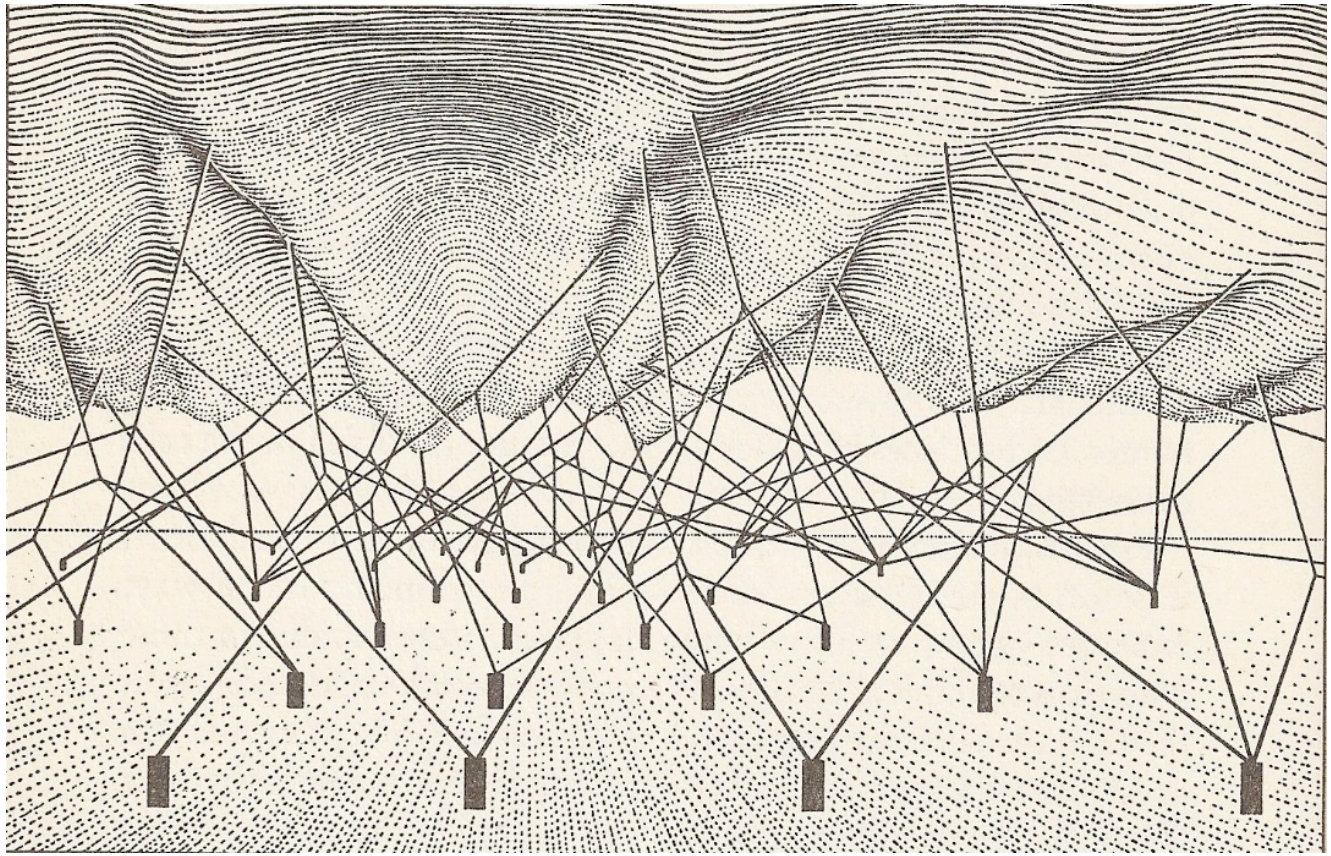


Figure: The tension of the string represents the chemical forces exerted by the genes

Next step

- We have :

$$\text{GRN} \longrightarrow \text{Coarse-grained model} \longrightarrow \text{Beta mixture}$$

- We want :

$$\text{Data} \longrightarrow \text{Beta mixture} \longrightarrow \text{GRN}$$

The problem of inference

- **Question** : Given a set of data X and an empirical distribution u_X, we would like to find the Θ such that the stationary distribution of the PDMP process u_Θ is the closest from u_X

$\Longrightarrow$ We assume the **non identifiability** of the problem, as the function $\Theta \longrightarrow u_\Theta$ itself is not injective.

- Problem 0 : u_Θ **is not explicitly known**

The problem of inference

- We denote : $\alpha = \left(\mu_z, (k_{z_i}, k_{off,i})_{i=1,\dots,n}\right)_{z \in Z}$, the parameters describing a beta mixture (associated to the PDMP)

- **New question** : Given a Beta mixture fitting the data set X, characterized by $\alpha_0 = \alpha(X)$, what GRN could have generated it (in a stationary way) ?

$\Longrightarrow$ Implicitly, we suppose that from a data set, we can not get more information that the ones given by a Beta mixture

The problem of inference

- We denote R_0 the risk minimized by the numerical procedure of the first Section, defining α_Θ from Θ (ideally, R_0 would be a KL divergence) :

$$\alpha_\Theta = \arg\min_\alpha R_0(\Theta, \alpha)$$

- **Reformulation** : Given α_0, we want to find $\hat{\Theta}$ such that :

$$\alpha_0 = \arg\min_\alpha R_0\left(\hat{\Theta}, \alpha\right)$$

The problem of inference

- Problem 1 : We have **no analytical link** between α_Θ and Θ, and it would be **time-consuming** to use the previous numerical method for computing the best α_Θ

- Problem 2 : It is difficult to know for a class of function $k_{on,\hat{\Theta}}$ if it exists $\hat{\Theta}$ such that $\alpha_0 = \arg\min_{\alpha} R_0\left(\hat{\Theta}, \alpha\right)$

For example, this is not the case for any α_0 for the sigmoid

The problem of inference

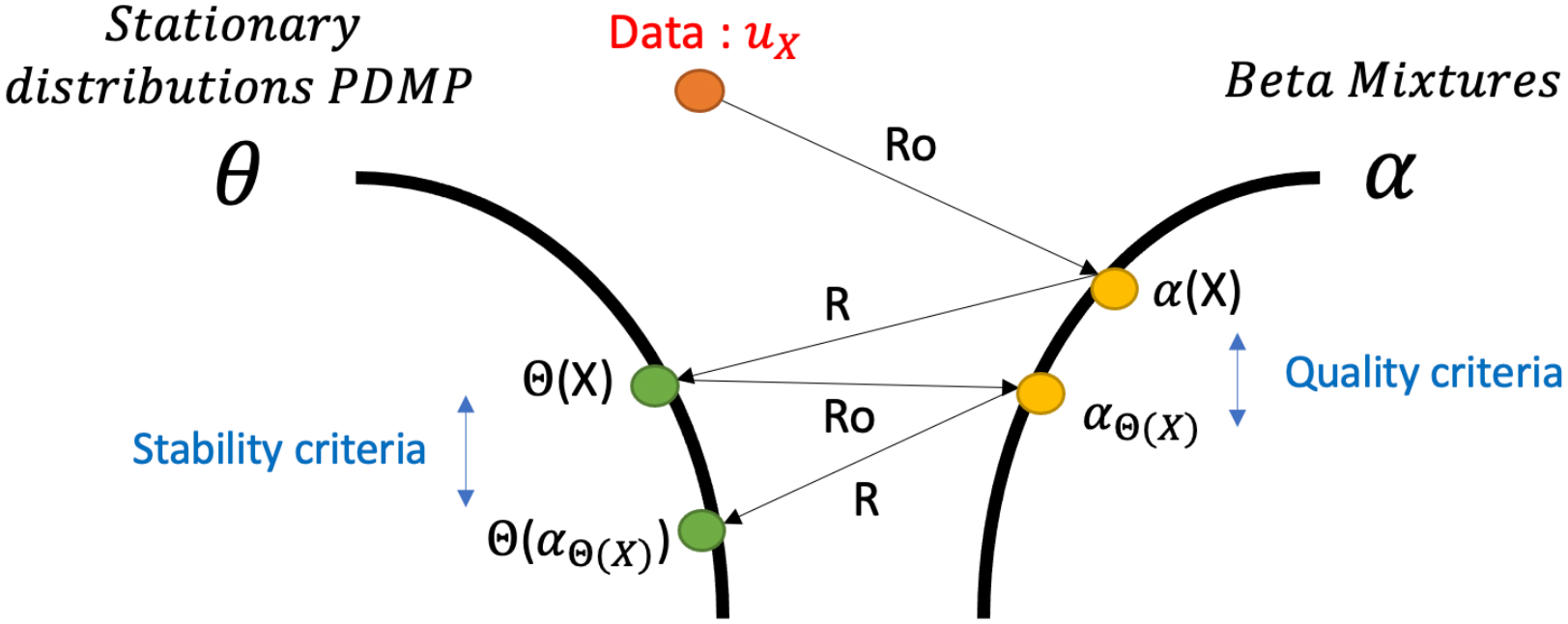
- New aim : find a risk R, accessible, such that :

$$\hat{\Theta} = \arg\min_{\Theta} R(\Theta, \alpha_0)$$

- Problem 2bis : We could rather ask :

$$\begin{cases} \hat{\Theta} \in \arg\min_{\Theta} R_0(\Theta, \alpha_0) & \text{quality condition} \\ \hat{\Theta} = \arg\min_{\Theta} R(\Theta, \alpha_{\hat{\Theta}}) & \text{stability condition} \end{cases}$$

The problem of inference



Focus on the problem 1 : find R

- With $k_{on,\Theta}$ and $k_{on,\alpha}$, we defined previously **two PDMP systems which have two close stationary distributions**, u_Θ and $u_{\alpha\Theta}$

$\Longrightarrow$ Could we build a risk R from the promoter frequency and not from the stationary distribution ?

$\Longrightarrow$ Does it mean that $k_{on,\theta}$ should be close than k_{on,α_0} for every x ?

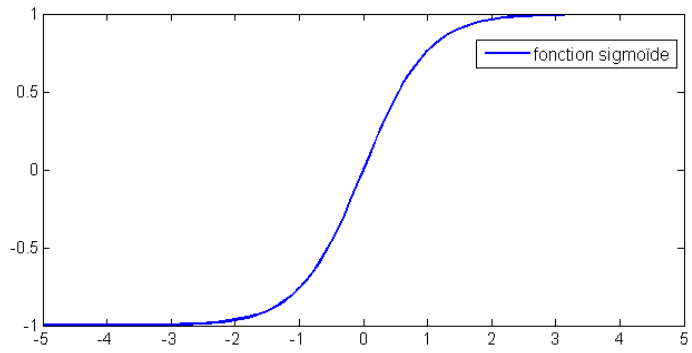
Importance of the function k_{on}

- As the basins are deep when $\varepsilon \ll 1$, the function $k_{on,\alpha_\Theta,i}$ are supposed to be steep, and appear closed to Hill functions

- Each $k_{on,\theta,i}$ can be represented by a **Hill function**. Intuitively, if the Hill function is sufficiently steep, it will be close on every point of the gene expression space to the function $k_{on,\alpha_\Theta,i}$

Example : the sigmoid

$$k_{on,\Theta,i}(X) = k_{1,i}\frac{e^{\beta_i+\sum\theta_{ji}X_j}}{1+e^{\beta_i+\sum\theta_{ji}X_j}}$$



Naive problem

- A first option is then to consider the risk :

$$R(\Theta, \alpha) = \mathbb{E}_X \left(\sum_{i=1}^{n} \mid k_{on,\Theta,i}(X) - k_{on,\alpha,i}(X) \mid \right)$$

- Then, from a data set $X = (X_1, \cdots, X_{n_c})$, we would compute $\alpha(X)$ and then minimize the risk

$$\hat{R}(\Theta, \alpha(X)) = \sum_{c=1}^{n_c} \sum_{i=1}^{n} \mid k_{on,\Theta,i}(X_c) - k_{on,\alpha(X),i}(X_c) \mid$$

WKB approximation

Now, we justify that this risk R indeed minimizes in a certain sens the distance between the associated distributions, and derive a new proposal.

- We seek a distribution of the form :

$$\forall e,\, u_e(x,t) = \zeta_e(x,t)\exp\left(-\frac{V(x,t)}{\varepsilon}\right)$$

WKB approximation

- We make the following Taylor expansion at the second order with respect to the scaling factor ε :

$$\begin{cases} \zeta = \zeta_0 + \varepsilon\zeta_1 + o(\varepsilon^2) \\ V = V_0 + \varepsilon V_1 + o(\varepsilon^2) \end{cases}$$

- V_0 appears as the solution of an **Hamilton-Jacobi** equation :

$$H_{k_{on}}(x, D_x V_0(x)) + \frac{\partial V_0}{\partial t} = 0$$

WKB approximation

- We denote $V_{k_{on}}$ the leading order term in ε of a solution to the stationary HJ equation for a certain k_{on} function

- We define a new risk :

$$\bar{R}(\Theta, \alpha) = \mathbb{E}_X \left(\mid H_{k_{on,\Theta}}(X, D_X V_{k_{on,\alpha}}(X)) \mid \right) = \int_\Omega \mid \frac{\partial}{\partial t} u_\alpha(X) \mid_0 dX$$

Formally, this quantity measures **how fast a PDMP process driven by $k_{on,\Theta}$ is going to evolve when distributed initially by u_α**

New proposal

- For any Θ such that $\nabla V_{k_{on,\Theta}}$ vanishes only on single points, we show that :

$$\bar{R}(\Theta, \alpha) = 0 \longleftrightarrow V_{k_{on,\Theta}} = V_{k_{on,\alpha}}$$

$\Longrightarrow$ It measures how far is the quasipotential $V_{k_{on,\Theta}}$ from $V_{k_{on,\alpha}}$.

- A large deviation analysis had shown **the importance of the quasipotential** to describe the dynamics of the process :

Elias Ventre et al. "Reduction of a stochastic model of gene expression: Lagrangian dynamics gives acces to basins of attraction as cell types and metastability". In: *bioRxiv* (2020).

New proposal

- As $\forall x,\ H_{k_{on,\alpha}}(x, p_{k_{on,\alpha}}(x)) = 0$, we can show that :

$$\overline{R}(\Theta, \alpha) \leq \mathbb{E}\left(\sum_{i=1}^{n} \mid k_{on,\alpha,i} - k_{on,\Theta,i} \mid + O\left(\sum_{i=1}^{n} (k_{on,\alpha,i} - k_{on,\Theta,i})^2\right)\right)$$

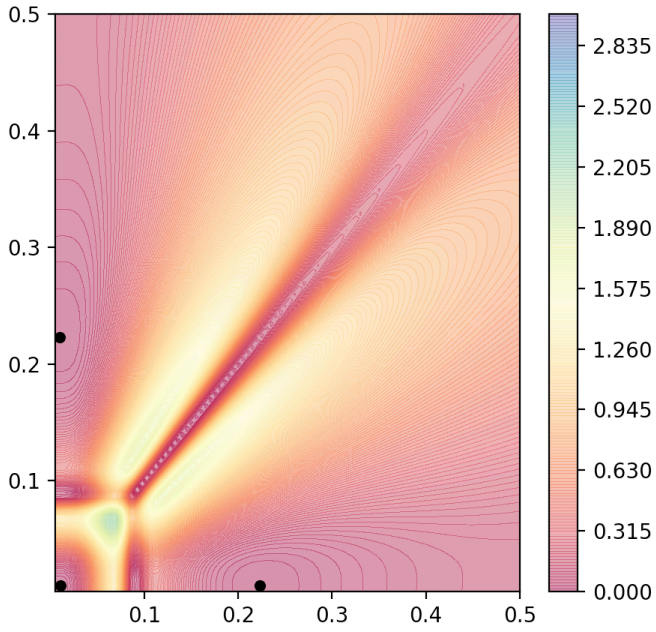
$\Longrightarrow$ **The previous naive proposal $R(\Theta, \alpha)$ minimizes an upper bound of $\overline{R}(\Theta, \alpha)$**

Intuitively, $\overline{R}$ is weaker than R : it allows more differences between the k_{on} without making worst the difference between the stationary distributions

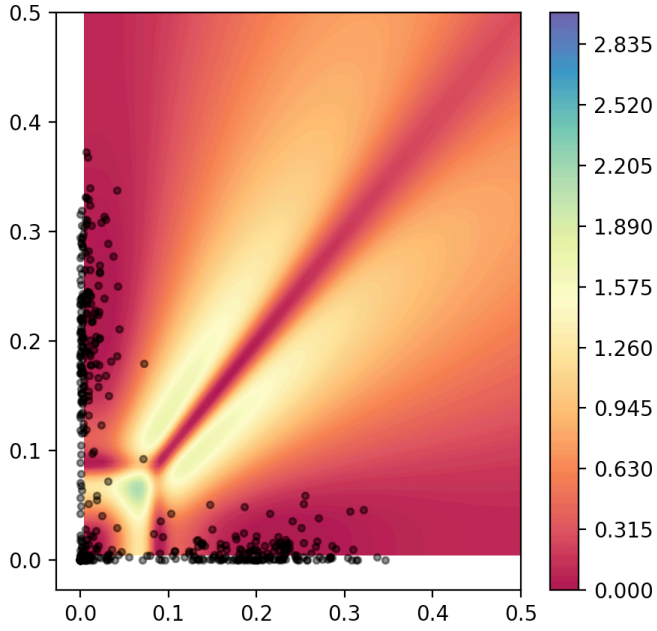
Analysis of the problem 2 : stability criteria

- We would like to verify that $\hat{\Theta}$:

$$\hat{\Theta} = \arg\min_{\Theta} \bar{R}(\Theta, \alpha_{\hat{\Theta}})$$



Values of $| H_{k_{on},\hat{\Theta}}(x, p_{k_{on},\alpha_{\hat{\Theta}}}(x)) |$ on the gene expression space



Cells concentrate where H is small : $\bar{R}(\hat{\Theta}, \alpha_{\hat{\Theta}})$ is then small

Analysis of the problem 2 : quality criteria

- We also would like to verify :

$$\hat{\Theta} \in \arg\min_{\Theta} R_0(\Theta, \alpha_0)$$

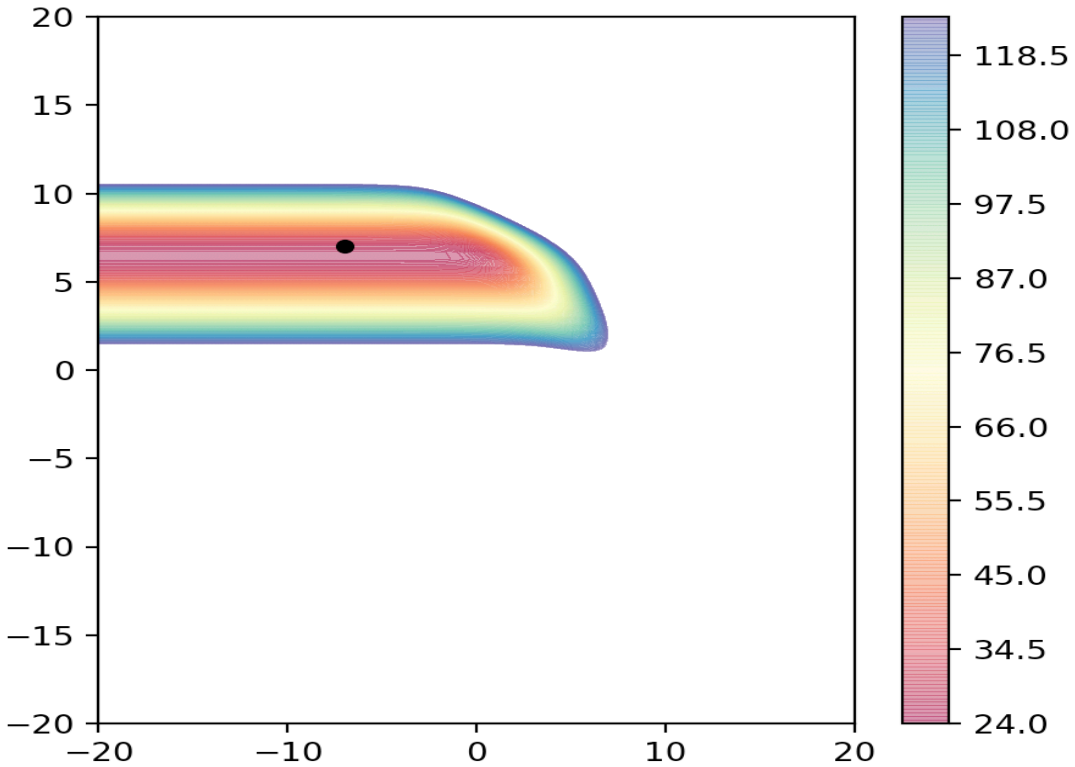
This not accessible but we could consider that

$$R_0(\Theta, \alpha) = KL\left(u_{\alpha_\Theta} \,||\, u_\alpha\right)$$

and then verify that $KL\left(u_{\alpha_{\hat{\Theta}}} \,||\, u_{\alpha_0}\right)$ is small

Non-identifiability

- In simple cases as the toggle switch, we see clearly that the problem is non identifiable : many Θ could lead to the same α



Example : Values of the risk $\overline{R}$ for a two genes network, by varying 2 values of theta, fixing the two others

Algorithm in practice

- Given a set of data $X = (X_1, \cdots, X_{n_c})$, find an $\alpha(X)$ fitting the data
- Compute :

$$\hat{\Theta}(X) = \alpha \arg\min_{\Theta} \hat{\bar{R}}(\Theta, \alpha(X))$$

- Find $\alpha_{\hat{\Theta}(X)}$ numerically.

Verify that the quality criteria $KL\left(u_{\alpha_{\hat{\Theta}(X)}} \,||\, u_{\alpha(X)}\right)$ is small

and that the stability criteria $\hat{\bar{R}}(\hat{\Theta}(X), \alpha_{\hat{\Theta}(X)})$ is small too.

Open questions

- For which type of k_{on} does it always exist, given any α_0, a matrix $\hat{\Theta}$ such that $\alpha_{\hat{\Theta}} = \alpha_0$?

- When this is the case, we would like to prove that the $\hat{\Theta}$ given by $\overline{R}$ verifies :

$$\alpha_0 = \arg\min_{\alpha} R_0(\hat{\Theta}, \alpha)$$

- When this is not the case, we need to quantify :

$$KL\left(u_{\alpha_{\hat{\Theta}}} \,||\, u_{\alpha_0}\right)$$

Work in progress

- The full model includes mRNAs :

$$\begin{cases} E(t) : 0 \xrightarrow{k_{on}(x)} 1, 1 \xrightarrow{k_{off}} 0, \\ M'(t) = s_0 E(t) - d_0 M(t), \\ P'(t) = s_1 M(t) - d_1 P(t). \end{cases}$$

$\longrightarrow$ Giving that the Hill function k_{on} is sufficiently steep, the mRNA distribution is also well approximated by a Beta mixture

$\longrightarrow$ We implement a specific **RJ-MCMC algorithm** to infer a set of parameters α from RNA-seq data

$\longrightarrow$ We obtain a collection of Θ !

To be continued…