# Nonparametric estimator of the distribution of fitness effects of new mutations

Guillaume Garnier
Under the supervision of Marie Doumic, Marc Hoffmann and Lydia Robert

25/01/2024

# Introduction

- All organisms are subject to mutations
- These new traits can change the selective value (fitness) of an individual
- *Fitness* : ability of an individual with a certain genome to survive and reproduce
- How these mutations affect selective value is a central question in evolutionary biology

- The density of the distribution of these effects is called the Distribution of Fitness Effect (DFE)

# Introduction

Why study the DFE ?

- DFE is important of these arising mutations define the range of possible evolutionary trajectories a population can follow
- *Study the effects of new mutations in an individual to see if they are beneficial or not*
- *Understanding and quantifying the genetic diversity of human diseases and its future evolution*
- *Predict the consequences of maintaining a small population of animals or plants, as in captive breeding programs*

# L'expérimentation

Goal :

*Inferring DFE from experimental measurements of selective value over time*

# What data ?

Two experimental protocols (Robert et al. 2018 [ROR+18])

- See in real time the appearance of mutations in e.coli
- New measurements of cell fitness

$\Longrightarrow$ New data to estimate the DFE

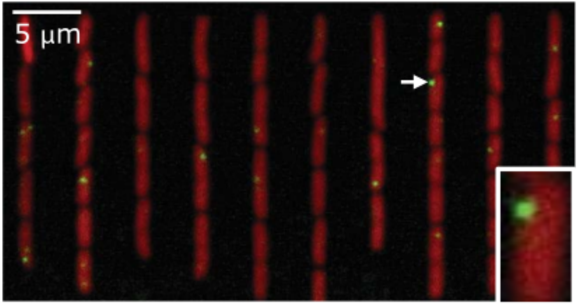


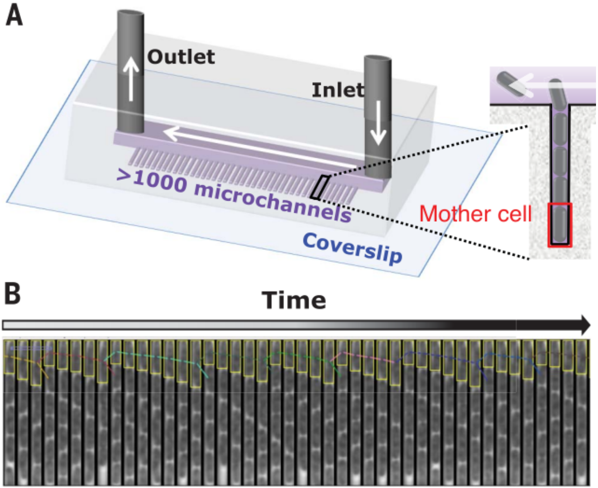

Figure – L. Robert and al, Science, 2018

# What data ?

cf. Video

# microfluidic Mutation Accumulation ($\mu$MA) experiment

- Measuring the fitness of cells
- 1476 parallel and independent channels



# microfluidic Mutation Accumulation ($\mu$MA) experiment

cf. Video

# Model Building

- A first model. (Robert, 18)
- The mutations are deleterious and appear according to a Poisson process $\mathcal{P}(\lambda t)$
- $(W_t)_{t \in \mathbb{R}^+}$ the selective value over time of an individual

$$\boxed{s_i = \frac{W_{t_{i-1}} - W_{t_i}}{W_{t_{i-1}}} \;,\; i > 0},$$

  $s_i$ effect of the $\{i\}$–i-th mutation on the fitness of the individual.
- If $(s_i)_i$ are i.i.d

$$\boxed{\frac{W_t}{W_0} = \prod_{i=1}^{N_t} (1 - s_i) \;,\; N_t \sim \mathcal{P}(\lambda t)}$$

- DFE = probability density of $s_i$

# Model Building

- By taking the logarithm, we have

$$\ln W_t = \sum_{i=1}^{N_t} \ln(1 - s_i)\,, \;\; N_t \sim \mathcal{P}(\lambda t)\,, \;\; \lambda > 0$$

- It is a compound Poisson process : $X_i \sim \ln(1 - s_i)$ et $Y_t \sim \ln W_t$,

$$\boxed{Y_t = \sum_{i=1}^{N_t} X_i}\,.$$

# Model Building

- We want to model the errors in the measurements

$$\frac{W_t}{W_0} = \prod_{i=1}^{N_t} (1 - s_i)\varepsilon_t\,, \;\; N_t \sim \mathcal{P}(\lambda t)\,, \;\; \lambda > 0\,,$$

- By taking the logarithm, we have (10). Dans ce cas on a

$$\boxed{Z_t := Y_t + \xi_t = \sum_{i=1}^{N_t} X_i + \xi_t}\,,$$

# Model Building

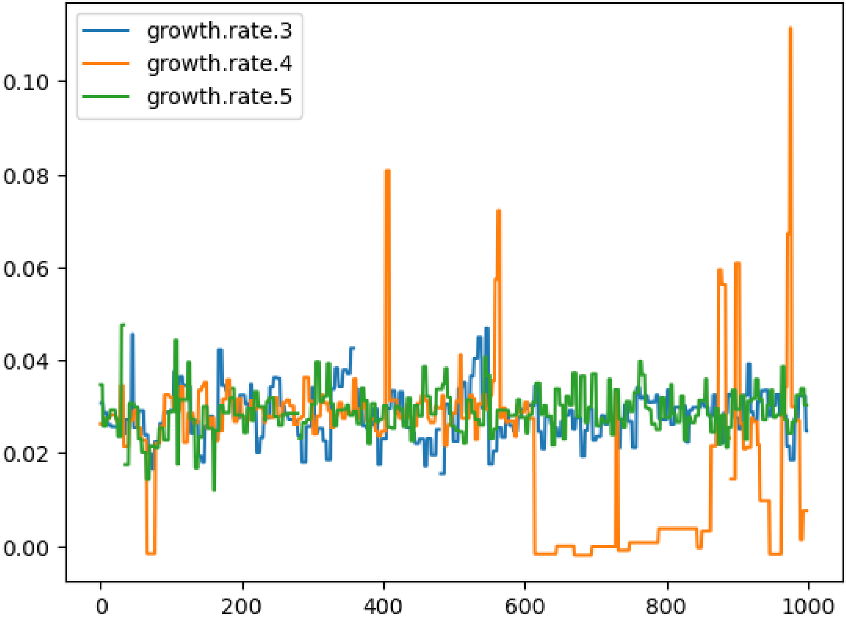
1. $Z_t^j$ : noisy measure in channel $j \in J$ at time $t$.
2. $N_t^j$ : number of mutation in channel $j$. $(N_j(t), j \geq 1)$ are *i.i.d* Poisson processes with intensity $\lambda \in (0, \infty)$.
3. $X_k^j$ jump of $k$-th mutation in channel $j$. $(X_i^j)_{i,j\geq 0}$ are *i.i.d* with density $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$.
4. $\varepsilon_t^j$ represents the measurement noise at time $t$ for channel $j$. $(\varepsilon_t^j)_{j\geq 0}$ are *i.i.d* and that $\mathbb{E}(\varepsilon_t^j) = 0$.

We consider a noisy compound Poisson process :

$$\boxed{Z_t^j = \left(\sum_{k=1}^{N_t^j} X_k^j\right) + \varepsilon_t^j,\ t \geq 0}.$$

cf. Video 3

| | time | generation | growth.rate | time.1 | generation.1 | growth.rate.1 | time.2 | generation.2 | growth.rate.2 | time.3 | ... | growth.rate.1247 | time.1248 | generation.1248 | growth.rate.1248 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0.015533 | 0 | 1 | NaN | 0 | 1 | NaN | 0 | ... | 0.028119 | 0 | 1 | 0.100498 |
| 1 | 4 | 1 | 0.015533 | 4 | 2 | 0.032278 | 4 | 2 | 0.030302 | 4 | ... | 0.028119 | 4 | 1 | 0.100498 |
| 2 | 8 | 2 | 0.031221 | 8 | 2 | 0.032278 | 8 | 2 | 0.030302 | 8 | ... | 0.026839 | 8 | 1 | 0.100498 |
| 3 | 12 | 2 | 0.031221 | 12 | 2 | 0.032278 | 12 | 2 | 0.030302 | 12 | ... | 0.026839 | 12 | 1 | 0.100498 |
| 4 | 16 | 2 | 0.031221 | 16 | 2 | 0.032278 | 16 | 2 | 0.030302 | 16 | ... | 0.026839 | 16 | 2 | 0.026508 |
| 5 | 20 | 2 | 0.031221 | 20 | 2 | 0.032278 | 20 | 2 | 0.030302 | 20 | ... | 0.026839 | 20 | 2 | 0.026508 |
| 6 | 24 | 2 | 0.031221 | 24 | 2 | 0.032278 | 24 | 2 | 0.030302 | 24 | ... | 0.026839 | 24 | 2 | 0.026508 |
| 7 | 28 | 3 | 0.029121 | 28 | 3 | 0.030657 | 28 | 2 | 0.030302 | 28 | ... | 0.026839 | 28 | 2 | 0.026508 |
| 8 | 32 | 3 | 0.029121 | 32 | 3 | 0.030657 | 32 | 2 | 0.030302 | 32 | ... | 0.026839 | 32 | 2 | 0.026508 |
| 9 | 36 | 3 | 0.029121 | 36 | 3 | 0.030657 | 36 | 2 | 0.030302 | 36 | ... | 0.041219 | 36 | 2 | 0.026508 |


growth.rate.3
growth.rate.4
growth.rate.5
0.10
0.08
0.06
0.04
0.02
0.00
0
200
400
600
800
1000

# Estimate the DFE

*In each model, we want to estimate the probability density of* $X_i$ *from the observations*

# Estimate the DFE

*In each model, we want to estimate the probability density of $X_i$ from the observations*

We want an approximation

$$\boxed{\sup_{f \in \mathcal{F}} \mathbb{E}_\theta[\|f_n, f\|^2] \leq C\psi_n^2}$$

# Strategy, Tools & Methods

**Strategy :** We want to estimate the characteristic function of $X$ :

**(heuristic)** If $\varphi_X(\xi) \simeq \widehat{\varphi}_X(\xi)$, then $f(x) \simeq \widehat{f}(x)$

Indeed, the characteristic function $\varphi_X \to$ Density $f$ of $X$ :

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \varphi_X(\xi) e^{-ix\xi} d\xi$$

# Building the estimator

We consider a noisy compound Poisson process :

$$\boxed{Z_t^j = \left( \sum_{k=1}^{N_t^j} X_k^j \right) + \varepsilon_t^j,\ t \geq 0}.$$

For a single channel $Z_t^j$, the characteristic function is :

$$\boxed{\forall u \in \mathbb{R},\ \varphi_{Z_t^j}(u) = e^{-\lambda t + \lambda t \varphi_X(u)} \cdot \varphi_\varepsilon(u)}$$

# Building the estimator

Consider two different times $0 < t_1 < t_2$, then

$$\frac{\varphi_{Z_{t_2}}}{\varphi_{Z_{t_1}}} = e^{-\lambda(t_2-t_1)+\lambda(t_2-t_1)\varphi_X(u)}$$

Then

$$\varphi_X(u) = 1 + \frac{1}{t_2 - t_1}\Big(\log\varphi_{Z_{t_2}}(u) - \log\varphi_{Z_{t_1}}(u)\Big)$$

# Building the estimator

Consider two different times $0 < t_1 < t_2$, then

$$\frac{\varphi_{Z_{t_2}}}{\varphi_{Z_{t_1}}} = e^{-\lambda(t_2-t_1)+\lambda(t_2-t_1)\varphi_X(u)}$$

Then

$$\varphi_X(u) = 1 + \frac{1}{t_2 - t_1}\Big(\log \varphi_{Z_{t_2}}(u) - \log \varphi_{Z_{t_1}}(u)\Big)$$

It leads us to define

$$\widehat{\varphi}^J_X(u) = 1 + \frac{1}{t_2 - t_1}\Big(\log \widehat{\varphi}^J_{Z_{t_2}}(u) - \log \widehat{\varphi}^J_{Z_{t_1}}(u)\Big)$$

with

$$\widehat{\varphi}'^J_{Z_\tau}(u) = \frac{1}{J}\sum_{j=1}^{J} iZ^j_\tau e^{iuZ^j_\tau}, \ \widehat{\varphi}^J_{Z_\tau}(u) = \frac{1}{J}\sum_{j=1}^{J} e^{iuZ^j_\tau},$$

$$\log \widehat{\varphi}^J_{Z_\tau}(u) = \int_0^u \frac{\widehat{\varphi}'^J_{Z_\tau}(z)}{\widehat{\varphi}^J_{Z_\tau}(z)} dz$$

# Building the estimator

As there is no guarantee that the previous quantities will not explode, a cut-off is added to ensure this.

$$\widetilde{\varphi}_X^J(u) = 1 + \frac{1}{t_2 - t_1}\Big\{\log \widehat{\varphi}_{Z_{t_2}}^J(u) \cdot \mathbf{1}_{|\log \widehat{\varphi}_{Z_{t_2}}^J(u)| \leq \ln(J)} - \log \widehat{\varphi}_{Z_{t_1}}^J(u) \cdot \mathbf{1}_{|\log \widehat{\varphi}_{Z_{t_1}}^J(u)| \leq \ln(J)}\Big\}$$

We estimate $f$ by Fourier inversion.
For any $m \in (0, \infty)$ ,

$$\widehat{f}_{m,J}(x) = \frac{1}{2\pi}\int_{-m}^{m} e^{-iux}\widetilde{\varphi}_X^J(u)du\,,\ x \in \mathbb{R}$$

# Building the estimator

As there is no guarantee that the previous quantities will not explode, a cut-off is added to ensure this.

$$\widetilde{\varphi}_X^J(u) = 1 + \frac{1}{t_2 - t_1}\left\{ \log \widehat{\varphi}_{Z_{t_2}}^J(u) \cdot \mathbf{1}_{|\log \widehat{\varphi}_{Z_{t_2}}^J(u)| \leq \ln(J)} - \log \widehat{\varphi}_{Z_{t_1}}^J(u) \cdot \mathbf{1}_{|\log \widehat{\varphi}_{Z_{t_1}}^J(u)| \leq \ln(J)} \right\}$$

We estimate $f$ by Fourier inversion.
For any $m \in (0, \infty)$ ,

$$\widehat{f}_{m,J}(x) = \frac{1}{2\pi} \int_{-m}^{m} e^{-iux} \widetilde{\varphi}_X^J(u) du \, , x \in \mathbb{R}$$

Here, the choice of $m$ is very important because it defines the frequencies that we keep to apply the inverse Fourier transformation

# Theorem : convergence of the estimator

For all reals $0 < t_1 < t_2$ such that $t_2 \leq \frac{1}{4}\log(Jt_2)$ $Jt_1 \to \infty, Jt_2 \to \infty$ as $J \to \infty$ and for any $m < C^J_{t_1,t_2}$, the following inequality holds

$$\boxed{\begin{aligned} \mathbb{E}\left(\|\widehat{f}_{m,J} - f\|^2\right) \leq \|f_m - f\|^2 + \sum_{i=1}^{2} \frac{4e^{4t_i}}{J(t_2-t_1)^2} \int_{-m}^{m} \frac{du}{|\varphi_\varepsilon(u)|^2} \\ + \frac{4K_{J,t_1,t_2}}{(t_2-t_1)^2} \cdot \left( \frac{\mathbb{E}[X_i^2]}{Jt_i} + \frac{\mathbb{E}[\varepsilon^2]}{Jt_i^2} + 4\frac{m}{(Jt_i)^2} \right) \end{aligned}}.$$

where $K_{J,t_1,t_2}$ and $C^J_{t_1,t_2}$ depends on $m, t_1, t_2$ and $\log \varphi_\varepsilon(\cdot)$.

# Theorem : adaptative estimator

Question : How to select $m$ ?

- The dominant terms :
  biais term : $\int_{u\in[-m,m]} |\varphi_X(u)|^2 \, \mathrm{d}u$
  variance term : $\frac{4e^{4t_2}}{J(t_2-t_1)^2} \int_{-m}^{m} \frac{\mathrm{d}u}{|\varphi_\varepsilon(u)|^2}$
- Through differentation, the optimal $\overline{m_J}$ satisfies

$$|\varphi_X(\overline{m_J})|^2 = \frac{4ae^{4t_2}}{J(t_2-t_1)^2}(1+\overline{m_J}^2)\,.$$

then

$$\boxed{\left|\frac{\varphi_X(\overline{m_J})}{\sqrt{(1+\overline{m_J}^2)}}\right|^2 = \frac{4ae^{4t_2}}{J(t_2-t_1)^2}}\,.$$

# Theorem : adaptive estimator

It leads us to define the empirical cutoff parameter

$$\boxed{\widehat{m}_J = \max\left\{u \geq 0 : \left|\frac{\overline{\varphi}_X(u)}{\sqrt{1+u^2}}\right| \geq \frac{\kappa_{J,t_1,t_2}}{\sqrt{J}(t_2-t_1)}\right\} \wedge \left(J(t_2-t_1)^2\right)^{\alpha}, \quad \alpha \in (0,1)}$$

where

$$\overline{\varphi}_X^J(u) = \widetilde{\varphi}_X^J(u) \cdot \mathbb{1}\left|\frac{\widetilde{\varphi}_X^J(u)}{\sqrt{1+u^2}}\right| \geq \frac{\kappa_{J,t_1,t_2}}{\sqrt{J}(t_2-t_1)}$$

and $\kappa_J = 2e^{2t_2} + \kappa\sqrt{\ln(J(t_2-t_1)^2)}$, $\kappa > 0$

# Theorem : adaptive estimator

For all reals $0 < t_1 < t_2$ such that $t_2 \leq \frac{1}{4}\log(Jt_2)$ and $(m)^\alpha < C^J_{t_1,t_2}$, $Jt_1 \to \infty$ , $Jt_2 \to \infty$ as $J \to \infty$. Then,

$$\mathbb{E}\Big[\|\overline{f}_{\widehat{m}_J} - f\|^2\Big] \leq \inf_{m\in[0,m^\alpha_m]}\left\{\|f_m - f\|^2 + C\frac{\ln(J(t_2-t_1)^2)\cdot m\cdot(1+m^2)}{J(t_2-t_1)^2} + \widetilde{C}A\right\}$$
$$+\left(2+\frac{2\log(J)}{(t_2-t_1)}\right)^2\cdot T_J$$

where $A, T_j$ and $c(\theta)$ satisfies good conditions.

# Numerical Result

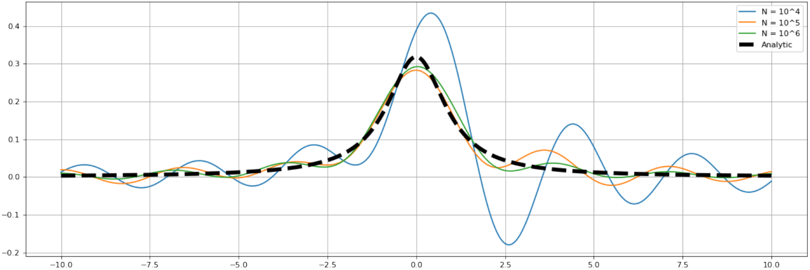



FIGURE – Reconstruction of the Cauchy $\mathcal{C}(0,1)$ distribution with J channels, corrupted by a Gaussian noise $\mathcal{N}(0,1)$ with $J \in 10^4, 10^5, 10^6$. $t_1 = 0.1, t_2 = 1, m = 2$

# Numerical Result

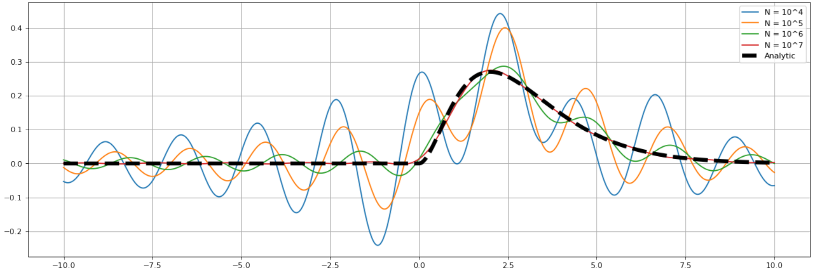



FIGURE – Reconstruction of the Gamma $\Gamma(3)$ distribution with J channels, corrupted by a Gaussian noise $\mathcal{J}(0,1)$ with $J \in 10^4, 10^5, 10^6, 10^7$. $t_1 = 0.1, t_2 = 1, m = 3$

# And now? An minimax estimator

A problem of nonparametric estimation is characterized by

- A class of functions $\mathcal{F}$ that contains $f$
- A family of probability measure $\{\mathbb{P}_f, f \in \mathcal{F}\}$ associated with the observations.

## Definition (Maximum risk)

$$r(\widehat{f_n}) = \sup_{f \in \mathcal{F}} \mathbb{E}_f(\|\widehat{f_n} - f\|^2)$$

**Previous goal : Obtain upper bounds on the maximum risk**, *i.e*

$$r(\widehat{f_n}) = \sup_{f \in \mathcal{F}} \mathbb{E}_f(\|\widehat{f_n} - f\|^2) \leq C\psi_n^2 \quad \text{where} \quad \psi_n \to 0\,.$$

# And now ? An minimax estimator

**How to know that we have the best possible estimator ?**

Definition (Minimal risk)

$$\boxed{\mathcal{R}_n^* = \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f(\|\widehat{f}_n - f\|^2)}$$

If you have a bound on the maximum risk

$$r(\widehat{f}_n) = \sup_{f \in \mathcal{F}} \mathbb{E}_f(\|\widehat{f}_n - f\|^2) \le C\psi_n^2 \quad \text{where} \quad \psi_n \to 0\,.$$

then

$$\boxed{\limsup_{n\to\infty} \psi_n^{-2} \mathcal{R}_n^* \le C}$$

# And now? An minimax estimator

## Definition (Optimal rate of convergence)

A positive sequence $(\psi_n)_n$ is called an optimal rate of convergence of estimators on $\mathcal{F}$ if there exists $c > 0$ and $C > 0$

$$\boxed{\limsup_{n\to\infty} \psi_n^{-2} \mathcal{R}_n^* \le C} \quad \text{and} \quad \boxed{\liminf_{n\to\infty} \psi_n^{-2} \mathcal{R}_n^* \ge c}.$$

## Definition (Rate optimal estimator)

An estimator $f_n^*$ satisfying

$$\sup_{f\in\mathcal{F}} \mathbb{E}_f(\|f_n^* - f\|) \le C^* \psi_n$$

where $(\psi_n)_n$ is an optimal rate of convergence of estimators

# And now? An minimax estimator

**Definition (Asymptotically efficient estimator)**

An estimator $f_n^*$ is called asymptotically efficient if

$$\lim_{n \to \infty} \frac{r(\theta_n^*)}{\mathcal{R}_n^*} = 1 \, .$$

# And now ? An minimax estimator

How to prove

$$\liminf_{n\to\infty} \psi_n^{-2}\mathcal{R}_n^* \geq c?$$

There is a classical "general scheme"

- Step 1. Reduction to bounds in probability ;
- Step 2. Reduction to a finite number of hypotheses ;
- Step 3. Choice of $2s$-separated hypotheses.

# And now ? An minimax estimator

**Step 1. Reduction to bounds in probability**
We obtain a lower bound on $\mathbb{E}_\theta(\psi_n^{-2}\|\widehat{f_n} - f\|^2)$ with the Markov inequality.

For any real $\alpha > 0$,

$$\begin{aligned}\mathbb{E}_\theta(\psi_n^{-2}\|\widehat{f_n} - f\|^2) &\geq \alpha^2 \mathbb{P}(\psi_n^{-1}\|\widehat{f_n} - f\| \geq \alpha) \\ &= \alpha^2 \mathbb{P}(\|\widehat{f_n} - f\| \geq s)\end{aligned}$$

where $s = s_n = A\psi_n$.

# And now? An minimax estimator

**Step 1. Reduction to bounds in probability**
We obtain a lower bound on $\mathbb{E}_\theta(\psi_n^{-2}\|\widehat{f_n} - f\|^2)$ with the Markov inequality.

For any real $\alpha > 0$,

$$\begin{aligned}\mathbb{E}_\theta(\psi_n^{-2}\|\widehat{f_n} - f\|^2) &\geq \alpha^2 \mathbb{P}(\psi_n^{-1}\|\widehat{f_n} - f\| \geq \alpha) \\ &= \alpha^2 \mathbb{P}(\|\widehat{f_n} - f\| \geq s)\end{aligned}$$

where $s = s_n = A\psi_n$.

It suffices to find a lower bound on the *minimax probabilities*

$$\boxed{\inf_{\widehat{f_n}} \sup_{f \in \mathcal{F}} \mathbb{P}(\|\widehat{f_n} - f\| \geq s)}$$

# And now? An minimax estimator

**Step 2. Reduction to a finite number of hypotheses**

It suffices to try a finite number of hypothesis.

$$\boxed{\inf_{\widehat{f_n}} \sup_{f \in \mathcal{F}} \mathbb{P}(\|\widehat{f_n} - f\| \geq s) \geq \inf_{\widehat{f_n}} \sup_{f \in \{f_0, \cdots, f_m\}} \mathbb{P}(\|\widehat{f_n} - f\| \geq s)}$$

# And now ? An minimax estimator

**Step 3. Choice of $2s$-separated hypotheses**

Assume that

$$\boxed{\|f_j - f_k\| \geq 2s}\ ,\quad \forall k, j : k \neq j.$$

Then for any estimator $\widehat{f_n}$

$$\mathbb{P}_{f_j}(\|\widehat{f_n} - f_j\| \geq s) \geq \mathbb{P}_{f_j}(\psi^* \neq j) \tag{1}$$

where $\psi : \mathcal{X} \to \{0, 1, \ldots, M\}$ is the *minimum distance test* defined by

$$\psi^* = \arg\min_{0 \leq k \leq M} (\|\widehat{f_n} - f_k\|).$$

# And now? An minimax estimator

**Step 3. Choice of $2s$-separated hypotheses**

Assume that

$$\boxed{\|f_j - f_k\| \geq 2s}, \quad \forall k, j : k \neq j.$$

Then for any estimator $\widehat{f_n}$

$$\mathbb{P}_{f_j}(\|\widehat{f_n} - f_j\| \geq s) \geq \mathbb{P}_{f_j}(\psi^* \neq j) \tag{1}$$

where $\psi : \mathcal{X} \to \{0, 1, \ldots, M\}$ is the *minimum distance test* defined by

$$\psi^* = \arg \min_{0 \leq k \leq M} (\|\widehat{f_n} - f_k\|).$$

It follows that

$$\boxed{\inf_{\widehat{f_n}} \sup_{f \in \mathcal{F}} \mathbb{P}(\|\widehat{f_n} - f\| \geq s) \geq p_{e,M} := \inf_{\psi} \max_{0 \leq j \leq M} \mathbb{P}_j(\psi \neq j)}$$

# And now? An minimax estimator

Goal : It suffices to obtain $c$ such that

$$\boxed{p_{e,M} := \inf_{\psi} \max_{0 \leq j \leq M} \mathbb{P}_j(\psi \neq j) \geq c}$$

# And now ? An minimax estimator

## Theorem (Tsybakov)

*Let* $f_0, \ldots, f_M$ *in* $\mathcal{F}$*, for some* $M \geq 1$ *such that*

1. $\|f_j - f_k\| \geq 2s$*, for all* $0 \leq j < k \leq M$*;*
2. $\mathbb{P}_j \ll \mathbb{P}_0$*,* $\forall j = 0, 1, \ldots, M$*, and*

$$\boxed{\frac{1}{M}\sum_{j=1}^{M} KL(\mathbb{P}_j^{\otimes n}, \mathbb{P}_0^{\otimes n}) \leq \alpha \log M} \quad or \quad \boxed{\sum_{j=1}^{M} \chi^2(\mathbb{P}_j^{\otimes n}, \mathbb{P}_0^{\otimes n}) \leq \alpha M}.$$

*with* $0 < \alpha < 1/8$ *and* $\mathbb{P}_j = \mathbb{P}_{f_j}$*,* $j = 0, 1, \ldots, M$*.*

*Then, for* $\psi = s/A$*, we have*

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}\left[\psi_n^{-2} \|\widehat{f_n} - f\|^2\right] \geq c(\alpha) A^2 .$$

# And now ? An minimax estimator

**Work in progress** : Is my estimator minimax ?

# Perspective

- Apply the numerical methods on experimental data.
- Is this estimator minimax ?
  *i.e* Is it the best estimator among all possible estimators ?
- Can this estimation be done through PDEs ?

# References I

Lydia Robert, Jean Ollion, Jérôme Robert, Xiaohu Song, Ivan Matic, and Marina Elez, *Mutation dynamics and fitness effects followed in single cells*, Science **359** (2018), no. 6381, 1283–1286.