# Calibration bayésienne par divergence KL discrète et distance de Wasserstein

*for Sorbonne Université*

Eyal Cohen[1], Eline Pot[1], Saïd Raïssi[1]
[1] LPSM

*Supervisor(s):*

Nicolas Bousquet[1]

## Table of Contents

Un modèle bayésien est la donnée, pour une variable aléatoire $X$ d'une loi conditionnelle et d'une loi *a priori*

$$X \sim f(X|\theta) \quad \theta \sim \pi$$

On cherche à représenter le comportement d'une variable aléatoire $X$ définie sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$, de densité supposée :

$$f(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta \tag{1}$$

avec :

- $x \mapsto f(x|\theta)$ est une mesure de probabilité paramétrique connue
- $\Theta$ est un espace probabilisé

$\rightarrow$ Quel choix de la loi *a priori* $\pi$ et de sa calibration ?

## Table of Contents

- Paramétrer $\pi$ avec des hyperparamètres $\xi$.
- Contraintes quantiles : $\mathbb{P}(X \leqslant x_i) = \alpha_i$

$$\pi(\theta) = \pi(\theta|\xi) \quad \text{avec } \xi \in \Xi \text{ un ensemble d'hyperparamètres dans } \Xi \text{ supposé de dimension fini}$$

Loi a priori prédicitve : $f(x) = f(x|\xi)$, de fonction de répartition $\mathbb{P}(X \leqslant x_i) = \mathbb{P}(X \leqslant x_i|\xi)$.

$\rightarrow$ Comment définir une règle de calibration de $\xi$ en fonction de nos contraintes ?
En introduisant les $\alpha_i(\xi) = \mathbb{P}(X \leqslant x_i|\xi)$, on cherche alors à résoudre le problème suivant :

$$\min_{\xi \in \Xi} \mathcal{D}(\alpha_i - \alpha_i(\xi) \quad \text{avec } \mathcal{D} \text{ une distance à définir}$$

On étudiera 3 types de perte :

- La perte $L2$
- La distance KL
- La distance de Wasserstein

# Cas de la perte $L2$



L'idée la plus intuitive est d'utiliser une perte quadratique, pondérée par des poids $(w_i)_{i=1}^{n}$ :

$$\min_{\xi \in \Xi} \sum_{i=1}^{p} w_i(\alpha_i(\xi) - \alpha_i)^2$$

**Definition**

Pour deux probabilités discrètes $P$ et $Q$ définies sur un ensemble $\mathcal{X}$, la divergence de Kullback-Leibler est définie par :

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

- Dans notre cas,la perte KL s'écrit :

$$\min_{\xi \in \Xi} \sum_{i=0}^{p} (\alpha_{i+1} - \alpha_i) \log\left(\frac{\alpha_{i+1} - \alpha_i}{\alpha_{i+1}(\xi) - \alpha_i(\xi)}\right)$$

où $\alpha_0(\xi) = \alpha_0 = 0$ et $\alpha_{p+1}(\xi) = \alpha_{p+1} = 1$

- Avec $p = 2$, le problème se réécrit :

$$\min_{\xi \in \Xi} \alpha_1 \log\left(\frac{\alpha_1}{\alpha_1(\xi)}\right) + (\alpha_2 - \alpha_1) \log\left(\frac{\alpha_2 - \alpha_1}{\alpha_2(\xi) - \alpha_1(\xi)}\right) + (1 - \alpha_2) \log\left(\frac{1 - \alpha_2}{1 - \alpha_2(\xi)}\right)$$

- Approche de résolution numérique

# Table of Contents

### Definition (Transport Optimal)

La relaxation de Kantorovitch du problème de Monge, donne une formulation du problème de transport optimal comme suit:

$$L_C(a,b) = \min_{\mathbf{P}\in\mathbf{U}(a,b)} \langle \mathbf{C},\mathbf{P}\rangle = \sum_{i,j} \mathbf{C}_{i,j}\mathbf{P}_{i,j}$$

$$\mathcal{L}_C(\alpha,\beta) = \inf_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y)$$

Where $\mathbf{U}(a,b) = \{\mathbf{P}\in\mathbb{R}_+^{n\times m} : \mathbf{P}1_m = a \text{ and } \mathbf{P}^T 1_n = b\}$ and $\mathcal{U}(\alpha,\beta) = \{\pi\in\mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y}) : P_{\mathcal{X}\#}\pi = \alpha$ and $P_{\mathcal{Y}\#}\pi = \beta\}$.

## Definition (Transport Optimal)

La relaxation de Kantorovitch du problème de Monge, donne une formulation du problème de transport optimal comme suit:

$$L_C(a,b) = \min_{\mathbf{P}\in\mathbf{U}(a,b)} \langle \mathbf{C}, \mathbf{P} \rangle = \sum_{i,j} \mathbf{C}_{i,j}\mathbf{P}_{i,j}$$

$$\mathcal{L}_C(\alpha,\beta) = \inf_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y)$$

Where $\mathbf{U}(a,b) = \{\mathbf{P} \in \mathbb{R}_+^{n\times m} : \mathbf{P}\mathbb{1}_m = a \text{ and } \mathbf{P}^T\mathbb{1}_n = b\}$ and $\mathcal{U}(\alpha,\beta) = \{\pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y}) : P_{\mathcal{X}\#}\pi = \alpha$ and $P_{\mathcal{Y}\#}\pi = \beta\}$.

## Definition

La distance de Wasserstein découle assez naturellement de ce problème:

$$W_p(a,b) = (L_{D^p}(a,b))^{\frac{1}{p}}$$

$$\mathcal{W}_p(\alpha,\beta) = (_{d^p}(\alpha,\beta))^{\frac{1}{p}}$$

Nous considérons un transport optimal avec pénalisation entropique.
On peut écrire: $P = diag(u)Kdiag(v)$ avec $K$ le noyau de Gibbs $exp(-C/\epsilon)$.
This leads to an alternating optimization algorithm:

**Algorithm** Sinkhorn Algorithm

**Require:** $\alpha, \beta$ distributions, $C$ une matrice de coup, $\epsilon$ le facteur de régularisation entropique, $T$ le nombre d'iterations maximal.

1: **return** $P$ un plan de transport optimal entre $\alpha$ et $\beta$
2: $u_0 \leftarrow ones(len(\alpha))$
3: $v_0 \leftarrow ones(len(\beta))$
4: $K \leftarrow \exp\left(-\frac{C}{\epsilon}\right)$
5: **for** $i \in \{1 : T\}$ **do**
6: $\quad u_{i+1} \leftarrow \frac{\alpha}{Kv}$
7: $\quad v_{i+1} \leftarrow \frac{\beta}{K^T u}$
8: **end for**
9: $P = diag(u_T)Kdiag(v_T)$

On peut le réécrire comme:

$$P_{i+1} = prox^{KL}_{\frac{1}{\epsilon}}(P_i) = argmin_{P \in \mathbb{R}_+^{n \times m}} KL(P||P_l) + \frac{1}{\epsilon}\langle P, \pi \rangle + \iota_{U(a,b)}(P)$$

En voyant cette formulation, nous avons pensé (et pensons toujours un peu) que chercher un algorithme de calibration basé sur cet algorithme serait prometteur. Malheureusement par manque de temps nous avons opté pour la résolution numérique que nous présentons plus tard.

# Table of Contents

Pipeline :

- On se place dans le cas où $X|\theta$ appartient à la famille exponentielle
$$f(x|\theta) = \exp(t(\theta)x - \phi(\theta))h(x)$$

# Famille Exponentielles



Pipeline :

- On se place dans le cas où $X|\theta$ appartient à la famille exponentielle
$$f(x|\theta) = \exp(t(\theta)x - \phi(\theta))h(x)$$
- On calcule
$$I(\theta)|^{1/2}$$

Pipeline :

- On se place dans le cas où $X|\theta$ appartient à la famille exponentielle
$$f(x|\theta) = \exp(t(\theta)x - \phi(\theta))h(x)$$
- On calcule
$$I(\theta)|^{1/2}$$
- On formule les prior à une constante près en fonction des $\xi$
$$\pi(\theta|\alpha,\beta) \propto \exp(\alpha.\theta - \beta\phi(\theta))|I(\theta)|^{1/2}$$
Avec :
$$\alpha = \sum_{i=1}^{m} t(\tilde{x}_i)$$
et $m = \beta$

Pipeline :

- On se place dans le cas où $X|\theta$ appartient à la famille exponentielle
$$f(x|\theta) = \exp(t(\theta)x - \phi(\theta))h(x)$$
- On calcule
$$I(\theta)|^{1/2}$$
- On formule les prior à une constante près en fonction des $\xi$
$$\pi(\theta|\alpha,\beta) \propto \exp(\alpha.\theta - \beta\phi(\theta))|I(\theta)|^{1/2}$$
Avec :
$$\alpha = \sum_{i=1}^{m} t(\tilde{x}_i)$$
et $m = \beta$
- Identification d'une loi adéquate en fonction de la forme du proportionnel

Pipeline :

- On se place dans le cas où $X|\theta$ appartient à la famille exponentielle
$$f(x|\theta) = \exp(t(\theta)x - \phi(\theta))h(x)$$
- On calcule
$$I(\theta)|^{1/2}$$
- On formule les prior à une constante près en fonction des $\xi$
$$\pi(\theta|\alpha,\beta) \propto \exp(\alpha.\theta - \beta\phi(\theta))|I(\theta)|^{1/2}$$
Avec :
$$\alpha = \sum_{i=1}^{m} t(\tilde{x}_i)$$
et $m = \beta$
- Identification d'une loi adéquate en fonction de la forme du proportionnel
- Formulation de la fonction de répartition à priori prédictive $P(X \leqslant x_i) = \int_\Theta \left( \int_{-\infty}^{x_i} f(u|\theta)du \right) \pi_\xi(\theta)d\theta$

Pipeline :

- On se place dans le cas où $X|\theta$ appartient à la famille exponentielle
$$f(x|\theta) = \exp(t(\theta)x - \phi(\theta))h(x)$$
- On calcule
$$I(\theta)|^{1/2}$$
- On formule les prior à une constante près en fonction des $\xi$
$$\pi(\theta|\alpha,\beta) \propto \exp(\alpha.\theta - \beta\phi(\theta))|I(\theta)|^{1/2}$$
  Avec :
$$\alpha = \sum_{i=1}^{m} t(\tilde{x}_i)$$
  et $m = \beta$
- Identification d'une loi adéquate en fonction de la forme du proportionnel
- Formulation de la fonction de répartition à priori prédictive $P(X \leqslant x_i) = \int_\Theta \left(\int_{-\infty}^{x_i} f(u|\theta)du\right) \pi_\xi(\theta)d\theta$
- Optimisation sur $\xi$ sur les différentes distances $L2, KL, W2$

- On a

$$t(x) = x, \quad \Phi(\theta) = -\log(-\theta), \quad h(x) = 1$$

- On a

$$t(x) = x, \quad \Phi(\theta) = -\log(-\theta), \quad h(x) = 1$$

- Puis :

$$\alpha = \sum_{i=1}^{m} \tilde{x}_i \quad \beta = m$$

- On a

$$t(x) = x, \quad \Phi(\theta) = -\log(-\theta), \quad h(x) = 1$$

- Puis :

$$\alpha = \sum_{i=1}^{m} \tilde{x}_i \quad \beta = m$$

- On a identifié :

$$\pi(\theta|\alpha,\beta) \sim \mathcal{G}(\beta,\alpha)$$

- On a

$$t(x) = x, \quad \Phi(\theta) = -\log(-\theta), \quad h(x) = 1$$

- Puis :

$$\alpha = \sum_{i=1}^{m} \tilde{x}_i \quad \beta = m$$

- On a identifié :

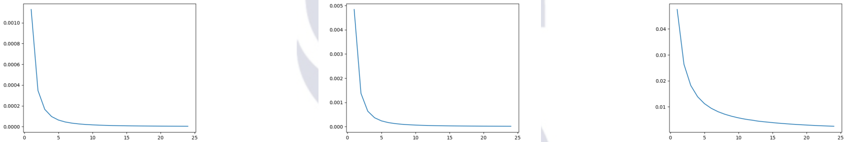
$$\pi(\theta|\alpha, \beta) \sim \mathcal{G}(\beta, \alpha)$$

- La fonction de répartition du prédictif à priori :

$$\begin{aligned} P(X \leqslant x_i) &= \int_{\Theta} \left( \int_{-\infty}^{x_i} f(u|\theta) du \right) \pi_{\xi}(\theta) d\theta \\ &= 1 - \frac{\beta^{\alpha}}{(\beta - x_i)^{\alpha}} \end{aligned} \tag{2}$$

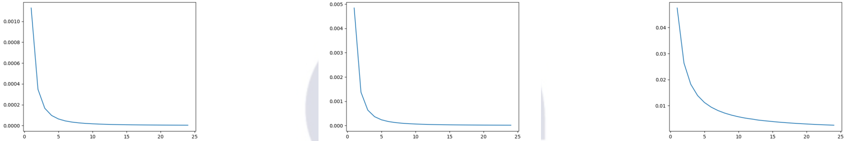
# Résultats pour la loi Exponentielle





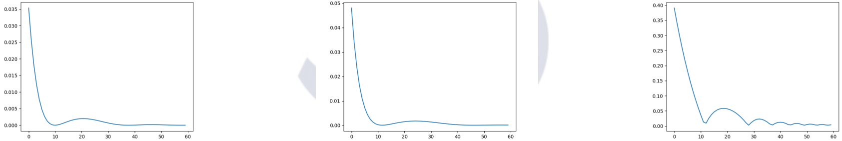

(a) L2 (b) KL (c) W2

Figure: 2 Quantiles - Loss w.r.t. m



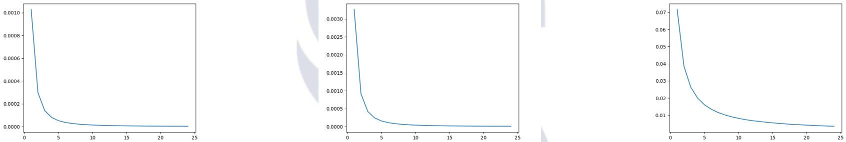

(a) L2 (b) KL (c) W2

Figure: 2 Quantiles - Loss w.r.t. m




(a) L2 (b) KL (c) W2

Figure: 2 Quantiles - Loss w.r.t. the $(\tilde{x}_i)_{i=1}^m$




(a) L2 (b) KL (c) W2

Figure: 5 Quantiles - Loss w.r.t. m

# Loi normale :



- On a $t(x) = \frac{x}{\sigma^2}, \quad \Phi(\theta) = \frac{\theta^2}{2\sigma^2}, \quad h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(x^2/(2\sigma^2))$

# Loi normale :



- On a $t(x) = \frac{x}{\sigma^2}, \quad \Phi(\theta) = \frac{\theta^2}{2\sigma^2}, \quad h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(x^2/(2\sigma^2))$
- On cherche à identifier $\mu$ avec $x|\theta \sim \mathcal{N}(\mu, \sigma^2)$

- On a $t(x) = \frac{x}{\sigma^2}, \quad \Phi(\theta) = \frac{\theta^2}{2\sigma^2}, \quad h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(x^2/(2\sigma^2))$
- On cherche à identifier $\mu$ avec $x|\theta \sim \mathcal{N}(\mu, \sigma^2)$
- On a identifié :

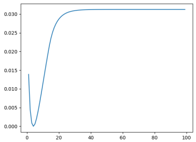
$$\pi(\theta|\alpha,\beta) \sim \mathcal{N}(\frac{\alpha}{\beta}, \frac{4}{\beta})$$

- On a $t(x) = \frac{x}{\sigma^2}$, $\quad \Phi(\theta) = \frac{\theta^2}{2\sigma^2}$, $\quad h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(x^2/(2\sigma^2))$
- On cherche à identifier $\mu$ avec $x|\theta \sim \mathcal{N}(\mu, \sigma^2)$
- On a identifié :

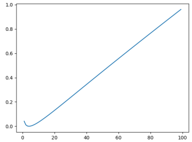
$$\pi(\theta|\alpha,\beta) \sim \mathcal{N}(\frac{\alpha}{\beta}, \frac{4}{\beta})$$

- La fonction de répartition du prédictif à priori :

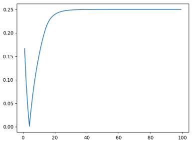
$$\begin{aligned} P(X \leqslant x_i) &= \int_\Theta \left( \int_{-\infty}^{x_i} f(u|\theta)du \right) \pi_\xi(\theta)d\theta \\ &= 0.5 \times \frac{1 + erf(x - \frac{\alpha}{\beta})}{\frac{4}{\beta} \times \sqrt{0.5}} \end{aligned} \tag{3}$$




(a) L2




(b) KL

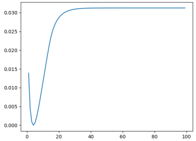



(c) W2

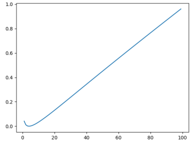
Figure: 2 Quantiles - Loss w.r.t. m

# Résultats pour la loi Normale

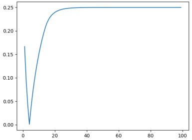





(a) L2

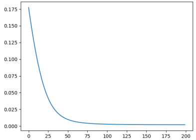



(b) KL

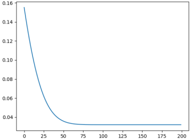



(c) W2

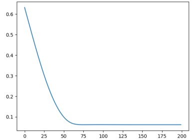
Figure: 2 Quantiles - Loss w.r.t. m




(a) L2




(b) KL




(c) W2

Figure: 2 Quantiles - Loss w.r.t. the $(\tilde{x}_i)_{i=1}^{m}$

# Table of Contents

On se place dans un cadre où notre loi prédictive est une loi de Gumbel.

On se place dans un cadre où notre loi prédictive est une loi de Gumbel.
Sa fonction de répartition s'écrit :

$$\mathbb{P}(X < x|\theta) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right) \quad \text{avec } \theta = (\sigma, \mu) \quad \sigma > 0, \mu \in \mathbb{R}, x \in \mathbb{R}$$

# Prior de Gumbel



On se place dans un cadre où notre loi prédictive est une loi de Gumbel.
Sa fonction de répartition s'écrit :

$$\mathbb{P}(X < x|\theta) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right) \quad \text{avec } \theta = (\sigma, \mu) \quad \sigma > 0, \mu \in \mathbb{R}, x \in \mathbb{R}$$

On considère la mesure *a priori* :

$$\pi(\theta) = \pi(\mu, \sigma) \quad \propto \quad \sigma^{-m} \exp\left(m\frac{\mu - \bar{\tilde{x}}_m}{\sigma} - \sum_{i=1}^{n} \exp\left(-\frac{\tilde{x}_i - \mu}{\sigma}\right)\right)$$

# Prior de Gumbel



On se place dans un cadre où notre loi prédictive est une loi de Gumbel.
Sa fonction de répartition s'écrit :

$$\mathbb{P}(X < x|\theta) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right) \quad \text{avec } \theta = (\sigma, \mu) \quad \sigma > 0, \mu \in \mathbb{R}, x \in \mathbb{R}$$

On considère la mesure *a priori* :

$$\pi(\theta) = \pi(\mu, \sigma) \quad \propto \quad \sigma^{-m} \exp\left(m\frac{\mu - \bar{\tilde{x}}_m}{\sigma} - \sum_{i=1}^{n} \exp\left(-\frac{\tilde{x}_i - \mu}{\sigma}\right)\right)$$

Ici, nos hyperparamètres sont : $\xi = (\lambda, m, \tilde{x}_1, \ldots, \tilde{x}_m)$. Le $m$ et les $(\tilde{x}_i)_{i=1}^{m}$ correspondent à la taille et aux données d'un échantillon virtuel.

# Prior de Gumbel



On se place dans un cadre où notre loi prédictive est une loi de Gumbel.
Sa fonction de répartition s'écrit :

$$\mathbb{P}(X < x|\theta) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right) \quad \text{avec } \theta = (\sigma, \mu) \quad \sigma > 0, \mu \in \mathbb{R}, x \in \mathbb{R}$$

On considère la mesure *a priori* :

$$\pi(\theta) = \pi(\mu, \sigma) \quad \propto \quad \sigma^{-m} \exp\left(m\frac{\mu - \bar{\tilde{x}}_m}{\sigma} - \sum_{i=1}^{n} \exp\left(-\frac{\tilde{x}_i - \mu}{\sigma}\right)\right)$$

Ici, nos hyperparamètres sont : $\xi = (\lambda, m, \tilde{x}_1, \ldots, \tilde{x}_m)$. Le $m$ et les $(\tilde{x}_i)_{i=1}^{m}$ correspondent à la taille et aux données d'un échantillon virtuel.

$\rightarrow$ Comment simuler la loi prédictive prédictive de $X$ en fonction des hyperparamètres $\xi$, afin de généner des quantiles?

La démarche ici est de simuler des $\mu, \sigma$ par Importance Sampling, puis de simuler $X$ avec une loi Gumbel de paramètres $(\mu, \sigma)$.

La démarche ici est de simuler des $\mu, \sigma$ par Importance Sampling, puis de simuler $X$ avec une loi Gumbel de paramètres $(\mu, \sigma)$. Pour l'Importance Sampling, on utilise comme lois instrumentales :

$$\mathcal{IG}\left(m-1, \sum_{i=1}^{m} \tilde{x}_i\right) \text{ pour } \sigma$$

et

$$\mathcal{E}_\mu(\lambda) \text{ pour } \mu$$

avec des poids d'importance de la forme:

$$w \quad \propto \quad \exp\left(\mu\left(\frac{m}{\sigma}+\lambda\right) - \sum_{i=1}^{m} \exp\left(-\frac{\tilde{x}-\mu}{\sigma}\right)\right)$$

La démarche ici est de simuler des $\mu, \sigma$ par Importance Sampling, puis de simuler $X$ avec une loi Gumbel de paramètres $(\mu, \sigma)$. Pour l'Importance Sampling, on utilise comme lois instrumentales :

$$\mathcal{IG}\left(m-1, \sum_{i=1}^{m} \tilde{x}_i\right) \text{ pour } \sigma$$

et

$$\mathcal{E}_\mu(\lambda) \text{ pour } \mu$$

avec des poids d'importance de la forme:

$$w \quad \propto \quad \exp\left(\mu\left(\frac{m}{\sigma}+\lambda\right)-\sum_{i=1}^{m}\exp\left(-\frac{\tilde{x}-\mu}{\sigma}\right)\right)$$

Pour trouver les meilleurs hyperparamètres, on procède par grid-search sur $m, \tilde{x}$ et $\lambda$.

| | **Loss L2** | **Loss KL** | **Loss W2** |
|---|---|---|---|
| $m$ | 3 | 1 | 1 |
| $\tilde{x}$ | [1, 2, 0.5] | 1.5 | 2 |
| $\lambda$ | 1 | 1 | 0.5 |
| *Loss* | **0.0226** | 0.0465 | 0.1701 |

Table: Meilleures loss trouvée pour divers essais de $\lambda$, $m$ et de $\tilde{x}$

# Calibration bayésienne par divergence KL discrète et distance de Wasserstein

Thank you for listening !
Any Questions ?