

INTRODUCTION À LA CALIBRATION  
RECENT ADVANCES IN MACHINE LEARNING

**Théo Lopès-Quintas**

BPCE Payment Services,  
Université Paris Dauphine

2023 - 2025

<b>1</b>	<b>Introduction . . . . .</b>	<b>1</b>
1.1	Confiance . . . . .	1
1.2	Calibration . . . . .	2
<b>2</b>	<b>Comment mesurer la distance à la calibration? . . . . .</b>	<b>6</b>
2.1	Vrai distance à la calibration . . . . .	6
2.2	Mesure de calibration consistante . . . . .	10
2.3	Plus petite distance à la calibration . . . . .	13
2.4	Smooth Calibration . . . . .	15
<b>3</b>	<b>Comment induire la calibration? . . . . .</b>	<b>21</b>
3.1	Cas particulier de la <i>Squared loss</i> . . . . .	23
3.2	Généralisation du lien entre pGap et smCE : par la dualité . . . . .	25

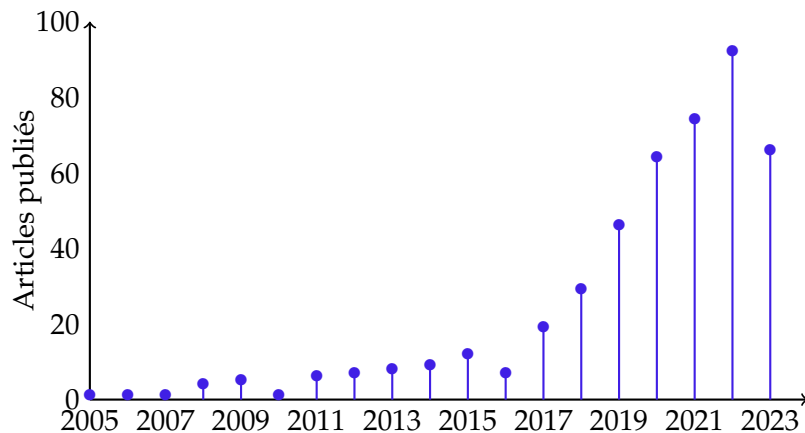
# INTRODUCTION

## CONFIANCE

- ▶ Joanna Bryson, *No One Should Trust Artificial Intelligence* [Bryson, 2018]
- ▶ Aurélie Jean, *Les algorithmes font-ils la loi ?* [Jean, 2021]
- ▶ Cathy O'Neil, *Weapons of Math Destruction* [O'Neil, 2016]

# INTRODUCTION

## CALIBRATION



**Figure** – Nombres d’articles Arxiv contenant le terme *calibration* dans le titre pour les catégories *cs.AI*, *cs.CL*, *cs.LG*, *math.ST*, *stat.AP*, *stat.CO*, *stat.ML* calculés en août 2023

# INTRODUCTION

## CALIBRATION

Domaine discret, espace des inputs

$(x, y) \sim \mathcal{D}$  avec  $\mathcal{D}$  une distribution sur  $\mathcal{X} \times \{0, 1\}$

On demande que  $\mathcal{X}$  soit discret, éventuellement très grand, pour éviter des difficultés théoriques<sup>1</sup>. Cette hypothèse est toujours vérifiée dans la pratique.

On appelle un *prédicteur* une fonction  $f : \mathcal{X} \rightarrow [0, 1]$  que l'on interprète comme une estimation de  $\mathbb{P}(y = 1|x)$ . C'est le résultat de l'entraînement de l'algorithme dont on souhaite mesurer la calibration. On considère la distribution induite par la paire prédiction-label  $(f(x), y) \in [0, 1] \times \{0, 1\}$  que l'on notera  $\mathcal{D}_f$ .

---

1. Les auteurs précisent que supposer que  $f$  soit mesurable devrait suffire pour traiter le cas où  $\mathcal{X}$  est infini.

# INTRODUCTION

## CALIBRATION

### Définition 1 (Parfaite calibration [Dawid, 1982])

*On dit qu'une distribution  $\Gamma$  sur  $[0, 1] \times \{0, 1\}$  est parfaitement calibrée si, et seulement si, on a :*

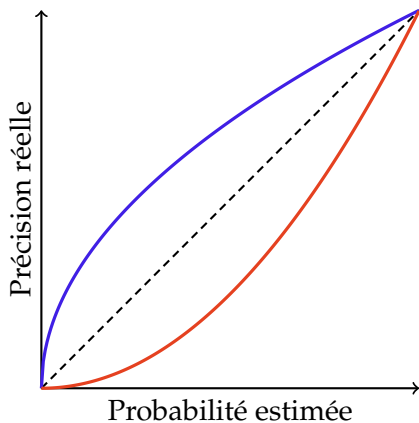
$$\mathbb{E}_{(v,y) \sim \Gamma} [y|v] = v$$

*On dit qu'un prédicteur  $f$  est parfaitement calibré par rapport à  $\mathcal{D}$  si  $\mathcal{D}_f$  est parfaitement calibrée. On note  $\text{cal}(\mathcal{D})$  l'ensemble des prédicteurs parfaitement calibrés par rapport à  $\mathcal{D}$ .*

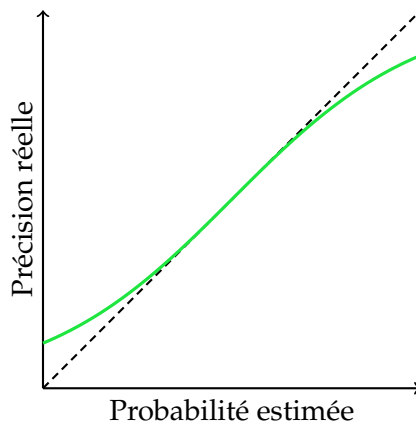
Remarquons que l'on peut également définir la parfaite calibration comme l'impossibilité de distinguer  $(f(x), y)$  et  $(f(x), y')$  où  $y' \sim \text{Ber}(f(x))$  quand on ne regarde que les prédictions !

# INTRODUCTION

## CALIBRATION



(a) Prédicteur **trop confiant** et **pas assez confiant**



(b) Prédicteur **non parfaitement calibré**

**Figure** – Comparaison de prédicteurs par rapport à un prédicteur parfaitement calibré (en pointillés)

# COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

## VRAI DISTANCE À LA CALIBRATION

Notre objectif est d'être capable de mesurer la distance à la calibration sans avoir à le faire visuellement. Nous serons donc amené à comparer les prédicteurs possibles entre eux. On considère la distance  $\ell_1$  sur cet ensemble :

Ensemble des prédicteurs  $f : \mathcal{X} \rightarrow [0, 1]$

$$\forall f, g \in \mathcal{F}_{\mathcal{X}}, \quad \ell_1(f, g) = \mathbb{E}_{\mathcal{D}} [|f(x) - g(x)|]$$

Notons que dans la définition précédente, nous calculons l'espérance par rapport à la distribution  $\mathcal{D}$  des observations  $(x, y)$  et non les distributions induites par  $f$  ou  $g$ . Si  $f$  et  $g$  sont distant alors  $D_f$  et  $D_g$  le sont aussi ?



# COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

## VRAI DISTANCE À LA CALIBRATION

### Exercice 1

Soit  $\mathcal{X} = \{0, 1\}$  et on considère une distribution uniforme sur  $\mathcal{X}$ .

1. Quel est le meilleur prédicteur que l'on puisse construire ?
2. On considère  $f(x) = x$  et  $g(x) = 1 - x$  comme prédicteurs. Que dire de  $D_f$  et  $D_g$  ?
3. Calculer  $\ell_1(f, g)$ . Commentez.

La distance ainsi définie permet de vraiment cibler la qualité d'un prédicteur, et non la distribution induite, bien que la définition de calibration se fasse par celle-ci. Nous avons donc un moyen bien défini pour comparer deux prédicteurs. Ce qui veut dire que nous avons une manière de comparer tous les prédicteurs entre eux.

# COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

## VRAI DISTANCE À LA CALIBRATION

### Définition 2 (Vrai distance à la calibration)

Pour une distribution  $\mathcal{D}$  donnée et une famille  $\mathcal{F}_{\mathcal{X}}$  donnée, on définit la vrai distance à la calibration comme :

$$\text{dCE}_{\mathcal{D}}(f) = \inf_{g \in \text{cal}(\mathcal{D})} \ell_1(f, g)$$

Cette distance serait une bonne candidate pour être une mesure de la distance à la calibration.

### Proposition 1

Soit  $\alpha \in [0, \frac{1}{2}]$ . Il existe un domaine  $\mathcal{X}$ , un prédicteur  $f \in \mathcal{F}_{\mathcal{X}}$  et deux distributions  $\mathcal{D}_f^1$  et  $\mathcal{D}_f^2$  telles que :

- ▶ Les distributions  $\mathcal{D}_f^1$  et  $\mathcal{D}_f^2$  sont identiques
- ▶  $\text{dCE}_{\mathcal{D}_f^1}(f) \leq 2\alpha^2$  tandis que  $\text{dCE}_{\mathcal{D}_f^2}(f) \geq \alpha$

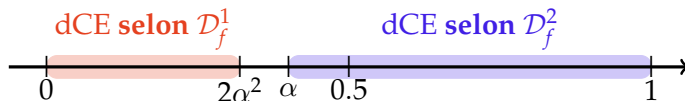


Figure – Inégalités de la proposition (1)

# COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

## RÉSUMÉ

On a accès à la paire prédiction-label  $(f(x), y) \in [0, 1] \times \{0, 1\}$  que l'on suppose issue d'une distribution.

1. On dit qu'un prédicteur  $f$  est **parfaitement calibré** si  $\mathbb{E}_{(v,y) \sim \Gamma} [y|v] = v$ .
2. On **compare** deux prédicteurs avec la **distance**  $\ell_1$  définie pour deux prédicteurs  $f$  et  $g$  :  
$$\ell_1(f, g) = \mathbb{E}_{\mathcal{D}} [|f(x) - g(x)|]$$
3. La **vraie distance à la calibration** est définie comme la plus petite distance  $\ell_1$  entre le prédicteur d'intérêt et un prédicteur calibré.

Cependant, la vraie distance à la calibration ne peut pas être calculé dans le cadre que l'on se fixe<sup>2</sup>.

---

2. N'avoir accès qu'à la paire prédiction-label.

# COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

## MESURE DE CALIBRATION CONSISTANTE

On attend d'une mesure de calibration qu'elle renvoie un nombre entre 0 et 1 à partir d'une distribution  $\mathcal{D}$  d'observations et d'un prédicteur  $f$ . On note cette valeur  $\mu_{\mathcal{D}}(f)$ . Au minimum, on souhaite avoir les propriétés suivantes :

$$\begin{aligned}\mu_{\mathcal{D}}(f) &= 0 && \text{si } f \in \text{cal}(\mathcal{D}) && \text{(Complétude)} \\ \mu_{\mathcal{D}}(f) &> 0 && \text{si } f \notin \text{cal}(\mathcal{D}) && \text{(Correction)}\end{aligned}$$

On préférerait que  $\mu_{\mathcal{D}}(f)$  soit petite si on est proche de la calibration parfaite et grande si on en est loin.

### Définition 3 (Mesure consistante de calibration)

Soit  $c \in \mathbb{R}_+$ , on dit qu'une mesure de calibration  $\mu$  satisfait la *c-robuste complétude* si :

$$\exists a \in \mathbb{R}_+, \forall \mathcal{D} \sim [0, 1] \times \{0, 1\}, \forall f \in \mathcal{F}_{\mathcal{X}}, \quad \mu_{\mathcal{D}}(f) \leq a (\text{dCE}_{\mathcal{D}}(f))^c$$

Soit  $s \in \mathbb{R}_+$ , on dit qu'une mesure de calibration  $\mu$  satisfait la *s-robuste correction* si :

$$\exists b \in \mathbb{R}_+, \forall \mathcal{D} \sim [0, 1] \times \{0, 1\}, \forall f \in \mathcal{F}_{\mathcal{X}}, \quad \mu_{\mathcal{D}}(f) \geq b (\text{dCE}_{\mathcal{D}}(f))^s$$

On dit que  $\mu$  est *(c, s)-consistante* si les deux conditions précédentes sont vérifiées.

# COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

## MESURE DE CALIBRATION CONSISTANTE

Dans la définition de consistance, nous n'avons pas pris en compte que la mesure de calibration puisse être calculable en ayant accès uniquement aux prédictions et aux labels. Voyons ce que cela impose comme conditions.

### Corollaire 1

Soit  $\mu$  une mesure de calibration  $(c, s)$ -consistante que l'on peut calculer à l'aide des prédictions et des labels. Alors  $s \geq 2c$

### Exercice 2 (Preuve)

Soit un domaine  $\mathcal{X}$ , un prédicteur  $f \in \mathcal{F}_{\mathcal{X}}$  et deux distributions identiques  $\mathcal{D}_f^1$  et  $\mathcal{D}_f^2$ . Soit  $\mu$  telle que définie par le corollaire.

1. Que peut-on dire de  $\mu_{\mathcal{D}^1}(f)$  et  $\mu_{\mathcal{D}^2}(f)$  ?
2. En utilisant la définition de la  $(c, s)$ -consistante, aboutir à une contradiction.

## COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

### EXPECTED CALIBRATION ERROR

Nous avons restreint les mesures de calibration éligibles, il nous faut maintenant en exhiber une !  
Commençons par la manière la plus intuitive de mesurer la distance à la calibration.

#### Définition 4 (Expected Calibration Error)

Soit  $\mathcal{D}$  une distribution sur  $[0, 1] \times \{0, 1\}$  et  $f : \mathcal{X} \rightarrow [0, 1]$  un prédicteur. On appelle Expected Calibration Error :

$$ECE_{\mathcal{D}}(f) = \mathbb{E}_{\mathcal{D}} \left[ \left| \mathbb{E}_{\mathcal{D}} [y \mid f(x)] - f(x) \right| \right]$$

Voyons si l'ECE est consistante.

#### Exercice 3

Soit  $f$  un prédicteur et  $\mathcal{S} = \text{level}(f)$ .

1. Montrer que  $ECE(f) = \ell_1(f, g_{\mathcal{S}})$  et en déduire que l'ECE est 1-robuste.
2. On considère une distribution uniforme sur  $\mathcal{X} = \{0, 1\}$  où le meilleur prédicteur est  $f^*(0) = 0$  et  $f^*(1) = 1$ . On considère le prédicteur  $f_{\varepsilon}$  avec  $\varepsilon \geq 0$  défini par  $f_{\varepsilon}(0) = \frac{1}{2} - \varepsilon$  et  $f_{\varepsilon}(1) = \frac{1}{2} + \varepsilon$ . Calculer  $ECE_{\mathcal{D}}(f_{\varepsilon})$  et montrer que l'ECE ne vérifie pas la robuste correction.

# COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

## PLUS PETITE DISTANCE À LA CALIBRATION

Revenons sur dCE. Nous ne pouvons pas l'exploiter dans notre cas. Pour obtenir des garanties sur les estimations que nous allons faire, définissons la distance inférieure à la calibration<sup>3</sup>.

### Définition 5

Soit  $\Gamma$  une distribution sur  $[0, 1] \times \{0, 1\}$ . On définit l'ensemble  $\text{ext}(\Gamma)$  comme l'ensemble des distributions  $\Pi$  des triplets  $(u, v, y) \in [0, 1] \times [0, 1] \times \{0, 1\}$  tel que :

- ▶ La distribution marginale de  $(v, y)$  est  $\Gamma$
- ▶ La distribution marginale de  $(u, y)$  est parfaitement calibré i.e.  $\mathbb{E}_{\Pi} [y|u] = u$ .

### Définition 6 (Distance inférieure à la calibration)

On définit la distance inférieure à la calibration par :

$$\underline{\text{dCE}}(\Gamma) = \inf_{\Pi \in \text{ext}(\Gamma)} \mathbb{E}_{(u,v,y) \sim \Pi} [|u - v|]$$

Pour une distribution  $\mathcal{D}$  et un prédicteur  $f$ , on définit  $\underline{\text{dCE}}_{\mathcal{D}}(f) = \underline{\text{dCE}}(\mathcal{D}_f)$

---

3. Dans l'article est également défini la distance supérieure à la calibration, mais nous n'en avons pas besoin dans le cadre de ce cours.

# COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

## PLUS PETITE DISTANCE À LA CALIBRATION

Dans l'article [Błasiok et al., 2023a], via des raisonnements que nous ne traiterons pas ici, est obtenue l'inégalité suivante :

### Proposition 2

*Pour toute distribution  $\mathcal{D}$  et tout prédicteur  $f$ , on a :*

$$\underline{\text{dCE}}_{\mathcal{D}}(f) \leq \text{dCE}_{\mathcal{D}}(f) \leq 4\sqrt{\underline{\text{dCE}}_{\mathcal{D}}(f)}$$

Cet encadrement ne permet pas une estimation précise de dCE mais informe que l'écart est au plus quadratique entre dCE et  $\underline{\text{dCE}}$ .



# COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

## SMOOTH CALIBRATION

Une alternative proposée à l'ECE est la smooth-calibration. Cette notion est un cas particulier d'une famille plus générale introduite en 2022 dans l'article *Low-degree multicalibration* [Gopalan et al., 2022].

### Définition 7 (Calibration pondérée)

Soit  $W$  une famille de fonctions  $w : [0, 1] \rightarrow \mathbb{R}$ . La calibration pondérée d'une distribution  $\Gamma$  sur  $[0, 1] \times \{0, 1\}$  est définie comme :

$$\text{wCE}^W(\Gamma) = \sup_{w \in W} \left| \mathbb{E}_{(v,y) \sim \Gamma} [(y - v)w(v)] \right|$$

Étant donné une distribution  $\mathcal{D}$  sur  $\mathcal{X} \times \{0, 1\}$  et un prédicteur  $f : \mathcal{X} \rightarrow [0, 1]$ , on appelle la calibration pondérée de  $f$  par rapport à  $\mathcal{D}$  :

$$\text{wCE}_{\mathcal{D}}^W(f) = \text{wCE}^W(\mathcal{D}_f) = \sup_{w \in W} \left| \mathbb{E}_{(x,y) \sim \mathcal{D}} [(y - f(x))w(f(x))] \right|$$

# COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

## SMOOTH CALIBRATION

### Exercice 4

*Montrer que la calibration pondérée vérifie la complétude.*

# COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

## SMOOTH CALIBRATION

### Exercice 4

*Montrer que la calibration pondérée vérifie la complétude.*

On considère une distribution  $\Gamma$  définie sur  $[0, 1] \times \{0, 1\}$  que l'on suppose parfaitement calibrée. Par définition,  $\mathbb{E}[y \mid v] = v$ . Exploitions cette information.

$$\begin{aligned}\mathbb{E}_{\Gamma}[(y - v)w(v)] &= \mathbb{E}_{\Gamma}\left[\mathbb{E}[(y - v)w(v) \mid v]\right] \quad \text{en conditionnant par } v \\ &= \mathbb{E}_{\Gamma}\left[\mathbb{E}[y \mid v]w(v) - v(w)\right] \\ &= 0 \quad \text{parce que } \Gamma \text{ est parfaitement calibrée}\end{aligned}$$

# COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

## SMOOTH CALIBRATION

Une famille importante est celle des fonctions 1-Lipschitzienne bornées. Elle est pour la première fois introduite dans ce contexte dans l'article [Kakade and Foster, 2004]. Nous suivrons cependant les notations de [Gopalan et al., 2022].

### Définition 8 (Smooth calibration)

Soit  $L$  la famille des fonctions 1-Lipschitzienne  $w : [0, 1] \rightarrow [-1, 1]$ . L'erreur de smooth calibration d'une distribution  $\Gamma$  sur  $[0, 1] \times \{0, 1\}$  est définie comme :

$$\text{smCE}(\Gamma) = \text{wCE}^L(\Gamma)$$

De manière similaire, pour une distribution  $\mathcal{D}$  et un prédicteur  $f$  on définit :

$$\text{smCE}_{\mathcal{D}}(f) = \text{smCE}(\mathcal{D}_f) = \text{wCE}^L(\mathcal{D}_f)$$

Nous avons déjà montré que la calibration pondérée en général vérifie la complétude. Il reste à identifier avec quel coefficient et savoir si elle vérifie la correction robuste.

# COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

## SMOOTH CALIBRATION

Pour cela, lions la smooth-calibration à la plus petite distance à la calibration.

### Théorème 1

Soit une distribution  $\Gamma$  sur  $[0, 1] \times \{0, 1\}$ . Alors,

$$\frac{1}{2} \underline{\text{dCE}}(\Gamma) \leq \text{smCE}(\Gamma) \leq 2 \underline{\text{dCE}}(\Gamma)$$

Cela induit immédiatement le corollaire :

### Corollaire 2

La mesure de calibration  $\text{smCE}$  est  $(1, 2)$ -consistante.

# COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

## SMOOTH CALIBRATION

### Exercice 5

Démontrer le corollaire 2 :  $\text{smCE}$  est  $(1, 2)$ -consistante.

On rappelle les résultats obtenus jusqu'ici :

- ▶  $c$ -robuste complétion :  $\mu_{\mathcal{D}}(f) \leq a (\text{dCE}_{\mathcal{D}}(f))^c$
- ▶  $s$ -robuste correction :  $\mu_{\mathcal{D}}(f) \geq b (\text{dCE}_{\mathcal{D}}(f))^s$
- ▶ Proposition (2) :  $\underline{\text{dCE}}_{\mathcal{D}}(f) \leq \text{dCE}_{\mathcal{D}}(f) \leq 4\sqrt{\underline{\text{dCE}}_{\mathcal{D}}(f)}$
- ▶ Théorème (1) :  $\frac{1}{2}\underline{\text{dCE}}(\Gamma) \leq \text{smCE}(\Gamma) \leq 2\underline{\text{dCE}}(\Gamma)$

# COMMENT MESURER LA DISTANCE À LA CALIBRATION ?

## RÉSUMÉ

On considère une distribution d'observations  $\mathcal{D}$  et un prédicteur  $f$ . On s'intéresse à une mesure de calibration  $\mu_{\mathcal{D}}(f)$

1. On dit que la mesure est **consistante** si elle vérifie :

- **Complétude** :  $\mu_{\mathcal{D}}(f)$  vaut une petite valeur si  $f$  est proche de la calibration
- **Correction** :  $\mu_{\mathcal{D}}(f)$  ne vaut pas une petite valeur si  $f$  est loin de la calibration

Nous avons défini la **robuste** consistance pour rendre plus souple la définition initiale, et déduis des conditions minimale à respecter.

2. La vraie distance à la calibration ne pouvant pas être exploitée dans notre contexte, nous avons définie la **distance inférieure à la calibration** et montré que l'écart était au plus quadratique entre les deux notions.

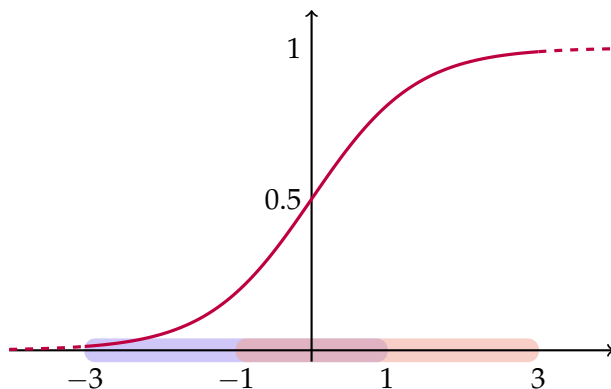
La manière naturelle de définir une mesure de calibration (ECE) ne satisfaisait pas la robuste consistance, donc nous avons définie une autre manière de mesurer la calibration (smCE).

# COMMENT INDUIRE LA CALIBRATION?

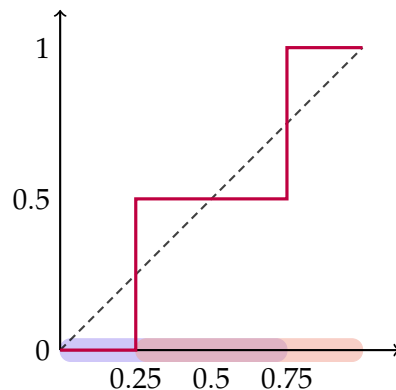
## INTRODUCTION

*Properly regularized logistic regression is well calibrated by default thanks to the use of the log-loss*

— Guide utilisateur scikit-learn (2007)



(a) Meilleur modèle obtenu



(b) Diagramme de calibration

**Figure** – Régression logistique pour la distribution  $(x, y) \sim \mathcal{D}$  avec  $y = 0$  pour  $x \sim \mathcal{U}([-3, 1])$  et  $y = 1$  pour  $x \sim \mathcal{U}([1, 3])$



# COMMENT INDUIRE LA CALIBRATION ?

## INTRODUCTION

Le lien entre performance d'optimisation de la fonction de perte et calibration n'est finalement pas si évident et contre-intuitif. C'est le travail réalisé dans [Błasiok et al., 2023b]. Commençons par éclaircir ce qu'on appelle *une fonction de perte bien définie*. On considère une distribution  $\mathcal{D}$  sur  $\mathcal{X} \rightarrow \{0, 1\}$  que nous omettrons de préciser pour simplifier la lecture, même dans la notation de smCE par exemple. On considère un prédicteur  $f : \mathcal{X} \rightarrow [0, 1]$ , sauf indication contraire.

### Définition 9 (Fonction de perte bien définie)

Soit  $V \subseteq [0, 1]$  un interval non vide. On dit qu'une fonction de perte  $\mathcal{L} : \{0, 1\} \times [0, 1]$  est bien définie si pour tout  $v \in V$ , on a que :

$$v \in \arg \min_{v' \in V} \mathbb{E}_{y \sim \text{Ber}(v)} [\mathcal{L}(y, v')]$$

Intéressons-nous dans un premier temps à la *Squared loss*  $\mathcal{L}_{\text{sq}}$  :

$$\forall y \in \{0, 1\}, \forall v \in V \subseteq [0, 1], \quad \mathcal{L}_{\text{sq}}(y, v) = (y - v)^2 \quad (\text{Squared loss})$$

## COMMENT INDUIRE LA CALIBRATION ?

### CAS PARTICULIER DE LA *Squared loss*

La notion centrale qu'introduit l'article est le *post-processing gap* :

#### Définition 10 (Post-processing gap)

Soit  $K$  une famille de fonctions de post-processing  $\kappa : [0, 1] \rightarrow [0, 1]$  telle que la fonction de mise à jour  $\eta(v) = \kappa(v) - v$  soit 1-Lipschitzienne.

On appelle *post-processing gap* de  $f$  par rapport à  $\mathcal{D}$  :

$$\text{pGap}_{\mathcal{D}}(f) = \mathbb{E} [\mathcal{L}_{\text{sq}}(y, f(x))] - \inf_{\kappa \in K} \mathbb{E} [\mathcal{L}_{\text{sq}}(y, \kappa \circ f(x))]$$

Le choix de la famille de fonction 1-Lipschitzienne, nous fait penser qu'on pourrait relier le post-processing gap avec la smooth calibration définie plus tôt :

#### Théorème 2

Pour la fonction de perte  $\mathcal{L}_{\text{sq}}$  :

$$\text{smCE}(f)^2 \leq \text{pGap}_{\mathcal{D}}(f) \leq 2 \text{smCE}(f)$$

## COMMENT INDUIRE LA CALIBRATION ?

### CAS PARTICULIER DE LA *Squared loss*

Prenons une autre fonction de perte, l'entropie croisée :

Observation positive

$$\mathcal{L}_{\text{xent}}(y, v) = - \left[ y \ln(v) + (1 - y) \ln(1 - v) \right]$$

Observation négative

On considère  $\mathcal{X} = \{x_0, x_1\}$  et  $\mathcal{D}$  une distribution sur  $\mathcal{X} \rightarrow \{0, 1\}$  telle que :

►  $\mathbb{E}_{(x,y) \in \mathcal{D}} [y \mid x = x_0] = 0.1$  et  $\mathbb{E}_{(x,y) \in \mathcal{D}} [y \mid x = x_1] = 0.9$

Soit  $f_\varepsilon : \mathcal{X} \rightarrow [0, 1]$  le prédicteur défini par  $f_\varepsilon(x_0) = \varepsilon$  et  $f_\varepsilon(x_1) = 1 - \varepsilon$  pour  $\varepsilon \in ]0, 0.1[$

Par la définition, on a que  $\lim_{\varepsilon \rightarrow 0} \text{smCE}(f_\varepsilon) = 0.05$ . Pourtant, puisque

$$\lim_{\varepsilon \rightarrow 0} \mathcal{L}_{\text{xent}}(1, \varepsilon) = \lim_{\varepsilon \rightarrow 0} \mathcal{L}_{\text{xent}}(0, 1 - \varepsilon) = +\infty$$
 invalidant le résultat obtenu précédemment pour  $\mathcal{L}_{\text{sq}}$ .

Nous ne pouvons donc pas *simplement* remplacer la fonction de perte. Il faut trouver une approche plus générale, en regardant la notion de fonction de perte d'une manière *différente*.

# COMMENT INDUIRE LA CALIBRATION?

## GÉNÉRALISATION DU LIEN ENTRE PGAP ET SMCE : PAR LA DUALITÉ

Les travaux de Leonard Savage en 1971 [Savage, 1971] et ceux de Tilmann Gneiting et Adrian Raftery en 2007 [Gneiting and Raftery, 2007] ont construit un pont entre les fonctions de pertes bien définies et les fonctions convexes.

### Proposition 3

Soit  $V \subseteq [0, 1]$  un intervalle non vide. Soit  $\mathcal{L} : \{0, 1\} \times V \rightarrow \mathbb{R}$  une fonction de perte bien définie. On a :

$$\begin{array}{c} \text{Fonction convexe } \psi : \mathbb{R} \rightarrow \mathbb{R} \\ \downarrow \\ \forall y \in \{0, 1\}, \forall v \in V, \quad \mathcal{L}(y, v) = \psi \left( \text{dual}(v) \right) - y \text{ dual}(v) \\ \uparrow \\ \forall v \in V, \text{dual}(v) = \mathcal{L}(0, v) - \mathcal{L}(1, v) \end{array}$$

De plus, si  $\psi$  est différentiable, alors  $\nabla \psi(t) \in [0, 1]$

# COMMENT INDUIRE LA CALIBRATION ?

GÉNÉRALISATION DU LIEN ENTRE PGAP ET SMCE : PAR LA DUALITÉ

## Définition 11 (Fonction de perte duale)

Soit une fonction  $\psi : \mathbb{R} \rightarrow \mathbb{R}$ . On définit une fonction de perte  $\mathcal{L}^{(\psi)} : \{0, 1\} \times \mathbb{R} \rightarrow \mathbb{R}$  comme :

$$\forall y \in \{0, 1\}, \forall t \in \mathbb{R}, \quad \mathcal{L}^{(\psi)}(y, t) = \psi(t) - yt$$

Ainsi, si une fonction de perte  $\mathcal{L} : \{0, 1\} \times V \rightarrow \mathbb{R}$  satisfait les conditions de la proposition 3 pour  $V \subseteq [0, 1]$  un intervalle non vide et  $\text{dual} : V \rightarrow \mathbb{R}$ , alors :

$$\forall y \in \{0, 1\}, \forall v \in V, \quad \mathcal{L}(y, v) = \mathcal{L}^{(\psi)}(y, \text{dual}(v))$$

The diagram illustrates the relationship between the loss function  $\mathcal{L}$  and the dual loss function  $\mathcal{L}^{(\psi)}$ . It shows the equation  $\mathcal{L}(y, v) = \mathcal{L}^{(\psi)}(y, \text{dual}(v))$ . A red arrow labeled "Induit avec dual" points from the variable  $v$  in the first term to the  $\text{dual}(v)$  term in the second term. A blue arrow labeled "Induit avec  $\mathcal{L}$ " points from the  $\text{dual}(v)$  term back to the variable  $v$ .

# COMMENT INDUIRE LA CALIBRATION ?

GÉNÉRALISATION DU LIEN ENTRE PGAP ET SMCE : PAR LA DUALITÉ

## Exercice 6 (Entropie croisée)

*On considère la fonction de perte entropie croisée définie comme :*

$$\mathcal{L}_{\text{xent}}(y, v) = -y \ln(v) - (1 - y) \ln(1 - v)$$

*On souhaite obtenir l'ensemble des informations nécessaires pour expliciter le mapping vers le monde dual. Identifier la fonction de perte duale,  $\psi$  et  $\text{dual}$ .*

Rappels :

- ▶  $\text{dual}(v) = \mathcal{L}(0, v) - \mathcal{L}(1, v)$
- ▶  $\mathcal{L}^{(\psi)}(y, t) = \psi(t) - yt$
- ▶  $\mathcal{L}(y, v) = \mathcal{L}^{(\psi)}(y, \text{dual}(v))$

## COMMENT INDUIRE LA CALIBRATION ?

GÉNÉRALISATION DU LIEN ENTRE PGAP ET SMCE : PAR LA DUALITÉ

### Définition 12 (Smooth calibration duale)

On note  $H_\lambda$  pour  $\lambda > 0$  l'ensemble des fonctions  $\eta : \mathbb{R} \rightarrow [-1, 1]$  qui sont  $\lambda$ -Lipschitzienne. Soit  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  une fonction différentiable avec pour tout  $t \in \mathbb{R}$ ,  $\nabla\psi(t) \in [0, 1]$ .

Pour une fonction  $g : \mathcal{X} \rightarrow \mathbb{R}$ , on définit le prédicteur  $f : \mathcal{X} \rightarrow [0, 1]$  tel que pour tout  $x \in \mathcal{X}$ ,  $f(x) = \nabla\psi(g(x))$ .

On appelle smooth calibration duale de  $g$  :

$$\text{smCE}^{(\psi, \lambda)}(g) = \sup_{\eta \in H_\lambda} \left| \mathbb{E}[(y - f(x))\eta \circ g(x)] \right|$$

## COMMENT INDUIRE LA CALIBRATION ?

GÉNÉRALISATION DU LIEN ENTRE PGAP ET SMCE : PAR LA DUALITÉ

Avant de continuer les développements, nous devons nous assurer que nous sommes capables de relier la définition classique de la smooth calibration à celle du monde dual.

### Proposition 4

Soit  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  une fonction convexe différentiable qui vérifie pour tout  $t \in \mathbb{R}$ ,  $\nabla\psi(t) \in [0, 1]$ . Soit  $\lambda > 0$ , on suppose que  $\psi$  est  $\lambda$ -smooth :

$$\forall x, y \in \mathbb{R}, \quad |\nabla\psi(x) - \nabla\psi(y)| \leq \lambda|x - y|$$

Pour  $g : \mathcal{X} \rightarrow \mathbb{R}$  on définit  $f : \mathcal{X} \rightarrow [0, 1]$  telle que  $f(x) = \nabla\psi(g(x))$  pour  $x \in \mathcal{X}$ . On a :

$$\text{smCE}(f) \leq \text{smCE}^{(\psi, \lambda)}(g)$$



# COMMENT INDUIRE LA CALIBRATION ?

GÉNÉRALISATION DU LIEN ENTRE PGAP ET SMCE : PAR LA DUALITÉ

## Théorème 3

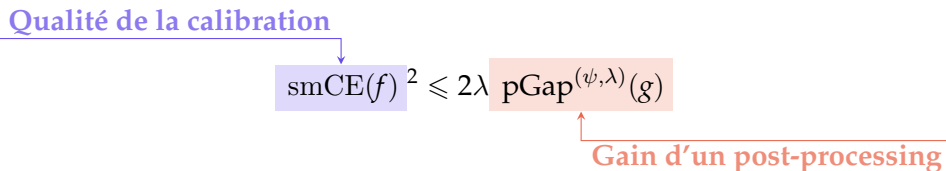
Soit  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  une fonction convexe différentiable qui vérifie pour tout  $t \in \mathbb{R}$ ,  $\nabla \psi(t) \in [0, 1]$ . Soit  $\lambda > 0$ , on suppose que  $\psi$  est  $\lambda$ -smooth.

Alors pour toute fonction  $g : \mathcal{X} \rightarrow \mathbb{R}$  et toute distribution  $\mathcal{D}$  sur  $\mathcal{X} \times \{0, 1\}$  :

$$\frac{1}{2} \text{smCE}^{(\psi, \lambda)}(g)^2 \leq \lambda \text{pGap}^{(\psi, \lambda)}(g) \leq \text{smCE}^{(\psi, \lambda)}(g)$$

L'écart entre les deux notions que l'on observe est à nouveau au plus quadratique, comme pour dCE et dCE. Maintenant que nous avons une borne dans le monde dual, développons-là dans le monde réel :

Qualité de la calibration


$$\text{smCE}(f)^2 \leq 2\lambda \text{pGap}^{(\psi, \lambda)}(g)$$

Gain d'un post-processing

# COMMENT INDUIRE LA CALIBRATION ?






## RÉSUMÉ

On cherche à savoir si une fonction de perte bien définie peut induire la calibration pour un prédicteur  $f$ .






1. Une **fonction de perte bien définie** permet d'atteindre le minimum que l'on cherche.
2. Le **post-processing gap** mesure le gain de performance que l'on peut obtenir en retravaillant uniquement les prédictions de  $f$ .
3. En traduisant une fonction de perte bien définie en version duale avec une paire de fonction  $(\psi, \text{dual})$ , *mutatis mutandis*, on obtient que la qualité de calibration est liée au gain d'un post-processing.

Autrement dit, pour qu'un prédicteur soit calibré, il faut qu'il apprenne à la fois la structure du problème et la fonction de post-processing. C'est une piste pour expliquer que les réseaux de neurones sont plus souvent calibré que les arbres par exemple.

## BIBLIOGRAPHIE I

-  [Błasiok, J., Gopalan, P., Hu, L., and Nakkiran, P. \(2023a\).](#)  
**A unifying theory of distance from calibration.**  
*In Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1727–1740.
-  [Błasiok, J., Gopalan, P., Hu, L., and Nakkiran, P. \(2023b\).](#)  
**When does optimizing a proper loss yield calibration?**  
*arXiv preprint arXiv :2305.18764*.
-  [Bryson, J. \(2018\).](#)  
**No one should trust artificial intelligence.**  
*Science & Technology : Innovation, Governance, Technology*.
-  [Dawid, A. P. \(1982\).](#)  
**The well-calibrated bayesian.**  
*Journal of the American Statistical Association*.
-  [Gneiting, T. and Raftery, A. E. \(2007\).](#)  
**Strictly proper scoring rules, prediction, and estimation.**  
*Journal of the American statistical Association*, 102(477) :359–378.

## BIBLIOGRAPHIE II

-  [Gopalan, P., Kim, M. P., Singhal, M. A., and Zhao, S. \(2022\).](#)  
**Low-degree multicalibration.**  
In *Conference on Learning Theory*, pages 3193–3234. PMLR.
-  [Jean, A. \(2021\).](#)  
***Les algorithmes font-ils la loi.***  
Edition de l’Observatoire.
-  [Kakade, S. M. and Foster, D. P. \(2004\).](#)  
**Deterministic calibration and nash equilibrium.**  
In *International Conference on Computational Learning Theory*, pages 33–48. Springer.
-  [O’Neil, C. \(2016\).](#)  
***Weapons of Math Destruction.***  
Crown Books.
-  [Savage, L. J. \(1971\).](#)  
**Elicitation of personal probabilities and expectations.**  
*Journal of the American Statistical Association*, 66(336) :783–801.