Bornes d'excès de risque asymptotiquement minimax pour l'estimation de densité et la régression logistique mal spécifiées

Jaouad Mourtada [1], Stéphane Gaïffas [2] & Erwan Scornet [1]

[1] *Centre de Mathématiques Appliquées, École Polytechnique*
UMR 7641, 91128 Palaiseau, France
{jaouad.mourtada,erwan.scornet}@polytechnique.edu
[2] *Laboratoire de Probabilités et Modèles Aléatoires, Université Paris-Diderot,*
UFR de Mathématiques, 75205 Paris
stephane.gaiffas@math.univ-paris-diderot.fr

Résumé. Nous introduisons un nouvelle procédure pour l'estimation de densité conditionnelle, qui satisfait une borne générale d'excès de risque prédictif pour la perte logarithmique. Cette borne reste valable dans le cas mal spécifié, et est d'ordre d/n dans de nombreux cas, où d est la dimension du modèle et n la taille de l'échantillon. En particulier, cette procédure est robuste au cas mal spécifié, contrairement à l'estimateur du maximum de vraisemblance qui peut se dégrader dans ce cas. Nous en déduisons une procédure *minimax au premier ordre* pour l'estimation de densité dans les familles exponentielles et la régression logistique mal spécifiées, avec un excès de risque asymptotique de $d/(2n) + o(1/n)$. Cette approche produit des estimateurs plus efficaces que celles utilisant des bornes sur le risque cumulées, et éliminent des facteurs logarithmiques superflus. Dans de nombreux cas (incluant la régression logistique), ces estimateurs sont calculables explicitement.

Mots-clés. Théorie de l'apprentissage statistique, Statistique mathématique.

Abstract. We introduce a new procedure for predictive conditional density estimation, which satisfies a general excess risk bound under logarithmic loss. This bound remains valid in the misspecified case, and scales as d/n in several cases, where d is the model dimension and n the sample size. In particular, this procedure is robust to misspecification, contrary to the maximum likelihood, which is sensitive to it. We deduce a *first-order minimax* procedure for misspecified density estimation in exponential families and logistic regression, with an asymptotic excess risk of $d/(2n) + o(1/n)$. Nous en déduisons en particulier une procédure *minimax au premier ordre* pour l'estimation de densité dans les familles exponentielles et la régression logistique mal spécifiées, avec un excès de risque asymptotique de $d/(2n) + o(1/n)$. This approach leads to more efficient estimators than those based on online-to-batch conversions, and eliminate suboptimal logarithmic terms from the bounds. In several cases (including logistic regression), these estimators can be computed explicitely.

Keywords. Statistical learning theory, Mathematical statistics.

1 Introduction

L'estimation de densité est un problème statistique classique. L'objectif est, étant donné un échantillon i.i.d. de loi inconnue, de produire une loi approchant la vraie loi des observations. Une approche standard consiste à postuler un modèle statistique, c'est-à-dire à supposer que la loi des observations appartient à une famille connue de lois de probabilité. Dans ce cas, le comportement de l'estimateur du maximum de vraisemblance (EMV) ou de procédures Bayésiennes est bien compris ; en particulier, de tels estimateurs sont asymptotiquement optimaux (Ibragimov et Has' Minskii, 1981; van der Vaart, 1998).

Une limitation importante de cette approche est qu'elle repose de manière cruciale sur l'hypothèse que le modèle est *bien spécifié*, c'est-à-dire que la loi des observations appartient au modèle utilisé. En général, cette hypothèse est très forte et n'a pas de raison d'être satisfaite, puisque le modèle est souvent un choix du statisticien traduisant ses connaissances et intuitions, et constitue une simplification du phénomène étudié. Cependant, un modèle même faux peut être utile, lorsque le modèle, bien que ne contenant pas la loi exacte (qui peut être complexe), capture certains aspects du phénomène étudié. Il est donc important de considérer le cas *mal spécifié* où la loi exacte n'appartient pas au modèle ; dans ce cas, il est naturel de chercher à obtenir une distribution qui approche la loi des observations presque aussi bien que la meilleure loi dans le modèle.

Nous considérons donc un problème d'estimation de densité mal spécifiée, où la qualité d'une loi estimée est mesurée par l'entropie relative (ou divergence de Kullback-Leibler). De manière équivalente, ce problème est un problème d'apprentissage statistique (Vapnik, 1998), où étant données une classe de prédicteurs et une fonction de perte, ainsi qu'un échantillon i.i.d., le but est de construire un prédicteur dont la perte moyenne est presque aussi faible que celle du meilleur élément dans la classe. Dans ce cas, la classe de prédicteurs est donnée par le modèle et la fonction de perte est la perte logarithmique $\ell(f, z) = -\log f(z)$ pour toute densité f et observation z.

En section 2, nous posons le problème et présentons les approches existantes pour ce problème, ainsi que leurs limites. En section 3, nous mentionnons une nouvelle procédure pour l'estimation de densité, que nous nommons *Sample Minimax Predictor* (SMP), qui est robuste au cas mal spécifié, ainsi que la variante HSMP, qui est asymptotiquement minimax au premier ordre pour les familles exponentielles. En section 4, nous discutons des bornes de SMP et HSMP pour le problème de la régression logistique.

2 Estimation de densité avec perte logarithmique

Définitions. Soit $\mathcal{Z}$ un espace mesurable, et μ une mesure sur $\mathcal{Z}$ (le choix de μ importe peu). On note $\mathcal{P}(\mathcal{Z})$ l'ensemble des densités de probabilités sur $\mathcal{Z}$ par rapport à μ, c'est-à-dire des fonctions mesurables positives $f : \mathcal{Z} \to \mathbf{R}^+$ telles que $\int_{\mathcal{Z}} f \mathrm{d}\mu = 1$. La *perte logarithmique* (ou : log-vraisemblance négative) est la fonction $\ell : \mathcal{P}(\mathcal{Z}) \times \mathcal{Z} \to \mathbf{R} \cup \{+\infty\}$ définie par $\ell(f, z) = -\log f(z)$.

Soit P une loi de probabilité sur $\mathcal{Z}$. Le risque d'une densité f est par définition sa perte moyenne sous P, soit $R(f) := \mathbb{E}[\ell(f, Z)]$ où $Z \sim P$. Pour toutes densités f, g, la quantité $\ell(f, z) - \ell(g, z)$ ne dépend que des lois $f \cdot \mu, g \cdot \mu$ et pas du choix de μ ; de plus, $R(f) - R(g) = \mathrm{KL}(P, f \cdot \mu) - \mathrm{KL}(P, g \cdot \mu)$, où $\mathrm{KL}(P, Q) = \mathbb{E}_{Z \sim P} \log \frac{\mathrm{d}P}{\mathrm{d}Q}(Z)$ est la divergence de Kullback-Leibler.

Soit $\mathcal{F}$ un modèle statistique sur $\mathcal{Y}$ dominé par μ, c'est-à-dire une famille de densités par rapport à μ. On dit que le modèle est bien spécifié lorsque $\frac{\mathrm{d}P}{\mathrm{d}\mu} \in \mathcal{F}$, et qu'il est mal spécifié dans le cas contraire. L'*excès de risque* de la densité g par rapport à la classe $\mathcal{F}$ est par définition

$$\mathcal{E}(g) = R(g) - \inf_{f \in \mathcal{F}} R(f)\,. \tag{1}$$

Soit $Z_1^n = (Z_1, \ldots, Z_n)$ un échantillon i.i.d. de loi P. Un *estimateur de densité* est par définition une fonction qui à Z_1^n associe une densité $\widehat{g}_n = \widehat{g}_n(Z_1, \ldots, Z_n)$. Dans ce qui suit, la qualité d'un estimateur sera mesurée par son excès de risque en espérance $\mathbb{E}[\mathcal{E}(\widehat{g}_n)] = \mathbb{E}[R(\widehat{g}_n)] - \inf_{f \in \mathcal{F}} R(f) = \mathbb{E}[\mathrm{KL}(P, \widehat{g}_n)] - \mathrm{KL}(P, f^*)$ où $f^* \in \arg\min R$.

Estimateur du maximum de vraisemblance. Un choix naturel d'estimateur est l'estimateur du maximum de vraisemblance (EMV) $\widehat{f}_n$, ou minimiseur du risque empirique :

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i) = \arg\max_{f \in \mathcal{F}} \prod_{i=1}^{n} f(Z_i)\,. \tag{2}$$

Dans le cas où le modèle est bien spécifié et régulier, l'EMV $\widehat{f}_n$ satisfait typiquement (au moins en supposant les densités minorées) $\mathbb{E}[\mathcal{E}(\widehat{f}_n)] \leqslant d/(2n) + o(1/n)$ (Ibragimov et Has' Minskii, 1981; van der Vaart, 1998). La vitesse $d/(2n)$ est asymptotiquement optimale (à la constante près), au sens où elle ne peut pas être améliorée uniformément, même localement ; on dit alors que l'EMV est *efficace*, ou *efficace au premier ordre*.

Cependant, le comportement de l'EMV peut se dégrader dans le cas mal spécifié. Une première limitation générale de l'EMV (qui tient au fait que la perte logarithmique est non bornée) est que les bornes non asymptotiques existantes requièrent souvent des conditions restrictives, comme des rapports de densités bornés (Birgé et Massart, 1998; Yang et Barron, 1998). Plus fondamentalement, même du point de vue asymptotique, l'excès de risque de l'EMV peut être sensiblement plus élevé que d/n en fonction de la vraie loi P. En effet, en notant $\mathcal{F} = \{f_\theta \,|\, \theta \in \Theta \subset \mathbf{R}^d\}$, sous des hypothèses de régularité classiques, en notant $\theta^* = \arg\min_{\theta \in \Theta} R(\theta)$, et

$$H = \nabla^2 R(\theta^*) = \mathbb{E}[\nabla^2 \ell(\theta^*, Z)], \qquad G = \mathbb{E}[\nabla \ell(\theta^*, Z) \nabla \ell(\theta^*, Z)^\top],$$

on a $\mathbb{E}[\mathcal{E}(\widehat{f}_n)] = \mathrm{tr}(H^{-1}G)/(2n) + o(1/n)$ (van der Vaart, 1998), où la quantité $\mathrm{tr}(H^{-1}G)$ peut en général être arbitrairement grande dans le cas mal spécifié.

Contrôle du risque cumulé et conversion « online to batch ». Il existe une théorie bien établie permettant de contrôler l'excès de risque *cumulé* pour la perte logarithmique (Merhav et Feder, 1998; Cesa-Bianchi et Lugosi, 2006), intimement liée au problème de la compression sans perte (Cover et Thomas, 2006) et au paradigme du *Minimum Description Length* (Rissanen, 1985; Grünwald, 2007). Plus précisément, étant donné un modèle $\mathcal{F}$ « borné » de dimension d, il existe des densités $\widehat{g}_t$ (ne dépendant que des observations $Z_1, \dots, Z_{t-1}$) telles que

$$\sum_{t=1}^{n} \ell(\widehat{g}_t, Z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f, Z_t) \leqslant \frac{d}{2} \log n + O(1) \tag{3}$$

quels que soient $Z_1, \dots, Z_n$. La borne de type (3) peut être obtenue en prenant pour $\widehat{g}_t$ un postérieur prédictif Bayésien selon une loi a priori π ayant pour support $\mathcal{F}$, *i.e.* $\widehat{g}_t(z) = \mathbb{E}_{f \sim \pi(\cdot \,|\, Z_1^{t-1})}[f(z)]$. Ceci implique une borne d'excès de risque *cumulé* $\mathbb{E}[\sum_{t=1}^{n} \mathcal{E}(\widehat{g}_t)] = (d/2) \log n + O(1)$. Par un procédé standard (appelé « conversion online to batch »), cette borne de risque cumulé implique une borne d'excès de risque (non cumulé) $\mathbb{E}[\mathcal{E}(\widehat{g}_n)] \leqslant (d \log n)/(2n) + O(1/n)$ pour la moyenne de Cesàro $\bar{g}_n = \frac{1}{n} \sum_{t=1}^{n} \widehat{g}_t$; lorsque $\widehat{g}_t$ est un postérieur de Bayésien, $\bar{g}_n$ est la *règle du mélange progressif* (Progressive Mixture Rule, PMR) (Catoni, 1997; Yang et Barron, 1998; Audibert, 2009). L'avantage de cette borne est qu'elle est valide dans le cas mal spécifié. En revanche, elle est sous-optimale d'un facteur $\log n$; cela correspond au fait qu'en pratique, cet estimateur est peu efficace car trop conservateur.

3 Un estimateur de densité minimax au premier ordre

Dans l'article Mourtada *et al.* (2019), nous introduisons une procédure générale pour l'estimation de densité, appelée *Sample Minimax Predictor* (SMP) ; cette procédure est *impropre*, au sens où l'estimateur $\widetilde{f}_n$ n'appartient en général pas à la classe $\mathcal{F}$. Cette procédure (et ses variantes pénalisées) admet une borne explicite d'excès de risque, dont on peut montrer asymptotiquement ou non-asymptotiquement dans divers cas qu'elle est d'ordre d/n, indépendamment de la loi P (Mourtada *et al.*, 2019). Nous introduisons également un autre estimateur, appelé *Half Sample Minimax Predictor*, pour lequel nous obtenons des bornes d'excès de risque asymptotiques minimax au premier ordre ; mentionnons que $\widetilde{h}_n = \frac{1}{2}(\widehat{f}_n + \widetilde{f}_n)$, où $\widehat{f}_n$ est l'EMV et $\widetilde{f}_n$ le SMP.

Théorème 1. *Supposons que $\mathcal{F}$ est une famille exponentielle $\mathcal{F} = \{f_\theta(y) = \exp(\langle \theta, y \rangle - \psi(\theta)) | \theta \in \Theta\}$, où μ une mesure sur $\mathcal{Y} = \mathbf{R}^d$, $\psi(\theta) = \log \int_{\mathbf{R}^d} e^{\langle \theta, y \rangle} \mu(\mathrm{d}y)$ et Θ est un sous-ensemble convexe compact de $\{\theta \in \mathbf{R}^d : \psi(\theta) < +\infty\}$.*

Alors, quelle que soit la loi P de Y, lorsque $n \to \infty$, l'excès de risque du SMP $\widetilde{f}_n$ satisfait

$$\mathbb{E}[\mathcal{E}(\widetilde{f}_n)] \leqslant \frac{d}{n} + o\left(\frac{1}{n}\right) , \tag{4}$$

et $\mathbb{E}[\mathcal{E}(\widetilde{f}_n)] \leqslant d/(2n) + o(1/n)$ *dans le cas bien spécifié. Par ailleurs, le HSMP satisfait dans le cas général* $\mathbb{E}[\mathcal{E}(\widetilde{h}_n)] \leqslant d/(2n) + o(1/n)$*, il est donc* minimax au premier ordre.

4 Régression logistique

Nous considérons maintenant le problème classique de la *régression logistique*. Il s'agit d'un problème d'estimation de densité conditionnelle, où les variables prédictives forment un vecteur $x \in \mathbf{R}^d$, tandis que la réponse $y \in \{-1, 1\}$ est binaire ; on cherche alors à prédire y à partir de x, ce qui revient à approcher la densité conditionnelle de Y sachant X. La perte d'une densité conditionnelle $f(y \,|\, x)$ au point $z = (x, y)$ est $\ell(f, z) = -\log f(y \,|\, x)$. Le risque de f est alors défini comme $R(f) = \mathbb{E}[\ell(f, Z)] = \mathbb{E}[-\log f(Y \,|\, X)]$. Dans le problème de la régression logistique, la classe $\mathcal{F}$ est donné par les prédicteurs linéaires, au sens où $\log[f(1 \,|\, x)/f(-1 \,|\, x)] = \langle \beta, x \rangle$ pour un certain $\beta \in \mathbf{R}^d$.

La régression logistique est un problème classique ; dans le cas bien spécifié (où la densité conditionnelle de $Y \,|\, X$ est linéaire), des procédures comme l'EMV atteignent la vitesse optimale de $d/(2n)+o(1/n)$ (van der Vaart, 1998). Dans le cas général, en revanche, le problème est intrinsèquement plus difficile. En effet, en notant R une borne sur la norme de X, la quantité $\mathrm{tr}(H^{-1}G)$ mentionnée précédemment peut être d'ordre de^R, avec une dépendance exponentielle en R (Bach et Moulines, 2013; Hazan *et al.*, 2014). Dans ce cas, l'EMV admet asymptotiquement un excès de risque de de^R/n. Cette limitation n'est pas propre à l'EMV, et affecte tous les estimateurs $\widehat{g}_n$ *propres* (c'est-à-dire appartenant à la classe $\mathcal{F}$, *i.e.* linéaires en la variable x) : en effet, une borne inférieure de Hazan *et al.* (2014) montre qu'un tel estimateur ne peut faire mieux que $O(\min(R/\sqrt{n}, de^R/n))$ dans le pire des cas. Récemment, en utilisant la règle de mélange progressif, Foster *et al.* (2018) montre qu'une borne de $O(d \log(Rn)/n)$ est possible par une procédure impropre ; en revanche, cet algorithme est difficile à implémenter en pratique, puisqu'il utilise un mélange continu de logistiques, approché par des méthodes de type MCMC.

Théorème 2. *Les prédictions du SMP peuvent être calculées de manière efficace, au coût de deux régressions logistiques. De plus, le SMP satisfait, sous des hypothèses peu contraignantes (moments finis de* X*, absence d'hyperplan séparateur) sur la loi de* (X, Y) *:* $\mathbb{E}[\mathcal{E}(\widetilde{f}_n)] \leqslant d/n + o\,(1/n)$ *; en outre, le HSMP satisfait, sous les mêmes hypothèses,* $\mathbb{E}[\mathcal{E}(\widetilde{h}_n)] \leqslant d/(2n) + o(1/n)$*.*

Nous obtenons également des bornes non asymptotiques pour le SMP et des variantes pénalisées (Mourtada *et al.*, 2019).

Références

Audibert, J.-Y. (2009). Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646.

Bach, F. et Moulines, É. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). *In Advances in Neural Information Processing Systems 26*, pages 773–781. Curran Associates, Inc.

Birgé, L. et Massart, P. (1998). Minimum contrast estimators on sieves : exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375.

Catoni, O. (1997). The mixture approach to universal model selection. Rapport technique, École Normale Supérieure.

Cesa-Bianchi, N. et Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, New York, USA.

Cover, T. M. et Thomas, J. A. (2006). *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, New York, USA, 2nd édition.

Foster, D. J., Kale, S., Luo, H., Mohri, M. et Sridharan, K. (2018). Logistic regression : The importance of being improper. *In Proceedings of the 31st Conference On Learning Theory (COLT)*, pages 167–208.

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. MIT Press.

Hazan, E., Koren, T. et Levy, K. Y. (2014). Logistic regression : Tight bounds for stochastic and online optimization. *In Proceedings of the 27th Conference on Learning Theory (COLT)*, pages 197–209.

Ibragimov, I. A. et Has' Minskii, R. Z. (1981). *Statistical estimation : asymptotic theory*, volume 16. Springer Science & Business Media.

Merhav, N. et Feder, M. (1998). Universal prediction. *IEEE Transactions on Information Theory*, 44:2124–2147.

Mourtada, J., Gaïffas, S. et Scornet, E. (2019). An excess risk bound for statistical learning, with applications to misspecified density estimation and logistic regression. *In preparation*.

Rissanen, J. J. (1985). *Minimum description length principle*. Wiley Online Library.

van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.

Yang, Y. et Barron, A. R. (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44(1):95–116.