Sous-échantillonnage de données pour les réseaux de neurones bayésiens

Eiji Kawasaki [1] & Markus Holzmann [2] & Lawrence Adu-Gyamfi [1]

[1] *Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France*
[2] *Univ. Grenoble Alpes, CNRS, LPMMC, 38000 Grenoble, France*

Résumé. La mise au point d'une méthode efficace de quantification d'incertitude en apprentissage profond est un tâche importante mais difficile car elle implique le calcul de la distribution prédictive en marginalisant l'ensemble des paramètres des réseaux de neurones. Dans ce contexte, les algorithmes Monte Carlo par chaînes de Markov ne s'adaptent pas bien aux grands volumes de données, ce qui entraîne des difficultés dans l'échantillonnage des distributions a posteriori des réseaux de neurones. En visant cet objectif d'inférence bayésienne, nous proposons de montrer qu'une généralisation de l'algorithme Metropolis-Hastings permet de restreindre l'évaluation de la vraisemblance à des sous-ensembles des données d'entraînement nommés mini-lot. Comme cette méthode nécessite le calcul d'une "pénalité de bruit" déterminée par la variance de la fonction de perte sur ces mini-lots, nous appelons cette stratégie de sous-échantillonnage des données "réseaux neuronaux bayésiens à pénalité de bruit".

Mots-clés. Réseaux de neurones bayésiens, Monte Carlo par chaînes de Markov

Abstract. The development of an effective uncertainty quantification method that computes the predictive distribution by marginalizing over Deep Neural Network parameter sets remains an important, challenging task. In this context, Markov Chain Monte Carlo algorithms do not scale well for large datasets leading to difficulties in Neural Network posterior sampling. We show that a generalization of the Metropolis Hastings algorithm allows to restrict the evaluation of the likelihood to small mini-batches in a Bayesian inference context. Since it requires the computation of a so-called "noise penalty" determined by the variance of the training loss function over the mini-batches, we refer to this data subsampling strategy as Penalty Bayesian Neural Networks.

Keywords. Bayesian Neural Network, Markov Chain Monte Carlo

1 Réseaux de neurones bayésien à pénalité de bruit

1.1 Introduction

Le développement d'une méthode efficace de quantification de l'incertitude de prédictions de modèles de réseaux de neurones profonds est une tâche importante et difficile [1]. Les méthodes d'inférence bayésienne permettent d'obtenir la distribution a posteriori des paramètres

en utilisant par exemple l'inférence variationnelle ou bien encore l'échantillonnage Monte Carlo. Les techniques de Monte Carlo par chaînes de Markov (MCMC) sont généralement considérées comme la référence en matière d'inférence bayésienne [2]. Cependant, l'exploration de l'espace des paramètres d'un réseau de neurones par une chaîne de Markov ne s'adapte pas bien aux grands volumes de données. En effet, elle nécessite l'évaluation de la log-vraisemblance du modèle sur l'ensemble des données à chaque étape d'itération. En pratique, cela constitue un grave obstacle à l'utilisation de l'échantillonnage des réseaux de neurones bayésiens. Le développement d'un algorithme d'échantillonnage MCMC pour des réseaux de neurones bayésiens capables de traiter des ensembles de données aux dimensions satisfaisantes dans le cadre de l'apprentissage profond reste donc un problème ouvert.

Par analogie avec les techniques de descente de gradient stochastique omniprésentes en apprentissage automatique, nous proposons une stratégie de sous-échantillonnage des données pour l'évaluation de la distribution a posteriori d'un réseau de neurones. Cela nous conduit à une variante du MCMC que nous appelons réseau de neurones bayésien avec pénalité de bruit. En effet, ne pas prendre en compte le bruit dû au sous-échantillonnage des données ne permet pas de correctement approcher la distribution par MCMC. Ce constat a été établi dans le contexte de l'inférence bayésienne : plusieurs méthodologies de sous-échantillonnage MCMC ont été proposées pour généraliser l'algorithme de Metropolis-Hastings et maîtriser ce biais [3]. Nous montrons, à la fois théoriquement et empiriquement, que notre approche originale permet un échantillonnage non baisé de la loi a posteriori en calculant explicitement la variance de la différence des fonctions de perte d'un mini-lot de données.

1.2 Distribution a posteriori et mini-lots

Nous considérons ici un vecteur θ qui décrit les paramètres d'un modèle, ce vecteur pourra notamment représenter les poids et les biais d'un réseau de neurones. Nous définissons $p(\theta)$ comme une distribution a priori sur cet ensemble de paramètres. Les distributions a priori usuelles en apprentissage profond sont les distributions gaussienne et de Laplace, qui correspondent respectivement aux régularisations L2 et L1. Dans le contexte de l'apprentissage supervisé, nous appelons $p(y|x,\theta)$ la probabilité d'une cible y compte tenu d'une donnée d'entrée x et d'un vecteur de paramètres θ. L'incertitude sur les paramètres θ étant donné un ensemble d'observations $\mathcal{D}$ est décrite par la distribution a posteriori qui est définie suivant,

$$p(\theta|\mathcal{D}) = \frac{p(\theta)\prod_{i=1}^{N} p(y_i|x_i,\theta)}{p(\mathcal{D})} \tag{1}$$

où $\mathcal{D} = \{(y_i, x_i)\}_{i=1}^{N}$. A une constante près, $p(\theta|\mathcal{D}) \propto e^{-\mathcal{L}_{\mathcal{D}}(\theta)}$ où la fonction de perte $\mathcal{L}_{\mathcal{D}}(\theta)$ correspond au log négatif de la distribution a posteriori

$$\mathcal{L}_{\mathcal{D}}(\theta) = -\log p(\theta) - \sum_{i=1}^{N} \log p(y_i|x_i,\theta) \tag{2}$$

Le dernier terme est la négative log-vraisemblance. Ce choix de fonction de perte n'est donné qu'à titre d'illustration et ne réduit pas la généralité de l'approche présentée ici, car nous

aurions également pu envisager une configuration non supervisée dans laquelle $\mathcal{D} = \{(x_i)\}_{i=1}^N$ et $\mathcal{L}_\mathcal{D}(\theta) = -\log p(\theta) - \sum_{i=1}^N \log p(x_i|\theta)$.

En apprentissage conventionnel, les paramètres du réseau sont généralement estimés par une descente de gradient stochastique qui cible le maximum de la distribution a posteriori en utilisant des mini-lots de données MB. Il en résulte une fonction de perte définie comme suit

$$\mathcal{L}_{\text{MB}}(\theta) = -\log p(\theta) - \frac{N}{n}\sum_{i=1}^{n} \log p(y_i|x_i, \theta) \tag{3}$$

où n correspond à la taille du mini-lot qui contient des données sous-échantillonnées sans remise, de sorte que par définition $\langle \mathcal{L}_{\text{MB}}(\theta) \rangle = \mathcal{L}_\mathcal{D}(\theta)$.

1.3 Metropolis-Hastings et bruit gaussien

Il est possible d'obtenir un ensemble i.i.d. d'échantillons de la distribution a posteriori définie dans l'équation 1 à l'aide d'un algorithme MCMC en explorant l'espace de définition de θ à l'aide de chaînes de Markov. Il est bien connu que l'équation de l'équilibre détaillé est une condition suffisante mais non nécessaire garantissant que ce processus de Markov possède une distribution stationnaire correspondant à l'équation 1. L'équilibre détaillé est donné par

$$A(\theta, \theta')q(\theta|\theta')e^{-\Delta(\theta',\theta)} = A(\theta', \theta)q(\theta'|\theta) \tag{4}$$

où $A(\theta', \theta)$ correspond à la probabilité d'accepter un déplacement d'un vecteur de paramètres θ vers θ'. Ce changement d'état est suggéré par la distribution $q(\theta'|\theta)$. Par souci de concision, nous considérons dans la suite une distribution de proposition symétrique $q(\theta'|\theta) = q(\theta|\theta')$. En utilisant l'algorithme de Metropolis-Hastings, l'acceptation s'écrit alors $A(\theta', \theta) = \min\left(1, e^{-\Delta(\theta',\theta)}\right)$ où $\Delta(\theta', \theta)$ correspond à la différence de fonctions de perte définie par

$$\Delta(\theta', \theta) = \mathcal{L}_\mathcal{D}(\theta') - \mathcal{L}_\mathcal{D}(\theta) \tag{5}$$

Nous souhaitons calculer les différences de fonction de perte sur des mini-lots aléatoires plutôt que sur l'ensemble des données $\mathcal{D}$. Par conséquent, nous introduisons une variable aléatoire $\delta(\theta', \theta)$ qui fournit une estimation non biaisée de $\Delta(\theta', \theta)$

$$\delta(\theta', \theta) \sim \mathcal{N}(\Delta(\theta', \theta), \sigma^2(\theta', \theta)) \tag{6}$$

ce qui signifie que nous pouvons écrire $\delta(\theta', \theta)$ comme étant égal à la différence de perte $\Delta(\theta', \theta)$ à laquelle on ajoute un bruit dont nous faisons l'hypothèse qu'il est distribué selon une gaussienne. Généralement, l'augmentation de la taille n des mini-lots diminue l'amplitude du bruit, c'est-à-dire qu'il réduit la variance $\sigma^2(\theta', \theta)$.

Comme le montre la figure 1, l'estimateur de la différence des fonctions de perte $\delta(\theta', \theta)$ empêche les algorithmes MCMC d'échantillonner la distribution a posteriori ciblée si le bruit qu'il introduit n'est pas correctement pris en compte. Dans le contexte de la physique statistique et de la chimie informatique, Ceperley et Dewing (1999) [5] ont généralisé l'algorithme de marche aléatoire de Metropolis-Hastings à la situation où la différence Δ (différence

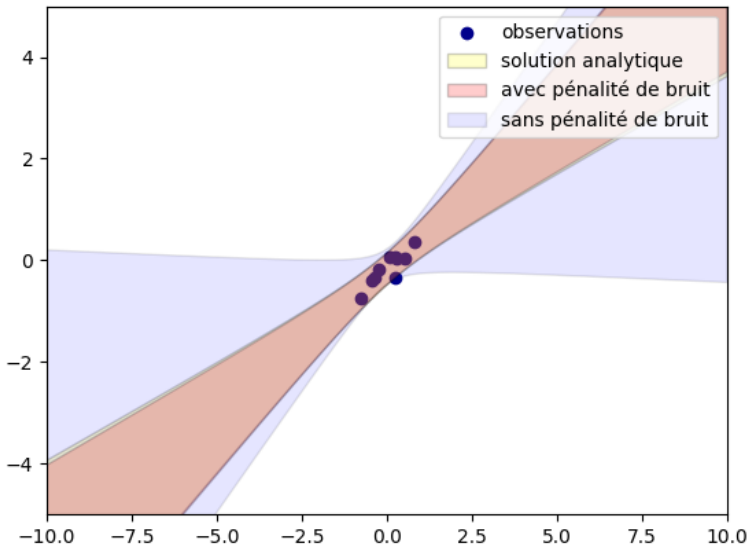

Figure 1: Tracé des distributions prédictives a posteriori calculées pour une régression linéaire univariée. Les zones colorées correspondent à la moyenne des distributions ± un écart-type. La courbe bleue est calculée en remplaçant naïvement Δ par δ dans l'algorithme MH. La différence bruitée δ est calculée sur un seul mini-lot contenant un sous-ensemble de 2 données. La courbe jaune inclut le terme de pénalité de bruit supplémentaire tel que défini dans l'eq 7. Nous remarquons qu'elle se superpose à la courbe rouge qui représente la solution analytique d'une régression linéaire bayésienne avec une distribution a priori gaussienne et une variance aléatorique connue [4].

d'énergies dans leur cas d'étude) est bruitée par un bruit gaussien et ne peut être qu'estimée. Ils ont montré qu'il est possible d'échantillonner la distribution ciblée malgré la présence d'un bruit. Pour cela, il est nécessaire de modifier la probabilité d'acceptation et d'appliquer une pénalité de bruit $-\sigma^2(\theta', \theta)/2$ à la différence de perte dans le ratio d'acceptation A, de telle sorte que :

$$A(\delta, \theta', \theta) = \min\left(1, e^{-\delta(\theta', \theta) - \sigma^2(\theta', \theta)/2}\right) \tag{7}$$

On peut alors montrer que l'équilibre détaillé est satisfait en moyenne suivant l'équation 8, ce qui est une condition suffisante pour que la chaîne de Markov échantillonne la distribution sans biais dans le régime stationnaire.

$$\begin{aligned} &\int d\delta A(\delta, \theta, \theta') q(\theta|\theta') \mathcal{N}(\delta; \Delta(\theta', \theta), \sigma^2(\theta', \theta)) e^{-\delta} \\ &= \int d\delta A(\delta, \theta', \theta) q(\theta'|\theta) \mathcal{N}(\delta; \Delta(\theta, \theta'), \sigma^2(\theta, \theta')) \end{aligned} \tag{8}$$

Notons que cette méthode de pénalité de bruit peut être étendue à une distribution de proposition non symétrique $q(\theta'|\theta)$. A ce titre, les réseaux de neurones bayésiens sont

souvent échantillonnés par dynamique de Langevin [6] ou encore par Monte Carlo hybride [7]. L'inconvénient de la pénalité de bruit est qu'elle entraîne une diminution exponentielle de l'acceptation $A(\delta, \theta, \theta')$, puisque la variance $\sigma^2(\theta', \theta)$ est toujours non négative. Notons en outre que dans le cas de l'échantillonnage a posteriori, $\sigma^2(\theta', \theta)$ n'est en général pas connu et ne peut qu'être estimé. Il est possible d'étendre ce raisonnement pour prendre en compte des variances bruitées [5].

1.4 La pénalité de bruit en pratique

Afin d'obtenir une acceptation moyenne raisonnable $A(\delta, \theta', \theta_t)$, c'est-à-dire qui ne tend pas vers 0, la différence de fonction de perte $\delta(\theta', \theta_t)$ doit dominer la variance $\sigma^2(\theta', \theta_t)/2$ qui est par définition toujours positive. Toutefois, en pratique, la pénalité de bruit domine souvent tout gain entre θ_t et θ' si cette différence n'est calculée que sur un seul petit mini-lot. Cela conduit à une dimininution exponentielle de l'acceptation et à de longs temps de corrélation de la chaîne de Markov.

Pour éviter cette situation, nous définissons $\delta(\theta', \theta)$ comme une moyenne empirique de la différence des fonctions de perte :

$$\delta(\theta', \theta) = \frac{1}{M} \sum_{j=1}^{M} \left(\mathcal{L}_{\mathrm{MB},j}(\theta') - \mathcal{L}_{\mathrm{MB},j}(\theta) \right) \tag{9}$$

où MB, j correspond à un mini-lot j choisi aléatoirement. Par définition, la moyenne est un estimateur non biaisé tel que $\langle \delta(\theta', \theta) \rangle = \Delta(\theta', \theta)$. Nous remarquons ici que le théorème central limite garantit que $\delta(\theta', \theta)$ est distribué selon une gaussienne dans la limite d'un nombre M infini.

La variance σ^2 de la variable aléatoire $\delta(\theta', \theta)$ diminue lorsque le nombre de mini-lots M augmente. Cette variance théorique nous est inconnue, mais nous pouvons en calculer une estimation non biaisée :

$$\sigma^2(\theta', \theta) \simeq \frac{1}{M(M-1)} \sum_{j=1}^{M} \left(\mathcal{L}_{\mathrm{MB},j}(\theta') - \mathcal{L}_{\mathrm{MB},j}(\theta) - \delta(\theta', \theta) \right)^2 \tag{10}$$

La figure 2 illustre une expérience numérique d'un réseaux de neurones bayésien à pénalité de bruit sur une tâche de régression à partir de données synthétiques. L'erreur sur l'estimation de la variance $\sigma^2(\theta', \theta)$ n'est pas prise en compte, nous prenons ici l'hypothèse que ses variations en fonction de θ' et θ dominent largement le bruit dû à son estimation. Les corrections d'ordre supérieur tenant compte de cette incertitude sont discutées dans [5].

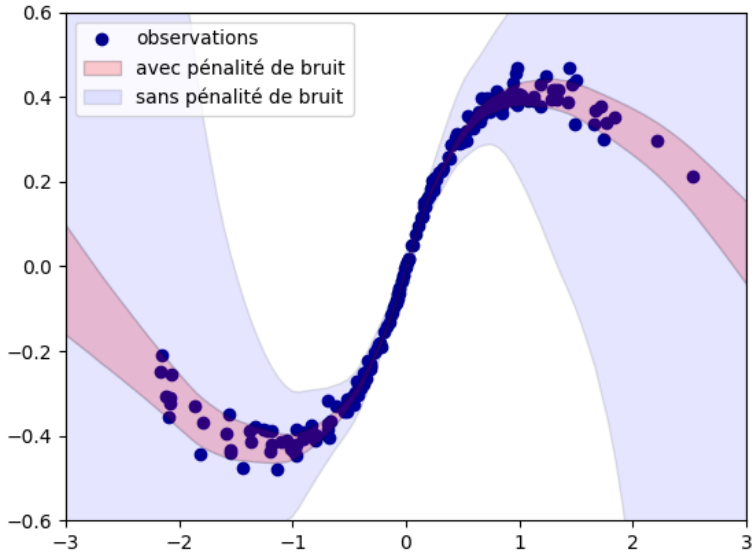

Figure 2: Tracé des distributions prédictives a posteriori calculées pour un problème de régression synthétique univariée. Le modèle de vraissemblance est une gaussienne dont la moyenne et la variance sont paramétrisés par un réseau de neurones [8].

Bibliographie

[1] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A Survey of Uncertainty in Deep Neural Networks. *arXiv:2107.03342 [cs, stat]*, July 2021. arXiv: 2107.03342.

[2] Pavel Izmailov, Sharad Vikram, Matthew D. Hoffman, and Andrew Gordon Wilson. What Are Bayesian Neural Network Posteriors Really Like? *arXiv:2104.14421 [cs, stat]*, April 2021. arXiv: 2104.14421.

[3] Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43, 2017.

[4] Christopher M Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[5] D. M. Ceperley and M. Dewing. The penalty method for random walks with uncertain energies. *The Journal of Chemical Physics*, 110(20):9812–9820, May 1999.

[6] Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. page 8.

[7] Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer New York, New York, NY, 1996.

[8] Christopher M. Bishop. Mixture density networks, 1994. Num Pages: 26 Place: Birmingham Publisher: Aston University.