entropy

MDPI

Editorial

Approximate Bayesian Inference

Pierre Alquier

Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo 103-0027, Japan; pierrealain.alquier@riken.jp

Received: 28 October 2020; Accepted: 6 November 2020; Published: 10 November 2020

Abstract: This is the Editorial article summarizing the scope of the Special Issue: Approximate Bayesian Inference.

Keywords: Bayesian statistics; machine learning; variational approximations; PAC-Bayes; expectation-propagation; Markov chain Monte Carlo; Langevin Monte Carlo; sequential Monte Carlo; Laplace approximations; approximate Bayesian computation; Gibbs posterior

1. Introduction

Extremely popular for statistical inference, Bayesian methods are gaining importance in machine learning and artificial intelligence problems. Indeed, in many applications, it is important for any device not only to predict well, but also to provide a quantification of the uncertainty of the prediction.

The main problem when one is to apply Bayesian statistics is that the computation of the estimators is expensive and sometimes not feasible. Bayesian estimators are based on the posterior distribution on parameters θ given by:

$$\pi(\theta|x) = \frac{\mathcal{L}(\theta;x)\pi(\theta)}{\int \mathcal{L}(\theta;x)\pi(\mathrm{d}\theta)} \tag{1}$$

where π is the prior, x the observations, and $\mathcal{L}(\theta;x)$ the likelihood function. For example, the computation of the posterior mean $\int \theta\pi(\mathrm{d}\theta|x)$ requires a difficult evaluation of the integrals. Thanks to the development of computational power, Bayesian estimation became feasible in the 1980s and the 1990s through Markov Chain Monte Carlo (MCMC) methods, such as the Metropolis–Hastings algorithm [1] and the Gibbs sampler [2,3]. These algorithms target the exact posterior distribution. They proved to be useful in many contexts and are still an active area of research. The performances and applicability of MCMC were improved by variants such as the Hamiltonian MCMC [4,5], adaptive MCMC [6–8], etc. We refer the reader to the review [9], the books [10–12], and Part III in [13] for detailed introductions to MCMC. The surveys [14,15] provide an overview on more recent advances. The asymptotic theory of Markov chains, ensuring the consistency of these algorithms, was covered in the monographs [16,17]. A few non-asymptotic results are also available [18].

Sequential Monte Carlo emerged in the 1990s as a way to update sequentially (that is, for each new data) samples from the posterior in hidden state models. They allow thus the computation of a Bayesian version of filters (such as the Kalman filter [19]). For this reason, they are also referred to as "particle filters". We refer the reader to [20] for the state-of-the-art of the early years and to the recent books [21,22] for pedagogical introductions and an overview of the most recent progress.

However, many modern models in statistics are simply too complex to use such methodologies. In machine learning, the volume of the data used in practice makes MCMC too slow to be used: first, each iteration of the algorithm requires accessing all the data, then the number of iterations required to reach convergence explodes when the dimension is large. In these cases, it seems that targeting the exact posterior is no longer a realistic objective. This motivated the development of many new methodologies, where the target is no longer the exact posterior, but simply a part of the information contained in it, or an approximation.

Before a short overview of these approximations techniques, let us mention two important examples where approximations were an essential ingredient in the application of Bayesian methods. In 2006, Netflix released a dataset containing movie ratings by its users and challenged the machine learning community to improve on its own predictions for movies that were not rated [23]. Many algorithms were proposed, including methods based on matrix factorization. Bayesian matrix factorization is computationally intensive. The first success at scaling Bayesian methods to the Netflix dataset was based on a mean-field variational approximation of the posterior by [24]. Such approximations will be discussed below.

In computer vision problems, the best performances are reached by deep neural networks [25]. Bayesian neural networks became a popular research direction. A new field of Bayesian deep learning has emerged that relies on approximate Bayesian inference to provide uncertainty estimates for neural networks without increasing the computation cost too much [26–29]. In particular, References [28,29] scaled these algorithms to the size of benchmark datasets such as CIFAR-10 and ImageNet.

2. Approximation in the Modelization

In many practical situations, the statistician is not interested in building a complete model describing the data, but simply in learning some aspects of it. One can think for example of a classification problem where one does not want to learn the full distribution of the data, but only a good classifier. A natural idea is to replace $\pi(\theta|x)$ in (1) by:

$$\tilde{\pi}(\theta|x) = \frac{\exp\left[-\ell(x;\theta)\right]\pi(\theta)}{\int \exp\left[-\ell(x;\theta)\right]\pi(\mathrm{d}\theta)} \tag{2}$$

where $\ell(x;\theta)$ is a Taylor loss function—for example, the classification error. When $\ell(x;\theta) = -\log \mathcal{L}(\theta;x)$, we recover (1) as a special case. When $\ell(x;\theta) = -\alpha \log \mathcal{L}(\theta;x)$ for some $\alpha \neq 1$, we obtain tempered posteriors, which appeared for various computational and theoretical reasons in the statistical literature; see [30–34], respectively. The use of the general form (2) was advocated to the statistical community by [35].

It appears that this idea was already popular in the machine learning theory community, where distributions like $\tilde{\pi}(\theta|x)$ are often referred to as Gibbs posteriors or aggregation rules. The PAC-Bayesian theory was developed to provide upper bounds on the prediction risk of such distributions [36–38]. We refer the reader to nice tutorials on PAC-Bayes bounds [39,40]. References [41–43] emphasized the connection to information theory. Note that the dropout technique used in deep learning to improve the performances of neural networks [44] was studied with PAC-Bayes bounds in [40]; see also [26]. Many publications in the past few years indeed confirmed that PAC-Bayes bounds are very well suited to analyze the performances of deep learning [45–51]. See [52] for a recent survey on PAC-Bayes bounds.

Such distributions were also well known in game theory and in prediction with expert advice since the 1990s [53,54]. We refer to the book [55], the recent work [56], and to connected problems such as bandits [57,58].

Finally, many aggregation procedures studied in high-dimensional statistics can also be written under the form of (2); see [59–64] with various regression or classification losses. References [65] used a Gibbs posterior based on the quantile loss to estimate a VaR (Value at Risk, a measure of risk in finance).

3. Approximation in the Computations

Many works have been done in the past few years to compute estimators based on $\pi(\theta|x)$ or $\tilde{\pi}(\theta|x)$ in complex problems, or with very large datasets. Very often, this is at the cost of targeting an approximation rather than the exact posterior. It is then important to analyze the accuracy of the approximation.

The nature and accuracy of these approximations are extremely different from one algorithm to the other, and some of them are not well understood theoretically. Below, we group these algorithms into three groups. In Section 3.1, we present methods that still essentially rely on simulations. In Section 3.2, we present asymptotic approximations. Finally, in Section 3.3, we present optimization based methods (this grouping is for the ease of exposition and is of course a little crude; each subsection mentions methods that have little to do with each other).

3.1. Non-Exact Monte Carlo Methods

Monte Carlo methods based on Langevin diffusions were introduced in physics in the 1970s [66]. Let $(U_t)_{t\geq 0}$ be a diffusion process given by the stochastic differential equation:

$$\mathrm{d}U_t = \nabla \log \pi(U_t|x)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t,$$

where $(W_t)_{t\geq 0}$ is a standard Brownian motion. It turns out that the invariant distribution of (U_t) is $\pi(\cdot|x)$. A discretization scheme with step $h > 0$ leads to the Markov chain $\tilde{U}_{n+1} = \tilde{U}_n + h\nabla \log \pi(U_n|x) + \sqrt{2h}\xi_n$, where the (ξ_n) are i.i.d standard Gaussian variables. However, it is important to note that (U_n) does not admit $\pi(\cdot|x)$ as an invariant distribution. Thus, the Langevin Monte Carlo method is not exact (it would become exact with $h \to 0$). Reference [67] proposed a correction of this method based on the Metropolis–Hastings algorithm, which leads to an exact algorithm, known as the MALA (the Monte Carlo Adjusted Langevin Algorithm). The Langevin Monte Carlo and MALA became popular in statistics and machine learning following [68]. This paper studies the asymptotic properties of both algorithms. Surprisingly, the exact method does not necessarily enjoy the best asymptotic guarantees. More recently, in the case where $\log \pi(U_n|x)$ is concave, non-asymptotic guarantees where proven for Langevin Monte Carlo with a running time that depends only polynomially on the dimension of the parameter θ; see [69–74]. Such results are usually not available for exact MCMC methods.

The implementation of the classical Metropolis–Hastings algorithm requires being able to compute the ratio $\mathcal{L}(\theta; x)/\mathcal{L}(\theta'|x)$ for any θ, θ'. In some models with complex likelihoods, or with intractable normalization constants, this is not possible. This led to a new direction, that is approximations of this likelihood ratio. A surprising and beautiful fact is that, if each likelihood is computed by an unbiased Monte Carlo estimator, the algorithm remains exact: this was studied under the name pseudo-marginal MCMC in [75]. Still, it sometimes requires much work to get unbiased estimates [76,77], when possible at all. Some authors proposed more general approximations of the likelihood ratio, leading to non-exact algorithms. References [78–81] proposed estimators based on subsampling when the data x are too large. Reference [82] proposed an estimator of the likelihood ratio when the likelihood has intractable constants, as in the exponential random graph model, and proved that, even if the resulting MCMC is inexact, it remains asymptotically close to the exact chain. A further theory was developed in [83–85]. More on MCMC for big data can be found in [86].

Finally, the ABC (Approximate Bayesian Computation) algorithm was proposed in population genetics for models where the likelihood is far too complex to be computed, but where it is relatively easy to sample from it [87,88]. It became extremely popular in some applications; we refer the reader to the survey [89], to Section 3 in [15], and more recently, to the book [90]. Some theoretical results were proven in [91]; we also refer the reader to [92–94] for some recent advances.

3.2. Asymptotic Approximations

Laplace's method provides a Gaussian approximation of the posterior centered on the Maximum Likelihood Estimator (MLE) and whose covariance matrix is the inverse of the Fisher information. This approximation can be theoretically justified in parametric models under appropriate regularity conditions thanks to the Bernstein–von Mises theorem. We refer the reader to Chapter 13 in [95] for a complete statement of this result. Integrated Nested Laplace Approximations (INLA) indeed became very popular in Gaussian latent models to compute approximations of the posterior marginals [96].

The extension of the Bernstein–von Mises theorem to nonparametric or semiparametric models is a quite technical and important research direction; see for example [97–101] and Chapter 10 in the monograph [102]. It is important to keep in mind that even in parametric models, when the assumptions of the theorem are not met, Laplace approximation can be wrong. The asymptotic of the posterior in such models was studied in detail in [103].

3.3. Approximations via Optimization

A huge number of methods are based on the idea of using optimization algorithms to find the best approximation of $\pi(\cdot|x)$, or $\tilde{\pi}(\cdot|x)$, in a set of probability distributions $\mathcal{Q}$ fixed by the statistician. The difference between the various methods is in the choice of the criterion used to define the "best" approximation. The set $\mathcal{Q}$ can be parametric (e.g., Gaussian distributions, inspired by Laplace's method) or not, the choice being prescribed by the feasibility of the optimization problem.

Variational approximations are based on the Kullback–Leibler divergence KL:

$$\hat{\pi}(\theta|x) = \operatorname*{argmin}_{q\in\mathcal{Q}} KL(q||\pi(\cdot|x)) \tag{3}$$

$$= \operatorname*{argmin}_{q\in\mathcal{Q}} \left\{ \mathbb{E}_{\theta\sim q}[-\log \mathcal{L}(\theta;x)] + KL(q||\pi) \right\}, \tag{4}$$

where we remind that $KL(q||p) = \int \log(\mathrm{d}q/\mathrm{d}p)\mathrm{d}p$ when q is absolutely continuous with respect to p, and $KL(q||p) = +\infty$ otherwise. We refer the reader to the seminal papers [104,105], to the tutorial [106], and to the recent review of the huge literature on variational approximations [107]. Note that the approximation used in [108] in the early days of neural networks can also be interpreted as a variational approximation. Besides the aforementioned applications to recommender systems and to deep learning, variational inference was successfully used in network data analysis [109], economics and econometrics [110–113], finance [114], natural language processing [115], and video processing [116], among others. A huge range of optimization algorithm were used, from the coordinate-wise optimization in the original publications to message passing [117], the gradient and stochastic gradient algorithm [27,115,118], and the natural gradient [119]. The convexity and smoothness of the minimization problem were discussed in [120]. The scope of these methods was extended to models with intractable likelihood in [121]. Reference [122] pointed out a connection between (4) and PAC-Bayes bounds, which led to the first generalization error bounds for variational inference for some Gibbs posteriors, as in (2). The analysis was extended to various settings, including regular posteriors, as in (1), by [123–131]. In particular, Reference [132] proved that variational inference leads to the optimal estimation of some classes of functions with deep learning. Note that even when $\mathcal{Q}$ is the set of all Gaussian distributions on the parameter space, the approximation can be very different from the Laplace approximation. Indeed, Reference [129] contains an example of a mixture model where the MLE is not consistent, but Gaussian variational inference is.

The choice of the Kullback–Leibler divergence in (3) and (4) was initially motivated by the tractability of the computational program to which it leads. Recently, many authors questioned that choice and proposed extended definitions of variational inference using other divergences; for a presentation of the most popular divergences in statistics, see the introduction to information geometry [133]. Note that if we replace KL by another divergence, (3) and (4) are in general no longer equivalent, which leads to two possible ways to extend the definition. Reference [134] extended (3) by replacing the KL term by a Rényi divergence, and Reference [135] used the χ^2 divergence. However, Reference [136] discussed the computational difficulties induced by these changes, which might outweigh the benefits. Reference [137] discussed other criteria, including the Wasserstein distance, and provided some theoretical guarantees. On the other hand, References [138–141] proposed to use more general divergences in (3). This can be related to the generalized exponential family of [142] and the PAC-Bayes bounds in [143,144].

The very popular Expectation Propagation algorithm (EP) was introduced by [145]. EP can be interpreted as the minimization of the reverse KL, $KL(\pi(\cdot|x)||q))$, instead of (3). This was detailed in [146], where the author also proposed an extension with α-divergences called power EP. Algorithmic issues were discussed in [147] and by [148], who proposed stochastic optimization methods. A first theoretical analysis of EP was proposed in [149]. Let us mention that the textbook [150], which is a generalist introduction to machine learning, contains a full chapter entirely devoted to a pedagogical introduction to variational approximations and EP. The paper [151] focuses on the application of EP to hierarchical models, but also contains a very nice introduction to EP and the conditions ensuring its stability.

Finally, let us mention approximations by discrete distributions, of the form $q = \frac{1}{M}\sum_{i=1}^{M}\delta_{\theta_i}$ where δ_x is the Dirac mass at x. Note that this is typically the kind of approximation provided by the MCMC and sequential Monte Carlo methods, but in these methods, the θ_i are sampled. It is also possible to try to minimize a distance criterion between q and $\pi(\cdot|x)$. Unfortunately, when $\pi(\cdot|x)$ is continuous, both $KL(\pi(\cdot|x)||q)) = KL(q||\pi(\cdot|x))) = +\infty$, so it is not possible to use variational inference or EP in this case. An energy based criterion was proposed in [152]. Reference [153] proposed to use Stein divergences between q and $\pi(\cdot|x)$, and the technique became quite successful [154–156]. Another possible research direction is to use the Wasserstein distance [157].

4. Scope of This Special Issue

The objective of this Special Issue is to provide the latest advances in approximate Monte Carlo methods and in approximations of the posterior: the design of efficient algorithms, the study of the statistical properties of these algorithms, and challenging applications.

Funding: This research received no external funding.

Acknowledgments: The author gratefully thanks Emtiyaz Khan (RIKEN AIP) and Nicolas Chopin (ENSAE Paris) for useful comments.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ABC	Approximate Bayesian Computation
EP	Expectation Propagation
MALA	Monte Carlo Adjusted Langevin Algorithm
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimator
PAC	Probably Approximately Correct
VaR	Value at Risk

References

1. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092. [CrossRef]
2. Geman, S.; Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, *6*, 721–741. [CrossRef] [PubMed]
3. Casella, G.; George, E.I. Explaining the Gibbs sampler. *Am. Stat.* **1992**, *46*, 167–174.
4. Duane, S.; Kennedy, A.D.; Pendleton, B.J.; Roweth, D. Hybrid Monte Carlo. *Phys. Lett. B* **1987**, *195*, 216–222. [CrossRef]
5. Neal, R. *Bayesian Learning for Neural Networks*; Springer Lecture Notes in Statistics; Springer: Berlin/Heidelberg, Germany, 1999; Volume 118.

6. Gilks, W.R.; Roberts, G.O.; Sahu, S.K. Adaptive Markov chain monte carlo through regeneration. *J. Am. Stat. Assoc.* **1998**, *93*, 1045–1054. [CrossRef]
7. Atchade, Y.; Fort, G.; Moulines, E.; Priouret, P. Adaptive Markov chain Monte Carlo: Theory and methods. In *Bayesian Time Series Models*; Cambridge University Press: Cambridge, UK, 2011; pp. 32–51.
8. Roberts, G.O.; Rosenthal, J.S. Examples of adaptive MCMC. *J. Comput. Graph. Stat.* **2009**, *18*, 349–367. [CrossRef]
9. Besag, J.; Green, P.; Higdon, D.; Mengersen, K. Bayesian Computation and Stochastic Systems. *Stat. Sci.* **1995**, *10*, 3–41. [CrossRef]
10. Andrieu, C.; De Freitas, N.; Doucet, A.; Jordan, M.I. An introduction to MCMC for machine learning. *Mach. Learn.* **2003**, *50*, 5–43. [CrossRef]
11. Brooks, S.; Gelman, A.; Jones, G.; Meng, X.L. (Eds.) *Handbook of Markov Chain Monte Carlo*; CRC Press: Boca Raton, FL, USA, 2011.
12. Robert, C.; Casella, G. *Monte Carlo Statistical Methods*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
13. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.B. *Bayesian Data Analysis*, 3rd ed.; CRC Press: Boca Raton, FL, USA, 2013.
14. Chopin, N.; Gadat, S.; Guedj, B.; Guyader, A.; Vernet, E. On some recent advances on high dimensional Bayesian statistics. *ESAIM Proc. Surv.* **2015**, *51*, 293–319. [CrossRef]
15. Green, P.J.; Łatuszyński, K.; Pereyra, M.; Robert, C.P. Bayesian computation: A summary of the current state, and samples backwards and forwards. *Stat. Comput.* **2015**, *25*, 835–862. [CrossRef]
16. Meyn, S.P.; Tweedie, R.L. *Markov Chains and Stochastic Stability*; Springer: Berlin/Heidelberg, Germany, 2012.
17. Douc, R.; Moulines, E.; Priouret, P.; Soulier, P. *Markov Chains*; Springer: Berlin, Germany, 2018.
18. Joulin, A.; Ollivier, Y. Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Probab.* **2010**, *38*, 2418–2442. [CrossRef]
19. Kalman, R.E. A New Approach to Linear Filtering and Prediction Problems. *Trans. ASM J. Basic Eng.* **1960**, *82*, 35–45. [CrossRef]
20. Doucet, A.; De Freitas, N.; Gordon, N. (Eds.) *Sequential Monte Carlo Methods in Practice*; Springer: Berlin/Heidelberg, Germany, 2001.
21. Chopin, N.; Papaspiliopoulos, O. *An Introduction to Sequential Monte Carlo*; Springer: Berlin/Heidelberg, Germany, 2020.
22. Naesseth, C.A.; Lindsten, F.; Schön, T.B. Elements of Sequential Monte Carlo. *Found. Trends Mach. Learn.* **2019**, *12*, 307–392. [CrossRef]
23. Bennett, J.; Lanning, S. The Netflix prize. In Proceedings of the KDD Cup and Workshop, Los Gatos, CA, USA, 12 August 2005; pp. 35–38.
24. Lim, Y.J.; Teh, Y.W. Variational Bayesian approach to movie rating prediction. In Proceedings of the KDD Cup and Workshop, Jose, CA, USA, 12 August 2007; pp. 15–21.
25. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
26. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1050–1059.
27. Mandt, S.; Hoffman, M.D.; Blei, D.M. Stochastic gradient descent as approximate Bayesian inference. *J. Mach. Learn. Res.* **2017**, *18*, 1–35.
28. Maddox, W.J.; Izmailov, P.; Garipov, T.; Vetrov, D.P.; Wilson, A.G. A simple baseline for Bayesian uncertainty in deep learning. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2019; pp. 13153–13164.
29. Osawa, K.; Swaroop, S.; Khan, M.E.; Jain, A.; Eschenhagen, R.; Turner, R.E.; Yokota, R. Practical deep learning with Bayesian principles. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 4287–4299.
30. Neal, R.M. Sampling from multimodal distributions using tempered transitions. *Stat. Comput.* **1996**, *6*, 353–366. [CrossRef]
31. Friel, N.; Pettitt, A.N. Marginal likelihood estimation via power posteriors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2008**, *70*, 589–607. [CrossRef]

32. Walker, S.; Hjort, N.L. On Bayesian consistency. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2001**, *63*, 811–821. [CrossRef]
33. Grünwald, P.D.; Van Ommen, T. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.* **2017**, *12*, 1069–1103. [CrossRef]
34. Bhattacharya, A.; Pati, D.; Yang, Y. Bayesian fractional posteriors. *Ann. Stat.* **2019**, *47*, 39–66. [CrossRef]
35. Bissiri, P.G.; Holmes, C.C.; Walker, S.G. A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2016**, *78*, 1103–1130. [CrossRef] [PubMed]
36. Shawe-Taylor, J.; Williamson, R.C. A PAC analysis of a Bayesian estimator. In Proceedings of the Tenth Annual Conference on Computational Learning Theory, Nashville, TN, USA, 6–9 July 1997; pp. 2–9.
37. McAllester, D.A. Some PAC-Bayesian theorems. *Mach. Learn.* **1999**, *37*, 355–363. [CrossRef]
38. Catoni, O. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*; Monograph Series 56; IMS Lecture Notes: Beachwood, OH, USA, 2007.
39. Van Erven, T. PAC-Bayes mini-tutorial: A continuous union bound. *arXiv* **2014**, arXiv:1405.1580.
40. McAllester, D.A. A PAC-Bayesian tutorial with a dropout bound. *arXiv* **2013**, arXiv:1307.2118.
41. Catoni, O. *Statistical Learning Theory and Stochastic Optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI-2001*; Springer: Berlin/Heidelberg, Germany, 2004.
42. Zhang, T. From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *Ann. Stat.* **2006**, *34*, 2180–2210. [CrossRef]
43. Grünwald, P.D.; Mehta, N.A. A tight excess risk bound via a unified PAC-Bayesian–Rademacher–Shtarkov–MDL complexity. *Conf. Algorithmic Learn.* **2019**, *98*, 433–465.
44. Deng, L.; Hinton, G.; Kingsbury, B. New types of deep neural network learning for speech recognition and related applications: An overview. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 8599–8603.
45. Neyshabur, B.; Bhojanapalli, S.; McAllester, D.; Srebro, N. Exploring generalization in deep learning. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5947–5956.
46. Dziugaite, G.K.; Roy, D. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv* **2017**, arXiv:1703.11008.
47. Dziugaite, G.K.; Roy, D. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors. In Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 1377–1386.
48. Amit, R.; Meir, R. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 205–214.
49. Nozawa, K.; Sato, I. PAC-Bayes Analysis of Sentence Representation. *arXiv* **2019**, arXiv:1902.04247.
50. Pitas, K. Better PAC-Bayes bounds for deep neural networks using the loss curvature. *arXiv* **2019**, arXiv:1909.03009.
51. Rivasplata, O.; Tankasali, V.M.; Szepesvari, C. PAC-Bayes with backprop. *arXiv* **2019**, arXiv:1908.07380 .
52. Guedj, B. A primer on PAC-Bayesian learning. In Proceedings of the Second Congress of the French Mathematical Society, Lille, France, 4–8 June 2018.
53. Vovk, V.G. Aggregating strategies. In Proceedings of the Third Annual Workshop on Computational Learning Theory, Rochester, NY, USA, 6–8 August 1990.
54. Littlestone, N.; Warmuth, M.K. The weighted majority algorithm. *Inf. Comput.* **1994**, *108*, 212–261. [CrossRef]
55. Cesa-Bianchi, N.; Lugosi, G. *Prediction, Learning, and Games*; Cambridge University Press: Cambridge, UK, 2006.
56. Besson, R.; Le Pennec, E.; Allassonnière, S. Learning from both experts and data. *Entropy* **2019**, *21*, 1208. [CrossRef]
57. Seldin, Y.; Auer, P.; Shawe-Taylor, J.S.; Ortner, R.; Laviolette, F. PAC-Bayesian analysis of contextual bandits. In Proceedings of the Advances in Neural Information Processing Systems, Granada, Spain, 12–14 December 2011; pp. 1683–1691.
58. Bubeck, S.; Cesa-Bianchi, N. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Found. Trends Mach. Learn.* **2012**, *5*, 1–122. [CrossRef]

59. Leung, G.; Barron, A.R. Information theory and mixing least-squares regressions. *IEEE Trans. Inf. Theory* **2006**, *52*, 3396–3410. [CrossRef]
60. Jiang, W.; Tanner, M.A. Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Stat.* **2008**, *36*, 2207–2231. [CrossRef]
61. Dalalyan, A.S.; Tsybakov, A.B. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. Syst. Sci.* **2012**, *78*, 1423–1443. [CrossRef]
62. Suzuki, T. PAC-Bayesian bound for Gaussian process regression and multiple kernel additive model. In Proceedings of the 25th Annual Conference on Learning Theory, Edinburgh, Scotland, 25–27 June 2012; pp. 8.1–8.20.
63. Dalalyan, A.S.; Salmon, J. Sharp oracle inequalities for aggregation of affine estimators. *Ann. Stat.* **2012**, *40*, 2327–2355. [CrossRef]
64. Dalalyan, A.S.; Grappin, E.; Paris, Q. On the exponentially weighted aggregate with the Laplace prior. *Ann. Stat.* **2018**, *46*, 2452–2478. [CrossRef]
65. Syring, N.; Hong, L.; Martin, R. Gibbs posterior inference on Value-At-Risk. *Scand. Actuar. J.* **2019**, *7*, 548–557. [CrossRef]
66. Ermak, D.L. A computer simulation of charged particles in solution. I. Technique and equilibrium properties. *J. Chem. Phys.* **1975**, *62*, 4189–4196. [CrossRef]
67. Rossky, P.J.; Doll, J.D.; Friedman, H.L. Brownian dynamics as smart Monte Carlo simulation. *J. Chem. Phys.* **1978**, *69*, 4628–4633. [CrossRef]
68. Roberts, G.O.; Tweedie, R.L. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **1996**, *2*, 341–363. [CrossRef]
69. Dalalyan, A.S. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In Proceedings of the 2017 Conference on Learning Theory, PMLR, Amsterdam, The Netherlands, 7–10 July 2017; pp. 678–689.
70. Raginsky, M.; Rakhlin, A.; Telgarsky, M. Non-convex learning via Stochastic Gradient Langevin Dynamics: A nonasymptotic analysis. In Proceedings of the 2017 Conference on Learning Theory, PMLR, Amsterdam, The Netherlands, 7–10 July 2017; pp. 1674–1703.
71. Cheng, X.; Chatterji, N.S.; Bartlett, P.L.; Jordan, M.I. Underdamped Langevin MCMC: A non-asymptotic analysis. In Proceedings of the 31st Conference on Learning Theory, PMLR, Stockholm, Sweden, 6–9 July 2018; pp. 300–323.
72. Dalalyan, A.S.; Riou-Durand, L.; Karagulyan, A. Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets. *arXiv* **2019**, arXiv:1906.08530.
73. Durmus, A.; Moulines, E. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli* **2019**, *25*, 2854–2882. [CrossRef]
74. Mou, W.; Flammarion, N.; Wainwright, M.J.; Bartlett, P.L. Improved bounds for discretization of Langevin diffusions: Near-optimal rates without convexity. *arXiv* **2019**, arXiv:1907.11331.
75. Andrieu, C.; Roberts, G.O. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Stat.* **2009**, *37*, 697–725. [CrossRef]
76. Lyne, A.M.; Girolami, M.; Atchadé, Y.; Strathmann, H.; Simpson, D. On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Stat. Sci.* **2015**, *30*, 443–467. [CrossRef]
77. Vats, D.; Gonçalves, F.; Łatuszyński, K.; Roberts, G.O. Efficient Bernoulli factory MCMC for intractable likelihoods. *arXiv* **2020**, arXiv:2004.07471.
78. Korattikara, A.; Chen, Y.; Welling, M. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 181–189.
79. Huggins, J.; Campbell, T.; Broderick, T. Coresets for Scalable Bayesian Logistic Regression. In Proceedings of the Advances in Neural Information Processing Systems 29, Barcelona, Spain, 5–10 December 2016; pp. 4080–4088.
80. Quiroz, M.; Kohn, R.; Villani, M.; Tran, M.N. Speeding up MCMC by efficient data subsampling. *J. Am. Stat. Assoc.* **2018**, *114*, 831–843. [CrossRef]
81. Maire, F.; Friel, N.; Alquier, P. Informed sub-sampling MCMC: Approximate Bayesian inference for large datasets. *Stat. Comput.* **2019**, *29*, 449–482. [CrossRef]

82. Alquier, P.; Friel, N.; Everitt, R.; Boland, A. Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Stat. Comput.* **2016**, *26*, 29–47. [CrossRef]
83. Medina-Aguayo, F.J.; Lee, A.; Roberts, G.O. Stability of noisy metropolis–hastings. *Stat. Comput.* **2016**, *26*, 1187–1211. [CrossRef] [PubMed]
84. Rudolf, D.; Schweizer, N. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli* **2018**, *24*, 2610–2639. [CrossRef]
85. Stoehr, J.; Benson, A.; Friel, N. Noisy Hamiltonian Monte Carlo for doubly intractable distributions. *J. Comput. Graph. Stat.* **2019**, *28*, 220–232. [CrossRef]
86. Bardenet, R.; Doucet, A.; Holmes, C. On Markov chain Monte Carlo methods for tall data. *J. Mach. Learn. Res.* **2017**, *18*, 1515–1557.
87. Tavaré, S.; Balding, D.; Griffith, R.; Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics* **1997**, *145*, 505–518.
88. Beaumont, M.A.; Zhang, W.; Balding, D.J. Approximate Bayesian computation in population genetics. *Genetics* **2002**, *162*, 2025–2035.
89. Marin, J.-M.; Pudlo, P.; Robert, C.P.; Ryder, R.J. Approximate Bayesian computational methods. *Stat. Comput.* **2012**, *22*, 1167–1180. [CrossRef]
90. Sisson, S.A.; Fan, Y.; Beaumont, M. (Eds.) *Handbook of Approximate Bayesian Computation*; CRC Press: Boca Raton, FL, USA, 2018.
91. Biau, G.; Cérou, F.; Guyader, A. New insights into approximate Bayesian computation. *Ann. De L'IHP Probab. Stat.* **2015**, *51*, 376–403. [CrossRef]
92. Bernton, E.; Jacob, P.E.; Gerber, M.; Robert, C.P. Approximate Bayesian computation with the Wasserstein distance. *J. R. Stat. Soc. Ser. B* **2019**, *81*, 235–269. [CrossRef]
93. Buchholz, A.; Chopin, N. Improving approximate Bayesian computation via quasi-Monte Carlo. *J. Comput. Graph. Stat.* **2019**, *28*, 205–219. [CrossRef]
94. Nguyen, H.D.; Arbel, J.; Lü, H.; Forbes, F. Approximate Bayesian computation via the energy statistic. *IEEE Access* **2020**, *8*, 131683–131698. [CrossRef]
95. Van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 2000.
96. Rue, H.; Martino, S.; Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2009**, *71*, 319–392. [CrossRef]
97. Freedman, D. Wald Lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Stat.* **1999**, *em 27*, 1119–1141. [CrossRef]
98. Boucheron, S.; Gassiat, E. A Bernstein-von Mises theorem for discrete probability distributions. *Electron. J. Stat.* **2009**, *3*, 114–148. [CrossRef]
99. Bickel, P.J.; Kleijn, B.J. The semiparametric Bernstein–von Mises theorem. *Ann. Stat.* **2012**, *40*, 206–237. [CrossRef]
100. Rivoirard, V.; Rousseau, J. Bernstein–von Mises theorem for linear functionals of the density. *Ann. Stat.* **2012**, *40*, 1489–1523. [CrossRef]
101. Castillo, I.; Nickl, R. On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Stat.* **2014**, *42*, 1941–1969. [CrossRef]
102. Ghosal, S.; Van der Vaart, A. *Fundamentals of Nonparametric Bayesian Inference*; Cambridge University Press: Cambridge, UK, 2017.
103. Watanabe, S. *Mathematical Theory of Bayesian Statistics*; CRC Press: Boca Raton, FL, USA, 2018.
104. Attias, H. Inferring parameters and structure of latent variable models byvariational Bayes. In Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, 30 July–1 August 1999; pp. 21–30.
105. Jordan, M.I.; Ghahramani, Z.; Jaakkola, T.S.; Saul, L.K. An introduction to variational methods for graphical models. *Mach. Learn.* **1999**, *37*, 183–233. [CrossRef]
106. Wainwright, M.J.; Jordan, M.I. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **2008**, *1*, 1–305. [CrossRef]
107. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]

108. Hinton, G.E.; Van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In Proceedings of the Sixth Annual Conference on Computational Learning Theory, Santa Cruz, CA, USA, 26–28 July 1993; pp. 5–13.
109. Salter-Townshend, M.; Murphy, T.B. Variational Bayesian inference for the latent position cluster model for network data. *Comput. Stat. Data Anal.* **2013**, *57*, 661–671. [CrossRef]
110. Braun, M.; McAuliffe, J. Variational inference for large-scale models of discrete choice. *J. Am. Stat. Assoc.* **2010**, *105*, 324–335. [CrossRef]
111. Wu, G. Fast and scalable variational Bayes estimation of spatial econometric models for Gaussian data. *Spat. Stat.* **2018**, *24*, 32–53. [CrossRef]
112. Baltagi, B.H.; Bresson, G.; Etienne, J.M. Carbon dioxide emissions and economic activities: A mean field variational Bayes semiparametric panel data model with random coefficients. *Ann. Econ. Stat.* **2019**, *134*, 43–77. [CrossRef]
113. Gefang, D.; Koop, G.; Poon, A. Computationally efficient inference in large Bayesian mixed frequency VARs. *Econ. Lett.* **2020**, *191*, 109120. [CrossRef]
114. Gunawan, D.; Kohn, R.; Nott, D. Variational Approximation of Factor Stochastic Volatility Models. *arXiv* **2020**, arXiv:2010.06738.
115. Hoffman, M.D.; Blei, D.M.; Wang, C.; Paisley, J. Stochastic variational inference. *J. Mach. Learn. Res.* **2013**, *14*, 1303–1347.
116. Li, X.; Zheng, Y. Patch-based video processing: A variational Bayesian approach. *IEEE Trans. Circuits Syst. Video Technol.* **2009**, *19*, 27–40.
117. Winn, J.; Bishop, C.M. Variational Message Passing. *J. Mach. Learn. Res.* **2005**, *6*, 661–694.
118. Broderick, T.; Boyd, N.; Wibisono, A.; Wilson, A.C.; Jordan, M.I. Streaming Variational Bayes. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 1727–1735.
119. Khan, M.E.; Lin, W. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Lauderdale, FL, USA, 20 April 2017; pp. 878–887.
120. Domke, J. Provable smoothness guarantees for black-box variational inference. *arXiv* **2019**, arXiv:1901.08431.
121. Tran, M.N.; Nott, D.J.; Kohn, R. Variational Bayes with intractable likelihood. *J. Comput. Graph. Stat.* **2017**, *26*, 873–882. [CrossRef]
122. Alquier, P.; Ridgway, J.; Chopin, N. On the properties of variational approximations of Gibbs posteriors. *J. Mach. Learn. Res.* **2016**, *17*, 8374–8414.
123. Sheth, R.; Khardon, R. Excess risk bounds for the Bayes risk using variational inference in latent Gaussian models. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 5151–5161.
124. Cottet, V.; Alquier, P. 1-Bit matrix completion: PAC-Bayesian analysis of a variational approximation. *Mach. Learn.* **2018**, *107*, 579–603. [CrossRef]
125. Wang, Y.; Blei, D.M. Frequentist consistency of variational Bayes. *J. Am. Stat. Assoc.* **2019**, *114*, 1147–1161. [CrossRef]
126. Chérief-Abdellatif, B.-E. Consistency of ELBO maximization for model selection. In Proceedings of the 1st Symposium on Advances in Approximate Bayesian Inference, PMLR, Montreal, QC, Canada, 2 December 2018; pp. 11–31.
127. Guha, B.S.; Bhattacharya, A.; Pati, D. Statistical Guarantees and Algorithmic Convergence Issues of Variational Boosting. *arXiv* **2020**, arXiv:2010.09540.
128. Chérief-Abdellatif, B.-E.; Alquier, P.; Khan, M.E. A Generalization Bound for Online Variational Inference. *arXiv* **2019**, arXiv:1904.03920.
129. Alquier, P.; Ridgway, J. Concentration of tempered posteriors and of their variational approximations. *Ann. Stat.* **2020**, *48*, 1475–1497. [CrossRef]
130. Yang, Y.; Pati, D.; Bhattacharya, A. α-variational inference with statistical guarantees. *Ann. Stat.* **2020**, *48*, 886–905. [CrossRef]
131. Zhang, F.; Gao, C. Convergence rates of variational posterior distributions. *Ann. Stat.* **2020**, *48*, 2180–2207. [CrossRef]
132. Chérief-Abdellatif, B.E. Convergence Rates of Variational Inference in Sparse Deep Learning. *arXiv* **2019**, arXiv:1908.04847.

133. Nielsen, F. An elementary introduction to information geometry. *Entropy* **2020**, *22*, 1110. [CrossRef]
134. Li, Y.; Turner, R.E. Rényi divergence variational inference. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 1073–1081.
135. Dieng, A.B.; Tran, D.; Ranganath, R.; Paisley, J.; Blei, D. Variational inference via χ-upper bound minimization. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 2732–2741.
136. Geffner, T.; Domke, J. On the Difficulty of Unbiased Alpha Divergence Minimization. *arXiv* **2019**, arXiv:2010.09541.
137. Huggins, J.; Kasprzak, M.; Campbell, T.; Broderick, T. Validated Variational Inference via Practical Posterior Error Bounds. In Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, Sicily, Italy, 3 June 2020; pp. 1792–180.
138. Reid, M.D.; Frongillo, R.M.; Williamson, R.C.; Mehta, N. Generalized mixability via entropic duality. In Proceedings of the 28th Conference on Learning Theory, Paris, France, 3–6 July 2015; pp. 1501–1522.
139. Knoblauch, J.; Jewson, J.; Damoulas, T. Generalized variational inference: Three arguments for deriving new posteriors. *arXiv* **2019**, arXiv:1904.02063.
140. Alemi, A.A. Variational Predictive Information Bottleneck. In Proceedings of the 2nd Symposium Advances Approximate Bayesian Inference, PMLR, Vancouver, BC, Canada, 8 December 2019; pp. 1–6.
141. Alquier, P. Non-exponentially weighted aggregation: Regret bounds for unbounded loss functions. *arXiv* **2020**, arXiv:2009.03017.
142. Grunwald, P.D.; Dawid, A.P. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Ann. Stat.* **2004**, *32*, 1367–1433. [CrossRef]
143. Bégin, L.; Germain, P.; Laviolette, F.; Roy, J.-F. PAC-Bayesian bounds based on the Rényi divergence. In Proceedings of the 19th International Conference Artificial Intelligence and Statistics PMLR, Cadiz, Spain, 9–11 May 2016; pp. 435–444.
144. Alquier, P.; Guedj, B. Simpler PAC-Bayesian bounds for hostile data. *Mach. Learn.* **2018**, *107*, 887–902. [CrossRef]
145. Minka, T.P. Expectation propagation for approximate Bayesian inference. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, Seattle, WA, USA, 2–5 August 2001; pp. 362–369.
146. Minka, T. *Divergence Measures and Message Passing*; Technical Report; Microsoft Research: Redmond, DC, USA, 2005.
147. Seeger, M.; Nickisch, H. Fast convergent algorithms for expectation propagation approximate Bayesian inference. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 652–660.
148. Li, Y.; Hernández-Lobato, J.M.; Turner, R.E. Stochastic expectation propagation. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 2323–2331.
149. Dehaene, G.P.; Barthelmé, S. Bounding errors of expectation-propagation. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 244–252.
150. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
151. Vehtari, A.; Gelman, A.; Sivula, T.; Jylänki, P.; Tran, D.; Sahai, S.; Blomstedt, P.; Cunningham, J.P.; Schiminovich, D.; Robert, C.P. Expectation Propagation as a Way of Life: A Framework for Bayesian Inference on Partitioned Data. *J. Mach. Learn. Res.* **2020**, *21*, 1–53.
152. Joseph, V.R.; Dasgupta, T.; Tuo, R.; Wu, C. Sequential exploration of complex surfaces using minimum energy designs. *Technometrics* **2015**, *57*, 64–74. [CrossRef]
153. Liu, Q.; Wang, D. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2378–2386.
154. Chen, W.Y.; Mackey, L.; Gorham, J.; Briol, F.-X.; Oates, C.J. Stein points. In Proceedings of the 35th International Conference on Machine Learningc PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 843–852.
155. Chen, W.Y.; Barp, A.; Briol, F.-X.; Gorham, J.; Girolami, M.; Mackey, L.; Oates, C. Stein Point Markov Chain Monte Carlo. In Proceedings of the 36th International Conference on Machine Learningc PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 1011–1021.

156. Kassab, R.; Simeone, O. Federated Generalized Bayesian Learning via Distributed Stein Variational Gradient Descent. *arXiv* **2020**, arXiv:2009.06419.
157. Nitanda, A.; Suzuki, T. Stochastic Particle Gradient Descent for Infinite Ensembles. *arXiv* **2017**, arXiv:1712.05438.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.