Towards the Statistical Foundations of Variational Inference

Marco Avella, José Montiel Olea, Cynthia Rush

November 23, 2020

Project Overview. Analyzing and learning from large datasets coming from scientific fields like astrophysics, computational genetics, finance, and image processing, are fundamental challenges of modern statistics and data science. A core problem is that the statistical models used to describe these large data sets typically make use of a large number of hidden parameters (or latent variables) to characterize the data generating process.

Variational inference is a modern machine learning technique for conducting inference about the hidden parameters of high-dimensional models that provides a computationally-efficient alternative to more classic statistical techniques (such as Markov-Chain Monte-Carlo methods) and its faster computation allows for scaling with modern datasets. However, variational inference lacks the sound theoretical framework – especially from a statistical point-of-view – that supports many of the more classical, but ultimately computationally infeasible, methods. For this reason, despite a tremendous impact in machine learning, the use of variational methods in fields such as statistics and economics has been limited.

This project aims to provide theoretical foundations for current variational methods that are in wide application in machine learning and to develop novel variational techniques supported by solid mathematical justification. This will be done by using insights from the areas of decision theory, robust statistics, penalized estimation, message passing, and statistical physics which are the areas of expertise of the PIs and Co-PI. In this project, we want to leverage the large amount of literature in the areas just described to study variational inference.

Methodologies. The basic variational inference framework is as follows. The statistician models observed data $\mathbf{x} = (x_1, \dots, x_n)$ through a joint probability distribution $p(\mathbf{x}, \mathbf{z})$ with some unobserved variables $\mathbf{z} = (z_1, \dots, z_m)$. As usual, we refer to $p(\mathbf{x}|\mathbf{z})$ as the likelihood and to $p(\mathbf{z})$ as the prior. The statistician is interested in estimation and inference about the hidden variables $\mathbf{z}$ given the observed data $\mathbf{x}$, and in particular, the key object of interest the posterior distribution $p(\mathbf{z}|\mathbf{x})$.

In many interesting problems, especially those where the latent variables are high-dimensional, one is unable to compute this posterior analytically or to approximate it using Monte-Carlo Markov-Chain methods. Variational inference allows the statistician to approximate the posterior distribution by solving an optimization problem. In more detail, doing so involves choosing a distribution $q(\mathbf{z})$, usually assumed to belong to some class of possible distributions. Variational inference aims to find the 'best' approximating distribution by minimizing the Kullback-Liebler (KL) divergence between possible $q(\mathbf{z})$ and $p(\mathbf{z}|\mathbf{x})$, defined as

$$D_{\mathsf{KL}}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) = \int q(\mathbf{z}) \log\left(\frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})}\right) d\mathbf{z}. \tag{1}$$

A key insight of the variation framework is that minimizing the above KL divergence is equivalent to maximizing the Evidence Lower Bound (ELBO) denoted as $\mathcal{L}(q)$ and defined by

$$\mathcal{L}(q) = \int q(\mathbf{z}) \log \left(\frac{p(\mathbf{z}, \mathbf{x})}{q(\mathbf{z})} \right) d\mathbf{z} = \int q(\mathbf{z}) \log p(\mathbf{x}|\mathbf{z}) d\mathbf{z} - D_{\mathsf{KL}}(q(\mathbf{z})||p(\mathbf{z})), \tag{2}$$

where we notice that the KL divergence in (2) is between the variational distribution, $q(\mathbf{z})$, and the *prior* distribution on the latent variables. Thus, the variational approximation can be written as

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z}) \in \mathcal{Q}}{\arg\max} \{\mathcal{L}(q)\}, \tag{3}$$

where the optimization is over all distributions $q(\mathbf{z})$ belonging to a class $\mathcal{Q}$.

In this project, we propose to study the variational methods through the lens of penalized (*randomized*) estimators, where one optimizes over distributions for the parameters describing the data. To get a feel for what we mean, notice that the ELBO objective function in (2) has exactly this form: it involves a data-fitting term (the average log-likelihood, $\int q(\mathbf{z}) \log p(\mathbf{x}|\mathbf{z}) d\mathbf{z}$) and a regularization or penalization term, $-D_{\mathsf{KL}}(q(\mathbf{z})||p(\mathbf{z}))$, that forces the distribution over parameters to be close to a baseline (e.g., a prior). In particular, this the proposed work will study approximations of the form

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z}) \in \mathcal{Q}}{\arg\max} \left\{ \int q(\mathbf{z}) \log p(\mathbf{x}|\mathbf{z}) d\mathbf{z} - \beta D(q(\mathbf{z})||p(\mathbf{z})) \right\}, \tag{4}$$

where we have generalized the optimization of the ELBO in (3) in a few ways by introducing a regularization parameter $\beta > 0$ that controls the amount of penalization and also by replacing the KL divergence with some general divergence $D(q(\mathbf{z})||p(\mathbf{z}))$. The non-trivial aspect of this penalized estimation problem, however, is that one is no longer optimizing over vectors, as in standard statistical estimation, but now optimizing over a class of probability distributions.

Related Literature. The optimization problem in (4) (with KL divergence) has been the subject of recent work in the representation learning literature; see the definition of β-Variational Auto Encoder in [4] and [8]. It has been argued that values of $\beta > 1$ result in 'disentangled' latent representations (which can be defined, loosely speaking, as distributions where 'single latent units are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors').

The optimization problem in (4) has been studied in detail in axiomatic decision theory; see, for example, the *multiplier preferences* introduced in [7] and their axiomatization in [12]. More generally, the objective function in (4) with an arbitrary divergence function is analogous to the so-called *divergence preferences* studied in [10]. Most of the decision theory literature has focused on understanding the causes and consequences of using these different representations. We think this literature could be potentially useful in understanding the role of the different divergence functions in penalizations, as well as the multiplier parameter β.

Objectives. To illustrate the usefulness of framing the variational inference problem in terms of penalized estimation, we study a few special cases of the optimization scheme in (4) and discuss their connections to robust decision theory, robust statistics, and the current variational inference literature. First we consider the case where $\mathcal{Q}$ contains all possible distributions – so there are no assumed constraints on the approximate distribution – and the divergence considered is the standard KL divergence, $D(q(\mathbf{z})||p(\mathbf{z})) = D_{\mathsf{KL}}(q(\mathbf{z})||p(\mathbf{z}))$. First, for $\beta = 0$, it is straightforward to see that (4) returns an approximating distribution that is the point mass at the MLE estimator, in other words, $q^*(\mathbf{z}) = \arg\max_{\mathbf{z}} \log p(\mathbf{x}|\mathbf{z})$. Now consider $\beta \in (0, \infty)$. In this case, Proposition 1.4.2 in [6] shows that the optimization in (4) returns an exponentially tilted approximating distribution, namely $q^*(\mathbf{z}) \propto \exp\{\frac{1}{\beta}\log p(\mathbf{x}|\mathbf{z})\}p(\mathbf{z}) = (p(\mathbf{x}|\mathbf{z}))^{1/\beta}p(\mathbf{z})$ where the proportionality constant is $\int p(\mathbf{x}|\mathbf{z}))^{1/\beta}p(\mathbf{z})d\mathbf{z}$. Notice that for $\beta = 1$, this says that $q^*(\mathbf{z}) \propto p(\mathbf{x}, \mathbf{z})$ giving $q^*(\mathbf{z}) = p(\mathbf{z}|\mathbf{x})$, which is clearly the solution that is needed since in this case the optimization in (4) is equivalent to minimizing the KL divergence in (1).

We note that the solution discussed above, namely $q^*(\mathbf{z}) \propto (p(\mathbf{x}|\mathbf{z}))^{1/\beta}p(\mathbf{z})$ when $\beta > 1$, has been studied in the statistics literature under various names including the *power*, *fractional*, or *tempered* posteriors. The work of [2] shows that these posteriors have better frequentist properties under weaker assumptions than the usual posterior $p(\mathbf{z}|\mathbf{x})$.[1] Tempered posteriors are natural approximations of the so-called *coarsened posteriors*, which can be thought of as robust counterparts of posteriors; see [11].[2]

Using this connection between coarsened posteriors and tempered posteriors, we can see that the penalized problem (4) induces some *statistical robustness* to the β-variational approximation of the posterior distribution. In particular, using the language of robust statistics, it is possible to show—using Theorem 3.8 in [11]—that the coarsened posterior is *qualitatively robust* (as defined in Section 1.3 of [9]), unlike the standard posterior. Loosely speaking, this result tells us that inference based on a power, fractional, or tempered posterior—which can be shown to solve the variational problem in (3) when $\mathcal{Q}$ is unrestricted—cannot be too affected by the presence of a few outliers. Some of this robustness properties have been informally documented in the representation learning literature, but we would like to establish a more explicit set of results and deepen our understanding of the interplay between robust statistics and variational inference.[3]

A usual assumption in the variational literature is the mean field assumption, where the constraining set $\mathcal{Q} = \mathcal{Q}_{\mathsf{MF}}$ assumes a factorization across the latent variables: $q(\mathbf{z}) \in \mathcal{Q}_{\mathsf{MF}}$ means that $q(\mathbf{z}) = \prod_{k=1}^m q_k(z_k)$. Now we consider the optimization scheme in (4) where $\mathcal{Q} = \mathcal{Q}_{\mathsf{MF}}$ and we again assume $D(q(\mathbf{z})||p(\mathbf{z})) = D_{KL}(q(\mathbf{z})||p(\mathbf{z}))$ is the standard KL divergence. In this case, we can show that a necessary condition for a variation distribution $q(\mathbf{z}) = \prod_{k=1}^m q_k(z_k)$ to maximize ELBO is that for any index $k \in \{1, 2, \ldots, m\}$ the distribution $q_k(z_k)$, given the other variational factors $q_\ell(z_\ell)$ for $\ell \neq k$, solves the problem

$$q^*_k(z_k) = \underset{q(z)}{\arg\min} \left\{ -\int q(z) \left(\mathbb{E}_{q_{-k}} \left[\log \left(\frac{p(\mathbf{x}|\mathbf{z})(p(\mathbf{z}))^\beta}{(p_k(z_k))^\beta} \right) \right] \right) dz + \beta D_{\mathsf{KL}}(q(z)||p_k(z_k)) \right\},$$

where we use the notation $\mathbb{E}_{q_{-k}}$ to mean the expectation with respect to $q_\ell(z_\ell)$ for $\ell \neq k$, and $p_k(z_k)$

[1] This motivated [1] to directly consider variational approximations to tempered posteriors, i.e. they replace the usual posterior in the divergence in (1) with a tempered posterior. Notice that in our penalized estimation framework, the tempered posterior is, in fact, the *solution* found by maximizing the ELBO (generalized with the β coefficient).

[2] Coarsened posteriors are constructed to take into account the fact that the statistical models that one uses to represent the data are merely idealized distributions of the actual data. Here the observed data $\mathbf{x}$ is viewed as a perturbed version of some idealized data $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ and instead of considering posteriors $p(\mathbf{z}|\mathbf{x} = \tilde{\mathbf{x}})$, the coarsened posterior considers $p(\mathbf{z}|d(\hat{F}_{\mathbf{x}}, \hat{F}_{\tilde{\mathbf{x}}}) < r)$ where $d(\cdot, \cdot)$ is some discrepancy and $\hat{F}_{\mathbf{x}}(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{x_i \leq x\}$ is the empirical distribution function of $\mathbf{x}$.

[3] A few examples of the questions we have in mind are as follows. First, we would like to understand whether solutions to (4) when $\mathcal{Q}$ is a strict subset of the space of probability measures inherit some of the robustness of tempered posteriors. Relatedly, one could try to understand whether variational approximations to tempered posteriors studied by [1] also have robustness properties. Finally, another interesting direction would be refine our understanding of the robustness properties of tempered posterior inference by deriving expressions for the influence function or the breakdown point, which are tools use to quantify the degree of robustness of statistical functionals. In a reweighted Bayesian model closely related to tempered posteriors, [13] studied the influence function of the posterior mean. While their result shows that their method is robust towards unlikely 'outlying' observations, in general it is also interesting to study the influence function over all the sample space. Indeed, the influence function is often used to assess the asymptotic efficiency of M-estimators.

is the k^{th} marginal of $p(\mathbf{z})$. The above implies

$$q_k^*(z_k) \propto \exp\left\{\mathbb{E}_{q_{-k}}\left[\log\left(\frac{p(\mathbf{x}|\mathbf{z})(p(\mathbf{z}))^{\beta}}{(p_k(z_k))^{\beta}}\right)\right]\right\} p_k(z_k) = \exp\left\{\mathbb{E}_{q_{-k}}\left[\log\left((p(\mathbf{x}|\mathbf{z}))^{1/\beta}p(\mathbf{z})\right)\right]\right\}.$$

Notice that when $\beta = 1$ this returns the usual mean-field variational approximation discussed, for example, in [3].

We hope that by studying the variational methods through the lens of a penalized statistical estimation procedure, we will be able to provide a better understanding of the connections between variational procedures and robust statistics – a field that provides a framework for understanding how sensitive statistical procedures can be to the presence of a small that fraction of outlying observations. We envision that connecting these two fields could lead to the development of general guiding principles for constructing outlier robust variational inference algorithms. Such considerations seem fundamental to variational inference as it is commonly deployed for the analysis of modern large scale data that are bound to have outliers.

In summary, through reframing the variational inference optimization as a functional version of the traditional statistical penalized estimation, as in (4), we hope both to **(a)** be able to study theoretical aspects of variational inference and **(b)** establish a direct connection to decision theory and robust statistics. This project will provide preliminary research in the area of variational inference that will allow the PIs and co-PI to apply for more traditional funding sources after its completion, for example from the Division of Mathematical Sciences at the National Science Foundation.

Significance. Variational inference is already having a tremendous impact in machine learning and computer science. Its success as algorithm is largely explained by its reliability and scalability. The use of variational methods in other data-driven disciplines –such as econometrics and statistics– is still not as widespread. This project will present a theoretical analysis of variational inference using popular and well-known tools of statistics and econometrics. If successful, we will be contributing to setting up theoretical foundations for variational methods. Once these foundations become available, we hope to see variational inference becoming the default method in statistics and econometrics for estimating the hidden parameters of high-dimensional statistical models.

Though variational inference as a tool for the estimation of distributions of hidden parameters or latent variables when datasets are large or high-dimensional, is both timely and of interest to a large community, we feel that this particular project is not fundable by more traditional funding sources because it would be deemed too risky or early stage, especially due to the fact that the PIs and Co-PI on the project do not have a demonstrated expertise in the area of variational inference through past work on this topic. However, we believe that this is the strength of the project for the RISE award: namely, we propose to use the areas of expertise of the project PIs and Co-PI – statistics, econometrics, decision theory, robust statistics, message passing, and statistical physics – as a lens through which to bring new ideas and fundamental change to a field that is otherwise of practical interest to machine learning researchers and practitioners.

Budget. For the first year of RISE support, the PIs and co-PI will use the funds to partially support two full-time PhD students working on the project ($32,000), to support variational inference working group meetings bringing together researchers from different departments and schools at Columbia ($8,000), and for conference and research travel related to the project for the PIs, the co-PI, and the two supported students ($40,000).

References

[1] Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *Annals of Statistics*, 48(3):1475–1497, 2020.

[2] Anirban Bhattacharya, Debdeep Pati, and Yun Yang. Bayesian fractional posteriors. *The Annals of Statistics*, 47(1):39–66, 2019.

[3] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[4] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

[5] Gary Chamberlain. Robust decision theory and econometrics. *Working Paper*, 2019.

[6] Paul Dupuis and Richard S Ellis. A weak convergence approach to the theory of large deviations. 1997.

[7] Lars Peter Hansen and Thomas J Sargent. Robust control and model uncertainty. *American Economic Review*, 91(2):60–66, 2001.

[8] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the 5th International Conference on Learning Representations*, volume 2, page 6, 2017.

[9] Peter J Huber and Elvezio M Ronchetti. *Robust statistics.* Wiley, New York, 2nd ed., 2009.

[10] Fabio Maccheroni, Massimo Marinacci, and Aldo Rustichini. Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, 74(6):1447–1498, 2006.

[11] Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019.

[12] Tomasz Strzalecki. Axiomatic foundations of multiplier preferences. *Econometrica*, 79(1):47–73, 2011.

[13] Yixin Wang, Alp Kucukelbir, and David M Blei. Robust probabilistic modeling with Bayesian data reweighting. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3646–3655, 2017.