A Deterministic Global Optimization Method for Variational Inference

Hachem Saddiki
Mathematics and Statistics
University of Massachusetts, Amherst
saddiki@math.umass.edu

Andrew C. Trapp
Operations and Industrial Engineering
Worcester Polytechnic Institute
atrapp@wpi.edu

Patrick Flaherty
Mathematics and Statistics
University of Massachusetts, Amherst
flaherty@math.umass.edu

Abstract

Variational inference methods for latent variable statistical models have gained popularity because they are relatively fast, can handle large data sets, and have deterministic convergence guarantees. However, in practice it is unclear whether the fixed point identified by the variational inference algorithm is a local or a global optimum. Here, we propose a method for constructing iterative optimization algorithms for variational inference problems that are guaranteed to converge to the ϵ-global variational lower bound on the log-likelihood. We derive inference algorithms for two variational approximations to a standard Bayesian Gaussian mixture model (BGMM). We present a minimal data set for empirically testing convergence and show that a variational inference algorithm frequently converges to a local optimum while our algorithm always converges to the globally optimal variational lower bound. We characterize the loss incurred by selecting a non-optimal variational approximation distribution, suggesting that selection of the approximating variational distribution deserves as much attention as the selection of the original statistical model for a given data set.

1 Introduction

Maximum likelihood estimation of latent (hidden) variable models is computationally challenging because one must integrate over the latent variables to compute the likelihood. Often, the integral is high-dimensional and computationally intractable so one must resort to approximation methods such as variational expectation-maximization [1].

The variational expectation-maximization algorithm alternates between maximizing an evidence lower bound (ELBO) on the log-likelihood with respect to the model parameters and maximizing the lower bound with respect to the variational distribution parameters. Variational expectation-maximization is popular because it is computationally efficient and performs deterministic coordinate ascent on the ELBO surface. However, variational inference has some statistical and computational issues. First, the ELBO is often multimodal, so the deterministic algorithm may converge to a local rather than a global optimum depending on the initial parameter estimates. Second, the variational distribution is often chosen for computational convenience rather than for accuracy with respect to the posterior distribution, so the lower bound may be far from tight. The magnitude of the cost of using a poor approximation is typically not known. Here, we develop a deterministic global optimization

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

algorithm for variational inference to address the first issue and quantify the optimal ELBO for an example model to address the second issue.

2 GOP Variational Inference

The general Global OPtimization (GOP) algorithm was first developed by Floudas et. al. [4, 5, 3]. They extended the decomposition ideas of Benders [2] and Geoffrion [7]. While the framework has been successfully used for problems in process design, control, and computational chemistry, it has not, to our knowledge, been applied to statistical inference.

Here, we apply the GOP algorithm for variational inference in hierarchical exponential family models. In Section 2.1, we state the problem conditions necessary for the GOP algorithm. Then, we briefly review the mathematical theory for the GOP algorithm in Section 2.2. The review of the GOP algorithm in this section is not novel and necessarily brief. A more complete presentation of the general algorithm can be found in [3].

2.1 Problem Statement

GOP addresses biconvex optimization problems that can be formulated as

$$\begin{aligned} \min_{\alpha,\beta} \quad & f(\alpha,\beta) \\ \text{s.t.} \quad & g(\alpha,\beta) \leq 0 \\ & h(\alpha,\beta) = 0 \\ & \alpha \in A, \beta \in B, \end{aligned} \tag{1}$$

where A and B are convex compact sets, and $g(\alpha,\beta)$ and $h(\alpha,\beta)$ are vectors of inequality and equality constraints respectively. We require the following conditions on (1):

1. $f(\alpha,\beta)$ is convex in α for every fixed β, and convex in β for every fixed α;
2. $g(\alpha,\beta)$ is convex in α for every fixed β, and convex in β for every fixed α;
3. $h(\alpha,\beta)$ is affine in α for every fixed β, and affine in β for every fixed α;
4. first-order constraints qualifications (e.g. Slater's conditions) are satisfied for every fixed β.

If these conditions are satisfied, we have a biconvex optimization problem and we can use the GOP algorithm to find the ϵ-global optimum.

Partitioning and Transformation of Decision Variables Variational inference is typically formulated as an alternating coordinate ascent algorithm where the evidence lower bound is iteratively maximized with respect to the model parameters and the variational distribution parameters. However, the objective function construed under that partitioning of the decision variables is not necessarily biconvex. Instead, we consider all of the decision variables (model parameters and variational parameters) in the variational inference optimization problem jointly. Then, we are free to partition variables as we choose to ensure the GOP problem conditions are satisfied.

One may be able to transform variables in the original variational optimization problem, adding equality constraints as necessary, to satisfy the GOP problem conditions. For larger problems and more complex models, the partitioning of the decision variables may not be apparent. Hansen [8] proposed an algorithm to bilinearize quadratic and polynomial function problems, rational polynomials, and problems involving hyperbolic functions.

Biconvexity for Exponential Family Hierarchical Latent Variable Models The variational EM problem is a biconvex optimization problem under certain conditions. We cast the variational EM problem into a biconvex convex optimization form by partitioning the model parameters, ϕ, and the variational parameters ξ into α and β such that the GOP conditions are satisfied. All of the functions are analytical and differentiable.

2.2 GOP Theory

Projection We can cast the original optimization problem (1) as inner and outer optimization problems using a *projection* of the problem onto the space of β variables:

$$\begin{aligned} \min_{\beta} \quad & v(\beta) \\ \text{s.t.} \quad & v(\beta) = \min_{\alpha \in A} f(\alpha, \beta) \\ & \qquad \text{s.t.} \quad g(\alpha, \beta) \leq 0 \\ & \qquad \qquad h(\alpha, \beta) = 0 \\ & \beta \in B \cap V \end{aligned} \tag{2}$$

where V = $\{\beta : h(\alpha, \beta) = 0,\ g(\alpha, \beta) \leq 0\}$ for some $\alpha \in A$.

For a given β^t, the inner minimization problem in (2) is called the primal problem. Therefore, the function $v(\beta)$ can be defined for a set of solutions of the primal problem for different values of β. However, the last two conditions in (2) define an implicit set of constraints making the solution of the problem difficult.

Relaxed Dual Problem Dropping the last two constraints gives a relaxed dual problem,

$$\begin{aligned} \min_{\beta \in B,\ v_B} \quad & v_B \\ \text{s.t.} \quad & v_B \geq \min_{\alpha \in A} L(\alpha, \beta, \lambda, \mu) \quad \mu \geq 0,\ \lambda \end{aligned} \tag{3}$$

where $v_B \in \mathbb{R}$ and the Lagrange function for the primal problem is

$$L(\alpha, \beta, \lambda, \mu) = f(\alpha, \beta) + \mu^\top g(\alpha, \beta) + \lambda^\top h(\alpha, \beta)$$

We call the minimization problem in (3) the *inner relaxed dual problem*.

In summary, the primal problem contains *more* constraints than the original problem because β is fixed and so provides an *upper bound* on the original problem. The relaxed dual problem (3) contains *fewer* constraints than the original problem and so provides a *lower bound* on the original problem. Because the primal and relaxed dual problems are formed from an inner-outer optimization problem, they can be used to form an iterative alternating algorithm to determine the global solution of the original problem. For the rest of the mathematical development, we make the algorithm iterations explicit by introducing a t index. The key here is that information is passed from the primal problem to the dual problem through the Lagrange multipliers, $\{\lambda^t, \mu^t\}$, and information is passed from the dual problem to the primal problem through the optimal dual variables, β^t.

3 Bayesian Gaussian Mixture Model - Point Mass Approximation

Here, we consider a variant of the Gaussian mixture model where the cluster mean is endowed with a Gaussian prior. We consider the variational model for the posterior distribution to be a point mass function for discrete latent random variables and a Dirac delta function for continuous random variables. Variational expectation-maximization with a point-mass approximating distribution is exactly classical expectation-maximization [6, p 337].

3.1 GOP Algorithm Derivation

Model Structure A Gaussian mixture model with a prior distribution on the cluster means is

$$\begin{aligned} M_k|\Gamma &\sim \text{Gaussian}(0, \Gamma) \text{ for } k = 1, \ldots, K, \\ Z_i|\pi &\sim \text{Categorial}(K, \pi), \text{ for } i = 1, \ldots, N, \\ Y_i|Z_i, M &\sim \text{Gaussian}(m_{z_i}, 1), \text{ for } i = 1, \ldots, N, \end{aligned} \tag{4}$$

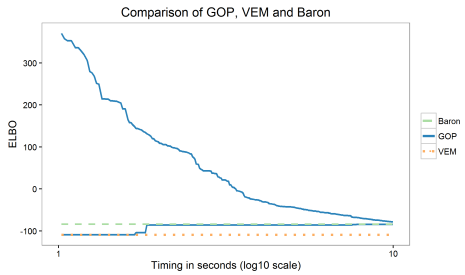
with model parameters $\phi \triangleq \{\pi, \Gamma\} = \{\pi_1, \ldots, \pi_K, \Gamma\}$, unobserved/latent variables $\{Z, M_1, \ldots, M_K\}$, and observed variable $Y = y$.

Our inferential aim is the joint posterior distribution $Z_i, M \mid Y_i, \hat{\phi}$, for $i = 1, \ldots, N$, where $\hat{\phi}$ is the maximum likelihood parameter estimate. The log likelihood for the model is $l_{\mathcal{D}}(\phi) \triangleq \log f(y|\phi)$.

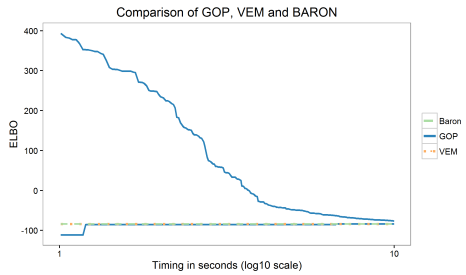
3.2 Experimental Results

Here, we examine the convergence of the upper and lower bounds on the ELBO for a small data set: $y = (-10, -10, 5, 25)$. We find that this minimal data set is useful for testing inference algorithms for the (Bayesian) Gaussian mixture model. This data set has two properties that promote local optima – repeated observations and outliers.

Convergence Experiment We examined the optimum ELBO values for GOP and variational EM (VEM) by randomly drawing initial model and variational parameters and running each algorithm to convergence. We drew 100 random initial values and observed that variational EM converged to a sub-optimal local fixed point in 87% of the trials while GOP converged to the global optimum in 100% of the trials. We verified the global optimal value using BARON, an independent global optimization algorithm [10, 11]. These results show that GOP, empirically, always converges to the global optimum. The VEM algorithm failed to converge to the global optimum more often than it succeeded. Figure 1 shows the VEM and GOP iterations with time along the x-axis and the ELBO along the y-axis for two random seeds. In Figure 1a VEM converges to a local optimum and is outside the bounds provided by GOP. Figure 1b shows a case where VEM converged to the global optimum. The locally optimal ELBO is -108.86 and the globally optimal ELBO is -84.04. These results quantify and agree with anecdotal evidence in the literature [9].



(a) GOP iterations for a parameter initialization where VEM converges to a *local* optimum.



(b) GOP iterations for a parameter initialization where VEM converges to a *global* optimum.

Figure 1: Comparison of GOP, VEM, and BARON for two random initializations.

4 Bayesian Gaussian Mixture Model - Gaussian Approximation

In this section we explore the quality of the approximation provided by two different variational models for the Bayesian Gaussian mixture model - a point mass distribution and a Gaussian distribution. We ran GOP, adapted to both approximations, using the small data set described. The globally optimal ELBO for point mass approximation and for Gaussian approximation is -84.04 and -82.75, respectively. We observe that the globally optimal ELBO for Gaussian approximation is higher than that of point mass approximation. This result shows that we do indeed incur a loss in accuracy when selecting a sub-optimal approximating distribution suggesting that careful consideration must be made when choosing an approximating distribution.

References

[1] Matthew J Beal and Zoubin Ghahramani. The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics*, 7, 2003.

[2] J. F. Benders. Partitioning procedures for solving mixed-variables programming problems. *Computational Management Science*, 2(1):3–19, January 1962.

[3] Christodoulos A Floudas. *Deterministic Global Optimization*. Theory, Methods and Applications. Springer US, 2000.

[4] Christodoulos A Floudas and V Visweswaran. A global optimization algorithm (gop) for certain classes of nonconvex nlps—i. theory. *Computers & chemical engineering*, 14(12):1397–1417, December 1990.

[5] Christodoulos A Floudas and Vishy Visweswaran. Primal-relaxed dual global optimization approach. *Journal of Optimization Theory and Applications*, 78(2):187–225, 1993.

[6] Andrew Gelman, J B Carlin, Stern Hal, Dunson David, Aki Vehtari, and Rubin Donald. *Bayesian Data Analysis*. Chapman and Hall, 3 edition, 2013.

[7] A. M. Geoffrion. Generalized benders decomposition. *Journal of optimization theory and applications*, 1972.

[8] P Hansen, B Jaumard, and Junjie Xiong. Decomposition and interval arithmetic applied to global minimization of polynomial and rational functions. *Journal of Computational Optimization*, 3(4):421–437, December 1993.

[9] G. D. Murray. Contribution to discussion of paper by A. P. Dempstar, N. M. Laird, and D. B. Rubin. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:27–28, 1977.

[10] Nikolaos V Sahinidis. Baron: A general purpose global optimization software package. *Journal of global optimization*, 8(2):201–205, 1996.

[11] M Tawarmalani and N V Sahinidis. A polyhedral banch-and-cut approach to global optimization. *Mathematical Programming*, 103(2):225–249, 2005.