CRVI: Convex Relaxation for Variational Inference

Ghazal Fazelnia [1] **John Paisley** [1]

Abstract

We present a new technique for solving non-convex variational inference optimization problems. Variational inference is a widely used method for posterior approximation in which the inference problem is transformed into an optimization problem. For most models, this optimization is highly non-convex and so hard to solve. In this paper, we introduce a new approach to solving the variational inference optimization based on convex relaxation and semidefinite programming. Our theoretical results guarantee very tight relaxation bounds that get nearer to the global optimal solution than traditional coordinate ascent. We evaluate the performance of our approach on regression and sparse coding.

1. Introduction

A major challenge of Bayesian modeling is posterior inference. For many models this requires calculating normalizing integrals that neither have a closed form, nor are solvable numerically in polynomial time. There are two fundamental approaches to addressing the posterior inference problem. One uses Markov chain Monte Carlo (MCMC) sampling techniques that are asymptotically exact. However, these methods tend to be slow compared with point-estimates and not scalable to large datasets (Hastings, 1970; Gelfand and Smith, 1990). Mean-field variational inference is another approach that approximates the posterior distribution by first defining a simpler family of distributions and then finding a member that is closest to the desired posterior (Jordan et al., 1999) according to the KullbackLeibler (KL) divergence. This turns the inference problem into an optimization problem. However, this introduces new challenges due to the resulting non-convex optimization.

In this paper, we present a method to deal with the non-convexities in variational inference (VI) optimization for conjugate models that achieve near globally optimal solutions. Our method is based on convex relaxation and semidefinite programming (SDP). In our approach, an SDP relaxation converts a non-convex polynomial optimization of vector parameters to a convex optimization with matrix parameters via a lifting technique. We call this approach convex relaxation for variational inference (CRVI). The exactness of the relaxation can then be interpreted as the existence of a low-rank solution to this SDP. Our main contribution is to solve this variational optimization problem in an accurate way and provide theoretical guarantees for the exactness of our solution using graph theoretic tools. To the best of our knowledge, this is the first time that a relaxation for variational inference could guarantee and produce optimal solutions that are either globally optimal solution or very close to it. Our experimental results demonstrate the effectiveness of CRVI compared with coordinate ascent for sparse regression and sparse coding models.

Convex optimization problems are one of the most important areas of optimization theory. They are guaranteed to have global optimal solutions that can be found with a numerical algorithm. On the other hand, there is no such theory for solving generic non-convex problems. Recent advances in the area of convex optimization provide a variety of methods for approaching and solving non-convex optimization problems exactly or approximately (Boyd and Vandenberghe, 2004; Yedidia et al., 2005; Wainwright and Jordan, 2008). For instance, several works have studied the existence of a low-rank solution to matrix optimizations with linear or nonlinear constraints (Pataki, 1998; Sturm and Zhang, 2003; Parrilo, 2003; Fazelnia et al., 2017; Madani et al., 2017). We build on the method in Madani et al. (2017) to obtain theoretical bounds for the exactness of CRVI.

There are a number of works that have addressed problems with probabilistic inference using convex optimization methods. These works have mostly focused on convex relaxation for maximum entropy and message passing algorithms (Guo and Schuurmans, 2008; Nickisch and Seeger, 2009; Seeger and Nickisch, 2011). In general, they lack control over the exactness of their approximations in that there is no estimate of the closeness of the solution of the relaxed problem to the optimal solution of the original problem.

In this work, we apply convex relaxation techniques to the

[1]Department of Electrical Engineering & Data Science Institute, Columbia University, New York, USA. Correspondence to: Ghazal Fazelnia <ghazal@ee.columbia.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

optimization problem introduced by variational inference with more focus on the cases where the hardness of the problem is due to quadratic or higher order polynomial terms. We first break down the objective function into two parts, one representing the polynomial and non-convex part and one for the rest of the objective function. In this method, we lift the domain of optimization from vectors to matrices, and capture all of non-convexities in the optimization within the transformed problem. As we show, tight relaxation bounds can be achieved to guarantee near-global optimal solution. We also observe that, in models with many parameters this matrix may be prohibitively large. In this case, we still demonstrate how CRVI can be beneficial by relaxing a locally non-convex problem over a subset of variational parameters.

In Section 2 we review variational inference and our proposed convex relaxation technique. In Section 3 we illustrate our method and discuss theoretical contributions. In Section 4, we show experimental results.

2. Background

2.1. Variational Inference

Variational inference approximates the posterior distribution of variables in a probabilistic model. Let $\mathcal{D}$ be a dataset that is analyzed with a model having variables in the set θ. The model assumption is $\mathcal{D}|\theta \sim p(\mathcal{D}|\theta)$, $\theta \sim p(\theta)$.

The goal is to calculate the posterior distribution $p(\theta|\mathcal{D})$ after observing the data. Due to complexities in most models, finding the true posterior distribution is a difficult task. Instead, we can approximate it by $q(\theta)$ such that this approximation is close to the true distribution according to some notion of similarity. For variational inference, this closeness is measured by the Kulback-Leibler (KL) divergence. To optimize the KL-divergence, one can observe that

$$\ln p(\mathcal{D}) \;= \underbrace{\mathbb{E}_q\Big[\ln \frac{p(\mathcal{D},\theta)}{q(\theta)}\Big]}_{\mathcal{L}(q(\theta))} + \underbrace{\mathbb{E}_q\Big[\ln \frac{q(\theta)}{p(\theta|\mathcal{D})}\Big]}_{\mathrm{KL}(q\|p)}, \qquad (1)$$

and since the LHS is constant, one can minimize KL by maximizing the variational objective function $\mathcal{L}$ over the parameters of a predefined distribution family $q(\theta)$. To define this family in a way that is amenable to optimization, one often assumes that $q(\theta)$ belongs to a family of distributions that factorizes over the variables in θ. Seeking to find parameters for this distribution, ϕ, results in optimizing the following problem,

$$\max_{\phi} \;\; \mathcal{L}(q(\theta)) \;\text{ subject to }\; \phi \in \text{feasible set}, \qquad (2)$$

where the feasible set is the intersection of possible regions for all of the constraints on the parameters. For a very large set of models, this optimization is non-convex or combinatorial, and hard to solve. Numerical algorithms are only able to achieve a local maximum, and most of the time there is no evaluation about how close this local optimum is to the global one.[1] In this paper, we consider the cases where this optimization is non-convex and NP-hard. While the global optimum for these optimizations might not be achievable, we aim to find a local optimum that is close to the global solution. Better local optima assure us that we obtain lower KL-divergence and a more accurate posterior approximation. Without loss of generality, we convert the problem to minimizing $-\mathcal{L}(q(\theta))$ over the same feasible set to make the problem more compatible with the convex optimization framework and notations.

We propose a new optimization approach to VI that we call convex relaxation for variational inference (CRVI). This technique approximates the optimization problem to overcome the issues related to non-convexities. As we will show, CRVI can result in near-global optimal solutions that are not only a better local optima compared to the standard coordinate ascent approach, but also provides a means for assessing closeness to the global optimum.

2.2. Convex Relaxation

We next present the general technique that we adopt and build on in this paper in its abstract representation. We then apply it to two specific variational inference optimization problems. Although there are exceptions, polynomial terms in an objective or constraint tend to add non-convexities and make the optimization intractable to solve. The technique that we use deals with these hard polynomial parts by converting them into near-exact tractable terms.

First we note that any polynomial function or expression can be represented as a quadratic function, possibly by introducing new variables (Berlekamp, 1970). This conversion is straightforward, and every high order term could be broken down into lower order terms by introducing new parameters and quadratic equality constraints. As a result, without loss of generality, we assume that all of the polynomial terms are quadratic. Let the following be a general polynomial optimization problem,

$$\begin{aligned} &\min_{x\in\mathbb{R}^d} \;\; f_0(x) \\ &\text{subject to } f_k(x) \le 0 \;\; \text{for } k = 1,\dots,K, \end{aligned} \qquad (3)$$

where $f_k = x^\top A_k x + b_k^\top x + c_k$ for $k = 0,\dots,K$. Since there are no limitations on the coefficient choices, the terms in (3) can represent any polynomial optimization or expression.

If all of the matrices $\{A_0, A_1, ..., A_K\}$ are positive semidef-

[1] We note that by this we do not mean how close $q(\theta)$ is to $p(\theta|\mathcal{D})$, but how close we are to optimizing the chosen $q(\theta)$.

inite, the optimization in (3) is convex. Otherwise, it is non-convex, and there is no numerical or analytical procedure that guarantees achieving a global optimum. We use a lifting technique that involves changing the variable space from vectors to matrices (Boyd and Vandenberghe, 2004). More specifically, define F_k and X_k as follows,

$$F_K = \begin{bmatrix} c_k & \frac{1}{2}b_k^\top \\ \frac{1}{2}b_k & A_k \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x^\top \\ x & xx^\top \end{bmatrix} \tag{4}$$

Then the equivalent optimization to (3) is

$$\begin{aligned} \min_{X \in \mathbb{R}^{(d+1)\times(d+1)}} \quad & \text{trace}(F_0 X) \\ \text{subject to} \quad & \text{trace}(F_k X) \leq 0 \ \text{ for } k = 1, .., K, \\ & X_{1,1} = 1, \ \ X \succeq 0, \\ & \text{rank}(X) = 1. \end{aligned} \tag{5}$$

The entry equal to 1 in matrix X is to ensure that we have a way to represent the terms that are linear with respect to x. It should be pointed out that matrix X is designed such that it replaces $[1\ x^\top]^\top \times [1\ x^\top]$. This transformation requires us to be able to decompose back the solution X of optimization (5) to get the vector x after solving it. To assure this, X needs to be positive semidefinite and have rank 1.

All terms in (5) are linear with respect to X and consequently convex, except for the last constraint on the rank of the matrix. To avoid this non-convex rank constraint, we can simply drop it. By dropping the rank constraint, we achieve an optimization that is linear in terms of a matrix variable that has to be positive semidefinite. As a result, we obtain a semidefinite program (SDP) relaxation for the optimization in (3) (Vandenberghe and Boyd, 1996). Although SDP methods may not be fast in general, by carefully designing them and avoiding redundancies, they can run in a reasonable amount of time. The following shows the relaxed optimization problem,

$$\begin{aligned} \min_{X \in \mathbb{R}^{(d+1)\times(d+1)}} \quad & \text{trace}(F_0 X) \\ \text{subject to} \quad & \text{trace}(F_k X) \leq 0 \ \text{ for } k = 1, .., K, \\ & X_{1,1} = 1, \ \ X \succeq 0. \end{aligned} \tag{6}$$

One of the important steps here is to quantify the exactness of this relaxation. Naturally we seek approximations that result in finding global optimal or near-global optimal solutions. The only constraint that we dropped is that the matrix has to be rank 1. Hence, in this relaxation, the final rank of X carries information on the exactness of this approximation. After solving the relaxed semidefinite program, if the rank of the optimal X is 1, we have found the global optimal solution for the original problem (3). Otherwise, we reach an approximate solution to the original problem. It should be noted that the lower the rank of the optimal solution of the relaxed problem, the closer the approximation to the global optimal solution of the original problem. Thus, the closer the rank of the optimal solution gets to 1, the closer we are to the global optimal solution. This rank of the relaxed problem helps us measure the closeness of the approximate solution to the global optimal solution of the original problem.

Fortunately, the rank of the solution of the relaxed problem cannot be arbitrary large, as shown by Madani et al. (2017). In fact, it is upper bounded by a property of a defined graph structure for the original problem which is its *treewidth*. The treewidth of an undirected graph is a number associated with the graph that is mainly used for complexity analysis of graphs. It can be calculated from the minimum size of largest node over all tree-decomposition of the graph or from the size of the largest clique in a chordal completion of the graph. The treewidth mainly parametrizes and describes the sparsity of a graph, meaning that sparser graphs tend to have smaller treewidths. The process is to first construct a graph from the original quadratic optimization problem (3), and then calculate an upper bound on the rank of the semidefinite relaxation using the treewidth of the constructed graph.

To build the graph, we need to assign a vertex to every entry of the vector $[1\ x^\top]^\top$ and add edges between vertices whose product appears in the objective function or any of the constraints of the original problem (3). All of the constants or non-variable coefficients are neglected in this process. For instance, if cross-term $x_i x_j$ appears somewhere in (3), we put an edge between vertices that correspond to entry x_i and x_j. Or if term x_k appears, we add an edge between vertices corresponding to x_k and 1 since $x_k = x_k \times 1$. Hence, every term in the optimization problem can be translated into a graph edge. Interestingly, one interpretation of adding entry '1' in the matrix definition (4) is to be able to represent linear terms as an edge here in the construction of the graph. The fewer the number of cross terms in the optimization, the fewer edges and the sparser the graph.

Now with the graph constructed, we can find an upper bound for the rank of the optimal solution of the relaxed problem in (6). The rank of the optimal solution to the relaxed problem is less than or equal to one plus the treewidth of its *enriched super-graph*. As a result, the lower the treewidth of the graph of the problem, the better approximation to the global optimal solution. As we show in the examples, no matter how large the dimensionality of the matrix X in (6), the rank of the optimal solution matrix will be smaller than or equal to the calculated upper bound.

Overall, in this relaxation and transformation, all approximations are pulled into the rank of the optimal solution. An important advantage of this is that if the structure of the sparsity graph of a problem is good enough for us to have a low upper bound, we can achieve a strong relaxation that

gives a near global optimal solution. To show how we use this in variational inference, we use a simple example model next. We then generalize it to other models.

3. Convex Relaxation for Variational Inference

3.1. CRVI for Bayesian Linear Regression

We first show the proposed CRVI method on two Bayesian linear regression models in which the posterior distribution is approximated with variational inference. We start with a simple model. Consider the dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$ with $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$, and the model,

$$\begin{aligned} y_i &\sim \text{Normal}(x_i^\top w, \alpha^{-1}), \\ w &\sim \text{Normal}(0, \lambda^{-1} I), \\ \alpha &\sim \text{Gamma}(a_0, b_0). \end{aligned} \tag{7}$$

The goal is to find $p(w, \alpha|\mathcal{D})$, the posterior distribution of the model parameters given the input data. Since the true posterior is hard to find, we apply variational inference to approximate it. Let $q(w, \alpha)$ denote the approximate posterior density and define

$$\begin{aligned} q(w, \alpha) &= q(w)q(\alpha) \\ &= \text{Normal}(w|\mu, \Sigma)\text{Gamma}(\alpha|a, b), \end{aligned} \tag{8}$$

where the factorization comes from the mean-field approximation. The variational objective $\mathcal{L}$ for this optimization problem is

$$\begin{aligned} \mathcal{L}(q) =& (a_0 - 1)(\psi(a) - \ln b) - b_0 \frac{a}{b} - \frac{\lambda}{2}(\mu^\top \mu + \text{trace}(\Sigma)) \\ &+ \frac{N}{2}(\psi(a) - \ln b) - \sum_{i=1}^N \frac{1}{2}\frac{a}{b}((y_i - x_i^\top \mu)^2 + x_i^\top \Sigma x_i) \\ &+ a - \ln b + \ln \Gamma(a) + (1 - a)\psi(a) + \frac{1}{2}\ln|\Sigma| + \text{const.} \end{aligned} \tag{9}$$

where 'const.' is a constant with respect to the variational parameters of this model, $\{a, b, \mu, \Sigma\}$, which this function should be maximized over. This objective function is non-concave with respect to its parameters and coordinate ascent variational updates—in which the parameters are cycled over and locally optimized holding the others fixed during each iteration—using arbitrary initialization will likely only achieve locally optimal solutions. We will next show how CRVI can significantly improve this result. We consider the variational inference optimization problem that minimizes $-\mathcal{L}$ subject to $a, b > 0, \ \Sigma \succeq 0$.

Our approach is to use the relaxation technique presented in the previous section on the polynomial part of this optimization that contains all of the non-convexities associated with this optimization problem. Consider the following reformulated optimization problem,

$$\begin{aligned} \min \ & \sum_{i=1}^N \frac{1}{2}((ey_i^2 - 2x_i^\top u + x_i^\top u\mu^\top x_i) + x_i^\top e\Sigma x_i) \\ & + \frac{\lambda}{2}(\mu^\top \mu + \text{trace}(\Sigma)) + b_0 e \\ & - (a_0 - 1)(\psi(a) + \ln c) - \frac{N}{2}(\psi(a) + \ln c) \\ & - a - \ln c - \ln \Gamma(a) - (1 - a)\psi(a) - \frac{1}{2}\ln|\Sigma| \\ \text{subject to } & a, c, e > 0, \quad \Sigma \succeq 0, \quad e = ac, \quad u = e\mu. \end{aligned} \tag{10}$$

This optimization is over the variables a, c, e, μ, u, Σ. Note that we introduced new variables c to replace $\frac{1}{b}$, e to represent ac and u to replace $e \times \mu$. This enables us to reformulate the polynomial part as a quadratic optimization problem. Hence, optimization problems (9) and (10) are identical. We refer to the first two lines of (10) as $f(a, c, e, \mu, u, \Sigma)$ which is in polynomial form and contains all of the non-convexities in this problem, while we refer to the rest as $g(a, c, \Sigma)$, which is non-linear and convex. This is due to convexity of negative ψ function for positive scalars as well as the convexity of the negative log and negative entropies. Therefore, by relaxing the first part, we get a convex relaxation for the optimization problem. In order to perform the relaxation, we need to rewrite $f(a, c, e, \mu, u, \Sigma)$ as a quadratic function of a vector variable. Based on the semidefinite relaxation construction in the previous section, we define the following vector

$$\nu = \begin{bmatrix} 1 & a & c & e & \mu^\top & u^\top & \Sigma_{1,1} & \Sigma_{1,2} & \cdots & \Sigma_{d,d} \end{bmatrix}^\top$$

It is easy to see that $f(a, c, e, \mu, u, \Sigma)$ is quadratic with respect to entries of ν. We reformulate the function f to use ν as an argument in f_{CR}. Thus the transformed optimization problem is as follows

$$\begin{aligned} \min_{\nu, a, c, \Sigma} \ & f_{CR}(\nu) + g(a, c, \Sigma) \\ \text{subject to } & a, c, e \geq 0, \ \ e = ac, \ \ u = e\mu, \ \ \Sigma \succeq 0 \\ & a = \nu_2, \ \ c = \nu_3, \\ & \text{vector}(\Sigma) = [\nu_{(5+2*d)} \ldots \nu_{(4+2d+d^2)}] \end{aligned} \tag{11}$$

where vector$(\cdot)$ vectorizes the matrix. Convex relaxation can now be defined for the optimization (11) by introducing new matrix variable $\mathbf{A} := \nu \times \nu^\top \in \mathbb{S}^{(4+2d+d^2)\times(4+2d+d^2)}$ and following the relaxation steps. $\mathbf{A}$ in this formulation plays the role of X in optimization (6). The following proposition gives our theoretical bounds for the exactness of this relaxation.

Proposition 1. The matrix solution obtained by CRVI for (11) has a rank less than or equal to 3.

Proof. Figure 1 shows the constructed graph for the original quadratic optimizations (9) on the left side, and its tree

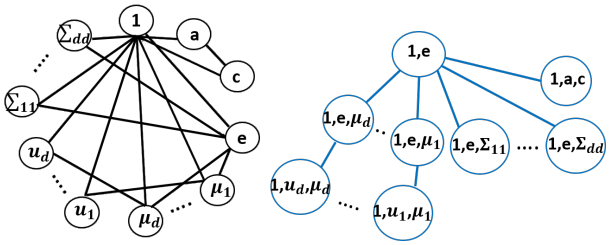

Figure 1. Constructed graph for the optimization problem (9) on the left side, and its tree decomposition on the right side. Some edges are removed for better legibility of the graphs.

decomposition on the right side. *Treewidth* is the cardinality of the largest vertex in a graph's tree decomposition minus 1 where its *enriched super-graph* is constructed. Since the cardinality of the largest vertex in its tree decomposition is 3, its treewidth is 2. This guarantees that the rank of the optimal solution of CRVI is upper bounded by 3. □

Note that in Figure (1) on the left side, vertex 1 is connected to e, all u entries and all Σ entries. Similarly, e is connected to all entries of μ and Σ. Big blue circles on the right side show the bag of nodes created in the tree decomposition construction.

Although the dimensionality of this optimization can be very large ($(4+2d+d^2)\times(4+2d+d^2)$), the rank of its solution is very low (upper bounded by 3 here). This indicates that the relaxation result will be in a close neighborhood of the global optimal solution considering the fact that a rank 1 solution specifies the global optimal solution. Furthermore, this bound exists regardless of dimensionality or scale of the input data.

3.2. Model Expansion Using Sparse Priors

We next generalize the Bayesian linear regression model by including dimension specific precisions to w that can be learned to prune irrelevant coefficients in a similar spirit as the Lasso (Tibshirani, 1996). This model is also known as the relevance vector machine or automatic relevance determination (Bishop, 2006). It modifies the Bayesian linear regression model by defining a separate prior on the diagonal entries of the covariance matrix of w as follows,

$$\begin{aligned} y_i &\sim \text{Normal}(x_i^\top w, \alpha^{-1}), \\ \alpha &\sim \text{Gamma}(a_0, b_0), \\ w &\sim \text{Normal}(0, \text{diag}(\lambda_1, \ldots, \lambda_d)^{-1}), \\ \lambda_k &\sim \text{Gamma}(m_0, l_0). \end{aligned} \tag{12}$$

Defining a posterior approximating variational distribution q as in the previous case, we now include $q(\lambda_k) = \text{Gamma}(m_k, l_k)$ for $k = 1, \ldots, d$. Calculating the objective results in the same form as before,

$$\begin{aligned} &\mathcal{L}(a, b, m_1, \ldots, m_d, l_1, \ldots, l_d, \mu, \Sigma) \; = \\ &-\sum_{i=1}^{N} \frac{1}{2}\frac{a}{b}((y_i - x_i^\top \mu)^2 + x_i^\top \Sigma x_i) + \frac{N}{2}(\psi(a) - \ln b) \\ &+\sum_{i=1}^{d}(\psi(m_i) - \ln(l_i)) - \frac{1}{2}(\mu^\top \text{diag}(\frac{m_1}{l_1}, ..., \frac{m_d}{l_d})\mu) \\ &-\frac{1}{2}\text{trace}(\text{diag}(\frac{m_1}{l_1}, ..., \frac{m_d}{l_d})\Sigma) \\ &+\sum_{i=1}^{d}(m_0 - 1)(\psi(m_i) - \ln(l_i)) - l_0\frac{m_i}{l_i} \\ &+(a_0 - 1)(\psi(a) - \ln b) - b_0\frac{a}{b} + \frac{1}{2}\ln\,|\Sigma| \\ &+a - \ln b + \ln\Gamma(a) + (1-a)\psi(a) \\ &+\sum_{i=1}^{d}(m_i - \ln l_i + \ln\Gamma(m_i) + (1-m_i)\psi(m_i)) + \text{const.} \end{aligned} \tag{13}$$

By reformulating this objective appropriately for convex relaxation, the procedure is very similar to the simpler model. We introduce new variables to replace high order polynomial terms. These new variables are

$$s_i = \frac{1}{l_i},\; r_i = m_i s_i,\;\; \zeta_i = r_i \mu_i \quad \text{for } i = 1, \ldots, d. \tag{14}$$

Repeating the relaxation steps described earlier, we achieve a convex relaxation for the optimization of (13). Similar to the simpler model, we can achieve the following theoretical result.

Proposition 2. The matrix solution obtained by CRVI for (13) has a rank less than or equal to 3.

The graph structure and tree decomposition for this problem is very similar to the simpler model in (3.1), and the same theoretical upper bounds are guaranteed. This strong upper bound exists regardless of the dimensionality of data or size of the input, even though this Bayesian model has a more complex prior structure and many more model parameters. Still, this is only a bound; as we will show in the experiments section the actual rank of the solution to the relaxed optimization is less than 3, and in fact is very close to 1. This means that although the theoretical bound assure us that the rank is less than or equal to 3, in practice on real data sets we can get almost exactly the global optimal solutions of the original problem.

3.3. CRVI for Nonparametric Factor Analysis

We illustrate CRVI on a more complex model, Bayesian nonparametric factor analysis (Paisley and Carin, 2009) of data $\mathcal{D} = \{x_i \in \mathbb{R}^d\}_{i=1}^N$. This will also allow us to propose another modification for the application of this framework due to the much larger number of parameters in the model.

The model is

$$\begin{aligned} x_i &\sim \text{Normal}(WZ_iC_i, \sigma^2 I), \\ C_i &\sim \text{Normal}(0, \lambda^{-1} I), \\ \pi_k &\sim \text{Beta}(\alpha\tfrac{\gamma}{K}, \alpha(1-\tfrac{\gamma}{K})), \\ z_{i,k} &\sim \text{Bernoulli}(\pi_k), \\ Z_i &= \text{diag}(z_{i,1}, \ldots, z_{i,K}), \end{aligned} \tag{15}$$

where $k = 1, ..., K$ are the latent factor indexes. In the limit $K \to \infty$ this converges to a nonparametric beta process model (Paisley and Jordan, 2016). In addition, due to the model specifications in (15), a sparse representation in enforced by beta-Bernoulli prior for Z.

Given a matrix $W \in \mathbb{R}^{d\times K}$, for each vector x_i we seek a sparse zero-one coding Z_i of this vector as well as weight coefficients C_i. The Z's specify which factors in W are used to represent the data, while the C's indicate the weights of those selected factors. In this model we will seek to find the posterior distribution of C as well as point estimates for Z as well as W. Therefore, the algorithm is actually EM and not variational inference since there is no forced factorization of q. However, we do this to focus on another area where CRVI may be useful, as described below.

For each data point i we define $q(C_i) = \text{Normal}(C_i|\mu, \Sigma)$. Here, we only focus on learning the local variables for a specific data point x_i, being Z_i, C_i. Therefore, we drop the subscripts below. The optimization problem corresponding to this part of the model is

$$\begin{aligned} \min_{Z,\mu,\Sigma} \quad & \frac{1}{2\sigma^2}(x - WZ\mu)^\top (x - WZ\mu) \\ & + \frac{1}{2\sigma^2}\text{trace}(WZ\Sigma ZW^\top) \\ & + \frac{\lambda}{2}\mu^\top\mu + \frac{\lambda}{2}\text{trace}(\Sigma) - \frac{1}{2}\log(|\Sigma|) + Z^\top h \\ \text{subject to} \quad & Z_{k,k} \in \{0, 1\}, \text{ for } k = 1, ..., K, \;\; \Sigma \succeq 0 \end{aligned} \tag{16}$$

where h is a constant vector with respect to optimization variables. Note that this optimization can be done in parallel for data points due to their independence. All of objective terms are polynomial with respect to the optimization variables. In addition, the log term is also convex with respect to Σ. To make all of the constraints quadratic, we replace the zero or one constraint for $Z_{k,k}$ with $Z_{k,k}^2 - Z_{k,k} = 0$. Therefore, we obtain a non-convex optimization with polynomial terms containing all of the non-convexities.

Motivation and discussion. Following the steps described in the previous section, we are able to define the convex relaxation optimization for this problem. Another novelty introduced here is that we have not relaxed the *entire* problem globally, which is computationally impossible for a model of this size (the dimensionality of X would be too massive). Instead, we only relaxed *locally* on the parameters for each observation. However, since optimizing over C and Z is both non-convex and combinatorially hard, we use this model to illustrate a proposed approach to *local relaxation* of the objective. Contrasting this with coordinate ascent, which would update one variable holding another fixed, we anticipate that this can find better local optimal values over *subsets* of parameters, and therefore hopefully over the entire objective function. After constructing the graph of this problem, we find that the rank of the optimal solution of the relaxed problem is upper bounded by 3. Accordingly, we anticipate to find near-global optimal solutions over these interacting local parameters.

3.4. CRVI in General Form

Following the ideas introduced by these examples, we present CRVI as a general framework. Let us consider the generic variational inference problem in (2). We split the objective into two functions, one containing polynomial terms, f, and one for the remaining parts, g. Transforming f to be a quadratic function, possibly by adding new constraints and variables, we get the optimization

$$\begin{aligned} \min_{\varphi^{(1)},\varphi^{(2)}} \quad & f(\varphi^{(1)}) + g(\varphi^{(2)}) \\ \text{subject to} \quad & \varphi^{(1)},\ \varphi^{(2)} \in \text{feasible set}. \end{aligned} \tag{17}$$

Note that $\varphi^{(1)}$ and $\varphi^{(2)}$ might have overlapping parameters. To complete the relaxation, we introduce a new matrix variable $\Phi^{(1)}$ and obtain CRVI for the general form,

$$\begin{aligned} \min_{\Phi^{(1)},\varphi^{2}} \quad & f(\Phi^{(1)}) + g(\varphi^{(2)}) \\ \text{subject to} \quad & \Phi^{(1)},\ \varphi^{(2)} \in \text{feasible set}, \\ & {\Phi^{(1)}}_{1,1} = 1, \;\; \Phi^{(1)} \succeq 0. \end{aligned} \tag{18}$$

If g is a convex function, (18) is a convex optimization problem solvable in polynomial time. By constructing the graph for this relaxation approximation bounds can be achieved. The lower the rank of the optimal solution $\Phi^{(1)}_{\text{opt}}$, the more exact the approximation. As seen in the above examples, variational inference do have this structure, for which low rank recovery and near-global optimal solutions are guaranteed. In the cases where g is non-convex, CRVI could be used to partially convexify the optimization problem. We can reduce the hardness related to f with this relaxation technique, get approximation bounds, and improve the results compared to the cases where we have to deal with both non-convex f and g.

Table 1. Information about the datasets, running time of the algorithms, and rank of the found solution using CRVI. We see that CRVI is slower than CAVI (coordinate ascent). However, the rank of the found CRVI solution is near 1 (and less than the theoretical upper bound of 3), indicating a solution nearer the global optimum. This is confirmed in Figure 2.

DataSet	Dim.	# of Samples	CAVI time (s)	CRVI time (s)	Rank
Birth Rate & Econ	4	30	0.281	1.115	1.11
Iris	4	150	0.231	1.807	1.20
Yacht	6	308	0.402	2.111	1.10
Pima Indian Diabetes	8	768	0.571	3.040	1.67
Bike Sharing	13	731	0.884	6.749	1.61
Parkinson	21	5875	0.962	7.309	1.98
WDBC	31	569	1.059	10.766	1.73
Online News Popularity	58	39644	9.341	15.223	1.52
Year Prediction Songs	90	515345	18.809	22.050	1.78

4. Experimental Results

4.1. CRVI for Sparse Bayesian Linear Regression

We focus on comparing the optimal value of the variational objective calculated by our method CRVI in Section (3.2), and using coordinate ascent variational inference (CAVI) which is the standard method for variational optimization. We implemented CRVI code using CVX, which is a package for specifying and solving convex programs (Grant and Boyd, 2014; 2008). We experiment on 9 datasets from the UCI repository with various sizes and dimensions. These data sets are: Iris, Birth rate and economic growth, Yacht, Pima Indian diabetes, Bike sharing, Parkinson data, Wisconsin breast cancer (WDBC), Online news popularity, Year of release prediction for a million songs. We experimented using 100 different hyper-parameter settings and initial values for each dataset. Table (1) shows some details about these datasets, as well as the average running time for our simulations and the average rank of the optimal solution found by CRVI.

As can be seen, CRVI is slower than CAVI, which is not unexpected. Although the actual dimensionality of the semidefinite matrix variables for these datasets varies from 28×28 to 8284×8284, the average ranks found show that, regardless of the size of the data, the rank remains small and close to 1. This means that the CRVI is able to find nearly-global optimal solutions, considering that a rank 1 solution gives the exact global optimum solution. To evaluate the improvement according to the variational objective function, for each simulation of each dataset we subtracted the local optimal value of CAVI from CRVI, and divided it by optimal value found by CAVI to get the relative improvement to the maximization problem. We show a summary of these results in a boxplot for each dataset in Figure 2. As can be seen, CRVI significantly improved the local optimal solution of the optimization over coordinate ascent, which can be interpreted as finding a more accurate posterior approximation.

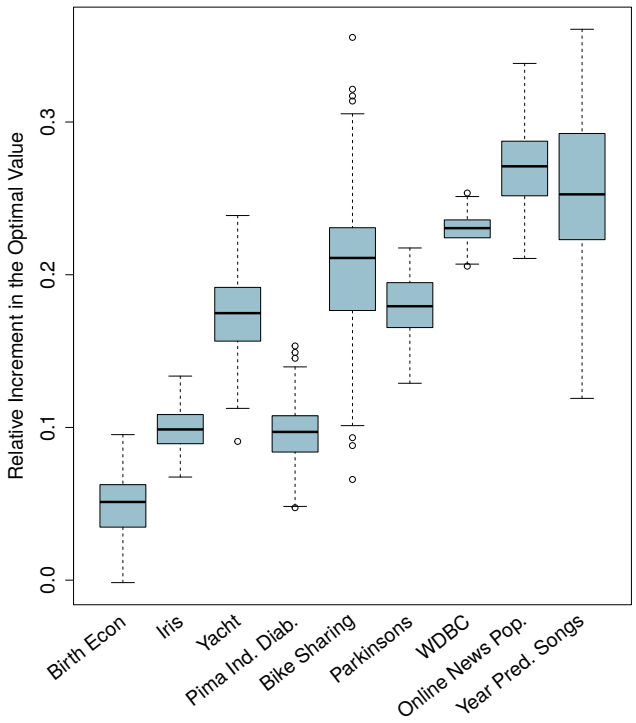

Figure 2. Boxplot of relative improvement in the calculated local optimal value of CRVI compared to CAVI. Each box represents the summary of the fractional improvement of CRVI over CAVI for 100 simulations using different prior hyper-parameters and initializations. After calculating the respective local optimal variational objective functions, the value found by CAVI is subtracted from the value from CRVI and divided by the values from CAVI to obtain the relative improvement score. As is evident, CRVI gave a significant improvement over CAVI.

4.2. CRVI for Nonparametric Factor Analysis

We also compare the accuracy of CRVI for sparse signal representation for dictionary learning with K-SVD (Aharon et al., 2006) on synthetic data. K-SVD uses orthogonal matching pursuits (OMP) to encode each signal in a dictionary (Tropp, 2004), which is also learned during the optimization process. Our goal is to compare the number of correctly recovered entries in the binary Z. We generate $N = 300$ observations of $D = 100$ dimensions and set $K = 100$ and $\lambda = 0.1$. We change the sparsity level of the generated Z over different simulations.

In Figure 3, the x-axis represents the probability of a '1' in each entry of Z when generating this binary encoding, while the y-axis shows the percentage of correctly recovered values in Z over the entire data set. As can be seen, CRVI is able to better learn the correct values for Z by finding the correct sparsity. Figure 4 shows the percentage of correctly recovered 1's for CRVI and KSVD. As can be seen from the figure, CRVI has a better performance in recovering the correct locations of 1's in the original Z matrix. Also, since we are focusing on the local optimal solution over Z and C as discussed in Section 3.3, we use the correct W in this experiment. Therefore, KSVD actually reduces to OMP in this experiment.

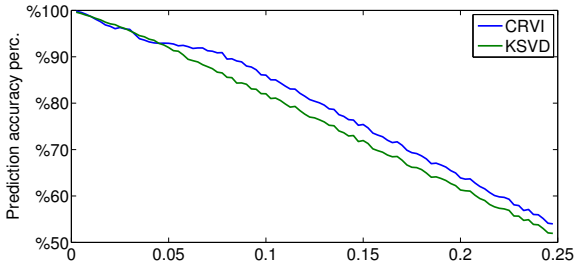

Figure 3. The fraction of agreement in the recovered Z's with original Z using CRVI and K-SVD (here, OMP). The x-axis shows the probability of a 1 in every entry of the original sparse matrix.

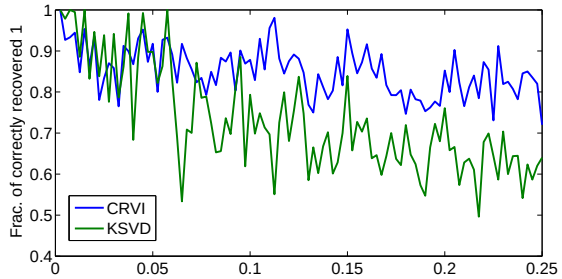

Figure 4. The fraction of correctly recovered 1's in the original Z using CRVI and K-SVD (here, OMP). The x-axis shows the probability of a 1 in every entry of original sparse matrix.

5. Discussion

Convex relaxations are a powerful technique for approximating (convexifying) hard optimization problems associated with variational inference. However, one of the caveats of this method is its runtime complexity, arising mostly from the positive semidefinite constraint. Fortunately, recent advances in this area have suggested faster ways to impose these types of constraints by breaking them into several smaller-sized semidefinite constraints. This significantly improves the running time of these types of relaxations (Kalbat and Lavaei, 2016). We expect that incorporating these techniques can improve the computational performance of this algorithm. Another future direction is to find tighter bounds for the relaxation exactness using the treewidth measure. Finding the exact treewidth of a graph is an NP-hard problem in general, and the bounds given in this paper used the treewidth's that were within our computational power. There may be better ways to reach smaller treewidth's and make the theoretical bounds tighter. The observed ranks in Table (1), smaller than the theoretical upper bound of 3, indicate that there is room for improvement in the theory in this direction.

6. Conclusion

We presented convex relaxation for variational inference (CRVI), a method to learn parameters of approximate posterior distributions using mean-field variational inference. We focused on Bayesian linear regression and sparse coding models. By lifting the domain of the optimization, we were able to relax the non-convex parts of the variational objective function and approximate the variational parameters. Graph theoretic tools enabled us to quantify the exactness of this approximation, and estimate the closeness of the obtained solution to the global optimal solution. We showed that CRVI can significantly improve the traditional coordinate ascent (CAVI) optimization technique on various datasets for sparse Bayesian linear regression and sparse coding for nonparametric factor analysis.

References

Aharon, M., Elad, M., and Bruckstein, A. (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322.

Berlekamp, E. R. (1970). Factoring polynomials over large finite fields. *Mathematics of Computation*.

Bishop, C. (2006). Pattern recognition and machine learning. *Springer*.

Boyd, S. and Vandenberghe, L. (2004). Convex optimization. *Cambridge University Press*.

Fazelnia, G., Madani, R., Kalbat, A., and Lavaei, J. (2017). Convex relaxation for optimal distributed control problems. *IEEE Transactions on Automatic Control*, 62(1):206–221.

Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*.

Grant, M. and Boyd, S. (2008). Graph implementations for nonsmooth convex programs. In Blondel, V., Boyd, S., and Kimura, H., editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited.

Grant, M. and Boyd, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx.

Guo, Y. and Schuurmans, D. (2008). Convex relaxations of latent variable training. *Advances in Neural Information Processing Systems*.

Hastings, W. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Kalbat, A. and Lavaei, J. (2016). A fast distributed algorithm for sparse semidefinite programs.

Madani, R., Fazelnia, G., Sojoudi, S., and Lavaei, J. (2017). Finding low-rank solutions of sparse linear matrix inequalities using convex optimization. *SIAM Journal on Optimization*.

Nickisch, H. and Seeger, M. W. (2009). Convex variational bayesian inference for large scale generalized linear models. *Proceedings of the 26th Annual International Conference on Machine Learning*.

Paisley, J. and Carin, L. (2009). Nonparametric factor analysis with beta process priors. In *International Conference on Machine Learning*.

Paisley, J. and Jordan, M. I. (2016). A constructive definition of the beta process. *arXiv preprint, arXiv: 1604.0068*.

Parrilo, P. (2003). Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming*.

Pataki, G. (1998). On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of Operations Research*, 23.

Seeger, M. W. and Nickisch, H. (2011). Large scale bayesian inference and experimental design for sparse linear models. *SIAM Journal on Imaging Sciences*, 4(1):166–199.

Sturm, J. F. and Zhang, S. (2003). On cones of nonnegative quadratic functions. *Mathematics of Operations Research*, 28.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*.

Tropp, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242.

Vandenberghe, L. and Boyd, S. (1996). Semidefinite programming. *SIAM Review*.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2005). Bethe free energy, kikuchi approximations, and belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312.