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 with a focus on Bayesian linear regression and sparse coding models. Variational inference is a widely used method for posterior inference in which the inference problem is transformed into an optimization problem. For most models, this optimization is highly non-convex and hard to solve. In this paper, we introduce a novel approach to solve variational inference optimization and get near global optimal solutions using convex relaxation and semidefinite programming. Our theoretical results guarantee very tight relaxation bounds. We evaluate the performance of our approach on various datasets for Bayesian regression and sparse coding, which show clear improvements over coordinate ascent methods.

1. Introduction

The main challenge in Bayesian models is posterior inference of model parameters. 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 solve the posterior inference problem. One approach uses Markov chain Monte Carlo (MCMC) sampling technique which is asymptotically exact. However, these methods tend to be slow compared to point-estimate techniques, and not scalable for 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 factorized distribution and then finding a member that is closest to the desired posterior (Jordan et al., 1999; Blei et al., 2017; Seeger and Wipf, 2010). It uses Kullback–Leibler (KL) divergence as a measure of closeness. Therefore, the inference problem becomes an optimization problem. This optimization however, imposes new types of challenges. Finding closed form expressions for the objective or solving hard and non-convex optimizations are instances of these problems. In this paper, we present a novel method to deal with the non-convexities in variational inference optimization for conjugate models and achieve near exact globally optimal solutions and posterior approximations.

Our method is based on convex relaxation and semidefinite programming (SDP) that optimizes variational inference over its parameters. In our approach, SDP relaxation converts a non-convex polynomial optimization with vector variables to a convex optimization with matrix variables 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 (e.g., rank-1) solution for SDP relaxation. Our main contribution is to solve the variational optimization problem in a very accurate way and provide theoretical guarantees for 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 global optimal solution or very close to it. Our experimental results demonstrate the effectiveness and performance of CRVI, and show significantly improved results for variational inference optimization compared to coordinate ascent. In this paper, we focus primarily on Bayesian regression models, an essential part of statistical analysis, as well as sparse coding problem for dictionary learning.

Convex optimization problems are one the most important sets of optimization problems. They are guaranteed to have a global optimal solution that is within reach with a numerical algorithm. On the other hand, there is no theory for solving non-convex problems in general. However, 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). 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; Madani et al., 2017). We use the method in (Madani et al., 2017) for obtain theoretical bounds for...
exactness of CRVI.

There are a number of works that have addressed the 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 on the closeness of the solution of the relaxed problem to the optimal solution of the original problem.

This paper is organized as follow. As background overview in section 2.1, we present an overview of variational inference method, and challenges raised in its inference process. Next in subsection 2.2 we present the convex relaxation method that we incorporate in solving the optimization problems. In section 3 we illustrate our method and theoretical contributions. In the last section, we show the experimental results.

Notation: Throughout this paper, we use superscript $\top$ to show transpose of a vector or a matrix. The space of $n \times n$ positive semidefinite matrices is shown by $S^n_{\geq 0}$ or with the inequality $\succeq 0$. Determinant of matrix $X$ is indicated by $|X|$, and $\text{diag}(x)$ shows a diagonal matrix whose diagonal entries is vector $x$.

2. Background

In this section, we provide a brief introduction to variational inference and convex relaxation methods.

2.1. Variational Inference

Variational inference is an inference technique that computes the posterior distribution for parameters of a probabilistic model. Let $\mathcal{D}$ be an observed dataset that is generated from a model with parameters belonging to the set $\theta$. We define the model as follow: $\mathcal{D} \sim p(\mathcal{D} | \theta), \theta \sim p(\theta)$.

The goal is to calculate the posterior distribution $p(\theta | \mathcal{D})$ and update the model parameters after observing the data. Due to complexities in most of the models, finding the true posterior distribution is a very hard task. Instead, we can approximate it by $q(\theta)$ such that this approximation is close to the true distribution according to some notion of distance measure. For variational inference, the closeness is measured by Kulback-Leibler (KL) divergence which is a non-symmetric and non-negative quantity used in calculating the difference between two probability distributions. To construct the KL-divergence, the first step is to expand the probability of observed data $\mathcal{D}$ using the model parameters.

$$
\ln p(\mathcal{D}) = \ln p(\mathcal{D}, \theta) - \ln p(\theta | \mathcal{D})
= E_q \left[ \ln \frac{p(\mathcal{D}, \theta)}{q(\theta)} \right] + E_q \left[ \ln \frac{q(\theta)}{p(\theta | \mathcal{D})} \right]. 
$$

Note that left hand side is invariant with respect to $q(\theta)$ and remains the same under the expectation. As can be seen from equation (1), calculating KL-divergence is intractable due to unknown true posterior $p(\theta | \mathcal{D})$. However, the goal is to find an approximation $q(\theta)$ that minimizes KL-divergence and ideally makes it zero. In equation (1), $p(\mathcal{D})$ is constant with respect to $q(\theta)$ which means that minimizing the KL-divergence is equivalent to maximizing $\mathcal{L}$ as a function of the parameters of $q(\theta)$ distribution. This interpretation would transform the original inference problem into an optimization problem over the parameters of $q(\theta)$. Hence, instead of minimizing KL-divergence, we can maximize $\mathcal{L}$ which is called evidence lower bound (ELBO) or variational objective function. To parametrize the problem, we assume that $q(\theta)$ belongs to a family of distributions factorized over the variables in $\theta$. Seeking to find parameters for this distribution, $\phi$, results in optimizing the following problem.

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

where 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 (Blei et al., 2017). Numerical algorithms are only able to achieve a local maximum when often times there is no evaluation about how close this local optimum is to the global one. In this paper, we consider the cases where this optimization is non-convex and NP-hard to solve. While the global optimum for these optimizations might not be achievable, we aim to find a local optimal that is very close to the global solution. Better local optimal assure us that we obtain lower KL-divergence and 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 to the convex optimization framework notations.

We propose a new technique based on a convex relaxation methodology, that we call convex relaxation for variational inference (CRVI), to approximate the original problem and overcome the issues related to non-convexities. As we will show, CRVI results in near-global optimal solution that is not only a better local optimal compared to other methods, but also it provides us with a valuable measurement to assess closeness to the global optimum and effectiveness of the relaxation.
2.2. Convex Relaxation

In this section, we explain the technique that we use in this paper in a completely general framework. Then, we apply it to specific variational inference optimization problems. Although there are exceptions, polynomial parts in objective or constraints of an optimization tend to add non-convexities and make the optimization intractable to solve in general. The technique that we use deals with these hard parts, and convert them into near-exact tractable terms.

First we note that any polynomial function or term could be represented as a quadratic function possibly with 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.

\[
\min_{x \in \mathbb{R}^d} f_0(x) \\
\text{subject to } f_k(x) \leq 0 \text{ for } k = 1, \ldots, K, \tag{3}
\]

where \( f_k = x^T A_k x + b_k^T x + c_k \) for \( k = 0, \ldots, K \). Since there are no limitations for the coefficient choices, the terms in optimization (3) can represent any polynomial optimizations or sections.

If all of \( \{A_0, A_1, \ldots, A_K\} \) matrices are positive semidefinite, optimization (3) is convex. Otherwise, it is non-convex, and there is no numerical or analytical procedure that guarantees to achieve the global optimum. The lifting technique that we use involves changing the variable space from vectors to matrices (Boyd and Vandenberghe, 2004). More specifically, define \( F_k \)'s and \( X_k \)'s as follow:

\[
F_K = \begin{bmatrix}
c_k & 1/2 b_k^T \\
1/2 b_k & A_k
\end{bmatrix}, X = \begin{bmatrix}
x \\
x^T x^T
\end{bmatrix} \tag{4}
\]

Then the equivalent optimization to (3) is:

\[
\min_{X \in \mathbb{R}^{(d+1) \times (d+1)}} \text{trace}(F_0 X) \\
\text{subject to } \text{trace}(F_k X) \leq 0 \text{ for } k = 1, \ldots, K, \tag{5}
\]

\[
X_{1,1} = 1, \ X \succeq 0, \ \text{rank}(X) = 1.
\]

The entry '1' on the upper left side of matrix \( X \) is to ensure that we have a way to represent the terms that are linear with respect to \( x \) as will be clearer later. It should be pointed out that matrix \( X \) is designed such that it replaces the \( [1 \ x^T]^T \times [1 \ x^T] \). 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 has rank 1.

All terms in optimization (5) are linear with respect to \( X \) and consequently convex, except for the last constraint on rank of the matrix. To deal with the non-convexity of the rank constraint, we 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 (3) (Vandenberghe and Boyd, 1996). Although SDP methods may not be fast in general, by carefully designing them and avoiding redundancies, they can perform in a reasonable amount of time. Following shows the relaxed optimization problem.

\[
\min_{X \in \mathbb{R}^{(d+1) \times (d+1)}} \text{trace}(F_0 X) \\
\text{subject to } \text{trace}(F_k X) \leq 0 \text{ for } k = 1, \ldots, K, \tag{6}
\]

\[
X_{1,1} = 1, \ X \succeq 0
\]

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, rank of optimal \( X \) carries the information on exactness of this approximation.

After solving the relaxed semidefinite program, if the rank of the optimum matrix is 1, we are guaranteed to have an exact relaxation, or equivalently, we have found the global optimal solution for the original problem (3). Otherwise, we reach to an approximation solution of the original problem. It should be noted that the lower the rank of the optimal solution of the relaxed problem, the more precise approximation to the global optimal solution of the original problem. Thus, the closer the rank of the optimal solution can get to 1, the closer we can get to the global optimal solution. This quantity helps us to measure the exactness of the relaxation as well as the closeness of the approximate solution to the global optimal solution of the original problem.

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

To build the graph, we need to assign a vertex to every entry of the vector \( [1 \ x^T]^T \), and add edges between vertices whose
We first start with a simple model. Consider the dataset

$$y_i \sim \text{Normal}(x_i^T w, \alpha^{-1}), \quad w \sim \text{Normal}(0, \lambda^{-1} I),$$

$$\alpha \sim \text{Gamma}(a_0, b_0)$$

(7)

The goal is to find $$P(w, \alpha | D)$$, 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

$$q(w, \alpha) = q(w)q(\alpha) = \text{Normal}(w|\mu, \Sigma)\text{Gamma}(\alpha|a, b),$$

where this factorization comes from the mean-field assumption. The variational objective $$L$$ for this optimization problem is

$$L(q) = (a_0 - 1)(\psi(a) - \ln b) - b_0 a - \frac{\lambda}{2} (\mu^T \mu + \text{trace}(\Sigma))$$

$$+ \frac{N}{2} (\psi(a) - \ln b) - \sum_{i=1}^{N} \frac{1}{2} a b ((y_i - x_i^T \mu)^2 + x_i^T \Sigma x_i)$$

$$+ a - b + \ln \Gamma(a) + (1 - a) \psi(a) + \frac{1}{2} \ln |\Sigma| + \text{const.}$$

(8)

where ‘const.’ is a constant with respect to parameters of this model, $$\{a, b, \mu, \Sigma\}$$. This maximizing objective function is non-concave with respect to its parameters (it is a non-convex optimization), and coordinate ascent variational updates with arbitrary initialization will likely only achieve locally optimal solutions. We will next show how CRVI can significantly improve this result. Without loss of generality, we consider the variational inference optimization problem that minimizes $$-L$$ 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 happens to carry all of the non-convexities associated with this optimization problem. Consider the reformulated optimization as follows

$$\min \sum_{i=1}^{N} \frac{1}{2} ((e y_i^2 - 2 x_i^T u + x_i^T u \mu^T x_i) + x_i^T e \Sigma x_i)$$

$$+ \frac{\lambda}{2} (\mu^T \mu + \text{trace}(\Sigma)) + b_0 c$$

$$- (a_0 - 1)(\psi(a) + \ln c) - \frac{N}{2} \psi(a) + \ln c$$

$$- a - b + \ln \Gamma(a) - (1 - a) \psi(a) - \frac{1}{2} \ln |\Sigma|$$

subject to $$a, c, d > 0, \Sigma \succeq 0, e = ac, u = e \mu$$.

(9)

This optimization is over the variables $$a, c, e, \mu, u, \Sigma$$. Note that we introduced new variables $$c$$ to replace $$\frac{1}{2}$$, $$e$$ 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 (8) and (9) are identical. We refer to the first two lines of (9) as $$f(a, c, e, \mu, u, \Sigma)$$ that is

$$\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$$ with $$x \in \mathbb{R}^d$$ and $$y \in \mathbb{R}$$, and the model:

product is appeared 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 $$x_i x_j$$ is appeared somewhere in the optimization (3), we put an edge between vertices corresponding to $$x_i$$ and $$x_j$$. If term $$x_k$$ is appeared, we would 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. As an observation in this framework, the fewer the number of cross terms in the optimization, the fewer edges in the graph and sparser graph.

Now with the graph constructed, we can find an upper bound for rank of the optimal solution of the relaxation (6). It is proved that the rank of the optimum solution for the relaxed problem is less than or equal to treewidth of its graph plus 1 (Madani et al., 2017). 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 variable $$X$$ in (6) is, 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 of approximations are pulled into the rank of the optimum solution. An important advantage of this is that if structure of sparsity graph of a problem is good enough for us to have a low upper bound, we could achieve a strong relaxation that gives us near global optimum solutions. To show how we use this in variational inference optimization, we use a simple example model next, and then we generalize it to other models as well.

3. Convex Relaxation for Variational Inference

We start off this section by presenting CRVI for simple models, and then we expand it for general models.

3.1. Convex Relaxation for Sparse Bayesian Linear Regression

We show CRVI method on a specific prior structure for the Bayesian linear regression model known as automatic relevance determination (Drugowitsch, 2013) in which posterior distributions are approximated with variational inference. We first start with a simple model. Consider the dataset
in a polynomial form and carries all of the non-convexities in this problem, while we refer to the rest of its objective as \( g(a, c, \Sigma) \), that is non-linear and convex. This is due to convexity of negative \( \psi \) function for positive scalars as well as convexity of negative log and negative entropies. Therefore, by relaxing the first part, we get a convex relaxation. In order to perform the relaxation, we need to reformulate \( f(a, c, e, \mu, u, \Sigma) \) as a quadratic function. Based on the semidefinite relaxation construction in the previous section, we define the following vector

\[
\nu = [1\ a\ c\ e\ \mu^\top\ u\ \Sigma_{1,1}\ \Sigma_{1,2}\ \ldots\ \Sigma_{d,d}]^\top
\]

It is easy to see that \( f(a, c, e, \mu, u, \Sigma) \) is quadratic with respect to entries of \( \nu \). We reformulate function \( f \) to use \( \nu \) as an argument in \( f_{CR} \).

Hence, the transformed optimization problem is as follows

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

where vector(\( \cdot \)) vectorizes the matrix. Convex relaxation can now be defined for the optimization (10) by introducing new matrix variable \( A := \nu \times \mu^\top \in \mathbb{S}^{(4+2d+d^2) \times (4+2d+d^2)} \) and following the relaxation steps. We call this method Convex relaxation for variational inference (CRVI). \( A \) in this formulation plays the role of \( X \) in optimization (6). The following proposition shows out theoretical bounds for exactness of this relaxation.

**Proposition 1.** The matrix solution obtained by CRVI for (10) has a rank less than or equal to 3.

**Proof.** Fig. (1) shows the constructed graph for the original quadratic optimizations (8) on the left side, and its tree decomposition on the right side. **Tree-width** is cardinality of the largest vertex in its tree decomposition minus 1. 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 that in Fig. (1) in the left side Fig., vertex 1 is connected to \( e \), all \( u \) entries and all \( \Sigma \) entries. Similarly, \( e \) is connected to all entries of \( \mu \) and \( \Sigma \). Big blue circles in the right side image show the bag of nodes created in tree decomposition construction.

Although the dimensionality of argument of this optimization could be very large, 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 the rank 1 solution specifies the exact relaxation. Furthermore, this bound exists regardless of dimensionality or scale of the input data.

### 3.2. Model Expansion Using Sparse Priors

In this subsection we generalize the model in the previous section. The base of this model is the same as the one described before with the addition of assigning an individual hyper-prior to each regression coefficient separately. These hyper-priors are adjusted to eventually prune irrelevant coefficients without the need for a separate validation set, unlike comparable sparsity inducing methods such as the Lasso (Tibshirani, 1996). Let us modify the model to define a separate prior on diagonal entries of covariance matrix of \( w \) as follows:

\[
\begin{align*}
\gamma_i & \sim \text{Normal}(x_i^\top \omega^{-1}, \omega) \quad i = 1, \ldots, d \\
\omega & \sim \text{Normal}(0, \text{diag}(\lambda_1, \ldots, \lambda_d)^{-1}) \\
\lambda_k & \sim \text{Gamma}(m_k, l_k) \quad k = 1, \ldots, d
\end{align*}
\]

Defining posterior approximation distributions \( q \) as in the previous case, we now add \( q(\lambda_k) = \text{Gamma}(m_k, l_k) \) for \( k = 1, \ldots, d \). Calculating the objective results in the same form as before,

\[
\begin{align*}
\mathcal{L}(a, b, m_1, \ldots, m_d, l_1, \ldots, l_d, \mu, \Sigma) & = \\
& \sum_{i=1}^{N} \frac{1}{2b} ((y_i - x_i^\top \mu)^2 + x_i^\top \Sigma x_i) + \frac{N}{2} (\psi(a) - \ln b) \\
& + \sum_{i=1}^{d} (\psi(m_i) - \ln(l_i)) - \frac{1}{2} (\mu^\top \text{diag}(m_1, \ldots, m_d) \mu) \\
& - \frac{1}{2} \text{trace} (\text{diag}(m_1, \ldots, m_d) \Sigma) \\
& + \sum_{i=1}^{d} (m_0 - 1)(\psi(m_i) - \ln(l_i)) - l_0 m_i \\
& + (a_0 - 1)(\psi(a) - \ln b) - b_0 \frac{a}{b} + \frac{1}{2} \ln |\Sigma| \\
& + a - \ln b + \ln \Gamma(a) + (1 - a)\psi(a) \\
& + \sum_{i=1}^{d} (m_i - \ln l_i + \ln \Gamma(m_i) + (1 - m_i)\psi(m_i)) + \text{const.}
\end{align*}
\]
By reformulating this objective appropriately for the convex relaxation, the procedure is very similar to the simpler model. We introduce new variables to replace high order polynomial terms. The new variables are as follows:

\[ s_i = \frac{1}{r_i}, r_i = m_is_i, \quad \zeta_i = r_i\mu_i \quad \text{for } i = 1, \ldots, d. \quad (13) \]

Repeating the relaxation steps described earlier, we achieve a convex relaxation for the optimization (12). Similar to the simpler model, we can achieve the following theoretical result.

**Proposition 2.** The matrix solution obtained by CRVI for (12) has a rank less than or equal to 3.

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. The strong upper bound exists regardless of the dimensionality of data or size of input even though this Bayesian model has a more complex prior structure and a lot more model parameters. As we will show in the experiment section, the obtained rank of the optimal matrix solution of the relaxed optimization is less than 3, and in fact, it 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 world data, we can get almost exact relaxation and global optimal solutions of the original problem.

### 3.3. CRVI for Sparse Data Representation and Dictionary Learning

In this section, we illustrate CRVI for a more complicated modeling framework. We consider a Bayesian nonparametric factor analysis for data \( D = \{ x_i \in \mathbb{R}^d \}_{i=1}^N \). Model description is as below:

\[ x_i \sim \text{Normal}(W^TZ_iC_i, \sigma^2 I), \quad C_i \sim \text{Normal}(0, \lambda^{-1} I), \]

\[ \pi_k \sim \text{Beta}(\alpha \frac{\gamma}{K}, \alpha (1 - \frac{\gamma}{K})), \quad z_{i,k} \sim \text{Bernoulli}(\pi_k), \]

\[ Z_i = \text{diag}(z_{i,1}, \ldots, z_{i,K}), \quad (14) \]

where \( k = 1, \ldots, K \) shows the factor indices. In the limit \( K \to \infty \) the random measure \( H_K = \sum_{k=1}^K \pi_k \delta_{w_k} \) converges to a beta process (Paisley and Jordan, 2016). In addition, due to the model specifications in (14), a sparse representation in enforced by beta Bernoulli prior for \( Z \).

Given \( W \in \mathbb{R}^{d \times K} \), for each input vector \( x_i \) we seek a sparse zero-one coding \( Z_i \) as well as weight coefficients \( C_i \). The \( Z \)'s specify which factors in \( W \) are used to represent the input data, while the \( C \)'s indicate the weighting of those selected factors. We seek for posterior parameters for \( C \) as well as point estimates for \( Z \).

For each data point \( i \) we define \( q(C_i) = \text{Normal}(C_i | \mu, \Sigma) \).

Here, we only focus on learning the parameters for a specific data point, \( x_i \), being \( Z_i, C_i \). (Therefore, we drop the subscripts below.) The optimization problem corresponding to this portion of the model is:

\[ \min_{Z,\mu,\Sigma} \frac{1}{2\sigma^2} (x - WZ\mu)^T (x - WZ\mu) \]

\[ + \frac{1}{2\sigma^2} \text{trace}(WZ\Sigma W^T) \]

\[ + \frac{1}{2} \lambda^T \mu + \frac{1}{2} \text{trace}(\Sigma) - \frac{1}{2} \log(|\Sigma|) + Z^T h \]

subject to \( Z_{k,k} = 0 \) or 1, \( \Sigma \succeq 0 \) for \( k = 1, \ldots, K \).

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 expression with respect to optimization variables. In addition, the log term is also convex with respect to \( \Sigma \). To make all of the constraint quadratic, we replace the zero or one constraint for \( Z_{k,k} \) to \( Z_{k,k}^2 - Z_{k,k} = 0 \). Therefore, we obtain a non-convex optimization with polynomial terms carrying all of the non-convexities.

Following the steps described in the previous section, we are able to define the convex relaxation optimization for this problem. The novelty introduced here is that we have not relaxed the entire problem globally, which is impossible with a model of this size, but only relaxed locally on the parameters for each observation. After constructing the graph of this problem, we deduced 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, and overcome the complexities issued in variational inference parameter estimations.

### 3.4. CRVI in General Form

Following the idea introduced by the examples, we present CRVI for general form. Let us consider the general variational inference problem in (2). We split the objective into two functions, one for polynomial terms, \( f \), and one for the rest, \( g \). Transforming \( f \) to be a quadratic function possibly with adding new constraints and variables, we would get:

\[ \min_{\phi^{(1)}, \phi^{(2)}} f(\phi^{(1)}) + g(\phi^{(2)}) \]

\[ \text{subject to } \phi^{(1)}, \phi^{(2)} \in \text{feasible set}. \]

(16)

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

\[ \min_{\Phi^{(1)}, \phi^{(2)}} f(\Phi^{(1)}) + g(\phi^{(2)}) \]

\[ \text{subject to } \Phi^{(1)}, \phi^{(2)} \in \text{feasible set}, \]

\[ \Phi^{(1)}_{1,1} = 1, \quad \phi^{(1)} \succeq 0. \]

(17)
If \( g \) is a convex function, (17) 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 \( \Phi^{(1)} \), the more exact approximation. As seen in the examples in the previous sections, many of variational inference problems have this structure, and low rank recovery and near-global optimal solution is guaranteed for them. For the cases where \( g \) is non-convex, CRVI still works fine. We can reduce the hardness related to \( f \) with this relaxation and lifting technique, get approximation bounds, and improve the results compared to the cases where we have to deal with both non-convex \( f \) and \( g \).

### 4. Experimental Results

In this section, we present experimental results on various datasets.

#### 4.1. CRVI for Bayesian Linear Regression with Sparse Priors

In this part, we focus on comparing the optimal value of the variational objective calculated by our method CRVI in section (3.2) and coordinate ascent variational inference (CAVI) which a standard method for variational optimization. We experiment on 9 datasets from UCI repository (Lichman, 2013) with various sizes: {Iris, Birth rate and economic growth, Yacht, Pima Indian diabetes, Bike sharing, Parkinsons data, Wisconsin breast cancer (WDBC), Online news popularity, Year of release prediction for a million songs}. We compared the variational lower bound for 100 different hyper-parameters and initial values for each dataset. Table (1) shows some details about these datasets, average elapsed time for the simulations and average rank of the optimal solution of the 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 \( 28 \times 28 \) to \( 8284 \times 8284 \), average ranks show that regardless of the size of the data, the rank remains small and close to 1. Meaning that the CRVI is able to find near-global optimal solutions noting that rank 1 solution shows an exact relaxation and global optimal solution. To evaluate the improvement for the variational objective function, for each simulation of each dataset, we subtracted the optimum value of CAVI from CRVI, and divided it by optimum value of CRVI to get the relevance improvement in the maximization problem. We show the summary of results in a boxplot for each dataset in Fig. (2). As can be seen, CRVI significantly improved the optimal solution of the optimization which can be interpreted as a more accurate posterior approximation.

#### 4.2. Convex Relaxation for Sparse Coding

In this experiment, we compare the exactness of sparse signal representation in dictionary learning for CRVI with K-SVD (Aharon et al., 2006) on synthetic data. K-SVD uses orthogonal matching pursuit (OMP) to encode each signal using the dictionary (Tropp, 2004). The goal is to compare the number of correct recovered \( Z \) entries where \( N = 300, D = 100, K = 100, \lambda = 0.1 \). We change the sparsity level of generated \( Z \) over different simulations.

In Fig. (3) the x-axis represents the probability of one in every entry of \( Z \) when generating this binary matrix with independent and identically distributed entries, while the y-axis shows the percentage of correctly recovery of \( Z \) entries when learning. As can be seen, CRVI is able to better learn the correct values for \( Z \) by higher accuracy in finding the correct sparsity.
**Table 1.** Information about the datasets, running time of the algorithms, and rank of the optimal solution in CRVI optimizations. Columns of the table represent the followings, respectively: dataset name, dimensionality of each data point, number of samples, average time spent for training CAVI, average time spent for training CVRI, average rank of the optimal matrix calculated in the CRVI optimization.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Dim.</th>
<th># of Samples</th>
<th>CAVI time (s)</th>
<th>CRVI time (s)</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth Rate &amp; Econ</td>
<td>4</td>
<td>30</td>
<td>0.281</td>
<td>1.115</td>
<td>1.11</td>
</tr>
<tr>
<td>Iris</td>
<td>4</td>
<td>150</td>
<td>0.231</td>
<td>1.807</td>
<td>1.20</td>
</tr>
<tr>
<td>Yacht</td>
<td>6</td>
<td>308</td>
<td>0.402</td>
<td>2.111</td>
<td>1.10</td>
</tr>
<tr>
<td>Pima Ind. Diab.</td>
<td>8</td>
<td>768</td>
<td>0.571</td>
<td>3.040</td>
<td>1.67</td>
</tr>
<tr>
<td>Bike Sharing</td>
<td>13</td>
<td>731</td>
<td>0.884</td>
<td>6.749</td>
<td>1.61</td>
</tr>
<tr>
<td>Parkinson</td>
<td>21</td>
<td>5875</td>
<td>0.962</td>
<td>7.309</td>
<td>1.98</td>
</tr>
<tr>
<td>WDBC</td>
<td>31</td>
<td>569</td>
<td>1.059</td>
<td>10.766</td>
<td>1.73</td>
</tr>
<tr>
<td>Online News Pop.</td>
<td>58</td>
<td>39644</td>
<td>9.341</td>
<td>15.223</td>
<td>1.52</td>
</tr>
<tr>
<td>Year Pred. Songs</td>
<td>90</td>
<td>515345</td>
<td>18.809</td>
<td>22.050</td>
<td>1.78</td>
</tr>
</tbody>
</table>

**Figure 3.** The fraction of agreement in recovered $Z$’s with original $Z$ using CRVI and K-SVD. The x-axis shows the probability of 1 in every entry of original matrix $Z$.

5. **Discussion**

CRVI method is a powerful way to approximate (convexify) hard optimization problems. One of the caveats of this method though is its time complexity. The positive semidefinite constraint is the source of this issue. Fortunately, recent advances in this area have suggested faster ways to impose these types of constraints by breaking it 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 as well as diagnosing problematic parts in the graph would significantly improve the performance of this algorithm. One of our future directions is to use the new methodologies to enhance time and space complexities. Another direction would be to make tighter bounds for the relaxation exactness using the treewidth. Finding the exact treewidth of a graph is a NP-hard problem in general, and the bounds given in this paper used the treewidth’s that was readily achievable for us. Needless to mention that there might be better ways to reach smaller treewidth’s and make the theoretical bounds tighter. Obtained ranks in table (1) motivate us to push the theory in this direction as well.

6. **Conclusion**

We presented convex relaxation for variational inference (CRVI) method to infer posterior distributions using mean-field variational inference with a focus on Bayesian linear regression and sparse coding models. By lifting the domain of the optimization, we were able to relax 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 closeness of the obtained solution to the global optimum one. We verified the performance of CRVI on various datasets for sparse Bayesian linear regression problem and sparse coding problem for dictionary learning.

**References**


CRVI: Convex Relaxation for Variational Inference


