Variance Control for Black Box Variational Inference Using The James-Stein Estimator

Dominic B. Dayta
Mathematical Informatics
Nara Institute of Science and Technology
dominic.dayta.da4@naist.ac.jp

Abstract

Black Box Variational Inference is a promising framework in a succession of recent efforts to make Variational Inference more “black box". However, in basic version it either fails to converge due to instability or requires some fine-tuning of the update steps prior to execution that hinder it from being completely general purpose. We propose a method for regulating its parameter updates by reframing stochastic gradient ascent as a multivariate estimation problem. We examine the properties of the James-Stein estimator as a replacement for the arithmetic mean of Monte Carlo estimates of the gradient of the evidence lower bound. The proposed method provides relatively weaker variance reduction than Rao-Blackwellization, but offers a tradeoff of being simpler and requiring no fine tuning on the part of the analyst. Performance on benchmark datasets also demonstrate a consistent performance at par or better than the Rao-Blackwellized approach in terms of model fit and time to convergence.

1 Introduction

Black Box Variational Inference (BBVI) [1] presents a promising alternative to MCMC-based techniques for fitting the posterior distribution in arbitrarily large Bayesian models. In line with the general framework of Variational Inference (VI) [2], BBVI works around the tendency of MCMC solutions to explode in computational complexity by providing an approximate, instead of an exact, solution to the optimal parameters defining the model.

Where exact MCMC methods aim to produce random samples of the posterior distribution, VI works by approximating the posterior with a family of tractable densities, indexed by free parameters known as *variational parameters*. In this way, VI changes the problem from one of sampling (as in MCMC) to one of optimization, to find the variational parameters that make the resulting density as close as possible (in terms of Kullback-Leibler divergence) to the target posterior distribution [2].

However, the process of finding the correct optimization algorithm for the variational parameters can lead to highly complex derivations that are prone to human error, still rendering the process of model exploration quite slow [2]. The promise of BBVI is in removing the need for such derivations, by providing a generalized algorithm for finding the variational parameters for models of any form and size.

Our contribution lies primarily in our demonstration of the variance reduction properties of using the James-Stein estimator in estimating noisy ELBO gradients in the BBVI Update Steps. Through simulation studies, we are able to conclude that the James-Stein estimator yields tightly controlled variance of the noisy ELBO gradients even as the complexity of the model grows.

The rest of this paper is organized as follows: section 2 presents the problem setting, introducing the general algorithm for BBVI as well as its improved version via Rao-Blackwellization. We then

Preprint. Under review.

present in section 3 our proposed improvement using the James-Stein estimator, casting BBVI (and in general, stochastic gradient ascent problems) in the language of classical point estimation theory. Finally, we provide empirical results in 4 through simulated and benchmark datasets involving finite mixtures of Gaussians.

2 Preliminaries

We briefly introduce the general algorithm for BBVI. Suppose we have data $y = \{y_1, y_2, ..., y_n\}$ with n observations for which we have posited some arbitrary model parameterized by θ, with prior density $p(\theta)$. Bayesian data analysis primarily makes use of the posterior density,

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y)$$

where $p(y|\theta)$ denotes the model likelihood, and the marginal distribution $p(y)$ is known as the "evidence".

For models arbitrarily large (or in the case of non-conjugacy), finding the form of $p(\theta|y)$ can be exceedingly complex. In many cases, $p(\theta|y)$ may not even belong to a known family of densities, and the complexity of the model makes applying exact MCMC inference too slow, consequently inhibiting frequent and comprehensive model criticism and exploration. VI proceeds by finding an approximating distribution $q(\theta|\lambda)$ that is much simpler in form, but analytically is nearly identical to the posterior distribution. This is achieved by allowing the parameter λ to vary freely, and finding its value such that the Kullback-Leibler divergence,

$$KL(q(\theta|\lambda)||p(\theta|y)) = \int_\theta \log\left(\frac{q(\theta|\lambda)}{p(\theta|y)}q(\theta|\lambda)\right)d\theta \tag{1}$$

is minimized.

Optimizing the KL-divergence directly can be intractable. Alternatively, one can make use of a quantity referred to in VI and Expectation Propagation literature as the *evidence lower-bound* (ELBO), maximizing on which is equivalent to minimizing the KL-divergence [1]. Via a common decomposition [3] of (1), we can write the logarithm of the evidence $p(y)$ as

$$\log p(y) = \mathcal{L}(\lambda) + KL(q(\theta|\lambda)||p(\theta|y))$$

where

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\log p(y,\theta) - \log q(\theta|\lambda)] \tag{2}$$

The form (2) is called the evidence lower-bound as $\log p(y) = \mathcal{L}(\lambda)$ when q equals the posterior distribution exactly (i.e., $KL(q(\theta|\lambda)||p(\theta|y)) = 0$), and $\log p(y) > \mathcal{L}(\lambda)$ otherwise. Thus, maximizing $\mathcal{L}(\lambda)$ is equivalent to minimizing on $KL(q(\theta|\lambda)||p(\theta|y))$.

2.1 Black Box Variational Inference

Performing VI typically requires finding the appropriate coordinate ascent algorithm needed specific to each model combination [2], but in BBVI this step is conveniently left out, opting instead of a general algorithm that works for most cases. It is shown [1] that the gradient of $\mathcal{L}(\lambda)$ is given by

$$\nabla_\lambda \mathcal{L}(\lambda) = \mathbb{E}_q[\nabla_\lambda \log q(\theta|\lambda)(\log p(y,\theta) - \log q(\theta|\lambda))] \tag{3}$$

Which can now be used in a general gradient ascent algorithm. A sample of S draws $\theta \sim q(\theta|\lambda)$ can be obtained and used to estimate the expectation using

$$\hat{\nabla}_\lambda \mathcal{L} = \frac{1}{S}\sum_{s=1}^{S} \nabla_\lambda \log q(\theta[s]|\lambda)(\log p(y,\theta[s]) - \log q(\theta[s]|\lambda)) \tag{4}$$

In simpler cases the gradient $\nabla_\lambda \log q(\theta|\lambda)$, also known in classical statistical theory as the *score function*, can be obtained analytically, but various autodifferentiation packages have become available for most computational environments such that the algorithm can truly be approached in a black-box manner. We can now present the "Naive" form of BBVI in terms of the stochastic gradient ascent formulation in Algorithm 2 in our Supplementary Materials.

2.2 Variance Control Using Rao-Blackwellization

However, this form of BBVI is noted for either failing to converge or find meaningful solutions within reasonable time due to a high variance in its sampling distribution. This problem is addressed in the original paper through Rao-Blackwellization [4]. First, we note that in most applications there will be more than one parameter $\theta_1, \theta_2, ..., \theta_p$ in the posterior distribution, and generally each of them are assigned their corresponding variational distribution $\lambda_1, \lambda_2, ..., \lambda_p$. The simplest and most commonly used [2] variational family is the *mean-field* family, defined as follows.

Definition 1. *Q is known as the* mean-field variational family *if for all* $q \in Q$*:*

$$q(\theta|\lambda) = \prod_{i=1}^{p} q(\theta_i|\lambda_i) \tag{5}$$

To be sure, the mean-field family is not the only one used when performing either BBVI or VI. Other forms of VI have been proposed and explored in the literature for specific cases where the mean-field assumption may be inappropriate [2]. For instance, *structured variational inference* [5, 6] removes the independence assumption that is inherent in the mean-field flavor family by specifically inducing dependencies between the variational parameters. Another approach is to expand the mean-field family with the addition of latent variables that encode these relationships [7]. We follow the mean-field assumption to facilitate a straightforward optimization problem in this paper, although our proposed black box approach does not specifically require this factorization to hold.

In performing Rao-Blackwellization this factorization is exploited through the average (4), specifically the difference between the log-joint distribution $\log p(y, \theta[s])$ and its approximation $\log q(\theta[s]|\lambda)$, being equal to

$$\log p(y, \theta[s]) - \log q(\theta[s]|\lambda) = \sum_{j=1}^{p} (\log p(y, \theta_{\lambda_j}[s]) - \log q(\theta_{\lambda_j}[s]|\lambda_j))$$

This grows linearly with the number of variational parameters p. Hence, the variance of the gradient estimate $\hat{\nabla}_\lambda \mathcal{L}$ grows linearly as well. More importantly, we see that this growth in variance is mostly unnecessary, as updating a particular parameter λ_j will be based on a gradient whose variance is composed of those for other parameters $\lambda_{j'}$, $j' \neq j$.

Using Rao-Blackwellization, the gradient is estimated for each λ_j parameter conditioned on current values of the other variational parameters. That is, given $q(\theta_{\lambda_j}|\lambda_j)$ as the terms of the approximation that depend only on λ_j, and $p(y, \theta_{\lambda_j})$ as the terms of the joint distribution keeping only θ_{λ_j} depending on λ_j. Using an analogy from graphical models, these remaining terms are referred to as the *Markov Blanket* [8] of λ_j. The update rule is modified in Algorithm 3 in our Supplementary Materials.

This results in best-in-class variance reduction for the algorithm. However, factorizing the joint and variational distributions to obtain the corresponding update steps, while straightforward, requires additional steps for the analyst before conducting BBVI, and can become unnecessarily tedious in the case of large, highly-layered Bayesian models. This step is a significant hurdle in achieving a truly *black-box* algorithm, and so in Section 3, we discuss an alternative that does not require finding the appropriate factorization. We first conclude our discussion of preliminaries with a brief overview of some related work and recent publications that have appeared since [1].

2.3 Related Work

BBVI has received increasing attention in the machine learning literature for its promise of a general algorithm that can be applicable in a wide variety of settings. Recent work have explored improvements to the algorithm by providing adaptive stopping criteria [9], as well as proving convergence guarantees within common expected scenarios [10].

A related work worth mentioning [11] examines using the reparameterization trick [12, 13] to reduce the variance of the gradient estimates used for the algorithm's update step. However, the reparametrization trick is not strictly a method for variance control in BBVI. Instead, it is a method for changing the parameters of a learning parameter to remove constraints that might hinder computation. This means that both Rao-Blackwellization and the James-Stein estimator can be used on the gradient estimators under reparametrization. We therefore focus primarily on Rao-Blackwellization as a benchmark for our analysis.

Similarly, our proposed methodology remains sufficiently general such that it should be straightforward to combine with other emerging practices being proposed and tested in more recent works. For instance, it is possible to apply our regularized ELBO gradient estimate with the automated stopping and estimate correction logic proposed by [9]. In the following paper, we have returned to the basics of BBVI to ensure that any variance reduction observed can confidently be attributed to our proposed estimation method, and not as a side effect of competing layers in the algorithm.

Our proposal for the use of the James-Stein estimator is motivated by the logic of biasing the noisy gradient estimate in stochastic gradient descent/ascent problems by preventing exploding gradients from straying parameter updates into problematic regions. This is not a new idea in the deep learning literature, where heuristics like gradient clipping [14, 3] are already established with proven convergence guarantees. We explore this connection further in section 3.2.

3 Variance Control Using The James-Stein Estimator

We now propose our James-Stein BBVI with the objective of performing general variational inference without requiring the analyst to find the necessary factorizations for BBVI-RB. This is achieved through the James-Stein estimator. To better understand the motivation behind this proposal, we recast the problem of gradient ascent in BBVI as an estimation problem.

3.1 Gradient Ascent As Estimation

In Algorithms 2 and 3, we perform gradient ascent of the form

$$\lambda^t = \lambda^{t-1} + \rho^t \hat{\nabla}_\lambda \mathcal{L}(\lambda^{t-1}) \tag{6}$$

where $\hat{\nabla}_\lambda \mathcal{L}(\lambda^{t-1})$ is obtained via Monte Carlo samples as Equation (3) is intractable. The noise in the ELBO estimate is due to this stochastic approach. Hence, we can consider this and the general stochastic gradient ascent/descent problem as one of estimating a fixed but unknown gradient $\mu = \nabla_\lambda \mathcal{L}$. In consequence, it is feasible to borrow established techniques from statistical estimation theory [15] for further constraining and regulating the behavior of (6).

We introduce the following necessary assumption, which is shared with the proof used by [11] to demonstrate the variance reduction properties of the reparametrization trick.

Assumption 1. *Given a sample z_s, for $s = 1, 2, ..., S$, the sample average*

$$\hat{\mu} = \frac{1}{S}\sum_{s=1}^{S} z_s \sim \mathcal{N}(\mu, \sigma^2)$$

for

$$z_s = \nabla_\lambda \log q(\theta[s]|\lambda)(\log p(y, \theta[s]) - \log q(\theta[s]|\lambda))$$

and $S \to +\infty$.

Our confidence in the applicability of Assumption 1 rests in the Central Limit Theorem. We observe that the estimator for the gradient is merely a simple average over independent, identically distributed observations of z_s. Thus, for $S \to \infty$ sufficiently large, normality can reasonably be expected to hold.

We see then that BBVI-Naive in Algorithm 2 can be recast as a maximum likelihood estimator, $\hat{\mu}_{MLE}$, as re-stated in the following theorem.

Theorem 1 (BBVI as MLE Estimator). *BBVI-Naive, which we now denote as $\hat{\mu}_{MLE}$ is the Maximum Likelihood estimator of $\mu = \nabla_\lambda \mathcal{L}$, where*

$$\hat{\mu}_{MLE} = \frac{1}{S}\sum_{s=1}^{S} z_s \tag{7}$$

for $z_s = \nabla_\lambda \log q(\theta[s]|\lambda)(\log p(y, \theta[s]) - \log q(\theta[s]|\lambda))$. Furthermore, $\hat{\mu}_{MLE}$ is unbiased to the true gradient μ.

The proof immediately follows from Assumption 1. [1] supports the unbiasedness of this estimator, showing that $E(\hat{\mu}_{MLE}) = \mu$. However, it is well known [15, 16] that for $p > 2$ dimensions, the MLE estimator is dominated in mean square error (MSE) by the James-Stein estimator, specifically the Positive Part James-Stein estimator.

Theorem 2 (Positive Part James-Stein Estimator)**.** *The Positive-Part James-Stein estimator $\hat{\mu}_{JS+}$ given by*

$$\hat{\mu}_{JS+} = \left(1 - \frac{(p-3)\sigma^2}{||\bar{z}||^2}\right)^+ \bar{z}$$

for $\bar{z} = \frac{1}{S}\sum_s z_s$ and $(g)^+ = gI_{[0,+\infty)}(g)$ dominates $\hat{\mu}_{MLE}$ in MSE.

The proof for Theorem 2 is already a canonical result in estimation theory and can be obtained from [15], while an Empirical Bayes approach can be found in [16]. The algorithm BBVI-JS+ is summarized in Algorithm 1.

Algorithm 1: Positive-Part James-Stein BBVI (BBVI-JS+)

```
Input   :Model, Monte Carlo Sample Size $S$, convergence threshold $\varepsilon$, learning rate $\rho^t$
Initialize $\lambda^0$ randomly, set $t = 0$ and $\Delta = \infty$
while $\Delta > \varepsilon$ do
    $t = t + 1$
    // Draw $S$ samples from $q(\theta|\lambda^{t-1})$
    for s = 1 to S do
        $\theta[s] \sim q(\theta|\lambda^{t-1})$
    end
    $\bar{z} = \frac{1}{S}\sum_{s=1}^{S} \nabla_\lambda \log q(\theta[s]|\lambda^{t-1})(\log p(y, \theta[s]) - \log q(\theta[s]|\lambda^{t-1}))$
    $\hat{\nabla}_\lambda \mathcal{L}(\lambda^{t-1}) = \left(1 - \frac{(p-3)\sigma^2}{||\bar{z}||^2}\right)^+ \bar{z}$
    $\lambda^t = \lambda^{t-1} + \rho^t \hat{\nabla}_\lambda \mathcal{L}(\lambda^{t-1})$
    $\Delta = \frac{||\lambda^t - \lambda^{t-1}||}{||\lambda^{t-1}||}$
end
```

Having defined our proposed estimator, we can now make quantitative statements about its relationship with both the Naive and Rao-Blackwellized versions of BBVI.

Theorem 3 (Variance Reduction of the James-Stein Estimator)**.** *Given the Naive BBVI/MLE Estimator $\hat{\mu}_{MLE}$ and the Positive-Part James-Stein estimator $\hat{\mu}_{JS+}$, then*

$$V(\hat{\mu}_{JS+}) < V(\hat{\mu}_{MLE})$$

Proof. Due to the MSE dominance of $\hat{\mu}_{JS+}$ over $\hat{\mu}_{MLE}$, and taking advantage of the Bias-Variance decomposition of MSE, we have that

$$V(\hat{\mu}_{JS+}) + \text{Bias}^2(\hat{\mu}_{JS+}) \leq V(\hat{\mu}_{MLE})$$

as $\hat{\mu}_{MLE}$ is an unbiased estimator. From the above proof we find that the James-Stein estimator, when applied to BBVI, should be able to control the sampling distribution of $\hat{\nabla}\mathcal{L}$ in Equation (6). The shrinkage factor allows the parameter λ^t to move when the gradient is relatively small, but forces it to remain near or at its previous value when the gradient explodes.

However, $\hat{\mu}_{JS+}$ generally performs worse than Rao-Blackwellization $\hat{\mu}_{RB}$ as shown in the following theorem.

Theorem 4. *Given the Positive-Part James-Stein estimator $\hat{\mu}_{JS+}$ and the Rao-Blackwellized BBVI $\hat{\mu}_{RB}$*

$$V(\hat{\mu}_{RB}) \leq V(\hat{\mu}_{JS+})$$

Proof. Under the factorized case (which is necessary anyway when using $\hat{\mu}_{RB}$), we can split the summation for some parameter λ_j

$$\sum_{s=1}^{S} \nabla_{\lambda} \log q(\theta[s]|\lambda^{t-1})(\log p(y, \theta[s]) - \log q(\theta[s]|\lambda^{t-1})) =$$

$$\sum_{s=1}^{S} \nabla_{\lambda_j} \log q(\theta_{\lambda_j}[s]|\lambda_j^{t-1})(\log p(y, \theta_{\lambda_j}[s]) - \log q(\theta_{\lambda_j}[s]|\lambda_j^{t-1}))$$

$$+ \sum_{s=1}^{S} \nabla_{\lambda_{-j}} \log q(\theta_{-\lambda_j}[s]|\lambda_{-j}^{t-1})(\log p(y, \theta_{\lambda_{-j}}[s]) - \log q(\theta_{-\lambda_j}[s]|\lambda_{-j}^{t-1}))$$

We can think of the summation in the second term of the right-hand side as an *anti-Markov blanket* to the parameter λ_j. Now, holding the norm $||\bar{z}||^2$ constant,

$$V(\hat{\mu}_{JS+}) = k \times V(\hat{\mu}_{RB}) + C$$

where

$$k = \left[\left(1 - \frac{(p-3)\sigma^2}{||\bar{z}||^2} \right)^+ \right]^2$$

is a value constrained to $[0, 1]$ and

$$C = k \times V\left(\frac{1}{S} \sum_{s=1}^{S} \nabla_{\lambda_{-j}} \log q(\theta_{-\lambda_j}[s]|\lambda_{-j}^{t-1})(\log p(y, \theta_{\lambda_{-j}}[s]) - \log q(\theta_{-\lambda_j}[s]|\lambda_{-j}^{t-1})) \right)$$

being the variance of the anti-Markov Blanket of λ_j approaches infinity with greater and greater p. Hence, the theorem holds for $C \to +\infty$. Our proof has the interesting implication that it is, in fact, possible for $\hat{\mu}_{JS+}$ to outperform $\hat{\mu}_{RB}$ in variance, provided that $C \to 0$. In practice this is only possible when there are only very small number of variational parameters, and should be very rare (if it ever occurs) in practice.

Also a consequence of this theorem is an interesting behavior that is achieved when applying the positive-part James-Stein shrinkage factor to the Rao-Blackwellized estimator.

Corollary 1. *Suppose we have a Positive-Part Rao-Blackwellized estimator $\hat{\mu}_{RB+}$ given by*

$$\hat{\mu}_{RB+} = \left(1 - \frac{(p-3)\sigma^2}{||\hat{\mu}_{RB}||^2} \right)^+ \hat{\mu}_{RB}$$

Then its variance

$$V(\hat{\mu}_{RB+}) \leq V(\hat{\mu}_{RB})$$

Proof. We note that

$$V(\hat{\mu}_{RB+}) = k \times V(\hat{\mu}_{RB})$$

This means that a positive-component James Stein estimator applied on the Rao-Blackwellized estimator further constricts the variance of the noisy ELBO gradient estimate.

3.2 Relationship to Gradient Clipping

The idea of regulating the path of estimated parameters in stochastic gradient ascent/descent problems (6) directly through the gradient estimate is not new within the deep learning literature. In the following section, we discuss the connection between the form we have proposed in Algorithm 1 with the method of gradient clipping in training deep neural networks.

Clipping mitigates the issue of exploding gradients, in which estimates of the gradients can become very large during training, leading to instability and straying the parameter updates from convergence.

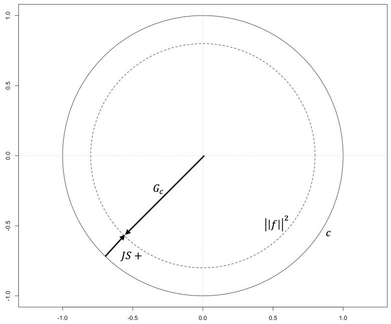

Figure 1: Relationship between gradient clipping and the James-Stein estimator. Gradient clipping G_c preserves values of the gradient only up to $||f||^2 \leq c$. The Positive Part James-Stein operator $JS+$ penalizes $||f||^2$ for being close to c and forces it towards zero.

The gradient estimate $\hat{\nabla}\mathcal{L}$ is effectively constricted by a pre-set radius c (which can be a learnable parameter), such that $\hat{\nabla}\mathcal{L}$ can only take on values up to c. Formally, we can define a clipping function

$$G_c(f) = \min\left(1, \frac{c}{||f||}\right)$$

and modify Equation (6) to

$$\lambda^t = \lambda^{t-1} + \rho^t G_c(\hat{\nabla}\mathcal{L}(\lambda^{t-1}))\hat{\nabla}\mathcal{L}(\lambda^{t-1})$$

The gradient used in each update step can only be a fraction $\frac{c}{||\hat{\nabla}\mathcal{L}(\lambda^{t-1})||}$ of the actual gradient value whenever $\hat{\nabla}\mathcal{L}(\lambda^{t-1}) > c$. In better-behaved iterations when $\hat{\nabla}\mathcal{L}(\lambda^{t-1}) < c$, the update step is allowed to use the full value of the gradient. With this constraint, clipping prevents the gradients from growing too large, thereby stabilizing the training process. In deep neural networks, clipping limits the influence of any single training sample or layer on the overall parameter updates, leading to more stable training.

We find that it is trivial to suppose a modified form given by

$$G_c(f) = \min\left(1, \frac{c}{||f||^2}\right) \tag{8}$$

which simply means re-scaling the radius c to be in units of the squared norm of f. We now observe the following relationship between clipping and our method.

Theorem 5. *The Positive-Part James-Stein estimator $\hat{\mu}_{JS+}$ represents a reversal of the modified gradient clipping function (8). That is, the shrinkage operator,*

$$JS(f) = \left(1 - \frac{(p-3)\sigma^2}{||f||^2}\right)^+ = 1 - G_c(f)$$

Proof. [16] show that the shrinkage operator

$$\left(1 - \frac{(p-3)\sigma^2}{||f||^2}\right)^+ = 1 - \min\left(1, 1 - \frac{(p-3)\sigma^2}{||f||^2}\right)$$

which for $c = ||f||^2 - (p-3)\sigma^2$ satisfies the theorem.

This connection is illustrated in Figure 1. Gradient clipping is not a shrinkage method, as it does not force the gradient to zero. Instead, clipping is concerned with keeping f within a region such that its squared norm $||f||^2 \leq c$. On the other hand, the James Stein estimator is explicitly a shrinkage method, and it forces the gradient to be as small as possible, imposing a penalty for being close to the limit c.

We can then contrast our method from clipping by framing it within a developing framework within the larger field of Bayesian Optimization [17] of keeping updates to stay as close as possible to previous observations [18]. Whereas clipping is largely a heuristic to prevent exploding gradients, application of the James-Stein estimator represents a prior belief that the correct update step is likely to be small.

3.3 Further Extensions

So far we have maintained ambiguity regarding the learning rate ρ^t. In practice, the exact value used for this learning rate contributes significantly to the success of any stochastic gradient ascent/descent problem. It is maintained [1] that BBVI will converge to its optimal values for a general class of ρ^t, in fact requiring only that it follow the Robbins-Monro [19] conditions. In their extensions, it is recommended that ρ^t can follow the AdaGrad algorithm [20, 3], as a way of adjusting the learning rate of each parameter dynamically during training. This is achieved by scaling the learning rate for each parameter based on the historical values of the gradient.

However, as a side effect of the biasing done by the James-Stein estimator on the gradient estimates, we find that AdaGrad may be too aggressive in reducing the learning rate towards the latter runs of the algorithm. As the biased gradient means that the algorithm is forced to consider smaller steps, monotonically scaling down the learning rate too fast may prevent the algorithm from reaching an optimal convergence point. Hence, we propose using RMSProp [3] instead, which adds an extra parameter β that exponentially decreases the impact of the earlier iterations,

$$G^t = \beta G^{t-1} + (1-\beta) g^t (g^t)^T$$
$$\rho^t = \eta \text{diag}(G^t + I\xi)^{-1/2}$$

where g^t represents the gradient estimate at the current step t, $G^0 = 0$, and ξ is some small positive number to prevent division by zero. The parameters β and η can both be considered as learnable parameters, though in practice $\beta = 0.9$ or some similarly high number. This decay factor limits the accumulation of historical gradients, and ensures that the learning rates do not become too small over time.

4 Experiments

We perform experiments using simulated and benchmark datasets for finite Gaussian mixture models to demonstrate the performance of our proposed algorithm. For demonstrating variance reduction properties, we use simulated data following a univariate mixture of Gaussians and show that the sampling distribution is constricted in BBVI-JS+. We also apply the algorithm to a set of benchmark datasets for clustering tasks provided in the FCPS [21] package. All computations were performed on R version 4.2.3 running under MacOS 14.4.1.

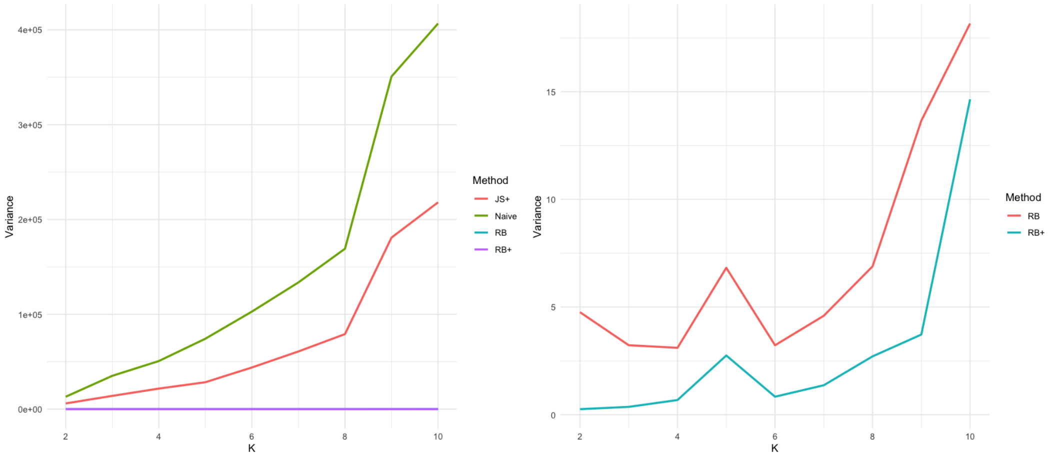

Figure 2: Resulting estimator variances from the Gaussian Mixture experiment with $K = 2$ to 10 components. BBVI-JS+ produces controlled variances in its sampling distribution for the ELBO gradient compared to BBVI-Naive, but relative to BBVI-RB still grows with the number of parameters. Interestingly, BBVI-RB+ provides even stricter variance control over BBVI-RB.

Simulated Data. We follow the formulation presented in [22] except the Dirichlet prior over the mixture components as well as the Inverse-Gamma prior on the variance τ^2. This model has been selected as it permits mimicking the behavior of more complex models easily by adding more components.

Figure 2 shows the resulting variances of the Naive, James-Stein (JS+), Rao-Blackwellized (RB) as well as James-Stein applied on the Rao-Blackwellized (RB+) BBVI. We find that as the number of components of the mixture distribution, the variance of BBVI-Naive increases linearly as expected. On the other hand, the variance of both RB and RB+ versions remain controlled to several orders of magnitude. Between these two bounds, we have BBVI-JS+ remaining within a level that is found to be between 38 to 46% of BBVI-Naive in terms of relative efficiency. This confirms our result in Theorems 3 and 4. Also in terms of relative efficiency, the performance of BBVI-RB is at around 0.00 to 0.04% of BBVI-Naive, while BBVI-RB+ is practically at 0.00%.

Benchmarks. Focusing now on the BBVI-JS+ and BBVI-RB algorithms, we make use of three benchmark datasets found in the `FCPS` [21] package: EngyTime, Lsun3D, and Tetra. To accommodate these datasets, we extend the univariate gaussian mixture model we used for our simulations to their multivariate counterparts. The results for these benchmarks are provided in Table 1.

Table 1: Time to convergence and fit criteria of Rao-Blackwellized and James-Stein BBVI in three benchmark datasets. For each run, we force the algorithm to take at least 100 iterations before assessing convergence to allow ample warm-up. Due to the differing magnitudes of the gradients, the parameter η for RMSProp between the Rao-Blackwellized and James-Stein have been adjusted.

Benchmark	**Method**	η	β	**K**	**p**	**Fit**				
						Iter	**Time**	**ELBO**	**LogLik**	**DIC**
EngyTime	BBVI-RB	1.0	0.9	2	2	200	13.09	-2,268.88	-2,516.83	4,535.48
	BBVI-JS+	0.1	0.9	2	2	101	4.27	-2,231.65	-2,363.15	4,459.48
Lsun3D	BBVI-RB	1.0	0.9	4	2	113	7.81	-1,921.70	-2,197.60	3,837.53
	BBVI-JS+	0.1	0.9	4	2	113	4.69	-1,859.93	-1,795.25	3,367.49
Tetra	BBVI-RB	1.0	0.9	4	3	398	24.34	-3,197.82	-4,238.86	6,389.14
	BBVI-JS+	0.1	0.9	4	3	149	5.80	-2,470.81	-2,812.26	4,556.51

Results on the benchmark show that BBVI-JS+ combined with the RMSProp learning rate generally reaches convergence faster than BBVI-RB even with its larger sampling distribution for the ELBO gradient estimate. This performance is attributed to two key factors: because the ELBO gradients are still computed as a whole, there is no need to cycle through the entire dataset for computing the difference factor in Equation (3). Moreover, the shrinkage penalty in BBVI-JS+ appear to have allowed the parameter to update slowly and with smaller steps per iteration, keeping the algorithm from performing massive U-Turns in resulting ELBO. Also seen in Table 1 is that BBVI-JS+ consistently resulted in higher ELBO, and lower DIC than BBVI-RB, although the differences are not very vast. At the very least, these results demonstrate that BBVI-JS+ is able to perform at least at the level of BBVI-RB in coming up with optimal approximations to target posterior densities.

5 Conclusions

We have proposed a method of controlling for the variance of the noisy ELBO gradient estimates in Black Box Variational Inference by first casting the stochastic gradient ascent problem as one of estimating a true gradient at each iteration. Borrowing from an established property of multivariate estimators, we proposed a shrinkage operator in the form of the Positive Part James-Stein estimator to bias the gradients towards zero. The result is the inclusion of a prior belief at each iteration that each parameter's update step should be small. Theoretical and empirical results confirm that such a behavior results in narrower sampling distributions for the estimated gradients, and in consequence more stable paths towards optimal values of the variational parameters.

References

[1] Rajesh Ranganath, Sean Gerrish, and David Blei. Black Box Variational Inference. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland, 22–25 Apr 2014. PMLR. URL `https://proceedings.mlr.press/v33/ranganath14.html`.

[2] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[4] George Casella and Christian P. Robert. Rao-blackwellisation of sampling schemes. *Biometrika*, 83(1): 81–94, 1996. ISSN 00063444. URL http://www.jstor.org/stable/2337434.

[5] Lawrence Saul and Michael Jordan. Exploiting tractable substructures in intractable networks. In D. Touretzky, M.C. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995. URL https://proceedings.neurips.cc/paper_files/paper/1995/file/285f89b802bcb2651801455c86d78f2a-Paper.pdf.

[6] David Barber and Wim Wiegerinck. Tractable variational structures for approximating graphical models. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1998. URL https://proceedings.neurips.cc/paper_files/paper/1998/file/297fa7777981f402dbba17e9f29e292d-Paper.pdf.

[7] Christopher Bishop, Neil Lawrence, Tommi Jaakkola, and Michael Jordan. Approximating posterior distributions in belief networks using mixtures. In M. Jordan, M. Kearns, and S. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1997. URL https://proceedings.neurips.cc/paper_files/paper/1997/file/c0826819636026dd1f3674774f06c51d-Paper.pdf.

[8] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 1558604790.

[9] Manushi Welandawe, Michael Riis Andersen, Aki Vehtari, and Jonathan H. Huggins. Robust, automated, and accurate black-box variational inference, 2022.

[10] Justin Domke. Provable smoothness guarantees for black-box variational inference. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2587–2596. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/domke20a.html.

[11] Ming Xu, Matias Quiroz, Robert Kohn, and Scott A. Sisson. Variance reduction properties of the reparameterization trick. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2711–2720. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/xu19a.html.

[12] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[13] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/rezende14.html.

[14] Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees, 2023.

[15] Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer-Verlag, New York, NY, USA, second edition, 1998.

[16] Bradley Efron and Carl Morris. Stein's estimation rule and its competitors–an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973. ISSN 01621459. URL http://www.jstor.org/stable/2284155.

[17] Roman Garnett. *Bayesian Optimization*. Cambridge University Press, 2023.

[18] Michael Y. Li and Ryan P. Adams. Explainability constraints for bayesian optimization. In *6th ICML Workshop on Automated Machine Learning*, 2019.

[19] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

[20] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL http://jmlr.org/papers/v12/duchi11a.html.

[21] Michael Christoph and Quirin Stier. Fundamental clustering algorithms suite. *SoftwareX*, 13:100642, 2021. ISSN 2352-7110. doi: 10.1016/j.softx.2020.100642. URL https://doi.org/10.3390/pr9101697.

[22] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

A Algorithm Listings

In order to preserve space, we have left out the complete algorithm listing for two already known variants of BBVI, particularly BBVI-Naive (Algorithm 2) and BBVI-RB (Algorithm 3).

Algorithm 2: Naive Black Box Variational Inference (BBVI-Naive) [1]

```
Input   :Model, Monte Carlo Sample Size $S$, convergence threshold $\varepsilon$, learning rate $\rho^t$
Initialize $\lambda^0$ randomly, set $t = 0$ and $\Delta = \infty$
while $\Delta > \varepsilon$ do
    $t = t + 1$
    // Draw $S$ samples from $q(\theta|\lambda^{t-1})$
    for $s$ = 1 to $S$ do
        $\theta[s] \sim q(\theta|\lambda^{t-1})$
    end
    $\hat{\nabla}_\lambda \mathcal{L}(\lambda^{t-1}) = \frac{1}{S} \sum_{s=1}^{S} \nabla_\lambda \log q(\theta[s]|\lambda^{t-1})(\log p(y, \theta[s]) - \log q(\theta[s]|\lambda^{t-1}))$
    $\lambda^t = \lambda^{t-1} + \rho^t \hat{\nabla}_\lambda \mathcal{L}(\lambda^{t-1})$
    $\Delta = \frac{||\lambda^t - \lambda^{t-1}||}{||\lambda^{t-1}||}$
end
```

Algorithm 3: Rao-Blackwellized BBVI (BBVI-RB) [1]

```
Input   :Model, Monte Carlo Sample Size $S$, convergence threshold $\varepsilon$, learning rate $\rho^t$
Initialize $\lambda^0$ randomly, set $t = 0$ and $\Delta = \infty$
while $\Delta > \varepsilon$ do
    $t = t + 1$
    // Draw $S$ samples from $q(\theta|\lambda^{t-1})$
    for $s = 1$ to $S$ do
        $\theta[s] \sim q(\theta|\lambda^{t-1})$
    end
    for $j = 1$ to $p$ do
        $\hat{\nabla}_{\lambda_j} \mathcal{L}(\lambda_j^{t-1}) = \frac{1}{S} \sum_{s=1}^{S} \nabla_{\lambda_j} \log q(\theta_{\lambda_j}[s]|\lambda_j^{t-1})(\log p(y, \theta_{\lambda_j}[s]) - \log q(\theta_{\lambda_j}[s]|\lambda_j^{t-1}))$
        $\lambda_j^t = \lambda_j^{t-1} + \rho^t \hat{\nabla}_{\lambda_j} \mathcal{L}(\lambda_j^{t-1})$
    end
    $\Delta = \frac{||\lambda^t - \lambda^{t-1}||}{||\lambda^{t-1}||}$
end
```

B Experimental Setup

All experiments reported in this paper were conducted on an M1 MacBook Air running `R` version 4.2.3 under MacOS 14.4.1.

Simulated Dataset. The simulated dataset for demonstrating variance reduction in BBVI-JS+ was done by sampling a total of $N = 200$ observations from a finite mixture of Gaussians with $K = 2$ to 10 components, means and variances deterministically set at $\mu_k = \{-5, -4, -3, -2, -1, 0, +1, +2, +3, +4\}$ and common $\sigma^2 = 3$, respectively. During Monte Carlo estimation, a total sample size of $S = 500$ were drawn, and the sampling distributions reported are based on bootstrap samples of size $B = 100$.

Our generative model is given by

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \tau^2) \\ z_i &\sim \text{Categorical}(1/K, ..., 1/K) \\ y_i | z_{ik} = 1, \mu, \tau^2 &\sim \mathcal{N}(\mu_k, \sigma^2) \end{aligned}$$

We can propose mean-field variational approximations

$$\begin{aligned} q(\mu_k | m_k, s_k^2) &= \mathcal{N}(m_k, s_k^2) \\ q(z_i | \phi_i) &= \text{Categorical}(\phi_i) \end{aligned}$$

Benchmark Datasets. To demonstrate convergence and final model fit statistics resulting from using BBVI-JS+, we use three datasets for clustering and Gaussian mixture model tasks provided in the `FCPS` package [21] for `R`. Specific information regarding these datasets can be found in the package documentation. A specific note is made for the `EngyTime` dataset, which originally contains 4,096 observations. To ease computation time, we use a random subset containing 400 randomly selected observations. Scatterplots of the datasets used are in Figure 3.

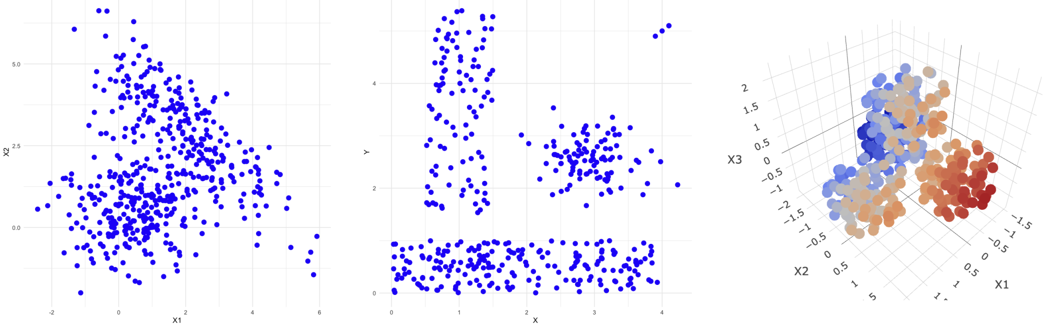

Figure 3: Scatterplots of benchmark datasets from the `FCPS` package [21]. From left to right: EngyTime, Lsun3D, and Tetra.

To prevent early convergence of the model at sub-optimal locations, we force the models to run for at least 100 iterations before assessing convergence. This is due to an observed tendency of BBVI in general to first oscillate within sub-optimal solutions, likely as a consequence of initializations, before taking a more consistent trajectory towards maximum ELBO. We note that selection of proper convergence criteria for BBVI is currently an active research problem [1, 9]. Nevertheless, if a model has already reached convergence before 100 iterations, then it is expected that it should stop within a few iterations after the warm-up. A convergence criterion of $\varepsilon = 0.01$ is used for the `EngyTime` dataset, and $\varepsilon = 0.1$ for the `Lsun3D` and `Tetra` datasets. In each dataset, the convergence criterion used for BBVI-JS+ and BBVI-RB are always the same.