Convergence of Proximal-Gradient Stochastic Variational Inference under Non-Decreasing Step-Size Sequence

Mohammad Emtiyaz Khan
Ecole Polytechnique Fédérale de Lausanne
Lausanne Switzerland
emtiyaz@gmail.com

Reza Babanezhad
University of British Columbia
Vancouver, Canada
babanezhad@gmail.com

Wu Lin
University of Waterloo
Waterloo, Canada
wu.lin@uwaterloo.ca

Mark Schmidt
University of British Columbia
Vancouver, Canada
schmidtm@cs.ubc.ca

Masashi Sugiyama
University of Tokyo
Tokyo, Japan
sugi@k.u-tokyo.ac.jp

Abstract

Stochastic approximation methods have recently gained popularity for variational inference, but many existing approaches treat them as "black-box" tools. Thus, they often do not take advantage of the geometry of the posterior and usually require a decreasing sequence of step-sizes (which converges slowly in practice). We introduce a new stochastic-approximation method that uses a proximal-gradient framework. Our method exploits the geometry and structure of the variational lower bound, and contains many existing methods (like stochastic variational inference) as special cases. We establish the convergence of our method under a *non-decreasing* step-size schedule, which has both theoretical and practical advantages. We consider setting the step-size based on the continuity of the objective and the geometry of the posterior, and experimentally show that our method gives a faster rate of convergence for variational-Gaussian inference than existing stochastic methods.

1 Introduction

Stochastic methods have recently gained popularity as a method for variational inference to maximize lower bounds to marginal likelihoods [1]. Stochastic-approximation gradient-descent (SGD) methods have now been extensively applied for variational inference in latent-variable models [2, 3, 4, 5, 6, 7]. These methods scale well to large datasets and are widely applicable due to their simplicity. However, such "black-box" approaches do not exploit the structure of the variational objective function and usually converge slowly. For example, the variational objective often consists of both convex and non-convex parts and exploiting this structure could improve convergence.

Another problem with existing black-box methods is that most of them ignore the geometry of the variational parameter space. One of the most popular methods, stochastic variational inference (SVI), does take the geometry into account by using natural gradients [1], but unfortunately can only be applied to a limited class of models, such as conditionally-conjugate exponential-family distributions. On the other hand, exploiting Bregman divergences to adapt to the structure of problems is quite popular in the optimization community, where it is known as mirror descent [8]. However, there are only a small number of variational methods that exploit the geometry of the problem in such a way (e.g. [9, 10]).

As a result, many existing stochastic methods lack a principled approach for step-size selection and usually suffer from slow convergence. In most cases, convergence is only guaranteed under a decreasing step-size sequence [11]. It is hard to find a good schedule of step sizes in practice and this typically leads to poor practical performance. Many approaches rely on automatic methods for step-size selection (e.g. AdaGrad [12] and ADADELTA [13]) which, when used as a black-box, do not exploit the structure and geometry of the problem.

In this paper, we propose a stochastic-approximation variational method based on a proximal-gradient framework. The proximal-gradient framework splits the variational bound into convex and non-convex parts, thereby taking the structure of the lower bound into account. In this framework we can use a divergence function, like the Kullback-Leibler (KL) divergence or other Bregman divergences, to incorporate the geometry of the posterior in the variational objective. By doing this, we get existing methods like SVI as special cases of our method. Each step in our method corresponds to solving a simple problem where the non-convex part is linearized and for which closed-form expressions often exist.

We establish the convergence of proximal-gradient stochastic variational methods under very general conditions. We prove that in many cases these stochastic methods can converge with a constant step-size and thus *do not require the decreasing step-size sequences* that destroy practical performance. For example, when the posterior belongs to an exponential family, the step-size can be set to be α_*/L. Here, L is the Lipschitz constant of the gradient of the non-convex part and α_* is a constant related to the partition function of the exponential family.

Background on variational inference and notation: We first briefly describe the model set-up. Consider a general latent variable model with a data vector $\mathbf{y}$ of length N and a latent vector $\mathbf{z}$ of length D. The joint distribution under the model is denoted by $p(\mathbf{y}, \mathbf{z})$. The evidence lower bound optimization (ELBO) approximates the posterior $p(\mathbf{z}|\mathbf{y})$ by a distribution $q(\mathbf{z}|\boldsymbol{\lambda})$ that maximizes a lower bound to the marginal likelihood as shown below:

$$\log p(\mathbf{y}) = \log \int q(\mathbf{z}|\boldsymbol{\lambda}) \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z}|\boldsymbol{\lambda})} \, d\mathbf{z} \geq \max_{\boldsymbol{\lambda} \in \mathcal{S}} \left\{ \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}[\log p(\mathbf{y}, \mathbf{z})] - \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}[\log q(\mathbf{z}|\boldsymbol{\lambda})] \right\}. \tag{1}$$

Here, $\boldsymbol{\lambda}$ is the set of variational parameters. We denote the term inside the max by $\underline{\mathcal{L}}(\boldsymbol{\lambda})$.

2 Proximal-Gradient Stochastic Variational Inference

In this paper, we propose a method to address the three issues discussed in the previous section: (1) exploiting the structure of the lower bound, (2) exploiting the geometry of the posterior, and (3) setting the step-size using the structure and geometry.

Composite structure of the lower bound: A function can always be expressed as the sum of convex and non-convex 'parts'. For the variational lower bound, such splits naturally occur. This is due to the presence of the second term of (1), which is the entropy of q. For the negative of the variational lower bound we denote convex part by h and non-convex part by f, as shown below: $-\underline{\mathcal{L}}(\lambda) := f(\boldsymbol{\lambda}) + h(\boldsymbol{\lambda})$. For example, for a conditionally-conjugate exponential family (using the notation of [14]), the second term is convex in the lower bound shown below:

$$-\underline{\mathcal{L}}(\boldsymbol{\lambda}_i) := (\boldsymbol{\lambda}_i^* - \boldsymbol{\lambda}_i)^T \bigtriangledown A_i(\boldsymbol{\lambda}_i) + A_i(\boldsymbol{\lambda}_i), \tag{2}$$

where $\boldsymbol{\lambda}_i^*$ is the mean-field update for the *variational* parameter $\boldsymbol{\lambda}_i$ of $q(\mathbf{z}_i|\boldsymbol{\lambda}_i)$, the i'th latent variable in a Bayesian network, and A_i is the partition function of the exponential family (see Appendix A.1 and A.2 in [14] for details).

Geometry of the posterior $q(\mathbf{z}|\boldsymbol{\lambda})$: The geometry of the posterior distribution can be incorporated using divergence measures. We will denote the divergence between two distributions $q(\mathbf{z}|\boldsymbol{\lambda})$ and $q(\mathbf{z}|\boldsymbol{\lambda}')$ by $\mathbb{D}(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$. For exponential family distributions, there are natural alternatives to using the squared Euclidean distance. For example, the KL divergence which is defined as:

$$\mathbb{D}_{KL}(q(\mathbf{z}|\boldsymbol{\lambda}) \,\|\, q(\mathbf{z}|\boldsymbol{\lambda}')) := A(\boldsymbol{\lambda}') - A(\boldsymbol{\lambda}) - \bigtriangledown A(\boldsymbol{\lambda})(\boldsymbol{\lambda}' - \boldsymbol{\lambda}). \tag{3}$$

The so-called 'Bregman' divergence defines another class of divergence functions. For exponential family distributions, it is equal to the KL divergence with swapped natural parameters: $\mathbb{D}_{Breg}(\boldsymbol{\lambda}' \| \boldsymbol{\lambda}) = \mathbb{D}_{KL}(\boldsymbol{\lambda} \,\|\, \boldsymbol{\lambda}')$. Finally, the symmetric-KL divergence $\mathbb{D}_{KL}^{sym}(\boldsymbol{\lambda} \| \boldsymbol{\lambda}')$ used in SVI is equal to the sum of the KL divergence and the 'Bregman' divergence for exponential family distributions. Note that we get back the standard *prox* operator if we use the Euclidean distance instead of a divergence function. Thus, introducing a divergence function can be viewed as using a different *prox* operator.

Stochastic-Approximation: We compute a stochastic approximation to the gradient of the non-convex $f(\boldsymbol{\lambda})$, denoting this approximation by $\hat{\mathbf{g}}(\boldsymbol{\lambda}_k, \boldsymbol{\xi}_k)$ where $\boldsymbol{\lambda}_k$ is $\boldsymbol{\lambda}$ at the k'th iteration and $\boldsymbol{\xi}_k$ is a random variable that represents the noise in the approximation. We assume that the approximation is unbiased and has a bounded variance,

$$\text{A1.} \quad \mathbb{E}[\hat{\mathbf{g}}(\boldsymbol{\lambda}_k, \boldsymbol{\xi}_k)] = \triangledown f(\boldsymbol{\lambda}_k) \quad , \quad \text{A2.} \quad \mathbb{E}[||\hat{\mathbf{g}}(\boldsymbol{\lambda}_k, \boldsymbol{\xi}_k) - \triangledown f(\boldsymbol{\lambda}_k)||^2] \leq \sigma^2, \tag{4}$$

where $\sigma > 0$ is a constant. We also assume (A3) that the gradient of f is L-Lipschitz continuous for any $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}' \in \mathcal{S}$.

This notation contains the doubly-stochastic approximation [4] as a special case. In particular, we may have stochasticity due to the mini-batch selection and also stochasticity due to the Monte-Carlo (MC) approximation to the intractable expectations with respect to q. The latter can be approximated using samples from q. In particular, for our approach we assume a mini-batch size of M_k. For the ith data example, we compute average gradient approximations to get an approximation to the gradient: $\hat{\mathbf{g}}_k := \frac{1}{M_k} \sum_{i=1}^{M_k} \hat{\mathbf{g}}(\boldsymbol{\lambda}_k, \xi_k^{(i)})$.

Our algorithm: Our proximal-gradient stochastic variational inference (PG-SVI) starts with a value $\boldsymbol{\lambda}_0$ and uses the following update at every iteration k using the gradient $\hat{\mathbf{g}}_k$ and divergence $\mathbb{D}(\boldsymbol{\lambda} \| \boldsymbol{\lambda}_k)$:

$$\boldsymbol{\lambda}_{k+1} = \arg\min_{\boldsymbol{\lambda} \in \mathcal{S}} \; \boldsymbol{\lambda}^T \hat{\mathbf{g}}_k + h(\boldsymbol{\lambda}) + \frac{1}{\beta_k} \mathbb{D}(\boldsymbol{\lambda} \,\|\, \boldsymbol{\lambda}_k) \tag{5}$$

Convergence: Our convergence results suggest that the algorithm can converge even with a constant step-size. Our proof techniques are based on the work of [15], but we need to assume that there exist a scalar $\alpha > 0$ such that for every subproblem (5),

$$\text{A4.} \quad (\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k)^T \triangledown_1 \mathbb{D}(\boldsymbol{\lambda}_{k+1} \,\|\, \boldsymbol{\lambda}_k) \geq \alpha \|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k\|^2, \tag{6}$$

where $\triangledown_1$ denotes the gradient of the first argument. But this condition only needs to hold at the solution of the subproblem. The following theorem gives us a bound on $\|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k\|$.

Theorem 1. (Convergence of PG-SVI) *Let α be the constant such that A4 is satisfied. Define $\alpha_* = \alpha - 1/(2c)$ where c is a constant such that $c > 1/(2\alpha)$. Now, let $k = 1, 2, \ldots, K$ where K is the total number of iterations, and let β_k be such that $0 < \beta_k \leq 2\alpha_*/L$ with $\beta_k < 2\alpha_*/L$ for at least one k. Suppose that we sample a discrete random variable $R \in \{1, 2, \ldots, K\}$ using the probability mass function*

$$P_R(k) := Prob(R = k) = \frac{\alpha_* \beta_k - L\beta_k^2/2}{\sum_{k=1}^{K} (\alpha_* \beta_k - L\beta_k^2/2)}.$$

Then, under assumption (A1-A4), we have the following result (where $\underline{\mathcal{L}}^*$ is the maximum):

$$\frac{1}{\beta_R} \mathbb{E}(\|\boldsymbol{\lambda}_R - \boldsymbol{\lambda}_{R-1}\|^2) \leq \frac{\underline{\mathcal{L}}^* - \underline{\mathcal{L}}(\boldsymbol{\lambda}_0) + \frac{1}{2} c\sigma^2 \sum_{k=1}^{K} (\beta_k / M_k)}{\sum_{k=1}^{K} \left(\alpha_* \beta_k - \frac{1}{2} L\beta_k^2\right)} \tag{7}$$

The bound depends on the noise variance σ^2, mini-batch size M_k, Lipschitz constant L, constant α for Assumption A4, step-size β_k, and the gap between the maximum and the starting point $\underline{\mathcal{L}}^* - \underline{\mathcal{L}}(\boldsymbol{\lambda}_0)$. In addition, a constant c needs to be chosen such that $c > 1/(2\alpha)$. When the step-size and mini-batch size are held constant, we get the following corollary:

Corollary 1. (Convergence under a constant step-size) *Let $\beta_k = \alpha_*/L$ and $M_k = M > 1$ for all k, then $\mathbb{E}(\|\boldsymbol{\lambda}_R - \boldsymbol{\lambda}_{R-1}\|^2)/\beta_R$ is bounded by, $\frac{2L}{K\alpha_*^2}[\underline{\mathcal{L}}^* - \underline{\mathcal{L}}(\boldsymbol{\lambda}^0)] + \frac{q\sigma^2}{M\alpha_*}$.*

We see that the bound gets tighter as the mini-batch size M and number of iterations K are increased, as expected. It also shows that the bound gets tighter as α_* is increased, establishing the usefulness of adding the divergence in our update. We also see a trade-off between the term depending on the Lipschitz constant L and the term depending on the variance σ^2. Most important of all, the

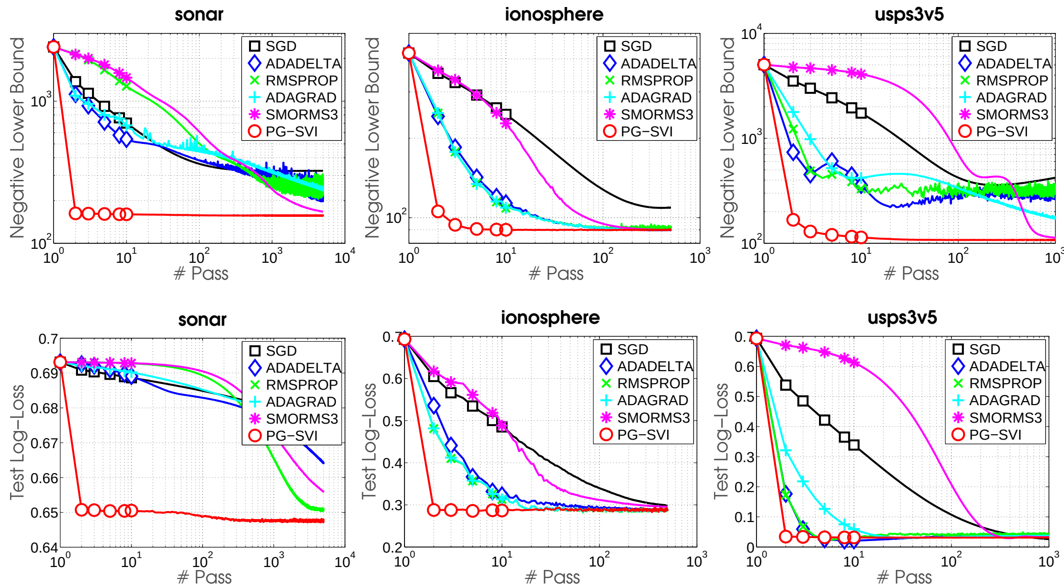

Figure 1: We show results for binary Gaussian process classification, using the setup of [16]. We compare our approach PG-SVI with SGD, ADADELTA, RMSprop [17], ADAGRAD [12], and SMORMS3 [18] on three datasets: sonar, ionosphere, and USPS-3vs5. Each column shows results for a dataset. The top row shows the negative of the lower bound, while the bottom row shows the test log-loss. In each plot, the x-axis shows the number of passes made through the data. Our method always converges within 10 passes through the data, while other methods take around 100 to 1000 passes. All the methods are compared within the GPML toolbox. We use a fixed mini-batch size M of 5, 5, and 20 respectively for the three datasets. The number of MC samples are set to 2000, 500, and 2000 respectively. For SGD, we use a schedule set according to $(1+k)^\tau$ with τ set to 0.80, 0.51 and 0.6 respectively. For ADADELTA, RMSprop, and ADAGRAD, we set $\epsilon = 10^{-8}$, while for SMORMS3 we set it to 10^{-16}. For these four methods, we choose the initial learning rate as follows: for ADADELTA we set it to 1.0, 0.1, and 1.0 respectively; for RMSprop we set it to 0.1, 0.04, and 0.1 respectively; for ADAGRAD we set it to 4.5, 4, and 8 respectively; for SMORMS3 we set it to 5, 5, and 5 respectively. For ADADELTA, we set the decay factor to $1 - 5 * 10^{-10}$, $1 - 10^{-11}$, and $1 - 10^{-12}$ respectively, and for RMSprop, we set it to 0.9, 0.9999, and 0.9 respectively. Finally, for PG-SVI, we set β_k to $0.2, 2.0,$ and 2.5 respectively. For the Gaussian processes, we use a mean function of zero a and squared-exponential covariance function. The hyperparameters were set to values that maximize the marginal-likelihood as suggested in [16].

corollary establishes the convergence of our algorithm under a constant step-size, which depends on the Lipschitz constant and the geometry of the posterior .Ghadimi et. al. discuss some strategies in [15] for tightening the second term by adapting the mini-batch size.

Existing methods as special cases: Many existing methods can be seen as special cases of our framework. Suppose that q is an exponential family distribution. When $\mathbb{D}$ is the Euclidean distance, we recover gradient descent updates. SVI can be obtained as a special case by setting $h \equiv 0$ and the divergence function to $(\boldsymbol{\lambda} - \boldsymbol{\lambda}_k)^T \bigtriangledown^2 A(\boldsymbol{\lambda}_k)(\boldsymbol{\lambda} - \boldsymbol{\lambda}_k)$. Methods based on the 'Bregman' divergence and KL divergence (e.g. [19, 20, 21, 22, 23]) are also special cases. For most of these methods, assumptions A1, A2, and A3 hold. A sufficient condition for Assumption A4 to hold is the strong-convexity of $A(\boldsymbol{\lambda})$, but our convergence results apply when the eigenvalues of $A(\boldsymbol{\lambda})$ are lower bounded at all $\boldsymbol{\lambda}_k$ that are solutions of subproblem of (5).

The parameter $\boldsymbol{\alpha}$: The parameter α can be shown to exist for many interesting distributions and problems. For example, Bernoulli and Multinomial distribution have $\alpha = 1$. For variational Gaussian approximations to latent-Gaussian models, α can be found using a lower bound on the eigenvalues of the prior covariance matrix for the Gaussian latent variable.

References

[1] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[2] Tim Salimans, David A Knowles, et al. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.

[3] Rajesh Ranganath, Sean Gerrish, and David M Blei. Black box variational inference. *arXiv preprint arXiv:1401.0118*, 2013.

[4] Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In *ICML*, 2014.

[5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[6] Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. *arXiv preprint arXiv:1402.0030*, 2014.

[7] Alp Kucukelbir, Rajesh Ranganath, Andrew Gelman, and David Blei. Fully automatic variational inference of differentiable probability models. In *NIPS Workshop on Probabilistic Programming*, 2014.

[8] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[9] Masa-Aki Sato. Online model selection based on the variational bayes. *Neural Computation*, 13(7):1649–1681, 2001.

[10] A. Honkela, T. Raiko, M. Kuusela, M. Tornio, and J. Karhunen. Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *JMLR*, 11:3235–3268, 2011.

[11] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 09 1951.

[12] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

[13] Matthew D Zeiler. ADADELTA: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[14] Ulrich Paquet. On the Convergence of Stochastic Variational Inference in Bayesian Networks. *NIPS Workshop on variational inference*, 2014.

[15] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, pages 1–39, 2014.

[16] M. Kuss and C. Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6:1679–1704, 2005.

[17] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning 4*, 2012.

[18] RMSprop loses to SMORMS3 - Beware the Epsilon! `http://sifter.org/~simon/journal/20150420.html`. Accessed: Dec. 4, 2015.

[19] Pradeep Ravikumar, Alekh Agarwal, and Martin J Wainwright. Message-passing for graph-structured linear programs: Proximal methods and rounding schemes. *The Journal of Machine Learning Research*, 11:1043–1080, 2010.

[20] Behnam Babagholami-Mohamadabadi, Sejong Yoon, and Vladimir Pavlovic. D-MFVI: Distributed Mean Field Variational Inference using Bregman ADMM. *arXiv preprint arXiv:1507.00824*, 2015.

[21] Bo Dai, Niao He, Hanjun Dai, and Le Song. Scalable Bayesian Inference via Particle Mirror Descent. *CoRR*, abs/1506.03101, 2015.

[22] Lucas Theis and Matthew D Hoffman. A trust-region method for stochastic variational inference with applications to streaming data. *arXiv preprint arXiv:1505.07649*, 2015.

[23] Mohammad Emtiyaz Khan, Pierre Baque, Francois Flueret, and Pascal Fua. Kullback-Leibler Proximal Variational Inference. In *Advances in Neural Information Processing Systems*, 2015.