Faster Stochastic Variational Inference using Proximal-Gradient Methods with General Divergence Functions

Mohammad Emtiyaz Khan[1], Reza Babanezhad[2], Wu Lin[3], Mark Schmidt[2], Masashi Sugiyama[4]

[1]Ecole Polytechnique Fédérale de Lausanne,[2] University of British Columbia, [3] University of Waterloo, [4] University of Tokyo

Introduction

Issues with existing methods:

- Most existing methods rely on "black-box" methods for optimization, and ignore the *geometry* of the variational-parameter space.
- Methods like stochastic variational inference (SVI) use *natural gradients* to exploit the geometry, but only apply to conjugate models.
- Theoretical convergence rate of natural-gradient methods for variational inference is unclear.

Contributions: We propose a proximal-gradient framework that unifies most of the existing approaches. Our method,

- can be stochastic to allow huge datasets,
- can exploit the geometry to improve performance,
- can yield a closed-form update even for non-conjugate models.

We also analyze the convergence rate of the proposed method, clearly showing the conditions under which a natural-gradient method can enable large steps than basic stochastic gradient methods.

Variational Inference

In Bayesian inference, we need to marginalize the unknowns $\mathbf{z}$ over the joint distribution $p(\mathbf{y},\mathbf{z})$ of the model, given data $\mathbf{y}$. Variational inference maximizes a lower bound to the integral w.r.t. a distribution $q(\mathbf{z}|\boldsymbol{\lambda})$ where $\boldsymbol{\lambda}$ is the vector of variational parameters.

$$\log \int p(\mathbf{y},\mathbf{z})d\mathbf{z} = \log \int q(\mathbf{z}|\boldsymbol{\lambda})\frac{p(\mathbf{y},\mathbf{z})}{q(\mathbf{z}|\boldsymbol{\lambda})}d\mathbf{z} \geq \max_{\boldsymbol{\lambda}} \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\left[\log \frac{p(\mathbf{y},\mathbf{z})}{q(\mathbf{z}|\boldsymbol{\lambda})}\right] := \underline{\mathcal{L}}(\boldsymbol{\lambda}).$$

Existing methods for lower-bound optimization differ from each other in the choice of approximation and divergence functions (highlighted in red).

$$\text{Gradient Descent: } \boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \beta_k \bigtriangledown \underline{\mathcal{L}}(\boldsymbol{\lambda}_k)$$

$$\iff \boldsymbol{\lambda}_{k+1} = \operatorname*{argmin}_{\boldsymbol{\lambda}} \; -\boldsymbol{\lambda}^T[\bigtriangledown \underline{\mathcal{L}}(\boldsymbol{\lambda}_k)] + \frac{1}{2\beta_k}\|\boldsymbol{\lambda}-\boldsymbol{\lambda}_k\|_2^2,$$

$$\text{Natural Gradient: } \boldsymbol{\lambda}_{k+1} = \operatorname*{argmin}_{\boldsymbol{\lambda}} \; -\boldsymbol{\lambda}^T[\bigtriangledown \underline{\mathcal{L}}(\boldsymbol{\lambda}_k)] + \frac{1}{\beta_k}\mathbb{D}_{KL}^{sym}[q(\mathbf{z}|\boldsymbol{\lambda}) \,\|\, q(\mathbf{z}|\boldsymbol{\lambda}_k)],$$

$$\text{Mirror Descent: } \boldsymbol{\lambda}_{k+1} = \operatorname*{argmin}_{\boldsymbol{\lambda}} \; -\boldsymbol{\lambda}^T[\bigtriangledown \underline{\mathcal{L}}(\boldsymbol{\lambda}_k)] + \frac{1}{\beta_k}\mathbb{D}_{Breg}(\boldsymbol{\lambda}\|\boldsymbol{\lambda}_k),$$

$$\text{Trust Region: } \boldsymbol{\lambda}_{k+1} = \operatorname*{argmin}_{\boldsymbol{\lambda}} \; -\underline{\mathcal{L}}(\boldsymbol{\lambda}) + \frac{1}{\beta_k}\mathbb{D}_{KL}[q(\mathbf{z}|\boldsymbol{\lambda}) \,\|\, q(\mathbf{z}|\boldsymbol{\lambda}_k)],$$

$$\text{KL proximal: } \boldsymbol{\lambda}_{k+1} = \operatorname*{argmin}_{\boldsymbol{\lambda}} \; \boldsymbol{\lambda}^T[\bigtriangledown f(\boldsymbol{\lambda}_k)] + h(\boldsymbol{\lambda}) + \frac{1}{\beta_k}\mathbb{D}_{KL}[q(\mathbf{z}|\boldsymbol{\lambda}) \,\|\, q(\mathbf{z}|\boldsymbol{\lambda}_k)].$$

Pros and cons:

- Gradient descent is generally applicable but ignores the geometry.
- Natural-gradient methods exploit the geometry but only apply to conjugate models. Mirror-descent does not cover all cases of natural-gradient and may not give a closed-form solution.
- Trust-region method may also lead to a difficult optimization problem.
- The KL-proximal method requires exact gradients.

Proximal-Gradient SVI

We propose a proximal-gradient framework that unifies many existing approaches. Specifically, our method generalizes the KL-proximal method by allowing stochastic gradients and general divergence functions.

We split the ratio $p(\mathbf{y},\mathbf{z})/q(\mathbf{z}|\boldsymbol{\lambda}) \equiv c\,\tilde{p}_d(\mathbf{z}|\boldsymbol{\lambda})\tilde{p}_e(\mathbf{z}|\boldsymbol{\lambda})$, where $\tilde{p}_d$ contains all factors that make the optimization difficult while $\tilde{p}_e$ contains the rest.

$$\underline{\mathcal{L}}(\boldsymbol{\lambda}) = \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\left[\log \frac{p(\mathbf{y},\mathbf{z}|\theta)}{q(\mathbf{z}|\boldsymbol{\lambda})}\right] := \underbrace{\mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}[\log \tilde{p}_d(\mathbf{z}|\boldsymbol{\lambda})]}_{-f(\boldsymbol{\lambda})} + \underbrace{\mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}[\log \tilde{p}_e(\mathbf{z}|\boldsymbol{\lambda})]}_{-h(\boldsymbol{\lambda})},$$

We linearize the difficult terms and solve the following subproblem:

$$\boldsymbol{\lambda}_{k+1} = \operatorname*{argmin}_{\boldsymbol{\lambda}\in\mathcal{S}} \left\{\boldsymbol{\lambda}^T\left[\widehat{\bigtriangledown} f(\boldsymbol{\lambda}_k)\right] + h(\boldsymbol{\lambda}) + \frac{1}{\beta_k}\mathbb{D}(\boldsymbol{\lambda}\,\|\,\boldsymbol{\lambda}_k)\right\}. \tag{1}$$

where $\widehat{\bigtriangledown} f(\boldsymbol{\lambda}_k)$ is the noisy gradient and $\mathbb{D}$ is a divergence function. We make the following assumptions:

- (A1-A2) f is non-convex and L-smooth and h is convex.
- (A3-A4) $\widehat{\bigtriangledown} f(\boldsymbol{\lambda}_k)$ is an unbiased estimate and its variance $\leq \sigma^2$.
- (A5-A6) $\mathbb{D}(\boldsymbol{\lambda}\,\|\,\boldsymbol{\lambda}') > 0, \forall \boldsymbol{\lambda} \neq \boldsymbol{\lambda}'$, and there exist an $\alpha > 0$ such that for all $\boldsymbol{\lambda}, \boldsymbol{\lambda}'$ generated by (1) we have: $(\boldsymbol{\lambda}-\boldsymbol{\lambda}')^T[\bigtriangledown_{\boldsymbol{\lambda}}\mathbb{D}(\boldsymbol{\lambda}\,\|\,\boldsymbol{\lambda}')] \geq \alpha\|\boldsymbol{\lambda}-\boldsymbol{\lambda}'\|^2$.

Convergence

Deterministic algorithm: Suppose A1, A2, A5, and A6 be satisfied and we run t iterations of (1) with a fixed step-size $\beta_k = \alpha/L$ for all k and an exact gradient $\nabla f(\boldsymbol{\lambda})$, then we have

$$\min_{k\in\{0,1,\ldots,t-1\}} \|\boldsymbol{\lambda}_{k+1}-\boldsymbol{\lambda}_k\|^2 \leq \frac{2C_0}{\alpha t}, \tag{2}$$

where $C_0 = \underline{\mathcal{L}}^* - \underline{\mathcal{L}}(\boldsymbol{\lambda}_0)$ is the initial (constant) sub-optimality.

Stochastic algorithm: Let A1-A6 be satisfied and we run t iterations of (1) for a fixed step-size $\beta_k = \gamma\alpha_*/L$ (where $0 < \gamma < 2$ is a scalar) and fixed batch-size $M_k = M$ for all k with a stochastic gradient $\widehat{\nabla} f(\boldsymbol{\lambda})$, then we have

$$\mathbb{E}_{R,\xi}(\|\boldsymbol{\lambda}_{R+1}-\boldsymbol{\lambda}_R\|^2) \leq \frac{1}{2-\gamma}\left[\frac{2C_0}{\alpha_* t} + \frac{\gamma c\sigma^2}{ML}\right]. \tag{3}$$

where c is a constant such that $c > 1/(2\alpha)$, $\alpha_* := \alpha - 1/(2c)$, and the expectation is taken with respect to the noise $\boldsymbol{\xi} := \{\xi_0, \xi_1, \ldots, \xi_{t-1}\}$ and a random variable R which follows the uniform distribution.

Example: Gaussian Process Models

We give an example for Gaussian-process models where we use $q(\mathbf{z}|\boldsymbol{\lambda}) := \mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})$, i.e. $\boldsymbol{\lambda} = \{\mathbf{m},\mathbf{V}\}$ with mean $\mathbf{m}$ and covariance $\mathbf{V}$. Consider N input-output pairs $\{y_n,\mathbf{x}_n\}$ indexed by n. Let $z_n := f(\mathbf{x}_n)$ be the latent function drawn from a GP with mean 0 and covariance $\mathbf{K}$. We use a non-Gaussian likelihood $p(y_n|z_n)$ to model the output. We use the split,

$$\frac{p(\mathbf{y},\mathbf{z})}{q(\mathbf{z}|\boldsymbol{\lambda})} = \underbrace{\prod_{n=1}^{N} p(y_n|z_n)}_{\tilde{p}_d(\mathbf{z}|\boldsymbol{\lambda})} \times \underbrace{\frac{\mathcal{N}(\mathbf{z}|0,\mathbf{K})}{\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})}}_{\tilde{p}_e(\mathbf{z}|\boldsymbol{\lambda})}. \tag{4}$$

By substituting in the lower bound, we obtain the following:

$$-\underline{\mathcal{L}}(\boldsymbol{\lambda}) := \underbrace{\sum_n \mathbb{E}_q[-\log p(y_n|z_n)]}_{f(\boldsymbol{\lambda})} + \underbrace{\mathbb{D}_{KL}[\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V}) \,\|\, \mathcal{N}(\mathbf{z}|0,\mathbf{K})]}_{h(\boldsymbol{\lambda})}. \tag{5}$$

We randomly pick a data-term $\mathbb{E}_q[-\log p(y_n|z_n)]$ and denote it by $f_n(m_n, v_n)$ where m_n and v_n are mean and variance of z_n. We compute its stochastic gradients and use a KL divergence function:

$$\boldsymbol{\lambda}^T\left[\widehat{\bigtriangledown} f(\boldsymbol{\lambda}_k)\right] + h(\boldsymbol{\lambda}) + \frac{1}{\beta_k}\mathbb{D}(\boldsymbol{\lambda}\|\boldsymbol{\lambda}_k) = \left[m_n\left\{\widehat{\nabla}_{m_n} f_n(\lambda_k)\right\} + v_n\left\{\widehat{\nabla}_{v_n} f_n(\boldsymbol{\lambda}_k)\right\}\right]$$
$$+ \mathbb{D}_{KL}[\mathcal{N}(\mathbf{m},\mathbf{V}) \,\|\, \mathcal{N}(0,\mathbf{K})] + \frac{1}{\beta_k}\mathbb{D}_{KL}[\mathcal{N}(\mathbf{m},\mathbf{V}) \,\|\, \mathcal{N}(\mathbf{m}_k,\mathbf{V}_k)],$$

The minimum of this function can be obtained as shown below:

$$\mathcal{N}(\mathbf{z}|\mathbf{m}_{k+1},\mathbf{V}_{k+1}) \propto \left[e^{z_n - c_k z_n^2} \times \mathcal{N}(\mathbf{z}|0,\mathbf{K})\right]^{1-r_k} [\mathcal{N}(\mathbf{z}|\mathbf{m}_k,\mathbf{V}_k)]^{r_k}$$

where $r_k := 1/(1+\beta_k)$. This equation can be implemented by solving two linear systems (see the paper for details).

Experiments

Results for GP classification and Correlated topic model. We show that, compared to existing methods, PG-SVI requires less number of passes through the data to converge.