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

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

Variational methods: optimization for high-dimensional Bayesian integral.

Drawbacks of existing approaches:

- "Black box": ignore the structure and geometry of the problem.
- Non-blackbox methods (SVI) do not extend to non-conjugate models.
- Slow convergence due to decreasing step-size sequence.

Contribution: Proximal-gradient method exploiting *structure*/*geometry*:

- Exploit the geometry with divergence functions (many existing methods as special cases).
- Exploit structure using convex/non-convex splitting.
- Convergence under a constant step size.
- Setting step-size using structure and geometry.

Variational Inference

Bayesian inference:

- Marginalize unknowns $\mathbf{z}$ over the joint $p(\mathbf{y},\mathbf{z}|\mathbf{x})$, given data $\{\mathbf{y},\mathbf{x}\}$.

Variational inference:

- Introduce distribution $q(\mathbf{z}|\boldsymbol{\lambda})$, maximize lower bound on integral,

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

where $\boldsymbol{\lambda}$ contains variational parameters.

Geometry and Natural Gradients

Gradient-Descent: $\boldsymbol{\lambda}_{k+1} = \arg\max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \bigtriangledown \underline{\mathcal{L}}(\boldsymbol{\lambda}_k)$, s.t. $\|\boldsymbol{\lambda}-\boldsymbol{\lambda}_k\|_2^2 \leq \epsilon_k$,

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \delta_k \bigtriangledown \underline{\mathcal{L}}(\boldsymbol{\lambda}_k)$$

Symmetric KL divergence because optimizing distribution parameters:

Natural-Gradient: $\boldsymbol{\lambda}_{k+1} = \arg\max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \bigtriangledown \underline{\mathcal{L}}(\boldsymbol{\lambda}_k)$, s.t. $\mathbb{D}_{KL}^{sym}[q(\mathbf{z}|\boldsymbol{\lambda}) \,\|\, q(\mathbf{z}|\boldsymbol{\lambda}_k)] \leq \epsilon_k$,

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \delta_k \mathbf{I}_k^{-1} \bigtriangledown \underline{\mathcal{L}}(\boldsymbol{\lambda}_k),$$

Equal to gradient multiplied by inverse of Fisher information $\mathbf{I}_k$ of $q(\mathbf{z}|\boldsymbol{\lambda}_k)$

Proximal-Gradient SVI

Split $\underline{\mathcal{L}}(\boldsymbol{\lambda})$ into difficult and easy parts,

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

Linearize the difficult terms to yield the iteration

$$\boldsymbol{\lambda}_{k+1} = \arg\min_{\boldsymbol{\lambda}} \; \boldsymbol{\lambda}^T \left(\sum_{k=1}^{M} \hat{\mathbf{g}}(\boldsymbol{\lambda}_k,\xi_k)\right) + h(\boldsymbol{\lambda}) - \frac{1}{\beta_k}\mathbb{D}\left[q(\mathbf{z}|\boldsymbol{\lambda})\|q(\mathbf{z}|\boldsymbol{\lambda}_k)\right].$$

where $\hat{\mathbf{g}}(\boldsymbol{\lambda}_k,\xi_k)$ is approximation of f at iteration k and with batch-size M.

We assume:

- f may be non-convex but ∇f is L-Lipschitz continuous, h is convex.
- $\hat{\mathbf{g}}(\boldsymbol{\lambda}_k,\boldsymbol{\xi}_k)$ is an unbiased estimate of $\bigtriangledown f$, with bounded variance σ^2.
- Divergence and q chosen so that

$$(\boldsymbol{\lambda}_{k+1}-\boldsymbol{\lambda}_k)^T \bigtriangledown_1 \mathcal{D}(\boldsymbol{\lambda}_{k+1}\,\|\,\boldsymbol{\lambda}_k) \geq \alpha||\boldsymbol{\lambda}_{k+1}-\boldsymbol{\lambda}_k||^2,$$

for some $\alpha > 0$.

Special cases when q is an exponential family distribution:

- Mirror descent by using:

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

- KL proximal variational inference by using:

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

- Stochastic variational inference (SVI) by using:

$$\mathbb{D}_{KL}^{sym}(\boldsymbol{\lambda}\|\boldsymbol{\lambda}') := \mathbb{D}_{KL}(\boldsymbol{\lambda}\|\boldsymbol{\lambda}') + \mathbb{D}_{Breg}(\boldsymbol{\lambda}\|\boldsymbol{\lambda}').$$

Examples of Splitting

Generalized linear model:

$$p(\mathbf{y},\mathbf{z}|\mathbf{x},\boldsymbol{\theta}) = \prod_{n=1}^{N} p(y_n|\mathbf{x}_n^T\mathbf{z})\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}), \quad q(\mathbf{z}|\boldsymbol{\lambda}) = \mathcal{N}(\mathbf{m},\mathbf{V})$$

$$\underline{\mathcal{L}}(\mathbf{m},\mathbf{V}) := \sum_{n=1}^{N} \mathbb{E}_{\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})}[\log p(y_n|\mathbf{x}_n^T\mathbf{z})] - \mathbb{D}_{KL}[\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})\,\|\,\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})]$$

Bayesian network with conditional-conjugacy:

$$p(\mathbf{z}|\boldsymbol{\eta}) := \prod_i p(\mathbf{z}_i|\mathrm{pa}_i) = \prod_i h_i(\mathbf{z}) \exp\left[\boldsymbol{\eta}_i^T \mathbf{T}_i(\mathbf{z}_i) - A_i(\boldsymbol{\eta}_i)\right]$$

$$q_i(\mathbf{z}_i|\boldsymbol{\lambda}_i) := h_i(\mathbf{z}) \exp\left[\boldsymbol{\lambda}_i^T \mathbf{T}_i(\mathbf{z}_i) - A_i(\boldsymbol{\lambda}_i)\right],$$

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

Convergence

Main result: Set the step-size β_k so that $0 < \beta_k \leq 2\alpha_*/L$, where $\alpha_* = \alpha - 1/(2c)$ and $c > 1/(2\alpha)$ is a constant. If K is the number of iterations and we sample $R \in \{1,2,\ldots,K\}$ with density:

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

then with $\underline{\mathcal{L}}^*$ the local maximum we have

$$\frac{1}{\beta_R}\mathbb{E}(\|\boldsymbol{\lambda}_R - \boldsymbol{\lambda}_{R-1}\|^2) \leq \frac{\underline{\mathcal{L}}^* - \underline{\mathcal{L}}(\boldsymbol{\lambda}^0) + \frac{1}{2}q\sigma^2\sum_{k=1}^{K}(\beta_k/M_k)}{\sum_{k=1}^{K}\left[\alpha_*\beta_k - \frac{1}{2}L\beta_k^2\right]}.$$

Constant step-size: If simply set $\beta_k = \alpha_*/L$ and $M_k = M$ then we have

$$\mathbb{E}(\|\boldsymbol{\lambda}_R - \boldsymbol{\lambda}_{R-1}\|^2)/\beta_R \leq \frac{2L}{K\alpha_*^2}[\underline{\mathcal{L}}^* - \underline{\mathcal{L}}(\boldsymbol{\lambda}^0)] + \frac{q\sigma^2}{M\alpha_*}$$

Experiments

Results for GP classification using negative log-likelihood (lower is better).