Provable Gradient Variance Guarantees for Black-Box Variational Inference

Justin Domke, University of Massachusetts Amherst

1. Overview

Motivation:

- Recent black box variational inference methods used reparameterization gradient estimators.
- Their variance is poorly understood.
- This means we don't really understand when they work!

Contributions:

1. If target distribution $\log p(\boldsymbol{z}, \boldsymbol{x})$ is M-smooth over $\boldsymbol{z}$, then
$$\mathbb{E}\,\|\mathsf{g}\|_2^2 \leq aM^2 \|\boldsymbol{w} - \bar{\boldsymbol{w}}\|_2^2$$
for some fixed $\bar{w}$.
2. This generalizes to the case where $\log p$ has different smoothness in different directions.
3. This generalizes to consider data subsampling.
4. All contributions unimprovable!

2. Background and Setup

2.1 Variational Inference

Goal is to maximize

$$\mathrm{ELBO}(\boldsymbol{w}) = \underbrace{\mathop{\mathbb{E}}_{\mathsf{z}\sim q_{\boldsymbol{w}}(\mathsf{z})} \log p(\mathsf{z}, \boldsymbol{x})}_{l(\boldsymbol{w})} + \underbrace{\mathop{\mathbb{E}}_{\mathsf{z}\sim q_{\boldsymbol{w}}(\mathsf{z})} (-\log q_{\boldsymbol{w}}(\mathsf{z}))}_{h(\boldsymbol{w})}.$$

Equivalent to minimizing $KL\left(q_{\boldsymbol{w}}(\boldsymbol{z}) \| p(\boldsymbol{z}|\boldsymbol{x})\right)$.

2.2 Reparameterization Estimators

- Choose $s(\boldsymbol{u})$ and $\mathcal{T}_{\boldsymbol{w}}(\boldsymbol{u})$ such that if $\mathsf{u} \sim s$, then $\mathcal{T}_{\boldsymbol{w}}(\mathsf{u})$ has same distribution as $q_{\boldsymbol{w}}(\boldsymbol{z})$.
- Then,
$$l(\boldsymbol{w}) = \mathop{\mathbb{E}}_{\mathsf{u}\sim s} f\left(\mathcal{T}_{\boldsymbol{w}}(\mathsf{u})\right)$$
where $f(\boldsymbol{z}) = \log p(\boldsymbol{z}, \boldsymbol{x})$.
- Estimator:
$$\mathsf{g} = \nabla_{\boldsymbol{w}} f\left(\mathcal{T}_{\boldsymbol{w}}(\mathsf{u})\right).$$
- Goal of this paper: Bound $\mathbb{E}\,\|\mathsf{g}\|_2^2$

2.3 Location-Scale Families

- A location scale family $q_{\boldsymbol{w}}(\boldsymbol{z})$ is the distribution that results from drawing $\mathsf{u} \sim s$ and returning $\mathcal{T}(\mathsf{u})$, where
$$\mathcal{T}_{\boldsymbol{w}}(\boldsymbol{u}) = C\boldsymbol{u} + \boldsymbol{m}.$$
- For example, if $s = \mathcal{N}(0, I)$ then $q_{\boldsymbol{w}}(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{m}, CC^\top)$.
- s "standardized" if $(\mathsf{u}_1, \cdots, \mathsf{u}_d) \sim s$ are i.i.d. with $\mathbb{E}\,\mathsf{u}_1 = \mathbb{E}\,\mathsf{u}_1^3 = 0$ and $\mathbb{V}\,\mathsf{u}_1 = 1$.
- Bounds will depend on $\kappa = \mathbb{E}[\mathsf{u}_1^4]$.

3. Main Results

3.1 Main Theorem

Suppose f is M-smooth, $\boldsymbol{z}^*$ is a stationary point of f, and s is standardized. Let $\mathsf{g} = \nabla_{\boldsymbol{w}} f\left(\mathcal{T}_{\boldsymbol{w}}(\mathsf{u})\right)$ for $\mathsf{u} \sim s$. Then,

$$\mathbb{E}\,\|\mathsf{g}\|_2^2 \leq M^2 \left((d+1)\,\|\boldsymbol{m} - \boldsymbol{z}^*\|_2^2 + (d+\kappa)\,\|C\|_F^2 \right).$$

This result is unimprovable.

3.2 Main Proof

$$\begin{aligned}
\mathbb{E}\,\|\mathsf{g}\|_2^2 &= \mathbb{E}\,\|\nabla_{\boldsymbol{w}} f(\mathcal{T}_{\boldsymbol{w}}(\mathsf{u}))\|_2^2 && \text{(Definition of g)}\\
&= \mathbb{E}\,\|\nabla f(\mathcal{T}_{\boldsymbol{w}}(\mathsf{u}))\|_2^2 (1 + \|\mathsf{u}\|_2^2) && \text{(First Technical Lemma)}\\
&= \mathbb{E}\,\|\nabla f(\mathcal{T}_{\boldsymbol{w}}(\mathsf{u})) - \nabla f(\boldsymbol{z}^*)\|_2^2 (1 + \|\mathsf{u}\|_2^2) && (\nabla f(\boldsymbol{z}^*) = 0)\\
&\leq \mathbb{E}\,M^2 \|\mathcal{T}_{\boldsymbol{w}}(\mathsf{u}) - \boldsymbol{z}^*\|_2^2 (1 + \|\mathsf{u}\|_2^2) && (f \text{ is smooth})\\
&= M^2 \left((d+1)\,\|\boldsymbol{m} - \boldsymbol{z}^*\|_2^2 + (d+\kappa)\,\|C\|_F^2 \right). && \text{(Second Technical Lemma)}
\end{aligned}$$

To prove first technical lemma:

- Substitute definition of $\mathcal{T}_{\boldsymbol{w}}$
- Compute all components $\|\nabla_{w_i} f(\mathcal{T}_{\boldsymbol{w}}(\mathsf{u}))\|_2^2$
- Sum and simplify.

To prove second technical lemma:

- Substitute definition of $\mathcal{T}_{\boldsymbol{w}}$.
- Resulting expression has expectations between order 0 and 4 in $\boldsymbol{u}$.
- Compute each of these and simplify using that s is standardized.

3.3 Generalized Theorem

Definition: f is M-**matrix-smooth** if $\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{z})\|_2 \leq \|M(\boldsymbol{y} - \boldsymbol{z})\|_2$ (for symmetric M).
Suppose f is M-matrix smooth, $\boldsymbol{z}^*$ is a stationary point of f, and s is standardized. Then,

$$\mathbb{E}\,\|\mathsf{g}\|_2^2 \leq (d+1)\,\|M(\boldsymbol{m} - \boldsymbol{z}^*)\|_2^2 + (d+\kappa)\,\|MC\|_F^2.$$

Unimprovable!

(Proof as above, with a trick of "absorbing" M into the parameters.)

3.4 Generalized Generalized Theorem

Suppose that $f(\boldsymbol{z}) = \sum_{n=1}^N f_n(\boldsymbol{z})$.
Suppose f_n is M_n-matrix-smooth, $\boldsymbol{z}_n^*$ is a stationary point of f_n, and s is standardized.
Let $\mathsf{g} = \frac{1}{\pi(\mathsf{n})} \nabla f_{\mathsf{n}}(\mathcal{T}_{\boldsymbol{w}}(\mathsf{u}))$ for $\mathsf{u} \sim s$ and $\mathsf{n} \sim \pi$ independent. Then,

$$\mathbb{E}\,\|\mathsf{g}\|_2^2 \leq \sum_{n=1}^N \frac{1}{\pi(n)} \left((d+1)\,\|M_n(\boldsymbol{m} - \boldsymbol{z}_n^*)\|_2^2 + (d+\kappa)\,\|M_n C\|_F^2 \right).$$

This result is unimprovable.

(Proof uses previous result as a lemma, takes expectation over n.)

4. Experiments

Dataset	Type	# data	# dims
boston	r	506	13
fires	r	517	12
cpusmall	r	8192	13
a1a	c	1695	124
ionosphere	c	351	35
australian	c	690	15
sonar	c	208	61
mushrooms	c	8124	113

Table 1: Regression (r) and classification (c) datasets