Provable Smoothness Guarantees for Black-Box Variational Inference

Justin Domke

College of Information and Computer Sciences, UMass Amherst

This paper in one slide

Variational Inference (VI): Approximate $p(\boldsymbol{z}|\boldsymbol{x})$ with $q_{\boldsymbol{w}}(\boldsymbol{z})$ by solving

$$\max_{w} \mathrm{ELBO}(\boldsymbol{w}), \quad -\mathrm{ELBO}(\boldsymbol{w}) = \underbrace{-\mathop{\mathbb{E}}_{\mathrm{z}\sim q_{\boldsymbol{w}}} \log p(\mathrm{z}, \boldsymbol{x})}_{\text{Energy term } l(\boldsymbol{w})} + \underbrace{\mathop{\mathbb{E}}_{\mathrm{z}\sim q_{\boldsymbol{w}}} \log q_{\boldsymbol{w}}(\mathrm{z})}_{\text{Neg-Entropy term } h(\boldsymbol{w})} .$$

This paper: If $p(\boldsymbol{z}, \boldsymbol{x})$ is *nice* then $l(\boldsymbol{w})$ is also *nice* (for Gaussian $q_{\boldsymbol{w}}$)

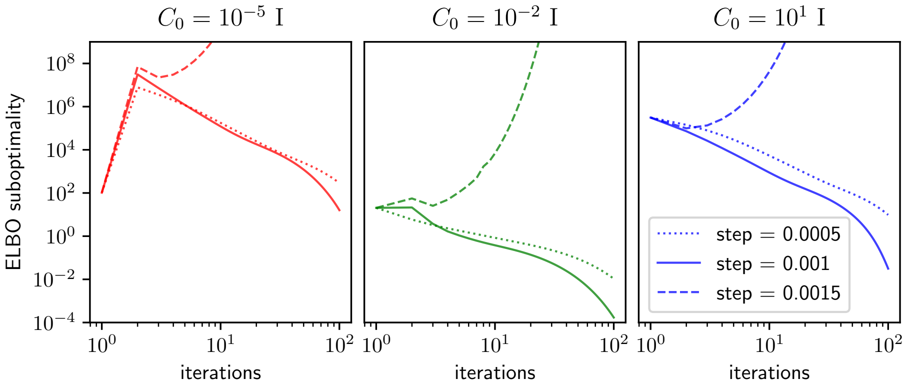
- $\log p(\boldsymbol{z}, \boldsymbol{x})$ smooth over $\boldsymbol{z}$ $\Rightarrow$ $l(\boldsymbol{w})$ smooth
- $\log p(\boldsymbol{z}, \boldsymbol{x})$ strongly concave over $\boldsymbol{z}$ $\Rightarrow$ $l(\boldsymbol{w})$ strongly convex

Implications: If you can do MAP inference, then you can do VI, *as long as you're careful.*

Motivation

Black-Box VI. Do SGD on $\text{ELBO}(\boldsymbol{w})$.

Example Problem: Three different initializations, three different step sizes. (Exact gradients)



Goals

Black-Box VI often works, but also often fails!

To give a convergence guarantee for SGD you need two things:

- A bound on the gradient estimator's variance.
- A proof that the objective is smooth or (strongly) convex (or both).

Main Result: Smoothness

- $\phi(\boldsymbol{x})$ is M-smooth if $\|\nabla\phi(\boldsymbol{x}) - \nabla\phi(\boldsymbol{x}')\|_2 \leq M\,\|\boldsymbol{x} - \boldsymbol{x}'\|_2$.

Theorem: Say $q_{\boldsymbol{w}}$ is a **location-scale family** with a **standardized base distribution** (e.g. a Gaussian) and $f(\boldsymbol{z})$ is M-smooth. Then,

$$l(\boldsymbol{w}) = \mathop{\mathbb{E}}_{\mathrm{z} \sim q_{\boldsymbol{w}}} f(\mathrm{z})$$

is also M-smooth.

Proof: Define inner-product space + Bessel's inequality + several laborious exact calculations for location-scale families.

Secondary Result: Strong Convexity

- $\phi(\boldsymbol{x})$ c-strongly convex if $\phi(\boldsymbol{y}) \geq \phi(\boldsymbol{x}) + \nabla \phi(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{c}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2$

Theorem: Say $q_{\boldsymbol{w}}$ is a **location-scale family** with a **standardized base distribution** (e.g. a Gaussian) and $f(\boldsymbol{z})$ is c-strongly convex. Then,

$$l(\boldsymbol{w}) = \mathop{\mathbb{E}}_{\mathsf{z} \sim q_{\boldsymbol{w}}} f(\mathsf{z})$$

is also c-strongly convex.

Proof: Comparatively easy.

Convergence Considerations

Say $\log p(\boldsymbol{z}, \boldsymbol{x})$ is M-smooth. Want to opt. $-\mathrm{ELBO}(\boldsymbol{w}) = l(\boldsymbol{w}) + h(\boldsymbol{w})$.

Main result: $l(\boldsymbol{w})$ is M-smooth.
Problem: $h(\boldsymbol{w})$ is *not* smooth.

One solution:

- Define $\mathscr{W}_M = \left\{ \boldsymbol{w} \middle| \mathrm{Cov\ of\ } q_{\boldsymbol{w}} \succeq \frac{1}{M} \right\}$.
- Result: Optimum of ELBO is in $\mathscr{W}_M$.
- Result: $h(\boldsymbol{w})$ is M-smooth over $\mathscr{W}_M$ (so $l + h$ is $2M$-smooth)
- So projected gradient descent works.

Another solution: Do proximal gradient descent.

Demonstration

Compare three algorithms:

- Projected optimization (step $1/(2M)$)
- Proximal optimization (step $1/M$)
- Naive optimization (step $1/M$)

Initialize $q_{\boldsymbol{w}}$ with mean 0 and covariance $\rho^2 I$ where ρ is a scaling factor.