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 $p(z|x)$ with $q_w(z)$ by solving

$$\max_w \text{ELBO}(w), \quad -\text{ELBO}(w) = -\mathbb{E}_{z \sim q_w} \log p(z, x) + \mathbb{E}_{z \sim q_w} \log q_w(z).$$

\[ \text{Energy term } l(w) \quad \text{Neg-Entropy term } h(w) \]

This paper: If $p(z, x)$ is nice then $l(w)$ is also nice (for Gaussian $q_w$)

- $\log p(z, x)$ smooth over $z \Rightarrow l(w)$ smooth
- $\log p(z, x)$ strongly concave over $z \Rightarrow l(w)$ 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 $\text{ELBO}(\mathbf{w})$.

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

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

Theorem: Say $q_w$ is a location-scale family with a standardized base distribution (e.g. a Gaussian) and $f(z)$ is $M$-smooth. Then,

$$l(w) = \mathbb{E}_{z \sim q_w} f(z)$$

is also $M$-smooth.

Proof: Define inner-product space + Bessel’s inequality + several laborious exact calculations for location-scale families.
Secondary Result: Strong Convexity

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

Theorem: Say \( q_w \) is a location-scale family with a standardized base distribution (e.g. a Gaussian) and \( f(z) \) is \( c \)-strongly convex. Then,

\[
I(w) = \mathbb{E}_{z \sim q_w} f(z)
\]

is also \( c \)-strongly convex.

Proof: Comparatively easy.
Convergence Considerations

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

Main result: $l(w)$ is $M$-smooth.
Problem: $h(w)$ is not smooth.

One solution:
- Define $\mathcal{W}_M = \left\{ w \mid \text{Cov of } q_w \succeq \frac{1}{M} \right\}$.
- Result: Optimum of ELBO is in $\mathcal{W}_M$.
- Result: $h(w)$ is $M$-smooth over $\mathcal{W}_M$ (so $l + h$ is $2M$-smooth)
- So projected gradient descent works.

Another solution: Do proximal gradient descent.
Demonstration

Compare three algorithms:
- Projected optimization \((\text{step } 1/(2M))\)
- Proximal optimization \((\text{step } 1/M)\)
- Naive optimization \((\text{step } 1/M)\)

Initialize \(q_w\) with mean 0 and covariance \(\rho^2 I\) where \(\rho\) is a scaling factor.

![Graphs comparing ELBO suboptimality for ionosphere dataset with different scalings: 2^{-14.0}, 2^{-4.0}, and 2^{4.0}](image)