A Divergence Bound for Hybrids of MCMC and VI and an application to Langevin Dynamics and SGVI

Justin Domke, UMass Amherst

---

**Introduction**

Variational inference (VI): 
\[ \min_{q} KL(q(Z|w)||p(Z)) \]

Markov chain Monte Carlo (MCMC): 
Sample from \( p(z) \)

This paper: 
Something in the middle

VI and MCMC both seek high probability \( z \).
Different coverage strategies.
- VI: include entropy \( H(w) = -\int q(z|w) \log q(z|w) \) in objective.
- MCMC inject randomness.

Idea: Random walk over \( w \). Trade off:
- "How random" the walk is
- "How much" \( H(w) \) is favored

Easy to imagine... but what are we doing?

---

**Minimizer of the bound**

Thm: Choose \( p(w|z) = r(w)q(z|w)/r_z \) where \( r_z = \int r(w)q(z|w) \) is constant. Then, \( D_p \) minimized by

\[
q^*(w) = \exp(u(w) - A)
\]

\[
s(w) = \log r(w) - \log \phi
\]

\[
A = \log \int \exp(u(w))
\]

Furthermore, the divergence at \( q^* \) is \( D_p^* = -\beta A \).

**Algorithms**

Langevin (MCMC): 
\[ z \rightarrow z + \frac{\lambda}{2} \nabla \log p(z) + \sqrt{2\eta} \]

(Stochastic) Gradient VI: 
\[ w \rightarrow w - \frac{\lambda}{2} \nabla \log p(Z|w) + [1 - \beta^{-1} - \log q(z|w)] \]

Hybrid (this paper): 
Apply Langevin to \( q^* \) and scale \( \epsilon \)

\[ w \rightarrow w + \frac{\lambda}{2} \left( -KL(q(Z|w)||p(Z)) - \beta H(w) + \beta \log q(w) \right) + \sqrt{2\eta} \]

**Intuition**

Becomes VI when \( \beta \rightarrow 0 \) (VI easy)
Becomes Langevin (on \( z \)) when \( \beta = 1 \)

- \( \beta \) likes \( w \) where \( q(Z|w) \) concentrates.

**Divergence Bounds**

Goal: Choose \( q(w) \) so \( q(z) = \int p(w)q(z|w) = p(z) \).

Impossible: minimize \( KL(q(Z)||p(Z)) = \int q(z)\log \frac{q(z)}{p(z)} \)

1st bound: (conditional divergence)

\[
KL(q(Z)||p(z)) \leq \int q(w) \int q(z|w) \log \frac{q(z|w)}{p(z|w)} = D_p
\]

2nd bound: (joint divergence) "Augment" with \( p(w|z) \).

\[
KL(q(Z)p(z)) \leq \int q(w) \int q(z|w) \log \frac{q(z|w)p(z)}{p(z|w)} = D_p
\]

Use convex combination: \( D_p = (1 - \beta)D_1 + \beta D_1 \)

**Algorithmic details**

- Use a diagonal Gaussian for \( q(z|w) \), with \( w = (\mu, \nu), v_i = \log \sigma_i \).
- To estimate gradient, use standard tricks from SGVI:
  - For Bayesian inference, estimate \( \nabla \log p(z) \) using subsampling.
  - Reparameterization trick: \( \nabla \log p(Z) = \nabla \log p_\beta(\log Z) \).
  - Then sample \( \beta \) and apply autodiff.
  - Use closed form for entropy \( q(w) = -\log q(w) \).

- Use (improper) \( \gamma(z|w) \propto \prod_{i=1}^N |v_i| |u_i| \). Numerically optimize \( u_i \) to minimize \( D_p^* \) when \( p(z) \) is a standard Gaussian.

**Toy 2-D Example**

Toy 1-D Visualization

**Logistic Regression**

- a1
- Australian
- ionosphere
- sonar