A Divergence Bound for Hybrids of MCMC and Variational Inference and an Application to Langevin Dynamics and SGVI

Justin Domke

University of Massachusetts, Amherst

ICML 2017
Outline

Overview

Divergence bound and its minimizer

Algorithm

Experiments
  One dimensional running example
  Logistic Regression examples

Discussion
Outline

Overview

Divergence bound and its minimizer

Algorithm

Experiments
  One dimensional running example
  Logistic Regression examples

Discussion
Target distribution: $p(z)$
Variational inference (VI): $\min_w KL(q(Z|w) \| p(Z))$
Makov chain Monte Carlo (MCMC): Sample from $p(z)$
This paper: A hybrid interpolating between VI and MCMC
Variational Inference
(Fast but approximate)

MCMC
(Slow but accurate)
Variational Inference
(Fast but approximate)

MCMC
(Slow but accurate)
Variational Inference (Fast but approximate) 0 1 MCMC (Slow but accurate)
**Intuition:** VI and MCMC both seek high probability $z$.

Different **coverage** strategies.

- VI: include entropy $H(w) = - \int_z q(z|w) \log q(z|w)$ 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?
**Intuition:** VI and MCMC both seek high probability $z$.

Different **coverage** strategies.
- VI: include entropy $H(w) = - \int_z q(z|w) \log q(z|w)$ 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?
**Intuition:** VI and MCMC both seek high probability $z$.

Different **coverage** strategies.

- VI: include entropy $H(w) = - \int_{z} q(z|w) \log q(z|w)$ 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?**
First contribution: Divergence bound and it’s minimizer

Define \( q(z) = \int_w q(w) q(z|w) \). 

\( q(z) \) distr. over \( w \) variational family

Impossible: Directly choose \( q(w) \) so that \( KL(q(z)||p(z)) \) is small.

Instead: upper-bound: \( KL(q(z)||p(z)) \leq D_{\beta} \).

Possible: Derive \( q^*(w) \) that minimizes \( D_{\beta} \).
First contribution: Divergence bound and it’s minimizer

Define \( q(z) = \int_w q(w) q(z|w) \). 

\[ \text{distr. over w} \quad \text{variational family} \]

**Impossible:** Directly choose \( q(w) \) so that \( KL(q(z)\|p(z)) \) is small.

**Instead:** upper-bound: \( KL(q(z)\|p(z)) \leq D_\beta \).

**Possible:** Derive \( q^*(w) \) that minimizes \( D_\beta \).
First contribution: Divergence bound and it’s minimizer

Define $q(z) = \int_w q(w) q(z|w)$.

Impossible: Directly choose $q(w)$ so that $KL(q(z)||p(z))$ is small.

Instead: upper-bound: $KL(q(z)||p(z)) \leq D_\beta$.

Possible: Derive $q^*(w)$ that minimizes $D_\beta$. 
First contribution: Divergence bound and it’s minimizer

Define \( q(z) = \int_w q(w) \underbrace{q(z|w)}_{\text{distr. over } w} \) \underbrace{\text{variational family}}_{\text{variational family}}. 

Impossible: Directly choose \( q(w) \) so that \( KL(q(z)||p(z)) \) is small.

Instead: upper-bound: \( KL(q(z)||p(z)) \leq D_\beta \).

Possible: Derive \( q^*(w) \) that minimizes \( D_\beta \).
Second contribution: How to sample from $q^*$?

Langevin (MCMC): Gradient ascent (in $z$) with injected noise on $\log p(z)$.

(Stochastic) Gradient VI: Gradient ascent on $-KL(q(Z|w)\|p(Z))$.

Hybrid: Gradient ascent (in $w$) with (varying) injected noise on $-KL(q(Z|w)\|p(Z)) - \beta H(w)$.

Becomes VI when $\beta \to 0$ VI (easy)

Becomes Langevin (on $z$) when $\beta \to 1$
Second contribution: How to sample from $q^*$?

**Langevin (MCMC)**: Gradient ascent (in $z$) with injected noise on $\log p(z)$.

**(Stochastic) Gradient VI**: Gradient ascent on $-KL(q(Z|w)\|p(Z))$.

**Hybrid**: Gradient ascent (in $w$) with (varying) injected noise on $-KL(q(Z|w)\|p(Z)) - \beta H(w)$.

Becomes VI when $\beta \to 0$ VI (easy)

Becomes Langevin (on $z$) when $\beta \to 1$
Second contribution: How to sample from $q^*$?

Langevin (MCMC): Gradient ascent (in $z$) with injected noise on $\log p(z)$.

(Stochastic) Gradient VI: Gradient ascent on $-KL(q(Z|w)\|p(Z))$.

Hybrid: Gradient ascent (in $w$) with (varying) injected noise on $-KL(q(Z|w)\|p(Z)) - \beta H(w)$.

Becomes VI when $\beta \to 0$ VI (easy)

Becomes Langevin (on $z$) when $\beta \to 1$
Overview

Divergence bound and its minimizer

Algorithm

Experiments
  One dimensional running example
  Logistic Regression examples

Discussion
Divergence Bounds

**Goal:** Choose $q(w)$ so $p(z) \approx q(z) = \int_w q(w) q(z|w)$.

**Impossible:** minimize $KL(q(Z)||p(Z)) = \int_z q(z) \log \frac{q(z)}{p(z)}$.

1st bound: (conditional divergence)

$$KL(q(Z)||p(z)) \leq KL(q(Z|W)||p(Z)) = D_0.$$  

2nd bound: (joint divergence) “Augment” $p(z)$ with $p(w|z)$.

$$KL(q(Z)||p(z)) \leq KL(q(Z,W)||p(Z,W)) = D_1.$$  

Use convex combination: $D_\beta = (1-\beta)D_0 + \beta D_1$
Divergence Bounds

Goal: Choose $q(w)$ so $p(z) \approx q(z) = \int_w q(w) \, q(z|w)$.

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

1st bound: (conditional divergence)

$$ KL(q(Z)||p(Z)) \leq KL(q(Z|W)||p(Z)) = D_0. $$

2nd bound: (joint divergence) “Augment” $p(z)$ with $p(w|z)$.

$$ KL(q(Z)||p(z)) \leq KL(q(Z, W)||p(Z, W)) = D_1. $$

Use convex combination: $D_\beta = (1 - \beta)D_0 + \beta D_1$
Divergence Bounds

Goal: Choose $q(w)$ so $p(z) \approx q(z) = \int_w q(w) \, q(z|w)$.

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

1st bound: (conditional divergence)
\[ KL(q(Z)\|p(z)) \leq KL(q(Z|W)\|p(Z)) = D_0. \]

2nd bound: (joint divergence) “Augment” $p(z)$ with $p(w|z)$.
\[ KL(q(Z)\|p(z)) \leq KL(q(Z,W)\|p(Z,W)) = D_1. \]

Use convex combination: $D_\beta = (1 - \beta)D_0 + \beta D_1$
Divergence Bounds

Goal: Choose $q(w)$ so $p(z) \approx q(z) = \int_w q(w) q(z|w)$.

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

1st bound: (conditional divergence)

$$KL(q(Z)\|p(z)) \leq KL(q(Z|W)\|p(Z)) = D_0.$$  

2nd bound: (joint divergence) “Augment” $p(z)$ with $p(w|z)$.

$$KL(q(Z)\|p(z)) \leq KL(q(Z,W)\|p(Z,W)) = D_1.$$  

Use convex combination: $D_\beta = (1-\beta)D_0 + \beta D_1$
Minimizer of the bound

**Thm:** Choose $p(w|z) \propto r(w)q(z|w)$. Then, $D_\beta$ is minimized by

$$\log q^*(w) = \log r(w) + \mathbb{E}_{q(Z|w)} \left[ \beta^{-1} \log p(Z) + (1 - \beta^{-1}) \log q(Z|w) \right] + C.$$

- Resembles a divergence (similar to VI).
- But defines a distribution (similar to MCMC).

$^1$ Assume $\int_w r(w)q(z|w)$ is constant.
Outline

Overview

Divergence bound and its minimizer

Algorithm

Experiments
  One dimensional running example
  Logistic Regression examples

Discussion
To get final algorithm: Apply Langevin dynamics to \( q^*(w) \) (and scale step-size with \( \beta \)). Leads to

\[
w \leftarrow w + \frac{\epsilon}{2} \left( \nabla_w \mathbb{E}[\log p(Z) + (\beta - 1) \log q(Z|w)] + \beta \log r_\beta(w) \right) + \sqrt{\epsilon \beta} \eta.
\]

Borrow “local reparamaterization trick” from VI to estimate grad.
Algorithm

To get final algorithm: Apply Langevin dynamics to $q^*(w)$ (and scale step-size with $\beta$). Leads to

$$w \leftarrow w + \frac{\varepsilon}{2} (\nabla_w \mathbb{E}[\log p(Z) + (\beta - 1) \log q(Z|w)] + \beta \log r_\beta(w)) + \sqrt{\varepsilon \beta} \eta.$$ 

Borrow “local reparamaterization trick” from VI to estimate grad.
Algorithm

To get final algorithm: Apply Langevin dynamics to $q^*(w)$ (and scale step-size with $\beta$). Leads to

$$w \leftarrow w + \frac{\epsilon}{2} \nabla_w \mathbb{E}[\log p(Z) + (\beta - 1) \log q(Z|w)] + \beta \log r_\beta(w) + \sqrt{\epsilon \beta} \eta.$$

Borrow “local reparamaterization trick” from VI to estimate grad.

Reductions:

- If $\beta \to 0$, becomes VI:
  $$w \leftarrow w + \frac{\epsilon}{2} \nabla_w \mathbb{E}[\log p(Z) - \log q(Z|w)]$$
Algorithm

To get final algorithm: Apply Langevin dynamics to $q^*(w)$ (and scale step-size with $\beta$). Leads to

$$w \leftarrow w + \frac{\epsilon}{2} \left( \nabla_w \mathbb{E}[\log p(Z) + (\beta - 1) \log q(Z|w)] + \beta \log r_\beta(w) \right) + \sqrt{\epsilon \beta} \eta.$$ 

Borrow “local reparamaterization trick” from VI to estimate grad.

Reductions:

- If $\beta \rightarrow 1$, becomes:

$$w \leftarrow w + \frac{\epsilon}{2} \left( \nabla_w \mathbb{E}[\log p(Z)] + \log r_\beta(w) \right) + \sqrt{\epsilon} \eta$$

But when $\beta \rightarrow 1$, $r_\beta(w)$ makes $q(z|w)$ concentrate around a single point. Thus, equivalent to Langevin dynamics on $z$. 

Outline

Overview

Divergence bound and its minimizer

Algorithm

Experiments
  One dimensional running example
  Logistic Regression examples

Discussion
Outline

Overview

Divergence bound and its minimizer

Algorithm

Experiments
  One dimensional running example
  Logistic Regression examples

Discussion
\[ \beta = 0 \]
\( \beta = 0.01 \)
$\beta = 0.05$
$\beta = 0.10$
\[ \beta = 0.25 \]
$\beta = 0.50$
$\beta = 0.75$
$\beta = 0.90$
$\beta = 1.00$
Overview

Divergence bound and its minimizer

Algorithm

Experiments

One dimensional running example

Logistic Regression examples

Discussion
a1a (1st 2 PCA components)

Top $10^4$ / middle $10^5$ / bottom $10^6$ iterations

\begin{align*}
\beta &= 0.0 & \beta &= 0.2 & \beta &= 0.4 & \beta &= 0.6 & \beta &= 0.8 & \beta &= 1.0 & \text{Stan}
\end{align*}
australian (1st 2 PCA components)

Top $10^4$ / middle $10^5$ / bottom $10^6$ iterations

\begin{align*}
\beta & = 0.0 & \beta & = 0.2 & \beta & = 0.4 & \beta & = 0.6 & \beta & = 0.8 & \beta & = 1.0 & \text{Stan}
\end{align*}
ionosphere (1st 2 PCA components)

Top $10^4$ / middle $10^5$ / bottom $10^6$ iterations

$\beta = 0.0$  $\beta = 0.2$  $\beta = 0.4$  $\beta = 0.6$  $\beta = 0.8$  $\beta = 1.0$  Stan
sonar (1st 2 PCA components)

Top $10^4$ / middle $10^5$ / bottom $10^6$ iterations

\[ \beta = 0.0 \quad \beta = 0.2 \quad \beta = 0.4 \quad \beta = 0.6 \quad \beta = 0.8 \quad \beta = 1.0 \quad \text{Stan} \]
a1a errors vs. time

MMD vs. iterations for different values of MMD:
- 0.0
- 0.2
- 0.4
- 0.6
- 0.8
- 1.0

The graph shows the decrease in MMD over iterations for each MMD value.
australian errors vs. time

MMD vs. iterations for different values of MMD (0.0, 0.2, 0.4, 0.6, 0.8, 1.0). The curves show the decrease in MMD over iterations.
sonar errors vs. time

![Graph showing the relationship between MMD and iterations for different values of MMD.]
Outline

Overview

Divergence bound and its minimizer

Algorithm

Experiments
  One dimensional running example
  Logistic Regression examples

Discussion
Summary:

- Parameterized bound $KL(q(z)\|p(z)) \leq D_\beta$
- Found $q^*(w)$ that minimizes $D_\beta$
- Alg. to sample from $q(w)$, simplifies to Langevin and SGVI.
- Experimental evidence beneficial time/accuracy tradeoff.

Questions:

- Theory for time/accuracy tradeoff?
- What is best $\beta$ for a situation?
- Other algorithmic pairs than Langevin / SGVI?
- Is this the “right” way to do things?
Summary:

- Parameterized bound $KL(q(z)\|p(z)) \leq D_\beta$
- Found $q^*(w)$ that minimizes $D_\beta$
- Alg. to sample from $q(w)$, simplifies to Langevin and SGVI.
- Experimental evidence beneficial time/accuracy tradeoff.

Questions:

- Theory for time/accuracy tradeoff?
- What is best $\beta$ for a situation?
- Other algorithmic pairs than Langevin / SGVI?
- Is this the “right” way to do things?
\[ \beta = 0 \quad \beta = 0.05 \quad \beta = 0.10 \quad \beta = 0.15 \quad \beta = 0.20 \]

\[ \beta = 0.25 \quad \beta = 0.30 \quad \beta = 0.35 \quad \beta = 0.40 \quad \beta = 0.45 \]

\[ \beta = 0.50 \quad \beta = 0.55 \quad \beta = 0.60 \quad \beta = 0.65 \quad \beta = 0.70 \]

\[ \beta = 0.75 \quad \beta = 0.80 \quad \beta = 0.85 \quad \beta = 0.90 \quad \beta = 0.95 \]

Thanks
Choosing $p(w|z)$ very similar to *auxiliary random variables* in VI. Differences:

- Here, $p(w|z)$ is fixed, numerically optimized at runtime in VI.
- Here, optimal $q^*(w)$ is found (with math), optimal $w^*$ found numerically with VI.