Concave Gaussian Variational Approximations for Bayesian GLMs

Edward Challis, David Barber

University College London
In Brief:

Bayesian Linear Models

Heavily used in ML and statistics (Inverse Modelling, TrueSkill, Collaborative filtering, ...)

The posterior distribution is typically non-Gaussian and intractable (EP, variational, sampling, ...)

We concentrate on variational methods that bound the marginal likelihood (for model comparison and hyperparameter learning)

Our Contribution

Relate Local and Variational Gaussian marginal likelihood bounds

VG is tighter than the Local bound

VG objective is concave

VG can be implemented cheaply
In Brief:

Bayesian Linear Models

- Heavily used in ML and statistics (Inverse Modelling, TrueSkill, Collaborative filtering, ...)

Our Contribution

- Relate Local and Variational Gaussian marginal likelihood bounds
- VG is tighter than the Local bound
- VG objective is concave
- VG can be implemented cheaply
In Brief:

Bayesian Linear Models

- Heavily used in ML and statistics (Inverse Modelling, TrueSkill, Collaborative filtering, . . . )
- The posterior distribution is typically non-Gaussian and intractable (EP, variational, sampling, . . . )
In Brief:

Bayesian Linear Models

- Heavily used in ML and statistics (Inverse Modelling, TrueSkill, Collaborative filtering, ...)
- The posterior distribution is typically non-Gaussian and intractable (EP, variational, sampling, ...)
- We concentrate on variational methods that bound the marginal likelihood (for model comparison and hyperparameter learning)
In Brief:

Bayesian Linear Models

- Heavily used in ML and statistics (Inverse Modelling, TrueSkill, Collaborative filtering, ...)
- The posterior distribution is typically non-Gaussian and intractable (EP, variational, sampling, ...)
- We concentrate on variational methods that bound the marginal likelihood (for model comparison and hyperparameter learning)

Our Contribution

Relate Local and Variational Gaussian marginal likelihood bounds
VG is tighter than the Local bound
VG objective is concave
VG can be implemented cheaply
In Brief:

Bayesian Linear Models

- Heavily used in ML and statistics (Inverse Modelling, TrueSkill, Collaborative filtering, . . .)
- The posterior distribution is typically non-Gaussian and intractable (EP, variational, sampling, . . .)
- We concentrate on variational methods that bound the marginal likelihood (for model comparison and hyperparameter learning)

Our Contribution

- Relate Local and Variational Gaussian marginal likelihood bounds
In Brief:

Bayesian Linear Models

- Heavily used in ML and statistics (Inverse Modelling, TrueSkill, Collaborative filtering, ...)
- The posterior distribution is typically non-Gaussian and intractable (EP, variational, sampling, ...)
- We concentrate on variational methods that bound the marginal likelihood (for model comparison and hyperparameter learning)

Our Contribution

- Relate Local and Variational Gaussian marginal likelihood bounds
- VG is tighter than the Local bound
In Brief:

Bayesian Linear Models

- Heavily used in ML and statistics (Inverse Modelling, TrueSkill, Collaborative filtering, ...)
- The posterior distribution is typically non-Gaussian and intractable (EP, variational, sampling, ...)
- We concentrate on variational methods that bound the marginal likelihood (for model comparison and hyperparameter learning)

Our Contribution

- Relate Local and Variational Gaussian marginal likelihood bounds
- VG is tighter than the Local bound
- VG objective is concave
In Brief:

Bayesian Linear Models

- Heavily used in ML and statistics (Inverse Modelling, TrueSkill, Collaborative filtering, . . .)
- The posterior distribution is typically non-Gaussian and intractable (EP, variational, sampling, . . .)
- We concentrate on variational methods that bound the marginal likelihood (for model comparison and hyperparameter learning)

Our Contribution

- Relate Local and Variational Gaussian marginal likelihood bounds
- VG is tighter than the Local bound
- VG objective is concave
- VG can be implemented cheaply
Bayesian Linear Models
Bayesian Models

For data $D$ and parameter $w$:

$$p(w|D) = \frac{p(D|w)p(w)}{Z} \quad \text{posterior}$$

$$Z = \int p(D|w)p(w)dw \quad \text{marginal likelihood}$$

likelihood prior
Bayesian Models

For data $D$ and parameter $w$:

$$p(w|D) = \frac{p(D|w)p(w)}{Z} \quad \text{posterior}$$

$$Z = \int p(D|w)p(w)dw \quad \text{marginal likelihood}$$

Bayesian Generalised Linear Models

$$p(w|D) = \frac{1}{Z} \mathcal{N}(w) \prod_n \phi(w^Th_n) \quad \text{Gaussian \ non-Gaussian}$$

$$Z = \int \mathcal{N}(w) \prod_n \phi(w^Th_n)dw \quad \text{typically intractable}$$
Bayesian Logistic Regression

For classes $s_n \in \{-1, +1\}$ and $\phi(x) = 1/(1 + \exp(-x))$:

$$p(w|D) = \frac{1}{Z} N(w) \prod_n \phi(s_n w^T x_n)$$

$w$ weight prior

$\prod_n \phi(s_n w^T x_n)$ likelihood

inputs
Bayesian Logistic Regression

Data

Prior

Likelihood

Posterior
Local and Variational Gaussian Bounds
Bounding $Z = \int \prod_n f_n(w) dw$
Bounding \( Z = \int \prod_{n} f_{n}(w) dw \)

Local (bound integrand) [Girolami, Jaakkola & Jordan, Seeger]

\[ f_{n}(w) \geq g_{n}(w; \xi_{n}) \Rightarrow \int \prod_{n} f_{n}(w) dw \geq \int \prod_{n} g_{n}(w; \xi_{n}) dw \]

\( Z \) \( G(\xi) \)
Bounding \( Z = \int \prod_n f_n(w) \, dw \)

Local (bound integrand) [Girolami, Jaakkola & Jordan, Seeger]

\[
f_n(w) \geq g_n(w; \xi_n) \Rightarrow \int \prod_n f_n(w) \, dw \geq \int \prod_n g_n(w; \xi_n) \, dw
\]

VG (bound integral) [Barber & Bishop, Opper & Archembeau]

\[
p(w|D) = \frac{1}{Z} \prod_n f_n(w); \quad \text{KL}(q(w) | p(w|D)) \geq 0
\]

\[
\log Z \geq - \int q(w) \log q(w) \, dw + \sum_n \int q(w) \log f_n(w) \, dw
\]
Bounding $Z = \int \prod_n f_n(w) \, dw$

Local (bound integrand) [Girolami, Jaakkola & Jordan, Seeger]

$$f_n(w) \geq g_n(w; \xi_n) \Rightarrow \int \prod_n f_n(w) \, dw \geq \int \prod_n g_n(w; \xi_n) \, dw$$

VG (bound integral) [Barber & Bishop, Opper & Archembeau]

$$p(w|D) = \frac{1}{Z} \prod_n f_n(w); \quad \text{KL}(q(w)|p(w|D)) \geq 0$$

$$\log Z \geq -\int q(w) \log q(w) \, dw + \sum_n \int q(w) \log f_n(w) \, dw$$

What is the relationship between the Local and VG bounds?
Local Variational Bounds

$$Z = \int \mathcal{N}(w) \prod_n \phi(w^T h_n) dw$$

$$\prod_n \phi(w^T h_n) \geq \psi(w, \xi) = c(\xi) e^{-\frac{1}{2} w^T F(\xi) w + w^T f(\xi)}$$

Optimise bound w.r.t. $\xi$

$$\log Z \geq B(\xi) \equiv \log \int \mathcal{N}(w) \psi(w, \xi) dw$$
Variational Gaussian Bounds

\[
\log Z \geq \mathcal{B}(\mathbf{m}, \mathbf{S}) = H[q(\mathbf{w})] + \langle \log \mathcal{N}(\mathbf{w}) \rangle_q + \sum_n \langle \log \phi(\mathbf{w}^T \mathbf{h}_n) \rangle_q
\]

For \( q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S}) \) the bound is tractable for all \( \phi \).
## Properties

### Known results:

<table>
<thead>
<tr>
<th>Local</th>
<th>VG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restricted to super Gaussian $\phi$</td>
<td>Generally applicable</td>
</tr>
<tr>
<td>Convex in $\xi$</td>
<td>?</td>
</tr>
<tr>
<td>Variational opt. scalable</td>
<td>?</td>
</tr>
<tr>
<td>Bound not scalable</td>
<td>?</td>
</tr>
</tbody>
</table>
## Properties

### Known results:

<table>
<thead>
<tr>
<th>Local</th>
<th>VG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restricted to super Gaussian $\phi$</td>
<td>Generally applicable</td>
</tr>
<tr>
<td>Convex in $\xi$</td>
<td>?</td>
</tr>
<tr>
<td>Variational opt. scalable</td>
<td>?</td>
</tr>
<tr>
<td>Bound not scalable</td>
<td>?</td>
</tr>
</tbody>
</table>

### New results:

<table>
<thead>
<tr>
<th>Local</th>
<th>VG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convex in $\xi$</td>
<td>Concave in $m, S$</td>
</tr>
<tr>
<td>Variational opt. scalable</td>
<td>Variational opt. scalable</td>
</tr>
<tr>
<td>Bound not scalable</td>
<td>Bound scalable</td>
</tr>
<tr>
<td>Bound is worse than VG</td>
<td>Bound is tighter than local</td>
</tr>
</tbody>
</table>
Relating the Local and VG Bounds

\[ \mathcal{B}(m, S) = H[q(w)] + \langle \log \mathcal{N}(w) \rangle + \left\langle \log \phi(w) \right\rangle \geq \tilde{\mathcal{B}}(m, S, \xi) \]
Scalability
Concavity of $\mathcal{B}(m, S = CC^T)$

Interested in concavity w.r.t $m, C$:

$$H[q(w)] + \langle \log \mathcal{N}(w) \rangle_q + \langle \log \phi(w^T h) \rangle_q$$

Concave

Concave

? 

$$\langle \log \phi(w^T h) \rangle_q = \langle \log \phi(\mu + z\sigma) \rangle_{\mathcal{N}(z|0,1)}, \quad \mu = m^T h, \sigma^2 = \|C^T h\|^2$$
Concavity of $\mathcal{B}(m, S = CC^T)$

Interested in concavity w.r.t $m, C$:

$$H[q(w)] + \langle \log \mathcal{N}(w) \rangle_q + \langle \log \phi(w^T h) \rangle_q$$

\[\text{concave} \quad \text{concave} \quad ?\]

$$\langle \log \phi(w^T h) \rangle_q = \langle \log \phi(\mu + z\sigma) \rangle_{\mathcal{N}(z|0,1)}, \quad \mu = m^T h, \sigma^2 = \|C^T h\|^2$$

---

log-concave $\phi$

Hence the VG bound is concave.
VG Optimisation: Full Cholesky

- $S = CC^T$
- Bound & Gradients scale $O(ND^2)$
VG Optimisation: Banded Cholesky

- $S = CC^T$
- Band width $= K$
- Bound & Gradients scale $O(NDK)$
VG Optimisation : Chevron Cholesky

- $S = CC^T$
- chevron width $= K$
- Bound & Gradients scale $O(NDK)$
Experiments
Bayesian Logistic Regression: realsim

- \( N = 36,000 \) training points
- \( D = 20,958 \) input dimension
- \( \approx 1.85 \times 10^6 \) non-zeros in the inputs (0.25% sparsity)

<table>
<thead>
<tr>
<th></th>
<th>VG Diag</th>
<th>VG Chev</th>
<th>Local</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K )</td>
<td>–</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>( B(m, S) )</td>
<td>–5,564</td>
<td>–5,551</td>
<td>–</td>
</tr>
<tr>
<td>CPU (s)</td>
<td>180</td>
<td>350</td>
<td>220</td>
</tr>
<tr>
<td>Test Acc. (%)</td>
<td>2.86</td>
<td>2.86</td>
<td>3.04</td>
</tr>
</tbody>
</table>
Inverse Modelling (MRI, gene reg. networks etc.)

\[ y = Mw + \eta, \quad \eta \sim \mathcal{N}(0, s^2 I) \]

\[ \phi(w) = \prod_{i=1}^{D} \frac{e^{-|w_i|/\tau_i}}{2\tau_i} \]

Model Likelihood:

\[ p(y|M, \tau) = \int \mathcal{N}(y|Mw, s^2 I) \phi(w) dw \]
Inverse Modelling

![Graph showing Bound log Z vs τ for different models: VG Full, VG Chev K=50, Local Full, Local K=50. The x-axis represents τ ranging from 0.05 to 0.2, and the y-axis represents Bound log Z ranging from -100 to -700. Each model is represented by a different marker and line style.](image-url)
Summary

Relating VG and Local bounds

- VG bounds are tighter than local bounds.
Summary

Relating VG and Local bounds

- VG bounds are tighter than local bounds.

Scalability

- VG bounds are concave in $m$ and $C$.
- Constrained Covariance provides fast/scalable inference.
- VG Bound is tractable in large systems.
Summary

Relating VG and Local bounds

- VG bounds are tighter than local bounds.

Scalability

- VG bounds are concave in $m$ and $C$.
- Constrained Covariance provides fast/scalable inference.
- VG Bound is tractable in large systems.

Implementation

- Straightforward - gradients with off-the-shelf optimisers.
Thankyou!

Code is available at:

https://mloss.org/software/view/308/