Deep Gaussian Processes with Importance-Weighted Variational Inference

Hugh Salimbeni
Vincent Dutordoir, James Hensman, Marc P Deisenroth
Problem setting
Problem setting
Problem setting
Problem setting

Skewness
Problem setting

- Skewness

- Bus arrival times
Problem setting

• Skewness

• Bus arrival times
• Confounding variables
A possible approach

\[ y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2) \]
\[ w_n \sim \mathcal{N}(0, 1) \]
A possible approach

\[ y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2) \]
\[ w_n \sim \mathcal{N}(0, 1) \]
A possible approach

\[ y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2) \]

\[ w_n \sim \mathcal{N}(0, 1) \]
A possible approach

\[ y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2) \]
\[ w_n \sim \mathcal{N}(0, 1) \]
A possible approach

\[ y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2) \]

\[ w_n \sim \mathcal{N}(0, 1) \]

Neural network

Latent variable (per point)

Concatenation with inputs

Neural network

Latent variable (per point)

Concatenation with inputs

Training data

Test samples
A possible approach

\[ y_n = \mathcal{N}(f_{\phi}([x_n, w_n]), \sigma^2) \]
\[ w_n \sim \mathcal{N}(0, 1) \]
A possible approach

\[
\begin{align*}
  y_n &= \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2) \\
  w_n &\sim \mathcal{N}(0, 1)
\end{align*}
\]
A possible approach

\[ y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2) \]

\[ w_n \sim \mathcal{N}(0, 1) \]
A possible approach

\[ y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2) \]
\[ w_n \sim \mathcal{N}(0, 1) \]
A possible approach

\[ y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2) \]

\[ w_n \sim \mathcal{N}(0, 1) \]
A possible approach

\[ y_n = \mathcal{N}(f_{\phi}([x_n, w_n]), \sigma^2) \]

\[ w_n \sim \mathcal{N}(0, 1) \]
A possible approach

$y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2)$

$w_n \sim \mathcal{N}(0, 1)$
A possible approach

$$y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2)$$

$$w_n \sim \mathcal{N}(0, 1)$$
Another possible approach

\[ y_n = \mathcal{N}(f([x_n, w_n]), \sigma^2) \]
\[ w_n \sim \mathcal{N}(0, 1) \]
\[ f \sim \mathcal{GP}(\mu, k) \]
Another possible approach

\[ y_n = N(f([x_n, w_n]), \sigma^2) \]
\[ w_n \sim N(0, 1) \]
\[ f \sim GP(\mu, k) \]

Non-parametric prior

\[ N \]
\[ f \]
\[ \infty \]
\[ x_n \rightarrow y_n \rightarrow w_n \]
Another possible approach

\[ y_n = \mathcal{N}(f([x_n, w_n]), \sigma^2) \]
\[ w_n \sim \mathcal{N}(0, 1) \]
\[ f \sim \mathcal{GP}(\mu, k) \]

Better extrapolation

Non-parametric prior
Another possible approach

\[ y_n = \mathcal{N}(f([x_n, w_n]), \sigma^2) \]

\[ w_n \sim \mathcal{N}(0, 1) \]

\[ f \sim \mathcal{GP}(\mu, k) \]

Non-parametric prior

Better extrapolation

Underfitting
Our model

\[ y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2) \]

\[ w_n \sim \mathcal{N}(0, 1) \]

\[ f \sim \mathcal{GP}(\mu_1, k_1) \]

\[ g \sim \mathcal{GP}(\mu_2, k_2) \]
Our model

\[ y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2) \]

\[ w_n \sim \mathcal{N}(0, 1) \]

\[ f \sim \mathcal{GP}(\mu_1, k_1) \]

\[ g \sim \mathcal{GP}(\mu_2, k_2) \]
Our model

\[ y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2) \]

\[ w_n \sim \mathcal{N}(0, 1) \]

\[ f \sim \mathcal{GP}(\mu_1, k_1) \]

\[ g \sim \mathcal{GP}(\mu_2, k_2) \]
Our model

\[ y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2) \]
\[ w_n \sim \mathcal{N}(0, 1) \]
\[ f \sim \mathcal{GP}(\mu_1, k_1) \]
\[ g \sim \mathcal{GP}(\mu_2, k_2) \]

Extrapolating gracefully
Our model

\[
y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2)
\]
\[
w_n \sim \mathcal{N}(0, 1)
\]
\[
f \sim \mathcal{GP}(\mu_1, k_1)
\]
\[
g \sim \mathcal{GP}(\mu_2, k_2)
\]
Contributions
Contributions

- **New architecture** - latent variables by concatenation, not addition
Contributions

- New architecture - latent variables by concatenation, not addition

- Importance-weighted variational inference, exploiting analytic results
Contributions

- **New architecture** - latent variables by concatenation, not addition
- **Importance-weighted** variational inference, exploiting analytic results
- Provide an extensive empirical comparison with all 41 UCI regression datasets
A few details

\[ y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2) \]

\[ w_n \sim \mathcal{N}(0, 1) \]

\[ f \sim \mathcal{GP}(\mu_1, k_1) \]

\[ g \sim \mathcal{GP}(\mu_2, k_2) \]
A few details

\[ y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2) \]

Importance weighting (Gaussian proposal)

\[ w_n \sim \mathcal{N}(0, 1) \]

\[ f \sim \mathcal{GP}(\mu_1, k_1) \]

\[ g \sim \mathcal{GP}(\mu_2, k_2) \]
A few details

\[ y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2) \]

Importance weighting (Gaussian proposal)

\[
\begin{align*}
\mathcal{N} & \sim \mathcal{N}(0, 1) \\
f & \sim \mathcal{GP}(\mu_1, k_1) \\
g & \sim \mathcal{GP}(\mu_2, k_2)
\end{align*}
\]

Variational inference (sparse GP posterior)
A few details

\[
y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2)
\]

Importance weighting (Gaussian proposal)

Variational inference (sparse GP posterior)

Our approach exploits analytic results, leading to a tighter bound
Results
Results

• **Latent variables** in the DGP are highly beneficial
- **Latent variables** in the DGP are highly beneficial.

- Sometimes **depth** is enough. Sometimes **latent variables** are enough. Some datasets need **both**.
Results

- **Latent variables** in the DGP are highly beneficial.

- Sometimes **depth** is enough. Sometimes **latent variables** are enough. Some datasets need **both**.

- **Importance-weighted VI** outperforms VI.
Results

- **Latent variables** in the DGP are highly beneficial.
- Sometimes depth is enough. Sometimes **latent variables** are enough. Some datasets need both.
- **Importance-weighted VI** outperforms VI.
Thanks for listening

*Poster #218*

- New architecture
- Importance-weighted
- 41 datasets