Deep Gaussian Processes with Importance-Weighted Variational Inference (and Latent Variables)

Hugh Salimbeni
10th October 2018
Outline:

• Why we want a Deep GP (2 reasons)

• The Deep GP model

• Inference in the Deep GP

• Why latent variables are important

• Inference over latent variables

• Results
Why we might want a DGP (1)
What’s wrong with deep learning?

- Bag of tricks (often) necessary
- No (calibrated) uncertainty
- Black-box (sometimes) not acceptable
- Weakness to adversarial attacks
Ambition:

- Win at deep learning tasks using fully Bayesian methods
- Get accurate uncertainty, adversarial robustness, principled model training and model selection etc

Not quite there yet…
Fundamental trade-off?

David Silver [Deep Learning Indaba 2018]:

• “Trust in experience as the sole source of knowledge”

• “Learning from experience always wins in the long run”

He is (probably) right

But asymptotics aren’t (always) what we care about
A personal view:

Two extremal options:

• The success of deep learning is evidence that we have infinite data

• The success of deep learning is attributable to a magically effective inductive bias

The truth is likely to lie somewhere between

• To do well in modern deep learning tasks, Bayesians need to think about both
Why aren’t we there yet?

- Not sufficiently scalable
- Insufficient understanding of probabilities in high dimensions

* scalability = \( \frac{d(\text{performance})}{d(\text{resource})} \)
Outline:

• Why we want a Deep GP (2 reasons)
  • The Deep GP model
  • Inference in the Deep GP
• Why latent variables are important
  • Inference over latent variables
• Results
Why we might want a DGP (2)
\( \text{cov}(a, b) = 0.5 \)
\( \text{cov}(a, c) = 0 \)
\( \text{cov}(d, e) = 0.9 \)
\( \text{cov}(f, g) = 0.1 \)
Good model
Bad model
Good model
What if we want to consider both?

???
A Deep GP posterior
Ambition:

- Form covariances hierarchically
- Get ‘GP-like’ behaviour, but allow more flexibility in the prior
Outline:

- Why we want a Deep GP (2 reasons)
- The Deep GP model
- Inference in the Deep GP
- Why latent variables are important
- Inference over latent variables
- Results
Key idea: form complex covariances with stationary kernels and input warping functions
Examples
To build a Deep GP:

\[ y_n \sim \mathcal{N}(f(g(x_n)), \sigma^2) \]
\[ f \sim \mathcal{GP}(m_1, k_1) \]
To build a Deep GP:

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

\[ f \sim \mathcal{GP}(m_1, k_1) \]

\[ g \sim \mathcal{GP}(m_2, k_2) \]
Model

\[ y_n \sim \mathcal{N}(f(g(x_n)), \sigma^2) \]
\[ f \sim \mathcal{GP}(m_1, k_1) \]
\[ g \sim \mathcal{GP}(m_2, k_2) \]

\[ m_1(x) = x \]
\[ m_2(x) = 0 \]

\( k_1, k_2 \) stationary RBF kernels
Outline:

- Why we want a Deep GP (2 reasons)
- The Deep GP model
- Inference in the Deep GP
- Why latent variables are important
- Inference over latent variables
- Results
Variational Inference

Fundamental identity for variational inference:

\[ \log p(y) = \mathbb{E}_{q(z)} \log \frac{p(y|z)p(z)}{q(z)} + \text{KL}(q(z) \| p(z|y)) \]

\[ \text{VI objective ('ELBO')} \]

- Fixed
- Maximize
- Minimize
Model:

\[ y_n \sim \mathcal{N}(f(g(x_n)), \sigma^2) \]
\[ f \sim \mathcal{GP}(m_1, k_1) \]
\[ g \sim \mathcal{GP}(m_2, k_2) \]

The VI identity in our case:

\[
\log p(y) = \mathbb{E}_{q(f,g)} \log \frac{p(y|f,g)p(f)p(g)}{q(f,g)} + \text{KL}(q(f,g) \| p(f,g|y))
\]

VI objective (‘ELBO’)}
Assumption 1 of 3

\[ q(f, g) = q(f)q(g) \]

\[
\text{ELBO} = \mathbb{E}_{q(f)q(g)} \log \frac{\prod_{n} p(y_{n} | x_{n}, f, g)}{\prod_{n} q(f)q(g)} \frac{p(f)p(g)}{q(f)q(g)}
\]

Data terms

KL terms

\[
\text{ELBO} = \sum_{n} \mathbb{E}_{q(f)q(g)} \log p(y_{n} | x_{n}, f, g) - \text{KL}(q(f) || p(f)) - \text{KL}(q(g) || p(g))
\]
The KL terms:

\[
KL(q(f) \| p(f)) = -\mathbb{E}_{q(f)} \log \frac{p(f)}{q(f)}
\]

\[
KL(q(f) \| p(f)) = -\mathbb{E}_{q(f)} \log \frac{p(f | \tilde{f})p(\tilde{f})}{q(f | \tilde{f})q(f)}
\]

Finite set of inducing points

\[
\tilde{f} = \{ f(\tilde{x}_i) \}_{i=1}^{M}
\]
Assumption 2 of 3

\[ q(f) = p(f|\tilde{f})q(\tilde{f}) \]

\[
\begin{align*}
\text{KL}(q(f) || p(f)) &= -\mathbb{E}_{q(f)} \log \frac{p(f|\tilde{f})p(\tilde{f})}{p(f|\tilde{f})q(\tilde{f})} \\
&= -\mathbb{E}_{q(\tilde{f})} \log \frac{p(\tilde{f})}{q(\tilde{f})} \\
&= \text{KL}(q(\tilde{f}) || p(\tilde{f}))
\end{align*}
\]

\[ p(\tilde{f}) = \mathcal{N}(0, \tilde{K}) \quad q(\tilde{f}) = \mathcal{N}(\tilde{m}, \tilde{S}) \]

Assumption 3 of 3
Assumption 2 of 3

\[ q(f) = p(f \mid \tilde{f}) q(\tilde{f}) \]

Assumption 3 of 3

\[ q(\tilde{f}) = \mathcal{N}(\tilde{m}, \tilde{S}) \]

It follows that:

\[ q(g) = \mathcal{GP}(\mu, \Sigma) \]

With:

\[ \mu(x) = k(x)^\top \tilde{K}^{-1} \tilde{m} \]

\[ \Sigma(x, x') = k(x, x') - k(x)^\top \tilde{K}^{-1} \left( \tilde{K} - \tilde{S} \right) \tilde{K}^{-1} k(x') \]

(NB: Temporary matrix notation)

NB:

\[ q(g(x)) = \mathcal{N}(\mu(x), \Sigma(x, x)) \]
The data terms:

$$\text{ELBO} = \sum_n \mathbb{E}_{q(f)q(g)} \log p(y_n | x_n, f, g) - \text{KL}(q(f) || p(f)) - \text{KL}(q(g) || p(g))$$

$$\mathcal{L}_n = \mathbb{E}_{q(f)q(g)} \log p(y_n | x_n, f, g)$$
$$= \mathbb{E}_{q(f)q(g)} \log p(y_n | f(g(x_n))))$$
$$= \mathbb{E}_{q(f)p(\epsilon)} \log p(y_n | f(z)), \quad z = \mu(x_n) + \epsilon \sqrt{\Sigma(x_n, x_n)}, \quad \epsilon \sim \mathcal{N}(0, 1)$$
Outline:

• Why we want a Deep GP (2 reasons)

• The Deep GP model

• Inference in the Deep GP

• Why latent variables are important

• Inference over latent variables

• Results
Noise-free
What’s wrong with this?

‘Epistemic uncertainty’ - uncertainty from lack of data
‘Aleatoric uncertainty’ - uncertainty from from inherent randomness

• GPs only model epistemic uncertainty, or marginal Gaussian aleatoric uncertainty for noisy kernels (noise = k(x, x) - k(x, x’) for limit x->x’)

• In noise-free case, we rely on epistemic uncertainty to get non-Gaussian marginals

• Noisy variables cannot be represented by our posterior, so the ELBO always favours the noise-free model
Additive noise
What’s wrong with this?

- Inference is difficult (cannot use inducing points)
- Modelling assumptions not clear (what does the noise mean?)
- Not easy to vary the dimensionality and strength of the noise
Single layer GP with ‘latent variables’

‘Latent variable’ = white noise GP

\[ y_n = \mathcal{N}(f([x, w_n]), \sigma^2) \]
Going deeper:

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

\[ y_n = \mathcal{N}(f(g(h([x, w_n]))), \sigma^2) \]
Latent variables in different places:

\[ y_n = \mathcal{N}(f([g(h(x)), w_n]), \sigma^2) \]
Outline:

- Why we want a Deep GP (2 reasons)
- The Deep GP model
- Inference in the Deep GP
- Why latent variables are important
- Inference over latent variables
- Results
Inference with latent variables

- Mean field for the latent variables
- This is reasonable as they are a priori independent
- We use variational inference or importance weighted variational inference for the latent variables
- Subtle modification to use the final layer analytic results

\[
p(y) = \mathbb{E}_{f,g,w} \left[ p(y|f,g,w) \frac{p(f)p(g)p(w)}{q(f)q(g)q(w)} \right]
\]

\[
\log p(y) \geq \sum_n (A_n - KL(w_n)) - KL_f - KL_g
\]

\[
A_n = \mathbb{E}_{f,g,w_n} \log p(y_n|f,g,w_n)
\]

\[
p(y) = \mathbb{E}_{f,g,w} \frac{1}{K} \sum_{k=1}^{K} p(y|f,g,w^{(k)}) \frac{p(w^{(k)})}{q(w^{(k)})} \frac{p(f)p(g)}{q(f)q(g)}
\]

\[
\log p(y) \geq \sum_{n=1}^{N} B_n - KL_f - KL_g
\]

\[
B_n = \mathbb{E}_{f,g,w_n} \frac{1}{K} \sum_k p(y_n|f,g,w_n^{(k)}) \frac{p(w_n^{(k)})}{q(w_n^{(k)})}
\]
Outline:

• Why we want a Deep GP (2 reasons)
• The Deep GP model
• Inference in the Deep GP
• Why latent variables are important
• Inference over latent variables
• Results
1D demo:

(a) GP

(b) GP-GP

(c) LV-GP

(d) LV-GP-GP

(a) LV-GP, VI

(b) LV-GP-GP, VI

(c) LV-GP-GP-GP, VI

(d) LV-GP-GP-GP, IW
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dataset</td>
<td>N</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fertility</td>
<td>100</td>
<td>9</td>
<td>-1.56</td>
<td>-1.40</td>
<td>-1.33</td>
<td>-1.27</td>
</tr>
<tr>
<td>concreteslump</td>
<td>103</td>
<td>7</td>
<td>1.59</td>
<td>1.87</td>
<td>1.94</td>
<td>1.58</td>
</tr>
<tr>
<td>autos</td>
<td>159</td>
<td>25</td>
<td>-0.32</td>
<td>-2.66</td>
<td>-3.16</td>
<td>-3.33</td>
</tr>
<tr>
<td>servo</td>
<td>167</td>
<td>4</td>
<td>-0.16</td>
<td>-0.12</td>
<td>-0.09</td>
<td>-0.05</td>
</tr>
<tr>
<td>breastcancer</td>
<td>194</td>
<td>33</td>
<td>-1.35</td>
<td>-12.28</td>
<td>-23.16</td>
<td>-1.31</td>
</tr>
<tr>
<td>machine</td>
<td>209</td>
<td>7</td>
<td>-0.56</td>
<td>-0.44</td>
<td>-0.47</td>
<td>-0.55</td>
</tr>
<tr>
<td>yacht</td>
<td>308</td>
<td>6</td>
<td>2.26</td>
<td>2.59</td>
<td>2.59</td>
<td>2.46</td>
</tr>
<tr>
<td>Antonpg</td>
<td>392</td>
<td>7</td>
<td>-0.34</td>
<td>-0.36</td>
<td>-0.35</td>
<td>-0.30</td>
</tr>
<tr>
<td>boston</td>
<td>506</td>
<td>13</td>
<td>-0.12</td>
<td>-0.10</td>
<td>-0.11</td>
<td>-0.12</td>
</tr>
<tr>
<td>forest</td>
<td>517</td>
<td>12</td>
<td>-1.37</td>
<td>-1.44</td>
<td>-1.59</td>
<td>-1.09</td>
</tr>
<tr>
<td>stock</td>
<td>536</td>
<td>11</td>
<td>-0.20</td>
<td>-0.20</td>
<td>-0.19</td>
<td>-0.18</td>
</tr>
<tr>
<td>pendulum</td>
<td>768</td>
<td>8</td>
<td>-0.43</td>
<td>-0.06</td>
<td>-0.05</td>
<td>-0.04</td>
</tr>
<tr>
<td>energy</td>
<td>1066</td>
<td>10</td>
<td>1.56</td>
<td>1.56</td>
<td>1.56</td>
<td>1.56</td>
</tr>
<tr>
<td>concrete</td>
<td>1030</td>
<td>8</td>
<td>-0.37</td>
<td>-0.37</td>
<td>-0.37</td>
<td>-0.37</td>
</tr>
<tr>
<td>solar</td>
<td>1066</td>
<td>10</td>
<td>-1.04</td>
<td>-0.31</td>
<td>-0.29</td>
<td>-0.29</td>
</tr>
<tr>
<td>airfoil</td>
<td>1503</td>
<td>5</td>
<td>0.41</td>
<td>0.05</td>
<td>0.13</td>
<td>0.01</td>
</tr>
<tr>
<td>winered</td>
<td>1599</td>
<td>11</td>
<td>-1.14</td>
<td>-1.15</td>
<td>-1.15</td>
<td>-1.13</td>
</tr>
<tr>
<td>gas</td>
<td>2565</td>
<td>128</td>
<td>0.69</td>
<td>1.02</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>skillcraft</td>
<td>3938</td>
<td>15</td>
<td>-0.97</td>
<td>-0.98</td>
<td>-0.98</td>
<td>-0.95</td>
</tr>
<tr>
<td>snl</td>
<td>4137</td>
<td>26</td>
<td>1.04</td>
<td>1.39</td>
<td>1.42</td>
<td>1.42</td>
</tr>
<tr>
<td>winewhite</td>
<td>4898</td>
<td>11</td>
<td>-1.14</td>
<td>-1.14</td>
<td>-1.14</td>
<td>-1.14</td>
</tr>
<tr>
<td>sml</td>
<td>5368</td>
<td>4</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>parkinsons</td>
<td>5875</td>
<td>20</td>
<td>0.48</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>kin8mn</td>
<td>8192</td>
<td>8</td>
<td>-0.35</td>
<td>-0.07</td>
<td>-0.02</td>
<td>-0.09</td>
</tr>
<tr>
<td>power</td>
<td>9568</td>
<td>4</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>naval</td>
<td>11934</td>
<td>14</td>
<td>3.84</td>
<td>3.11</td>
<td>3.51</td>
<td>3.34</td>
</tr>
<tr>
<td>pol</td>
<td>15000</td>
<td>26</td>
<td>0.17</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>elevators</td>
<td>16559</td>
<td>18</td>
<td>-0.45</td>
<td>-0.35</td>
<td>-0.40</td>
<td>-0.31</td>
</tr>
<tr>
<td>bike</td>
<td>17379</td>
<td>17</td>
<td>0.66</td>
<td>1.85</td>
<td>2.36</td>
<td>2.74</td>
</tr>
<tr>
<td>kin40k</td>
<td>40000</td>
<td>6</td>
<td>0.27</td>
<td>0.47</td>
<td>1.00</td>
<td>0.42</td>
</tr>
<tr>
<td>protein</td>
<td>45730</td>
<td>9</td>
<td>1.13</td>
<td>1.08</td>
<td>1.06</td>
<td>0.79</td>
</tr>
<tr>
<td>tamielectric</td>
<td>45781</td>
<td>3</td>
<td>-1.42</td>
<td>-1.42</td>
<td>-1.42</td>
<td>-1.27</td>
</tr>
<tr>
<td>kaggdirected</td>
<td>48827</td>
<td>20</td>
<td>0.04</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>slice</td>
<td>53500</td>
<td>385</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>kaggundirected</td>
<td>53608</td>
<td>27</td>
<td>0.69</td>
<td>0.72</td>
<td>0.73</td>
<td>0.71</td>
</tr>
<tr>
<td>3droad</td>
<td>434874</td>
<td>3</td>
<td>-1.03</td>
<td>-0.93</td>
<td>-0.87</td>
<td>-0.77</td>
</tr>
<tr>
<td>song</td>
<td>515345</td>
<td>90</td>
<td>-1.21</td>
<td>-1.18</td>
<td>-1.18</td>
<td>-1.18</td>
</tr>
<tr>
<td>buzz</td>
<td>583250</td>
<td>77</td>
<td>-0.26</td>
<td>-0.23</td>
<td>-0.23</td>
<td>-0.23</td>
</tr>
<tr>
<td>nytaxi</td>
<td>1420068</td>
<td>8</td>
<td>-0.78</td>
<td>-0.68</td>
<td>-0.68</td>
<td>-0.61</td>
</tr>
<tr>
<td>houseelectric</td>
<td>2046280</td>
<td>11</td>
<td>1.29</td>
<td>1.51</td>
<td>1.51</td>
<td>1.50</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>-0.28</td>
<td>0.09</td>
<td>-0.09</td>
<td>-0.29</td>
<td>-0.27</td>
<td>-0.00</td>
</tr>
<tr>
<td><strong>Median difference from GP</strong></td>
<td>0.00</td>
<td>0.39</td>
<td>0.46</td>
<td>0.07</td>
<td>0.11</td>
<td>0.32</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>-0.09</td>
<td>0.31</td>
<td>-0.37</td>
<td>0.15</td>
<td>0.20</td>
<td>0.41</td>
</tr>
<tr>
<td><strong>Mean difference from GP</strong></td>
<td>0.00</td>
<td>0.21</td>
<td>0.20</td>
<td>0.08</td>
<td>0.15</td>
<td>0.47</td>
</tr>
<tr>
<td><strong>Average rank</strong></td>
<td>2.67</td>
<td>4.29</td>
<td>5.04</td>
<td>3.32</td>
<td>4.30</td>
<td>5.26</td>
</tr>
<tr>
<td>--------------------</td>
<td>----------------------</td>
<td>------</td>
<td>-------</td>
<td>----------</td>
<td>-------</td>
<td>----------</td>
</tr>
<tr>
<td>solar</td>
<td></td>
<td>-1.34 (0.07)</td>
<td>-1.31 (0.08)</td>
<td>-1.34 (0.08)</td>
<td>1.35 (0.05)</td>
<td>1.53 (0.04)</td>
</tr>
<tr>
<td>winered</td>
<td></td>
<td>-1.14 (0.04)</td>
<td>-1.15 (0.04)</td>
<td>-1.15 (0.05)</td>
<td>-1.13 (0.04)</td>
<td>-1.13 (0.04)</td>
</tr>
<tr>
<td>bike</td>
<td></td>
<td>0.66 (0.02)</td>
<td>1.85 (0.02)</td>
<td>2.36 (0.05)</td>
<td>1.29 (0.01)</td>
<td>1.35 (0.01)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td>-0.28</td>
<td>-0.09</td>
<td>-0.09</td>
<td>-0.29</td>
<td>-0.27</td>
</tr>
<tr>
<td>Median difference from GP</td>
<td>0</td>
<td>0.08</td>
<td>0.10</td>
<td>0.02</td>
<td>0.05</td>
<td>0.20</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>-0.09 (0.19)</td>
<td>-0.23 (0.39)</td>
<td>-0.46 (0.69)</td>
<td>0.07 (0.19)</td>
<td>0.11 (0.19)</td>
</tr>
<tr>
<td>Mean difference from GP</td>
<td>0</td>
<td>-0.14 (0.31)</td>
<td>-0.37 (0.63)</td>
<td>0.15 (0.08)</td>
<td>0.20 (0.08)</td>
<td>0.41 (0.15)</td>
</tr>
<tr>
<td>Average ranks</td>
<td></td>
<td>2.67 (0.35)</td>
<td>4.29 (0.37)</td>
<td>5.04 (0.45)</td>
<td>3.32 (0.36)</td>
<td>4.30 (0.36)</td>
</tr>
</tbody>
</table>

Test log likelihoods (standard errors)

...an additional 36 rows can be found in the supplementary material...

(a) solar GP  
(b) solar GP-GP  
(c) solar LV-GP  
(d) solar LV-GP-GP
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dataset</td>
<td>N</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>solar</td>
<td>1066</td>
<td>10</td>
<td>-1.34 (0.07)</td>
<td>-1.31 (0.08)</td>
<td>-1.34 (0.08)</td>
<td>1.35 (0.05)</td>
<td>1.53 (0.04)</td>
</tr>
<tr>
<td>winered</td>
<td>1599</td>
<td>11</td>
<td>-1.14 (0.04)</td>
<td>-1.15 (0.04)</td>
<td>-1.15 (0.05)</td>
<td>-1.13 (0.04)</td>
<td>-1.13 (0.04)</td>
</tr>
<tr>
<td>bike</td>
<td>17379</td>
<td>17</td>
<td>0.66 (0.02)</td>
<td>1.85 (0.02)</td>
<td>2.36 (0.05)</td>
<td>1.29 (0.01)</td>
<td>1.35 (0.01)</td>
</tr>
</tbody>
</table>

...an additional 36 rows can be found in the supplementary material...

<table>
<thead>
<tr>
<th></th>
<th>Median</th>
<th>Median difference from GP</th>
<th>Mean</th>
<th>Mean difference from GP</th>
<th>Average ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.28</td>
<td>-0.09</td>
<td>-0.09</td>
<td>-0.29</td>
<td>2.67 (0.35)</td>
</tr>
<tr>
<td></td>
<td>-0.09</td>
<td>-0.23 (0.39)</td>
<td>0.10</td>
<td>-0.02</td>
<td>4.29 (0.37)</td>
</tr>
<tr>
<td></td>
<td>-0.09</td>
<td>-0.23 (0.39)</td>
<td>0.10</td>
<td>-0.02</td>
<td>5.04 (0.45)</td>
</tr>
<tr>
<td></td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.05</td>
<td>3.32 (0.36)</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.02</td>
<td>0.05</td>
<td>0.20</td>
<td>4.30 (0.36)</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.22</td>
<td>0.22</td>
<td>0.25</td>
<td>5.26 (0.34)</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.22</td>
<td>0.25</td>
<td>0.22</td>
<td>6.62 (0.32)</td>
</tr>
<tr>
<td></td>
<td>0.22</td>
<td>0.25</td>
<td>0.22</td>
<td>0.22</td>
<td>6.21 (0.32)</td>
</tr>
<tr>
<td></td>
<td>0.22</td>
<td>0.25</td>
<td>0.22</td>
<td>0.22</td>
<td>7.29 (0.33)</td>
</tr>
</tbody>
</table>

(e) bike GP   (f) bike GP-GP   (g) bike LV-GP   (h) bike LV-GP-GP
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dataset</td>
<td>N</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>solar</td>
<td>1066</td>
<td>10</td>
<td>-1.34 (0.07)</td>
<td>-1.31 (0.08)</td>
<td>-1.34 (0.08)</td>
<td>1.35 (0.05)</td>
<td>1.53 (0.04)</td>
</tr>
<tr>
<td>winered</td>
<td>1599</td>
<td>11</td>
<td>-1.14 (0.04)</td>
<td>-1.15 (0.04)</td>
<td>-1.15 (0.05)</td>
<td>-1.13 (0.04)</td>
<td>-1.13 (0.04)</td>
</tr>
<tr>
<td>bike</td>
<td>17379</td>
<td>17</td>
<td>0.66 (0.02)</td>
<td>1.85 (0.02)</td>
<td>2.36 (0.05)</td>
<td>1.29 (0.01)</td>
<td>1.35 (0.01)</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test log likelihoods (standard errors)

... an additional 36 rows can be found in the supplementary material...

<table>
<thead>
<tr>
<th>Median</th>
<th>Median difference from GP</th>
<th>Mean</th>
<th>Mean difference from GP</th>
<th>Average ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.28</td>
<td>-0.09</td>
<td>-0.09</td>
<td>-0.29</td>
<td>2.67 (0.35)</td>
</tr>
<tr>
<td>0</td>
<td>0.08</td>
<td>0.10</td>
<td>0.02</td>
<td>4.29 (0.37)</td>
</tr>
<tr>
<td>0.07</td>
<td>0.10</td>
<td>0.05</td>
<td>0.05</td>
<td>5.04 (0.45)</td>
</tr>
<tr>
<td>0.01</td>
<td>0.05</td>
<td>0.20</td>
<td>0.20</td>
<td>3.32 (0.36)</td>
</tr>
<tr>
<td>0.15</td>
<td>0.37</td>
<td>0.47</td>
<td>0.47</td>
<td>4.30 (0.36)</td>
</tr>
<tr>
<td>0.41</td>
<td>0.63</td>
<td>0.62</td>
<td>0.62</td>
<td>5.26 (0.34)</td>
</tr>
<tr>
<td>0.47</td>
<td>0.28</td>
<td>0.66</td>
<td>0.66</td>
<td>6.62 (0.32)</td>
</tr>
<tr>
<td>0.47</td>
<td>0.13</td>
<td>0.72</td>
<td>0.72</td>
<td>6.21 (0.32)</td>
</tr>
<tr>
<td>0.57</td>
<td>0.15</td>
<td>0.73</td>
<td>0.73</td>
<td>7.29 (0.33)</td>
</tr>
</tbody>
</table>
ELBO vs test log-likelihood
Summary:

- Deep GP gives a more useful prior than GP
- Need latent variables to get non-Gaussian marginals
- Variational inference appears to be effective in the noise-free case, and importance-weighted variational inference in the latent variable case
- Real data supports the hypothesis that both depth and latent variables are useful in practice
Further work:

- We haven’t broken into Deep Learning territory (yet)
- We’ve been thinking about scalability the wrong way
- We need more parameters in our variational distribution
- We need more specialised structures (e.g., convolutions)
Thanks for listening
Variational Inference

\[ p(y) = \mathbb{E}_{f,g,w} \left[ p(y|f,g,w) \frac{p(f)p(g)p(w)}{q(f)q(g)q(w)} \right] \]

\[ \log p(y) \geq \sum_n (A_n - \text{KL}_{w_n}) - \text{KL}_f - \text{KL}_g \]

\[ A_n = \mathbb{E}_{f,g,w_n} \log p(y_n|f,g,w_n) \]

\[ w_n = a_n + \epsilon_1 \sqrt{b_n} \]

\[ g([x_n, w_n]) = \mu_2([x_n, w_n]) + \epsilon_2 \sqrt{k_2([x_n, w_n], [x_n, w_n])} \]

\[ \epsilon_1, \epsilon_2 \sim N(0, 1) \]
Naive importance weighting

\[ p(y) = \mathbb{E}_{f,g,w} \frac{1}{K} \sum_{k=1}^{K} p(y|f,g,w^{(k)}) \frac{p(w^{(k)})}{q(w^{(k)})} \frac{p(f)p(g)}{q(f)q(g)} \]

\[ \log p(y) \geq \sum_{n=1}^{N} B_n - \text{KL}_f - \text{KL}_g \]

\[ B_n = \mathbb{E}_{f,g,w_n} \log \frac{1}{K} \sum_{k} p(y_n|f,g,w_n^{(k)}) \frac{p(w_n^{(k)})}{q(w_n^{(k)})} \]
Better importance weighting

\[ p(y) = \mathbb{E}_{g,w} p(y|g,w) \frac{p(w)p(g)}{q(w)q(g)} \]

\[ \log p(y|g,w) \geq \sum_n L_n(g,w_n) - KL_f \]

\[ L_n(g,w_n) = \mathbb{E}_f \log p(y_n|f,g,w_n) \]

\[ p(y|g,w) \geq \exp \left[ \sum_n L_n(g,w_n) - KL_f \right] \]

\[ p(y) \geq \mathbb{E}_{g,w} \exp \left[ \sum_n L_n(g,w_n) - KL_f \right] \frac{p(w)p(g)}{q(w)q(g)} \]

\[ \log p(y) \geq \sum_n \mathbb{E}_g \log \mathbb{E}_w \frac{e^{L_n(g,w_n)}p(w_n)}{q(w_n)} - KL_f - KL_g \]

\[ T_n(g) \]
\[
\log p(y) \geq \sum_n \mathbb{E}_g \log \mathbb{E}_w \frac{e^{L_n(g,w_n)}p(w_n)}{q(w_n)} - \text{KL}_f - \text{KL}_g
\]

\[
T_n(g) = \log \mathbb{E}_{w_n} \frac{1}{K} \sum_k e^{L_n(g,w_n^{(k)})}p(w_n^{(k)}) \frac{p(w_n^{(k)})}{q(w_n^{(k)})}
\]

\[
\sum_n \mathbb{E} \log \frac{1}{K} \sum_k e^{L_n(g,w_n^{(k)})}p(w_n^{(k)}) \frac{p(w_n^{(k)})}{q(w_n^{(k)})} - \text{KL}_f - \text{KL}_g
\]

\[
\log p(y) \geq \sum_{n=1}^N B_n - \text{KL}_f - \text{KL}_g
\]

\[
B_n = \mathbb{E} \log \frac{1}{K} \sum_k p(y_n | f, g, w_n^{(k)}) \frac{p(w_n^{(k)})}{q(w_n^{(k)})}
\]