SOME CONCERNS ABOUT SPARSE APPROXIMATIONS FOR GAUSSIAN PROCESS REGRESSION

Joaquin Quiñonero Candela
Max Planck Institute for Biological Cybernetics

Gaussian Process Round Table
Sheffield, June 9 and 10, 2005
• Concerns about the quality of the predictive distributions
• Augmentation: a bit more expensive, but gooood ...
• Dude, where’s my prior?
• A short tale about sparse greedy support set selection
The Regression Task

- Simplest case, additive independent Gaussian noise of variance $\sigma^2$
- Gaussian process prior over functions:

\[
p(y|f) \sim \mathcal{N}(f, \sigma^2 I), \quad p(f) \sim \mathcal{N}(0, K)
\]

- Task: obtain the predictive distribution of $f_*$ at the new input $x_*$:

\[
p(f_*|x_*, y) = \int p(f_*|x_*, f) p(f|y) \, df
\]

- Need to compute the posterior distribution (expensive):

\[
p(f|y) \sim \mathcal{N}(K (K + \sigma^2 I)^{-1} y, \sigma^2 K (K + \sigma^2 I)^{-1})
\]

- ... and integrate $f$ from the conditional distribution of $f_*$:

\[
p(f_*|x_*, f) \sim \mathcal{N}(K_* (K_*^{-1} y, K_{*,*} - K_* K_*^{-1} K_*^\top))
\]
Usual Reduced Set Approximations

- Consider some very common approximations
  - Naïve process approximation on subset of the data
  - Subset of regressors (Wahba, Smola and Bartlett...)
  - Sparse online GPs (Csató and Opper)
  - Fast Sparse Projected Process Approx (Seeger et al.)
  - Relevance Vector Machines (Tipping)
  - Augmented Reduced Rank GPs (Rasmussen, Quiñonero Candela)

- All based on considering only a subset $I$ of the latent variables

\[
p(f_*|x_*, y) = \int p(f_*|x_*, f_I) p(f_I|y) \, df_I
\]

- However they differ in:
  - the way the support set $I$ and the hyperparameters are learnt
  - the likelihood and/or predictive distribution approximations

- This has important consequences on the resulting predictive distribution
  - risk of over-fitting
  - degenerate approximations with nonsense predictive uncertainties
Naïve Process Approximation

• Extremely simple idea: throw away all the data outside $I$!

• The posterior only benefits from the information contained in $y_I$:

$$p(f_I | y_I) \sim \mathcal{N}(K_I(K_I + \sigma^2 I)^{-1} y_I, \sigma^2 K_I(K_I + \sigma^2 I)^{-1})$$

• The model underfits and is under-confident:

$$p(f_* | x_*, y_I) \sim \mathcal{N}(\mu_*, \sigma_*^2)$$

$$\mu_* = K_{*,I} (K_I + \sigma^2 I)^{-1} y_I, \quad \sigma_*^2 = K_{*,*} - K_{*,I} (K_I + \sigma^2 I)^{-1} K_{*,I}$$

• Training scales with $m^3$, predicting with $m$ and $m^2$ (mean and var)

• Baseline approximation: we want higher accuracy and confidence
Subset Of Regressors

- Finite linear model with peculiar prior on the weights:

\[ f_* = K_{*,I} \alpha_I, \quad \alpha_I \sim \mathcal{N}(0, K_I^{-1}) \quad \Rightarrow \quad f_* = K_{*,I} K_I^{-1} f_I, \quad f_I \sim \mathcal{N}(0, K_I) \]

- Posterior now benefits from all of \( y \):

\[ q(f_I | y) \propto \mathcal{N}(K_I^\top, K_I^{-1} f_I | y, \sigma^2 I) \cdot \mathcal{N}(f_I | 0, K_I), \]

\[ \sim \mathcal{N}(K_I [K_I, K_I^\top, + \sigma^2 K_I]^{-1} K_I, y, \sigma^2 K_I [K_I, K_I^\top, + \sigma^2 K_I]^{-1} K_I) \]

- The conditional distribution of \( f_* \) is degenerate!

\[ p(f_* | f_I) \sim \mathcal{N}(K_{*,I} K_I^{-1} f_I, 0)^\top \]

- The predictive distribution produces nonsense errorbars

\[ \mu_* = K_{*,I} [K_I, K_I^\top, + \sigma^2 K_I]^{-1} K_I, y, \]

\[ \sigma_*^2 = \sigma^2 K_{*,I} [K_I, K_I^\top, + \sigma^2 K_I]^{-1} K_{*,I}^\top \]

- Under the prior, only functions with \( m \) degrees of freedom
Projected Process (Seeger et al)

- Basic principle: likelihood approximation

\[ p(y|f_I) \sim (K_I^\top, K_I^{-1} f_I, \sigma^2 I) \]

- Leads to exactly the same posterior as for Subset of Regressors

- But the conditional distribution is now non-degenerate (process approximation)

\[ p(f_*|f_I) \sim \mathcal{N}(K_{*,I} K_I^{-1} f_I, K_{*,*} - K_{*,I} K_I^{-1} K_{*,I})^\top \]

- Predictive distribution with same mean as Subset of Regressors, but with way under-confident predictive variance!

\[
\begin{align*}
\mu_* &= K_{*,I} \left[ K_I, K_I^\top + \sigma^2 K_I \right]^{-1} K_I,. y \\
\sigma^2_* &= K_{*,*} - K_{*,I} K_I^{-1} K_{*,I} + \sigma^2 K_{*,I} \left[ K_I, K_I^\top + \sigma^2 K_I \right]^{-1} K_{*,I}^\top 
\end{align*}
\]
Augmented Subset Of Regressors

• For each $x_\star$, augment $f_I$ with $f_\star$; new active set $I_\star$

• Augmented posterior:
  $$q \left( \begin{bmatrix} f_I \\ f_\star \end{bmatrix} \bigg| y \right)$$

• ... at a cost of $O(nm)$ per test case: need to compute $K_\star, K_I^{-1}$

• aSoR:

  $$\mu_\star = K_\star, \cdot \left[ Q + \frac{v_\star v_\star^\top}{c_\star} \right]^{-1} y$$

  $$\sigma^2_\star = K_{\star,\star} - K_\star, \cdot \left[ Q + \frac{v_\star v_\star^\top}{c_\star} \right]^{-1} K_{\star,\star}^\top$$

  with the usual approximate covariance:

  $$Q = K_{I,\star}^\top K_I^{-1} K_{I,\star} + \sigma^2 I$$

  with the difference between actual and projected covariance of $f_\star$ and $f$:

  $$v_\star = K_{\star,\star}^\top - K_{I,\star}^\top K_I^{-1} K_{I,\star}$$

  with the difference between the prior variance of $f_\star$ and the projected:

  $$c_\star = K_{\star,\star} - K_{I,\star}^\top K_I^{-1} K_{I,\star}$$
Dude, where's my prior?
The Priors

The equivalent prior on $[f, f_*]^\top$ is $\mathcal{N}(0, P)$ with:

$$Q = K_{I,*}^\top K_{I, I}^{-1} K_{I,*}.$$

Subset of Regressors:

$$P = \begin{bmatrix} Q & K_{I,*}^\top K_{I, I}^{-1} K_{I,*} \\ K_{I,*}^\top K_{I, I}^{-1} K_{I,*} & K_{I,*}^\top K_{I, I}^{-1} K_{I,*} \end{bmatrix}$$

Nyström: (positive definiteness!)

$$P = \begin{bmatrix} Q & K_{I,*}^\top \cdot K_{I,*} \\ K_{I,*}^\top \cdot K_{I,*} & K_{I,*}^\top \cdot K_{I,*} \end{bmatrix}$$

Projected Process

$$P = \begin{bmatrix} Q & K_{I,*}^\top K_{I, I}^{-1} K_{I,*} \\ K_{I,*}^\top K_{I, I}^{-1} K_{I,*} & K_{I,*}^\top \cdot K_{I,*} \end{bmatrix}$$

Ed and Zoubin’s funky thing

$$P = \begin{bmatrix} Q + \Lambda & K_{I,*}^\top K_{I, I}^{-1} K_{I,*} \\ K_{I,*}^\top K_{I, I}^{-1} K_{I,*} & K_{I,*}^\top \cdot K_{I,*} \end{bmatrix}$$

$$\Lambda = \text{diag}(K_\cdot) - \text{diag}(Q)$$

Augmented Subset of Regressors:

$$P = \begin{bmatrix} Q + \frac{v_* v_*^\top}{c_*} & K_{I,*}^\top \\ K_{I,*}^\top & K_{I,*}^\top \cdot K_{I,*} \end{bmatrix}$$

with:

$$v_* = K_{I,*} - K_{I, I}^\top K_{I, I}^{-1} K_{I,*}, \quad c_* = K_{I,*}^\top - K_{I,*}^\top K_{I, I}^{-1} K_{I,*}$$
• Here’s a way of looking at it: the prior is a posterior process

\[ f_\ast | f_I = \mathcal{N}(K_{*}, I K_I^{-1} f_I, K_{*,*} - K_{*,I} K_I^{-1} K_{*,I}^T) , \]

... well, almost: \( E[f_+, f_\ast | f_I] = 0 \)

• And then of course \( f_I \sim \mathcal{N}(0, K_I) \)

• The corresponding prior is

\[ p(f) = \mathcal{N}(0, K_{*,*} I + Q - \text{diag}(Q)) , \quad Q = K_I, K_I^{-1} K_I^T. \]

• With a bit of algebra you recover the marginal likelihood and the predictive distribution

• I finished this 30 minutes ago, which is why I won’t show figures on it! (well, I now may)

• but ...
Naïve Process Approximation
Subset of Regressors (degenerate)
Projected Process Approximation
Ed and Zoubin’s Projected Process Method
Augmented SoR (pred scales with \textit{nm})
Comparing the Predictive Uncertainties

- Naive
- SR
- Seeger
- EdZoubin
- Augm
Smola and Bartlett’s Greedy Selection

- Test squared error:

- Size of support set, m, logarithmic scale

- Neg log evidence

- Min neg log ev.

- Min neg log post.

- Upper bound on neg log post.

- Lower bound on neg log post.

- Gap = 0.025
Wrap Up

- Training: from $O(n^3)$ to $O(nm^2)$
- Predicting: from $O(n^2)$ to $O(m^2)$ (or $O(nm)$)
- Be sparse if you must, but only then
- Beware of over-fitting prone greedy selection methods
- Do worry about the prior implied by the approximation!
Appendix: Healing the RVM by Augmentation
(joint work with Carl Rasmussen)
Finite Linear Model
Augmentation?

- Train once your $m$-dimensional model
- At each new test point add a new basis function
- Update the $m + 1$-dimensional model (update posterior)
- Testing is now more expensive
Wait a minute ...

I don’t care about probabilistic predictions!
## Another Symptom: Underfitting

### Abalone

<table>
<thead>
<tr>
<th></th>
<th>Squared error loss</th>
<th>Absolute error loss</th>
<th>- log test density loss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RVM</td>
<td>RVM*</td>
<td>GP</td>
</tr>
<tr>
<td>Loss:</td>
<td>0.138</td>
<td>0.135</td>
<td>0.092</td>
</tr>
<tr>
<td>RVM</td>
<td>0.259</td>
<td>0.253</td>
<td>0.209</td>
</tr>
<tr>
<td>RVM*</td>
<td>-</td>
<td>0.07</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>GP</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- not sig. $< 0.01$
- $< 0.01$
- $< 0.01$
- $< 0.01$
- $< 0.01$
- $< 0.01$

### Robot Arm

<table>
<thead>
<tr>
<th></th>
<th>Squared error loss</th>
<th>Absolute error loss</th>
<th>- log test density loss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RVM</td>
<td>RVM*</td>
<td>GP</td>
</tr>
<tr>
<td>Loss:</td>
<td>0.0043</td>
<td>0.0040</td>
<td>0.0024</td>
</tr>
<tr>
<td>RVM</td>
<td>0.0482</td>
<td>0.0467</td>
<td>0.0334</td>
</tr>
<tr>
<td>RVM*</td>
<td>-</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>GP</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- $< 0.01$
- $< 0.01$
- $< 0.01$
- $< 0.01$
- $< 0.01$
- $< 0.01$

- $< 0.01$
- $< 0.01$
- $< 0.01$
- $< 0.01$
- $< 0.01$
- $< 0.01$

- $< 0.01$
- $< 0.01$
- $< 0.01$
- $< 0.01$
- $< 0.01$
- $< 0.01$

### Notes
- **GP (Gaussian Process):** infinitely augmented linear model
- Beats finite linear models in all datasets I’ve looked at
Interlude

None of this happens with non-localized basis functions
A Bad Probabilistic Model
The Healing: Augmentation
Appendix: Augmentation in Sparse GPs

- $O(nm^2)$ sparse approx. to Gaussian Processes (Smola and Bartlett, 2001)
- Augmentation: same training, more expensive testing
- Better mean based and probabilistic performance

<table>
<thead>
<tr>
<th>method</th>
<th>tr. neg ev.</th>
<th>non-augmented</th>
<th>augmented</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGGP</td>
<td>—</td>
<td>0.0481</td>
<td>0.0048</td>
</tr>
<tr>
<td>SGEV</td>
<td>-1.1555</td>
<td>0.0484</td>
<td>0.0049</td>
</tr>
<tr>
<td>HPEV-rand</td>
<td>-1.0978</td>
<td>0.0503</td>
<td>0.0047</td>
</tr>
<tr>
<td>HPEV-SGEV</td>
<td>-1.3234</td>
<td>0.0425</td>
<td>0.0036</td>
</tr>
<tr>
<td>HPEV-SGGP</td>
<td>-1.3274</td>
<td>0.0425</td>
<td>0.0036</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2000 training - 2000 test</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGEV</td>
</tr>
<tr>
<td>HPEV-rand</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>36000 training - 4000 test</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPEV-SGEV</td>
</tr>
</tbody>
</table>
Thanks a lot to Sheffield and to Neil!