Improving the Gaussian Process Sparse Spectrum Approximation by Representing Uncertainty in Frequency Inputs

Yarin Gal • Richard Turner

yg279@cam.ac.uk
The Gaussian process (GP)

- Is awesome

- ... but with a great computational cost – $O(N^3)$ time complexity for $N$ data points:

\[
p(Y|X) = \mathcal{N}(Y; 0, K(X, X) + \tau^{-1}I_N)
\]

with $Q$ dimensional input $x$, $D$ dimensional output $y$, and stationary covariance function $K$. 
Many Approximations

Full GP:

- Sparse pseudo-input cannot handle complex functions well:

- Sparse spectrum is known to over-fit:
Main Idea:

- **Variational Sparse Spectrum GP (VSSGP)**
  - use variational inference for the sparse spectrum approximation
  - avoids over-fitting, efficiently captures globally complex behaviour

- **In short—**
  - we replace the GP covariance function with a finite Monte Carlo approximation
  - we view this as a random covariance function
  - conditioned on data this random variable has an intractable posterior
  - we approximate this posterior with variational inference
Main Idea:

- **Variational Sparse Spectrum GP (VSSGP)**
  - use variational inference for the sparse spectrum approximation
  - avoids over-fitting, efficiently captures globally complex behaviour

- **In short—**
  - we replace the GP covariance function with a finite Monte Carlo approximation
    - we view this as a random covariance function
    - conditioned on data this random variable has an intractable posterior
    - we approximate this posterior with variational inference
Main Idea:

- **Variational Sparse Spectrum GP (VSSGP)**
  - use variational inference for the sparse spectrum approximation
  - avoids over-fitting, efficiently captures globally complex behaviour

- **In short—**
  - we replace the GP covariance function with a finite Monte Carlo approximation
  - we view this as a *random covariance function*
  - conditioned on data this random variable has an intractable posterior
  - we approximate this posterior with variational inference
Main Idea:

- **Variational Sparse Spectrum GP (VSSGP)**
  - use variational inference for the sparse spectrum approximation
  - avoids over-fitting, efficiently captures globally complex behaviour

- **In short—**
  - we replace the GP covariance function with a finite Monte Carlo approximation
  - we view this as a **random covariance function**
  - conditioned on data this random variable has an intractable posterior
  - we approximate this posterior with variational inference
Main Idea:

- **Variational Sparse Spectrum GP (VSSGP)**
  - use variational inference for the sparse spectrum approximation
  - avoids over-fitting, efficiently captures globally complex behaviour

- **In short—**
  - we replace the GP covariance function with a finite Monte Carlo approximation
  - we view this as a *random covariance function*
  - conditioned on data this random variable has an intractable posterior
  - we approximate this posterior with variational inference
In more detail (with a squared exponential covariance function)—

Given Fourier transform of the covariance function:

\[
K(x - y) = \sigma^2 e^{-\frac{(x-y)^T(x-y)}{2}}
\]

\[
= \sigma^2 \int \mathcal{N}(w; 0, I_Q) \cos(2\pi w^T(x - y)) \, dw.
\]
Fourier transform of the squared exponential covariance function:

$$K(x - y) = \sigma^2 \int \mathcal{N}(w; 0, I_Q) \cos(2\pi w^T(x - y)) \, dw,$$

**Auxiliary variable $b$:**

$$K(x - y) = 2\sigma^2 \int \mathcal{N}(w; 0, I_Q) \text{Unif}[0, 2\pi] \cos(2\pi w^T x + b) \cos(2\pi w^T y + b) \, dw \, db.$$
Auxiliary variable $b$:

$$K(x - y) = 2\sigma^2 \int \mathcal{N}(w; 0, I_Q)\text{Unif}[0, 2\pi] \cos(2\pi w^T x + b) \cos(2\pi w^T y + b) \, dw \, db,$$

Monte Carlo integration with $K$ terms:

$$\hat{K}(x - y) = \frac{2\sigma^2}{K} \sum_{k=1}^{K} \cos(2\pi w_k^T x + b_k) \cos(2\pi w_k^T y + b_k)$$

with $w_k \sim \mathcal{N}(0, I_Q), b_k \sim \text{Unif}[0, 2\pi]$.

This is a random covariance function.
Monte Carlo integration with $K$ terms:

$$
\hat{K}(x - y) = \frac{2\sigma^2}{K} \sum_{k=1}^{K} \cos \left(2\pi w_k^T x + b_k\right) \cos \left(2\pi w_k^T y + b_k\right),
$$

Rewrite the covariance function with $\Phi \in \mathbb{R}^{N \times K}$

$$
\begin{align*}
\mathbf{w}_k &\sim \mathcal{N}(0, I_Q), \quad b_k \sim \text{Unif}[0, 2\pi], \quad \omega = \{\mathbf{w}_k, b_k\}_{k=1}^{K} \\
\Phi_{n,k}(\omega) &= \sqrt{\frac{2\sigma^2}{K}} \cos \left(2\pi \mathbf{w}_k^T \mathbf{x}_n + b_k\right), \\
\hat{K}(x - y) &= \Phi(\omega) \Phi(\omega)^T.
\end{align*}
$$
Rewrite the covariance function with $\Phi \in \mathbb{R}^{N \times K}$

$\mathbf{w}_k \sim \mathcal{N}(0, \mathbf{I}_Q), \quad b_k \sim \text{Unif}[0, 2\pi], \quad \omega = \{\mathbf{w}_k, b_k\}_{k=1}^K$

$\Phi_{n,k}(\omega) = \sqrt{\frac{2\sigma^2}{K}} \cos\left(2\pi \mathbf{w}_k^T \mathbf{x}_n + b_k\right),$

$\hat{\mathbf{K}}(\mathbf{x} - \mathbf{y}) = \Phi(\omega) \Phi(\omega)^T.$

Integrate the GP over the random covariance function

$\mathbf{w}_k \sim \mathcal{N}(0, \mathbf{I}_Q), \quad b_k \sim \text{Unif}[0, 2\pi], \quad \omega = \{\mathbf{w}_k, b_k\}_{k=1}^K$

$p(\mathbf{Y}|\mathbf{X}, \omega) = \mathcal{N}(\mathbf{Y}; \mathbf{0}, \Phi(\omega) \Phi(\omega)^T + \tau^{-1} \mathbf{I}_N)$

$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{X}, \omega) p(\omega) d\omega$

$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) = \int p(\mathbf{y}^*|\mathbf{x}^*, \omega) p(\omega|\mathbf{X}, \mathbf{Y}) d\omega.$
Integrate the GP over the random covariance function

\[ w_k \sim \mathcal{N}(0, I_Q), \quad b_k \sim \text{Unif}[0, 2\pi], \quad \omega = \{w_k, b_k\}_{k=1}^K \]

\[ p(Y|X, \omega) = \mathcal{N}(Y; 0, \Phi(\omega) \Phi(\omega)^T + \tau^{-1} I_N) \]

\[ p(Y|X) = \int p(Y|X, \omega)p(\omega)d\omega \]

\[ p(y^*|x^*, X, Y) = \int p(y^*|x^*, \omega)p(\omega|X, Y)d\omega. \]

Use variational distribution \( q(\omega) = \prod q(w_k)q(b_k) \) to approximate posterior \( p(\omega|X, Y) \):

\[ q(w_k) = \mathcal{N}(\mu_k, \Sigma_K), \quad q(b_k) = \text{Unif}(\alpha_k, \beta_k), \]

with \( \Sigma_K \) diagonal.
Maximise log evidence lower bound

\[ \mathcal{L}_{VSSGP} = \frac{1}{2} \sum_{d=1}^{D} \left( \log(|\tau^{-1} \Sigma|) + \tau \mathbf{y}_d^T E_{q(\omega)}(\Phi) \Sigma E_{q(\omega)}(\Phi^T) \mathbf{y}_d + \ldots \right) \]

\[ - \text{KL}(q(\omega)||p(\omega)) \]

with \[ \Sigma = (E_{q(\omega)}(\Phi^T \Phi) + \tau^{-1} I)^{-1}. \]
Maximise log evidence lower bound

\[
\mathcal{L}_{VSSGP} = \frac{1}{2} \sum_{d=1}^{D} \left( \log(|\tau^{-1} \Sigma|) + \tau y_d^T E_q(\omega)(\Phi) \Sigma E_q(\omega)(\Phi^T) y_d + \ldots \right) - \text{KL}(q(\omega)||p(\omega))
\]

with \( \Sigma = (E_q(\omega)(\Phi^T \Phi) + \tau^{-1} I)^{-1} \).

KL and expectations analytical with

\[
E_q(w)(\cos(w^T x + b)) = e^{-\frac{1}{2} x^T \Sigma x} \cos(\mu^T x + b).
\]

Requires \( O(NK^2 + K^3) \) time complexity.
Parallel inference with \( T \) workers: \( O\left(\frac{NK^2}{T} + K^3\right) \).
Factorised VSSGP (fVSSGP)

- We often use large $K$.

- $K$ by $K$ matrix inversion is still slow: $\mathcal{O}(K^3)$.

- It is silly to invert the whole matrix every time — slightly changing the parameters we expect the inverse to not change too much.

- We can do better with an additional auxiliary variable.
We integrated the GP over the random covariance function

\[ w_k \sim \mathcal{N}(0, I_Q), \quad b_k \sim \text{Unif}[0, 2\pi], \quad \omega = \{w_k, b_k\}_{k=1}^K \]

\[ p(Y|X, \omega) = \mathcal{N}(Y; 0, \Phi(\omega)\Phi(\omega)^T + \tau^{-1}I_N) \]

\[ p(Y|X) = \int p(Y|X, \omega)p(\omega)d\omega, \]

**Introduce auxiliary random variables** \( A \in \mathbb{R}^{K \times D} \)

\[ A \sim \mathcal{N}(0, I_{K \times D}), \]

\[ p(Y|X, A, \omega) = \mathcal{N}(Y; \Phi(\omega)A, \tau^{-1}I_N) \]

\[ p(Y|X) = \int p(Y|X, A, \omega)p(A)p(\omega)d\omega dA. \]
Introduce auxiliary random variables $A \in \mathbb{R}^{K \times D}$

$$A \sim \mathcal{N}(0, I_{K \times D}),$$

$$p(Y|X, A, \omega) = \mathcal{N}(Y; \Phi(\omega)A, \tau^{-1}I_N)$$

$$p(Y|X) = \int p(Y|X, A, \omega)p(A)p(\omega)d\omega dA,$$

Use variational distribution $q(\omega) = \prod q(w_k)q(b_k)\prod q(a_d)$ to approximate posterior $p(\omega, A|X, Y)$:

$$q(a_d) = \mathcal{N}(m_d, s_d)$$

over the rows of $A$ with $s_d$ diagonal.
Maximise log evidence lower bound

\[ \mathcal{L}_{fVSSGP} = \sum_{d=1}^{D} \left( \tau y_d^T E_{q(\omega)}(\Phi) m_d - \frac{\tau}{2} \text{tr} \left( E_{q(\omega)}(\Phi^T \Phi) \left( s_d + m_d m_d^T \right) \right) \right) \\
+ \ldots - \text{KL}(q(A) \| p(A)) - \text{KL}(q(\omega) \| p(\omega)). \]

Requires \( O(NK^2) \) time complexity — no matrix inversion.
Parallel inference with \( T \) workers \( \sim O\left( \frac{NK^2}{T} \right) \).
Stochastic Factorised VSSGP (sfVSSGP)

- We often use large $N$.

- $N$ matrix products of size $K \times K$ is still slow: $O(NK^2)$.

- It is silly to evaluate the objective over the entire dataset — might have redundant data.

- We can do even better with stochastic optimisation.
Maximise log evidence lower bound

$$\mathcal{L}_{sfVSSGP} \approx \frac{N}{|S|} \sum_{n \in S} \sum_{d=1}^{D} \mathcal{L}_{nd} - KL(q(A)||p(A)) - KL(q(\omega)||p(\omega))$$

with random data subset $S$ this is an unbiased estimator of $\mathcal{L}_{fVSSGP}$.

$O(SK^2)$ time complexity with $S << N$ size of random subset.
## Time Complexity Comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full GP</td>
<td>$O(N^3)$</td>
</tr>
<tr>
<td>SPGP / SSGP / VSSGP</td>
<td>$O(NK^2 + K^3)$</td>
</tr>
<tr>
<td>Factorised SPGP</td>
<td>$O(NK^2 + K^3)$</td>
</tr>
<tr>
<td>Factorised VSSGP</td>
<td>$O(NK^2)$</td>
</tr>
<tr>
<td>Stochastic SPGP</td>
<td>$O(SK^2 + K^3)$, $S &lt;&lt; N$</td>
</tr>
<tr>
<td>Stochastic fVSSGP</td>
<td>$O(SK^2)$, $S &lt;&lt; N$</td>
</tr>
</tbody>
</table>

with $K$ number of inducing points.
Example results

Interpolation on the reconstructed solar irradiance dataset (SE covariance function, \( K = 50 \) inducing inputs):

*Sparse pseudo-input GP*

*Sparse Spectrum GP*

*Random Projections (\( K = 500 \))*

*Variational Sparse Spectrum GP*

*Full GP*

<table>
<thead>
<tr>
<th>Model</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solar SPGP</td>
<td>0.23</td>
<td>0.61</td>
</tr>
<tr>
<td>SSGP</td>
<td>0.15</td>
<td>0.63</td>
</tr>
<tr>
<td>( \text{RP}_1 )</td>
<td>0.32</td>
<td>0.65</td>
</tr>
<tr>
<td>( \text{RP}_2 )</td>
<td>0.04</td>
<td>0.76</td>
</tr>
<tr>
<td>GP</td>
<td>0.08</td>
<td>0.50</td>
</tr>
<tr>
<td>VSSGP</td>
<td>0.13</td>
<td>0.41</td>
</tr>
</tbody>
</table>

Interpolation RMSE on train / test sets
Many more results

Many more results in the paper:

- Variational SSGP properties (uncertainty increases far from data)

- From SSGP to variational SSGP (big improvement)

- VSSGP, factorised VSSGP, and stochastic factorised VSSGP (more assumptions = worse results; still better than SSGP)

- Stochastic variational inference comparison (better than SPGP SVI)

- Speed-accuracy trade-offs (better accuracy with larger $K$)
Many more results in the paper:

- Variational SSGP properties (uncertainty increases far from data)
- From SSGP to variational SSGP (big improvement)
  - VSSGP, factorised VSSGP, and stochastic factorised VSSGP (more assumptions = worse results; still better than SSGP)
  - Stochastic variational inference comparison (better than SPGP SVI)
- Speed-accuracy trade-offs (better accuracy with larger $K$)
Many more results in the paper:

- Variational SSGP properties (uncertainty increases far from data)
- From SSGP to variational SSGP (big improvement)
- VSSGP, factorised VSSGP, and stochastic factorised VSSGP (more assumptions = worse results; still better than SSGP)
- Stochastic variational inference comparison (better than SPGP SVI)
- Speed-accuracy trade-offs (better accuracy with larger $K$)
Many more results

Many more results in the paper:

- Variational SSGP properties (uncertainty increases far from data)
- From SSGP to variational SSGP (big improvement)
- VSSGP, factorised VSSGP, and stochastic factorised VSSGP (more assumptions = worse results; still better than SSGP)
- Stochastic variational inference comparison (better than SPGP SVI)
- Speed-accuracy trade-offs (better accuracy with larger $K$)
Many more results in the paper:

- Variational SSGP properties (uncertainty increases far from data)
- From SSGP to variational SSGP (big improvement)
- VSSGP, factorised VSSGP, and stochastic factorised VSSGP (more assumptions = worse results; still better than SSGP)
- Stochastic variational inference comparison (better than SPGP SVI)
- Speed-accuracy trade-offs (better accuracy with larger $K$)
Many more results in the paper:

▶ Variational SSGP properties (uncertainty increases far from data)

▶ From SSGP to variational SSGP (big improvement)

▶ VSSGP, factorised VSSGP, and stochastic factorised VSSGP (more assumptions = worse results; still better than SSGP)

▶ Stochastic variational inference comparison (better than SPGP SVI)

▶ Speed-accuracy trade-offs (better accuracy with larger $K$)

Thank you