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

Yarin Gal, Richard Turner

University of Cambridge
{yg279, ret26}@cam.ac.uk

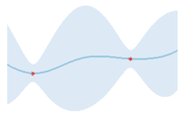
## In short:

- Gaussian process sparse pseudo-input approximations **cannot handle complex functions well**.
- Alternatively, sparse spectrum approximations are **known to over-fit**.

We develop **variational inference for the sparse spectrum approximation** avoiding both issues, **treating the covariance function as a random variable**.

## Background

**What is a Gaussian process (GP)?**



- A powerful tool in statistics, robust to over-fitting.
- Models distributions over functions.
- Supervised/unsupervised, regression/classification.
- Offers uncertainty estimates over the function values (in blue).

- Given training inputs $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N \in \mathbb{R}^{N\times Q}$ and outputs $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N \in \mathbb{R}^{N\times D}$, estimate a function $\mathbf{y} = \mathbf{f}(\mathbf{x})$ that is **likely to have generated** $\mathbf{Y}$.
- We place a joint Gaussian distribution over all function values:
$$p(\mathbf{Y} \mid \mathbf{X}) = \mathcal{N}(\mathbf{0}, \mathbf{K}(\mathbf{X},\mathbf{X}) + \tau^{-1}\mathbf{I}_N)$$
with precision hyper-parameter $\tau$ and covariance function $\mathbf{K}(\mathbf{X},\mathbf{X})$.

**What is variational inference?**

- Condition the model on a finite set of random variables $\boldsymbol{\omega}$.
- The predictive distribution for a new input point $\mathbf{x}^*$
$$p(\mathbf{y}^*|\mathbf{x}^*,\mathbf{X},\mathbf{Y}) = \int p(\mathbf{y}^*|\mathbf{x}^*,\boldsymbol{\omega}) p(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})\ \mathrm{d}\boldsymbol{\omega},$$
- The distribution $p(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})$ cannot be evaluated analytically — define an "easier" approximating *variational* distribution $q(\boldsymbol{\omega})$.
- Minimise the Kullback–Leibler (KL) divergence: $\mathrm{KL}(q(\boldsymbol{\omega}) \mid p(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y}))$.
- Minimising the KL divergence = maximising *log evidence lower bound*,
$$\mathcal{L}_{\mathrm{VI}} := \int q(\boldsymbol{\omega}) \log p(\mathbf{Y}|\mathbf{X},\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} - \mathrm{KL}(q(\boldsymbol{\omega})||p(\boldsymbol{\omega}))$$
with respect to the variational parameters defining $q(\boldsymbol{\omega})$.

## Variational Sparse Spectrum Approximation

1. We are given the Fourier transform of the squared exponential covariance function
$$\mathbf{K}(\mathbf{x}-\mathbf{y}) = \int \sigma^2 \mathcal{N}(\mathbf{w}; 0, \mathbf{I}_Q) \cos\big(2\pi \mathbf{w}^T(\mathbf{x}-\mathbf{y})\big)\mathrm{d}\mathbf{w}$$
2. Introduce auxiliary variable $b$:
$$= \int 2\sigma^2 \mathcal{N}(\mathbf{w}; 0, \mathbf{I}_Q)\mathrm{Unif}[0,2\pi] \cos\big(2\pi \mathbf{w}^T\mathbf{x} + b\big) \cos\big(2\pi \mathbf{w}^T\mathbf{y} + b\big)\mathrm{d}\mathbf{w}\mathrm{d}b$$
3. Approximate with Monte Carlo integration with $K$ terms:
$$\widehat{\mathbf{K}}(\mathbf{x}-\mathbf{y}) = \frac{\sigma^2}{K}\sum_{k=1}^K \sqrt{2}\cos\big(2\pi \mathbf{w}_k^T(\mathbf{x}-\mathbf{z}_k) + b_k\big)\sqrt{2}\cos\big(2\pi \mathbf{w}_k^T(\mathbf{y}-\mathbf{z}_k) + b_k\big)$$
with $\mathbf{w}_k \sim \mathcal{N}(0, \mathbf{I}_Q)$, $b_k \sim \mathrm{Unif}[0, 2\pi]$, and variational parameters $\mathbf{z}_k$. This is a **random covariance function**.
4. The GP is re-parametrised as
$$\mathbf{w}_k \sim \mathcal{N}(0, \mathbf{I}_Q), \quad b_k \sim \mathrm{Unif}[0, 2\pi], \quad \boldsymbol{\omega} = \{\mathbf{w}_k, b_k\}_{k=1}^K$$
$$\Phi_{n,k}(\boldsymbol{\omega}) = \sqrt{\frac{2\sigma^2}{K}}\cos\big(2\pi \mathbf{w}_k^T(\mathbf{x}_n - \mathbf{z}_k) + b_k\big), \quad \Phi \in \mathbb{R}^{N\times K}$$
$$p(\mathbf{Y}|\mathbf{X},\boldsymbol{\omega}) = \mathcal{N}\big(\mathbf{Y}; \mathbf{0}, \Phi(\boldsymbol{\omega})\Phi(\boldsymbol{\omega})^T + \tau^{-1}\mathbf{I}_N\big)$$
$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{X},\boldsymbol{\omega}) p(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega}$$
5. Use variational distribution $q(\boldsymbol{\omega}) = \prod_{k=1}^K q(\mathbf{w}_k) q(b_k)$ to approximate posterior $p(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})$:
$$q(\mathbf{w}_k) = \mathcal{N}(\mu_k, \Sigma_K), \quad q(b_k) = \mathrm{Unif}(\alpha_k, \beta_k),$$
with $\Sigma_K$ diagonal.
6. We maximise the log evidence lower bound:
$$\mathcal{L}_{VSSGP} = \sum_{d=1}^D \Big( -\frac{N}{2}\log(2\pi\tau^{-1}) - \frac{\tau}{2}\mathbf{y}_d^T\mathbf{y}_d + \frac{1}{2}\log(|\tau^{-1}\boldsymbol{\Sigma}|) + \frac{1}{2}\tau \mathbf{y}_d^T E_{q(\boldsymbol{\omega})}(\Phi)\boldsymbol{\Sigma} E_{q(\boldsymbol{\omega})}(\Phi^T)\mathbf{y}_d\Big) - \mathrm{KL}(q(\boldsymbol{\omega})||p(\boldsymbol{\omega}))$$
with $\boldsymbol{\Sigma} = (E_{q(\boldsymbol{\omega})}(\Phi^T\Phi) + \tau^{-1}I)^{-1}$. We can evaluate the KL and the expectations analytically using the identity
$$E_{q(\mathbf{w})}\big(\cos(\mathbf{w}^T\mathbf{x} + b)\big) = e^{-\frac{1}{2}\mathbf{x}^T\Sigma\mathbf{x}}\cos(\mu^T\mathbf{x} + b).$$
7. This requires $\mathcal{O}(NK^2 + K^3)$ time complexity. Using parallel inference with $T$ workers this reduces to $\mathcal{O}(\frac{NK^2}{T} + K^3)$.

## Factorised Variational Sparse Spectrum Approximation

We often use large $K$, for which the above is still slow. We can do better.

1. We introduce auxiliary random variables $\mathbf{a}_d$. The GP is re-parametrised as
$$\mathbf{a}_d \sim \mathcal{N}(0, \mathbf{I}_K), \quad \mathbf{A} = [\mathbf{a}_d]_{d=1}^D \in \mathbb{R}^{K\times D}$$
$$p(\mathbf{Y}|\mathbf{X},\mathbf{A},\boldsymbol{\omega}) = \mathcal{N}\big(\mathbf{Y}; \Phi(\boldsymbol{\omega})\mathbf{A}, \tau^{-1}\mathbf{I}_N\big)$$
$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{X},A,\boldsymbol{\omega}) p(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega}\mathrm{d}\mathbf{A}$$
2. Use variational distribution $q(\boldsymbol{\omega}) = \prod_{k=1}^K q(\mathbf{w}_k) q(b_k) \prod_{d=1}^D q(\mathbf{a}_d)$ to approximate posterior $p(\boldsymbol{\omega}, \mathbf{A}|\mathbf{X},\mathbf{Y})$:
$$q(\mathbf{a}_d) = \mathcal{N}(\mathbf{m}_d, \mathbf{s}_d)$$
with $\mathbf{s}_K$ diagonal.
3. This requires $\mathcal{O}(NK^2)$ time complexity — no matrix inverse. Using parallel inference with $T$ workers this reduces to $\mathcal{O}(\frac{NK^2}{T})$.

## Stochastic Factorised Variational Sparse Spectrum Approximation

With stochastic optimisation, we can do even better.

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

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

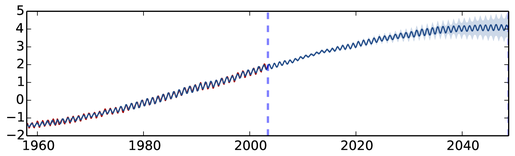
- Requires $\mathcal{O}(SK^2)$ time complexity with $S << N$ size of random subset, compared to $\mathcal{O}(SK^2 + K^3)$ of GP SVI using sparse pseudo-input approximation.

## Approximation Properties

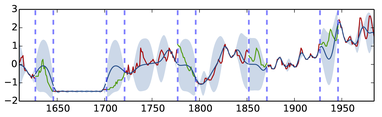
### Model uncertainty

- Model uncertainty in interpolation and extrapolation tasks

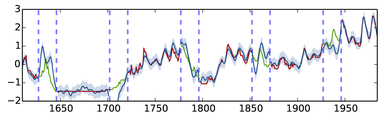
Extrapolation on the Mauna Loa $CO_2$ concentrations dataset (sum of two SE covariance functions, $K = 10$ inducing inputs):



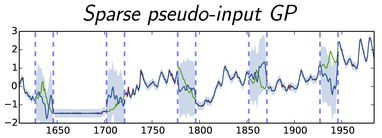
Interpolation on the reconstructed solar irradiance dataset (SE covariance function, $K = 50$ inducing inputs):



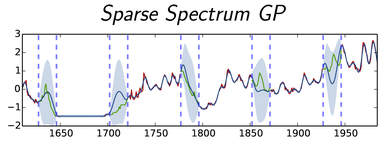
*Sparse pseudo-input GP*



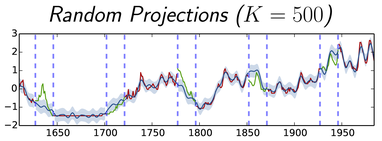
*Sparse Spectrum GP*



*Random Projections ($K = 500$)*



*Full GP*



**Variational Sparse Spectrum GP**

| Solar | SPGP | SSGP | $RP_1$ | $RP_2$ | GP | **VSSGP** |
|---|---|---|---|---|---|---|
| Train | 0.23 | 0.15 | 0.32 | 0.04 | 0.08 | **0.13** |
| Test | 0.61 | 0.63 | 0.65 | 0.76 | 0.50 | **0.41** |

*Interpolation RMSE on train / test sets*

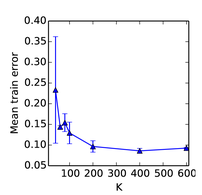
### Imputation accuracy

- Imputation RMSE on train / test sets, for a speech signal of length 1K ($K = 100$):

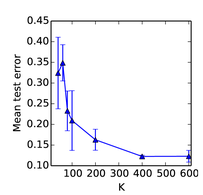
| Audio 1K | SSGP | VSSGP | fVSSGP | sfVSSGP |
|---|---|---|---|---|
| Train | $0.0091 \pm 0.0042$ | $\mathbf{0.0062 \pm 0.00048}$ | $\mathbf{0.0054 \pm 0.00083}$ | $\mathbf{0.005 \pm 0.003}$ |
| Test | $0.088 \pm 0.033$ | $\mathbf{0.034 \pm 0.0043}$ | $\mathbf{0.038 \pm 0.0049}$ | $\mathbf{0.04 \pm 0.0066}$ |

### Speed-Accuracy trade-off

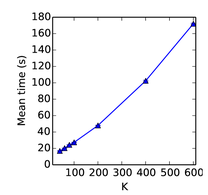
- Train error, test error, and running time as functions of number of inducing points ($K$) for a speech signal of length 4K:




*Train error*




*Test error*




*Running time*

**Code available at `http://github.com/yaringal/VSSGP`.**