Extended and Unscented Kitchen Sinks

Edwin V. Bonilla
e.bonilla@unsw.edu.au
The University of New South Wales

Daniel Steinberg
Daniel.Steinberg@nicta.com.au
NICTA

Alistair Reid
Alistair.Reid@nicta.com.au
NICTA

Gaussian Process Models

We consider models of the form $\mathbf{y} = \mathbf{g}(\mathbf{f}) + \boldsymbol{\epsilon}$, where $\mathbf{f}$ is drawn from a Gaussian process (GP):

- Standard supervised learning settings
- Inversion problems

Key challenges:

1. Scalability on the number observations
2. Muti-task settings
3. Nonlinear likelihoods $\mathbf{g}(\mathbf{f}) + \boldsymbol{\epsilon}$

Our solution considers:

- Random feature approximations to the covariance function (1);
- Affine transformations of latent processes (2); and
- Local and adaptive linearizations (3);

all within a single *variational inference* framework.

Multi-output Setting

We consider the supervised learning problem:

- Data: $\{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^N$, described compactly as $\{\mathbf{X}, \mathbf{Y}\}$, where $\mathbf{X} \in \mathbb{R}^{N \times d}$ and $\mathbf{Y} \in \mathbb{R}^{N \times P}$
- Prior: Q latent functions $\{f_q\}$ drawn from independent GP priors with covariance $k_q(\cdot,\cdot)$:

$$p(\mathbf{F}) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{f}_{\cdot q}; \mathbf{0}, \mathbf{K}_q) \tag{1}$$

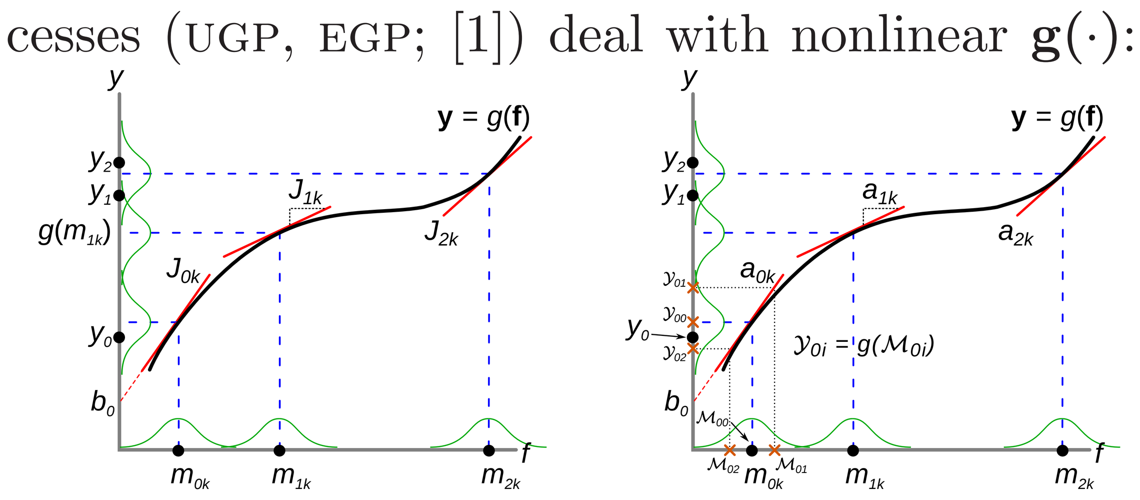
- Non-linear forward model: $\mathbf{g} : \mathbb{R}^Q \to \mathbb{R}^P$ and likelihood:

$$p(\mathbf{Y}|\mathbf{F}) = \prod_{n=1}^{N} p(\mathbf{y}_n | \mathbf{g}(\mathbf{f}_{n\cdot})), \tag{2}$$

Goal: Probabilistic predictions and posterior estimation $p(\mathbf{F}|\mathbf{Y})$

Extended and Unscented GPs

The Extended and Unscented Gaussian processes (UGP, EGP; [1]) deal with nonlinear $\mathbf{g}(\cdot)$:



EGP — Taylor expansion UGP — stat. linearization

☺ Approximation is local and adaptive
☹ Single output/task, $Q = 1$
☹ Non-scalable inference, $\mathcal{O}(N^3)$ in time

Random Kitchen Sinks

To achieve scalability, we use Random Kitchen Sinks (RKS, [2]) approximations to the kernel:

- Exploit Fourier duality of covariance function of stationary process and its spectral density:

$$k(\boldsymbol{\tau}) = \int S(\mathbf{s}) e^{2\pi i \mathbf{s}^T \boldsymbol{\tau}} d\mathbf{s} \longleftrightarrow S(\mathbf{s}) = \int k(\boldsymbol{\tau}) e^{-2\pi i \mathbf{s}^T \boldsymbol{\tau}} d\boldsymbol{\tau}. \tag{3}$$

- Approximate the above kernel by explicitly constructing "suitable" random features and (Monte Carlo) averaging over samples from $S(\mathbf{s})$:

$$k(\mathbf{x} - \mathbf{x}') = k(\boldsymbol{\tau}) \approx \frac{1}{D} \sum_{i=1}^{D} \phi_i(\mathbf{x}) \, \phi_i(\mathbf{x}'), \tag{4}$$

- For example, $[\phi_i(\mathbf{x}), \phi_{D+i}(\mathbf{x})] = \frac{1}{\sqrt{D}} [\cos(2\pi \mathbf{s}_i^T \mathbf{x}), \sin(2\pi \mathbf{s}_i^T \mathbf{x})]$ with $\mathbf{s}_i \sim \mathcal{N}(\mathbf{s}_i | \mathbf{0}, \sigma_\phi^2 \mathbf{I}_d)$, for $i = 1, \ldots, D$, converges in expectation to the (isotropic) squared exponential kernel.

Approximate Model

Using RKS bases such that $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{E}[\phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)]$, we approximate our GP models with:

$$p(\mathbf{W}) = \prod_{q=1}^{Q} \mathcal{N}\left(\mathbf{w}_q | \mathbf{0}, \omega_q^2 \mathbf{I}_D\right), \tag{5}$$

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{y}_n | \mathbf{g}(\mathbf{W}\boldsymbol{\phi}_n), \boldsymbol{\Sigma}), \text{ where} \tag{6}$$

- $\boldsymbol{\phi}_n \stackrel{\text{def}}{=} \phi(\mathbf{x}_n)$ is the D-dimensional vector of features corresponding to datapoint n;
- $\mathbf{w}_q \in \mathbb{R}^D$; $\mathbf{W} \in \mathbb{R}^{Q \times D}$; ω_q^2 is the prior variance over the weights; and
- $\boldsymbol{\Sigma} = \text{diag}\left(\left[\sigma_1^2, \ldots, \sigma_P^2\right]\right)$ is the noise variance.

Note that, effectively, we are making $\mathbf{f}_q = \boldsymbol{\Phi}\mathbf{w}_q$, with $\boldsymbol{\Phi} \stackrel{\text{def}}{=} \phi(\mathbf{X})$ being the $N \times D$ matrix of features.

Posterior Inference

To deal with the nonlinear likelihood in Eq. (6), we use variational inference with the approximate posterior:

$$\tilde{q}_{\mathbf{W}} \stackrel{\text{def}}{=} \tilde{q}(\mathbf{W}) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{w}_q | \mathbf{m}_q, \mathbf{C}_q), \tag{7}$$

thereby yielding the posterior latent tasks,

$$\tilde{q}(\mathbf{F}) = \prod_{q=1}^{Q} \mathcal{N}\left(\mathbf{f}_{\cdot q} | \boldsymbol{\Phi}\mathbf{m}_q, \boldsymbol{\Phi}\mathbf{C}_q\boldsymbol{\Phi}^T\right). \tag{8}$$

Variational Objective: The variational log-evidence lower bound is,

$$\mathcal{L} = \langle \log p(\mathbf{Y}|\mathbf{W}) \rangle_{\tilde{q}_{\mathbf{W}}} - \text{KL}[\tilde{q}(\mathbf{W}) \,\|\, p(\mathbf{W})]. \tag{9}$$

While the KL term is straightforward, the log likelihood term involves an expectation of a nonlinear function:

$$\left\langle (\mathbf{y}_n - \mathbf{g}(\mathbf{W}\boldsymbol{\phi}_n))^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_n - \mathbf{g}(\mathbf{W}\boldsymbol{\phi}_n)) \right\rangle_{\tilde{q}_{\mathbf{W}}}, \tag{10}$$

which we approximate using:

$$\mathbf{g}(\mathbf{W}\boldsymbol{\phi}_n) \approx \mathbf{A}_n \mathbf{W}\boldsymbol{\phi}_n + \mathbf{b}_n. \tag{11}$$

- The objective factorizes over the data → parallel or stochastic gradient algorithms easily applicable.
- Methods: how to linearize (set $\mathbf{A}_n$, $\mathbf{b}_n$)? → EKS vs. UKS

Extended Kitchen Sinks

EKS uses a first-order Taylor series,

$$\mathbf{g}(\mathbf{W}\boldsymbol{\phi}_n) \approx \mathbf{g}(\mathbf{M}\boldsymbol{\phi}_n) + \mathbf{J}_n (\mathbf{W} - \mathbf{M}) \boldsymbol{\phi}_n, \tag{12}$$

where $\mathbf{J}_n = \left.\frac{\partial \mathbf{g}(\mathbf{f}_{n\cdot})}{\partial \mathbf{f}_{n\cdot}}\right|_{\mathbf{f}_{n\cdot} = \mathbf{M}\boldsymbol{\phi}_n}$, obtaining:

$$\mathbf{A}_n = \mathbf{J}_n \text{ and } \mathbf{b}_n = \mathbf{g}(\mathbf{M}\boldsymbol{\phi}_n) - \mathbf{J}_n \mathbf{M}\boldsymbol{\phi}_n. \tag{13}$$

Unscented Kitchen Sinks

UKS estimates the linearization parameters using deterministic samples given by the unscented transform:

1. Exploit the structure from the marginal posterior $\tilde{q}(\mathbf{f}_{n\cdot}) = \mathcal{N}(\mathbf{f}_{n\cdot} | \boldsymbol{\mu}_n, \mathbf{E}_n)$
2. Define $2Q + 1$ so-called sigma-points $\mathcal{F}_{i,n}$, labels $\mathcal{Y}_{i,n} = \mathbf{g}(\mathcal{F}_{i,n})$ and weights u_i
3. Solve the weighted linear least squares problems with inputs, outputs, and weights $\{\mathcal{F}_{i,n}, \mathcal{Y}_{i,n}, u_i\}$:

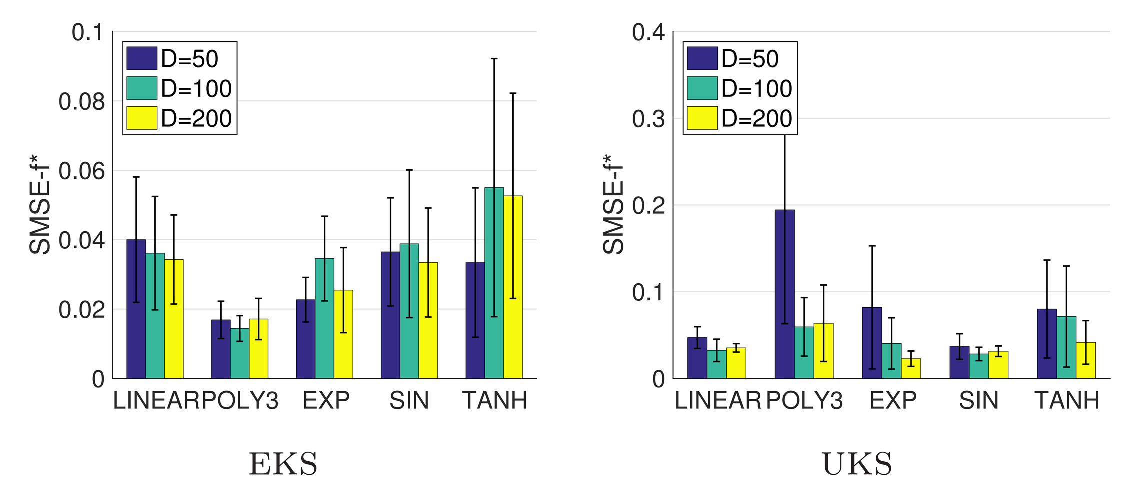
$$\mathbf{b}_n = \bar{\mathbf{y}}_n - \mathbf{A}_n \mathbf{M}\boldsymbol{\phi}_n \text{ and } \mathbf{A}_n = \boldsymbol{\Gamma}_n \mathbf{E}_n^{-1}, \tag{14}$$

where $\bar{\mathbf{y}}_n$ and $\boldsymbol{\Gamma}_n$ are the sufficient statistics.

UKS *is truly a 'black-box' method*

Experiments

Synthetic inversion problems:



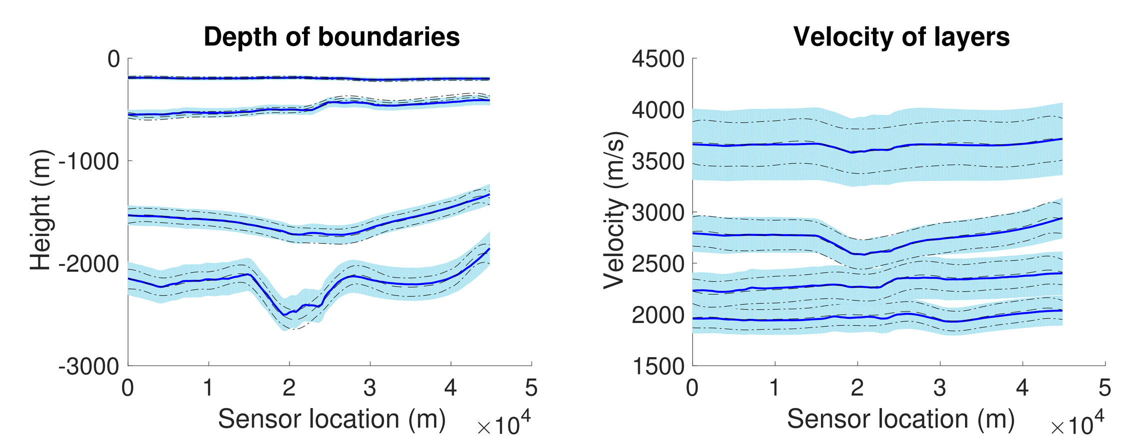
→ EKS and UKS perform similarly to original EGP and UGP and are competitive with the state-of-the-art (not shown)

Odd digits vs even digits on MNIST:

	NLP		Error Rate	
	$D = 1000$	$D = 2000$	$D = 1000$	$D = 2000$
EKS	0.129	0.088	0.043	0.026
UKS	0.129	0.088	0.043	0.026
[3]	0.069		0.022	
[4]	0.068		0.022	

→ Similar performance to recently developed inducing-point approximations.

Seismic inversion:



→ Similar solution to long-running MCMC simulation.

References

[1] D. M. Steinberg and E. V. Bonilla, "Extended and unscented Gaussian processes," in *NIPS*, 2014.
[2] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *NIPS*, 2008.
[3] J. Hensman, A. Matthews, and Z. Ghahramani, "Scalable variational Gaussian process classification," in *AISTATS*, 2015.
[4] A. Dezfouli and E. V. Bonilla, "Scalable inference for Gaussian process models with black-box likelihoods," in *NIPS*, 2015.