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 \( y = g(f) + \epsilon \), where \( f \) is drawn from a Gaussian process (GP):
- Standard supervised learning settings
- Inversion problems

Key challenges:
1. Scalability on the number observations
2. Multi-task settings
3. Nonlinear likelihoods \( g(f) + \epsilon \)

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

all within a single variational inference framework.

Multi-output Setting

We consider the supervised learning problem:
- Data: \( \{x_n, y_n\}_{n=1}^N \), described compactly as \( \{X, Y\} \), where \( X \in \mathbb{R}^{N \times d} \) and \( Y \in \mathbb{R}^{N \times r} \);
- Prior: \( Q \) latent functions \( \{f_q\} \) drawn from independent GP priors with covariance \( \kappa_q(\cdot, \cdot) \);

\[
p(F) = \prod_{q=1}^Q N(f_q; 0, K_q) \tag{1}
\]
- Non-linear forward model: \( g: \mathbb{R}^Q \rightarrow \mathbb{R}^P \) and likelihood:

\[
p(Y|F) = \prod_{n=1}^N p(y_n|g(f_n)) \tag{2}
\]

Goal: Probabilistic predictions and posterior estimation \( p(F|Y) \).

Approximate Model

Using RKs bases such that \( k(x_n, x_m) = E[\phi(x_n)\phi(x_m)] \), we approximate our GP models with:

\[
p(W) = \prod_{q=1}^Q N(w_q; 0, \omega_q^2 I_D),
\]

\[
p(Y|W) = \prod_{n=1}^N N(y_n|g(W\phi_n), \Sigma),
\]

- \( \phi_n \sim \phi(x_n) \) is the \( D \)-dimensional vector of features corresponding to datapoint \( n \);
- \( w_q \in \mathbb{R}^{D_q}; W \in \mathbb{R}^{Q \times D}; \omega_q^2 \) is the prior variance over the weights; and

\[
\Sigma = \text{diag}(\sigma_f^2, \ldots, \sigma_f^2) \text{ is the noise variance.}
\]

Note that, effectively, we are making \( f_q = \Phi w_q \), with \( \Phi \sim \tilde{\Phi}(X) \) 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:

\[
\tilde{q}(W) = \prod_{q=1}^Q N(w_q|m_q, C_q),
\]

thereby yielding the posterior latent tasks,

\[
\tilde{q}(W) = \prod_{q=1}^Q N(f_q|\Phi m_q, \Phi C_q \Phi^T).
\]

VARIATIONAL OBJECTIVE: The variational log-evidence lower bound is,

\[
\mathcal{L} = \langle \log p(Y|W) \rangle_{\tilde{q}(W)} - KL[\tilde{q}(W) || p(W)].
\]

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

\[
\langle (y_n - g(W\phi_n)^\top \Sigma^{-1} (y_n - g(W\phi_n)) \rangle_{\tilde{q}(W)},
\]

which we approximate using:

\[
g(W\phi_n) = A_n W\phi_n + b_n.
\]

- The objective factorizes over the data \( \rightarrow \) parallel or stochastic gradient algorithms easily applicable.
- Methods: how to linearize (set \( A_n, b_n \) )?

\( \rightarrow \) EKS vs. UKS

Extended and Unscented GPs

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

\[
\rightarrow \text{Approximation is local and adaptive}
\]

\[
\rightarrow \text{Single output/task, } Q = 1
\]

\[
\rightarrow \text{Non-scalable inference, } \mathcal{O}(N^3) \text{ in time}
\]

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:

\[
k(\tau) = \int S(s)e^{-2\pi i \tau s}ds \quad \rightarrow \quad S(s) = \int k(\tau)e^{-2\pi i \tau \tau}d\tau.
\]

- Approximate the above kernel by explicitly constructing “suitable” random features and (Monte Carlo) averaging over samples from \( S(s) \):

\[
k(x - x') = k(\tau) = \frac{1}{D} \sum_{d=1}^D \phi_i(x)\phi_i(x'),
\]

- For example, \( \phi_i(x) = \frac{1}{\sqrt{D}}[\cos(2\pi s_i^T x), \sin(2\pi s_i^T x)] \) with \( s_i \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_t) \), for \( i = 1, \ldots, D \), converges in expectation to the (isotropic) squared exponential kernel.

Extended Kitchen Sinks

EKS uses a first-order Taylor series,

\[
g(W\phi_n) = g(M\phi_n) + J_n(W - M)\phi_n,
\]

where \( J_n = \frac{\partial g(f_n)}{\partial f_n} |_{f_n=M\phi_n} \) obtaining:

\[
A_n = J_n \text{ and } b_n = g(M\phi_n) - J_n M\phi_n;
\]

Unscented Kitchen Sinks

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

1. Exploit the structure from the marginal posterior \( q(f_n) = N(f_n|\mu_n, S_n) \).
2. Define 2\( Q \) + 1 so-called sigma-points \( \mathcal{F}_{i,n} \), labels \( \mathcal{Y}_{i,n} = g(F_{i,n}) \) and weights \( \alpha \).
3. Solve the weighted linear least squares problems with inputs, outputs, and weights \( \{F_{i,n}, Y_{i,n}, W_\alpha\} \):

\[
b_n = Y_{n} - A_n M\phi_n \text{ and } A_n = \Gamma_n E_n, \tag{14}
\]

where \( y_n \) and \( \Gamma_n \) are the sufficient statistics.

UKS is truly a ‘black-box’ method.

Experiments

SYNTHETIC INVERSION PROBLEMS:

\begin{align*}
\text{NLP} & \quad D = 1000 \quad D = 2000 \quad D = 1000 \quad D = 2000 \\
\text{EKS} & \quad 0.129 \quad 0.088 \quad 0.043 \quad 0.026 \\
\text{UKS} & \quad 0.129 \quad 0.088 \quad 0.043 \quad 0.026 \\
\end{align*}