Probabilistic Approximate Least-Squares (APPENDIX)

Simon Bartels **Philipp Hennig**
Max Planck Institute for Intelligent Systems
Tübingen Germany

An Upper Bound on the Approximation Error

Let $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^N$ be arbitrary vectors. The Gaussian measure (11) on $\boldsymbol{H}$ implies that the scalar $\hat{\mu} := \boldsymbol{a}^\intercal \boldsymbol{H} \boldsymbol{b}$ is Gaussian distributed as well, with mean $\boldsymbol{a}^\intercal \boldsymbol{H}_M \boldsymbol{b}$ and a variance we denote with $\hat{\epsilon}^2$ (derived in the proof below).

Theorem 1. *The absolute error* $|\hat{\mu} - \boldsymbol{a}^\intercal \boldsymbol{H} \boldsymbol{b}|$ *divided by the standard deviation* $\hat{\epsilon}$ *is always less than* 1*:*

$$\frac{|\boldsymbol{a}^\intercal \boldsymbol{H}_M \boldsymbol{b} - \boldsymbol{a}^\intercal \boldsymbol{H} \boldsymbol{b}|}{\hat{\epsilon}} < 1 \tag{32}$$

Proof. If $\boldsymbol{v} \in \mathbb{R}^N$ is a Gaussian random vector $\mathcal{N}(\boldsymbol{v}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ is of rank M then $\boldsymbol{A}\boldsymbol{v}$ is also Gaussian with $\mathcal{N}(\boldsymbol{A}\boldsymbol{v}; \boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}^\intercal \boldsymbol{\Sigma} \boldsymbol{A})$. We can rewrite $\boldsymbol{a}^\intercal \boldsymbol{H}_M \boldsymbol{b}$ as $\overrightarrow{\boldsymbol{a}^\intercal \boldsymbol{H}_M \boldsymbol{b}} = (\boldsymbol{a}^\intercal \otimes \boldsymbol{b}^\intercal) \overrightarrow{\boldsymbol{H}_M}$. and it follows that $\hat{\epsilon}$ has the form

$$\hat{\epsilon}^2 = (\boldsymbol{a}^\intercal \otimes \boldsymbol{b}^\intercal)(\boldsymbol{W}_M \otimes \boldsymbol{W}_M)(\boldsymbol{a} \otimes \boldsymbol{b}) \tag{33}$$

To simplify this expression we use that $\boldsymbol{a} \otimes \boldsymbol{b}$ is an N^2 dimensional vector and thus $\boldsymbol{a} \otimes \boldsymbol{b} = (\boldsymbol{a} \otimes \boldsymbol{b}) \overrightarrow{1} = \overrightarrow{\boldsymbol{a} 1 \boldsymbol{b}^\intercal} = \overrightarrow{\boldsymbol{a}\boldsymbol{b}^\intercal}$. Therefore $\hat{\epsilon}^2$ reduces to

$$\begin{aligned}
\hat{\epsilon}^2 &= (\boldsymbol{a}^\intercal \otimes \boldsymbol{b}^\intercal) \boldsymbol{\Gamma} (\boldsymbol{W}_M \otimes \boldsymbol{W}_M) \boldsymbol{\Gamma} \overrightarrow{\boldsymbol{a}\boldsymbol{b}^\intercal} && (34)\\
&= \frac{1}{2} (\boldsymbol{a}^\intercal \otimes \boldsymbol{b}^\intercal) \boldsymbol{\Gamma} (\boldsymbol{W}_M \otimes \boldsymbol{W}_M) \overrightarrow{\boldsymbol{a}\boldsymbol{b}^\intercal + \boldsymbol{b}\boldsymbol{a}^\intercal} && (35)\\
&= \frac{1}{2} (\boldsymbol{a}^\intercal \otimes \boldsymbol{b}^\intercal) \boldsymbol{\Gamma} \overrightarrow{\boldsymbol{W}_M \boldsymbol{a}\boldsymbol{b}^\intercal \boldsymbol{W}_M + \boldsymbol{W}_M \boldsymbol{b}\boldsymbol{a}^\intercal \boldsymbol{W}_M} && (36)\\
&= \frac{1}{2} (\boldsymbol{a}^\intercal \otimes \boldsymbol{b}^\intercal) \overrightarrow{\boldsymbol{W}_M \boldsymbol{a}\boldsymbol{b}^\intercal \boldsymbol{W}_M + \boldsymbol{W}_M \boldsymbol{b}\boldsymbol{a}^\intercal \boldsymbol{W}_M} && (37)\\
&= \frac{1}{2} \overrightarrow{\boldsymbol{a}^\intercal \boldsymbol{W}_M \boldsymbol{a}\boldsymbol{b}^\intercal \boldsymbol{W}_M \boldsymbol{b} + \boldsymbol{a}^\intercal \boldsymbol{W}_M \boldsymbol{b}\boldsymbol{a}^\intercal \boldsymbol{W}_M \boldsymbol{b}} && (38)\\
&= \frac{1}{2} \left(\boldsymbol{a}^\intercal \boldsymbol{W}_M \boldsymbol{a}\boldsymbol{b}^\intercal \boldsymbol{W}_M \boldsymbol{b} + (\boldsymbol{a}^\intercal \boldsymbol{W}_M \boldsymbol{b})^2\right) && (39)
\end{aligned}$$

If we now choose $\boldsymbol{W} = \sqrt{2}\boldsymbol{H}$, and therefore $\boldsymbol{W}_M = \sqrt{2}(\boldsymbol{H} - \boldsymbol{H}_M)$, then

$$\begin{aligned}
&\frac{|\boldsymbol{a}^\intercal (\boldsymbol{H} - \boldsymbol{H}_M) \boldsymbol{b}|}{\sqrt{\hat{\epsilon}^2}} && (40)\\
=& \frac{|\boldsymbol{a}^\intercal \boldsymbol{W}_M \boldsymbol{b}|}{\sqrt{2}\sqrt{\frac{1}{2}\boldsymbol{a}^\intercal \boldsymbol{W}_M \boldsymbol{a} \cdot \boldsymbol{b}^\intercal \boldsymbol{W}_M \boldsymbol{b} + \frac{1}{2}(\boldsymbol{a}^\intercal \boldsymbol{W}_M \boldsymbol{b})^2}} && (41)\\
=& \frac{|\boldsymbol{a}^\intercal \boldsymbol{W}_M \boldsymbol{b}|}{\sqrt{\boldsymbol{a}^\intercal \boldsymbol{W}_M \boldsymbol{a} \cdot \boldsymbol{b}^\intercal \boldsymbol{W}_M \boldsymbol{b} + (\boldsymbol{a}^\intercal \boldsymbol{W}_M \boldsymbol{b})^2}} && (42)\\
<& \frac{|\boldsymbol{a}^\intercal \boldsymbol{W}_M \boldsymbol{b}|}{\sqrt{(\boldsymbol{a}^\intercal \boldsymbol{W}_M \boldsymbol{b})^2}} = 1 \qquad \text{(as } \boldsymbol{W}_M \text{ is s.p.d.)} && (43)
\end{aligned}$$

□

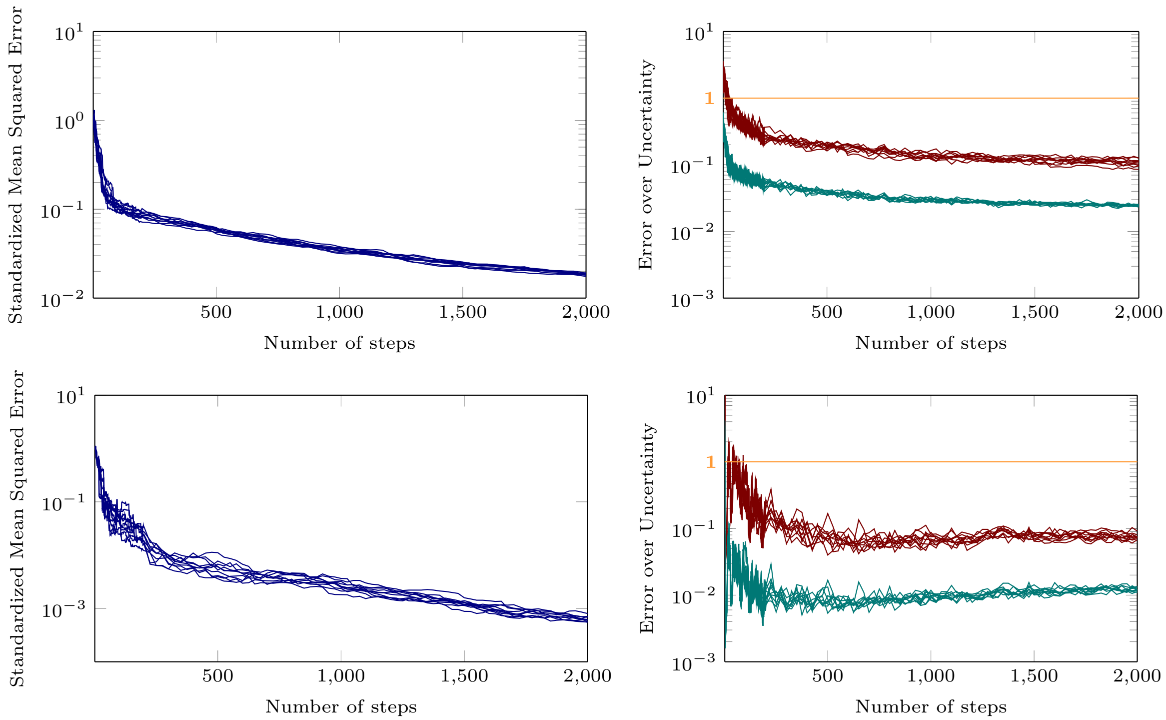

Figure 1: Ten Farthest Point Clustering initializations of the probabilistic subset of data approximation on the PUMADYN (**top row**) and CPU (**bottom row**) data sets, using the ARD Squared Exponential kernel. **Left:** standardized mean squared error for Subset of Data. **Right:** ratio between absolute error and uncertainty. The upper lines are the maximum, the lower lines the average over all test inputs. The horizontal line shows the theoretical bound at 1 that would be guaranteed if $\boldsymbol{W} = \sqrt{2}\boldsymbol{H}$ where estimated exactly.

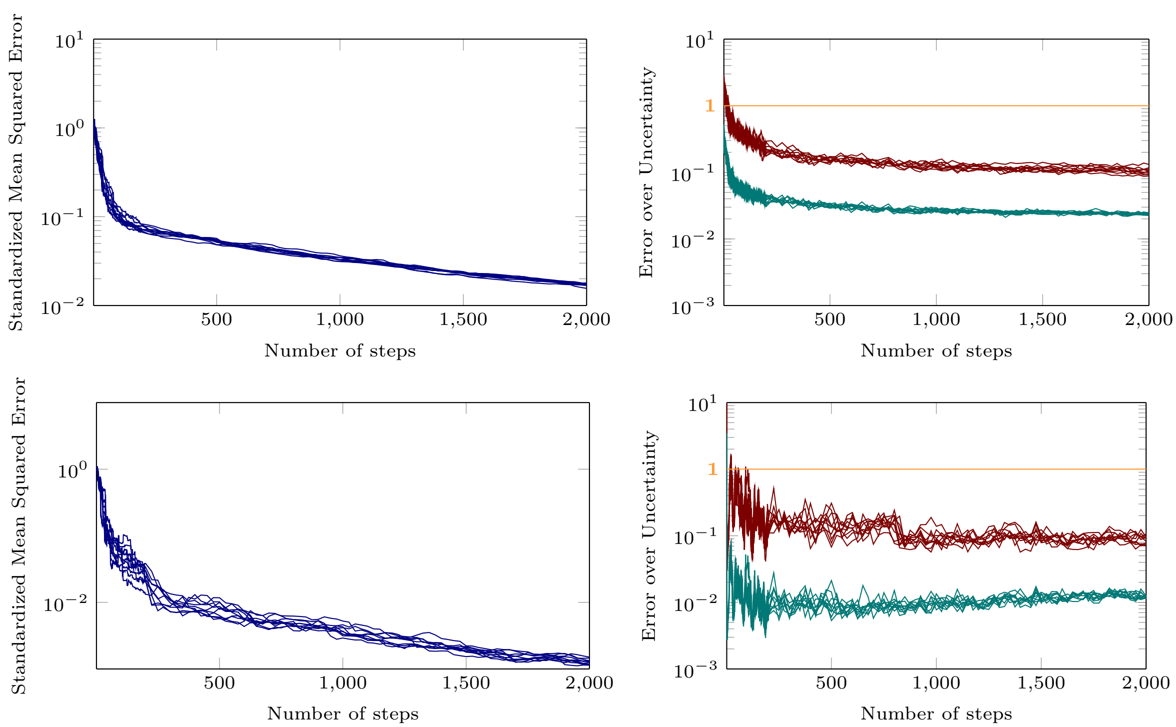

Figure 2: Same setup as Figure 1, but using the ARD Matérn $^5/_2$ kernel.