Appendices

A. Proofs

A.1. Proof of Proposition 1

Proof. Because $\boldsymbol{\eta}_1$ is drawn from an isotropic distribution, we can let $\boldsymbol{\tau} = \begin{bmatrix} x & 0 & 0 & ... & 0 \end{bmatrix}^\top$ without loss of generality. Then, define the expected kernel value at $\boldsymbol{\tau}$ as the expectation

$$\mathbb{E}[k(\boldsymbol{\eta}_1^\top \boldsymbol{\tau})] = \mathbb{E}[k(\eta_{11}||\boldsymbol{\tau}||_2)] =: k_e(||\boldsymbol{\tau}||_2)$$

Then, since k is bounded, by the strong law of large numbers, the empirical mean $\frac{1}{J}\sum_{j=1}^{J} k(\boldsymbol{\eta}_j^\top \boldsymbol{\tau})$ converges to the expectation $k_e(||\boldsymbol{\tau}||_2)$ almost surely as $J \to \infty$. □

A.2. Proof of Corollary 1

Proof. Let us write $r = ||\boldsymbol{\tau}||_2$ for simplicity.

We have $\boldsymbol{\eta}_1 \sim \mathcal{N}(0, I_d)$ which implies $\eta_{11} \sim \mathcal{N}(0, 1)$. Then,

$$\begin{aligned}
\mathbb{E}[e^{-\frac{1}{2}\eta_{11}^2 r^2}] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta_{11}^2 r^2} e^{-\frac{1}{2}\eta_{11}^2} d\eta_{11} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta_{11}^2 (1+r^2)} \\
&= \frac{1}{\sqrt{1+r^2}}.
\end{aligned}$$

□

A.3. Proof of Corollary 2

Proof. Let us write $r = ||\boldsymbol{\tau}||_2$.

We have again $\boldsymbol{\eta}_1 \sim \mathcal{N}(0, I_d)$ which implies $\eta_{11} \sim \mathcal{N}(0, 1)$. Then,

$$\begin{aligned}
\mathbb{E}[\cos(\eta_{11} r)] &= \mathbb{E}\left[\sum_{j=0}^{\infty} \frac{(-1)^j}{(2j)!} (\eta_{11} r)^{2j}\right] \\
&= \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j)!} r^{2j} \mathbb{E}[\eta_{11}^{2j}] \\
&= \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j)!} r^{2j} (2j-1)!! \\
&= \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j)!!} r^{2j} \\
&= \sum_{j=0}^{\infty} \frac{(-1)^j}{2^j j!} (r^2)^j \\
&= e^{-\frac{1}{2}r^2}.
\end{aligned}$$

The first equality is a Maclaurin expansion of cosine, the second equality is true by the Dominated Convergence Theorem. In particular, the series is upper bounded by $\sum_{j=0}^{\infty}((r\eta)^2)^j/(j!) = e^{r^2\eta^2}$. The third equality uses that pth moment of a standard normal random variable is $(p-1)!!$ if p is even. □

A.4. Proof of Proposition 2

Proof. From Proposition 1, we have an empirical mean of 1-dimensional kernels approximating its mean. We can apply concentration inequalities. We choose Bernstein's inequality to explain the effect of the projection variance on convergence; writing $\mathrm{var}(\phi(\boldsymbol{\eta}_1^\top \boldsymbol{\tau})) = v(\boldsymbol{\tau})$ and using that ϕ is bounded by 1,

$$\begin{aligned}
&P\left(\left|\frac{1}{J}\sum_{j=1}^{J} \phi(\boldsymbol{\eta}_j^\top \boldsymbol{\tau}) - k_{exp}(\boldsymbol{\tau})\right| \geq \epsilon\right) \\
\leq &2\exp\left(\frac{-\epsilon^2 J}{2v(\boldsymbol{\tau}) + 4\epsilon/3}\right) \qquad (8)
\end{aligned}$$

Letting the right-hand size equal $\delta > 0$ and solving for ϵ, we have

$$\begin{aligned}
\delta &= 2\exp\left(\frac{-\epsilon^2 J}{2v(\boldsymbol{\tau}) + 4\epsilon/3}\right) \\
\log(2/\delta) &= \frac{\epsilon^2 J}{2v(\boldsymbol{\tau}) + 4\epsilon/3} \\
\left(\frac{1}{J}\log(2/\delta)\right)(2v(\boldsymbol{\tau}) + 4\epsilon/3) &= \epsilon^2 \\
\frac{2v(\boldsymbol{\tau})}{J}\log(2/\delta) + \frac{4}{3J}\log(2/\delta)\epsilon &= \epsilon^2 \\
\epsilon^2 + b\epsilon + c &= 0
\end{aligned}$$

for $b = -4/(3J)\log(1/\delta)$, $c = -2v(\boldsymbol{\tau})/J\log(1/\delta)$. Then,

$$\begin{aligned}
\epsilon &= \frac{-b \pm \sqrt{b^2 - 4c}}{2a} \leq \frac{-b + \sqrt{-c}}{a} \\
&= \frac{4}{3J}\log(2/\delta) + \sqrt{\frac{2v(\boldsymbol{\tau})}{J}\log(2/\delta)} \qquad (9)
\end{aligned}$$

Finally, to derive the uniform convergence bound, applying union bounds to equation 8, we have

$$\begin{aligned}
&P\left(\left|\frac{1}{J}\sum_{j=1}^{J} \phi(\boldsymbol{\eta}_j^\top \boldsymbol{\tau}_{i,j}) - k_{exp}(\boldsymbol{\tau}_{i,j})\right| \geq \epsilon \forall i, j \in [n]\right) \\
\leq &2n^2 \exp\left(\frac{-\epsilon^2 J}{2\sup_{\boldsymbol{\tau}} v(\boldsymbol{\tau}) + 4\epsilon/3}\right).
\end{aligned}$$

We simply replace δ in (9) with δ/n^2 to arrive at the bound

$$\epsilon \leq \frac{4}{3J}\log(2n^2/\delta) + \sqrt{\frac{2v(\boldsymbol{\tau})}{J}\log(2n^2/\delta)}$$

□

B. Supplementary experimental results

We show the regression performance of additional models from Section 5.1 in Table 1. We also show a comparison of negative log likelihood in Table 2.

For each model, and for each data set, we perform 10-fold cross validation twice to accurately measure the performance of stochastic methods. For each fold, we normalize the features and target function by mean and standard deviation as computed on the training folds, so predicting the mean results in ≈ 1 RMSE. For each fold, we fit the kernel hyperparameters by maximizing the log-marginal likelihood. We use Adam (Kingma and Ba, 2014) with learning rate 0.1 for at most 1000 iterations, stopping if log marginal likelihood improves by less than 0.0001 over 20 iterations, smoothing with a moving average. We use a smoothed box prior over the likelihood noise parameter to encourage numerical stability. We compute performance statistics for each fold and report the mean and standard error of the mean.

C. DPA performance as the number of projections varies

Figure 8 illustrates how the test RMSE for DPA-ARD relative to RBF-ARD changes across different data sets as J varies. For small to medium-sized data sets, $J < d$ is typically sufficient to achieve test RMSE within five percent of RBF-ARD.

D. Tests on very high-dimensional data sets

We present the full set of results on very high-dimensional data sets in Figure 7, including the Olivetti face orientation data set and CoEPrA genomics data sets.

dataset	n	d	RBF-ARD	IMQ-ARD	Single-RP	RPA-GP-2	RPA-GP-3
challenger	23	4	$\mathbf{1.04 \pm 0.14}$	$\mathbf{1.00 \pm 0.15}$	$\mathbf{0.97 \pm 0.13}$	$\mathbf{1.16 \pm 0.17}$	$\mathbf{1.24 \pm 0.18}$
fertility	100	9	$\mathbf{1.02 \pm 0.05}$	$\mathbf{0.94 \pm 0.04}$	$\mathbf{0.99 \pm 0.05}$	$\mathbf{0.98 \pm 0.05}$	$\mathbf{1.05 \pm 0.05}$
concreteslump	103	7	$\mathbf{0.11 \pm 0.01}$	$\mathbf{0.10 \pm 0.01}$	0.99 ± 0.07	$\mathbf{0.11 \pm 0.01}$	$\mathbf{0.09 \pm 0.01}$
autos	159	24	$\mathbf{0.36 \pm 0.03}$	$\mathbf{0.34 \pm 0.02}$	0.83 ± 0.03	$\mathbf{0.33 \pm 0.03}$	$\mathbf{0.35 \pm 0.02}$
servo	167	4	$\mathbf{0.31 \pm 0.02}$	$\mathbf{0.31 \pm 0.02}$	0.90 ± 0.02	$\mathbf{0.32 \pm 0.02}$	$\mathbf{0.35 \pm 0.02}$
breastcancer	194	33	$\mathbf{0.98 \pm 0.04}$	$\mathbf{0.90 \pm 0.03}$	0.99 ± 0.02	$\mathbf{0.94 \pm 0.03}$	0.98 ± 0.03
machine	209	7	$\mathbf{0.40 \pm 0.01}$	$\mathbf{0.39 \pm 0.01}$	0.82 ± 0.05	$\mathbf{0.39 \pm 0.01}$	$\mathbf{0.38 \pm 0.02}$
yacht	308	6	$\mathbf{0.08 \pm 0.01}$	0.08 ± 0.01	0.87 ± 0.04	$\mathbf{0.07 \pm 0.01}$	$\mathbf{0.07 \pm 0.01}$
autompg	392	7	$\mathbf{0.34 \pm 0.01}$	$\mathbf{0.34 \pm 0.01}$	0.71 ± 0.04	$\mathbf{0.35 \pm 0.01}$	$\mathbf{0.34 \pm 0.02}$
housing	506	13	$\mathbf{0.31 \pm 0.01}$	$\mathbf{0.29 \pm 0.01}$	0.93 ± 0.04	0.37 ± 0.02	0.36 ± 0.02
forest	517	12	$\mathbf{1.06 \pm 0.04}$	$\mathbf{1.02 \pm 0.04}$	$\mathbf{0.99 \pm 0.04}$	$\mathbf{1.06 \pm 0.04}$	$\mathbf{1.07 \pm 0.04}$
stock	536	11	$\mathbf{0.32 \pm 0.01}$	$\mathbf{0.32 \pm 0.01}$	0.84 ± 0.04	$\mathbf{0.32 \pm 0.01}$	$\mathbf{0.33 \pm 0.01}$
energy	768	8	$\mathbf{0.05 \pm 0.00}$	0.05 ± 0.00	0.84 ± 0.04	0.07 ± 0.01	$\mathbf{0.04 \pm 0.00}$
concrete	1030	8	$\mathbf{0.49 \pm 0.05}$	$\mathbf{0.44 \pm 0.04}$	0.94 ± 0.07	$\mathbf{0.58 \pm 0.08}$	0.53 ± 0.03
airfoil	1503	5	0.23 ± 0.01	$\mathbf{0.20 \pm 0.00}$	0.92 ± 0.02	0.24 ± 0.01	$\mathbf{0.20 \pm 0.00}$
gas	2565	128	$\mathbf{0.11 \pm 0.01}$	$\mathbf{0.10 \pm 0.01}$	0.68 ± 0.05	0.13 ± 0.01	$\mathbf{0.12 \pm 0.01}$

dataset	n	d	RPA-GP-1	DPA-GP	RPA-GP-ARD	DPA-GP-ARD	DPA-GP-ARD-SKI
challenger	23	4	$\mathbf{0.93 \pm 0.16}$	$\mathbf{1.02 \pm 0.16}$	$\mathbf{1.02 \pm 0.14}$	$\mathbf{0.98 \pm 0.14}$	$\mathbf{0.98 \pm 0.14}$
fertility	100	9	1.09 ± 0.06	$\mathbf{1.02 \pm 0.05}$	$\mathbf{1.05 \pm 0.05}$	$\mathbf{0.95 \pm 0.05}$	$\mathbf{0.99 \pm 0.03}$
concreteslump	103	7	$\mathbf{0.10 \pm 0.01}$	0.09 ± 0.01	0.14 ± 0.03	$\mathbf{0.10 \pm 0.01}$	$\mathbf{0.10 \pm 0.01}$
autos	159	24	$\mathbf{0.37 \pm 0.02}$	$\mathbf{0.34 \pm 0.01}$	$\mathbf{0.36 \pm 0.02}$	$\mathbf{0.37 \pm 0.02}$	$\mathbf{0.36 \pm 0.01}$
servo	167	4	$\mathbf{0.35 \pm 0.02}$	0.35 ± 0.02	$\mathbf{0.34 \pm 0.02}$	$\mathbf{0.32 \pm 0.02}$	$\mathbf{0.34 \pm 0.01}$
breastcancer	194	33	1.03 ± 0.03	$\mathbf{0.90 \pm 0.04}$	1.04 ± 0.03	1.13 ± 0.03	1.00 ± 0.02
machine	209	7	$\mathbf{0.41 \pm 0.02}$	$\mathbf{0.39 \pm 0.01}$	$\mathbf{0.40 \pm 0.01}$	$\mathbf{0.41 \pm 0.01}$	$\mathbf{0.40 \pm 0.01}$
yacht	308	6	$\mathbf{0.10 \pm 0.01}$	0.11 ± 0.01	$\mathbf{0.09 \pm 0.01}$	$\mathbf{0.09 \pm 0.02}$	0.10 ± 0.01
autompg	392	7	$\mathbf{0.35 \pm 0.02}$	$\mathbf{0.35 \pm 0.01}$	$\mathbf{0.36 \pm 0.02}$	$\mathbf{0.34 \pm 0.01}$	$\mathbf{0.34 \pm 0.01}$
housing	506	13	0.41 ± 0.02	0.41 ± 0.02	0.38 ± 0.01	0.34 ± 0.01	0.38 ± 0.01
forest	517	12	$\mathbf{1.03 \pm 0.04}$	$\mathbf{1.01 \pm 0.04}$	$\mathbf{1.06 \pm 0.04}$	$\mathbf{1.05 \pm 0.04}$	$\mathbf{1.01 \pm 0.03}$
stock	536	11	$\mathbf{0.32 \pm 0.01}$	$\mathbf{0.32 \pm 0.01}$	$\mathbf{0.32 \pm 0.01}$	$\mathbf{0.32 \pm 0.01}$	$\mathbf{0.32 \pm 0.01}$
energy	768	8	0.18 ± 0.01	0.13 ± 0.01	0.05 ± 0.00	$\mathbf{0.05 \pm 0.00}$	0.06 ± 0.00
concrete	1030	8	0.58 ± 0.06	0.53 ± 0.04	$\mathbf{0.46 \pm 0.04}$	$\mathbf{0.47 \pm 0.03}$	$\mathbf{0.49 \pm 0.03}$
airfoil	1503	5	0.44 ± 0.01	0.44 ± 0.01	0.32 ± 0.01	0.31 ± 0.01	0.32 ± 0.01
gas	2565	128	0.17 ± 0.02	0.16 ± 0.01	0.13 ± 0.01	0.13 ± 0.01	0.22 ± 0.05

Table 1. Average RMSE on UCI regression data sets with SEMs. For each data set, we bold the best model and models whose means are not statistically significantly different according to a 1-sided t-test against the best model. Models are described in Table B. RBF-ARD and IMQ-ARD are generally the best models and perform similarly. A single random projection is handily beaten by all additive random projection methods. Among 1-dimensional random projections, there are slight benefits to using a diverse projected additive (DPA) GP. There is a large benefit to applying length-scales before projection (-ARD). There is little to no performance loss using SKI. The last experiment, DPA-GP-ARD-SKI on gas, contains an outlier instance where the GP was fit to only predict the mean, resulting in RMSE of 1. The mean excluding this instance and the median are both 0.13. As expected, from RPA-GP-2 and RPA-GP-3, we see we get closer to full kernels by for adding more random projections and sub-kernels of higher degrees.

dataset	n	d	RBF	RBF-ARD	IMQ-ARD	GAM-GP	RPA-GP	DPA-GP	DPA-GP-ARD
autos	159	24	0.34 ± 0.04	0.48 ± 0.06	0.40 ± 0.05	0.38 ± 0.06	0.59 ± 0.06	0.62 ± 0.08	0.47 ± 0.06
servo	167	4	0.46 ± 0.05	0.43 ± 0.06	0.43 ± 0.06	0.71 ± 0.06	0.54 ± 0.07	0.52 ± 0.07	0.47 ± 0.05
machine	209	7	0.62 ± 0.04	0.61 ± 0.04	0.58 ± 0.04	0.62 ± 0.04	0.64 ± 0.04	0.62 ± 0.05	0.60 ± 0.04
yacht	308	6	-0.51 ± 0.50	-0.06 ± 0.73	-0.04 ± 0.79	-0.32 ± 0.31	-0.34 ± 0.39	-0.56 ± 0.27	-0.07 ± 0.80
autompg	392	7	0.38 ± 0.05	0.39 ± 0.04	0.40 ± 0.04	0.47 ± 0.05	0.46 ± 0.05	0.43 ± 0.05	0.39 ± 0.04
housing	506	13	0.30 ± 0.05	0.25 ± 0.05	0.22 ± 0.04	0.62 ± 0.06	0.54 ± 0.06	0.53 ± 0.06	0.24 ± 0.05
stock	536	11	0.29 ± 0.03	0.30 ± 0.03	0.31 ± 0.03	0.31 ± 0.03	0.31 ± 0.04	0.32 ± 0.03	0.30 ± 0.03
energy	768	8	-1.67 ± 0.04	-1.69 ± 0.03	-1.76 ± 0.03	-0.82 ± 0.03	-0.31 ± 0.07	-0.67 ± 0.09	-1.70 ± 0.03
concrete	1030	8	0.38 ± 0.07	0.38 ± 0.09	0.34 ± 0.10	0.48 ± 0.07	0.80 ± 0.08	0.80 ± 0.08	0.38 ± 0.09
airfoil	1503	5	0.35 ± 0.02	-0.19 ± 0.02	-0.28 ± 0.02	1.03 ± 0.01	0.61 ± 0.03	0.57 ± 0.02	-0.19 ± 0.02
gas	2565	128	-0.66 ± 0.16	-0.70 ± 0.23	-0.65 ± 0.25	-0.66 ± 0.21	-0.51 ± 0.14	-0.43 ± 0.14	-0.70 ± 0.21
skillcraft	3338	19	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	0.98 ± 0.01	1.02 ± 0.01	1.00 ± 0.01	1.00 ± 0.01
sml	4137	22	-1.94 ± 0.01	-2.55 ± 0.02	-2.68 ± 0.03	-0.76 ± 0.01	-0.34 ± 0.02	-0.31 ± 0.02	-1.86 ± 0.19
pol	15000	24	-0.42 ± 0.00	-1.21 ± 0.01	0.91 ± 0.97	1.90 ± 0.87	0.68 ± 0.01	0.65 ± 0.01	-0.02 ± 0.01
elevators	16599	16	0.44 ± 0.01	0.38 ± 0.01	0.38 ± 0.01	0.52 ± 0.01	0.56 ± 0.01	0.55 ± 0.01	0.42 ± 0.01

Table 2. Average test negative log likelihood on UCI regression data sets with standard error of the mean. We see a similar trend as with RMSE: DPA-GP-ARD is is the best of the random projection methods, with predictive performance comparable with the limiting kernel (the inverse multiquadratic kernel) and the RBF kernel with ARD, even for high-dimensional (e.g. gas) and large data sets (e.g. elevators).

model	sub-kernel	projection method	sub-kernel degrees	pre-scale?
RBF-ARD	RBF	-	$1 \times d$	-
IMQ-ARD	Inverse Multiquadratic	-	$1 \times d$	-
Single-RP	RBF	Gaussian	1×1	No
RPA-GP-2	RBF	Gaussian	$4 \times 1, 4 \times 2, 4 \times 3$	No
RPA-GP-3	RBF	Gaussian	$3 \times 1, 3 \times 2, 3 \times 3, 2 \times 4, 2 \times 5, 1 \times 6$	No
RPA-GP-1	RBF	Gaussian	20×1	No
RPA-GP-SKI	RBF	Gaussian	20×1	No
DPA-GP	RBF	Maximize Eq. (7)	20×1	No
RPA-GP-ARD	RBF	Gaussian	20×1	Yes
DPA-GP-ARD	RBF	Maximize Eq. (7)	20×1	Yes
DPA-GP-ARD-SKI	RBF	Maximize Eq. (7)	20×1	Yes

Table 3. Summary of each evaluated model in Table 1.

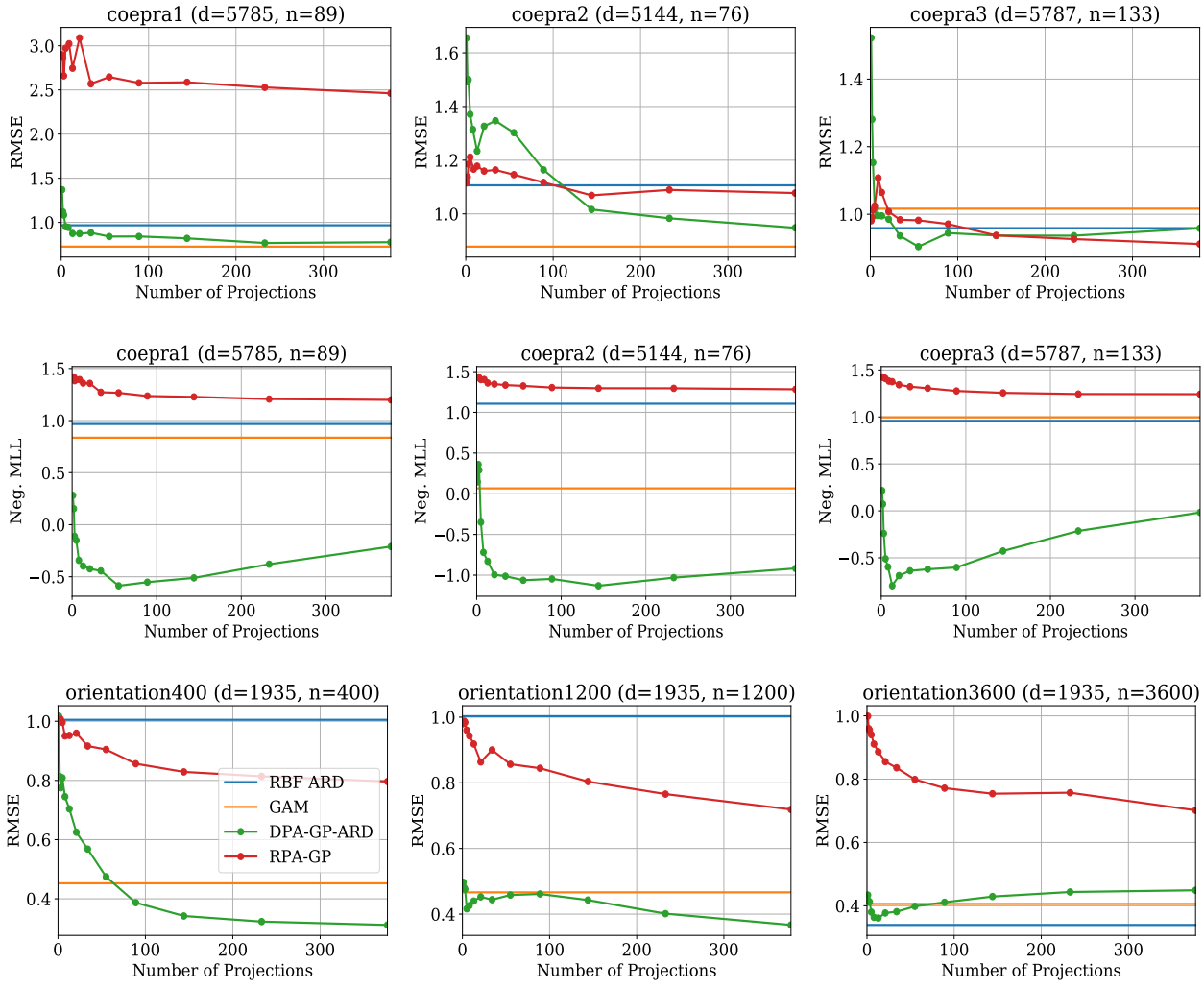

Figure 7. **Top**: RBF-ARD GP, GAM GP, RPA-GP, and DPA-GP-ARD average RMSE on the high dimensional CoEPrA data sets. DPA-GP-ARD is competitive with GAM on these data sets for the optimal number of projections. **Middle**: training negative marginal log likelihood for the same models. For each case, marginal likelihood greatly favors DPA-GP-ARD with few projections. **Bottom**: RBF-ARD GP, GAM GP, RPA-GP, and DPA-GP-ARD average RMSE on Olivetti face orientation data set. RBF-ARD explains all data with noise for $n = 400, 1200$, and DPA-GP-ARD achieves low error with less data than either GAM or RBF-ARD.

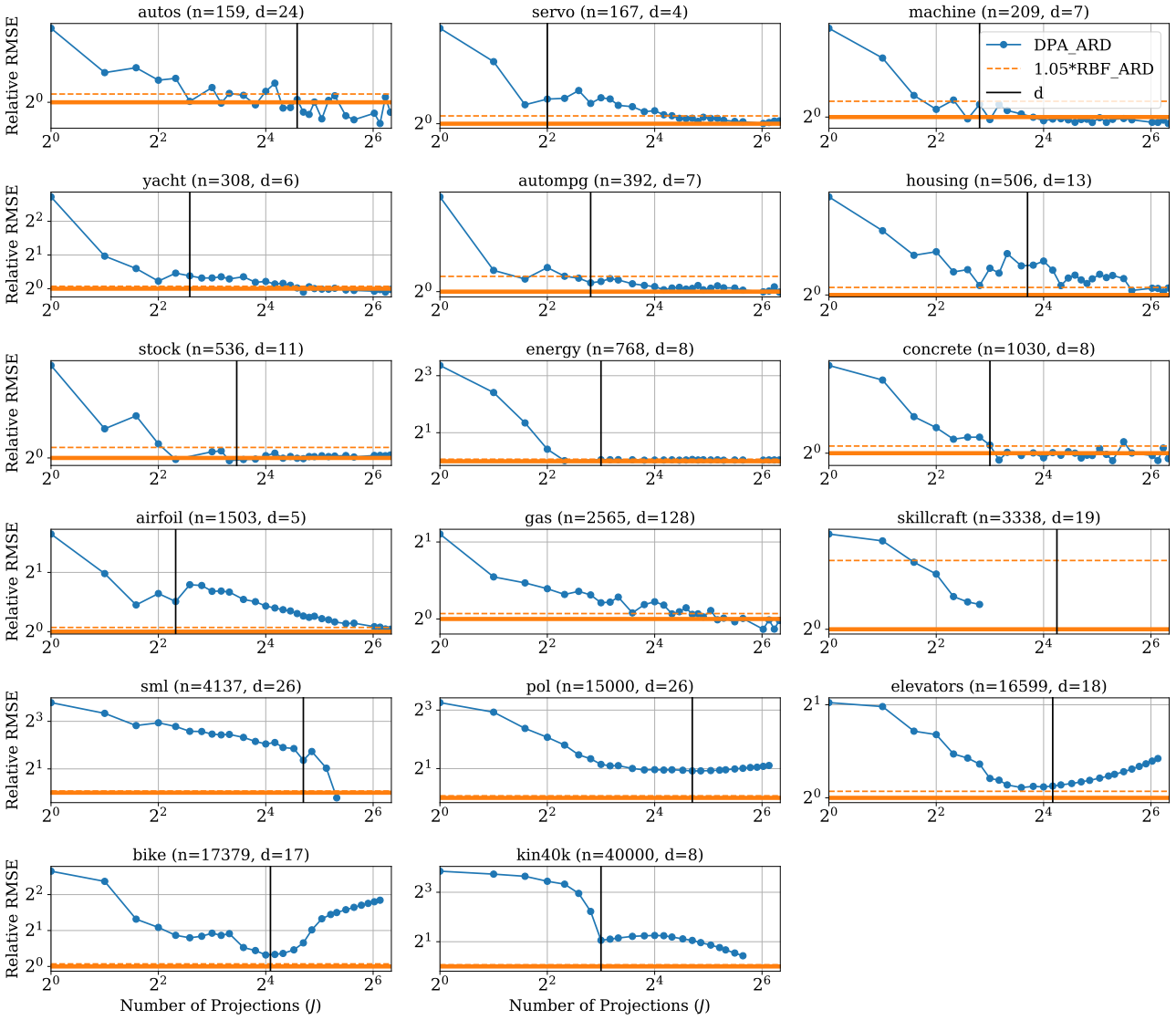

Figure 8. Test RMSE relative to the performance of an RBF-ARD GP as the number of projections varies. Small to medium datasets achieve RMSE within five percent of RBF-ARD with $J < d$ projections. We note that the increasing RMSE for bike and elevators is explainable by a failure of Adam to optimally optimize hyperparameters; the RMSE using L-BFGS-B instead decreases as with higher J expected.