A. Properties of Product-of-Experts

We now discuss the properties of PoEs, and provide formal proofs for the theoretical statements made throughout the paper. We first prove Proposition 2, which states that when weights are normalized, gPoE and rBCM are equivalent. We then prove Proposition 1, which states that in the infinite temperature limit, under our proposed softmax-variance weighting, gPoE, rBCM and the barycenter of GPs are equivalent.

Proof of Proposition 2:

If weights are normalised, i.e. $\sum_j \beta_j(\boldsymbol{x}_*) = 1$ for all $\boldsymbol{x}_*$ then it holds

$$\sigma^{-2}_{rbcm}(\boldsymbol{x}_*) = \sum_{j=1}^{J} \beta_j(\boldsymbol{x}_*)\sigma_j^{-2}(\boldsymbol{x}_*) = \sigma^{-2}_{gpoe}(\boldsymbol{x}_*), \tag{17}$$

and the aggregated means are thus also equal $m_{gpoe}(\boldsymbol{x}_*) = m_{rbcm}(\boldsymbol{x}_*)$ for all $\boldsymbol{x}_*$.

Proof of Proposition 1:

We consider the limit $T \to \infty$, and we consider weighting the experts using $\beta_j(\boldsymbol{x}_*) = \frac{e^{-T\sigma_j^2(\boldsymbol{x}_*)}}{\sum_k e^{-T\sigma_k^2(\boldsymbol{x}_*)}}$ where $\sigma_j^2(\boldsymbol{x}_*)$ is the predictive variance of expert j at $\boldsymbol{x}_*$. We set $\sigma^2_{\min}(\boldsymbol{x}_*) = \min\{\sigma_1^2(\boldsymbol{x}_*), ..., \sigma_J^2(\boldsymbol{x}_*)\}$, and define K as

$$K = |\{k \in \{1, ..., J\} : \sigma_j^2(\boldsymbol{x}_*) = \sigma^2_{\min}(\boldsymbol{x}_*)\}|. \tag{18}$$

First, we consider the case $\sigma_j^2(\boldsymbol{x}_*) = \sigma^2_{\min}(\boldsymbol{x}_*)$. Then

$$\lim_{T\to\infty} \beta_j(\boldsymbol{x}_*) = \lim_{T\to\infty} \frac{1}{\sum_k e^{T(\sigma_j^2(\boldsymbol{x}_*) - \sigma_k^2(\boldsymbol{x}_*))}} = \lim_{T\to\infty} \frac{1}{K e^{T(\sigma_j^2(\boldsymbol{x}_*) - \sigma^2_{\min}(\boldsymbol{x}_*))}} = \frac{1}{K}. \tag{19}$$

Second, we consider the case that $\sigma_j^2(\boldsymbol{x}_*) > \sigma^2_{\min}(\boldsymbol{x}_*)$. Then

$$\lim_{T\to\infty} \beta_j(\boldsymbol{x}_*) = \lim_{T\to\infty} \frac{e^{-T\sigma_j^2(\boldsymbol{x}_*)}}{\sum_l e^{-T\sigma_l^2(\boldsymbol{x}_*)}} = \lim_{T\to\infty} \frac{1}{\sum_l e^{T(\sigma_j^2(\boldsymbol{x}_*) - \sigma_l^2(\boldsymbol{x}_*))}} = 0 \tag{20}$$

Since at least one expert has variance $\sigma^2_{\min}(\boldsymbol{x}_*) < \sigma_j^2(\boldsymbol{x}_*)$, the term in the denominator goes to infinity.

Therefore, only experts that have minimum predictive variance $\sigma_j^2(\boldsymbol{x}_*) = \sigma^2_{\min}(\boldsymbol{x}_*)$ have (uniform) weight $\frac{1}{K}$ where K is the number of experts having minimum variance.

From Proposition 2, we know that under this weighting scheme, the gPoE and the rBCM are equivalent. We now show that the gPoE is equivalent to the barycenter of GPs, which will prove the result. Firstly, we have

$$\sigma^{-2}_{gpoe}(\boldsymbol{x}_*) = \sum_j \beta_j(\boldsymbol{x}_*)\sigma_j^{-2}(\boldsymbol{x}_*) = \frac{1}{K} \sum_{j:\sigma_j^2 = \sigma^2_{\min}(\boldsymbol{x}_*)} \sigma_j^{-2}(\boldsymbol{x}_*) = \sigma^{-2}_{\min}(\boldsymbol{x}_*), \tag{21}$$

$$\sigma^{2}_{bar}(\boldsymbol{x}_*) = \sum_j \beta_j(\boldsymbol{x}_*)\sigma_j^{2}(\boldsymbol{x}_*) = \frac{1}{K} \sum_{j:\sigma_j^2 = \sigma^2_{\min}(\boldsymbol{x}_*)} \sigma_j^{2}(\boldsymbol{x}_*) = \sigma^{2}_{\min}(\boldsymbol{x}_*). \tag{22}$$

Thus, for $T \to \infty$ and with normalised weights, the gPoE, BarGP and the rBCM have the same predictive variance. We now show the equivalence of the predictive means:

$$\mu_{bar}(\boldsymbol{x}_*) = \sum_{j=1}^{J} \beta_j(\boldsymbol{x}_*)m_j(\boldsymbol{x}_*) = \frac{1}{K} \sum_{j:\sigma_j^2(\boldsymbol{x}_*) = \sigma^2_{\min}(\boldsymbol{x}_*)} m_j(\boldsymbol{x}_*) \tag{23}$$

$$\mu_{gpoe}(\boldsymbol{x}_*) = \sigma^2_{\min}(\boldsymbol{x}_*) \sum_{j:\sigma_j^2(\boldsymbol{x}_*) = \sigma^2_{\min}(\boldsymbol{x}_*)} \frac{1}{K}\sigma_j^{-2}(\boldsymbol{x}_*)m_j(\boldsymbol{x}_*) \tag{24}$$

$$= \sigma^2_{\min}\sigma^{-2}_{\min}(\boldsymbol{x}_*)\frac{1}{K} \sum_{j:\sigma_j^2(\boldsymbol{x}_*) = \sigma^2_{\min}(\boldsymbol{x}_*)} m_j(\boldsymbol{x}_*) = \frac{1}{K} \sum_{j:\sigma_j^2(\boldsymbol{x}_*) = \sigma^2_{\min}(\boldsymbol{x}_*)} m_j(\boldsymbol{x}_*), \tag{25}$$

which proves the result. □

Finally, we provide a results that illustrates a failure of rBCM with unnormalised weights, and provides intuition for the erratic behaviours of rBCM in the transitioning region.

Proposition 3. *If* $J = 1$, $0 < \beta_1(\boldsymbol{x}_*) \neq 1$, *and* $\sigma^2_{**} > \sigma^2_1(\boldsymbol{x}_*) > 0$ *the rBCM is not equivalent to the full GP with identical hyperparameters.*

Proof of Proposition 3: If $J = 1$, the predictive precision of the rBCM is of the form

$$\sigma^{-2}_{rbcm}(\boldsymbol{x}_*) = \beta_1(\boldsymbol{x}_*)\sigma_1^{-2}(\boldsymbol{x}_*) + (1 - \beta_1(\boldsymbol{x}_*))\sigma^{-2}_{**} \neq \sigma_1^{-2}(\boldsymbol{x}_*). \tag{26}$$

Therefore, the mean is not identical either as the variance terms do not cancel. □

B. Extra Experimental Details

All experts share the same hyperparameters, which are trained jointly by maximising the marginal likelihood for full GPs (regression) or the ELBO for SVGP experts. The classification SVGP experts use a multiclass likelihood with a robustmax link function, and RBF kernels with ARD. We use L-BFGS-B with a maximum of 100 iterations for full GP, and ADAM for SVGP experts.

We compare against SVGP (Hensman et al., 2013) with 500 inducing points. For classification, we use random partitioning such that each expert is allocated to clusters with 500 training datapoints.

C. Additional Experimental Results

C.1. Random Data Partitioning Results

Table 3 shows the regression results when using random data partitioning.

dataset	N	D	gPoE_Unif	gPoE/rBCM_var	rBCM_entr	BAR_var	linear	SVGP_{500}
Airfoil	1503	8	0.820 (0.540)	0.767 (0.512)	0.795 (0.530)	0.774 (0.515)	1.096 (0.721)	0.409 (0.353)
Concrete	1030	5	0.614 (0.453)	0.558 (0.426)	0.630 (0.447)	0.562 (0.428)	0.953 (0.626)	0.289 (0.338)
Kin40k	40000	8	1.079 (0.717)	0.517 (0.408)	3.947 (0.470)	0.513 (0.405)	1.419 (1.000)	0.124 (0.263)
Parkinsons	5878	20	1.030 (0.669)	0.521 (0.415)	1.245 (0.481)	0.520 (0.415)	1.282 (0.871)	0.554 (0.412)
Power	9568	4	0.032 (0.249)	0.030 (0.248)	0.028 (0.249)	0.032 (0.249)	0.098 (0.267)	-0.044 (0.231)
Protein	45730	9	1.270 (0.857)	1.251 (0.841)	1.252 (0.846)	1.253 (0.843)	1.254 (0.848)	1.083 (0.715)
average			0.808 (0.581)	0.607 (0.475)	1.316 (0.504)	0.609 (0.476)	1.017 (0.722)	0.402 (0.385)

Table 3. Average NLPD (RMSE) values across small (1K+ points) and large-scale (10k+ points) benchmarks when using random data partitioning.

C.2. Robustness to the Temperature Parameter

The performance of the expert models tends to vary depending on the temperature parameter T.Tables 4–6 show similar performance metrics of the variance weighted expert models. A key observation is that the performance of the variance-weighted barGP and gPoE experts stabilise at temperatures above 10. This is desirable behaviour in terms of weight allocation across experts.

C.3. Sensitivity to the Number of Points per Expert

Tables 7 shows the performances of the various expert models across different settings of the the number of initial points assigned per expert under posterior predictive variance weights. The results show that the overall performance tends to improve as the number of points per expert increases while relative performance between models remains constant.

dataset	BAR_var$_{T=1}$	BAR_var$_{T=10}$	BAR_var$_{T=25}$	BAR_var$_{T=50}$	BAR_var$_{T=75}$	BAR_var$_{T=100}$	BAR_var$_{T=150}$
Airfoil	1.219 (0.729)	0.411 (0.350)	0.411 (0.349)	0.411 (0.350)	0.411 (0.350)	0.411 (0.350)	0.411 (0.350)
Concrete	1.019 (0.488)	0.293 (0.344)	0.289 (0.343)	0.289 (0.342)	0.288 (0.342)	0.288 (0.342)	0.288 (0.342)
Kin40k	1.217 (0.807)	0.106 (0.201)	-0.291 (0.158)	-0.365 (0.170)	-0.354 (0.178)	-0.339 (0.183)	-0.319 (0.190)
Parkinsons	1.342 (0.920)	0.208 (0.353)	0.093 (0.333)	0.093 (0.335)	0.097 (0.336)	0.100 (0.337)	0.103 (0.339)
Power	1.017 (0.690)	0.239 (0.291)	0.025 (0.243)	-0.047 (0.229)	-0.068 (0.225)	-0.076 (0.224)	-0.082 (0.222)
Protein	1.410 (0.992)	0.911 (0.676)	0.781 (0.602)	0.776 (0.587)	0.775 (0.583)	0.775 (0.582)	0.775 (0.580)
average	1.204 (0.771)	0.361 (0.369)	0.218 (0.338)	0.193 (0.335)	0.192 (0.336)	0.193 (0.336)	0.196 (0.337)

Table 4. Average NLPD (RMSE) values across benchmark datasets when using the posterior predictive variance expert weights across various settings of the temperature parameter for the Barycenter prediction aggregation method.

dataset	gPoE_var$_{T=1}$	gPoE_var$_{T=10}$	gPoE_var$_{T=25}$	gPoE_var$_{T=50}$	gPoE_var$_{T=75}$	gPoE_var$_{T=100}$	gPoE_var$_{T=150}$
Airfoil	0.510 (0.395)	0.412 (0.350)	0.411 (0.349)	0.411 (0.350)	0.411 (0.350)	0.411 (0.350)	0.411(0.350)
Concrete	0.309 (0.346)	0.290(0.343)	0.289 (0.342)	0.288 (0.342)	0.288 (0.342)	0.288 (0.342)	0.288 (0.342)
Kin40k	0.882 (0.550)	-0.240 (0.161)	-0.364 (0.164)	-0.359 (0.176)	-0.342 (0.182)	-0.329 (0.186)	-0.313 (0.191)
Parkinsons	0.855 (0.599)	0.094 (0.338)	0.088 (0.334)	0.094 (0.335)	0.098 (0.337)	0.101 (0.338)	0.104 (0.339)
Power	0.192 (0.279)	-0.052 (0.226)	-0.076 (0.223)	-0.082 (0.222)	-0.083 (0.222)	-0.084 (0.222)	-0.084 (0.222)
Protein	1.344 (0.932)	0.821 (0.645)	0.782 (0.601)	0.776 (0.587)	0.775 (0.583)	0.775 (0.582)	0.775 (0.580)
average	0.682 (0.517)	0.221 (0.344)	0.188 (0.336)	0.188 (0.335)	0.191 (0.336)	0.194 (0.336)	0.197 (0.337)

Table 5. Average NLPD (RMSE) values across benchmark datasets when using the posterior predictive variance expert weights across various settings of the temperature parameter for the gPoE prediction aggregation method.

dataset	rBCM_var$_{T=1}$	rBCM_var$_{T=10}$	rBCM_var$_{T=25}$	rBCM_var$_{T=50}$	rBCM_var$_{T=75}$	rBCM_var$_{T=100}$	rBCM_var$_{T=150}$
Airfoil	0.412 (0.354)	0.497 (0.451)	0.583 (0.544)	0.674 (0.614)	0.754 (0.663)	0.833 (0.702)	0.985 (0.768)
Concrete	0.287 (0.344)	0.352 (0.405)	0.431 (0.470)	0.547 (0.522)	0.665 (0.559)	0.775 (0.602)	0.947 (0.675)
Kin40k	2.397 (0.290)	-0.539 (0.153)	-0.271 (0.207)	0.124 (0.372)	0.423 (0.521)	0.640 (0.629)	0.916 (0.761)
Parkinsons	0.302 (0.350)	0.209 (0.418)	0.410 (0.528)	0.653 (0.651)	0.818 (0.731)	0.936 (0.785)	1.093 (0.856)
Power	-0.064 (0.226)	-0.076 (0.223)	-0.080 (0.222)	-0.079 (0.222)	-0.074 (0.224)	-0.067 (0.227)	-0.049 (0.239)
Protein	0.784 (0.564)	0.890 (0.696)	1.001 (0.769)	1.106 (0.828)	1.168 (0.860)	1.209 (0.881)	1.263 (0.909)
average	0.686 (0.355)	0.222 (0.391)	0.346 (0.457)	0.504 (0.535)	0.626 (0.593)	0.721 (0.638)	0.859 (0.701)

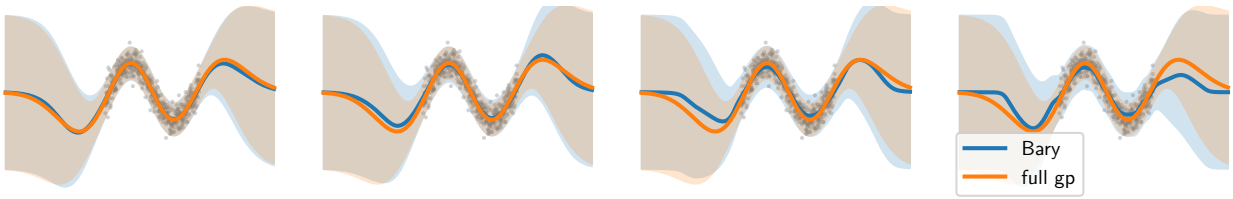
Table 6. Average NLPD (RMSE) values across benchmark datasets when using the posterior predictive variance expert weights across various settings of the temperature parameter for the rBCM prediction aggregation method with unnormalised weights.

Dataset	BAR_var_100p	BAR_var_200p	BAR_var_500p	gPoE_var_100p	gPoE_var_200	gPoE_var_500p	rBCM_var_100p	rBCM_var_200p	rBCM_var_500p
Airfoil	0.411 (0.350)	0.401 (0.344)	0.363 (0.332)	0.411 (0.350)	0.401 (0.344)	0.363 (0.332)	0.674 (0.614)	0.641 (0.598)	0.536 (0.515)
Concrete	0.289 (0.342)	0.277 (0.345)	0.261 (0.330)	0.288 (0.342)	0.277 (0.345)	0.261 (0.330)	0.547 (0.522)	0.503 (0.500)	0.381 (0.405)
Kin40k	-0.365 (0.170)	-0.545 (0.140)	-0.720 (0.117)	-0.359 (0.176)	-0.557 (0.145)	-0.765 (0.120)	0.124 (0.372)	-0.322 (0.233)	-0.679 (0.164)
Parkinsons	0.093 (0.335)	0.101 (0.331)	0.078 (0.333)	0.094 (0.335)	0.103 (0.332)	0.079 (0.333)	0.653 (0.651)	0.673 (0.657)	0.686 (0.679)
Power	-0.047 (0.229)	-0.039 (0.231)	-0.044 (0.231)	-0.082 (0.222)	-0.072 (0.225)	-0.066 (0.226)	-0.079 (0.222)	-0.070 (0.225)	-0.064 (0.226)
Protein	0.776 (0.587)	0.805 (0.705)	0.814 (0.673)	0.776 (0.587)	0.805 (0.704)	0.814 (0.673)	1.106 (0.828)	1.085 (0.891)	1.113 (0.888)
average	0.193 (0.335)	0.167 (0.349)	0.125 (0.336)	0.188 (0.335)	0.159 (0.349)	0.114 (0.336)	0.504 (0.535)	0.418 (0.517)	0.329 (0.480)

Table 7. Average NLPD (RMSE) values across benchmark datasets when using the posterior predictive variance expert weights across different settings for the number of initial points per expert.

Concrete	Airfoil	Power	Kin40k	Protein
0.352 (0.360)	0.506 (0.396)	-0.007 (0.238)	-0.320 (0.153)	0.838 (0.741)

Table 8. NLPDs/RMSEs of the grBCM with y-averaging.



(a) soft-var, 100 pts/exp (b) soft-var, 20 pts/exp (c) soft-var, 4 pts/exp (d) soft-var, 2 pts/exp

Figure 7. Full GP baseline (orange) and barycenter of GPs model (blue) trained on synthetic data with a decreasing number of points per experts (Left to Right), using softmax-variance weighting.