A Choice of $r(\cdot,\cdot)$ to ensure $\mu_{\mathsf{P}} \in \mathcal{H}_k$

We need to choose an appropriate covariance function r, such that $\mu_{\mathsf{P}} \in \mathcal{H}_k$, where $\mu_{\mathsf{P}} \sim \mathcal{GP}(0, r(\cdot,\cdot))$. In particular, it is for infinite-dimensional RKHSs not sufficient to define $r(\cdot,\cdot) = k(\cdot,\cdot)$, as draws from this particular prior are no longer in $\mathcal{H}_k$ (Wahba, 1990) (but see below). However, we can construct

$$r(x,y) = \int k(x,z)k(z,y)\nu(dz) \tag{8}$$

where ν is any finite measure on $\mathcal{X}$. This then ensures $\mu_{\mathsf{P}} \in \mathcal{H}_k$ with probability 1 by the nuclear dominance (Lukić and Beder, 2001; Pillai et al., 2007) for any stationary kernel k. In particular, Flaxman et al. (2016) provides details when k is a squared exponential kernel defined by

$$k(x,y) = \exp(-\frac{1}{2}(x-y)^\top \Sigma_k^{-1}(x-y)) \qquad x, y \in \mathbb{R}^p$$

and $\nu(dz) = \exp\left(-\frac{||z||_2^2}{2\ell^2}\right)dz$, i.e. it is proportional to a Gaussian measure on $\mathbb{R}^d$, which provides $r(\cdot,\cdot)$ with a non-stationary component. In this paper, we take $\Sigma_k = \sigma^2 I_p$, where σ^2 and ℓ are tuning parameters, or parameters that we learn.

Here, the above holds for a general set of stationary kernels, but note that by taking a convolution of a kernel with itself, it might make the space of functions that we consider overly smooth (i.e. concentrated on a small part of $\mathcal{H}_k$). In this work, however, we consider only the Gaussian RBF kernel k. In fact, recent work (Steinwart, 2017, Theorem 4.2) actually shows that in this case, the sample paths almost surely belong to (interpolation) spaces which are infinitesimally larger than the RKHS of the Gaussian RBF kernel. This suggests that we can choose r to be an RBF kernel with a length scale that is infinitesimally bigger than that of k; thus, in practice, taking $r = k$ would suffice and we do observe that it actually performs better (Fig. 4).

B Framework for Binary Classification

Suppose that our labels $y_i \in \{0, 1\}$, i.e. we are in a binary classification framework. Then a simple approach to accounting for uncertainty in the regression parameters is to use bayesian logistic regression, putting priors on β, i.e.

$$\begin{aligned} \beta &\sim \mathcal{N}(0, \rho^2) \\ y_i &\sim Ber(\pi_i), \text{ where } \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta^\top \hat{\mu}_i \end{aligned}$$

however for the mean shrinkage pooling model, if we use the above $y_i \mid \mu_i, \alpha$, we would not be able to obtain an analytical solution for $p(y_i|\mathbf{x_i}, \alpha)$. Instead we use the probit link function, as given by:

$$Pr(y_i = 1|\mu_i, \alpha) = \Phi\left(\alpha^\top \mu_i(\mathbf{z})\right)$$

where Φ denotes the Cumulative Distribution Function (CDF) of a standard normal distribution, with $\mu_i(\mathbf{z}) = [\mu_i(z_1), \ldots, \mu_i(z_s)]^\top$. Then as before we have

$$\mu_i(\mathbf{z}) \mid \mathbf{x_i} \sim \mathcal{N}(M_i, C_i)$$

with M_i and C_i as defined in section 3.3. Hence, as before

$$\begin{aligned} Pr(y_i = 1|\mathbf{x}_i, \alpha) &= \int Pr(y_i = 1|\mu_i, \alpha)p(\mu_i(\mathbf{z})|\mathbf{x}_i)d\mu_i(\mathbf{z}) \\ &= c\int \Phi(\alpha^\top \mu_i(\mathbf{z}))\exp\{-\frac{1}{2}(\mu_i(\mathbf{z}) - M_i)^\top C_i^{-1}(\mu_i(\mathbf{z}) - M_i)\}d\mu_i(\mathbf{z}) \\ (\text{with } l_i = \mu_i(\mathbf{z}) - M_i) &= c\int \Phi(\alpha^\top (l_i + M_i))\exp\{-\frac{1}{2}(l_i)^\top C_i^{-1}(l_i)\}dl_i \\ &= Pr(Y \le \alpha^\top (l_i + M_i)) \end{aligned}$$

Note here $Y \sim \mathcal{N}(0,1)$ and $l_i \sim \mathcal{N}(0, \Sigma_i)$ Then expanding and rearranging

$$Pr(y_i = 1|\mathbf{x}_i, \alpha) \quad = \quad Pr(Y - \alpha^\top l_i \leq \alpha^\top M_i)$$

Note that since Y and l_i independent normal r.v., $Y - \alpha^\top l_i \sim \mathcal{N}(0, 1 + \alpha^\top C_i \alpha^\top)$. Let T be standard normal, then we have:

$$\begin{aligned} Pr(y_i = 1|\mathbf{x}_i, \alpha) &= Pr(\sqrt{1 + \alpha^\top C_i \alpha}\, T \leq \alpha^\top M_i) \\ &= Pr(T \leq \frac{\alpha^\top M_i}{\sqrt{1 + \alpha^\top C_i \alpha}}) \\ &= \Phi\left(\frac{\alpha^\top M_i}{\sqrt{1 + \alpha^\top C_i \alpha}}\right) \end{aligned}$$

Hence, we also have:

$$Pr(y_i = 0|\mathbf{x}_i, \alpha) \quad = \quad 1 - \Phi\left(\frac{\alpha^\top M_i}{\sqrt{1 + \alpha^\top C_i \alpha}}\right)$$

Now placing the prior $\alpha \sim \mathcal{N}(0, \rho^2 K_{\mathbf{z}}^{-1})$, we have the following MAP objective:

$$\begin{aligned} J(\alpha) &= \log\left[p(\alpha)\prod_{i=1}^{n} p(y_i|\mathbf{x}_i, \alpha)\right] \\ &= \sum_{i=1}^{n}(1 - y_i)\log(1 - \Phi\left(\frac{\alpha^\top M_i}{\sqrt{1 + \alpha^\top C_i \alpha}}\right)) \\ &\quad + y_i \log(\Phi\left(\frac{\alpha^\top M_i}{\sqrt{1 + \alpha^\top C_i \alpha}}\right)) + \frac{1}{\rho^2}\alpha^\top K_{\mathbf{z}} \alpha \end{aligned}$$

Since we have an analytical solution for $Pr(y_i = 0|\mathbf{x}_i, \alpha)$, we can also use this in HMC for BDR.

C Some more intuition on the shrinkage estimator

In this section, we provide some intuition behind the shrinkage estimator in section 3.3. Here, for simplicity, we choose $\Sigma_i = \tau^2 I$ for all bag i, and $m_0 = 0$, and consider the case where $\mathbf{z} = \mathbf{u}$, i.e. $R = R_{\mathbf{z}} = R_{\mathbf{zz}}$. We can then see that if R has eigendecomposition $U\Lambda U^T$, with $\Lambda = \mathrm{diag}(\lambda_k)$, the posterior mean is

$$U\,\mathrm{diag}\left(\frac{\lambda_k}{\lambda_k + \tau^2/N_i}\right)U^T(\hat{\mu}_i),$$

so that large eigenvalues, $\lambda_k \gg \tau^2/N_i$, are essentially unchanged, while small eigenvalues, $\lambda_k \ll \tau^2/N_i$, are shrunk towards 0. Likewise, the posterior variance is

$$U\,\mathrm{diag}\left(\lambda_k - \frac{\lambda_k^2}{\lambda_k + \frac{\tau^2}{N_i}}\right)U^T = U\,\mathrm{diag}\left(\frac{1}{\frac{N_i}{\tau^2} + \frac{1}{\lambda_k}}\right)U^T;$$

its eigenvalues also decrease as N_i/τ^2 increases.

D Alternative Motivation for choice of f

Here we provide an alternative motivation for the choice of $f = \sum_{s=1}^{k} \alpha_s k(\cdot, z_s)$. First, consider the following Bayesian model with a linear kernel K on μ_i, where $f : \mathcal{H}_k \to \mathbb{R}$:

$$y_i \mid \mu_i, f \quad \sim \quad \mathcal{N}\left(f(\mu_i), \sigma^2\right).$$

Now considering the log-likelihood of $\{\mu, Y\} = \{\mu_i, y_i\}_{i=1}^n$ (supposing we have these exact embeddings), we obtain:

$$\log p(Y|\mu, f) = \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(y_i - f(\mu_i))^2$$

To avoid over-fitting, we place a Gaussian prior on f, i.e. $-\log p(f) = \lambda||f||_{\mathcal{H}_k} + c$. Minimizing the negative log-likelihood over $f \in \mathcal{H}_k$, we have:

$$f^* = \text{argmin}_{f\in\mathcal{H}_k} \sum_{i=1}^{n} \frac{1}{2\sigma^2}(y_i - f(\mu_i))^2 + \lambda||f||_{\mathcal{H}_k}$$

Now this is in the form of an empirical risk minimisation problem. Hence using the representer theorem (Schölkopf et al., 2001), we have that:

$$f = \sum_{j=1}^{n} \gamma_j K(., \mu_j)$$

i.e. we have a finite-dimensional problem to solve. Thus since K is a linear kernel:

$$y_i \mid \mu_i, \{\mu_j\}_{j=1}^n, \gamma \quad \sim \quad \mathcal{N}\left(\sum_{j=1}^{n} \gamma_j \langle \mu_i, \mu_j \rangle_{\mathcal{H}_k}, \sigma^2\right).$$

where $\langle \mu_i, \mu_j \rangle_{\mathcal{H}_k}$ can be thought of as the similarity between distributions.

Now we have the same $\mathcal{GP}$ posterior as in Section 3.3, and we would like to compute $p(y_i|\mathbf{x_i}, \gamma)$. This suggests we need to integrate out $\mu_1, \dots \mu_n$. But it is unclear how to perform this integration, since the μ_i follow Gaussian process distributions. Hence we can take an approximation to f, i.e. $f = \sum_{s=1}^{k} \alpha_s k(\cdot, z_s)$, which would essentially give us a dual method with a sparse approximation to f.