7 APPENDIX: derivations

7.1 Finite parametrization of GP

We describe here in more details how to get the equivalent finite dimensional parametrization of GP for regression (used in Sec. 4.1). We recall that $f_\mathcal{D} = (f(x_1), \ldots, f(x_N))^\top$, and let f_{rest} be the values of f on the complement of $\mathcal{D}$. Because of our conditional independence assumptions, we have that the posterior factorizes: $p(f|\mathcal{D}) = p(f_{\text{rest}}|f_\mathcal{D})p(f_\mathcal{D}|\mathcal{D})$. By using the linearity of expectations and interchanging the order of integration, the posterior risk thus becomes:

$$\mathcal{R}_{p_\mathcal{D}}(h) = \int_{\mathbb{R}^N} p(f_\mathcal{D}|\mathcal{D}) \left(\int_{\mathcal{X},\mathcal{Y}} p(x)\tilde{p}(y|x, f_\mathcal{D})\ell(y, h(x))dydx \right) df_\mathcal{D}, \tag{30}$$

where we have defined:

$$\begin{aligned} \tilde{p}(y|x, f_\mathcal{D}) &\doteq \int p(y|x, f)p(f_{\text{rest}}|f_\mathcal{D})df_{\text{rest}} \\ &= \mathcal{N}\left(y|K_{x\mathcal{D}}K_{\mathcal{D}\mathcal{D}}^{-1}f_\mathcal{D}, \sigma_x^2\right). \end{aligned} \tag{31}$$

The Gaussian expression in (31) is from standard properties of GP (basically coming from conditional independence and the conditioning formula for multivariate normals); by doing the change of variable $\theta = K_{\mathcal{D}\mathcal{D}}^{-1}f_\mathcal{D}$, we get the expressions that we gave in (18). We can then use the loss $L(\theta, h)$ defined in terms of $p(y|x, \theta)$ instead of $L(f, h)$ defined in term of $p(y|x, f)$ and do an equivalent analysis.

7.2 GP regression equations

The posterior $p_\mathcal{D}$ is a Gaussian with mean $\mu_{p_\mathcal{D}} = (K_{\mathcal{D}\mathcal{D}} + \sigma^2 I)^{-1}\mathbf{y}$ and covariance $\Sigma_{p_\mathcal{D}} = K_{\mathcal{D}\mathcal{D}}^{-1} - (K_{\mathcal{D}\mathcal{D}} + \sigma^2 I)^{-1}$ (recall that we did the change of variable $\theta = K_{\mathcal{D}\mathcal{D}}^{-1}f_\mathcal{D}$) where $\mathbf{y}$ is the vector of outputs $(y_1, \ldots, y_N)^\top$. By using the block matrix inversion lemma, we can get that $\Sigma_{p_\mathcal{D}}^{-1} = K_{\mathcal{D}\mathcal{D}} + \sigma^{-2}K_{\mathcal{D}\mathcal{D}}^2$ and so is different from Λ from (21). Even if we use the empirical distribution on $\mathcal{D}$ as the test distribution $p(x)$, then we get $\Lambda = K_{\mathcal{D}\mathcal{D}}^2/N$, which is still missing an additive $K_{\mathcal{D}\mathcal{D}}$ to become proportional to $\Sigma_{p_\mathcal{D}}^{-1}$.

We now derive the μ_q which minimizes the KL expression given in (22) subject to the sparsity constraint. We partition the set of indices of the dataset into a fixed set S of size k for the non-zero coefficient of μ_q and T for the set of coefficients that we constraint to zero. Writing $\tilde{\Lambda} \doteq \Sigma_{p_\mathcal{D}}^{-1}$ and setting the derivative to zero, we get that the non-zero components of μ_q (on the set S) are given by:

$$\mu_{q_{\text{sp}}^{\text{KL}}} = \tilde{\Lambda}_{SS}^{-1}\tilde{\Lambda}_{S\mathcal{D}}\mu_{p_\mathcal{D}}. \tag{32}$$

Substituting $\Sigma_{p_\mathcal{D}}^{-1} = K_{\mathcal{D}\mathcal{D}} + \sigma^{-2}K_{\mathcal{D}\mathcal{D}}^2$ and $\mu_{p_\mathcal{D}} = (K_{\mathcal{D}\mathcal{D}} + \sigma^2 I)^{-1}\mathbf{y}$, we have that:

$$\begin{aligned} \tilde{\Lambda}_{S\mathcal{D}}\mu_{p_\mathcal{D}} &= K_{S\mathcal{D}}(\mathcal{I} + \sigma^{-2}K_{\mathcal{D}\mathcal{D}})(K_{\mathcal{D}\mathcal{D}} + \sigma^2 I)^{-1}\mathbf{y} \\ &= \sigma^{-2}K_{S\mathcal{D}}\mathbf{y}, \end{aligned} \tag{33}$$

which is the convenient cancellation that enables us to avoid the inversion of the $N \times N$ matrix $K_{\mathcal{D}\mathcal{D}}$ which was previously needed to compute $\mu_{p_\mathcal{D}}$. Substituting (33) into (32) and expanding $\tilde{\Lambda}_{SS}$, we get

$$\mu_{q_{\text{sp}}^{\text{KL}}} = \left(\sigma^2 K_{SS} + K_{S\mathcal{D}}K_{\mathcal{D}S}\right)^{-1} K_{S\mathcal{D}}\,\mathbf{y}, \tag{34}$$

which only requires the inversion of a $k \times k$ matrix and so is computable in $O(k^3 + Nk^2)$ time.

On the other hand, the minimizer of d_L in (20) with sparse constraints is $\mu_{q_{\text{sp}}^{\text{opt}}} = \Lambda_{SS}^{-1}\Lambda_{S\mathcal{D}}\mu_{p_\mathcal{D}}$ which does not yield similar cancellations and so does not seem efficiently computable. It is clear in this case though that $\mu_{q_{\text{sp}}^{\text{opt}}} \neq \mu_{q_{\text{sp}}^{\text{KL}}}$ (unless $S = \mathcal{D}$) and so it leaves open how to obtain efficiently an approximate sparse solution with lower Bayesian risk.

7.3 Derivation of h_q for GPC

We provide a derivation here for (26). The q-conditional-risk, which we want to minimize pointwise, takes in this case the form:

$$\begin{aligned} \mathcal{R}_q(y'|x) = \; & \mathbb{I}_{\{y'=+1\}}c_+\Phi\left(\frac{-K_{x\mathcal{D}}\mu_q}{\sigma_q(x)}\right) \\ & + \mathbb{I}_{\{y'=-1\}}c_-\Phi\left(\frac{K_{x\mathcal{D}}\mu_q}{\sigma_q(x)}\right). \end{aligned} \tag{35}$$

So to minimize it pointwise, we want to choose $y' = +1$ when:

$$c_+\Phi\left(\frac{-K_{x\mathcal{D}}\mu_q}{\sigma_q(x)}\right) < c_-\Phi\left(\frac{K_{x\mathcal{D}}\mu_q}{\sigma_q(x)}\right).$$

Using the fact that $\Phi(-a) = 1 - \Phi(a)$ and rearranging the terms give the choice function (26).