A $\mathcal{GP}$-NC FOR SCALABLE $\mathcal{GP}$ METHODS

We can replace the NLL term in Algorithm (1) by the log likelihood of the different scalable $\mathcal{GP}$ methods. We have a scalable implementation of the D_{KL} update, so the entire Algorithm scales well with the input data size. It is straightforward to plug-in the class of scalable and Sparse $\mathcal{GP}$ regression models in the likelihood term of Algorithm (1) to account for the negative datapairs in their formulation. In particular we review the SVGP model by (Hensman et al., 2013), which is a popular scalable implementation of $\mathcal{GP}$s. We also investigate a recent parametric Gaussian Process regressors (PPGPR) method by (Jankowiak et al., 2019). In this section, we follow the notations given in their respective research works and give their derivations of the log likelihood function here for the sake of completeness.

A.1 SVGP REGRESSION MODEL

(Hensman et al., 2013) proposed the Scalable Variational GP (SVGP) method. The key technical innovation was the development of inducing point methods which we now review. By introducing inducing variables $\mathbf{u}$ that depend on variational parameters $\{\mathbf{z}_m\}_{m=1}^M$, where $M = \dim(\mathbf{u}) \ll N$ and with each $\mathbf{z}_m \in \mathbb{R}^d$, we augment the GP prior as follows:

$$p(\mathbf{f}|X) \rightarrow p(\mathbf{f}|\mathbf{u}, X, Z)p(\mathbf{u}|Z)$$

We then appeal to Jensen's inequality and lower bound the log joint density over the targets and inducing variables:

$$\begin{aligned}\log p(\mathbf{y}, \mathbf{u}|X, Z) &= \log \int d\mathbf{f} p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u}) \\ &\geq \mathbb{E}_{p(\mathbf{f}|\mathbf{u})}\left[\log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{u})\right] \\ &= \sum_{i=1}^{N} \log \mathcal{N}(y_i|\mathbf{k}_i^T K_{MM}^{-1}\mathbf{u}, \sigma_{\mathrm{obs}}^2) - \tfrac{1}{2\sigma_{\mathrm{obs}}^2}\mathrm{Tr}\, Kt_{NN} + \log p(\mathbf{u})\end{aligned} \tag{8}$$

where $\mathbf{k}_i = k(\mathbf{x}_i, Z)$, $K_{MM} = k(Z, Z)$ and Kt_{NN} is given by

$$Kt_{NN} = K_{NN} - K_{NM}K_{MM}^{-1}K_{MN} \tag{9}$$

with $K_{NM} = K_{MN}^{\mathrm{T}} = k(X, Z)$. The essential characteristics of Eqn. 8 are that: i) it replaces expensive computations involving K_{NN} with cheaper computations like K_{MM}^{-1} that scale as $\mathcal{O}(M^3)$; and ii) it is amenable to data subsampling, since the log likelihood and trace terms factorize as sums over datapoints $(y_i, \mathbf{x}_i)$.

A.1.1 SVGP LIKELIHOOD FUNCTION

SVGP proceeds by introducing a multivariate Normal variational distribution $q(\mathbf{u}) = \mathcal{N}(\boldsymbol{m}, S)$. The parameters $\boldsymbol{m}$ and S are optimized using the ELBO (evidence lower bound), which is the expectation of Eqn. 8 w.r.t. $q(\mathbf{u})$ plus an entropy term term $H[q(\mathbf{u})]$:

$$\begin{aligned}\mathcal{L}_{\mathrm{svgp}} &= \mathbb{E}_{q(\mathbf{u})}\left[\log p(\mathbf{y}, \mathbf{u}|X, Z)\right] + H[q(\mathbf{u})] \\ &= \sum_{i=1}^{N}\left\{\log \mathcal{N}(y_i|\mu_{\mathbf{f}}(\mathbf{x}_i), \sigma_{\mathrm{obs}}^2) - \frac{\sigma_{\mathbf{f}}(\mathbf{x}_i)^2}{2\sigma_{\mathrm{obs}}^2}\right\} - D_{\mathrm{KL}}(q(\mathbf{u})|p(\mathbf{u}))\end{aligned} \tag{10}$$

where KL denotes the Kullback-Leibler divergence, $\mu_{\mathbf{f}}(\mathbf{x}_i)$ is the predictive mean function given by $\mu_{\mathbf{f}}(\mathbf{x}_i) = \mathbf{k}_i^T K_{MM}^{-1}\boldsymbol{m}$ and $\sigma_{\mathbf{f}}(\mathbf{x}_i)^2 \equiv \mathrm{Var}[f_i|\mathbf{x}_i] = Kt_{ii} + \mathbf{k}_i^T K_{MM}^{-1} S K_{MM}^{-1}\mathbf{k}_i$ denotes the latent function variance.

$\mathcal{L}_{\mathrm{svgp}}$, which depends on $\boldsymbol{m}, S, Z, \sigma_{\mathrm{obs}}$ and the various kernel hyperparameters θ, can then be maximized with gradient methods. We refer to the resulting GP regression method as SVGP.

A.2 PPGPR-NC REGRESSION MODEL: LIKELIHOOD FUNCTION

Jankowiak et al. (2019) recently proposed a parametric Gaussian Process regressors (PPGPR) method. We defer the reader to Section (3.2) of their paper for details about their likelihood function.