Accurate and Scalable Stochastic Gaussian Process Regression via Learnable Coreset-based Variational Inference

Mert Ketenci[1] **Adler Perotte**[2] **Noémie Elhadad**[1,2] **Iñigo Urteaga**[3,4]

[1]Department of Computer Science, Columbia University , New York, USA
[2]Department of Biomedical Informatics, Columbia University , New York, USA
[3]BCAM — Basque Center for Applied Mathematics, Bilbao, Spain
[4]IKERBASQUE — Basque Foundation for Science, Bilbao, Spain

Abstract

We introduce a novel stochastic variational inference method for Gaussian process ($\mathcal{GP}$) regression, by deriving a $\mathcal{GP}$ posterior over a learnable set of coresets: i.e., over pseudo-input/output, weighted pairs. Unlike former free-form variational families for stochastic inference, our coreset-based variational $\mathcal{GP}$ (CVGP)[a] is defined in terms of the $\mathcal{GP}$ prior and the (weighted) data likelihood. This formulation naturally incorporates inductive biases of the prior, and ensures its kernel and likelihood dependencies are shared with the posterior. We derive a variational lower-bound on the log-marginal likelihood by marginalizing over the latent $\mathcal{GP}$ coreset variables, and show that CVGP's lower-bound is amenable to stochastic optimization. CVGP reduces the dimensionality of the variational parameter search space to linear $\mathcal{O}\left(M\right)$ complexity, while ensuring numerical stability at $\mathcal{O}\left(M^3\right)$ time complexity and $\mathcal{O}\left(M^2\right)$ space complexity. Evaluations on real-world and simulated regression problems demonstrate that CVGP achieves superior inference and predictive performance than state-of-the-art, stochastic sparse $\mathcal{GP}$ approximation methods.

[a]Code is publicly available at https://github.com/iurteagalab/cvgp_regression.

1 INTRODUCTION

Training $\mathcal{GP}$s efficiently with large datasets has been a long-standing challenge, as exact inference complexities grow $O(N^3)$ in time and $O(N^2)$ in space requirements. Successful state-of-the-art (SOTA) methods to scale $\mathcal{GP}$s —a detailed review can be found in [Liu et al., 2020]— are based on sparse and low-rank approximations [Williams and Seeger, 2000, Snelson and Ghahramani, 2005, Quinonero-Candela and Rasmussen, 2005], often using inducing random variables [Naish-Guzman and Holden, 2007, Titsias, 2009, Hensman et al., 2013, Wilson and Nickisch, 2015].

Amongst these techniques, variational learning of inducing variables by Titsias [2009] allows for time and space complexities of $O(NM^2)$ and $O(NM)$, with clear benefits when inducing point size M is small, i.e., $M \leq N$. However, in real-world applications, N can be in the order of millions, making model learning impractical. More recently, Hensman et al. [2013] introduced stochastic variational inference for Gaussian processes (SVGP), which reduces the time and space complexities to $\mathcal{O}(M^3)$ and $\mathcal{O}(M^2)$, respectively. This method has become the standard for training $\mathcal{GP}$ models on large datasets. However, SVGP's scalability comes at a cost: it requires learning additional $\mathcal{O}(M^2)$ parameters, resulting in an optimization problem that scales quadratically with the number of inducing points.

In this work, we propose a coreset-based variational $\mathcal{GP}$ (CVGP) technique that is amenable to stochastic optimization (i.e., scalable to big datasets) at reduced $\mathcal{O}\left(M\right)$ parameter complexity (see Table 1), and demonstrate its accurate inference and predictive performance in a wide range of real-datasets (see results in Section 4).

We take inspiration from Titsias [2009]'s optimal variational posterior, and ensure that CVGP's variational family also obeys (1) the $\mathcal{GP}$s' prior-conditional structure, and (2) the $\mathcal{GP}$ prior's dependencies in its posterior, all achieved via Bayesian coreset principles [Huggins et al., 2016, Zhang et al., 2021]. Specifically, we design and learn a variational distribution for a $\mathcal{GP}$-based probabilistic model, defined through a subset of learnable pseudo-points and a weighted likelihood function, in line with the Black-Box Bayesian coreset framework [Manousakas et al., 2020, 2022].

CVGP's coreset-based variational $\mathcal{GP}$ posterior, learnable via stochastic maximization of a lower-bound of the log-marginal data likelihood, enables not only a more accurate approximation to the true $\mathcal{GP}$ regression posterior, but a more efficient optimization process.

In summary, our contribution is a novel, coreset-based stochastic variational $\mathcal{GP}$ inference (CVGP) algorithm that:

1. Finds a coreset-based, sparse variational posterior to faithfully approximate the true $\mathcal{GP}$ posterior, enabling up- and down-weighting the influence of pseudo-points during learning (Section 4.4 and Appendix D.1,D.3);
2. Maximizes a lower-bound over the marginal log-likelihood that is amenable to efficient stochastic optimization (Section 3.2);
3. Provides a numerically stable algorithm requiring only $\mathcal{O}(M)$ parameters to be learned, at computational and memory complexities of $\mathcal{O}\left(M^3\right)$ and $\mathcal{O}\left(M^2\right)$ (Table 1);
4. Outperforms SOTA stochastic variational $\mathcal{GP}$ inference alternatives on real-world regression datasets (Section 4): CVGP not only provides improved predictive performance (Section 4.2), but achieves a tighter lower variational bound than alternatives (Section 4.3).

2 BACKGROUND

We introduce the notation and foundations of $\mathcal{GP}$ regression in Section 2.1, and describe sparse approximations for scalable $\mathcal{GP}$ inference in Section 2.2. We review the variational inducing point-based foundational work of Titsias [2009] and Hensman et al. [2013], in Sections 2.3 and 2.4, respectively. These are SOTA $\mathcal{GP}$ algorithms that will serve as competitive baselines for the experiments in Section 4.

2.1 $\mathcal{GP}$ REGRESSION

A (univariate) $\mathcal{GP}$ is a non-parametric prior over functions from input domain $\mathbf{x} \in \mathcal{X}$ into scalar space $y \in \mathcal{Y}$, denoted as $f(\mathbf{x};\Theta) \sim \mathcal{GP}\left(m(\mathbf{x};\theta_m), k(\mathbf{x},\mathbf{x};\theta_k)\right)$. A $\mathcal{GP}$ is specified by its mean, $m(\mathbf{x};\theta_m) : \mathcal{X} \to I\!R$, and covariance (kernel), $k(\mathbf{x},\mathbf{x};\theta_k) : \mathcal{X}\times\mathcal{X} \to I\!R$, functions with parameters θ_m and θ_k that are jointly referred to as $\mathcal{GP}$ hyperparameters $\Theta = \{\theta_m, \theta_k\}$.

The function f is a mapping from $\mathcal{X}$ to the real numbers and we may equivalently write $f \in I\!R^{\mathcal{X}}$, viewing functions as (infinite-dimensional) vectors with elements indexed by members of $\mathcal{X}$. Using vector notation, we define $\mathbf{f} = f(\mathbf{X})$ as the vector containing the $\mathcal{GP}$ prior values at a collection of points $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$. The $\mathcal{GP}$ prior evaluated at any subset of points $\mathbf{X}$ follows a multivariate Gaussian distribution $p(\mathbf{f};\Theta) \sim \mathcal{N}\left(\mathbf{f} \mid \mathbf{0}, \mathbf{K}_{NN}\right)$, with $\mathbf{K}_{NN} = \left(k(\mathbf{x}_i,\mathbf{x}_j)\right)_{1\le i,j\le N}$. We assume zero mean $\mathcal{GP}$ priors without loss of generality, and suppress explicit dependence on input points $\mathbf{X}$ to avoid notation clutter.

In $\mathcal{GP}$ regression with observations subject to Gaussian noise, i.e., $y = f(\mathbf{x};\Theta) + \epsilon$, $\epsilon \sim \mathcal{N}\left(\epsilon|0,\sigma^2\right)$, the data marginal likelihood is given by

$$p(\mathbf{y}) = \int_{\mathbf{f}} p(\mathbf{y}|\mathbf{f};\sigma)p(\mathbf{f};\Theta)\,\mathrm{d}\mathbf{f} = \mathcal{N}\left(\mathbf{y}|\mathbf{0},\sigma^2\mathbf{I}+\mathbf{K}_{NN}\right). \tag{1}$$

Given some observed data $\mathbf{X}$, the posterior over the $\mathcal{GP}$ function at any input $\mathbf{x}^\star$, $f^\star = f(\mathbf{x}^\star)$, is a Gaussian distribution computable in closed form, i.e.,

$$\begin{aligned} &p(f^\star \mid \mathbf{x}^\star,\mathbf{y}) = \mathcal{N}\left(f \mid m_{f^\star|\mathbf{x}^\star}, k_{f^\star|\mathbf{x}^\star}\right), \text{ with} \\ &m_{f^\star|\mathbf{x}^\star} = \mathbf{k}_{*N}\left(\sigma^2\mathbf{I}+\mathbf{K}_{NN}\right)^{-1}\mathbf{y}\,, \\ &k_{f^\star|\mathbf{x}^\star} = k(\mathbf{x}^\star,\mathbf{x}^\star) - \mathbf{k}_{*N}\left(\sigma^2\mathbf{I}+\mathbf{K}_{NN}\right)^{-1}\mathbf{k}_{N*}\,. \end{aligned} \tag{2}$$

where $\mathbf{k}_{*N}$ is the N-dimensional row vector of kernel function values between a new input $\mathbf{x}^\star$ and observed data $\mathbf{X}$.

2.2 SPARSE GAUSSIAN PROCESS REGRESSION

Even though posterior statistics in Equation (2) are analytically tractable, they raise computational challenges for big data, as they require computation of the inverse of $N \times N$ matrices with, in general, $\mathcal{O}\left(N^3\right)$ time and $\mathcal{O}\left(N^2\right)$ space complexity. An overview of sparse approximations to reduce such computational burden for $\mathcal{GP}$ regression can be found in [Rasmussen et al., 2006, Chapter 8], with a unifying view presented in [Quinonero-Candela and Rasmussen, 2005], summarized below. The innovation in sparse $\mathcal{GP}$s is to design *approximate* posteriors over $\mathcal{GP}$ function values $\mathbf{f}_M = f(\mathbf{X}_M)$ at a subset of M inducing inputs $\mathbf{X}_M$.

Quinonero-Candela and Rasmussen [2005] presented the Fully Independent Training Conditional (FITC) technique, as a unifying framework for many of the sparse $\mathcal{GP}$ formulations that had previously been presented, e.g., [Csató and Opper, 2002, Smola and Bartlett, 2000, Snelson and Ghahramani, 2005]. FITC, which was later connected to methods that approximate the $\mathcal{GP}$ posterior via Expectation Propagation [Snelson, 2008, Yuan et al., 2012, Bui et al., 2017], uses —unlike previous methods [Csató and Opper, 2002, Seeger et al., 2003]— the marginal likelihood to jointly learn the hyperparameters and the inducing points [Snelson and Ghahramani, 2005]. This relaxes the constraint of having the inducing points limited to a subset of the dataset, and turns a discrete inducing point selection problem into a continuous optimization one. Careful inspection of these sparse methodologies and, in particular, FITC [Quinonero-Candela and Rasmussen, 2005, Bauer et al., 2016] pointed out several limitations related to their tendency to overestimate marginal likelihood, which motivated Titsias [2009] to propose a variational formulation for sparse $\mathcal{GP}$ regression.

2.3 VARIATIONAL SPARSE $\mathcal{GP}$

Titsias [2009] revisited sparse $\mathcal{GP}$ inference and pose it as a variational optimization problem on jointly learning

M inducing inputs $\mathbf{X}_M$ (and $\mathcal{GP}$ hyperparameters Θ), by maximizing a lower-bound of the log-marginal likelihood:

$$\log p(\mathbf{y}) \geq \mathcal{L}_{SparseGP} = \mathbb{E}_{q(\mathbf{f},\mathbf{f}_M)} \left\{ \log \frac{p(\mathbf{y},\mathbf{f},\mathbf{f}_M)}{q(\mathbf{f},\mathbf{f}_M)} \right\}, \tag{3}$$

which is equivalent to minimizing the Kullback–Leibler (KL) divergence between the variational family $q \in \mathcal{Q}$ and the $\mathcal{GP}$ posterior, i.e., $\mathrm{KL}\left[q(\mathbf{f},\mathbf{f}_M) \| p(\mathbf{f},\mathbf{f}_M|\mathbf{y})\right]$.

Titsias [2009] showed that, for a factorization of the variational family of the $q(\mathbf{f},\mathbf{f}_M) = p(\mathbf{f} \mid \mathbf{f}_M) q(\mathbf{f}_M)$ form, one can marginalize over the $\mathcal{GP}$ inducing variables $\mathbf{f}_M = f(\mathbf{X}_M)$, to derive the following analytical lower-bound

$$\begin{aligned}\mathcal{L}_{SparseGP} = &\log \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \sigma^2\mathbf{I} + \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right) \\ &- \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\},\end{aligned} \tag{4}$$

which can be computed in $\mathcal{O}\left(NM^2\right)$ time and $\mathcal{O}\left(NM\right)$ space complexity. Equation (4) is the result of integrating out the *optimal* Gaussian variational posterior $q^*(\mathbf{f}_M)$, available in closed form for a set of inducing points $\mathbf{X}_M$, and expressed in terms of the prior modeling choices of kernel and likelihood noise, only used implicitly in inference.

2.4 STOCHASTIC VARIATIONAL $\mathcal{GP}$

Hensman et al. [2013] revisited Titsias [2009]'s evidence lower-bound (ELBO), and showed that it can be amenable to stochastic variational inference for $\mathcal{GP}$s (SVGP), by reorganizing it and avoiding direct marginalization over inducing variables $\mathbf{X}_M$, i.e.,

$$\begin{aligned}\log p(\mathbf{y}) \geq \mathcal{L}_{SVGP} = &\mathbb{E}_{q(\mathbf{f}_M)}\left\{\mathbb{E}_{q(\mathbf{f}|\mathbf{f}_M)}\left\{\log p(\mathbf{y}|\mathbf{f})\right\}\right\} \\ &- \mathrm{KL}\left[q(\mathbf{f}_M) \| p(\mathbf{f}_M)\right] .\end{aligned} \tag{5}$$

SVGP proceeds by defining a free-form variational family $q(\mathbf{f}_M) = \mathcal{N}\left(\mathbf{f}_M \mid \mathbf{m}, \mathbf{S}\right)$ and analytically computing the revised ELBO:

$$\begin{aligned}\mathcal{L}_{SVGP} = &\log \mathcal{N}\left(\mathbf{y} \mid \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{m}, \sigma^2\mathbf{I}\right) \\ &- \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} \\ &- \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{S}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} \\ &- \mathrm{KL}\left[q(\mathbf{f}_M) \| p(\mathbf{f}_M)\right] .\end{aligned} \tag{6}$$

Equation (6) allows for data subsampling, hence enabling stochastic optimization to learn the free variational parameters $\{\mathbf{m}, \mathbf{S}\}$ in $q(\mathbf{f}_M) = \mathcal{N}\left(\mathbf{f}_M \mid \mathbf{m}, \mathbf{S}\right)$, of order $\mathcal{O}\left(2M + M^2\right)$, where an unbiased estimate of the SVGP loss can be computed with $\mathcal{O}\left(M^3\right)$ time- and $\mathcal{O}\left(M^2\right)$ space-complexity.

The optimum of Equation (6) matches that of Equation (4), yet the latter directly leverages the optimal variational distribution $q^*(\mathbf{f}_M)$, while the former resorts to stochastic optimization of its free-form, variational $\mathcal{O}\left(M^2\right)$ parameters to find it. Namely, SparseGP operates by maximizing a tight —based on the optimal $q^*(\mathbf{f}_M)$)— lower-bound, with the disadvantage of not being able to use stochastic optimization.

Our goal here is to leverage the best of each world and to design a variational posterior that incorporates the dependencies set by the prior $\mathcal{GP}$ model (i.e., the kernel and the likelihood noise) for approximate $\mathcal{GP}$ inference that is amenable to stochastic optimization.

3 CORESET-BASED VARIATIONAL POSTERIOR $\mathcal{GP}$ (CVGP)

We use Bayesian coreset principles to derive an sparse approximation to the true $\mathcal{GP}$ regression posterior that is learnable via *stochastic* variational inference.

Bayesian coresets search for samples from a smaller data subset that can, via weighted likelihoods, approximate otherwise hard to compute posterior distributions [Huggins et al., 2016, Campbell and Broderick, 2018, 2019, Jubran et al., 2019]. From an optimization perspective, Bayesian coresets can also be understood as a set of *learnable* (observed or unobserved) points selected to minimize some divergence to a distribution of interest [Manousakas et al., 2020, 2022].

Inspired by such framework, we posit a coreset-based, variational posterior distribution for $\mathcal{GP}$s (CVGP): i.e., we learn a small subset of *pseudo-inputs* $\mathbf{X}_M = \{\mathbf{x}_1, \cdots, \mathbf{x}_M\}$, and *pseudo-observations* $\mathbf{y}_M = \{y_1, \cdots, y_M\}$, that if reweighted appropriately with parameters $\boldsymbol{\beta}_M = (\beta_1, \cdots, \beta_M)$, approximate the $\mathcal{GP}$ posterior accurately. Contrary to standard Bayesian coreset methodology, the coreset tuple $\{\mathbf{X}_M, \mathbf{y}_M\}$ is composed by *learnable pseudo-points* in the input-output data space —not restricted to the observed empirical data.

For accurate approximation of the posterior, and inspired by Titsias [2009]'s optimal solution, we ensure that CVGP's posterior obeys the $\mathcal{GP}$ prior-conditional and it's inductive biases (see Section 3.1). We learn the CVGP posterior by formulating a variational lower-bound objective that is amenable to its stochastic maximization (see Section 3.2).

3.1 THE CORESET-BASED $\mathcal{GP}$ POSTERIOR

CVGP's key novelty is a coreset-based distribution designed to incorporate the $\mathcal{GP}$'s prior model and likelihood characterizations into the CVGP posterior.

We formulate a coreset-based distribution $q(\mathbf{f}_M)$ over $\mathcal{GP}$ variables $\mathbf{f}_M = f(\mathbf{X}_M)$ at pseudo-inputs $\mathbf{X}_M = \{\mathbf{x}_m\}_{m=1}^M$

and associated pseudo-observations $\mathbf{y}_M = \{y_m\}_{m=1}^M$ as

$$q(\mathbf{f}_M \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M) = \frac{q(\mathbf{y}_M \mid \mathbf{f}_M, \boldsymbol{\beta}_M) p(\mathbf{f}_M \mid \mathbf{X}_M)}{p(\mathbf{y}_M \mid \mathbf{X}_M, \boldsymbol{\beta}_M)} = \frac{\left(\prod_{m=1}^M p(y_m|f_m)^{\beta_m}\right) p(\mathbf{f}_M|\mathbf{X}_M)}{p(\mathbf{y}_M|\mathbf{X}_M, \boldsymbol{\beta}_M)} , \quad (7)$$

where the data likelihood for each pseudo-observation $p(y_m|f_m), m \in \{1, \cdots, M\}$, is raised to the power of learnable parameters $\boldsymbol{\beta}_M = (\beta_1, \cdots, \beta_M)$. The CVGP posterior is a tempered distribution, which can be understood as if a small subset $M \leq N$ of pseudo-input/output pairs $\{\mathbf{X}_m, y_m\}$ are each drawn $\beta_m \geq 0$ times.

For a Gaussian observation likelihood,[1] we derive in Appendix Section A.1.1 the closed-form multivariate Gaussian distribution of CVGP's posterior over $\mathcal{GP}$ function variables $\mathbf{f}_M$, given coreset triplet $\{\mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M\}$:

$$q(\mathbf{f}_M \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M) = \mathcal{N}\left(\mathbf{f}_M | \mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}, \mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\right) , \quad (8)$$

$$\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M} = \mathbf{K}_{MM}(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M})^{-1}\mathbf{y}_M,$$
$$\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M} = \mathbf{K}_{MM} - \mathbf{K}_{MM}(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M})^{-1}\mathbf{K}_{MM},$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_M} = \sigma^2 \cdot \mathrm{diag}\left\{\boldsymbol{\beta}_M^{-1}\right\}$.

With this coreset-based distribution over coreset $\mathcal{GP}$ values $q(\mathbf{f}_M)$,[2] we now accommodate the $\mathcal{GP}$ prior's conditional dependency, $q(\mathbf{f}, \mathbf{f}_M) = p(\mathbf{f} \mid \mathbf{f}_M) q(\mathbf{f}_M)$, where

$$p(\mathbf{f} \mid \mathbf{f}_M) = \mathcal{N}\left(\mathbf{f}|\mathbf{m}_{\mathbf{f}|\mathbf{f}_M}, \mathbf{K}_{\mathbf{f}|\mathbf{f}_M}\right) \text{ , with} \quad (9)$$
$$\mathbf{m}_{\mathbf{f}|\mathbf{f}_M} = \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{f}_M,$$
$$\mathbf{K}_{\mathbf{f}|\mathbf{f}_M} = \mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN},$$

and compute the variational posterior of interest, $q(\mathbf{f} \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M)$ over $\mathcal{GP}$ function values $\mathbf{f} = f(\mathbf{X})$, by marginalizing the coreset-based, tempered posterior of Equation (8) from the joint distribution $q(\mathbf{f}, \mathbf{f}_M)$. The resulting CVGP coreset-based variational posterior is

$$q(\mathbf{f} \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M) = \mathcal{N}\left(\mathbf{f}|\mathbf{m}_{\mathbf{f}|\mathbf{y}_M}, \mathbf{K}_{\mathbf{f}|\mathbf{y}_M}\right) , \quad (10)$$
$$\mathbf{m}_{\mathbf{f}|\mathbf{y}_M} = \mathbf{K}_{NM}\left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{y}_M ,$$
$$\mathbf{K}_{\mathbf{f}|\mathbf{y}_M} = \mathbf{K}_{NN} - \mathbf{K}_{NM}\left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MN}.$$

If one were to follow standard Bayesian coreset procedures, we would directly aim to learn the coresets that best approximate $q(\mathbf{f} \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M)$ to the true posterior —which requires computing the $\mathcal{GP}$ posterior in Equation (2) of $\mathcal{O}\left(N^3\right)$ complexity [Manousakas et al., 2022]. On the contrary, we learn the coreset triplet $\{\mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M\}$ using a variational objective that aims to minimize the divergence between such two distributions at reduced computational cost, and in a form amenable to its stochastic minimization.

[1]Derivation of closed-form, coreset-based posteriors for non-Gaussian likelihoods is part of future investigations.

[2]The interested reader can find the complementary weight-space derivations in Appendix Section A.2.

3.2 CVGP'S VARIATIONAL LOWER-BOUND

We denote with $q(\cdot)$ a generic variational family of distributions over a $\mathcal{GP}$. Whenever $q(\mathbf{f}) \neq p(\mathbf{f} \mid \mathbf{y})$, we can lower-bound the log-marginal distribution,

$$\log p(\mathbf{y}) \geq \mathcal{L} = \mathbb{E}_{q(\mathbf{f})}\left\{\log p(\mathbf{y} \mid \mathbf{f})\right\} - \mathrm{KL}\left[q(\mathbf{f}) \| p(\mathbf{f})\right] ,$$

incurring on a gap determined by the Kullback–Leibler (KL) divergence between the variational distribution and the $\mathcal{GP}$ posterior. Hence, maximizing the loss $\mathcal{L}$ is equivalent to minimizing the KL divergence between the variational family $q(\mathbf{f})$ and the true posterior $p(\mathbf{f}|\mathbf{y})$, i.e., minimizing the gap $\Delta(\mathbf{f}) = \mathrm{KL}\left[q(\mathbf{f}) \| p(\mathbf{f} \mid \mathbf{y})\right]$.

In CVGP, we use the coreset-based posteriors of Equations (8) and (10) to maximize the lower-bound, i.e.,

$$\mathcal{L}_{CVGP} = \mathbb{E}_{q(\mathbf{f}|\mathbf{y}_M, \mathbf{X}_M, \boldsymbol{\beta}_M)}\left\{\log p(\mathbf{y}|\mathbf{f})\right\} - \mathrm{KL}\left[q(\mathbf{f}_M|\mathbf{y}_M, \mathbf{X}_M, \boldsymbol{\beta}_M) \| p(\mathbf{f}_M)\right] , \quad (11)$$

which has the following analytical solution:

$$\mathcal{L}_{CVGP} = \log \mathcal{N}\left(\mathbf{y}|\mathbf{m}_{\mathbf{f}|\mathbf{y}_M}, \sigma^2\mathbf{I}\right) - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{\mathbf{f}|\mathbf{y}_M}\right\} + \frac{1}{2}\left[\mathrm{tr}\left\{\mathbf{A}\mathbf{K}_{MM}\right\} - \mathbf{y}_M^\top\mathbf{A}\mathbf{K}_{MM}\mathbf{A}\mathbf{y}_M + \ln\left|\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}\right|\right] , \quad (12)$$

where $\mathbf{A} = \left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}$. Full details of the derivation are provided in Appendix Section A.

CVGP's lower-bound $\mathcal{L}_{CVGP}$: *optimality at reduced complexity and increased numerical stability*. Maximization of the variational lower-bound in Equation (12) to learn CVGP's coreset-based posterior in Equation(10) gives rise to the following desirable properties:

1. The maximum of Equation (12) is identical to the loss in Equation (4) derived by Titsias [2009]: i.e., SparseGP and CVGP *have the same optimum* —see proofs in Appendix Section A.3.
2. The lower-bound in Equation (12) is amenable to data-subsampling. Due to the uncorrelated Gaussian likelihood term and properties of the trace, we can apply *stochastic optimization* for its maximization, computing unbiased loss estimates with reduced (a single) data sample.
3. The algorithmic complexity of CVGP, for coreset size M, is $\mathcal{O}\left(M^3\right)$ in computational time and $\mathcal{O}\left(M^2\right)$ in space complexity. Importantly, CVGP's parameter complexity is of *reduced $\mathcal{O}\left(M\right)$ order*, as it only requires learning coreset triplets $(\mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M)$, each of size M —see Table 1 for a full comparison.
4. CVGP's posterior and lower-bound inherently provide a numerically stable stochastic algorithm, as all matrix inverse operations in Equation(12) involve $\mathbf{A} =$

$\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}$: the sum of a diagonal matrix ($\mathbf{\Sigma}_{\boldsymbol{\beta}_M}$) defined by positive coreset weights $\boldsymbol{\beta}_M \geq 0$ and positive definite matrix $\mathbf{K}_{MM}$.[3]

Algorithm	Complexities		
	Time	Space	Parameter
SparseGP	$\mathcal{O}\left(NM^2\right)$	$\mathcal{O}\left(NM^2\right)$	$\boldsymbol{\mathcal{O}\left(M\right)}$
SVGP	$\boldsymbol{\mathcal{O}\left(M^3\right)}$	$\boldsymbol{\mathcal{O}\left(M^2\right)}$	$\mathcal{O}\left(M^2\right)$
CVGP	$\boldsymbol{\mathcal{O}\left(M^3\right)}$	$\boldsymbol{\mathcal{O}\left(M^2\right)}$	$\boldsymbol{\mathcal{O}\left(M\right)}$

Table 1: Computational analysis of CVGP and sparse variational $\mathcal{GP}$ alternatives: time and space complexities for obtaining an unbiased estimate of their objectives. Desirable complexities are highlighted in **bold**. CVGP enjoys same time and space complexity as SVGP, at a reduced variational parameter dimensionality.

3.3 A COMPARISON TO ALTERNATIVES

CVGP is, to the best of our knowledge, the first variational $\mathcal{GP}$ inference method that leverages a coreset-based posterior for efficiency and scalability. It diverges from alternative sparse $\mathcal{GP}$ inference techniques in that its posterior is based on a coreset triplet $\{\mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M\}$:

- CVGP is not restricted to a sparse selection of observed inputs: $\mathbf{X}_M$ is a vector of free parameters, within the data domain, but **not restricted to the empirical data**.
- CVGP does not learn inducing variables $\mathbf{m} = \mathbb{E}_{q(\mathbf{f}_M)}\{\mathbf{f}_M\}$, i.e., posterior $\mathcal{GP}$ mean function values evaluated at inducing points $\mathbf{X}_M$. Instead, it **learns pseudo-observations** $\mathbf{y}_M$ that encapsulate (i.e., capture the characteristics of) the observed data (e.g., Figure 5).
- CVGP is the only existing $\mathcal{GP}$ method that **reweights the pseudo-observations** with learnable parameters $\boldsymbol{\beta}_M$, for flexibility and explainability of its coreset-based posterior: i.e., it learns which pseudo-points are important for accurate $\mathcal{GP}$ posterior approximation (Figures 4 and 5).

Comparison to non-variational sparse $\mathcal{GP}$s. Selection of $\mathcal{GP}$ inputs from within the training data involves a prohibitive combinatorial optimization that may require greedy optimization [Csató and Opper, 2002], based on posterior maximization [Smola and Bartlett, 2000], maximum information gain [Seeger et al., 2003], matching pursuit [Keerthi and Chu, 2005], or other techniques [Quinonero-Candela and Rasmussen, 2005]. On the contrary, CVGP leverages *stochastic optimization* to find a weighted subset of pseudo-points that efficiently approximate the $\mathcal{GP}$ posterior, sharing resemblance with the pioneer work of Snelson and Ghahramani [2005]. To circumvent overestimation of the marginal likelihood and under-estimation of the noise variance as reported by Titsias [2009], Bauer et al. [2016], CVGP resorts to variational inference. Hence, CVGP shares the variational formulation of Titsias [2009] and Hensman et al. [2013], yet is distinct in several important aspects.

Comparison to variational sparse $\mathcal{GP}$s. CVGP aligns with the approach by Titsias [2009] in the use of a variational lower-bound on the marginal log-likelihood that leverages the $\mathcal{GP}$ prior's conditional dependency, i.e., $q(\mathbf{f}, \mathbf{f}_M) = p(\mathbf{f} \mid \mathbf{f}_M)q(\mathbf{f}_M)$, and analytically marginalizes $q(\mathbf{f}_M)$. In contrast, SVGP does not marginalize this distribution and devises a different lower-bound for stochastic optimization. As a result, SparseGP and CVGP posteriors directly incorporate the $\mathcal{GP}$ prior's inductive biases and the likelihood model. The main difference is in the choice of $q(\mathbf{f}_M)$:

- SparseGP derives *the optimum distribution* at inputs $\mathbf{X}_M$ over function values $\mathbf{f}_M$, given observed data $\mathbf{y}$:

$$q^\star(\mathbf{f}_M) = \mathcal{N}\left(\mathbf{f}_M; \mathbf{m}^*_{\mathbf{f}_M}, \mathbf{K}^*_{\mathbf{f}_M,\mathbf{f}_M}\right), \text{ with} \tag{13}$$
$$\begin{cases} \mathbf{m}^*_{\mathbf{f}_M} = \mathbf{K}_{MM}\left(\sigma^2\mathbf{K}_{MM} + \mathbf{K}_{MN}\mathbf{K}_{NM}\right)^{-1}\mathbf{K}_{MN}\mathbf{y} \\ \mathbf{K}^*_{\mathbf{f}_M,\mathbf{f}_M} = \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \frac{1}{\sigma^2}\mathbf{K}_{MN}\mathbf{K}_{NM}\right)^{-1}\mathbf{K}_{MM} \end{cases}$$

- CVGP defines *a learnable distribution* $q(\mathbf{X}_M)$ with free coreset parameter triplet $\{\mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M\}$:

$$q(\mathbf{f}_M) = \mathcal{N}\left(\mathbf{f}_M; \mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}, \mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\right), \text{ with} \tag{14}$$
$$\begin{cases} \mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M} = \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{y}_M \\ \mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M} = \mathbf{K}_{MM}\left[\mathbf{K}_{MM}^{-1} - \left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\right]\mathbf{K}_{MM} \end{cases}$$

We note that the building blocks of CVGP's coreset based posterior are analogous to SparseGP's optimal posterior: CVGP's learned pseudo-observations $\mathbf{y}_M$ can be viewed as a weighted combination of observed datapoints, i.e., the $\mathbf{K}_{MN}\mathbf{y}$ term in SparsedGP's posterior mean. In addition, CVGP pseudo-observations $\mathbf{y}_M$ are modulated by the $\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}$ term in its posterior mean; in SparseGP, the $\left(\sigma^2\mathbf{K}_{MM} + \mathbf{K}_{MN}\mathbf{K}_{NM}\right)^{-1}$ term similarly weights the transformed observations $\mathbf{K}_{MN}\mathbf{y}$. In both posterior distributions, these terms in red are responsible for balancing the prior inductive biases with the information provided by observed data: i.e., the posterior means interpolate between the prior and observations. A similar dependency between the prior and the information provided by data is observed in the posterior covariances: i.e., the blue terms in both posteriors adapt the prior covariance to account for the uncertainty reduction due to observations. In CVGP, this balance is adjusted through the learnable matrix $\mathbf{\Sigma}_{\boldsymbol{\beta}_M}$, whereas in SparseGP, it is determined by the fixed dependency set by the prior covariance and the likelihood noise, i.e., $\frac{1}{\sigma^2}\mathbf{K}_{MN}\mathbf{K}_{NM}$.

[3] In theory, $\mathbf{K}_{MM}$ is positive definite and invertible. However, numerical issues can cause instability when inverted in practice.

Notably, as shown in Appendix Section A.3, when CVGP matrix $\mathbf{\Sigma}_{\boldsymbol{\beta}_M}$ matches the appropriate weighting, the optimum of SparseGP and CVGP's loss-functions are identical. Hence, the learned solutions match with $\mathbf{y}_M = \sigma^{-2}\mathbf{\Sigma}^*_{\boldsymbol{\beta}_M}\mathbf{y}$ and $\mathbf{\Sigma}^*_{\boldsymbol{\beta}_M} = \sigma^2\mathbf{K}_{MM}\left(\mathbf{K}_{MN}\mathbf{K}_{NM}\right)^{-1}\mathbf{K}_{MM}$, recovering Titsias [2009]'s optimal solution. We empirically showcase CVGP's ability to quickly and **efficiently close the gap to ExactGP's marginal log-likelihood** in Section 4.3.

Contrary to SparseGP, CVGP's loss in Equation (12) is amenable to stochastic optimization, making sparse $\mathcal{GP}$ regression scalable at reduced complexity. CVGP matches SVGP's scalability [Hensman et al., 2013], yet offers two key advantages: linear parameter complexity of order $\mathcal{O}(M)$, and a distinct optimization landscape. These arise from different design choices over $q(\mathbf{f}_M)$: whereas SVGP's free-form $q(\mathbf{f}_M) = \mathcal{N}(\mathbf{f}_M \mid \mathbf{m}, \mathbf{S})$ requires $\mathcal{O}\left(M^2\right)$ parameters and yields statistics $(\mathbf{m}, \mathbf{S})$ not directly tied to the model or data likelihood; CVGP's posterior in Equation (7) leverages the model's inductive biases, acting as a **natural interpolation between the $\mathcal{GP}$ prior and the data likelihood.**[4] These structural differences produce distinct loss landscapes, with SVGP's higher-dimensional optimization often struggling to converge, as shown in Section 4.3.

3.4 CVGP AS BAYESIAN CORESET LEARNING

CVGP enables a complementary Bayesian coreset learning-based perspective on sparse $\mathcal{GP}$ inference. Methodologically, CVGP maximizes the loss in Equation (12) for $\mathcal{GP}$ posterior inference; i.e., it maximizes the variational lower-bound $\mathcal{L}_{CVGP}$ with respect to CVGP parameters $\{\mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M\}$, encouraging approximations that minimize the gap to the true $\mathcal{GP}$ posterior. We note that, $\mathcal{L}_{CVGP} \rightarrow \mathcal{L} = \log p(\mathbf{y})$ implies $\Delta_{CVGP} = \mathrm{KL}\left[q(\mathbf{f}, \mathbf{f}_M \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M) \| p(\mathbf{f}, \mathbf{f}_M \mid \mathbf{y})\right] \rightarrow 0$. Hence, CVGP learns coresets that minimize the distance between its variational distribution and the true $\mathcal{GP}$ posterior.

To do so, it finds —indirectly, yet efficiently— a sparse representation of the data (i.e., the coreset triplet) that captures as much information as the $\mathcal{GP}$ posterior of interest, measured by the KL divergence between the true and CVGP's posterior. Initial estimates of the coreset triplet $\{\mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M\}$ can be selected randomly or using k-means (we evaluate CVGP's robustness to coreset initialization in Appendix C.3 and D) and recommend the latter.

Importantly, CVGP's learning procedure enables an **automatic relevance determination of pseudo-points** $\{\mathbf{X}_M, \mathbf{y}_M\}$ via adaptation of their $\boldsymbol{\beta}_M$ values: i.e., CVGP has the inherent flexibility to up- or down-weight ("*ignore*") the pseudo-points that are deemed (or not) important to describe the observed data —see experiments in Section 4.4.

[4]We analyze CVGP's prior to posterior noise adaptation as a function of observation noise levels in Appendix C.4.

Namely, inspection of $q(\mathbf{f} \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M)$ elucidates which learned coreset tuples $\{\mathbf{X}_M, \mathbf{y}_M\}$ weighted by $\boldsymbol{\beta}_M$, help describe the $\mathcal{GP}$ posterior best —as illustrated in Figure 4. We note that CVGP's coreset-based variational posteriors, when derived from the function-space and weight-space views of $\mathcal{GP}$s —see Appendix Section A for both derivations— provide complementary posterior insights.

4 EXPERIMENTS

We demonstrate CVGP's superior predictive performance in real-datasets in Section 4.2, before delving into its inference advantages in Section 4.3. We showcase the quality and explainability of the learned CVGP posteriors in Section 4.4.

4.1 EXPERIMENTAL SETUP

We compare CVGP against benchmark $\mathcal{GP}$ alternatives described in Section 2: ExactGP [Rasmussen et al., 2006], SparseGP [Titsias, 2009], and SVGP Hensman et al. [2013]. We also incorporate Parametric Gaussian Process Regressors (PPGPR) by Jankowiak et al. [2020] as a strong predictive baseline. We implement CVGP using Pytorch and GPyTorch libraries, and use benchmark GPytorch implementations [Gardner et al., 2018] for the baselines.

We use a zero-mean $\mathcal{GP}$ prior with a Radial basis kernel function (RBF) in all experiments, as the goal is to compare —for the same GP model— which approximate $\mathcal{GP}$ technique provides better inference and predictive performance. We evaluate different coreset (CVGP) and inducing point sizes M for sparse $\mathcal{GP}$ baselines (SparseGP, SVGP and PPGPR), all initialized with k-means [Hartigan and Wong, 1979]. We employ 5-fold cross-validation to compute and report each technique's predictive root-mean-squared error (RMSE) and posterior predictive log-likelihood (PPLL), over held out test splits. We enforce best validation RMSE performance as early stopping criteria. All details for the reproducibility of the experiments are provided in Appendix Section B.3.

CVGP predictive and inference experiments of Sections 4.2 and 4.3 are based on real-world regression datasets from the UCI machine learning repository data [Asuncion and Newman, 2007]. We use simulated datasets to showcase learned predictive posteriors in Section 4.4, with all dataset details described in Appendix Section B.1.

4.2 PREDICTIVE PERFORMANCE

We assess the predictive performance of all sparse $\mathcal{GP}$ methods for a variety of real-datasets, and illustrate the performance of ExactGP —when computationally possible— as the optimal benchmark, in Figure 1.

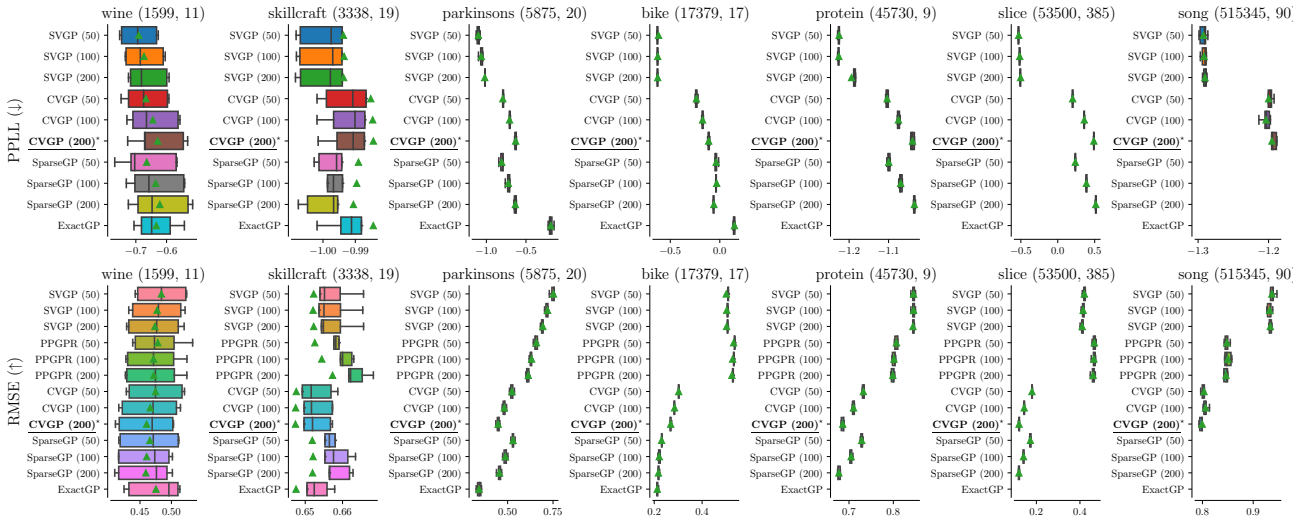


Figure 1: Box-and-whisker diagrams of predictive metrics (RMSE and PPLL) on real datasets. The titles denote the dataset and in parenthesis, its size and feature-dimensionality. Arrows indicate the desirable metric direction: higher PPLL (to the right) and lower RMSE (to the left). CVGP outperforms SVGP and PPGPR, and is on-par with SparseGP, with as few as 50 coresets. The best performing *stochastic gradient* model mean statistic (▲) is **emphasized***. SparseGP and ExactGP results are omitted for the largest datasets due to computational complexities.

Figure 1 demonstrates how CVGP outperforms (higher PPPL, lower RMSE) *stochastic* sparse $\mathcal{GP}$ alternatives (SVGP and PPGPR) consistently, with performance on par with SparseGP across all predictive metrics —we inspect the learning and inference gaps between methods in Section 4.3 and Appendix C.2.

Although CVGP, SVGP and SparseGP share the same theoretical optimum, empirical predictive performance in Figure 1 showcases that SVGP rarely reaches the desirable performance of SparseGP, while CVGP's is consistently similar to SparseGP —recall that SparseGP does not allow for stochastic optimization, while CVGP does.

CVGP's performance improves with increaset coreset size and —with as little as 50 coresets— consistently outperforms alternative stochastic methods, even when these baselines use *4-times* more inducing points, i.e., SVGP (200) and PPGPR (200). CVGP's predictive performance is also better than PPGPR, an approximate $\mathcal{GP}$ algorithm specifically designed for predictive performance.

We showcase in Figure 2 and Appendix C.1 the evolution of RMSE and PPLL across training, where training of models does not stop until there are no RMSE improvements. Notice that, while CVGP metrics improve consistently over training, the RMSE for PPGPR improves, while its PPLL deteriorates over training epochs. [5]

We demonstrate CVGP's predictive performance robustness to initialization in Figure 8 in Appendix Section C.3. We notice k-means and randomly initialized CVGPs' performance to be similar across metrics and datasets, which is likely due to the coreset-based posteriors' flexibility to up- and down-weight pseudo-input/output pairs via $\boldsymbol{\beta}_M$, *a property other methods do not pose*.

[5]Due to the large negative PPLL values of PPGPR, we have not reported them in Figure 1, see Appendix C.1.

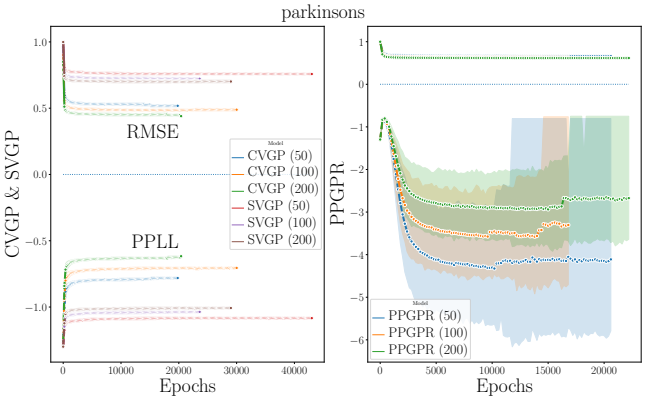


Figure 2: Evolution of RMSE and PPLL across training epochs. CVGP and SVGP's RMSE and PPLL consistently decrease with training epochs. Even though PPGPR's RMSE improves over epochs, its PPLL deteriorates — indicating some form of overfitting.

4.3 INFERENCE PERFORMANCE

We investigate why CVGP approximates the $\mathcal{GP}$ posterior predictive distributions more accurately, by studying the relationship between the variational lower-bounds ($\mathcal{L}$)

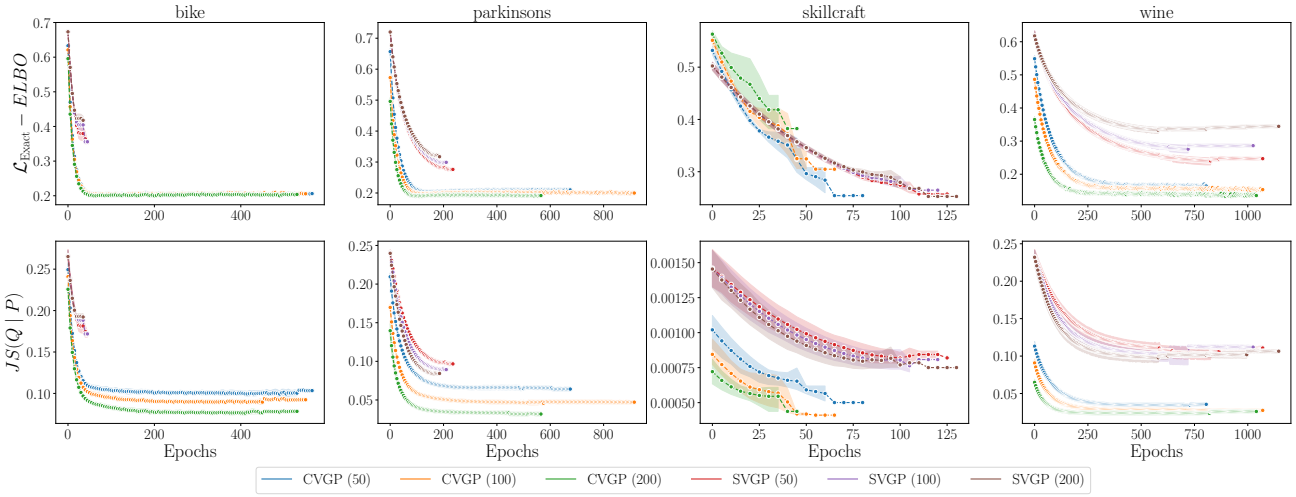


Figure 3: Learning and inference gaps for sparse $\mathcal{GP}$ methods over training, as measured by (top-row) the difference between the log-marginal of ExactGP and the variational bound for SVGP and CVGP; and (bottom-row) the Jensen-Shannon divergence between the exact posterior predictive and each method's approximate posterior predictive. CVGP provides a better approximation to the exact $\mathcal{GP}$ posterior, consistently optimizing a tighter lower-bound. A more detailed figure including PPGPR is available in Appendix C.2.

of sparse $\mathcal{GP}$ alternatives and the true $\mathcal{GP}$ marginal log-likelihood in Equation (1).

To that end, we depict in Figure 3 the difference between the log-marginal of ExactGP and the variational loss optimized by SVGP and CVGP. We also show the inference gap of these methods while in training, over held-out datasets, using the Jensen-Shannon divergence between the exact posterior predictive distribution $p(\mathbf{f}^\star \mid \mathbf{x}^\star, \mathbf{y}) = \int p(\mathbf{f}^\star \mid \mathbf{x}^\star, \mathbf{f}) p(\mathbf{f} \mid \mathbf{y}) \, \mathrm{d}\mathbf{f}$ in Equation (2), and each method's approximate posterior predictive $q(\mathbf{f}^\star \mid \mathbf{x}^\star) = \int p(\mathbf{f}^\star \mid \mathbf{x}^\star, \mathbf{f}_M) q(\mathbf{f}_M) \, \mathrm{d}\mathbf{f}_M$. We employ fixed, equal $\mathcal{GP}$ prior hyperparameters for all models.

Results in Figure 3 demonstrate how CVGP better closes the learning gap with ExactGP. In contrast, SVGP offers a looser bound even if, in theory, both loss functions have the same optimum. Moreover, smaller divergence from CVGP's posterior to that of ExactGP suggests that CVGP better approximates the $\mathcal{GP}$ posterior of interest, at only $\mathcal{O}(M)$ parameter, $\mathcal{O}(M^3)$ time and $\mathcal{O}(M^2)$ space complexities.

This notable inference improvement is attained with as little as 50 coresets, performance not reached by SVGP even with 200 inducing points. We argue that this performance gap is the result of the distinct optimization landscapes of the former compared to the latter, induced by the lower-dimensionality of CVGP's optimization problem and the explicit inductive biases present in CVGP's posterior: (i) its ability to interpolate easily between prior and posterior (see Appendix C.4 for more experiments), and (ii) its ability to learn informative pseudo-points —further investigated in what follows and in Appendix D.

4.4 CVGP AS BAYESIAN CORESET LEARNING

We illustrate posterior predictive distributions learned by stochastic sparse $\mathcal{GP}$ methods in Figure 4, for a 1-dimensional synthetic dataset.

We observe CVGP's approximate posterior to be closest to the exact predictive posterior, both in predicted mean and uncertainty quantification. On the contrary, SVGP and PPGPR encounter difficulties in accurately modeling the function of interest and their uncertainty in the $x \in (0, 2)$ range: SVGP computes a *low-uncertainty*, *smooth* posterior predictive mean, while PPGPR captures the mean but overestimates uncertainty for $x \in (1, 2)$. CVGP, regardless of initialization, better handles this noisy region, matching ExactGP's mean and uncertainty by learning coreset triplets $\{\mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M\}$ with up-weighted $\mathbf{y}_M$ that mitigate posterior $\mathcal{GP}$ uncertainty overestimation.

The input locations $\mathbf{X}_M$ learned by all sparse $\mathcal{GP}$ methods spread across the range of observed data $\mathbf{X}$ in Figure 4. While inducing-points methods SVGP and PPGPR learn $\{\mathbf{X}_M, \mathbf{m}_M = \mathbb{E}_q\{\mathbf{f}_M\}\}$ pairs, CVGP learns pseudo-points $\{\mathbf{X}_M, \mathbf{y}_M\}$ with pseudo-observations $\mathbf{y}_M$ in the observation space $\mathcal{Y}$. Hence, CVGP can learn pseudo-observations $\mathbf{y}_M$ that are correlated with observed data $\mathbf{y}$. Notice how, in Figure 5, CVGP's posterior is based on pseudo-outputs $\mathbf{y}_M$ that are far from the $\mathcal{GP}$ latent values $\mathbf{f}$ in the $x \in (1, 2)$ range, which are up-weighted (i.e., green colored dots), where the observations are subject to heteroskedasticity.

Figure 5 also shows CVGP's learned histograms of $\boldsymbol{\beta}_M$, where we compare CVGP with k-means and random initializations (RandomCVGP).

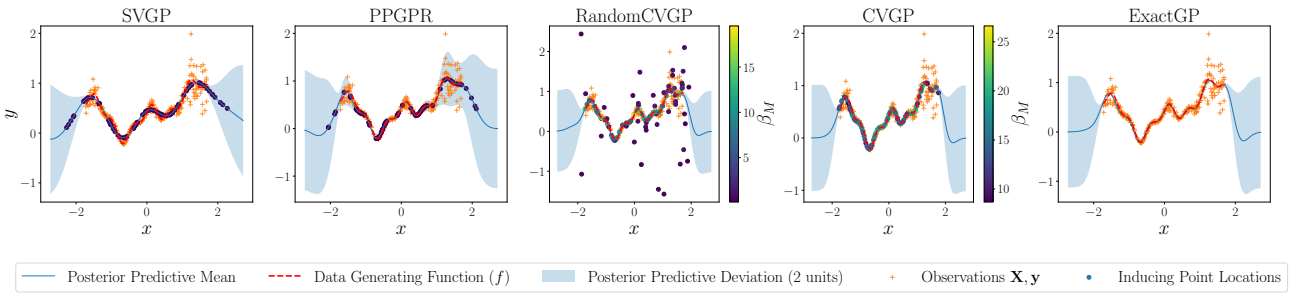


Figure 4: True data generating function (—), posterior predictive mean (—), and 2-unit credible intervals (shaded) for stochastic $\mathcal{GP}$ methods. We indicate the inducing variables $\{\mathbf{X}_M, \mathbf{m}_M = \mathbb{E}_q\{\mathbf{f}_M\}\}$ learned by SVGP and PPGPR; and for CVGP, the learned coreset pseudo-points $\{\mathbf{X}_M, \mathbf{y}_M\}$, with each pseudo-point's color intensity weighted by the learned β_m on the right hand-side bars. Notice CVGP's high-quality posterior, most similar to that of ExactGP, which serves as gold standard. All methods revert to the prior for ranges where *no data* has been observed.

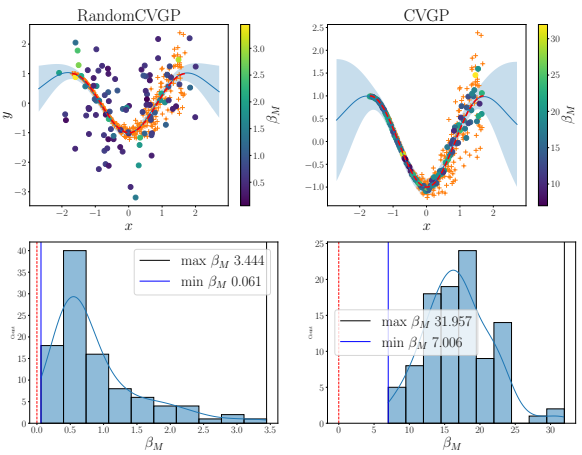


Figure 5: Learned coresets (top) and histograms of their weights (bottom) for CVGP with random (RandomCVGP) and k-means initialization, with legend as in Figure 4. CVGP down-weights uninformative data, yielding many $\boldsymbol{\beta}_M \approx 0$ for RandomCVGP (removing unhelpful points from its posterior). *Unlike other inducing-point methods —which must learn good locations— CVGP can eliminate (down-weight) points that do not converge to plausible values.*

Figure 5 illustrates CVGP's ability to up- and down-weight pseudo-input/output pairs, for both initializations. RandomCVGP drives many β_m to 0 for uninformative data regions, effectively ignoring those pseudo-points, while up-weighting more informative ones, improving posterior efficiency —recall that, in coreset-based posteriors, $\beta_m \geq 0$ corresponds to drawing β_m samples for each pseudo-point $\{\mathbf{X}_m, \mathbf{y}_m\}$.

We argue that it is CVGP's coreset-based distribution that enables efficient and accurate approximation of $\mathcal{GP}$ posteriors at a lower parameter complexity: i.e., better predictive posterior, based on fewer pseudo-points M. Additional benefits of CVGP coreset-based posteriors, namely posterior explainability and compact, informative representations of datasets are illustrated in Appendix Section D.2.

5 CONCLUSION

We introduced CVGP, the first $\mathcal{GP}$ inference method that leverages a coreset-based, variational posterior for accurate and scalable $\mathcal{GP}$ inference. CVGP enables stochastic optimization of its variational lower-bound to the $\mathcal{GP}$'s marginal log-likelihood, after marginalization of latent $\mathcal{GP}$ variables, at reduced $\mathcal{O}(M)$ parameter complexity, with $\mathcal{O}(M^3)$ time- and $\mathcal{O}(M^2)$ space-requirements.

Experimental results demonstrate that CVGP provides improved inference and predictive capabilities, outperforming stochastic variational inference-based alternatives. CVGP provides a high-quality $\mathcal{GP}$ posterior approximation that effectively interpolates between the $\mathcal{GP}$ prior and the data likelihood, learned via an efficient lower-dimensional stochastic optimization problem that results in CVGP achieving a tighter lower-bound than stochastic variational alternatives. Overall, CVGP's coreset-based posterior accurately approximates the true $\mathcal{GP}$ posterior, providing a sparse and explainable representation of the $\mathcal{GP}$ posterior, with added flexibility to adjust (and discard) pseudo-input/output pairs.

Building upon CVGP's formulation for $\mathcal{GP}$ regression, we embark on follow-up work with other data-likelihoods (e.g., for $\mathcal{GP}$-based classification), and envision methods that leverage the data-compression benefits of CVGP's coreset-based posterior.

Acknowledgements

Mert Ketenci acknowledges this research is supported by NHLBI award R01HL148248. Iñigo Urteaga acknowledges the support of "la Caixa" foundation's LCF/BQ/PI22/11910028 award, as well as funds by MICIU/AEI/10.13039/501100011033 and the BERC 2022-2025 program funded by the Basque Government.

References

Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

Matthias Bauer, Mark van der Wilk, and Carl Edward Rasmussen. Understanding Probabilistic Sparse Gaussian Process Approximations. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

T. Bertin-Mahieux. Year Prediction MSD. UCI Machine Learning Repository, 2011. DOI: https://doi.org/10.24432/C50K61.

Thang D. Bui, Josiah Yan, and Richard E. Turner. A Unifying Framework for Gaussian Process Pseudo-Point Approximations using Power Expectation Propagation. *Journal of Machine Learning Research*, 18(104):1–72, 2017. ISSN 1533-7928. URL `http://jmlr.org/papers/v18/16-603.html`.

Trevor Campbell and Boyan Beronov. Sparse variational inference: Bayesian coresets from scratch. *Advances in Neural Information Processing Systems*, 32, 2019.

Trevor Campbell and Tamara Broderick. Bayesian coreset construction via greedy iterative geodesic ascent. In *International Conference on Machine Learning*, pages 698–706. PMLR, 2018.

Trevor Campbell and Tamara Broderick. Automated scalable bayesian inference via hilbert coresets. *The Journal of Machine Learning Research*, 20(1):551–588, 2019.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553, 2009.

Lehel Csató and Manfred Opper. Sparse on-line gaussian processes. *Neural computation*, 14(3):641–668, 2002.

Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pages 1–15, 2013. ISSN 2192-6352. doi: 10.1007/s13748-013-0040-3. URL `[WebLink]`.

Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in neural information processing systems*, 31, 2018.

F Graf, HP Kriegel, M Schubert, S Poelsterl, and A Cavallaro. Relative location of ct slices on axial axis data set. *UCI Mach. Learn. Repository*, 2011.

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *Uncertainty in Artificial Intelligence*, 2013.

Jonathan Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable bayesian logistic regression. *Advances in Neural Information Processing Systems*, 29, 2016.

Martin Jankowiak, Geoff Pleiss, and Jacob Gardner. Parametric gaussian process regressors. In *International Conference on Machine Learning*, pages 4702–4712. PMLR, 2020.

Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. UCI Machine Learning Repository: Heart Disease Data Set. `https://archive.ics.uci.edu/ml/datasets/heart+disease`. (Accessed on 07/08/2021).

Ibrahim Jubran, Alaa Maalouf, and Dan Feldman. Introduction to Coresets: Accurate Coresets, October 2019.

Sathiya Keerthi and Wei Chu. A matching pursuit approach to sparse Gaussian process regression. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *International Conference for Learning Representations*, 2014.

Max Little, Patrick Mcsharry, Stephen Roberts, Declan Costello, and Irene Moroz. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Nature Precedings*, pages 1–1, 2007.

Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. When gaussian process meets big data: A review of scalable gps. *IEEE transactions on neural networks and learning systems*, 31(11):4405–4423, 2020.

Dionysis Manousakas, Zuheng Xu, Cecilia Mascolo, and Trevor Campbell. Bayesian pseudocoresets. *Advances in Neural Information Processing Systems*, 33:14950–14960, 2020.

Dionysis Manousakas, Hippolyt Ritter, and Theofanis Karaletsos. Black-box coreset variational inference. *Advances in Neural Information Processing Systems*, 35: 34175–34187, 2022.

Andrew Naish-Guzman and Sean Holden. The generalized fitc approximation. *Advances in neural information processing systems*, 20, 2007.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Joaquin Quinonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

Prashant S Rana. Physicochemical properties of protein tertiary structure data set. *UCI Machine Learning Repository*, 2013.

Carl Edward Rasmussen, Christopher KI Williams, et al. *Gaussian processes for machine learning*, volume 1. Springer, 2006.

Matthias W Seeger, Christopher KI Williams, and Neil D Lawrence. Fast forward selection to speed up sparse gaussian process regression. In *International Workshop on Artificial Intelligence and Statistics*, pages 254–261. PMLR, 2003.

Alex Smola and Peter Bartlett. Sparse Greedy Gaussian Process Regression. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18, 2005.

Edward Lloyd Snelson. *Flexible and efficient Gaussian process models for machine learning*. University of London, University College London (United Kingdom), 2008.

Joseph J Thompson, Mark R Blair, Lihan Chen, and Andrew J Henrey. Video game telemetry as a critical tool in the study of complex skill learning. *PloS one*, 8(9): e75129, 2013.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR, 2009.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

Andrew Wilson and Hannes Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *International conference on machine learning*, pages 1775–1784. PMLR, 2015.

Yuan, Qi, Ahmed H. Abdel-Gawad, and Thomas P. Minka. Sparse-posterior Gaussian Processes for general likelihoods. *arXiv preprint arXiv:1203.3507*, 2012. doi: 10.48550/arXiv.1203.3507. URL `http://arxiv.org/abs/1203.3507`.

Jacky Zhang, Rajiv Khanna, Anastasios Kyrillidis, and Sanmi Koyejo. Bayesian coresets: Revisiting the nonconvex optimization perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 2782–2790. PMLR, 2021.

Supplementary Material: Accurate and Scalable Stochastic Gaussian Process Regression via Learnable Coreset-based Variational Inference

Mert Ketenci[1] **Adler Perotte**[2] **Noémie Elhadad**[1,2] **Iñigo Urteaga**[3,4]

[1]Department of Computer Science, Columbia University , New York, USA
[2]Department of Biomedical Informatics, Columbia University , New York, USA
[3]BCAM — Basque Center for Applied Mathematics, Bilbao, Spain
[4]IKERBASQUE — Basque Foundation for Science, Bilbao, Spain

A CVGP DERIVATION DETAILS

We derive CVGP's coreset-based posterior and the log-marginal likelihood's variational lower-bound, first from the function-space view of $\mathcal{GP}$s in Section A.1, and then from the complementary weight-space view in Section A.2. Independently of the route taken, the attained variational lower-bounds are identical, yet the weight- and function-space coreset-based variational posteriors enable complementary understanding of CVGP's inference procedure and posterior distribution.

A.1 FUNCTION-SPACE DERIVATION OF CVGP

We derive below, under the assumption of standard Gaussian, uncorrelated observation noise, i.e., $\mathbf{y} = \mathbf{f} + \epsilon$, $\epsilon \sim \mathcal{N}\left(\epsilon \mid \mathbf{0}, \sigma^2 \mathbf{I}_N\right)$, the coreset-based tempered posterior over $\mathcal{GP}$ coreset function values $q(\mathbf{f}_M)$. The derivations are equivalent for non-zero mean and/or correlated noise functions.

A.1.1 CVGP's Coreset-based Posterior

To be able to accurately approximate the full $\mathcal{GP}$ posterior with a coreset $\{\mathbf{X}_M, \mathbf{y}_M\}$, we propose to weight with $\beta_c \geq 0$ [1] the likelihood of each psuedo-point when computing their corresponding coreset $\mathcal{GP}$ value $\mathbf{f}_M$, i.e.,

$$q(\mathbf{f}_M \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M) = \frac{q(\mathbf{y}_M \mid \mathbf{f}_M, \boldsymbol{\beta}_M) p(\mathbf{f}_M \mid \mathbf{X}_M)}{q(\mathbf{y}_M \mid \mathbf{X}_M, \boldsymbol{\beta}_M)} = \frac{\left(\prod_{m=1}^{M} p(y_m \mid f_m)^{\beta_m}\right) p(\mathbf{f}_M \mid \mathbf{X}_M)}{q(\mathbf{y}_M \mid \mathbf{X}_M, \boldsymbol{\beta}_M)} \,. \tag{15}$$

We start by deriving a closed form expression for the $\boldsymbol{\beta}_M$-weighted likelihood function $q(\mathbf{y}_M | \mathbf{f}_M, \boldsymbol{\beta}_M)$, by considering each coreset pair $\{\mathbf{x}_m, y_m\}$, for $m = 1, \cdots, M$, independently:

[1]In practice, we ensure positive β_m using the $\mathrm{softplus}(.)$ function.

$$
\begin{aligned}
q(y_m \mid f_m, \beta_m) &= p(y_m \mid f_m)^{\beta_m} = \mathcal{N}\left(y_m \mid f_m, \sigma^2\right)^{\beta_m} && (16)\\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(y_m - f_m)\sigma^{-2}(y_m - f_m)}\right)^{\beta_m} && (17)\\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\beta_m} e^{-\frac{1}{2}(y_m - f_m)(\beta_m^{-1}\sigma^2)^{-1}(y_m - f_m)} && (18)\\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\beta_m} \left(\frac{\sqrt{2\pi\beta_m^{-1}\sigma^2}}{\sqrt{2\pi\beta_m^{-1}\sigma^2}}\right) \exp\left\{-\frac{1}{2}(y_m - f_m)(\beta_m^{-1}\sigma^2)^{-1}(y_m - f_m)\right\} && (19)\\
&= \frac{\sqrt{2\pi\beta_m^{-1}\sigma^2}}{\left(\sqrt{2\pi\sigma^2}\right)^{\beta_m}} \left(\frac{1}{\sqrt{2\pi\beta_m^{-1}\sigma^2}}\right) \exp\left\{-\frac{1}{2}(y_m - f_m)(\beta_m^{-1}\sigma^2)^{-1}(y_m - f_m)\right\} && (20)\\
&= \frac{\sqrt{2\pi\beta_m^{-1}\sigma^2}}{\left(\sqrt{2\pi\sigma^2}\right)^{\beta_m}} \cdot \mathcal{N}\left(y_m | f_m, \beta_m^{-1}\sigma^2\right) && (21)\\
&= Q_c \cdot \mathcal{N}\left(y_m | f_m, \beta_m^{-1}\sigma^2\right) \ , \text{ with } Q_c = \frac{\sqrt{2\pi\beta_m^{-1}\sigma^2}}{\left(\sqrt{2\pi\sigma^2}\right)^{\beta_m}} \ . && (22)
\end{aligned}
$$

We write the joint over the full coreset pseudo-observations as a product over each likelihood term:

$$
\begin{aligned}
q(\mathbf{y}_M | \mathbf{f}_M, \boldsymbol{\beta}_M) &= \prod_{m=1}^{M} p(y_m | f_m)^{\beta_m} && (23)\\
&= \prod_{m=1}^{M} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\beta_m} \left(\frac{\sqrt{2\pi\beta_m^{-1}\sigma^2}}{\sqrt{2\pi\beta_m^{-1}\sigma^2}}\right) \exp\left\{-\frac{1}{2}(y_m - f_m)(\beta_m^{-1}\sigma^2)^{-1}(y_m - f_m)\right\} && (24)\\
&= \prod_{m=1}^{M} Q_m \cdot \mathcal{N}\left(y_m | f_m, \beta_m^{-1}\sigma^2\right) && (25)\\
&= Q_M \cdot \mathcal{N}\left(\mathbf{y}_M \mid \mathbf{f}_M, \Sigma_{\boldsymbol{\beta}_M}\right) \ , \text{ with } \begin{cases} Q_M = \prod_{m=1}^{M} \frac{\sqrt{2\pi\beta_m^{-1}\sigma^2}}{\left(\sqrt{2\pi\sigma^2}\right)^{\beta_m}} \\ \Sigma_{\boldsymbol{\beta}_M} = \sigma^2 \cdot \operatorname{diag}\left\{\boldsymbol{\beta}_M^{-1}\right\}. \end{cases} && (26)
\end{aligned}
$$

We derive the marginalized pseudo-observation coreset distribution

$$
\begin{aligned}
q(\mathbf{y}_M | \mathbf{X}_M, \boldsymbol{\beta}_M) &= \int_{\mathbf{f}_M} q(\mathbf{y}_M, \mathbf{f}_M | \mathbf{X}_M, \boldsymbol{\beta}_M) \, \mathrm{d}\mathbf{f}_M = \int_{\mathbf{f}_M} q(\mathbf{y}_M | \mathbf{f}_M, \boldsymbol{\beta}_M) p(\mathbf{f}_M | \mathbf{X}_M) \, \mathrm{d}\mathbf{f}_M && (27)\\
&= \int_{\mathbf{f}_M} Q_M \cdot \mathcal{N}\left(\mathbf{y}_M \mid \mathbf{f}_M, \Sigma_{\boldsymbol{\beta}_M}\right) \cdot \mathcal{N}\left(\mathbf{f}_M \mid \mathbf{0}, \mathbf{K}_{MM}\right) \mathrm{d}\mathbf{f}_M && (28)\\
&= Q_M \int_{\mathbf{f}_M} \mathcal{N}\left(\mathbf{y}_M \mid \mathbf{f}_M, \Sigma_{\boldsymbol{\beta}_M}\right) \cdot \mathcal{N}\left(\mathbf{f}_M \mid \mathbf{0}, \mathbf{K}_{MM}\right) \mathrm{d}\mathbf{f}_M && (29)\\
&= Q_M \cdot \mathcal{N}\left(\mathbf{y}_M \mid \mathbf{0}, \mathbf{K}_{MM} + \Sigma_{\boldsymbol{\beta}_M}\right) \ . && (30)
\end{aligned}
$$

We leverage the above distributions to derive the coreset-based, tempered $\mathcal{GP}$ posterior

$$
\begin{aligned}
q(\mathbf{f}_M|\mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M) &= \frac{q(\mathbf{y}_M|\mathbf{f}_M, \boldsymbol{\beta}_M) q(\mathbf{f}_M|\mathbf{X}_M)}{q(\mathbf{y}_M|\mathbf{X}_M, \boldsymbol{\beta}_M)} && (31)\\
&= \frac{Q_M \cdot \mathcal{N}\left(\mathbf{y}_M \mid \mathbf{f}_M, \Sigma_{\boldsymbol{\beta}_M}\right) \mathcal{N}\left(\mathbf{f}_M \mid \mathbf{0}, \mathbf{K}_{MM}\right)}{Q_M \cdot \mathcal{N}\left(\mathbf{y}_M \mid \mathbf{0}, \mathbf{K}_{MM} + \Sigma_{\boldsymbol{\beta}_M}\right)} && (32)\\
&= \frac{\mathcal{N}\left(\mathbf{y}_M \mid \mathbf{f}_M, \Sigma_{\boldsymbol{\beta}_M}\right) \mathcal{N}\left(\mathbf{f}_M \mid \mathbf{0}, \mathbf{K}_{MM}\right)}{\mathcal{N}\left(\mathbf{y}_M \mid \mathbf{0}, \mathbf{K}_{MM} + \Sigma_{\boldsymbol{\beta}_M}\right)} && (33)\\
&= \mathcal{N}\left(\mathbf{f}_M \mid \mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}, \mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\right) \ , \text{ with } \begin{cases} \mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M} = \mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\left(\Sigma_{\boldsymbol{\beta}_M}^{-1}\mathbf{y}_M\right) \\ \mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M} = \left(\mathbf{K}_{MM}^{-1} + \Sigma_{\boldsymbol{\beta}_M}^{-1}\right)^{-1} \end{cases}. && (34)
\end{aligned}
$$

The sufficient statistics of the coreset-based, tempered distibution above can be rewritten as

$$
\begin{aligned}
q(\mathbf{f}_M) = q(\mathbf{f}_M|\mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M) &= \mathcal{N}\left(\mathbf{f}_M \mid \mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}, \mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\right) \ , && (35)\\
&\text{with} \begin{cases} \mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M} = \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \Sigma_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{y}_M \\ \mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M} = \mathbf{K}_{MM} - \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \Sigma_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM} \\ \qquad \text{by Woodbury matrix identity .} \end{cases} && (36)
\end{aligned}
$$

The coreset-based posterior over $\mathcal{GP}$ function values. We now compute the posterior over $\mathcal{GP}$ values for any given data point $\mathbf{X}$, by marginalizing the $\mathcal{GP}$'s prior-conditional over the coreset-based distribution, i.e.,

$$
q(\mathbf{f}|\mathbf{X}_M, \mathbf{y}_M) = \int_{\mathbf{f}_M} p(\mathbf{f}|\mathbf{f}_M) q(\mathbf{f}_M|\mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M) \, \mathrm{d}\mathbf{f}_M \ . \tag{37}
$$

The above is analytically solvable due to all the distributions being Gaussian:

$$
\begin{aligned}
q(\mathbf{f}_M|\mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M) &= \mathcal{N}\left(\mathbf{f}_M \mid \mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}, \mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\right) \ , && (38)\\
&\text{with} \begin{cases} \mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M} = \mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\left(\Sigma_{\boldsymbol{\beta}_M}^{-1}\mathbf{y}_M\right) \\ \mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M} = \mathbf{K}_{MM} - \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \Sigma_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}, \end{cases} && (39)\\
p(\mathbf{f}|\mathbf{f}_M) &= \mathcal{N}\left(\mathbf{f} \mid \mathbf{m}_{\mathbf{f}|\mathbf{f}_M}, \mathbf{K}_{\mathbf{f}|\mathbf{f}_M}\right) \ , && (40)\\
&\text{with} \begin{cases} \mathbf{m}_{\mathbf{f}|\mathbf{f}_M} = \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{f}_M \\ \mathbf{K}_{\mathbf{f}|\mathbf{f}_M} = \mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}, \end{cases} && (41)\\
q(\mathbf{f}) = q(\mathbf{f}|\mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M) &= \mathcal{N}\left(\mathbf{f} \mid \mathbf{m}_{\mathbf{f}|\mathbf{y}_M}, \mathbf{K}_{\mathbf{f}|\mathbf{y}_M}\right) \ , && (42)\\
&\text{with} \begin{cases} \mathbf{m}_{\mathbf{f}|\mathbf{y}_M} = \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M} \\ \mathbf{K}_{\mathbf{f}|\mathbf{y}_M} = \mathbf{K}_{\mathbf{f}|\mathbf{f}_M} + \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}. \end{cases} && (43)
\end{aligned}
$$

We elaborate on the sufficient statistics of $q(\mathbf{f}|\mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M)$.

First, we rewrite the expected value as

$$
\begin{aligned}
\mathbf{m}_{\mathbf{f}|\mathbf{y}_M} &= \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M} && (44)\\
&= \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\Sigma_{\boldsymbol{\beta}_M}^{-1}\mathbf{y}_M && (45)\\
&= \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\left(\mathbf{K}_{MM}^{-1} + \Sigma_{\boldsymbol{\beta}_M}^{-1}\right)^{-1}\Sigma_{\boldsymbol{\beta}_M}^{-1}\mathbf{y}_M && (46)\\
&\qquad \text{by using equivalence in Equation (48)} \\
&= \mathbf{K}_{NM}\left(\mathbf{K}_{MM} + \Sigma_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{y}_M \ , && (47)
\end{aligned}
$$

where we have made use of the following equivalences,

$$\mathbf{K}_{MM}^{-1}\left(\mathbf{K}_{MM}^{-1}+\Sigma_{\boldsymbol{\beta}_M}^{-1}\right)^{-1}\Sigma_{\boldsymbol{\beta}_M}^{-1}=\mathbf{K}_{MM}^{-1}\left(\Sigma_{\boldsymbol{\beta}_M}\mathbf{K}_{MM}^{-1}+\mathbf{I}_M\right)^{-1}=\left(\Sigma_{\boldsymbol{\beta}_M}+\mathbf{K}_{MM}\right)^{-1}\ , \tag{48}$$

$$\Sigma_{\boldsymbol{\beta}_M}^{-1}\left(\mathbf{K}_{MM}^{-1}+\Sigma_{\boldsymbol{\beta}_M}^{-1}\right)^{-1}\mathbf{K}_{MM}^{-1}=\Sigma_{\boldsymbol{\beta}_M}^{-1}\left(\mathbf{I}_M+\mathbf{K}_{MM}\Sigma_{\boldsymbol{\beta}_M}^{-1}\right)^{-1}=\left(\Sigma_{\boldsymbol{\beta}_M}+\mathbf{K}_{MM}\right)^{-1}\ . \tag{49}$$

Second, for the covariance matrix, we write

$$\begin{aligned}
\mathbf{K}_{\mathbf{f}|\mathbf{y}_M} &= \mathbf{K}_{\mathbf{f}|\mathbf{f}_M}+\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN} && (50)\\
&= \mathbf{K}_{NN}-\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}+\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\left(\mathbf{K}_{MM}^{-1}+\Sigma_{\boldsymbol{\beta}_M}^{-1}\right)^{-1}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN} && (51)\\
&\qquad \text{by using the Woodbury matrix identity for } \left(\mathbf{K}_{MM}^{-1}+\Sigma_{\boldsymbol{\beta}_M}^{-1}\right)^{-1}\\
&= \mathbf{K}_{NN}-\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}+\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\left(\mathbf{K}_{MM}-\mathbf{K}_{MM}\left(\mathbf{K}_{MM}+\Sigma_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\right)\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN} && (52)\\
&= \mathbf{K}_{NN}-\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}+\mathbf{K}_{NM}\left(\mathbf{I}_M-\left(\mathbf{K}_{MM}+\Sigma_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\right)\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN} && (53)\\
&= \mathbf{K}_{NN}-\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}+\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}-\mathbf{K}_{NM}\left(\mathbf{K}_{MM}+\Sigma_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN} && (54)\\
&= \mathbf{K}_{NN}-\mathbf{K}_{NM}\left(\mathbf{K}_{MM}+\Sigma_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MN}\ . && (55)
\end{aligned}$$

A.1.2 CVGP's Variational Lower-bound

We derive the variational lower-bound by writing everything in terms of sufficient statistics of $q(\mathbf{f}_M)$:

$$
\begin{aligned}
\mathcal{L}_{CVGP} &= \mathbb{E}_{q(\mathbf{f})}\left\{\log p(\mathbf{y}|\mathbf{f})\right\} - \mathrm{KL}\left[q(\mathbf{f}_M)\|p(\mathbf{f}_M)\right] && (56)\\
&= \log \mathcal{N}\left(\mathbf{y}|\mathbf{m}_{\mathbf{f}|\mathbf{y}_M}, \sigma^2\mathbf{I}_N\right) - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{\mathbf{f}|\mathbf{y}_M}\right\} \\
&\quad - \frac{1}{2}\left(\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\right\} - M + \mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}^\top \mathbf{K}_{MM}^{-1}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M} + \log\frac{|\mathbf{K}_{MM}|}{|\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}|}\right) && (57)\\
&= \log \mathcal{N}\left(\mathbf{y}|\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}, \sigma^2\mathbf{I}_N\right) \\
&\quad - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN} + \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} \\
&\quad - \frac{1}{2}\left(\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\right\} - M + \mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}^\top \mathbf{K}_{MM}^{-1}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M} + \log\frac{|\mathbf{K}_{MM}|}{|\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}|}\right) && (58)\\
&= \left(-\frac{N}{2}\log(2\pi) - \frac{1}{2}\log\left|\sigma^2\mathbf{I}_N\right| - \frac{1}{2}(\mathbf{y} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M})^\top\sigma^{-2}\mathbf{I}_N(\mathbf{y} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M})\right) \\
&\quad - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} \\
&\quad - \frac{1}{2}\left(\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\right\} - M + \mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}^\top \mathbf{K}_{MM}^{-1}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M} + \log\frac{|\mathbf{K}_{MM}|}{|\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}|}\right) && (59)\\
&= -\frac{N}{2}\log(2\pi) + \frac{M}{2} - \frac{1}{2}\log\left|\sigma^2\mathbf{I}_N\right| - \frac{1}{2}\log|\mathbf{K}_{MM}| \\
&- \frac{1}{2}\mathbf{y}^\top\sigma^{-2}\mathbf{y} + \sigma^{-2}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}^\top\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\mathbf{y} && (60)\\
&- \frac{1}{2}\sigma^{-2}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}^\top\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M} - \frac{1}{2}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}^\top\mathbf{K}_{MM}^{-1}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M} \\
&- \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} \\
&- \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\right\} - \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\right\} + \frac{1}{2}\log\left|\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\right| && (61)\\
&= -\frac{N}{2}\log(2\pi) + \frac{M}{2} - \frac{1}{2}\log\left|\sigma^2\mathbf{I}_N\right| - \frac{1}{2}\log\mathbf{K}_{MM} - \frac{1}{2}\mathbf{y}^\top\sigma^{-2}\mathbf{y} \\
&+ \sigma^{-2}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}^\top\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\mathbf{y} - \frac{1}{2}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}^\top\left(\mathbf{K}_{MM}^{-1} + \sigma^{-2}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\right)\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M} \\
&- \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} \\
&- \frac{1}{2}\mathrm{tr}\left\{\left(\mathbf{K}_{MM}^{-1} + \sigma^{-2}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\right)\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\right\} + \frac{1}{2}\log\left|\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\right| && (62)\\
&= -\frac{N}{2}\log(2\pi) + \frac{M}{2} - \frac{1}{2}\log\left|\sigma^2\mathbf{I}_N\right| - \frac{1}{2}\log\mathbf{K}_{MM} - \frac{1}{2}\mathbf{y}^\top\sigma^{-2}\mathbf{y} \\
&+ \sigma^{-2}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}^\top\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\mathbf{y} - \frac{1}{2}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}^\top\mathbf{K}_{MM}^{-1}\left(\mathbf{K}_{MM}^{-1} - \left(\mathbf{K}_{MM} - \Sigma_{\boldsymbol{\beta}_M}\right)^{-1}\right)^{-1}\mathbf{K}_{MM}^{-1}\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M} \\
&- \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} \\
&- \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\left(\mathbf{K}_{MM}^{-1} - \left(\mathbf{K}_{MM} - \Sigma_{\boldsymbol{\beta}_M}\right)^{-1}\right)^{-1}\mathbf{K}_{MM}^{-1}\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\right\} + \frac{1}{2}\log\left|\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}\right| && (63)\\
&= \log \mathcal{N}\left(\mathbf{y}|\mathbf{m}_{\mathbf{f}|\mathbf{y}_M}, \sigma^2\mathbf{I}_N\right) - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{\mathbf{f}|\mathbf{y}_M}\right\} \\
&- \frac{1}{2}\left[-\mathrm{tr}\left\{\mathbf{A}\mathbf{K}_{MM}\right\} + \mathbf{y}_M^\top\mathbf{A}\mathbf{K}_{MM}\mathbf{A}\mathbf{y}_M - \ln|\mathbf{A}| - \ln\left|\boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}\right|\right] , && (64)
\end{aligned}
$$

where $\mathbf{A} = \left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}$.

A.2 WEIGHT-SPACE DERIVATION OF CVGP

For completeness and a complementary perspective, we derive CVGP inference from the weight-space view of $\mathcal{GP}$s, again under the assumption of standard Gaussian, uncorrelated observation noise, i.e., $\mathbf{y} = \mathbf{f} + \epsilon$, $\epsilon \sim \mathcal{N}\left(\epsilon \mid \mathbf{0}, \sigma^2 \mathbf{I}_N\right)$.

Recall the weight-space definition of $\mathcal{GP}$s, [Rasmussen et al., 2006]:

$$y_i \mid \mathbf{w}, \mathbf{x}_i \sim \mathcal{N}\left(\mathbf{y} \mid \Phi(\mathbf{x}_i)^\top \mathbf{w}, \sigma^2 \mathbf{I}_N\right) \quad \text{with } \mathbf{w} \sim \mathcal{N}\left(\mathbf{w} \mid \mathbf{0}, \mathbf{I}_D\right) \ , \tag{65}$$

where $\Phi(\cdot) : \mathcal{X} \to \mathcal{H}$ is a feature map with associated kernel $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to I\!R$ and Hilbert space $\mathcal{H}$. Namely, a $\mathcal{GP}$ can be viewed as a Bayesian linear regression where the covariates $\mathcal{X}$ are embedded into a potentially infinite dimensional Hilbert space $\mathcal{H}$. An advantage of the weight-space view is that it allows for conditional independence between different data points $\mathbf{x}$ given $\mathbf{w}$; the key property we leverage in the following derivations.

A.2.1 CVGP's coreset-based tempered-posterior

We aim for a small subset of psuedo-points $\{\mathbf{X}_M, \mathbf{y}_M\}$ that, if drawn $\beta_m \geq 0$ times, approximate the true weight posterior:

$$\underbrace{\frac{p(\mathbf{w}) \prod_{m=1}^{M} p(y_c \mid \mathbf{w}, \mathbf{x}_m)^{\beta_m}}{Z_q}}_{q(\mathbf{w}|\mathbf{y}_M, \mathbf{X}_M, \boldsymbol{\beta}_M)} \approx \underbrace{\frac{p(\mathbf{w}) \prod_{i=1}^{N} p(y_i \mid \mathbf{w}, \mathbf{x}_i)}{Z_p}}_{p(\mathbf{w}|\mathbf{y}, \mathbf{X})} \ . \tag{66}$$

Typically, the objective of the coreset problem is to learn vector $\beta^\star = \operatorname{argmin}_\beta \operatorname{Dist}\left(q(\mathbf{w}), p(\mathbf{w})\right)$, where $\operatorname{Dist}(.)$ is a distance metric such as the KL divergence [Campbell and Beronov, 2019].

We derive the coreset-based tempered posterior over $\mathcal{GP}$ weights $q(\mathbf{w} \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M)$, by noting it is proportional to

$$\exp\left\{-\frac{1}{2}\mathbf{w}^\top \mathbf{w}\right\} \prod_{m=1}^{M} \exp\left\{-\frac{1}{2}\sigma^{-2}\beta_m \left(y_c - \Phi\left(\mathbf{x}_m\right)^\top \mathbf{w}\right)^2\right\} \tag{67}$$

$$\propto \exp\left\{-\frac{1}{2}\left(-2\sigma^{-2}\mathbf{w}^\top \sum_{m=1}^{M} y_c \beta_m \Phi\left(\mathbf{x}_m\right) + \mathbf{w}^\top \underbrace{\left(\sigma^{-2} \sum_{m=1}^{M} \Phi\left(\mathbf{x}_m\right) \beta_m \Phi\left(\mathbf{x}_m\right)^\top + \mathbf{I}_D\right)}_{\mathbf{S}_{\mathbf{w}|\mathbf{y}_M}^{-1}} \mathbf{w}\right)\right\}, \tag{68}$$

which is a Gaussian distribution with covariance matrix:

$$\mathbf{S}_{\mathbf{w}|\mathbf{y}_M} = \left(\sigma^{-2} \sum_{m=1}^{M} \Phi\left(\mathbf{x}_m\right) \beta_m \Phi\left(\mathbf{x}_m\right)^\top + \mathbf{I}_D\right)^{-1} = \left(\Phi\left(\mathbf{X}_M\right)^\top \Sigma_{\boldsymbol{\beta}_M}^{-1} \Phi\left(\mathbf{X}_M\right) + \mathbf{I}_D\right)^{-1} \ , \tag{69}$$

with $\Sigma_{\boldsymbol{\beta}_M} = \sigma^2 \cdot \operatorname{diag}\left\{\boldsymbol{\beta}_M^{-1}\right\}$. Let us define $\Sigma_{\boldsymbol{\beta}_M}^{-1} = C^{1/2} C^{1/2}$, with $\Phi\left(\mathbf{X}_M\right)^\top C^{1/2} = \Phi\left(\mathbf{X}_M\right)'^\top$ and $C^{1/2}\Phi\left(\mathbf{X}_M\right) = \Phi\left(\mathbf{X}_M\right)'$. Then we can write the covariance matrix as

$$\mathbf{S}_{\mathbf{w}|\mathbf{y}_M} = \left(\Phi\left(\mathbf{X}_M\right)'^\top \Phi\left(\mathbf{X}_M\right)' + \mathbf{I}_D\right)^{-1} \tag{70}$$

$$= \mathbf{I}_D - \Phi\left(\mathbf{X}_M\right)'^\top \left(\mathbf{I}_D + \Phi\left(\mathbf{X}_M\right)' \Phi\left(\mathbf{X}_M\right)'^\top\right)^{-1} \Phi\left(\mathbf{X}_M\right)' \tag{71}$$

$$= \mathbf{I}_D - \Phi\left(\mathbf{X}_M\right)^\top \left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \Phi\left(\mathbf{X}_M\right) \ . \tag{72}$$

We revisit Equation (68) to identify the mean of the Gaussian distribution as follows, where we use $\mathbf{y}_M{}' = C^{1/2}\mathbf{y}_M$:

$$
\begin{aligned}
\mathbf{m}_{\mathbf{w}|\mathbf{y}_M} &= \mathbf{S}_{\mathbf{w}|\mathbf{y}_M}\left(\sigma^{-2}\sum_{m=1}^{M} y_c\beta_m\Phi\left(\mathbf{x}_m\right)\right) && (73)\\
&= \mathbf{S}_{\mathbf{w}|\mathbf{y}_M}\left(\Phi\left(\mathbf{X}_M\right)^\top\Sigma_{\boldsymbol{\beta}_M}^{-1}\mathbf{y}_M\right) && (74)\\
&= \mathbf{S}_{\mathbf{w}|\mathbf{y}_M}\left(\Phi\left(\mathbf{X}_M\right)^\top C^{1/2}C^{1/2}\mathbf{y}_M\right) && (75)\\
&= \mathbf{S}_{\mathbf{w}|\mathbf{y}_M}\left(\Phi\left(\mathbf{X}_M\right)'^\top\mathbf{y}_M{}'\right) && (76)\\
&= \left(\mathbf{I}_D - \Phi\left(\mathbf{X}_M\right)'^\top\left(\mathbf{I}_D + \Phi\left(\mathbf{X}_M\right)'\Phi\left(\mathbf{X}_M\right)'^\top\right)^{-1}\Phi\left(\mathbf{X}_M\right)'\right)\left(\Phi\left(\mathbf{X}_M\right)'^\top\mathbf{y}_M{}'\right) && (77)\\
&= \left(\Phi\left(\mathbf{X}_M\right)'^\top - \Phi\left(\mathbf{X}_M\right)'^\top\left(\mathbf{I}_D + \Phi\left(\mathbf{X}_M\right)'\Phi\left(\mathbf{X}_M\right)'^\top\right)^{-1}\Phi\left(\mathbf{X}_M\right)'\Phi\left(\mathbf{X}_M\right)'^\top\right)\mathbf{y}_M{}' && (78)\\
&= \left(\Phi\left(\mathbf{X}_M\right)'^\top - \Phi\left(\mathbf{X}_M\right)'^\top\left(\left(\Phi\left(\mathbf{X}_M\right)'\Phi\left(\mathbf{X}_M\right)'^\top\right)^{-1} + \mathbf{I}_D\right)^{-1}\right)\mathbf{y}_M{}' && (79)\\
&= \Phi\left(\mathbf{X}_M\right)'^\top\left(\mathbf{I}_D - \left(\left(\Phi\left(\mathbf{X}_M\right)'\Phi\left(\mathbf{X}_M\right)'^\top\right)^{-1} + \mathbf{I}_D\right)^{-1}\right)\mathbf{y}_M{}' && (80)\\
&= \Phi\left(\mathbf{X}_M\right)'^\top\left(\mathbf{I}_D - \Phi\left(\mathbf{X}_M\right)'\Phi\left(\mathbf{X}_M\right)'^\top\left(\underbrace{\mathbf{I}_D + \Phi\left(\mathbf{X}_M\right)'\Phi\left(\mathbf{X}_M\right)'^\top}_{D}\right)^{-1}\right)\mathbf{y}_M{}' && (81)\\
&= \Phi\left(\mathbf{X}_M\right)'^\top\left(DD^{-1} - \Phi\left(\mathbf{X}_M\right)'\Phi\left(\mathbf{X}_M\right)'^\top D^{-1}\right)\mathbf{y}_M{}' && (82)\\
&= \Phi\left(\mathbf{X}_M\right)'^\top\left(\cancel{D} - \cancel{\Phi\left(\mathbf{X}_M\right)'\Phi\left(\mathbf{X}_M\right)'^\top}\right)D^{-1}\mathbf{y}_M{}' && (83)\\
&= \Phi\left(\mathbf{X}_M\right)'^\top D^{-1}\mathbf{y}_M{}' && (84)\\
&= \Phi\left(\mathbf{X}_M\right)^\top C^{1/2}\left(\mathbf{I}_D + C^{1/2}\Phi\left(\mathbf{X}_M\right)\Phi\left(\mathbf{X}_M\right)^\top C^{1/2}\right)^{-1}C^{1/2}\mathbf{y}_M && (85)\\
&= \Phi\left(\mathbf{X}_M\right)^\top\left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1}\mathbf{y}_M\ . && (86)
\end{aligned}
$$

All in all, we have

$$
q(\mathbf{w}|\mathbf{X}_M,\mathbf{y}_M,\boldsymbol{\beta}_M) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_{\mathbf{w}|\mathbf{y}_M},\mathbf{S}_{\mathbf{w}|\mathbf{y}_M}\right)\ ,\ \text{with}\begin{cases}\mathbf{m}_{\mathbf{w}|\mathbf{y}_M} = \Phi\left(\mathbf{X}_M\right)^\top\left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1}\mathbf{y}_M\\ \mathbf{S}_{\mathbf{w}|\mathbf{y}_M} = \mathbf{I}_D - \Phi\left(\mathbf{X}_M\right)^\top\left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1}\Phi\left(\mathbf{X}_M\right).\end{cases} \tag{87}
$$

A.2.2 CVGP's Weight-space Variational Lower-bound

We write the variational lower-bound of the log-marginal likelihood as:

$$
\begin{aligned}
\log p(\mathbf{y}\mid\mathbf{X}) &= \int q(\mathbf{w})\log\left\{\frac{p(\mathbf{y},\mathbf{w}\mid\mathbf{X})q(\mathbf{w})}{p(\mathbf{w}\mid\mathbf{X},\mathbf{y})q(\mathbf{w})}\right\}\mathrm{d}\mathbf{w} && (88)\\
&= \mathbb{E}_{q(\mathbf{w})}\left\{\log p(\mathbf{y}\mid\mathbf{w},\mathbf{X})\right\} - \mathrm{KL}\left[q(\mathbf{w})\|p(\mathbf{w})\right] + \mathrm{KL}\left[q(\mathbf{w})\|p(\mathbf{w}\mid\mathbf{X},\mathbf{y})\right] && (89)\\
&\geq \underbrace{\mathbb{E}_{q(\mathbf{w})}\left\{\log p(\mathbf{y}\mid\mathbf{w},\mathbf{X})\right\} - \mathrm{KL}\left[q(\mathbf{w})\|p(\mathbf{w})\right]}_{\mathcal{L}_{CVGP}}\ , && (90)
\end{aligned}
$$

which is the lower-bound of the weight-space view of CVGP, where we set $q(\mathbf{w}) = p(\mathbf{w} \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M)$ and derive

$$\mathcal{L}_{CVGP} = \mathbb{E}_{p(\mathbf{w}|\mathbf{X}_M,\mathbf{y}_M,\boldsymbol{\beta}_M)} \left\{\log p(\mathbf{y} \mid \mathbf{w}, \mathbf{X})\right\} - \mathrm{KL}\left[p(\mathbf{w} \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M) \| p(\mathbf{w})\right] \tag{91}$$

$$= \mathbb{E}_{p(\mathbf{w}|\mathbf{X}_M,\mathbf{y}_M,\boldsymbol{\beta}_M)} \left\{\sum_{i=1}^{N} \log p(y_i \mid \mathbf{w}, \mathbf{x}_i)\right\} - \mathrm{KL}\left[p(\mathbf{w} \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M) \| p(\mathbf{w})\right] \tag{92}$$

$$= \sum_{i=1}^{N} \underbrace{\mathbb{E}_{p(\mathbf{w}|\mathbf{X}_M,\mathbf{y}_M,\boldsymbol{\beta}_M)} \left\{\log p(y_i \mid \mathbf{w}, \mathbf{x}_i)\right\}}_{\ell_i} - \mathrm{KL}\left[p(\mathbf{w} \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M) \| p(\mathbf{w})\right]. \tag{93}$$

We compute below the analytical expressions for ℓ_i and $\mathrm{KL}\left[p(\mathbf{y} \mid \mathbf{w}, \mathbf{X}) \| p(\mathbf{w})\right]$.

We start with ℓ_i:

$$\ell_i = \int p(\mathbf{w} \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M) \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - \Phi\left(\mathbf{x}_i\right)^\top \mathbf{w}\right)^2\right\} \mathrm{d}\mathbf{w} \tag{94}$$

$$= -\frac{1}{2}\left(\log 2\pi\sigma^2 + \sigma^{-2} \int p(\mathbf{w} \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M)\left(y_i^2 - 2y_i \Phi\left(\mathbf{x}_i\right)^\top \mathbf{w} + \Phi\left(\mathbf{x}_i\right)^\top \mathbf{w}\mathbf{w}^\top \Phi\left(\mathbf{x}_i\right)\right) \mathrm{d}\mathbf{w}\right) \tag{95}$$

We need to compute $\int \mathbf{w} p(\mathbf{w} \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M)\, \mathrm{d}\mathbf{w}$ and $\int \mathbf{w}\mathbf{w}^\top p(\mathbf{w} \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M)\, \mathrm{d}\mathbf{w}$. Note that the former is $\mathbf{m}_{\mathbf{w}|\mathbf{y}_M}$ and the latter is $\mathbf{S}_{\mathbf{w}|\mathbf{y}_M} + \mathbf{m}_{\mathbf{w}|\mathbf{y}_M}\mathbf{m}_{\mathbf{w}|\mathbf{y}_M}^\top$. Hence,

$$\ell_i = -\frac{1}{2}\left(\log 2\pi\sigma^2 + \sigma^{-2}\left(y_i^2 - 2y_i \Phi\left(\mathbf{x}_i\right)^\top \mathbf{m}_{\mathbf{w}|\mathbf{y}_M} + \Phi\left(\mathbf{x}_i\right)^\top (\mathbf{S}_{\mathbf{w}|\mathbf{y}_M} + \mathbf{m}_{\mathbf{w}|\mathbf{y}_M}\mathbf{m}_{\mathbf{w}|\mathbf{y}_M}^\top)\Phi\left(\mathbf{x}_i\right)\right)\right) . \tag{96}$$

Let us now define

$$m_{f_i|\mathbf{y}_M} = \Phi\left(\mathbf{x}_i\right)^\top \mathbf{m}_{\mathbf{w}|\mathbf{y}_M} \tag{97}$$

$$= \Phi\left(\mathbf{x}_i\right)^\top \Phi\left(\mathbf{X}_M\right)^\top \left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \mathbf{y}_M \tag{98}$$

$$= \mathbf{k}_{iM}\left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \mathbf{y}_M , \tag{99}$$

and

$$k_{f_i|\mathbf{y}_M} = \Phi\left(\mathbf{x}_i\right)^\top \mathbf{S}_{\mathbf{w}|\mathbf{y}_M} \Phi\left(\mathbf{x}_i\right) \tag{100}$$

$$= k_{ii} - \mathbf{k}_{iM}\left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \mathbf{k}_{\mathbf{X}_M,\mathbf{x}_i} , \tag{101}$$

where the above relate to the $\mathcal{GP}$ function values via transformation of the weights by the feature vectors, i.e., $f_i = f(\mathbf{x}_i) = \Phi\left(\mathbf{x}_i\right)^\top \mathbf{w}$. Notice how the above expressions match those in Equation (47) and (55). We can therefore write

$$\ell_i = -\frac{1}{2}\left(\log 2\pi\sigma^2 + \sigma^{-2}\left(y_i^2 - 2y_i m_{f_i|\mathbf{y}_M} + k_{f_i|\mathbf{y}_M} + {m_{f_i|\mathbf{y}_M}}^2\right)\right) \tag{102}$$

$$= \log \mathcal{N}\left(y_i \mid m_{f_i|\mathbf{y}_M}, \sigma^2\right) \exp\left\{-\frac{1}{2}\sigma^{-2} k_{f_i|\mathbf{y}_M}\right\} . \tag{103}$$

We continue with the KL divergence term, recalling $p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w} \mid \mathbf{0}, \mathbf{I}_D\right)$, and write

$$\mathrm{KL}\left[p(\mathbf{w} \mid \mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M) \| p(\mathbf{w})\right] = \frac{1}{2}\left(\mathbf{m}_{\mathbf{w}|\mathbf{y}_M}^\top \mathbf{I}_D^{-1} \mathbf{m}_{\mathbf{w}|\mathbf{y}_M} + \mathrm{tr}\left\{\mathbf{I}_D^{-1}\mathbf{S}_{\mathbf{w}|\mathbf{y}_M}\right\} + \log \mid \mathbf{I}_D \mid - \log \mid \mathbf{S}_{\mathbf{w}|\mathbf{y}_M} \mid - \mathrm{tr}\left\{\mathbf{I}_D\right\}\right) \tag{104}$$

$$= \frac{1}{2}\left(\mathbf{m}_{\mathbf{w}|\mathbf{y}_M}^\top \mathbf{m}_{\mathbf{w}|\mathbf{y}_M} + \mathrm{tr}\left\{\mathbf{S}_{\mathbf{w}|\mathbf{y}_M}\right\} - \log \mid \mathbf{S}_{\mathbf{w}|\mathbf{y}_M} \mid - \mathrm{tr}\left\{\mathbf{I}_D\right\}\right) . \tag{105}$$

We first compute

$$\mathbf{m}_{\mathbf{w}|\mathbf{y}_M}^\top \mathbf{m}_{\mathbf{w}|\mathbf{y}_M} = \left(\Phi\left(\mathbf{X}_M\right)^\top \left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \mathbf{y}_M\right)^\top \left(\Phi\left(\mathbf{X}_M\right)^\top \left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \mathbf{y}_M\right) \tag{106}$$

$$= \mathbf{y}_M^\top \left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \Phi\left(\mathbf{X}_M\right) \Phi\left(\mathbf{X}_M\right)^\top \left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \mathbf{y}_M \tag{107}$$

$$= \mathbf{y}_M^\top \left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \mathbf{K}_{MM} \left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \mathbf{y}_M \ , \tag{108}$$

then,

$$\mathrm{tr}\left\{\mathbf{S}_{\mathbf{w}|\mathbf{y}_M}\right\} = \mathrm{tr}\left\{\mathbf{I}_D - \Phi\left(\mathbf{X}_M\right)^\top \left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \Phi\left(\mathbf{X}_M\right)\right\} \tag{109}$$

$$= \mathrm{tr}\left\{\mathbf{I}_D\right\} - \mathrm{tr}\left\{\Phi\left(\mathbf{X}_M\right)^\top \left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \Phi\left(\mathbf{X}_M\right)\right\} \tag{110}$$

$$= \mathrm{tr}\left\{\mathbf{I}_D\right\} - \mathrm{tr}\left\{\left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \Phi\left(\mathbf{X}_M\right) \Phi\left(\mathbf{X}_M\right)^\top\right\} \tag{111}$$

$$= \mathrm{tr}\left\{\mathbf{I}_D\right\} - \mathrm{tr}\left\{\left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \mathbf{K}_{MM}\right\} \ , \tag{112}$$

and finally,

$$\log \mid \mathbf{S}_{\mathbf{w}|\mathbf{y}_M} \mid = \log\left|\left(\Phi\left(\mathbf{X}_M\right)'^\top \Phi\left(\mathbf{X}_M\right)' + \mathbf{I}_D\right)^{-1}\right| \tag{113}$$

$$= -\log\left|\left(\Phi\left(\mathbf{X}_M\right)'^\top \Phi\left(\mathbf{X}_M\right)' + \mathbf{I}_D\right)\right| \tag{114}$$

$$= -\log\left|\left(\Phi\left(\mathbf{X}_M\right)^\top C \Phi\left(\mathbf{X}_M\right) + \mathbf{I}_D\right)\right| \tag{115}$$

$$= -\log\left|\Sigma_{\boldsymbol{\beta}_M} + \Phi\left(\mathbf{X}_M\right) \Phi\left(\mathbf{X}_M\right)^\top\right| \left|\Sigma_{\boldsymbol{\beta}_M}^{-1}\right| \cancel{|\mathbf{I}_D|} \tag{116}$$

using metrix determinant lemma (117)

$$= -\log\left|\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right| - \log\left|\Sigma_{\boldsymbol{\beta}_M}^{-1}\right| \ . \tag{118}$$

We put it all together for the analytical, weight-space variational lower-bound of CVGP,

$$\begin{aligned}
\mathcal{L}_{CVGP} = \sum_{i=1}^{N} & \left(\log \mathcal{N}\left(y_i \mid m_{f_i|\mathbf{y}_M}, \sigma^2\right) \exp\left\{-\frac{1}{2}\sigma^{-2} k_{f_i|\mathbf{y}_M}\right\}\right) \\
& - \frac{1}{2}\Big(+\mathbf{y}_M \left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \mathbf{K}_{MM} \left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \mathbf{y}_M \\
& \quad + \cancel{\mathrm{tr}\{\mathbf{I}_D\}} - \mathrm{tr}\left\{\left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \mathbf{K}_{MM}\right\} \\
& \quad + \log\left|\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right| + \log\left|\Sigma_{\boldsymbol{\beta}_M}^{-1}\right| - \cancel{\mathrm{tr}\{\mathbf{I}_D\}}\Big)
\end{aligned} \tag{119}$$

$$\begin{aligned}
= \sum_{i=1}^{N} & \left(\log \mathcal{N}\left(y_i \mid m_{f_i|\mathbf{y}_M}, \sigma^2\right) - \frac{1}{2}\sigma^{-2} k_{f_i|\mathbf{y}_M}\right) \\
& - \frac{1}{2}\Big(-\mathrm{tr}\left\{\left(\mathbf{K}_{MM} + \Sigma_{\boldsymbol{\beta}_M}\right)^{-1} \mathbf{K}_{MM}\right\} \\
& \quad + \mathbf{y}_M \left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \mathbf{K}_{MM} \left(\Sigma_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1} \mathbf{y}_M \\
& \quad + \log\left|\mathbf{K}_{MM} + \Sigma_{\boldsymbol{\beta}_M}\right| - \log\left|\Sigma_{\boldsymbol{\beta}_M}\right|\Big)
\end{aligned} \tag{120}$$

$$\begin{aligned}
& = \log \mathcal{N}\left(\mathbf{y} | \mathbf{m}_{\mathbf{f}|\mathbf{y}_M}, \sigma^2 \mathbf{I}_D\right) - \frac{1}{2\sigma^2} \mathrm{tr}\left\{\mathbf{K}_{\mathbf{f}|\mathbf{y}_M}\right\} \\
& - \frac{1}{2}\left[-\mathrm{tr}\left\{\mathbf{A}\mathbf{K}_{MM}\right\} + \mathbf{y}_M^\top \mathbf{A}\mathbf{K}_{MM}\mathbf{A}\mathbf{y}_M - \ln|\mathbf{A}| - \ln\left|\boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}\right|\right] \ ,
\end{aligned} \tag{121}$$

where $\mathbf{A} = \left(\mathbf{\Sigma}_{\boldsymbol{\beta}_M} + \mathbf{K}_{MM}\right)^{-1}$ and we have combined (a) the sum over N scalar likelihoods into a single, multivariate Gaussian with mean $\mathbf{m}_{\mathbf{f}|\mathbf{y}_M}$ (composed of $\mathbf{m}_{f_i|\mathbf{y}_M}, \forall i$) and diagonal unit covariance; and (b) all $\mathbf{k}_{\mathbf{f}_i|\mathbf{y}_M}$ terms into a diagonal matrix $\mathbf{K}_{\mathbf{f}|\mathbf{y}_M ii} = \mathbf{k}_{\mathbf{f}_i|\mathbf{y}_M}$.

A.3 CVGP'S LOWER-BOUND AND ITS OPTIMUM

Before expanding CVGP's lower-bound in Equation (64), we defining some auxiliary quantities

$$\mathbf{A} = \left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1} = \mathbf{K}_{MM}^{-1} - \mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1} \tag{122}$$

$$\text{where } \mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1} = \mathbf{K}_{MM}^{-1} - \left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1} \tag{123}$$

to write it explicitly in terms of its parameters:

$$\begin{aligned}
\mathcal{L}_{CVGP} &= -\frac{N}{2}\log(2\pi) + \frac{M}{2} - \frac{1}{2}\log\left|\sigma^2\mathbf{I}_N\right| - \frac{1}{2}\log\left|\mathbf{K}_{MM}\right| - \frac{1}{2}\mathbf{y}^\top\sigma^{-2}\mathbf{y} \\
&+ \sigma^{-2}\widetilde{\mathbf{m}}_{\mathbf{f}_M}^\top\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\mathbf{y} - \frac{1}{2}\widetilde{\mathbf{m}}_{\mathbf{f}_M}^\top\mathbf{K}_{MM}^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\mathbf{K}_{MM}^{-1}\widetilde{\mathbf{m}}_{\mathbf{f}_M} \\
&- \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} \\
&- \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\mathbf{K}_{MM}^{-1}\widetilde{\mathbf{K}}_{\mathbf{f}_M,\mathbf{f}_M}\right\} + \frac{1}{2}\log\left|\widetilde{\mathbf{K}}_{\mathbf{f}_M,\mathbf{f}_M}\right| && (124) \\
&= -\frac{N}{2}\log(2\pi) + \frac{M}{2} - \frac{1}{2}\log\left|\sigma^2\mathbf{I}_N\right| - \frac{1}{2}\log\left|\mathbf{K}_{MM}\right| - \frac{1}{2}\mathbf{y}^\top\sigma^{-2}\mathbf{y} \\
&+ \sigma^{-2}\mathbf{y}_M^\top\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\mathbf{y} && (125) \\
&- \frac{1}{2}\mathbf{y}_M^\top\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\mathbf{K}_{MM}^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{y}_M \\
&- \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} \\
&- \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\mathbf{K}_{MM}^{-1}\left(\mathbf{K}_{MM} - \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\right)\right\} && (126) \\
&+ \frac{1}{2}\log\left|\mathbf{K}_{MM} - \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\right| && (127) \\
&= -\frac{N}{2}\log(2\pi) + \frac{M}{2} - \frac{1}{2}\log\left|\sigma^2\mathbf{I}_N\right| - \frac{1}{2}\log\left|\mathbf{K}_{MM}\right| - \frac{1}{2}\mathbf{y}^\top\sigma^{-2}\mathbf{y} \\
&+ \sigma^{-2}\mathbf{y}_M^\top\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MN}\mathbf{y} && (128) \\
&- \frac{1}{2}\mathbf{y}_M^\top\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{y}_M \\
&- \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} \\
&- \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\right\} && (129) \\
&+ \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\right\} && (130) \\
&+ \frac{1}{2}\log\left|\mathbf{K}_{MM} - \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\right| . && (131)
\end{aligned}$$

We now compute the derivatives with respect to its free parameters

$$\frac{\partial\mathcal{L}_{CVGP}}{\partial\mathbf{y}_M} = \sigma^{-2}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MN}\mathbf{y} - \left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{y}_M , \tag{132}$$

$$\begin{aligned}
\frac{\partial\mathcal{L}_{CVGP}}{\partial\mathbf{A}} &= \sigma^{-2}\mathbf{y}_M^\top\mathbf{K}_{MN}\mathbf{y} - \mathbf{y}_M^\top\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}A\mathbf{y}_M \\
&+ \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\mathbf{K}_{MM}\right\} \\
&- \frac{1}{2}\mathrm{tr}\left\{\left(\mathbf{K}_{MM} - \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\right)^{-1}\mathbf{K}_{MM}\mathbf{K}_{MM}\right\} .
\end{aligned} \tag{133}$$

We can readily resolve that

$$\mathbf{y}_M^* = \sigma^{-2}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\mathbf{K}_{MN}\mathbf{y} \tag{134}$$

$$= \sigma^{-2}A^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\mathbf{K}_{MN}\mathbf{y} \tag{135}$$

$$\tag{136}$$

and replace it in the covariance expression

$$\begin{aligned}
0 = {} & \sigma^{-2}\sigma^{-2}\mathbf{y}^\top\mathbf{K}_{NM}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}A^{-1}\mathbf{K}_{MN}\mathbf{y} \\
& - \sigma^{-2}\mathbf{y}^\top\mathbf{K}_{NM}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}A^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}A\sigma^{-2}A^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\mathbf{K}_{MN}\mathbf{y} \\
& + \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\mathbf{K}_{MM}\right\} \\
& - \frac{1}{2}\mathrm{tr}\left\{\left(\mathbf{K}_{MM} - \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\right)^{-1}\mathbf{K}_{MM}\mathbf{K}_{MM}\right\}
\end{aligned} \tag{137}$$

$$\begin{aligned}
0 = {} & \sigma^{-2}\sigma^{-2}\mathbf{y}^\top\mathbf{K}_{NM}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}A^{-1}\mathbf{K}_{MN}\mathbf{y} \\
& - \sigma^{-2}\sigma^{-2}\mathbf{y}^\top\mathbf{K}_{NM}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}A^{-1}\mathbf{K}_{MN}\mathbf{y} \\
& + \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\mathbf{K}_{MM}\right\} \\
& - \frac{1}{2}\mathrm{tr}\left\{\left(\mathbf{K}_{MM} - \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\right)^{-1}\mathbf{K}_{MM}\mathbf{K}_{MM}\right\}
\end{aligned} \tag{138}$$

$$\begin{aligned}
0 = {} & \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\mathbf{K}_{MM}\right\} \\
& - \frac{1}{2}\mathrm{tr}\left\{\left(\mathbf{K}_{MM} - \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\right)^{-1}\mathbf{K}_{MM}\mathbf{K}_{MM}\right\}
\end{aligned} \tag{139}$$

Equating the matrices inside the traces, we have

$$\mathbf{K}_{MM}^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\mathbf{K}_{MM} = \left(\mathbf{K}_{MM} - \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\right)^{-1}\mathbf{K}_{MM}\mathbf{K}_{MM} \tag{140}$$

$$\mathbf{K}_{MM}^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M} = \left(\mathbf{K}_{MM} - \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\right)^{-1}\mathbf{K}_{MM} \tag{141}$$

$$\mathbf{K}_{MM}^{-1}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\mathbf{K}_{MM}^{-1} = \left(\mathbf{K}_{MM} - \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\right)^{-1} \tag{142}$$

$$\mathbf{K}_{MM}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\mathbf{K}_{MM} = \left(\mathbf{K}_{MM} - \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\mathbf{K}_{MM}\right) \tag{143}$$

$$\mathbf{K}_{MM}\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\mathbf{K}_{MM} = \mathbf{K}_{MM}\left(\mathbf{K}_{MM}^{-1} - \left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\right)\mathbf{K}_{MM} \tag{144}$$

$$\mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1} = \left(\mathbf{K}_{MM}^{-1} - \left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\right) \tag{145}$$

$$\left(\mathbf{K}_{MM} + \mathbf{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1} = \left(\mathbf{K}_{MM}^{-1} - \mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\right) \tag{146}$$

$$\mathbf{K}_{MM}^{-1} - \mathbf{K}_{MM}^{-1}\left(\mathbf{\Sigma}_{\boldsymbol{\beta}_M}^{-1} + \mathbf{K}_{MM}^{-1}\right)^{-1}\mathbf{K}_{MM}^{-1} = \left(\mathbf{K}_{MM}^{-1} - \mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\right) \tag{147}$$

$$\mathbf{K}_{MM}^{-1}\left(\mathbf{\Sigma}_{\boldsymbol{\beta}_M}^{-1} + \mathbf{K}_{MM}^{-1}\right)^{-1}\mathbf{K}_{MM}^{-1} = \mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1} \tag{148}$$

$$\mathbf{K}_{MM}\left(\mathbf{\Sigma}_{\boldsymbol{\beta}_M}^{-1} + \mathbf{K}_{MM}^{-1}\right)\mathbf{K}_{MM} = \mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M} \tag{149}$$

$$\mathbf{K}_{MM}\mathbf{\Sigma}_{\boldsymbol{\beta}_M}^{-1}\mathbf{K}_{MM} + \mathbf{K}_{MM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MM} = \mathbf{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M} = \mathbf{K}_{MM} + \frac{1}{\sigma^2}\mathbf{K}_{MN}\mathbf{K}_{NM} \tag{150}$$

$$\mathbf{K}_{MM}\mathbf{\Sigma}_{\boldsymbol{\beta}_M}^{-1}\mathbf{K}_{MM} = \frac{1}{\sigma^2}\mathbf{K}_{MN}\mathbf{K}_{NM} \tag{151}$$

$$\mathbf{\Sigma}_{\boldsymbol{\beta}_M}^{-1} = \frac{1}{\sigma^2}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1} \tag{152}$$

$$\mathbf{\Sigma}_{\boldsymbol{\beta}_M}^{*} = \sigma^2\mathbf{K}_{MM}\left(\mathbf{K}_{MN}\mathbf{K}_{NM}\right)^{-1}\mathbf{K}_{MM} \tag{153}$$

We now elaborate on the optimal values for CVGP's pseudo-coresets, rewriting CVGP's optimal pseudo-observations as

$$\mathbf{y}_M^* = \sigma^{-2} \left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}\right) \boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1} \mathbf{K}_{MN}\mathbf{y} \tag{154}$$

$$= \sigma^{-2} \left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}\right) \left[\mathbf{K}_{MM}^{-1} - \left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}\right)^{-1}\right] \mathbf{K}_{MN}\mathbf{y} \tag{155}$$

$$= \sigma^{-2} \left[\left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}\right) \mathbf{K}_{MM}^{-1} - \mathbf{I}_M\right] \mathbf{K}_{MN}\mathbf{y} \tag{156}$$

$$= \sigma^{-2} \left[\mathbf{K}_{MM} \left[\mathbf{K}_{MM}^{-1} + \sigma^2 \left(\mathbf{K}_{MN}\mathbf{K}_{NM}\right)^{-1}\right] \mathbf{K}_{MM}\mathbf{K}_{MM}^{-1} - \mathbf{I}_M\right] \mathbf{K}_{MN}\mathbf{y} \tag{157}$$

$$= \sigma^{-2} \left[\mathbf{I}_M + \sigma^2 \mathbf{K}_{MM} \left(\mathbf{K}_{MN}\mathbf{K}_{NM}\right)^{-1} - \mathbf{I}_M\right] \mathbf{K}_{MN}\mathbf{y} \tag{158}$$

$$= \sigma^{-2}(\sigma^2 \mathbf{K}_{MM} \left(\mathbf{K}_{MN}\mathbf{K}_{NM}\right)^{-1} \mathbf{K}_{MN})\mathbf{y} \tag{159}$$

$$= \sigma^{-2} \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}^* \mathbf{y} \tag{160}$$

$$= \mathbf{K}_{MM} \left(\mathbf{K}_{MN}\mathbf{K}_{NM}\right)^{-1} \mathbf{K}_{MN}\mathbf{y} \tag{161}$$

With this optimal values, we can now rewrite the lower-bound at its maxima

$$
\begin{aligned}
\mathcal{L}_{CVGP}(\mathbf{y}_M^*, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}^*) &= -\frac{N}{2}\log(2\pi) + \frac{M}{2} - \frac{1}{2}\log\left|\sigma^2\mathbf{I}_N\right| - \frac{1}{2}\log\left|\mathbf{K}_{MM}\right| - \frac{1}{2}\mathbf{y}^\top\sigma^{-2}\mathbf{y} \\
&+ \sigma^{-2}\mathbf{y}_M^{*\top}\left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}^*\right)^{-1}\mathbf{K}_{MN}\mathbf{y} && (162)\\
&- \frac{1}{2}\mathbf{y}_M^{*\top}\left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}^*\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}^*\right)^{-1}\mathbf{y}_M^* \\
&- \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} \\
&- \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\right\} + \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}^*\right)^{-1}\mathbf{K}_{MM}\right\} && (163)\\
&+ \frac{1}{2}\log\left|\mathbf{K}_{MM} - \mathbf{K}_{MM}\left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}^*\right)^{-1}\mathbf{K}_{MM}\right| && (164)\\
&= -\frac{N}{2}\log(2\pi) + \frac{M}{2} - \frac{1}{2}\log\left|\sigma^2\mathbf{I}_N\right| - \frac{1}{2}\log\left|\mathbf{K}_{MM}\right| - \frac{1}{2}\mathbf{y}^\top\sigma^{-2}\mathbf{y} \\
&+ \sigma^{-2}\sigma^{-2}\mathbf{y}^\top\mathbf{K}_{NM}\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}\right)\left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}^*\right)^{-1}\mathbf{K}_{MN}\mathbf{y} && (165)\\
&- \frac{1}{2}\sigma^{-2}\mathbf{y}^\top\mathbf{K}_{NM}\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}\right)\left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}^*\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\left(\mathbf{K}_{MM}\right. && (166)\\
&\qquad \left.+\boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}^*\right)^{-1}\sigma^{-2}\left(\mathbf{K}_{MM} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_M}\right)\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\mathbf{K}_{MN}\mathbf{y} \\
&- \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} \\
&- \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\right\} + \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\left(\mathbf{K}_{MM}^{-1} - \boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\right)\mathbf{K}_{MM}\right\} && (167)\\
&+ \frac{1}{2}\log\left|\mathbf{K}_{MM} - \mathbf{K}_{MM}\left(\mathbf{K}_{MM}^{-1} - \boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\right)\mathbf{K}_{MM}\right| && (168)\\
&= -\frac{N}{2}\log(2\pi) + \frac{M}{2} - \frac{1}{2}\log\left|\sigma^2\mathbf{I}_N\right| - \frac{1}{2}\log\left|\mathbf{K}_{MM}\right| - \frac{1}{2}\mathbf{y}^\top\sigma^{-2}\mathbf{y} \\
&+ \sigma^{-2}\sigma^{-2}\mathbf{y}^\top\mathbf{K}_{NM}\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\mathbf{K}_{MN}\mathbf{y} && (169)\\
&- \frac{1}{2}\sigma^{-2}\mathbf{y}^\top\mathbf{K}_{NM}\sigma^{-2}\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\mathbf{K}_{MN}\mathbf{y} \\
&- \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} \\
&- \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\right\} + \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}\right\} - \frac{1}{2}\mathrm{tr}\left\{\mathbf{K}_{MM}^{-1}\mathbf{K}_{MM}\right\} && (170)\\
&+ \frac{1}{2}\log\left|\mathbf{K}_{MM}\left(\mathbf{K}_{MM}^{-1} - \left(\mathbf{K}_{MM}^{-1} - \boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\right)\right)\mathbf{K}_{MM}\right| && (171)\\
&= -\frac{N}{2}\log(2\pi) + \frac{M}{2} - \frac{1}{2}\log\left|\sigma^2\mathbf{I}_N\right| - \frac{1}{2}\log\left|\mathbf{K}_{MM}\right| - \frac{1}{2}\mathbf{y}^\top\sigma^{-2}\mathbf{y} \\
&+ \frac{1}{2}\sigma^{-2}\sigma^{-2}\mathbf{y}^\top\mathbf{K}_{NM}\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\mathbf{K}_{MN}\mathbf{y} \\
&- \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} - \frac{1}{2}M \\
&+ \frac{1}{2}\log\left|\mathbf{K}_{MM}\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\mathbf{K}_{MM}\right| && (172)\\
&= -\frac{N}{2}\log(2\pi) \\
&- \frac{1}{2}\mathbf{y}^\top\sigma^{-2}\left(\mathbf{I}_M + \sigma^{-2}\mathbf{K}_{NM}\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\mathbf{K}_{MN}\right)\mathbf{y} \\
&- \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\right\} \\
&- \frac{1}{2}\log\left|\sigma^2\mathbf{I}_N\right| + \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\mathbf{f}_M,\mathbf{f}_M}^{-1}\mathbf{K}_{MM}\right| && (173)\\
&= \log\mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \sigma^2\mathbf{I}_N + \mathbf{Q}_{\mathbf{f}_M,\mathbf{f}_M}\right) - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{NN} - \mathbf{Q}_{\mathbf{f}_M,\mathbf{f}_M}\right\} && (174)
\end{aligned}
$$

which, for $M = Z$, and $\mathbf{X}_M = \mathbf{X}_Z$[2], corresponds with the same lower-bound as demonstrated by Titsias [2009] for SparseGP. $\mathcal{L}_{CVGP} \leq \mathcal{L}_{SparseGP}$ and the bound is tight with equality when $\mathbf{K}_{MM} \left(\mathbf{K}_{MN}\mathbf{K}_{NM}\right)^{-1} \mathbf{K}_{MM}$ is diagonal.

[2]Matching notations.

A.4 COMPLEXITIES OF CVGP

CVGP maintains the time and space complexity of SVGP with less parameters. This is because CVGP does not need to learn a free-form covariance matrix $\mathbf{S}$, but only the coreset values $\mathbf{X}_M, \mathbf{y}_M$ and their weights $\boldsymbol{\beta}_M$, which is more efficient —note that these three are M-dimensional vectors in the scalar case. We describe the complexities of benchmarks and their parameters in Table 2 below.

Inference technique	Complexities			Parameters
	Time	Space	# Parameter	
SparseGP [Titsias, 2009]	$\mathcal{O}\left(NM^2\right)$	$\mathcal{O}\left(NM^2\right)$	$\mathcal{O}\left(M\right)$	$\mathbf{X}_M$
SVGP [Hensman et al., 2013]	$\mathcal{O}\left(M^3\right)$	$\mathcal{O}\left(M^2\right)$	$\mathcal{O}\left(M^2\right)$	$\mathbf{X}_M, \mathbf{m}, \mathbf{S}$
CVGP	$\mathcal{O}\left(M^3\right)$	$\mathcal{O}\left(M^2\right)$	$\mathcal{O}\left(M\right)$	$\mathbf{X}_M, \mathbf{y}_M, \boldsymbol{\beta}_M$

Table 2: Computational analysis of CVGP and sparse $\mathcal{GP}$ alternatives: time and space complexities for obtaining an unbiased estimate of the log-marginal likelihood. CVGP enjoys same time and space complexity as SVGP, yet with a reduced variational parameter dimensionality. Contrary to SVGP, CVGP does not learn a free-form covariance parameter $\mathbf{S}$, but only tempering-parameters $\boldsymbol{\beta}_M$ of same size as $\mathbf{X}_M, \mathbf{y}_M$.

B EXPERIMENTS: SET-UP AND ADDITIONAL DETAILS

B.1 DATASETS

In this section, we describe the simulated and real-world datasets used in our experiments. We use UCI machine learning repository for real-world datasets [Janosi et al.], as described in Section B.1.1. The generative processes of simulated data are explained below in Section B.1.2. For all datasets, $\mathbf{X}$ are normalized (0 centered and unit variance) before training.

B.1.1 Real-world Datasets

Physicochemical properties of protein tertiary structure dataset (protein). A physicochemical data collection containing the properties of protein tertiary structure, specifically sourced from CASP 5-9. The dataset includes 45730 data points and 9 features [Rana, 2013].

Bike sharing dataset (bike). A bike sharing dataset comprised of 17 features and 17379 data points [Fanaee-T and Gama, 2013].

Parkinsons telemonitoring dataset (parkinsons). A biomedical voice measurements dataset obtained from 42 individuals in the early stages of Parkinson's disease. These individuals were enrolled in a six-month trial for remote symptom progression monitoring, using a telemonitoring device [Little et al., 2007]. There are 20 features and 5875 datapoints.

SkillCraft1 master table dataset (skillcraft). A video gaming telemetry data collection consisting of 12 features and 3338 data points [Thompson et al., 2013].

Wine quality dataset (wine). A collection of red wine samples with 11 features that are used to predict the wine's quality. In total, there are 1600 data points available for analysis [Cortez et al., 2009].

Year Prediction MSD (song). A collection of audio features. The goal is to predict the year a song is released [Bertin-Mahieux, 2011].

Relative location of CT slices on axial axis (slice). 53500 CT images from 74 different patients. The goal is to predict the relative location of the CT slice [Graf et al., 2011].

B.1.2 Simulated Datasets

We generate 1000 examples for each of the following synthetic datasets.

Synthetic 1. A 1-dimensional dataset following the below generative process:

$$f = \frac{2}{5}\left(\sin 3x \cos 2x + \sin\frac{x}{2} + \cos 2x + \exp\left\{-x^2\right\} + |x|\right), \qquad x \sim U(-4,4)\ , \tag{175}$$

$$y = f + \epsilon \sin 2\pi f\ , \qquad \epsilon \sim \mathcal{N}\left(\epsilon \mid 0, 3\times 10^{-1}\right)\ . \tag{176}$$

Synthetic 2. A 1-dimensional dataset following the below generative process:

$$f = \sin x^2 + \cos x^2 + \sin 3x + \cos 5x + \frac{\sqrt{|x|}}{2}\ , \qquad x \sim U(-4,4)\ , \tag{177}$$

$$y = f + \epsilon \sin 2\pi f\ , \qquad \epsilon \sim \mathcal{N}\left(\epsilon \mid 0, 3\times 10^{-1}\right)\ . \tag{178}$$

Synthetic 3. A 1-dimensional dataset following the below generative process:

$$f = \cos 2\pi x\ , \qquad x \sim U(0,2)\ , \tag{179}$$

$$y = f + \epsilon x^3\ , \qquad \epsilon \sim \mathcal{N}\left(\epsilon \mid 0, 1\right)\ . \tag{180}$$

Synthetic 4. A 2-dimensional dataset following the below generative process:

$$\mathbf{x} \sim \texttt{MakeBlobs}(centers = 3, std = 0.4) \ , \tag{181}$$
$$f_1 = 4 \sin x_1 + 2 \sin 2x_1 \ , \tag{182}$$
$$f_2 = 3 \cos 3x_2 + 4 \sin 5x_2 \ , \tag{183}$$
$$f_{12} = \exp\left\{-(x_1 + x_2)^2\right\} \ , \tag{184}$$
$$y = f_1 + f_2 + f_{12} + \epsilon \ , \qquad \epsilon \sim \mathcal{N}\left(\epsilon \mid 0, 2 \times 10^{-1}\right) \ . \tag{185}$$

where the function `MakeBlobs` is implemented as in Pedregosa et al. [2011].

Synthetic 5. A 2-dimensional dataset following the below generative process:

$$\mathbf{x} \sim \texttt{MakeMoons}(noise = 0.05) \tag{186}$$
$$f_1 = \frac{x_1}{2} + \sin 2x_1 \tag{187}$$
$$f_2 = \frac{x_2}{2} + \cos 5x_2 \tag{188}$$
$$f_{12} = \frac{\exp\left\{-(x_1 + x_2)^2\right\}}{2} \tag{189}$$
$$y = f_1 + f_2 + f_{12} + \epsilon \ , \qquad \epsilon \sim \mathcal{N}\left(\epsilon \mid 0, 2 \times 10^{-1}\right) \tag{190}$$

where the function `MakeMoons` is implemented by Pedregosa et al. [2011].

B.2 BASELINES

We use the GPytorch [Gardner et al., 2018] implementation of SparseGP, SGVP, and PPGPR. For ExactGP, we simply use the derivation of Rasmussen et al. [2006] implemented using the `MultivariateNormal` method of Pytorch [Paszke et al., 2019].

SparseGP. Introduced by [Titsias, 2009], SparseGP offers a variational solution to inducing point methods. In particular, SparseGP minimizes the KL divergence between an approximate and true posterior distribution. The loss function is derived by finding and plugging the optimal posterior variational distribution, which can be derived in terms of the $\mathcal{GP}$ kernel parameters.

SVGP. SVGP minimizes the KL divergence between an approximate and true posterior distribution where the posterior distribution is explicitly defined [Hensman et al., 2013]. The parameters of the posterior and model are learned jointly. SVGP is an stochastic approximation to SparseGP, which allows computationally efficient learning.

PPGPR. PPGPR is a variational predictive method for $\mathcal{GP}$s that, instead of lower-bounding the prior-predictive distribution as the methods above, proposed to optimize a lower-bound over the posterior predictive [Jankowiak et al., 2020]. This method provides predictive uncertainty estimates that model the variance of the observed data more accurately. PPGPR results in Jankowiak et al. [2020] were based on 400 epochs, which we found insufficient for convergence to optimal RMSE in our experiments. Although longer training helps with better predictive RMSE performance, it also causes severe overfitting on noise —a behavior not observed in other $\mathcal{GP}$ algorithms.

ExactGP. The exact learning of Gaussian processes as described by Rasmussen et al. [2006]. We compute the marginal log-likelihood (prior predictive) by integrating the likelihood over the latent function-space (with respect to prior distribution) to learn model hyperparameters, at a computational complexity of $\mathcal{O}\left(N^3\right)$.

B.3 EXPERIMENT DETAILS FOR REPRODUCIBILITY

We employ 5-fold cross-validation to compute and report each variational technique's (lower-bound) objective $\mathcal{L}$ in inference, as well as their predictive root-mean-squared error (RMSE) and posterior predictive log-likelihood (PPLL) over held out test splits.

We do not scale the KL-divergence terms in each model's objective, for them to be valid lower-bounds. We use a fixed random seed over all datasets to ensure that the folds (with 70%-30% train and validations splits) for different models are the same.

We use Adam optimizer with a learning rate of 10^{-3} for all methods and single precision floating point [Kingma and Ba, 2014]. For the techniques amenable to stochastic optimization (SVGP, PPGPR, and CVGP), we use a batch size of 512. Each model is run on a single NVIDIA® GeForce® RTX 20 series graphics card.

To leverage full model capacity and achieve full optimization performance, we train for 10^5 epochs maximum, and stop training only if there are no RMSE improvements for 3×10^3 consecutive epochs over the validation set. We early stop with respect to the best *held-out* validation set RMSE metric attained.

C ADDITIONAL EXPERIMENTS

We showcase here additional and different performance findings. For box plots below, unless otherwise stated, the best performing stochastic model —in comparison to other stochastic sub-sampling methods— is showcased in bold.

C.1 EVOLUTION OF PERFORMANCE METRICS BY TRAINING EPOCHS

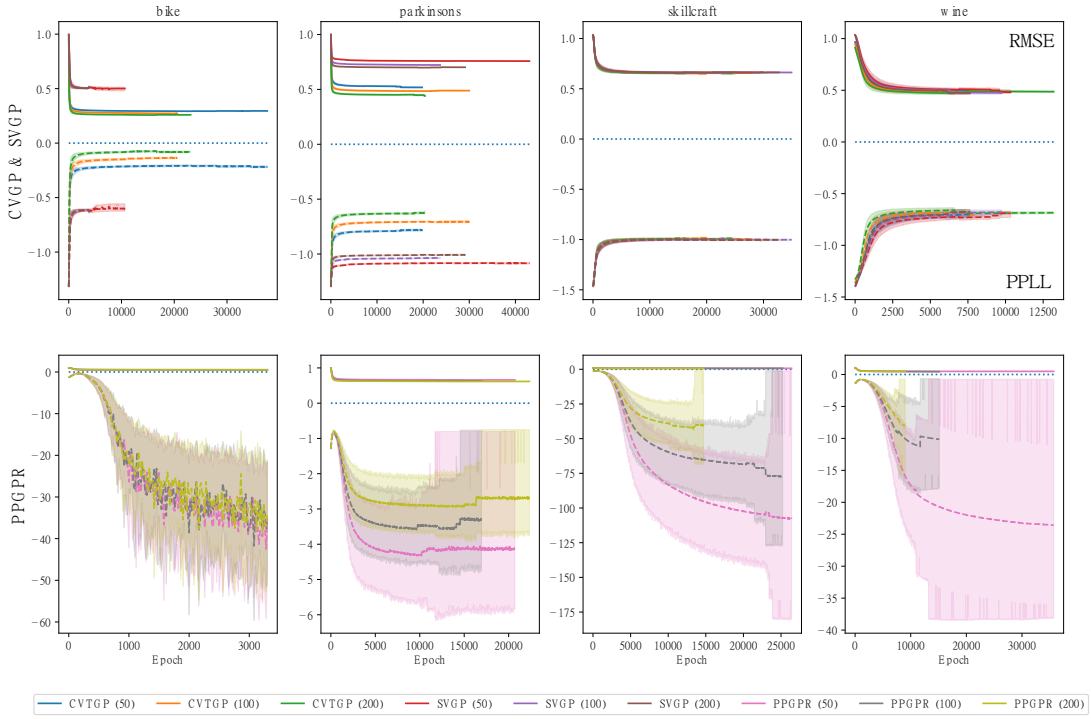


Figure 6: Evolution of RMSE and PPLL across training epochs. For CVGP and SVGP, validation RMSE and PPLL consistently decrease —showing no critical indication of overfitting. In contrast, PPGPR's RMSE improves, but PPLL worsens, indicating overfitting of noise and cross-correlation, leading to suboptimal PPLL. Training stops only when RMSE no longer improves. Large negative PPLL values prevent reporting PPGPR's PPLL in Figure 1.

C.2 MODEL LEARNING AND INFERENCE GAPS

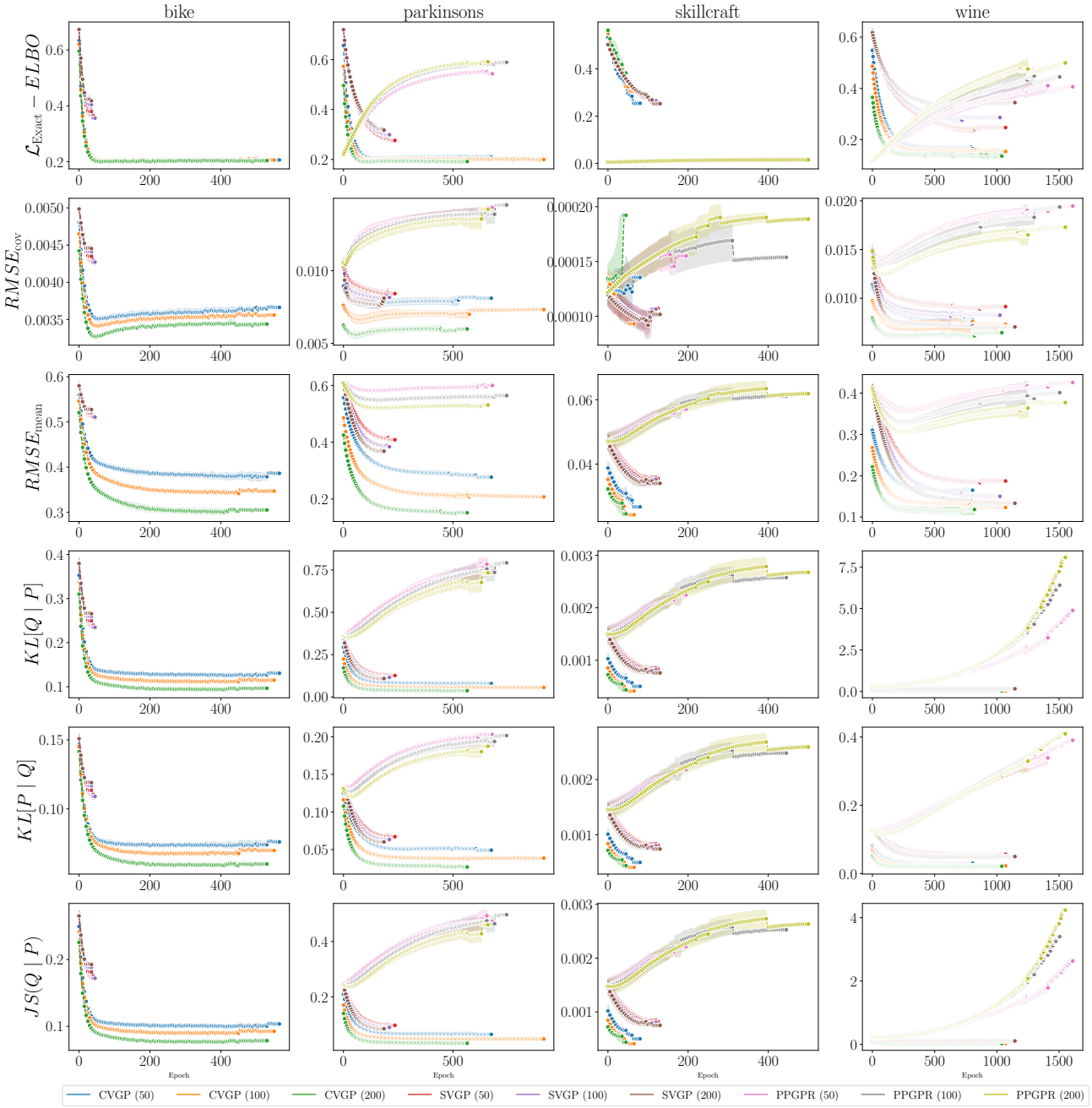


Figure 7: Divergence from the true posterior $p(f^\star \mid \mathbf{x}^\star, \mathbf{X}, \mathbf{y})$ and lower-bound tightness relative to the exact solution (i.e., the difference in log-marginal and ELBO, always positive). PPGPR does not lower-bound ExactGP and diverges, as shown in the figure. By directly fitting noisy observations, PPGPR overfits early, while SVGP and CVGP filter noise.

C.3 ROBUSTNESS TO INITIALIZATION

Below, we demonstrate CVGP's robustness to random initialization. We observe that RandomCVGP (CVGP initialized with white Gaussian noise) performs on par with CVGP in almost all cases and metrics. For very big datasets with many input-features (e.g., Song), a random initialization over high-dimensional input-output spaces is a clear disadvantage. Hence, we recommend, in general, to initialize CVGP with k-means.

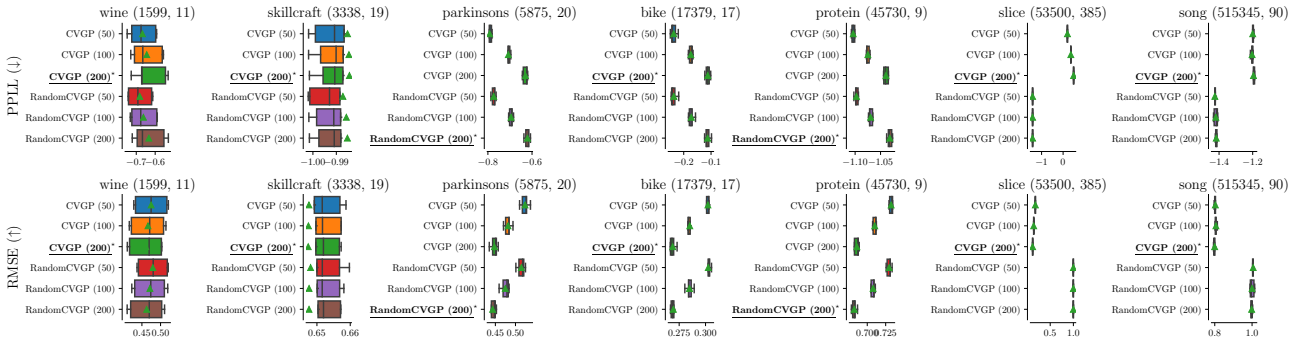


Figure 8: Predictive performance comparison for CVGP when initialized with random points (RandomCVGP) or k-means over the observed (real) datasets. CVGP is robust to random initializations. The best performing initiation mean statistic (▲) is **emphasized***.

C.4 POSTERIOR-PRIOR INTERPOLATION: NOISY REAL-WORLD DATA

We discuss all sparse $\mathcal{GP}$ method's ability for their posterior to interpolate between the model prior and the information provided by observations.

For CVGP, as the observation noise increases ($\sigma^2 \to \infty$), its posterior mean $\mathbf{m}_{\mathbf{f}_M|\mathbf{y}_M}$ converges to $\mathbf{0}$ (the $\mathcal{GP}$ prior mean), and its posterior covariance $\mathbf{K}_{\mathbf{f}_M|\mathbf{y}_M}$ converges to the prior covariance $\mathbf{K}_{MM}$; i.e., the observations are noninformative and CVGP's posterior reverts to the $\mathcal{GP}$ prior. Conversely, for noiseless data ($\sigma^2 \to 0$), CVGP's posterior mean approaches $\mathbf{y}_M$, and its posterior covariance diminishes to $\mathbf{0}$ (see Equation 14). On the contrary, SVGP's posterior statistics $(\mathbf{m}, \mathbf{S})$ have no explicit model dependencies, and therefore, are adjusted based purely on variational parameter optimization.

We run an empirical experiment below, where we take a real-world dataset and progressively add noise to the true regression values, before training SVGP, PPGPR, and CVGP on these extra-noisy versions of the datasets. As in any Bayesian model, we expect that for low noise regimes, the posterior should diverge from the prior to capture the information provided by observations; while for high noise regimes (uninformative data), the posterior should remain similar to the prior. Below, we notice that CVGP effectively resorts to the prior under uninformative data, a behavior exhibited by ExactGP, while PPGPR does not recover the prior —fitting the noisy data.

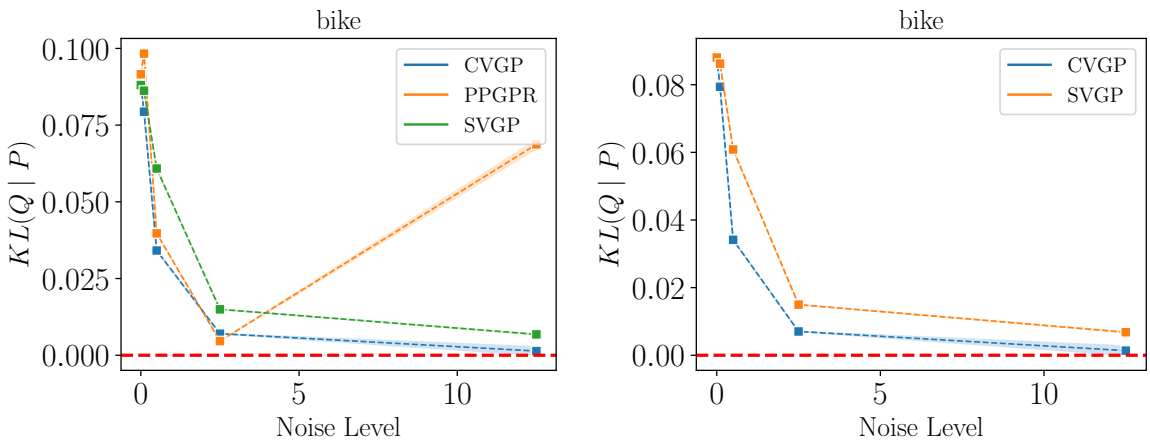


Figure 9: Study of the difference between sparse $\mathcal{GP}$ approximate posteriors and model prior for the Bike dataset, as measured by the KL-divergence between prior and approximate variational posterior ($\mathrm{KL}\left[q(\mathbf{f}_M) \| p(\mathbf{f}_M)\right]$) across different observation noise regimes. Left: CVGP, PPGPR, and SVGP, right: CVGP and SVGP. We see that PPGPR diverges from prior vastly while SVGP and CVGP retains the Gaussian prior-likelihood conjugacy (i.e., as noise increase they do not diverge from the prior vastly).

D QUALITATIVE STUDY

D.1 QUALITATIVE EVALUATION OF POSTERIOR PREDICTIVE

Below, we showcase the predictive distributions of each trained $\mathcal{GP}$ model, for the synthetic 1D datasets, i.e., synthetic 1 in Figure 10, synthetic 2 in Figure 11, and synthetic 3 in Figure 12.

Note that RandomCVGP is initialized with Gaussian white noise and faces a significantly more challenging task in fitting the data compared to SVGP, PPGPR, and CVGP. In fact, some of its learned inducing points/coresets fall off-the-grid and appear unrelated to the data. In these cases, the corresponding coreset weights are low (indicated in purple).

Conversely, points that effectively capture the $y|x$ relationship are shown in yellow-green, while those that do not are depicted in purple —and are consequently disregarded during posterior inference.

On the contrary, SVGP and PPGPR do not have this behavior. Hence, we use a single color for their coresets. This coreset-based posterior design endows CVGP with significant flexibility: if an inducing point proves unhelpful for predictions, its influence can be driven to 0 (with the help of its corresponding β_m). In contrast, both PPGPR and SVGP must select inducing points that consistently capture the data structure.

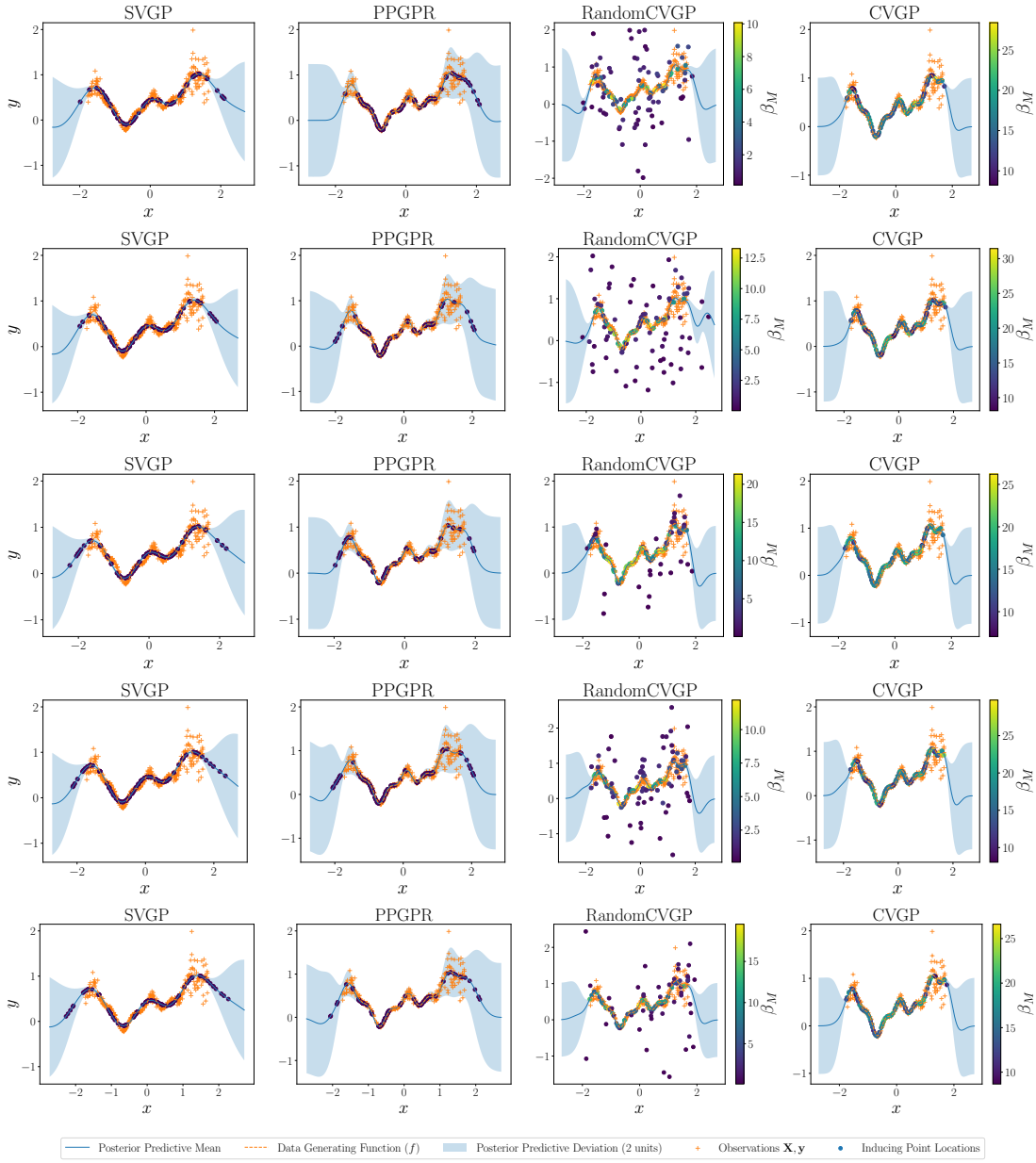


Figure 10: Posterior predictive distribution for the synthetic 1 dataset across different 5 folds, with 100 inducing points.

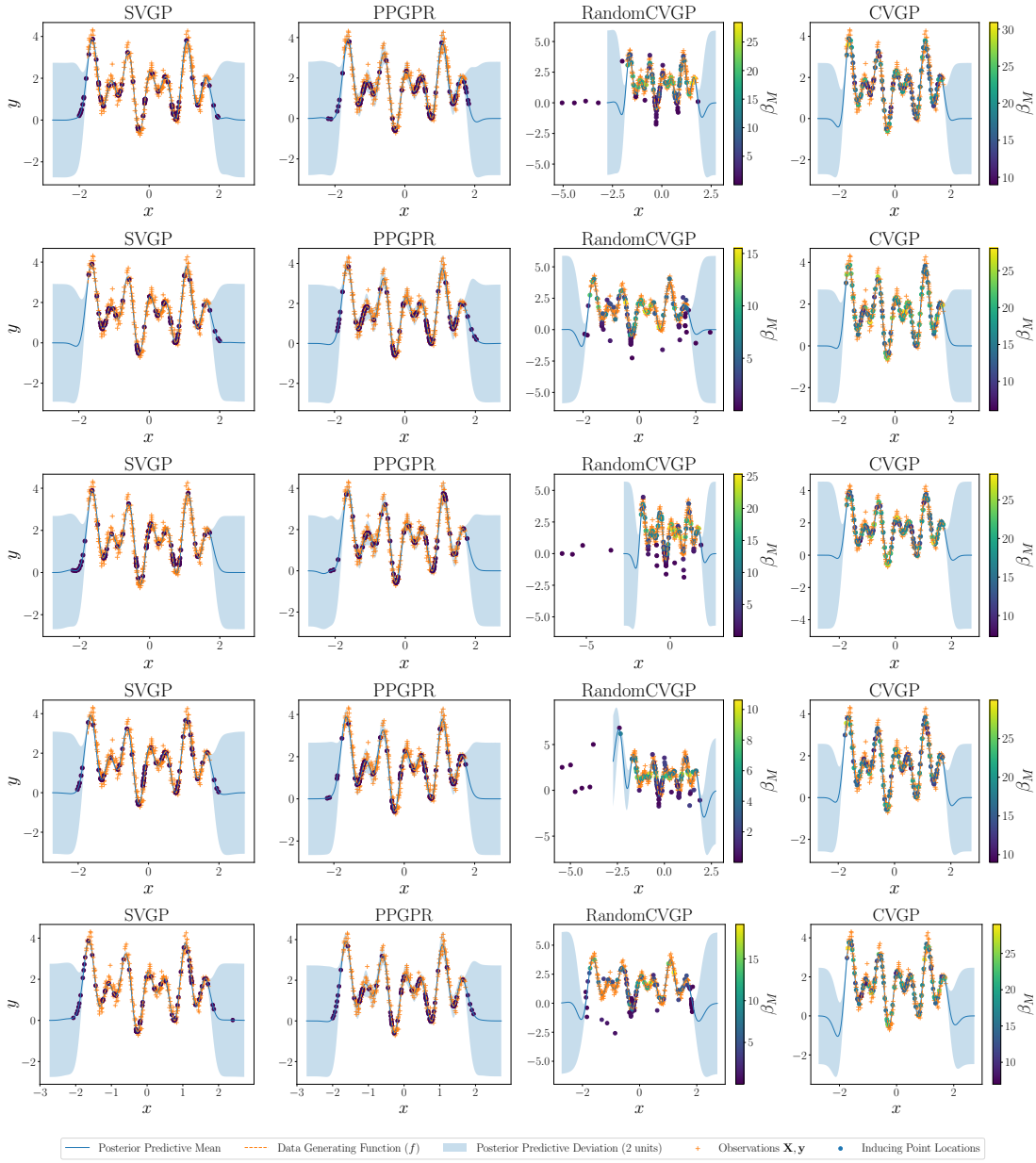


Figure 11: Posterior predictive distribution for the synthetic 2 dataset across different 5 folds, with 100 inducing points.

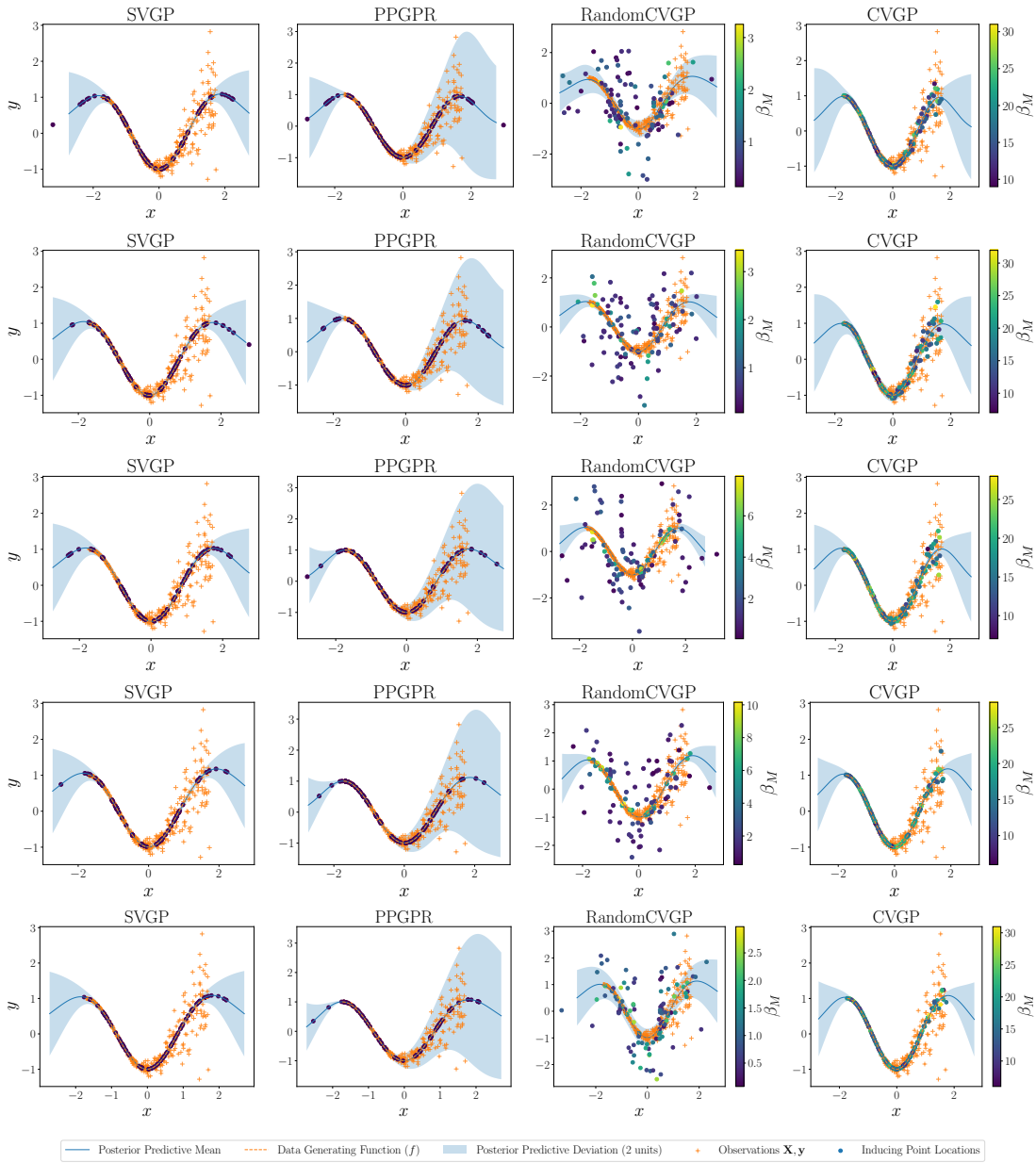
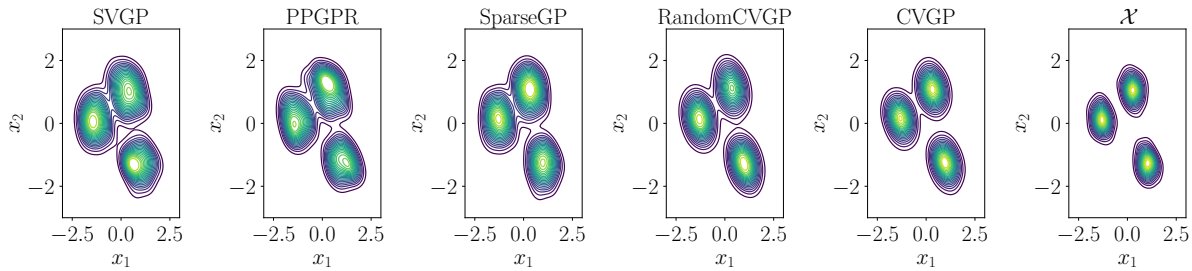


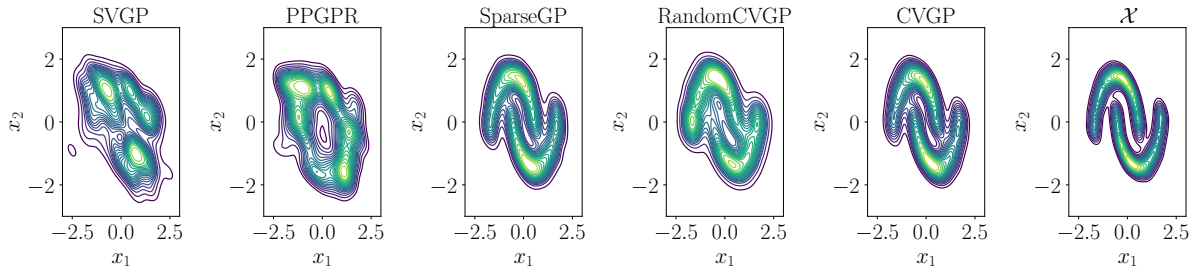
Figure 12: Posterior predictive distribution for the synthetic 3 dataset across different 5 folds, with 100 inducing points. We observe that PPGPR captures heteroscholastic uncertainty. *Although this seems like a plausible property, the noise in y can sometimes be pure noise, and could lead PPGPR to overfit as we discussed and demonstrated in Figures 2 and 6.*

D.2 STUDY OF INDUCING POINTS ($\mathbf{X}_M$)

We showcase the density of $\mathbf{X}_M$s learned by CVGP (weighted by $\boldsymbol{\beta}_M$), and the $\mathbf{X}_M$ points learned by other sparse $\mathcal{GP}$ methods on the 2-dimensional synthetic `Blobs` (Figure 14) and `TwoMoons` (Figure 15) datasets, across different folds of the training data. Notice how CVGP consistently learns meaningful data representations over all folds.



(a) Synthetic 4 dataset where $y = f(\mathbf{x}) + \epsilon$, and $\mathbf{x} \sim$ `MakeBlobs`$(.)$



(b) Synthetic 5 dataset where $y = f(\mathbf{x}) + \epsilon$, and $\mathbf{x} \sim$ `MakeMoons`$(.)$

Figure 13: Kernel density estimation (KDE) plots for $\mathbf{X}_M$ learned by CVGP and $\mathbf{X}_M$ for sparse baselines, on (a) synthetic 4 and (b) synthetic 5 datasets. For CVGP we use $\boldsymbol{\beta}_M$-weighted KDE plots, not possible for alternatives. All methods capture the clustered `Blobs` empirical distribution in Figure 13a, yet CVGP models the bi-modal nature of data more clearly. CVGP, RandomCVGP, and SparseGP adeptly capture the distinctive `TwoMoons` shape exhibited by the empirical data distribution in Figure 13b, in contrast to other stochastic sparse inference alternatives.

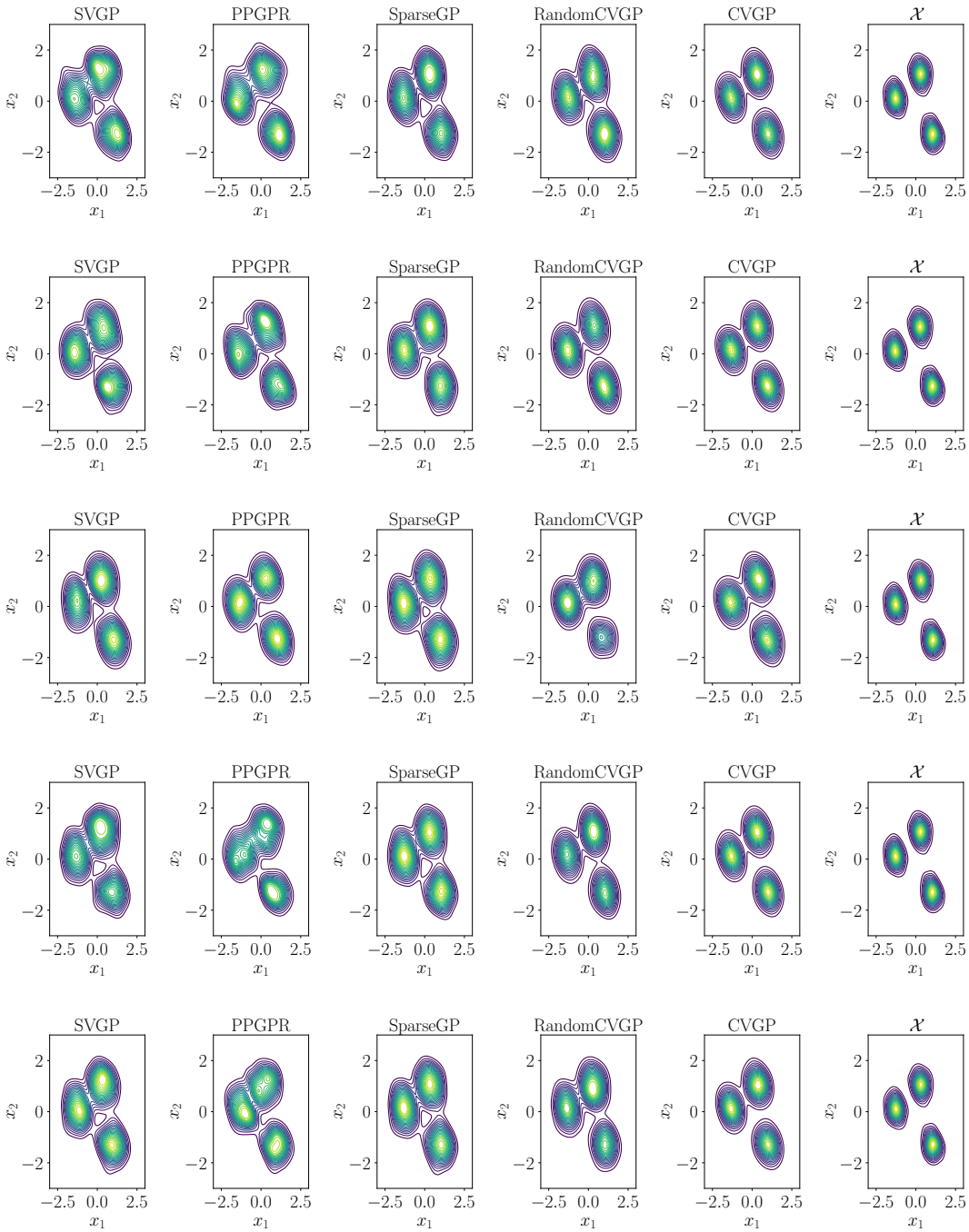


Figure 14: Learned representations for synthetic 4 dataset over 5 different folds with 100 inducing points. CVGP learns meaningful representations over different folds. RandomCVGP is more noisy than CVGP as it is initialized with Gaussian white noise.

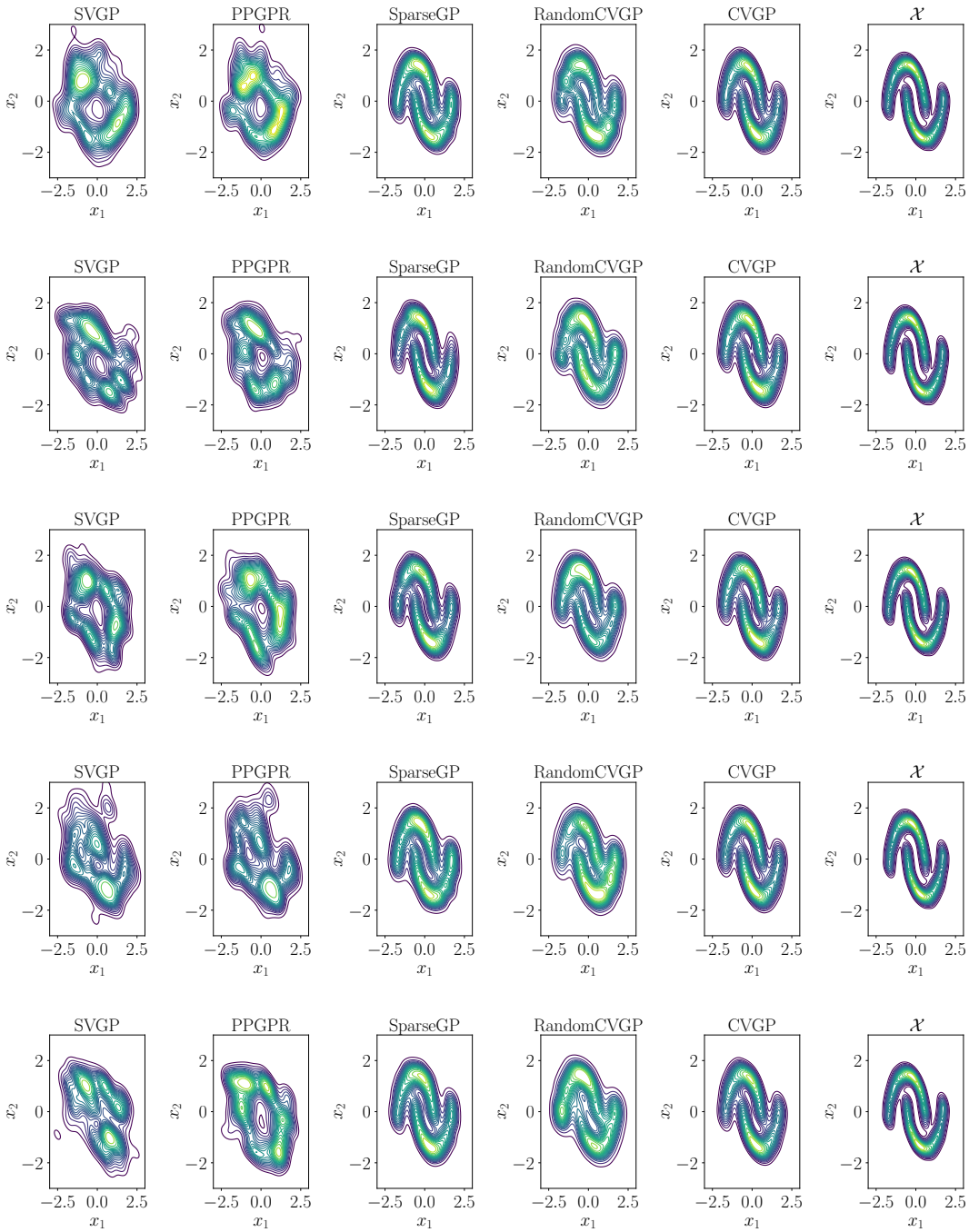


Figure 15: Learned representations for synthetic 5 dataset over 5 different folds. CVGP learns meaningful representations over the different folds while other models, except for SparseGP, struggle to capture the empirical distribution. RandomCVGP is more noisy than CVGP as it is initialized with Gaussian white noise.

D.3 LEARNED CORESET WEIGHT DISTRIBUTION FOR K-MEANS AND RANDOM INITIALIZATIONS

We show below the histogram of CVGP's learned coreset's weights across all synthetic datasets. Note that, all learned coresets have nonzero weights $\beta_m > 0,\ \forall m$, with very different histograms depending on the dataset: for some datasets, some pseudo input-output points $\{\mathbf{X}_M, \mathbf{y}_M\}$ are considerably up-weighted. We observe that RandomCVGP consistently drives the weight of unplausible inducing points to 0.

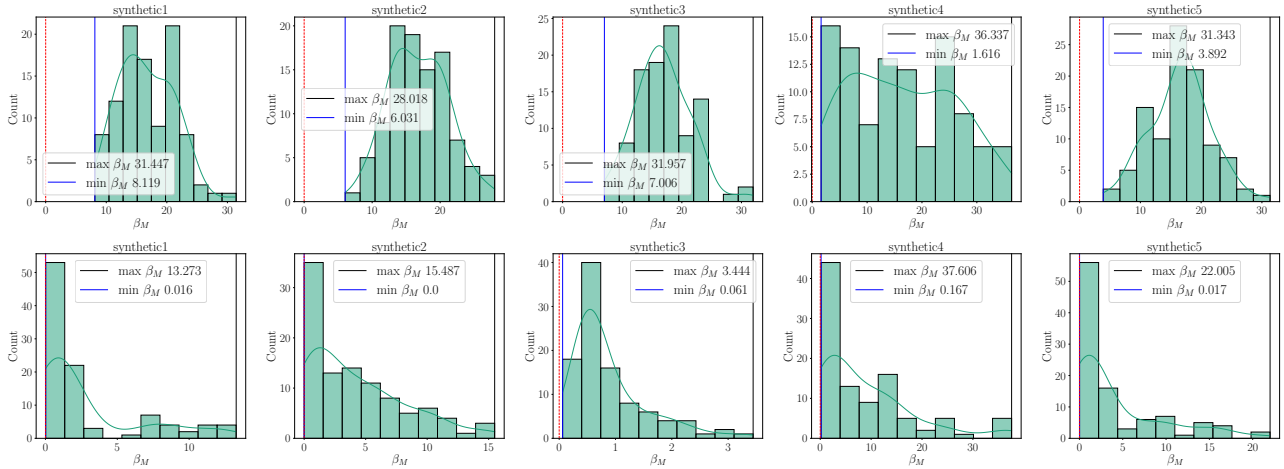


Figure 16: Histogram of learned CVGP coreset weights $\boldsymbol{\beta}_M$. Top CVGP, bottom RandomCVGP (i.e., initialization with white noise). We see that some of weights of RandomCVGP go to 0 while almost all weight values of CVGP are non-zero (i.e., no coreset tuple $\{\mathbf{X}_M, \mathbf{y}_M\}$ is discarded ($\beta_m > 0,\ \forall m$)).