Quantile Propagation for Wasserstein-Approximate Gaussian Processes

Rui Zhang [§♭] **Christian J. Walder** [§♭] **Edwin V. Bonilla** [♭] **Marian-Andrei Rizoiu** [♮♭] **Lexing Xie** [§♭]
[§]The Australian National University [♭]CSIRO's Data61 [♮]University of Technology Sydney

Abstract

In this work, we develop a new approximation method to solve the analytically intractable Bayesian inference for Gaussian process models with factorizable Gaussian likelihoods and single-output latent functions. Our method – dubbed QP – is similar to the expectation propagation (EP), however it minimizes the L^2 Wasserstein distance instead of the Kullback-Leibler (KL) divergence. We consider the specific case in which the non-Gaussian likelihood is approximated by the Gaussian likelihood. We show that QP has the following properties: (1) QP matches quantile functions rather than moments in EP; (2) QP and EP have the same local update for the mean of the approximate Gaussian likelihood; (3) the local variance estimate for the approximate likelihood is smaller for QP than for EP's, addressing EP's over-estimation of the variance; (4) the optimal approximate Gaussian likelihood enjoys a univariate parameterization, reducing memory consumption and computation time. Furthermore, we provide a unified interpretations of EP and QP – both are coordinate descent algorithms of a KL and an L^2 Wasserstein global objective function respectively, under the same assumptions. In the performed experiments, we employ eight real world datasets and we show that QP outperforms EP for the task of Gaussian process binary classification.

1 Introduction

Gaussian Process Models and Expectation Propagation. Gaussian process (GP) models have attracted the attention of the machine learning community due to their flexibility and their capacity to measure uncertainty. They have been widely applied to learning tasks such as regression (Williams and Rasmussen, 1996; Snelson et al., 2004), classifications (Williams and Barber, 1998; Hensman et al., 2015) and stochastic point processes modeling (Møller et al., 1998). However, one shortcoming of GP models with non-Gaussian likelihoods is the analytic intractability of the Bayesian inference. To address this issue, various approximate Bayesian inference methods were proposed, such as the Monte Carlo sampling (Neal, 1997), Laplace approximation (Williams and Barber, 1998), variational inference (Jordan et al., 1999) and expectation propagation (EP) (Opper and Winther, 2000; Minka, 2001b). The existing approach most relevant to this work is EP, which approximates the non-Gaussian likelihoods with the Gaussian likelihoods by iteratively minimizing local forward Kullback-Leibler (KL) divergences. As described by Gelman et al. (2017), EP enjoys high scalability on large datasets. However, EP is not guaranteed to converge, and it is known to over-estimate the GP variance (Minka, 2005; Heess et al., 2013; Hernández-Lobato et al., 2016).

Wasserstein Distance. Optimal transport divergences – such as the Wasserstein distance – have recently gained substantial popularity. The Wasserstein distance is a natural measure between two distributions, and it has been successfully employed for a number of learning tasks, such as image retrieval (Rubner et al., 2000), text classification (Huang et al., 2016) and image fusion (Courty et al., 2016). More recent works focus on using the Wasserstein distance for inference, including designing Wasserstein generative adversarial networks (Arjovsky et al., 2017), Wasserstein variational inference (Ambrogioni et al., 2018) and Wasserstein auto-encoders (Tolstikhin et al., 2017). In spite of its appealing intuitive formulation and excellent performance, the Wasserstein distance has prohibitive computation cost (Cuturi, 2013), especially for high dimensional distributions (Bonneel et al., 2015).

Contributions. In this work, we overcome some of the shortcomings of EP by developing an efficient Bayesian inference approximation method that minimizes the L^2 Wasserstein distance. Here below we detail the four main contributions of this paper.

First, in Sec. 4, we develop QP, an approximate inference algorithm similar in spirit to EP for GP models with fac-

Preprint. Work in progress.

torizable non-Gaussian likelihoods and single-output latent functions. Similar to EP, QP does not require to directly minimize the global L^2 Wasserstein distance between the true and the approximate joint posteriors. Instead, it iteratively minimizes the local L^2 Wasserstein distances between two kinds of approximate marginal distributions, employing true likelihoods or not, and in turn matches the quantile functions (rather than moments in EP), for which it is dubbed *quantile propagation* (QP). We further provide the update formulas for the mean and the variance of the Gaussian likelihood. The estimate of the mean in QP is equal to EP's, however its variance is lower than EP's, therefore improving EP's over-estimation of the variance (Minka, 2005; Heess et al., 2013; Hernández-Lobato et al., 2016). Importantly, although employing the more complex L^2 Wasserstein distance compared to the KL divergence, our method enjoys the same computational time complexity as EP's, with the help of two lookup tables.

Second, in Sec. 5, we show that similar to EP, the optimal approximate Gaussian likelihood in QP enjoys an economic parameterization, i.e. its form relies solely on a single latent variable. Compared with a general full parameterization on all latent variables, this property allows to reduce the memory consumption by a factor of N (the size of the data). It also consistently reduces the computation time by optimizing fewer parameters in each local update – $O(1)$ for the economic parametrization vs $O(N^2)$ for the full parameterization.

Third, in Sec. 6, we provide an unified interpretations of EP and QP. We show that both methods are instances of coordinate descent algorithms to a KL divergence and an L^2 Wasserstein global objective function respectively with the same assumptions.

Finally, in our experiments in Sec. 7, we compare EP and QP for the task of Gaussian process binary classifications on eight real world datasets. The results show that our method outperforms EP in both predictive accuracy and uncertainty quantification, which validates that QP alleviates EP's over-estimation of the variance.

2 Related Work

Expectation propagation (EP) was first proposed for the GP model (Opper and Winther, 2000) and then generalized by Minka (2001a,b). Power EP (an extension of EP) (Minka, 2004, 2005) exploits the more general α-divergence (a value of $\alpha = 1$ corresponds to the forward KL divergence in EP) and was recently used in conjuncture with the GP pseudo-point approximation (Bui et al., 2017). Although, generally not guaranteed to converge, performing local updates using fixed-point iterations was shown to perform well in practice for GP regression and classification (Bui et al., 2017). By comparison, our approach provides guarantees of convergent local updates for the class of GP models equipped with the general L^p ($p \geq 1$) Wasserstein distance. Moreover and for the same GP class with the L^2 Wasserstein distance, we show the optimal approximate likelihood to rely solely on a single latent variable – as opposed to all latent variables. A similar result was previously shown for the GP with a forward KL divergence (Seeger, 2005).

Without guarantees of convergence or the explicit global objective function, interpreting EP has proven to be a hard task. As a result, a number of works have instead attempted to directly minimize the divergences between true and approximate joint posteriors, employing such as KL (Jordan et al., 1999; Dezfouli and Bonilla, 2015), renyi (Li and Turner, 2016), α (Hernández-Lobato et al., 2016) and optimal transport divergences (Ambrogioni et al., 2018). To deal with the infinity issue of KL (and more general of the renyi and α divergences) raised by different supports (Montavon et al., 2016; Arjovsky et al., 2017; Gulrajani et al., 2017), Hensman et al. (2014) employs the product of tilted distributions as an approximation. A number of variants for EP have also been proposed, including the convergent double loop algorithm (Opper and Winther, 2005), distributed EP (Xu et al., 2014; Gelman et al., 2017) built on partitioned datasets, averaged EP (Dehaene and Barthelmé, 2018) assuming that all approximate likelihoods contribute similarly, and stochastic EP (Li et al., 2015) regarded as sequential averaged EP.

The **$\mathbf{L^2}$ Wasserstein distance** between two Gaussian distributions has the closed form expression (Dowson and Landau, 1982). A detailed research on the Wasserstein geometry of the Gaussian distribution is conducted by Takatsu (2011). Recently, the formula is applied to a robust Kalman filtering (Shafieezadeh-Abadeh et al., 2018) and to Gaussian processes (Mallasto and Feragen, 2017). A more general extension to elliptically contoured distributions is provided by Gelbrich (1990), which they employ to compute probabilistic embeddings for words (Muzellec and Cuturi, 2018). Moreover, a geodesic interpretation can be attributed to the L^2 Wasserstein distance using any distribution (Benamou and Brenier, 2000), and this has already been exploited to develop approximate Bayesian inferences (El Moselhy and Marzouk, 2012). Our work builds on the L^2 Wasserstein distance, and it does not exploit any of these closed form expressions (or their assumptions); it enjoys computational efficiency by levering the EP framework.

3 Prerequisites

In this section, we review GP models with non-Gaussian likelihoods and one-output latent functions, the expectation propagation algorithm and the Wasserstein distance.

3.1 Gaussian Process Models

Consider a dataset of N samples $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the input vector and $y_i \in \mathbb{R}$ is the output scalar. The basic idea behind the GP model is establishing the mapping from inputs to outputs via a latent function $f : \mathbb{R}^d \to \mathbb{R}$ which is assigned a GP prior. Specifically, we assume a zero-mean GP prior $p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, K)$, where $\boldsymbol{f} = \{f_i\}_{i=1}^N$, with $f_i \equiv f(\boldsymbol{x}_i)$, is the set of latent function values and K is the covariance matrix induced by evaluating the covariance function $k(\cdot,\cdot)$ at every pair of inputs. This work exploits the commonly used squared exponential covariance function $k(\boldsymbol{x}, \boldsymbol{x}') = \gamma \exp\big[-\sum_{i=1}^d (x_i - x_i')^2/(2\alpha_i^2)\big]$. We denote the GP hyper-parameters as $\boldsymbol{\theta} = \{\gamma, \alpha_1, \cdots, \alpha_d\}$, and for notational simplification, we will omit conditioning on $\boldsymbol{\theta}$. Along with the prior, we assume a factorized likelihood $p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^N p(y_i|f_i)$ where $\boldsymbol{y}$ is the set of all outputs. Given the above, the posterior $\boldsymbol{f}$ is expressed as:

$$p(\boldsymbol{f}|\mathcal{D}) = \frac{p(\boldsymbol{f})\prod_{i=1}^N p(y_i|f_i)}{p(\mathcal{D})},$$

where the model evidence $p(\mathcal{D}) = \int p(\boldsymbol{f}) \prod_{i=1}^N p(y_i|f_i)\, \mathrm{d}\boldsymbol{f}$ is often analytically intractable. Interestingly, numerous problems can be well modeled in this way: binary classification (Williams and Barber, 1998; Kuss and Rasmussen, 2005), one-output regression (Williams and Rasmussen, 1996; Jylänki et al., 2011), log Gaussian cox processes (Møller et al., 1998) and warped Gaussian processes (Snelson et al., 2004).

3.2 Expectation Propagation

In this section, we review the application of expectation propagation (EP) to the above described GP models. EP deals with the analytical intractability by exploiting the Gaussian likelihood to approximate the individual non-Gaussian likelihood:

$$p(y_i|f_i) \approx t_i(f_i) \equiv \widetilde{Z}_i \mathcal{N}(f_i|\widetilde{\mu}_i, \widetilde{\sigma}_i^2).$$

The function t_i is often called the *site function* and is specified by *site parameters*: the scale $\widetilde{Z}_i$, the mean $\widetilde{\mu}_i$ and the variance $\widetilde{\sigma}_i^2$. Notably, in the case of GP models with factorized likelihoods, each local approximation relies merely on f_i instead of all variables $\boldsymbol{f}$. We refer to this property as an economic parameterization. Seeger (2005) showed that such a parameterization is equivalent to a more general one that uses all variables. Interestingly, although our method employs a more complex Wasserstein distance, this property still holds, as elaborated in Sec. 5.2.

Consequently, the intractable posterior distribution $p(\boldsymbol{f}|\mathcal{D})$ is approximated by a Gaussian distribution $q(\boldsymbol{f})$:

$$q(\boldsymbol{f}) = \frac{p(\boldsymbol{f})\prod_{i=1}^N t_i(f_i)}{q(\mathcal{D})} \equiv \mathcal{N}(\boldsymbol{f}|\boldsymbol{\mu}, \Sigma),$$

$$\boldsymbol{\mu} = \Sigma\widetilde{\Sigma}^{-1}\widetilde{\boldsymbol{\mu}},\ \ \Sigma = (K^{-1} + \widetilde{\Sigma}^{-1})^{-1},$$

where conditioning on $\mathcal{D}$ is omitted from $q(\boldsymbol{f})$ for notational simplification, $\widetilde{\boldsymbol{\mu}}$ is the vector of $\widetilde{\mu}_i$, $\widetilde{\Sigma}$ is diagonal with $\widetilde{\Sigma}_{ii} = \widetilde{\sigma}_i^2$, and the log approximate model evidence $\log q(\mathcal{D})$ has the below closed form expression:

$$\log q(\mathcal{D}) = \sum_{i=1}^N \log \widetilde{Z}_i - \frac{1}{2}\log|K + \widetilde{\Sigma}| - \frac{1}{2}\widetilde{\boldsymbol{\mu}}^{\mathrm{T}}(K + \widetilde{\Sigma})^{-1}\widetilde{\boldsymbol{\mu}} - \frac{N}{2}\log(2\pi).$$

This approximate model evidence is further employed to optimize GP hyper-parameters $\boldsymbol{\theta}^\star = \mathrm{argmax}_{\boldsymbol{\theta}} \log q(\mathcal{D})$.

The core of EP is the optimization for a site function $t_i(f_i)$. A naive way is minimizing the forward KL divergence between the true posterior $p(\boldsymbol{f}|\mathcal{D})$ and the distribution $\widetilde{p}(\boldsymbol{f}|\mathcal{D}) \propto p(\boldsymbol{f}|\mathcal{D})t_i(f_i)/p(y_i|f_i)$ formed by replacing the factor $p(y_i|f_i)$ by $t_i(f_i)$, namely, $\mathrm{argmin}_{t_i}\, \mathrm{KL}(p(\boldsymbol{f}|\mathcal{D})\|\widetilde{p}(\boldsymbol{f}|\mathcal{D}))$. However, this is still intractable. Instead, EP substitutes $q^{\setminus i}(\boldsymbol{f}) \propto q(\boldsymbol{f})/t_i(f_i)$ (cavity distribution) for $p^{\setminus i}(\boldsymbol{f}|\mathcal{D}) \propto p(\boldsymbol{f}|\mathcal{D})/p(y_i|f_i)$ (Li et al., 2015; Bui et al., 2017), expressed as:

$$q^{\setminus i}(\boldsymbol{f}) = p(\boldsymbol{f})\Big(\prod_{j\neq i} t_j(f_j)\Big)/q^{\setminus i}(\mathcal{D})$$
$$\approx p^{\setminus i}(\boldsymbol{f}) = p(\boldsymbol{f})\Big(\prod_{j\neq i} p(y_j|f_j)\Big)/p^{\setminus i}(\mathcal{D}),$$

where $q^{\setminus i}(\mathcal{D})$ and $p^{\setminus i}(\mathcal{D})$ are normalizers. This assumption in turn leads to an approximation $\widetilde{q}(\boldsymbol{f})$ (tilted distribution) to the true posterior $p(\boldsymbol{f}|\mathcal{D})$ and a new KL objective function to minimize:

$$p(\boldsymbol{f}|\mathcal{D}) \approx \widetilde{q}(\boldsymbol{f}) = \frac{q(\boldsymbol{f})p(y_i|f_i)p^{\setminus i}(\mathcal{D})q(\mathcal{D})}{t_i(f_i)q^{\setminus i}(\mathcal{D})p(\mathcal{D})},$$
$$\mathrm{KL}(\widetilde{q}(\boldsymbol{f})\|q(\boldsymbol{f})) = \mathrm{KL}(\widetilde{q}(f_i)\|q(f_i)).$$

The minimization of the KL objective function is realized by projecting the tilted distribution $\widetilde{q}(f_i)$ onto the Gaussian distribution space $\mathrm{proj}_{\mathrm{KL}}(\widetilde{q}(f_i)) = \mathrm{argmin}_{\mathcal{N}} \mathrm{KL}(\widetilde{q}(f_i)\|\mathcal{N}(f_i)) = \mathcal{N}(f_i|\mu^\star, \sigma^{\star 2})$, and then using the optimal Gaussian distribution to update $t_i(f_i) \propto \mathcal{N}(f_i)/q^{\setminus i}(f_i)$, where the optimum parameters $\mu^\star$ and $\sigma^{\star 2}$ can be found by simply matching its moments with $\widetilde{q}(f_i)$'s,

$$\mu^\star = \mu_{\widetilde{q}_i}, \quad \sigma^{\star 2} = \sigma_{\widetilde{q}_i}^2,$$

where $\mu_{\widetilde{q}_i}$ and $\sigma_{\widetilde{q}_i}^2$ are the mean and the variance of $\widetilde{q}(f_i)$. We summarize EP in Algorithm 1. In Sec. 4, we propose a new approximation algorithm similar to EP while exploiting the L^2 Wasserstein distance for local updates instead of the KL divergence.

3.3 Wasserstein Distance

We use $\mathcal{M}_+^1(\Omega)$ to denote the set of all probability measures on Ω. This work considers probability measures on

the d-dimensional real space $\mathbb{R}^d$. Intuitively speaking, the Wasserstein distance between two probability distributions $\mu, \nu \in \mathcal{M}_+^1(\mathbb{R}^d)$ is defined as the cost of transporting probability mass from one to the other. We are particularly interested in the subclass of L^p Wasserstein distances and its formal definition is presented as below.

Definition 1 (L^p Wasserstein distances)**.** *Consider the set of all probability measures on the product space $\mathbb{R}^d \times \mathbb{R}^d$, whose marginal measures are μ and ν respectively, denoted as $U(\mu, \nu)$. The L^p Wasserstein distance between μ and ν is defined as*

$$W_p(\mu,\nu) \equiv \left(\inf_{\pi \in U(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\boldsymbol{x}-\boldsymbol{z}\|_p^p \, d\pi(\boldsymbol{x},\boldsymbol{z}) \right)^{1/p}$$

where $p \in [1, \infty)$ and $\| \cdot \|_p$ is the L^p norm.

The Wasserstein distance has a number of important properties. Like the KL divergence, the Wasserstein distance has a minimum value of zero, achieved when two distributions are equivalent, but different from KL, it is symmetric. Another essential property, we exploit to provide our method with computational efficiency, is:

Proposition 1. *(Peyré et al., 2019, Remark 2.30) The L^p Wasserstein distance between one-dimensional distribution functions μ and $\nu \in \mathcal{M}_+^1(\mathbb{R})$ equals the L^p distance between the quantile functions of μ and ν*

$$W_p^p(\mu,\nu) = \int_0^1 |F_\mu^{-1}(y) - F_\nu^{-1}(y)|^p \, dy,$$

where $F_\mu : \mathbb{R} \to [0,1]$ is the cumulative distribution function (CDF) of μ, defined as $F_\mu(x) = \int_{-\infty}^{x} d\mu$, and F_μ^{-1} is the pseudoinverse or quantile funcntion, defined as $F_\mu^{-1}(y) = \min_x \{x \in \mathbb{R} \cup \{-\infty\} : F_\mu(x) \geq y\}$. Replacing μ with ν in these definitions yields F_ν and F_ν^{-1}.

We will often use probability density functions into the notation $\mathrm{W}(\cdot,\cdot)$. Besides, another important feature of the L^2 Wasserstein distance is on translating random variables. We exploit it in the proof of the economic parameterization.

Proposition 2. *(Peyré et al., 2019, Remark 2.19) Consider the L^2 Wasserstein distance defined for μ and $\nu \in \mathcal{M}_+^1(\mathbb{R}^d)$, and let $f_{\boldsymbol{\tau}}(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{\tau}$, $\boldsymbol{\tau} \in \mathbb{R}^d$, be a translation operator. If $\mu_{\boldsymbol{\tau}}$ and $\nu_{\boldsymbol{\tau}'}$ denote the probability measures of translated random variables $f_{\boldsymbol{\tau}}(\boldsymbol{x})$, $\boldsymbol{x} \sim \mu$, and $f_{\boldsymbol{\tau}'}(\boldsymbol{x})$, $\boldsymbol{x} \sim \nu$, respectively, then we have $W_2^2(\mu_{\boldsymbol{\tau}}, \nu_{\boldsymbol{\tau}'}) = W_2^2(\mu,\nu) - 2(\boldsymbol{\tau} - \boldsymbol{\tau}')^T(\boldsymbol{m}_\mu - \boldsymbol{m}_\nu) + \|\boldsymbol{\tau} - \boldsymbol{\tau}'\|_2^2$ where $\boldsymbol{m}_\mu$ and $\boldsymbol{m}_\nu$ are means of μ and ν respectively. In particular when $\boldsymbol{\tau} = \boldsymbol{m}_\mu$ and $\boldsymbol{\tau}' = \boldsymbol{m}_\nu$, $\mu_{\boldsymbol{\tau}}$ and $\nu_{\boldsymbol{\tau}'}$ become zero-mean measures, and there is*

$$W_2^2(\mu_{\boldsymbol{\tau}}, \nu_{\boldsymbol{\tau}'}) = W_2^2(\mu,\nu) - \|\boldsymbol{m}_\mu - \boldsymbol{m}_\nu\|_2^2.$$

4 Quantile Propagation

In this section, we propose a new approximation algorithm, which matches local tilted distributions by minimizing the L^2 Wasserstein distance instead of EP's forward KL divergence. As summarized in Algorithm 1, the difference lies in the L^2 Wasserstein distance based projection $\mathrm{proj}_{\mathrm{W}}(\widetilde{q}(f_i)) \equiv \mathrm{argmin}_{\mathcal{N}} \mathrm{W}_2^2(\widetilde{q}(f_i), \mathcal{N}(f_i))$. Although employing a more complicated divergence, our method enjoys the same computational time complexity as EP's and alleviates EP's over-estimation of variances.

Algorithm 1 Expectation (Quantile) Propagation

Input: $p(\boldsymbol{f}), p(y_i|f_i), t_i(f_i), i = 1, \cdots, N, \boldsymbol{\theta}$
Output: $q(\boldsymbol{f})$ approximate posterior

repeat
  repeat
    for $i = 1$ to N **do**
      $q^{\backslash i}(f_i) \propto q(f_i)/t_i(f_i)$ cavity distribution
      $\widetilde{q}(f_i) \propto q^{\backslash i}(f_i) p(y_i|f_i)$ tilted distribution
      $t_i(f_i) \leftarrow \mathrm{proj}_{\mathrm{KL}}[\widetilde{q}(f_i)]/q^{\backslash i}(f_i)$ by (3.2)
      (QP: $t_i(f_i) \leftarrow \mathrm{proj}_{\mathrm{W}}[\widetilde{q}(f_i)]/q^{\backslash i}(f_i)$) by (4.2)(4.2)
      $q(\boldsymbol{f}) \propto p(\boldsymbol{f}) \prod_i t_i(f_i)$ by (3.2)
    end for
  until convergence
  $\boldsymbol{\theta} = \mathrm{argmax}_{\boldsymbol{\theta}} \log q(\mathcal{D})$ by (3.2)
until convergence
return $q(\boldsymbol{f})$

More as per Proposition 1, minimizing $\mathrm{W}_2^2(\widetilde{q}(f_i), \mathcal{N}(f_i))$ is equivalent to minimizing the L^2 distance between quantile functions of $\widetilde{q}(f_i)$ and $\mathcal{N}(f_i)$, so we refer to our method as quantile propagation (QP). This section focuses on deriving formulas for local updates with approximating the non-Gaussian likelilhood $p(y_i|f_i)$ with an effficiently parameterized Gaussian likelihood $t_i(f_i)$. We later (in Sec. 5) show that such a parameterization is equivalent to a more general one using all $\boldsymbol{f}$.

4.1 Convexity of L^p Wasserstein Distance

We first show $\mathrm{W}_p^p(\widetilde{q}, \mathcal{N}(\mu, \sigma^2))$ to be strictly convex about μ and σ. Provided the convexity, we can resort to either elegant closed form expressions if available or for example the gradient descent to optimize μ and σ. Formally, we want to show the following theorem:

Theorem 1. *Given two univariate distributions: Gaussian $\mathcal{N}(\mu, \sigma^2)$, $\sigma \in \mathbb{R}_+$ [1], and arbitrary $\widetilde{q}$, the L^p Wasserstein distance $W_p^p(\widetilde{q}, \mathcal{N})$ is strictly convex about μ and σ.*

Proof. Let $F_{\widetilde{q}}^{-1}$ and $F_{\mathcal{N}}^{-1}(y) = \mu + \sqrt{2}\sigma \mathrm{erf}^{-1}(2y-1)$ be quantile functions of $\widetilde{q}$ and the Gaussian $\mathcal{N}$, where erf is the error function. Then, we consider two distinct Gaussian distributions $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ with $\sigma_1,\ \sigma_2 \in \mathbb{R}_+$ and a convex combination w.r.t. their parameters $\mathcal{N}(a_1\mu_1 + a_2\mu_2, (a_1\sigma_1 + a_2\sigma_2)^2)$ with $a_1, a_2 \in \mathbb{R}_+$ and $a_1 + a_2 = 1$.

[1] In this paper, $\mathbb{R}_+$ represents the set of positive real numbers.

Given the above, we further define $\varepsilon_k(y) = F_{\widetilde{q}}^{-1}(y) - \mu_k - \sigma_k\sqrt{2}\text{erf}^{-1}(2y-1)$, $k = 1, 2$, for notational simplification, and derive the convexity as below:

$$
\begin{aligned}
&\mathrm{W}_p^p(\widetilde{q}, \mathcal{N}(a_1\mu_1 + a_2\mu_2, (a_1\sigma_1 + a_2\sigma_2)^2)) \\
&\overset{(a)}{=} \int_0^1 |a_1\varepsilon_1(y) + a_2\varepsilon_2(y)|^p \,\mathrm{d}y \\
&\overset{(b)}{\leq} \int_0^1 \Big|a_1|\varepsilon_1(y)| + a_2|\varepsilon_2(y)|\Big|^p \,\mathrm{d}y \\
&\quad\overset{(c)}{\leq} a_1 \int_0^1 |\varepsilon_1(y)|^p \,\mathrm{d}y + a_2 \int_0^1 |\varepsilon_2(y)|^p \,\mathrm{d}y \\
&\quad\quad = a_1\mathrm{W}_p^p(\widetilde{q}, \mathcal{N}(\mu_1, \sigma_1^2)) + a_2\mathrm{W}_p^p(\widetilde{q}, \mathcal{N}(\mu_2, \sigma_2^2)),
\end{aligned}
$$

where (a) and (c) are due to Proposition 1 and the convexity of $f(x) = x^p$, $p \geq 1$, over $\mathbb{R}_+$ respectively. the equality at (b) holds iff (condition 1) $F_q^{-1}(y) = 1/2\sum_{k=1}^2(\mu_k + \sigma_k\sqrt{2}\text{erf}(2y-1))$, and (c)'s equality holds iff (condition 2) $\varepsilon_k(y) \geq 0, k = 1, 2, \forall y \in [0, 1]$. Two conditions can't be attained at the same time as the condition 1 is equivalent to $\varepsilon_1(y) = -\varepsilon_2(y)$ and then in terms of the condition 2, there is $\varepsilon_1(y) = \varepsilon_2(y)$, which contracts the fact that $\mathcal{N}(\mu_1, \sigma_1^2)$ is different from $\mathcal{N}(\mu_2, \sigma_2^2)$. Therefore, the L^p Wasserstein distance $\mathrm{W}_p^p(\widetilde{q}, \mathcal{N})$, $p \geq 1$, is strictly convex about μ and σ, which guarantees the uniqueness of minimum parameters. □

4.2 Minimization of L^2 Wasserstein Distance

Fortunately, the L^2 Wasserstein distance $\mathrm{W}_2^2(\widetilde{q}, \mathcal{N}(\mu, \sigma^2))$ has almost closed form expressions for optimal μ and σ, so yielding efficient local updates. We leave extensions of our method on other L^p, with $p \neq 2$, Wasserstein distances to the future work.

We obtain minimum parameters $\mu^\star$ and $\sigma^\star$ of the L^2 Wasserstein distance $\mathrm{W}_2^2(\widetilde{q}, \mathcal{N}(\mu, \sigma^2))$ based on Proposition 1. Specifically, we employ the quantile function based reformulation (Eqn. (1)) of $\mathrm{W}_2^2(\widetilde{q}, \mathcal{N}(\mu, \sigma^2))$, and zero its derivatives w.r.t. μ and σ. The results are directly provided with derivations left to Appx. A:

$$
\begin{aligned}
\mu^\star &= \mu_{\widetilde{q}}, \\
\sigma^\star &= \sqrt{2}\int_0^1 F_{\widetilde{q}}^{-1}(y)\text{erf}^{-1}(2y-1)\,\mathrm{d}y, \\
&= \sqrt{2}\int_{-\infty}^{\infty} x\text{erf}^{-1}(2F_{\widetilde{q}}(x) - 1)\widetilde{q}(x)\,\mathrm{d}x,
\end{aligned}
$$

where $F_{\widetilde{q}}$ is the CDF of $\widetilde{q}$. Eqn. (4.2) is used for analysis of QP's properties, and we turn to computate the local variance based on Eqn. (4.2) to get rid of the usually analytically intractable quantile functions. Interestingly, the updating equation for μ is same as EP's: both equal the mean of $\widetilde{q}$. If $\widetilde{q}$ has multiple modes, EP and QP match a Gaussian distribution with the average of modes. In terms of the local variance estimate, the optimum provided by QP, i.e., Eqn. (4.2), denoted as σ_{QP}, is always less or equal to EP's optimum, denoted as σ_{EP} and calculated as Eqn. (3.2):

Theorem 2. *$\sigma_{QP} \leq \sigma_{EP}$ where QP employs the L^2 Wasserstein distance.*

Proof. Let $\mathcal{N}(\mu_{\text{QP}}, \sigma_{\text{QP}}^2)$ be the optimal Gaussian in the QP algorithm with the L^2 Wasserstein distance. As per Proposition 1, we reformulate $\mathrm{W}_2^2(\widetilde{q}, \mathcal{N}(\mu_{\text{QP}}, \sigma_{\text{QP}}^2))$ in quantile functions:

$$
\begin{aligned}
&\mathrm{W}_2^2(\widetilde{q}, \mathcal{N}(\mu_{\text{QP}}, \sigma_{\text{QP}}^2)) \\
&= \int_0^1 |F_{\widetilde{q}}^{-1}(y) - \mu_{\text{QP}} - \sqrt{2}\sigma_{\text{QP}}\text{erf}^{-1}(2y-1)|^2 \,\mathrm{d}y \\
&\overset{(a)}{=} \int_0^1 \underbrace{(F_{\widetilde{q}}^{-1}(y) - \mu_{\text{QP}})^2}_{\text{(A)}} + \underbrace{(\sqrt{2}\sigma_{\text{QP}}\text{erf}^{-1}(2y-1))^2}_{\text{(B)}} \\
&\quad \underbrace{-2(F_{\widetilde{q}}^{-1}(y) - \mu_{\text{QP}})\sqrt{2}\sigma_{\text{QP}}\text{erf}^{-1}(2y-1)}_{\text{(C)}} \,\mathrm{d}y \\
&= \sigma_{\text{EP}}^2 - \sigma_{\text{QP}}^2.
\end{aligned}
$$

where the step (a) decomposes the Wasserstein distance into three components: the integral of the term (A) equals the variance estimated by EP, σ_{EP}^2, due to the equivalent mean estimates $\mu_{\text{QP}} = \mu_{\text{EP}} = \mu_{\widetilde{q}}$; calculation of the integral of (B) is straightforward and yields σ_{QP}^2; in (C), $\int \mu_{\text{QP}}\sigma_{\text{QP}}\text{erf}^{-1}(2y-1)\,\mathrm{d}y$ is zero and the left can be rewritten as $2\sigma_{\text{QP}}^2$ with utility of Eqn. (4.2). Therefore, the last line is obtained, and further due to the non-negativity of the Wasserstein distance, we have $\sigma_{\text{EP}}^2 \leq \sigma_{\text{QP}}^2$. Equality holds iff $\widetilde{q}$ is Gaussian. □

4.3 Implementations of Updating Formulas

Updates of the mean and the variance based on Eqn. (4.2) and (4.2) require the probability density function (PDF) and the CDF of the tilted distribution $\widetilde{q}$. As per the book of (Rasmussen and Williams, 2005), the cavity distribution is a Gaussian denoted as $q^{\setminus i}(f) = \mathcal{N}(f|\mu, \sigma^2)$, and the PDF of the tilted distribution is therefore expressed as $\widetilde{q}(f) \propto q^{\setminus i}(f)p(y|f)$. Typical choices of likelihoods include the probit likelihood $p(y|f) = \Phi(yf)$ for binary classification, where $\Phi(x) = \int_{-\infty}^{x}\mathcal{N}(x|0, 1)\,\mathrm{d}x$ is the cumulative distribution function of the standard Gaussian and $y \in \{-1, 1\}$, the Gaussian $p(y|f) = \mathcal{N}(y|f, \sigma_n^2)$ with some fixed variance σ_n^2 and $y \in \mathbb{R}$ for one-output regression, the Poisson likelihood $p(y|\lambda = \exp(f)) = \lambda^y\exp(-\lambda)/y!$ with $y \in \mathbb{N}$ for log Gaussian cox processes.

For the binary classification, the PDF of the tilted distribution $\widetilde{q}(f)$ with the probit likelihood is provided by Rasmussen and Williams (2005):

$$
\widetilde{q}(f) = Z^{-1}\Phi(yf)\mathcal{N}(f|\mu, \sigma^2),
$$

where $Z \equiv \Phi(ky)$ and $k \equiv \mu/\sqrt{1+\sigma^2}$, and the mean estimate also has a closed form expression:

$$\mu^\star = \mu_{\widetilde{q}} = \mu + \frac{\sigma^2 y \mathcal{N}(ky)}{\Phi(ky)\sqrt{1+\sigma^2}}.$$

The computation of the optimal $\sigma^\star$ requires the CDF of $\widetilde{q}$. Fortunately, it has a almost closed form expression:

$$F_{\widetilde{q}}(x) = Z^{-1}\left[\frac{1}{2}\Phi(h) - yT\left(h, \frac{k+\rho h}{h\sqrt{1-\rho^2}}\right) + \frac{y}{2}\Phi(k) - yT\left(k, \frac{h+\rho k}{k\sqrt{1-\rho^2}}\right) + y\eta\right],$$

where Z, k have been defined in Eqn. (4.3), $h \equiv (x-\mu)/\sigma$, $\rho \equiv 1/\sqrt{1+\sigma^{-2}}$, $T(\cdot,\cdot)$ is the Owen's T function:

$$T(h,a) = \frac{1}{2\pi}\int_0^a \frac{\exp\big[-(1+x^2)h^2/2\big]}{1+x^2}\,\mathrm{d}x,$$

and η is defined as

$$\eta = \begin{cases} 0 & hk > 0 \text{ or } (hk = 0 \text{ and } h + k \geq 0), \\ -0.5 & \text{otherwise.} \end{cases}$$

Provided the above, the optimal $\sigma^\star$ can be computed by Monte Carlo sampling:

$$\sigma^\star \approx \frac{\sqrt{2}}{ZN}\sum_{j=1}^{J} x_j \mathrm{erf}^{-1}(2F_{\widetilde{q}}(x_j) - 1)p(y|x_j),$$

where x_j, $j = 1, \cdots, J$, are sampled from the Gaussian $q^{\setminus i}(f|\mu, \sigma^2)$. We observe that $\sigma^\star$ is a function of three parameters: the cavity distribution's μ and σ, and the output y, so we resort to two lookup tables to accelerate QP in practice. Two tables are built for $y = 1$ and $y = -1$ respectively, and each table is two-dimensional, containing values of the optimal $\sigma^\star$ over a regular grid of μ and σ. As a result, the local updates of QP and EP can be implemented in the same computational time complexity of $O(1)$.

5 Economic Parameterization

In this section, we prove the efficient parameterization for QP, namely, the optimal site function t_i only depends on the local latent variable f_i instead of all latent variables $\boldsymbol{f}$.

5.1 Review: Proof for EP

We first review the proof for EP (Seeger, 2005). Let's define a more general site function $t_i(\boldsymbol{f})$ approximating the likelihood $p(y_i|f_i)$, and the cavity and the tilted distributions are expressed as $q^{\setminus i}(\boldsymbol{f}) \propto p(\boldsymbol{f})\prod_{j\neq i}\widetilde{t}_j(\boldsymbol{f})$ and $\widetilde{q}(\boldsymbol{f}) \propto q^{\setminus i}(\boldsymbol{f})p(y_i|f_i)$. To update $t_i(\boldsymbol{f})$, EP matches a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{f})$ to $\widetilde{q}(\boldsymbol{f})$ by minimizing the KL divergence $\mathrm{KL}(\widetilde{q}\|\mathcal{N})$ which is rewritten as (see details in Appx. C.1):

$$\mathrm{KL}\big(\widetilde{q}\|\mathcal{N}\big) = \mathrm{KL}\big(\widetilde{q}_i\|\mathcal{N}_i\big) + \mathbb{E}_{\widetilde{q}_i}\Big[\mathrm{KL}\big(q^{\setminus i}_{\setminus i|i}\|\mathcal{N}_{\setminus i|i}\big)\Big],$$

where and hereinafter, $\setminus i|i$ represents the conditional distribution of $\boldsymbol{f}_{\setminus i}$ (leaving f_i out from $\boldsymbol{f}$) given f_i. In other words, $q^{\setminus i}_{\setminus i|i} = q^{\setminus i}(\boldsymbol{f}_{\setminus i}|f_i)$ and $\mathcal{N}_{\setminus i|i} = \mathcal{N}(\boldsymbol{f}_{\setminus i}|f_i)$. We minimize this KL divegence by setting $q^{\setminus i}_{\setminus i|i} = \mathcal{N}_{\setminus i|i}$ to zero the term $\mathbb{E}_{\widetilde{q}_i}[\mathrm{KL}\big(q^{\setminus i}_{\setminus i|i}\|\mathcal{N}_{\setminus i|i}\big)]$ as $q^{\setminus i}_{\setminus i|i}$ is Gaussian and matching moments between $\widetilde{q}_i$ and $\mathcal{N}_i$ to minimize $\mathrm{KL}\big(\widetilde{q}_i\|\mathcal{N}_i\big)$. Finally, the site function t_i is updated by dividing the optimal Gaussian $\mathcal{N}(\boldsymbol{f})$ by the cavity distribution $q^{\setminus i}(\boldsymbol{f})$:

$$\begin{aligned} t_i(\boldsymbol{f}) &\propto \mathcal{N}(\boldsymbol{f})/q^{\setminus i}(\boldsymbol{f}) \\ &= \cancel{\mathcal{N}(\boldsymbol{f}_{\setminus i}|f_i)}\mathcal{N}(f_i)/(\cancel{q^{\setminus i}(\boldsymbol{f}_{\setminus i}|f_i)}q^{\setminus i}(f_i)) \\ &= \mathcal{N}(f_i)/q^{\setminus i}(f_i). \end{aligned}$$

Thus, the optimal site function t_i relies solely on the local latent variable f_i.

5.2 Proof for QP

This section proves the economic parameterization for QP. We first show the following theorem, and then follow the same steps in Eqn. (5.1), it is simple to show that the optimal site function t_i replies on only f_i.

Theorem 3. *Minimization of* $\mathrm{W}_2^2(\widetilde{q}(\boldsymbol{f}), \mathcal{N}(\boldsymbol{f}))$ *w.r.t.* $\mathcal{N}(\boldsymbol{f})$ *results in*

$$q^{\setminus i}(\boldsymbol{f}_{\setminus i}|f_i) = \mathcal{N}(\boldsymbol{f}_{\setminus i}|f_i).$$

Proof. **(a)** We rewrite the $\mathrm{W}_2^2(\widetilde{q}(\boldsymbol{f}), \mathcal{N}(\boldsymbol{f}))$ as below (see detailed derivations in Appx. C.2):

$$\mathrm{W}_2^2(\widetilde{q},\mathcal{N}) = \inf_{\pi_i}\mathbb{E}_{\pi_i}\Big[\|f_i - f_i'\|_2^2 + \mathrm{W}_2^2(q^{\setminus i}_{\setminus i|i}, \mathcal{N}_{\setminus i|i})\Big],$$

where the prime indicates that the variable is from the Gaussian $\mathcal{N}$, $\pi_i = \pi(f_i, f_i')$ and the inf is over $U(\widetilde{q}_i, \mathcal{N}_i)$. $q^{\setminus i}(\boldsymbol{f})$ is known to be Gaussian and WLOG defined as

$$q^{\setminus i}(\boldsymbol{f}) \equiv \mathcal{N}\left(\begin{bmatrix}\boldsymbol{f}_{\setminus i} \\ f_i\end{bmatrix} \,\middle|\, \begin{bmatrix}\boldsymbol{m}_{\setminus i} \\ m_i\end{bmatrix}, \begin{bmatrix}\boldsymbol{S}_{\setminus i} & \boldsymbol{S}_{\setminus ii} \\ \boldsymbol{S}_{\setminus ii}^{\mathrm{T}} & S_i\end{bmatrix}\right),$$

where $\boldsymbol{m}_{\setminus i}$ and $\boldsymbol{S}_{\setminus i}$ are $\boldsymbol{f}_{\setminus i}$'s mean and covariance, and m_i and S_i are f_i's. It follows that the conditional $q^{\setminus i}(\boldsymbol{f}_{\setminus i}|f_i)$ is expressed as:

$$\begin{aligned} &q^{\setminus i}(\boldsymbol{f}_{\setminus i}|f_i) = \mathcal{N}(\boldsymbol{f}_{\setminus i}|\boldsymbol{m}_{\setminus i|i}, \boldsymbol{S}_{\setminus i|i}), \\ &\boldsymbol{m}_{\setminus i|i} = \boldsymbol{m}_{\setminus i} + \boldsymbol{S}_{\setminus ii}S_i^{-1}(f_i - m_i) \equiv \boldsymbol{a}f_i + \boldsymbol{b}, \\ &\boldsymbol{S}_{\setminus i|i} = \boldsymbol{S}_{\setminus i} - \boldsymbol{S}_{\setminus ii}S_i^{-1}\boldsymbol{S}_{\setminus ii}^{\mathrm{T}}. \end{aligned}$$

Similar to $q^{\backslash i}(\boldsymbol{f})$, the Gaussian $\mathcal{N}$ is written as:

$$\mathcal{N}(\boldsymbol{f}') \equiv \mathcal{N}\left(\begin{bmatrix}\boldsymbol{f}'_{\backslash i}\\ f'_i\end{bmatrix} \middle| \begin{bmatrix}\boldsymbol{m}'_{\backslash i}\\ m'_i\end{bmatrix}, \begin{bmatrix}\boldsymbol{S}'_{\backslash i} & \boldsymbol{S}'_{\backslash ii}\\ \boldsymbol{S}'^{\mathrm{T}}_{\backslash ii} & S'_i\end{bmatrix}\right)$$

and the conditional is correspondingly expressed as:

$$\begin{aligned}
&\mathcal{N}(\boldsymbol{f}'_{\backslash i}|f'_i) = \mathcal{N}(\boldsymbol{m}'_{\backslash i|i}, \boldsymbol{S}'_{\backslash i|i}),\\
&\boldsymbol{m}'_{\backslash i|i} = \boldsymbol{m}'_{\backslash i} + \boldsymbol{S}'_{\backslash ii}S'^{-1}_i(f'_i - m'_i) \equiv \boldsymbol{a}'f'_i + \boldsymbol{b}',\\
&\boldsymbol{S}'_{\backslash i|i} = \boldsymbol{S}'_{\backslash i} - \boldsymbol{S}'_{\backslash ii}S'^{-1}_i\boldsymbol{S}'^{\mathrm{T}}_{\backslash ii}.
\end{aligned}$$

Given the above expressions, we exploit Proposition 2 to rewrite Eqn. (5.2) as:

$$\begin{aligned}
\mathrm{W}_2^2(\widetilde{q}, \mathcal{N}) = \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i - f'_i\|_2^2 + \|\boldsymbol{m}_{\backslash i|i} - \boldsymbol{m}'_{\backslash i|i}\|_2^2\Big]\\
+ \underbrace{\mathrm{W}_2^2\Big(\mathcal{N}(\boldsymbol{0}, \boldsymbol{S}_{\backslash i|i}), \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}'_{\backslash i|i})\Big)}_{\text{(A)}}.
\end{aligned}$$

To minimize this expression, we need to optimize m'_i,$\boldsymbol{m}'_{\backslash i}$, S'_i, $\boldsymbol{S}'_{\backslash i}$ and $\boldsymbol{S}'_{\backslash ii}$. Note that among them, $\boldsymbol{S}'_{\backslash i}$ is only contained in the term (A) and is optimized by simply setting

$$\boldsymbol{S}'_{\backslash i|i} = \boldsymbol{S}_{\backslash i|i} \overset{\text{Eqn. (5.2)}}{\Longrightarrow} \boldsymbol{S}^{(n)*}_{\backslash i} = \boldsymbol{S}_{\backslash i|i} + \boldsymbol{S}'_{\backslash ii}S'^{-1}_i\boldsymbol{S}'^{\mathrm{T}}_{\backslash ii}.$$

As a result, Part A is minimized to zero.

Next, we plug in expressions of $\boldsymbol{m}_{\backslash i|i}$ and $\boldsymbol{m}'_{\backslash i|i}$ (Eqs. (5.2) and (5.2)) into optimized Eqn. (5.2):

$$\begin{aligned}
\min_{\boldsymbol{S}'_{\backslash i}} \text{Eqn. (5.2)} = \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i - f'_i\|_2^2 +\\
\|\boldsymbol{a}f_i - \boldsymbol{a}'f'_i + \boldsymbol{b} - \boldsymbol{b}'\|_2^2\Big].
\end{aligned}$$

Note that $\boldsymbol{m}'_{\backslash i}$ is only contained in $\boldsymbol{b}'$, so we optimize it by zeroing the above function's derivative about $\boldsymbol{m}'_{\backslash i}$, which results in:

$$\begin{aligned}
&\boldsymbol{b}' = \boldsymbol{b} + \boldsymbol{a}\mu_{\widetilde{q}_i} - \boldsymbol{a}'m'_i \overset{\text{Eqn. (5.2)}}{\Longrightarrow}\\
&\boldsymbol{m}^{(n)*}_{\backslash i} = \boldsymbol{S}'_{\backslash ii}S'^{-1}_i m'_i + \boldsymbol{b} + \boldsymbol{a}\mu_{\widetilde{q}_i} - \boldsymbol{a}'m'_i,
\end{aligned}$$

where $\mu_{\widetilde{q}_i}$ is the mean of $\widetilde{q}(f_i)$. Consequently, the minimum value of Eqn. (5.2) becomes (see detailed derivations in Appx. C.3):

$$\begin{aligned}
\min_{\boldsymbol{m}'_{\backslash i}} \text{Eqn. (5.2)} = (1 + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}')\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2\sigma^2_{\widetilde{q}_i} +\\
\|\boldsymbol{a}'\|_2^2 S'_i - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}'\Big[\sigma^2_{\widetilde{q}_i} + S'_i + (\mu_{\widetilde{q}_i} - m'_i)^2\Big],
\end{aligned}$$

where $\sigma^2_{\widetilde{q}_i}$ is the variance of $\widetilde{q}(f_i)$. We note that $\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i)$ is the Wasserstein distance between two univariate distributions, so we exploit Proposition 1 to reformulate it:

$$\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) = \sigma^2_{\widetilde{q}_i} + (\mu_{\widetilde{q}_i} - m'_i)^2 + S'_i - 2\sqrt{2S'_i}c_{\widetilde{q}_i},$$

where $c_{\widetilde{q}_i} \equiv \int_0^1 F^{-1}_{\widetilde{q}_i}(y)\mathrm{erf}^{-1}(2y-1)\,\mathrm{d}y$. Eqn. (5.2) is therefore simplified by plugging the above reformulation into it (see detailed derivations in Appx. C.4):

$$\begin{aligned}
\text{Eqn. (5.2)} = \mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2\sigma^2_{\widetilde{q}_i} -\\
\underbrace{2\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}'c_{\widetilde{q}_i}\sqrt{2S'_i} + \|\boldsymbol{a}'\|_2^2 S'_i}_{\text{(C)}}.
\end{aligned}$$

Now, we are left with optimizing m'_i, S'_i and $\boldsymbol{S}'_{\backslash ii}$. We optimize $\boldsymbol{S}'_{\backslash ii}$, only existing in the above term (C), by zeroing the derivative of the above function w.r.t. $\boldsymbol{S}'_{\backslash ii}$:

$$\boldsymbol{a}'^{*} = \sqrt{2}(S'_i)^{-1/2}c_{\widetilde{q}_i}\boldsymbol{a} \overset{\text{Eqn. (5.2)}}{\Longrightarrow} \boldsymbol{S}'^{*}_{\backslash ii} = \sqrt{2}(S'_i)^{1/2}c_{\widetilde{q}_i}\boldsymbol{a}.$$

The corresponding minimum value of Eqn. (5.2) is

$$\min_{\boldsymbol{S}'_{\backslash ii}} \text{Eqn. (5.2)} = \mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2(\sigma^2_{\widetilde{q}_i} - 2c^2_{\widetilde{q}_i}).$$

The results of optimizing m'_i and S'_i have already been provided in Eqs. (4.2) and (4.2): $m'^{*}_i = \mu_{\widetilde{q}_i}$ and $S'^{*}_i = 2c^2_{\widetilde{q}_i}$. Plug them into Eqs. (5.2) and (5.2), and we have $\boldsymbol{a}'^{*} = \boldsymbol{a}$ and $\boldsymbol{b}'^{*} = \boldsymbol{b}$. Based on the two equations and Eqn. (5.2), we have $q^{\backslash i}(\boldsymbol{f}_{\backslash i}|f_i)(= \mathcal{N}(\boldsymbol{f}_{\backslash i}|\boldsymbol{a}f_i + \boldsymbol{b}, \boldsymbol{S}_{\backslash i|i})) = \mathcal{N}(\boldsymbol{f}_{\backslash i}|f_i)(= \mathcal{N}(\boldsymbol{f}_{\backslash i}|\boldsymbol{a}'f_i + \boldsymbol{b}', \boldsymbol{S}'_{\backslash i|i}))$. Furthermore, based on this conclusion and taking the same steps in Eqn. (5.1), the economic parameterization is proved. □

5.3 Benefits from Economic Parameterization

Consider the full parameterization for the approximate likelihood $t_i(\boldsymbol{f}) = \widetilde{Z}_i\mathcal{N}(\boldsymbol{f}|\widetilde{\boldsymbol{\mu}}_i, \widetilde{\boldsymbol{\Sigma}}_i)$. It has N random variables $\boldsymbol{f}$, a N-dimenstional mean $\widetilde{\boldsymbol{\mu}}_i$ and a $N \times N$ covariance matrix $\widetilde{\boldsymbol{\Sigma}}_i$. As a result, the approximate posterior distribution $q(\boldsymbol{f})$ with N likelihoods owns $O(N^3)$ indeterminate parameters and consumes $O(N^3)$ memory. And, in each local update, $O(N^2)$ parameters need optimizing.

In constrast, the economic parameterization for the approximate likelihood $t_i(f_i) = \widetilde{Z}_i\mathcal{N}(f_i|\widetilde{\mu}_i, \widetilde{\sigma}_i^2)$ has only one random variables f_i, one mean $\widetilde{\mu}_i$ and one variance $\widetilde{\sigma}_i^2$. Consequently, the approximate posterior distribution has $O(N)$ indeterminate parameters and consumes $O(N^2)$ memory dominated by the covariance matrix of the prior distribution. In terms of a local update, less computational time is required due to $O(1)$ parameters being optimized, which is significantly less than $O(N^2)$ parameters of the full parameterization.

6 Unified Interpretations

In this section, we intend to provide unified interpretations of EP and QP algorithms.

6.1 Interpretation of EP

Consider an intractable global objective function of the forward KL divergence between the true and the approx-

imate posterior distributions $\mathrm{KL}(p(\boldsymbol{f}|\mathcal{D})\|q(\boldsymbol{f}))$. EP randomly selects a factor $t_i(f_i)$ to optimize by minimizing this function. For tractability, EP makes the assumption of $p^{\backslash i}(\boldsymbol{f}|\mathcal{D}) \approx q^{\backslash i}(\boldsymbol{f})$ as mentioned in Sec. 3.2, which results in two more approximations: the true posterior being approximated by the tilted distribution $p(\boldsymbol{f}|\mathcal{D}) \approx \widetilde{q}(\boldsymbol{f})$ and $\mathrm{KL}(p(\boldsymbol{f}|\mathcal{D})\|q(\boldsymbol{f})) \approx \mathrm{KL}(\widetilde{q}(\boldsymbol{f})\|q(\boldsymbol{f}))$. Further as per the simplification in Eqn. (3.2), EP turns out to minimize the local KL objective function $\mathrm{KL}(\widetilde{q}(f_i)\|q(f_i))$. Concretely, EP matches a Gaussian $\mathcal{N}(f_i)$ to $\widetilde{q}(f_i)$ by minimizing the forward KL divergence and then uses the Gaussian to update $t_i(f_i) \propto \mathcal{N}(f_i)/q^{\backslash i}(f_i)$. In this way, we can regard EP as a coordinate descent algorithm to the global objective function $\mathrm{KL}(p(\boldsymbol{f}|\mathcal{D})\|q(\boldsymbol{f}))$ under the assumption $p^{\backslash i}(\boldsymbol{f}|\mathcal{D}) \approx q^{\backslash i}(\boldsymbol{f})$.

6.2 Interpretation of QP

Similar to EP, QP considers an intractable global objective function of the L^2 Wasserstein distance between the true and the approximate posterior distributions $\mathrm{W}_2^2(p(\boldsymbol{f}|\mathcal{D}), q(\boldsymbol{f}))$. QP also randomly selects a factor $t_i(f_i)$ to optimize and makes the same assumption of $p^{\backslash i}(\boldsymbol{f}|\mathcal{D}) \approx q^{\backslash i}(\boldsymbol{f})$, which causes an approximation to the L^2 Wasserstein objective function $\mathrm{W}_2^2(p(\boldsymbol{f}|\mathcal{D}), q(\boldsymbol{f})) \approx \mathrm{W}_2^2(\widetilde{q}(\boldsymbol{f}), q(\boldsymbol{f}))$. To optimize $t_i(f_i)$, QP fits a Gaussian $\mathcal{N}(\boldsymbol{f})$ to $\widetilde{q}(\boldsymbol{f})$ by minimizing the L^2 Wasserstein distance, and then updates $t_i(f_i) \propto \mathcal{N}(\boldsymbol{f})/q^{\backslash i}(\boldsymbol{f})$. As per what has been concluded in Theorem 5.2, QP finally solves $\mathcal{N}^\star(f_i) = \mathrm{argmin}_{\mathcal{N}_i} \mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i)$ (Eqn. (5.2)), and updates $t_i(f_i) \propto \mathcal{N}^\star(f_i)/q^{\backslash i}(f_i)$. In this way, we can regard QP as a coordinate descent algorithm to the global objective function $\mathrm{W}_2^2(p(\boldsymbol{f}|\mathcal{D}), q(\boldsymbol{f}))$ under the assumption $p^{\backslash i}(\boldsymbol{f}|\mathcal{D}) \approx q^{\backslash i}(\boldsymbol{f})$. Interestingly, such analysis provides unified interpretations for EP and QP. That is, EP and QP are coordinate descent algorithms to a forward KL and a L^2 Wasserstein objective function respectively under the same assumption.

7 Experiments

In this section, we compare QP and EP algorithms on the GP binary classification. The experiments employ various real world data and aim to compare relative accuracy of two different approximations rather than optimizing the absolute performance.

Performance Evaluation To evaluate the performance, we employ two measures: the error rate (**E**) and the test log-likelihood (**TLL**). E quantifies the prediction accuracy for the binary classification while TLL measures the prediction uncertainty.

Benchmark Data We use the six real life datasets employed in the work of Kuss and Rasmussen (2005): Ionosphere, Wisconsin Breast Cancer, Sonar (Dua and Graff, 2017), Leptograpsus Crabs, Pima Indians Diabetes (Ripley, 1996) and the USPS digit data (Hull, 1994). The USPS digit data consist of 0 - 9 and images of 3 and 5 are used for the binary classification. Besides, we use extra UCI open datasets: Iris and Wine (Dua and Graff, 2017). Iris and Wine have three classes and experiments of binary classification are conduct on every two different classes. We summarize the dataset size and data dimensions in Table 1.

Prediction In the training stage, EP and QP optimize the site parameters $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\boldsymbol{\Sigma}}$ and GP hyper-parameters $\boldsymbol{\theta}$. Given these, we compute the approximate predictive distribution for the binary target y_* with the input of $\boldsymbol{x}_*$ in the test stage. Let K_* denote the $N \times 1$ covariance matrix evaluated at all pairs of training and test inputs, $(\boldsymbol{x}_i, \boldsymbol{x}_*)$, $i = 1, \cdots, N$, and $k_* = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$, and then we have the following predictive distribution:

$$q(y_* = 1|\mathcal{D}) = \Phi\left(\frac{K_*^{\mathrm{T}}(K + \widetilde{\Sigma})^{-1}\widetilde{\boldsymbol{\mu}}}{\sqrt{1 + k_* - K_*^{\mathrm{T}}(K + \widetilde{\Sigma})^{-1}K_*}}\right).$$

Experiment Settings In the experiment, we split the each dataset into 10 folds and each time we use 1 fold as the test set and other 9 folds as the training set. As a result, totally 10 experiments are run on each dataset and the averaged results are summarized in Table 1. For EP and QP, we set the maximal iteration of the inner repeat loop of Algorithm 1 as 10, and for the outer loop, we use the minimization optimizer implemented in the book of Rasmussen and Williams (2005) with the maximal iteration of 40.

Results The result table (Table 1) consists of two sections. The top one shows the results on the same datasets employed by Kuss and Rasmussen (2005), and our results about EP are close to theirs. Compared with EP, QP gains lower error rates on a half of datasets and same eror rates on one third of datasets, while performs worse on the Pima Indians dataset. In perspective of the TLL, QP surpasses EP on all datasets except the Pima Indians dataset. The bottom section employs additional datasets to provide more evidence of QP's alleviation of EP's over-estimation of variances. Although the error rates are not improved significantly on these experiments, the uncertainty measures of QP are always better than EP's.

8 Conclusions

In this work, we propose an approximate Bayesian inference for Gaussian process models with factorizable Gaussian likelihoods and one-output latent functions. Like EP, the proposed method approximates the non-Gaussian likelihood by the Gaussian likelihood and optimizes the Gaussian likelihood in the local updates, but it minimizes the L^2

			E (%)		TLL	
Dataset	n	m	EP	QP	EP	QP
Ionosphere	351	34	10.71	**6.79**	-13.70	**-9.01**
Cancer	683	9	3.24	3.24	-7.29	**-6.74**
Pima Indians	732	7	**14.29**	17.42	**-22.72**	-30.46
Crabs	200	7	3.0	3.0	-1.92	**-1.78**
Sonar	208	60	14.12	**13.53**	-16.46	**-11.94**
USPS (3 vs 5)	1540	256	3.44	**2.92**	-12.24	**-10.63**
Iris (setosa vs versicolor)	100	4	0	0	-4.97×10^{-2}	-4.97×10^{-2}
Iris (setosa vs virginica)	100	4	0	0	-3.28×10^{-2}	$\mathbf{-3.23\times10^{-2}}$
Iris (versicolor vs virginica)	100	4	7	7	-2.46	**-2.07**
Wine (classes 1 vs 2)	130	13	**3.08**	3.85	-1.34	**-1.27**
Wine (classes 1 vs 3)	107	13	0	0	-4.99×10^{-2}	$\mathbf{-4.87\times10^{-2}}$
Wine (classes 2 vs 3)	119	13	3.64	**2.73**	-8.19×10^{-1}	$\mathbf{-8.17\times10^{-1}}$

Table 1: Results on Benchmark Datasets. The first three columns give the name of the dataset, the number of instances m and the number of features n. For two methods, the table records the average error rate (E) and the test log likelihood (TLL). The top section is on the benchmark datasets employed by Kuss and Rasmussen (2005) and the bottom section uses additional datasets to further demonstrate the performance of our method.

Wasserstein distance instead of the KL divergence. Consequently, the local update matches the quantile functions of local one-dimensional distributions rather than moments by EP. Interestingly, both of our method and EP update the Gaussian mean to be the mean of the tilted distribution, while the local variance estimate by our method is smaller than EP's, which results in alleviating EP's deficiency of over-estimating variances. We further show the economic parameterization property of our method – the optimal approximate likelihood relies on merely a single latent variable, and such a property reduces memory consumption by a factor N, i.e., the number of data, and computational time through optimizing a much less number ($O(1)$, vs $O(N^2)$ for the full parameterization) of parameters in each local update. Also, we deliver unified interpretations of EP and the proposed method respectively – we regard both methods as coordinate descent algorithms to a KL and a L^2 Wasserstein global objective function under the same approximation assumption. Our experiments employing a wide range of real world datasets show that our method outperforms EP on the task of Gaussian process binary classification, and also provides empirical evidence for alleviation of over-estimation of variances.

References

Ambrogioni, L., Güçlü, U., Güçlütürk, Y., Hinne, M., van Gerven, M. A., and Maris, E. (2018). Wasserstein variational inference. In *Advances in Neural Information Processing Systems*, pages 2473–2482.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Benamou, J.-D. and Brenier, Y. (2000). A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015). Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45.

Bui, T. D., Yan, J., and Turner, R. E. (2017). A unifying framework for gaussian process pseudo-point approximations using power expectation propagation. *J. Mach. Learn. Res.*, 18(1):3649–3720.

Courty, N., Flamary, R., Tuia, D., and Corpetti, T. (2016). Optimal transport for data fusion in remote sensing. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3571–3574. IEEE.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300.

Dehaene, G. and Barthelmé, S. (2018). Expectation propagation in the large data limit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):199–217.

Dezfouli, A. and Bonilla, E. V. (2015). Scalable inference for gaussian process models with black-box likelihoods. In *Advances in Neural Information Processing Systems*, pages 1414–1422.

Dowson, D. and Landau, B. (1982). The frchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450 – 455.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

El Moselhy, T. A. and Marzouk, Y. M. (2012). Bayesian

inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850.

Gelbrich, M. (1990). On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203.

Gelman, A., Vehtari, A., Jylänki, P., Sivula, T., Tran, D., Sahai, S., Blomstedt, P., Cunningham, J. P., Schiminovich, D., and Robert, C. (2017). Expectation propagation as a way of life: A framework for bayesian inference on partitioned data. *arXiv preprint arXiv:1412.4869*.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777.

Heess, N., Tarlow, D., and Winn, J. (2013). Learning to pass expectation propagation messages. In *Advances in Neural Information Processing Systems*, pages 3219–3227.

Hensman, J., Matthews, A., and Ghahramani, Z. (2015). Scalable variational gaussian process classification. *Journal of Machine Learning Research*.

Hensman, J., Zwießele, M., and Lawrence, N. (2014). Tilted variational bayes. In *Artificial Intelligence and Statistics*, pages 356–364.

Hernández-Lobato, J. M., Li, Y., Rowland, M., Hernández-Lobato, D., Bui, T., and Turner, R. (2016). Black-box α-divergence minimization. *International Conference on Machine Learning*.

Huang, G., Guo, C., Kusner, M. J., Sun, Y., Sha, F., and Weinberger, K. Q. (2016). Supervised word mover's distance. In *Advances in Neural Information Processing Systems*, pages 4862–4870.

Hull, J. J. (1994). A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Jylänki, P., Vanhatalo, J., and Vehtari, A. (2011). Robust gaussian process regression with a student-t likelihood. *Journal of Machine Learning Research*, 12(Nov):3227–3257.

Kuss, M. and Rasmussen, C. E. (2005). Assessing approximate inference for binary gaussian process classification. *Journal of machine learning research*, 6(Oct):1679–1704.

Li, Y., Hernández-Lobato, J. M., and Turner, R. E. (2015). Stochastic expectation propagation. In *Advances in neural information processing systems*, pages 2323–2331.

Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081.

Mallasto, A. and Feragen, A. (2017). Learning from uncertain curves: The 2-wasserstein metric for gaussian processes. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5660–5670. Curran Associates, Inc.

Minka, T. (2004). Power ep. *Dep. Statistics, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep*.

Minka, T. (2005). Divergence measures and message passing. Technical report, Microsoft Research.

Minka, T. P. (2001a). The ep energy function and minimization schemes. Technical report, Technical report.

Minka, T. P. (2001b). Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482.

Montavon, G., Müller, K.-R., and Cuturi, M. (2016). Wasserstein training of restricted boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 3718–3726.

Muzellec, B. and Cuturi, M. (2018). Generalizing point embeddings using the wasserstein space of elliptical distributions. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 10258–10269, USA. Curran Associates Inc.

Neal, R. M. (1997). Monte carlo implementation of gaussian process models for bayesian regression and classification. *arXiv preprint physics/9701026*.

Opper, M. and Winther, O. (2000). Gaussian processes for classification: Mean-field algorithms. *Neural computation*, 12(11):2655–2684.

Opper, M. and Winther, O. (2005). Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6(Dec):2177–2204.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121.

Seeger, M. (2005). Expectation propagation for exponential families. Technical report, Department of EECS, University of California at Berkeley.

Shafieezadeh-Abadeh, S., Nguyen, V. A., Kuhn, D., and Esfahani, P. M. (2018). Wasserstein distributionally robust kalman filtering. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 8483–8492, USA. Curran Associates Inc.

Snelson, E., Ghahramani, Z., and Rasmussen, C. E. (2004). Warped gaussian processes. In Thrun, S., Saul, L. K., and Scholkopf, B., editors, *Advances in Neural Information Processing Systems 16*, pages 337–344. MIT Press.

Takatsu, A. (2011). Wasserstein geometry of gaussian measures. *Osaka J. Math.*, 48(4):1005–1026.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. (2017). Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*.

Williams, C. K. and Barber, D. (1998). Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351.

Williams, C. K. and Rasmussen, C. E. (1996). Gaussian processes for regression. In *Advances in neural information processing systems*, pages 514–520.

Williams, C. K. I. and Barber, D. (1998). Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351.

Xu, M., Lakshminarayanan, B., Teh, Y. W., Zhu, J., and Zhang, B. (2014). Distributed bayesian posterior sampling via moment sharing. In *Advances in Neural Information Processing Systems*.

Accompanying the submission *Quantile Propagation for Wasserstein-Approximate Gaussian Processes*.

A Minimum of L^2asserstein Distance between Univariate Gaussian and Non-Gaussian Distributions

In this section, we prove the formulas (Eqs. (4.2) and (4.2)) of the minimum μ^* and σ^* for $p = 2$. Recall the optimization problem: we want to use a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ to fit a non-Gaussian distribution q by minimizing the L^2 Wasserstein distance between them:

$$\min_{\mu,\sigma} \mathrm{W}_2^2(q, \mathcal{N}) = \min_{\mu,\sigma} \int_0^1 \left| F_q^{-1}(y) - \mu - \sqrt{2}\sigma \mathrm{erf}^{-1}(2y-1) \right|^2 \mathrm{d}y,$$

where F_q^{-1} is the quantile function of the non-Gaussian distribution q, namely the pseudoinverse function of the corresponding cumulative distribution function F_q.

To solve this problem, we first calculate derivatives about μ and σ:

$$\begin{aligned}
\frac{\partial \mathrm{W}_2^2}{\partial \mu} &= -2 \int_0^1 F_q^{-1}(y) - \mu - \sqrt{2}\sigma \mathrm{erf}^{-1}(2y-1) \,\mathrm{d}y, \\
\frac{\partial \mathrm{W}_2^2}{\partial \sigma} &= -2 \int_0^1 (F_q^{-1}(y) - \mu - \sqrt{2}\sigma \mathrm{erf}^{-1}(2y-1)) \sqrt{2} \mathrm{erf}^{-1}(2y-1) \,\mathrm{d}y.
\end{aligned}$$

Then, by zeroing derivatives, we obtain the optimal parameters:

$$\begin{aligned}
\mu^* &= \int_0^1 F_q^{-1}(y) - \sqrt{2}\sigma \mathrm{erf}^{-1}(2y-1) \,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} x q(x) \,\mathrm{d}x - \frac{\sqrt{2}}{2}\sigma \int_{-1}^{1} \mathrm{erf}^{-1}(y) \,\mathrm{d}y \\
&= \mu_q - \sqrt{2}\sigma \int_{-\infty}^{\infty} x \mathcal{N}(x|0, 1/2) \,\mathrm{d}x \\
&= \mu_q, \\
\sigma^* &= \sqrt{2} \int_0^1 (F_q^{-1}(y) - \mu) \mathrm{erf}^{-1}(2y-1) \,\mathrm{d}y \Big/ \int_0^1 2(\mathrm{erf}^{-1})^2(2y-1) \,\mathrm{d}y \\
&= \sqrt{2} \int_0^1 F_q^{-1}(y) \mathrm{erf}^{-1}(2y-1) \,\mathrm{d}y \Big/ \underbrace{\int_{-\infty}^{\infty} 2x^2 \mathcal{N}(x|0, 1/2) \,\mathrm{d}x}_{=1} \\
&= \sqrt{2} \int_0^1 F_q^{-1}(y) \mathrm{erf}^{-1}(2y-1) \,\mathrm{d}y.
\end{aligned}$$

B Minimization of W_p^p

When using other W_p^p to get different performances, we can apply the gradient descent to obtain optimal μ^* and σ^* (elegant updating equations are unavailable). For simplification, we define $\varepsilon(y) = F_{\widetilde{q}}^{-1}(y) - \mu - \sqrt{2}\sigma \mathrm{erf}^{-1}(2y-1)$ and $\eta(x) = x - \mu - \sqrt{2}\sigma \mathrm{erf}^{-1}(2F_{\widetilde{q}}(x) - 1)$, and derivatives can be directly computed as:

$$\begin{aligned}
\partial_\mu \mathrm{W}_p^p &= -p \int_0^1 |\varepsilon(y)|^{p-1} \mathrm{sgn}(\varepsilon(y)) \,\mathrm{d}y = -p \int_{-\infty}^{\infty} |\eta(x)|^{p-1} \mathrm{sgn}(\eta(x)) \widetilde{q}(x) \,\mathrm{d}x, \\
\partial_\sigma \mathrm{W}_p^p &= -p \int_0^1 |\varepsilon(y)|^{p-1} \mathrm{sgn}(\varepsilon(y)) \mathrm{erf}^{-1}(2y-1) \,\mathrm{d}y = -p \int_{-\infty}^{\infty} |\eta(x)|^{p-1} \mathrm{sgn}(\eta(x)) \mathrm{erf}^{-1}(2F_{\widetilde{q}}(x) - 1) \widetilde{q}(x) \,\mathrm{d}x.
\end{aligned}$$

Furthermore, we resort to Monte Carlo sampling to calculate the derivatives. Consider $\widetilde{q}(x) = q^{\setminus i}(x) p(y_i|x)/Z$ defined as Eqn. (4.3) and we approximate $\partial_\mu \mathrm{W}_p^p$ as:

$$\partial_\mu \mathrm{W}_p^p = -p \int_{-\infty}^{\infty} \big|\eta(x)\big|^{p-1} \mathrm{sgn}\big(\eta(x)\big) p(y_i|x) q^{\setminus i}(x)/Z \,\mathrm{d}x,$$

$$
\begin{aligned}
&\approx -\frac{p}{J}\sum_{j=1}^{J}\big|\eta(x_j)\big|^{p-1}\mathrm{sgn}\big(\eta(x_j)\big))p(y_i|x_j)/Z,\\
\partial_\sigma \mathrm{W}_p^p &= -p\int_{-\infty}^{\infty}\big|\eta(x)\big|^{p-1}\mathrm{sgn}\big(\eta(x)\big)\mathrm{erf}^{-1}(2F_{\widetilde{q}}(x)-1)p(y_i|x)q^{\setminus i}(x)/Z\,\mathrm{d}x,\\
&\approx -\frac{p}{J}\sum_{j=1}^{J}\big|\eta(x_j)\big|^{p-1}\mathrm{sgn}\big(\eta(x_j)\big)\mathrm{erf}^{-1}(2F_{\widetilde{q}}(x_j)-1)p(y_i|x_j)/Z,
\end{aligned}
$$

where x_j are i.i.d. sampled from the Gaussian distribution $q^{\setminus i}(x)$.

C Details of Economic Parameterization

C.1 Decomposition of the KL divergence in Proof for EP

$$
\begin{aligned}
\mathrm{KL}(\widetilde{q}(\boldsymbol{f})\|\mathcal{N}(\boldsymbol{f})) &= \int \widetilde{q}(\boldsymbol{f})\log\frac{\widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i)\widetilde{q}(f_i)}{\mathcal{N}(\boldsymbol{f}_{\setminus i}|f_i)\mathcal{N}(f_i)}\,\mathrm{d}\boldsymbol{f}\\
&= \int \widetilde{q}(f_i)\log\frac{\widetilde{q}(f_i)}{\mathcal{N}(f_i)}\,\mathrm{d}f_i + \int \widetilde{q}(f_i)\int \widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i)\log\frac{\widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i)}{\mathcal{N}(\boldsymbol{f}_{\setminus i}|f_i)}\,\mathrm{d}\boldsymbol{f}_{\setminus i}\,\mathrm{d}f_i\\
&= \mathrm{KL}\big(\widetilde{q}(f_i)\|\mathcal{N}(f_i)\big) + \mathbb{E}_{\widetilde{q}(f_i)}\Big[\mathrm{KL}\big(\widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i)\|\mathcal{N}(\boldsymbol{f}_{\setminus i}|f_i)\big)\Big]
\end{aligned}
$$

C.2 Details of Eqn. (5.2)

$$
\begin{aligned}
\mathrm{W}_2^2\left(\widetilde{q}(\boldsymbol{f}),\mathcal{N}(\boldsymbol{f})\right) &\equiv \inf_{\pi\in U(\widetilde{q},\mathcal{N})}\mathbb{E}_\pi\left(\|\boldsymbol{f}-\boldsymbol{f}^{(n)}\|_2^2\right)\\
&= \inf_{\pi\in U(\widetilde{q},\mathcal{N})}\mathbb{E}_\pi\left(\|f_i-f_i^{(n)}\|_2^2\right)+\mathbb{E}_\pi\left(\|\boldsymbol{f}_{\setminus i}-\boldsymbol{f}_{\setminus i}^{(n)}\|_2^2\right)\\
&\overset{\text{(a)}}{=} \inf_{\pi\in U(\widetilde{q},\mathcal{N})}\mathbb{E}_{\pi_i}\Big[\|f_i-f_i^{(n)}\|_2^2+\mathbb{E}_{\pi_{\setminus i|i}}\left(\|\boldsymbol{f}_{\setminus i}-\boldsymbol{f}_{\setminus i}^{(n)}\|_2^2\right)\Big]\\
&\overset{\text{(b)}}{=} \inf_{\pi_i}\mathbb{E}_{\pi_i}\Big[\|f_i-f_i^{(n)}\|_2^2+\inf_{\pi_{\setminus i|i}}\mathbb{E}_{\pi_{\setminus i|i}}\left(\|\boldsymbol{f}_{\setminus i}-\boldsymbol{f}_{\setminus i}^{(n)}\|_2^2\right)\Big]\\
&= \inf_{\pi_i}\mathbb{E}_{\pi_i}\Big[\|f_i-f_i^{(n)}\|_2^2+\mathrm{W}_2^2(\widetilde{q}_{\setminus i|i},\mathcal{N}_{\setminus i|i})\Big]\\
&\overset{\text{(c)}}{=} \inf_{\pi_i}\mathbb{E}_{\pi_i}\Big[\|f_i-f_i^{(n)}\|_2^2+\mathrm{W}_2^2(q_{\setminus i|i}^{\setminus i},\mathcal{N}_{\setminus i|i})\Big],
\end{aligned}
$$

where the superscript (n) indicates that the variable is from the Gaussian $\mathcal{N}$. In (a), $\pi_i = \pi(f_i, f_i^{(n)})$ and $\pi_{\setminus i|i} = \pi(\boldsymbol{f}_{\setminus i}, \boldsymbol{f}_{\setminus i}^{(n)}|f_i, f_i^{(n)})$. In (b), the first and the second inf are over $U(\widetilde{q}_i,\mathcal{N}_i)$ and $U(\widetilde{q}_{\setminus i|i},\mathcal{N}_{\setminus i|i})$ respectively. (c) is due to $\widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i)$ being equal to $q^{\setminus i}(\boldsymbol{f}_{\setminus i}|f_i)$:

$$
\begin{aligned}
\widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i) = \frac{\widetilde{q}(\boldsymbol{f})}{\widetilde{q}(f_i)} &\propto \frac{p(\boldsymbol{f})\cancel{p(y_i|f_i)}\prod_{j\neq i}t_j(\boldsymbol{f})}{q^{\setminus i}(f_i)\cancel{p(y_i|f_i)}}\\
&= q^{\setminus i}(\boldsymbol{f}_{\setminus i}|f_i).
\end{aligned}
$$

C.3 Details of Eqn. (5.2)

$$
\begin{aligned}
&\min_{\boldsymbol{m}_{\setminus i}^{(n)}}\ \text{Eqn. (5.2)}\\
&= \inf_{\pi_i}\mathbb{E}_{\pi_i}\Big[\|f_i-f_i^{(n)}\|_2^2+\|\boldsymbol{a}(f_i-\mu_{\widetilde{q}_i})-\boldsymbol{a}^{(n)}(f_i^{(n)}-m_i^{(n)})\|_2^2\Big]
\end{aligned}
$$

$$
\begin{aligned}
&= \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i - f_i^{(n)}\|_2^2\Big] + \|\boldsymbol{a}\|_2^2\sigma_{\widetilde{q}_i}^2 + \|\boldsymbol{a}^{(n)}\|_2^2 S_i^{(n)} - 2\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}^{(n)}\mathbb{E}_{\pi_i}\Big(f_i f_i^{(n)} - \mu_{\widetilde{q}_i} m_i^{(n)}\Big) \\
&= \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i - f_i^{(n)}\|_2^2\Big] + \|\boldsymbol{a}\|_2^2\sigma_{\widetilde{q}_i}^2 + \|\boldsymbol{a}^{(n)}\|_2^2 S_i^{(n)} + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}^{(n)}\mathbb{E}_{\pi_i}\Big(\|f_i - f_i^{(n)}\|_2^2 - f_i^2 - (f_i^{(n)})^2 + 2\mu_{\widetilde{q}_i} m_i^{(n)}\Big) \\
&= \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i - f_i^{(n)}\|_2^2\Big] + \|\boldsymbol{a}\|_2^2\sigma_{\widetilde{q}_i}^2 + \|\boldsymbol{a}^{(n)}\|_2^2 S_i^{(n)} + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}^{(n)}\mathbb{E}_{\pi_i}\Big(\|f_i - f_i^{(n)}\|_2^2 - (f_i - \mu_{\widetilde{q}_i})^2 - \\
&\quad 2f_i\mu_{\widetilde{q}_i} + \mu_{\widetilde{q}_i}^2 - (f_i^{(n)} - m_i^{(n)})^2 - 2f_i^{(n)} m_i^{(n)} + (m_i^{(n)})^2 + 2\mu_{\widetilde{q}_i} m_i^{(n)}\Big) \\
&= (1 + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}^{(n)})\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2\sigma_{\widetilde{q}_i}^2 + \|\boldsymbol{a}^{(n)}\|_2^2 S_i^{(n)} - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}^{(n)}\Big(\sigma_{\widetilde{q}_i}^2 + \mu_{\widetilde{q}_i}^2 + S_i^{(n)} + (m_i^{(n)})^2 - 2\mu_{\widetilde{q}_i} m_i^{(n)}\Big) \\
&= (1 + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}^{(n)})\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2\sigma_{\widetilde{q}_i}^2 + \|\boldsymbol{a}^{(n)}\|_2^2 S_i^{(n)} - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}^{(n)}\Big[\sigma_{\widetilde{q}_i}^2 + S_i^{(n)} + (\mu_{\widetilde{q}_i} - m_i^{(n)})^2\Big]
\end{aligned}
$$

C.4 Details of Eqn. (5.2)

$$
\begin{aligned}
\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) &= \int_0^1 \big(F_{\widetilde{q}_i}^{-1}(y) - m_i' - \sqrt{2S_i'}\mathrm{erf}^{-1}(2y-1)\big)^2 \,\mathrm{d}y, \\
&= \int_0^1 (F_{\widetilde{q}_i}^{-1}(y) - m_i')^2 + 2S_i'\mathrm{erf}^{-1}(2y-1)^2 - 2\sqrt{2S_i'}\mathrm{erf}^{-1}(2y-1)(F_{\widetilde{q}_i}^{-1}(y) - m_i') \,\mathrm{d}y, \\
&= \int_0^1 (F_{\widetilde{q}_i}^{-1}(y) - \mu_{\widetilde{q}_i} + \mu_{\widetilde{q}_i} - m_i')^2 \,\mathrm{d}y + S_i' - 2\sqrt{2S_i'}c_{\widetilde{q}_i}, \\
&= \sigma_{\widetilde{q}_i}^2 + (\mu_{\widetilde{q}_i} - m_i')^2 + S_i' - 2c_{\widetilde{q}_i}\sqrt{2S_i^{(n)}},
\end{aligned}
$$

where $F_{\widetilde{q}_i}^{-1}(y)$ is the quantile function of $\widetilde{q}(f_i)$ and $c_{\widetilde{q}_i} \equiv \int_0^1 F_{\widetilde{q}_i}^{-1}(y)\mathrm{erf}^{-1}(2y-1)\,\mathrm{d}y$.