Quantile Propagation for Wasserstein-Approximate Gaussian Processes

Rui Zhang [1 2] Christian J. Walder [1 2] Edwin V. Bonilla [2] Marian-Andrei Rizoiu [3 2] Lexing Xie [1 2]

Abstract

We develop a new approximate Bayesian inference method for Gaussian process models with factorized non-Gaussian likelihoods. Our method—dubbed Quantile Propagation (QP)—is similar to expectation propagation (EP) but minimizes the L_2 Wasserstein distance rather than the Kullback-Leibler (KL) divergence. We consider the case where likelihood factors are approximated by a Gaussian form. We show that QP matches quantile functions rather than moments as in EP and has the same mean update but a smaller variance update than EP, thereby alleviating the over-estimation of the posterior variance exhibited by EP. Crucially, QP has the same favorable locality property as EP, and thereby admits an efficient algorithm. Experiments on classification and Poisson regression tasks demonstrate that QP outperforms both EP and variational Bayes.

1. Introduction

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 (Matheron, 1963), classification (Williams & Barber, 1998; Hensman et al., 2015) and stochastic point process modeling (Møller et al., 1998; Zhang et al., 2019). However, exact Bayesian inference for GP models is intractable for all but the Gaussian likelihood function. To address this issue, various approximate Bayesian inference methods have been proposed, such as Markov Chain Monte Carlo (MCMC, see *e.g.* Neal, 1997), the Laplace approximation (Williams & Barber, 1998), variational inference (Jordan et al., 1999; Opper & Archambeau, 2009) and expectation propagation (Opper & Winther, 2000; Minka, 2001b).

The existing approach most relevant to this work is expectation propagation (EP), which approximates each non-Gaussian likelihood term with a Gaussian by iteratively minimizing a set of local forward Kullback-Leibler (KL) divergences. As shown by Gelman et al. (2017), EP can scale to very large datasets. However, EP is not guaranteed to converge, and is known to over-estimate posterior variances (Minka, 2005; Jylänki et al., 2011; Heess et al., 2013).

This latter deficiency of over-estimation of the variances, is an inherent drawback of using the forward KL divergence for approximate Bayesian inference. More generally, several authors have pointed out that using the KL divergence can yield undesirable results such as (but not limited to) over-dispersed or under-dispersed posteriors (Dieng et al., 2017; Li & Turner, 2016; Hensman et al., 2014).

As an alternative to the KL divergence, optimal transport divergences—such as the Wasserstein distance (WD, Villani, 2008, §6)—have recently gained substantial popularity. The WD is a natural measure between two distributions, and has been successfully employed in 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). Recent work has begun to employ the WD for inference, as in 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 intuitive formulation and excellent performance, in contrast to the KL divergence the WD is computationally challenging (Cuturi, 2013), especially in high dimensions (Bonneel et al., 2015).

Contributions. In this work, we overcome some of the shortcomings of the KL divergence for approximate inference with GP models, by developing an efficient approximate Bayesian scheme that minimizes a specific class of WD distances, which we refer to as the L_2 WD. Below we detail the three main contributions of this paper.

First, in section 4, we develop quantile propagation (QP), an approximate inference algorithm for models with GP priors and factorized likelihoods. Like EP, QP does not directly minimize global distances between high-dimensional distributions. Instead, QP estimates a fully coupled Gaussian posterior by iteratively minimizing *local* divergences between two particular marginal distributions; as these marginals are univariate, it turns out that QP yields quantile function

[1] The Australian National University, Canberra, Australia [2] CSIRO's Data61, Australia [3] University of Technology Sydney, Sydney, Australia. Correspondence to: Rui Zhang <rui.zhang@anu.edu.au>.

matching (rather than moment matching as in EP), hence we term our inference scheme *quantile propagation*. We derive the updates for the approximate likelihood terms and show that while the QP mean estimates match those of EP, the variance estimates are lower for QP.

Second, in section 5 we show that like EP, QP satisfies the locality property, meaning that it is sufficient to employ *univariate* approximate likelihood terms, and that the updates can thereby be performed efficiently only using marginal distributions. Consequently, although our method employs a more complex divergence than that of EP (L_2 WD vs KL), it has the same computational complexity.

Finally, in section 6 we employ eight real-world datasets and compare our method to EP and variational Bayes (VB) on the tasks of binary classification and Poisson regression. We find that in terms of predictive accuracy, QP performs similarly to EP but is superior to VB. In terms of predictive uncertainty, however, we find QP superior to both EP and VB, thereby supporting our claim that QP alleviates variance over-estimation associated with the KL divergence (Minka, 2005; Jylänki et al., 2011; Heess et al., 2013).

2. Related Work

The basis of the EP algorithm for GP models was first proposed by Opper & Winther (2000) and then generalized by Minka (2001a;b). Power EP (Minka, 2004; 2005) is an extension of EP that exploits the more general α-divergence (with $\alpha = 1$ corresponding to the forward KL divergence in EP) and has been recently used in conjunction with GP pseudo-point approximations (Bui et al., 2017). Although generally not guaranteed to converge locally or globally, Power EP uses fixed-point iterations for its local updates and has been shown to perform well in practice for GP regression and classification (Bui et al., 2017). In comparison, our approach uses the L_2 WD, and like EP, it yields convex local optimizations for GP models with factorized likelihoods. This convexity benefits the convergence of the local update, and is retained even with the general L_p ($p \geq 1$) WD as shown in Theorem 1. Moreover, for the same class of GP models, both EP and our approach have the locality property (Seeger, 2005) and can be unified in the generic message passing framework (Minka, 2005).

Without the guarantee of convergence for the explicit global objective function, interpreting EP has proven to be a challenging task. As a result, a number of works have instead attempted to directly minimize divergences between the true and approximate joint posteriors, for divergences such as the KL (Jordan et al., 1999; Dezfouli & Bonilla, 2015), Rényi (Li & Turner, 2016), α (Hernández-Lobato et al., 2016) and optimal transport divergences (Ambrogioni et al., 2018). To deal with the infinity issue of the KL (and more generally the Rényi and α divergences) which arises from differing distribution 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 of EP have also been proposed, including the convergent double loop algorithm (Opper & Winther, 2005), parallel EP (Minka, 2001c), distributed EP built on partitioned datasets (Xu et al., 2014; Gelman et al., 2017), averaged EP assuming that all approximate likelihoods contribute similarly (Dehaene & Barthelmé, 2018), and stochastic EP which may be regarded as sequential averaged EP (Li et al., 2015).

The L_2 WD between two Gaussian distributions has a closed form expression (Dowson & Landau, 1982). Detailed research on the Wasserstein geometry of the Gaussian distribution is conducted by Takatsu (2011). Recently, the closed form expression has been applied to robust Kalman filtering (Shafieezadeh-Abadeh et al., 2018) and to the analysis of populations of GPs (Mallasto & Feragen, 2017). A more general extension to elliptically contoured distributions is provided by Gelbrich (1990) and used to compute probabilistic word embeddings (Muzellec & Cuturi, 2018). A geodesic interpretation for the L_2 WD between any distributions (Benamou & Brenier, 2000), and this has already been exploited to develop approximate Bayesian inference schemes (El Moselhy & Marzouk, 2012). Our approach builds on the L_2 WD without exploiting any of these closed form expressions; it enjoys computational efficiency by leveraging the EP framework and using the quantile function form of the L_2 WD for univariate distributions.

3. Prerequisites

In this section, we review GP models with factorized non-Gaussian likelihoods, the EP algorithm and the WD.

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 noisy output. Our goal is to establish the mapping from inputs to outputs via a latent function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ which is assigned a GP prior. Specifically, assuming a zero-mean GP prior with covariance function $k(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the GP hyper-parameters, we have that $p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, K)$, 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 at every pair of inputs. In this work we use the squared exponential covariance function $k(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta}) = \gamma \exp\big[-\sum_{i=1}^d (x_i - x_i')^2/(2\alpha_i^2)\big]$, where $\boldsymbol{\theta} = \{\gamma, \alpha_1, \cdots, \alpha_d\}$. For notational simplification, we will omit the conditioning on $\boldsymbol{\theta}$ in the rest of this paper.

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

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

where the normalizer $p(\mathcal{D}) = \int p(\boldsymbol{f})\prod_{i=1}^N p(y_i|f_i)\,\mathrm{d}\boldsymbol{f}$ is often analytically intractable. Numerous problems may be modeled in this way: binary classification (Williams & Barber, 1998), single-output regression with Gaussian likelihood (Matheron, 1963), Student's-t likelihood (Jylänki et al., 2011) or Poisson likelihood (Zou, 2004), and warped Gaussian processes (Snelson et al., 2004).

3.2. Expectation Propagation

In this section we review the application of EP to the GP models described above. EP deals with the analytical intractability by using Gaussian approximations to the individual non-Gaussian likelihoods:

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

The function t_i is often called the *site function* and is specified by the *site parameters*: the scale $\widetilde{Z}_i$, the mean $\widetilde{\mu}_i$ and the variance $\widetilde{\sigma}_i^2$. Notably, it is sufficient to use univariate site functions given that the local update can be efficiently computed using the marginal distribution only (Seeger, 2005). We refer to this as the *locality property*. Although in this work we employ a more complex L_2 WD, our approach retains this property, as we elaborate in subsection 5.2.

Given the site functions, one can approximate the intractable posterior distribution $p(\boldsymbol{f}|\mathcal{D})$ using a Gaussian $q(\boldsymbol{f})$:

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

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

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

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

This 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 to optimize site functions $\{t_i(f_i)\}_{i=1}^N$. Ideally, one would seek to minimize the global KL divergence between the true and approximate posterior distributions $\mathrm{KL}(p(\boldsymbol{f}|\mathcal{D})\|q(\boldsymbol{f}))$, however this is intractable. Instead, EP is built based on the assumption that the global divergence can be approximated by the local divergence $\mathrm{KL}(\widetilde{q}(\boldsymbol{f})\|q(\boldsymbol{f}))$, where $\widetilde{q}(\boldsymbol{f}) \propto q^{\setminus i}(\boldsymbol{f})p(y_i|f_i)$ and $q^{\setminus i}(\boldsymbol{f}) \propto q(\boldsymbol{f})/t_i(f_i)$ are refered to as the tilted and cavity distributions, respectively. Note that the cavity distribution is Gaussian while the tiltled distribution is usually not. The local divergence can be simplified from multi-dimensional to univariate, $\mathrm{KL}(\widetilde{q}(\boldsymbol{f})\|q(\boldsymbol{f})) = \mathrm{KL}(\widetilde{q}(f_i)\|q(f_i))$ (detailed in Appendix E), and $t_i(f_i)$ is optimized by minimizing it.

The minimization is realized by projecting the tilted distribution $\widetilde{q}(f_i)$ onto the Gaussian family, with the projected Gaussian denoted $\mathrm{proj}_{\mathrm{KL}}(\widetilde{q}(f_i)) \equiv \mathrm{argmin}_{\mathcal{N}}\,\mathrm{KL}(\widetilde{q}(f_i)\|\mathcal{N}(f_i))$. Then the projected Gaussian is used to update $t_i(f_i) \propto \mathrm{proj}_{\mathrm{KL}}(\widetilde{q}(f_i))/q^{\setminus i}(f_i)$. The mean and the variance of $\mathrm{proj}_{\mathrm{KL}}(\widetilde{q}(f_i)) \equiv \mathcal{N}(\mu^\star, \sigma^{\star 2})$ match the moments of $\widetilde{q}(f_i)$ and are used to update $t_i(f_i)$'s parameters:

$$\mu^\star = \mu_{\widetilde{q}_i}, \quad \sigma^{\star 2} = \sigma^2_{\widetilde{q}_i}, \quad (3)$$

$$\widetilde{\mu}_i = \widetilde{\sigma}_i^2\left(\mu^\star(\sigma^\star)^{-2} - \mu_{\setminus i}\sigma_{\setminus i}^{-2}\right),\ \ \widetilde{\sigma}_i^{-2} = (\sigma^\star)^{-2} - \sigma_{\setminus i}^{-2}, \quad (4)$$

where $\mu_{\widetilde{q}_i}$ and $\sigma^2_{\widetilde{q}_i}$ are the mean and the variance of $\widetilde{q}(f_i)$, and $\mu_{\setminus i}$ and $\sigma^2_{\setminus i}$ are the mean and the variance of $q^{\setminus i}(f_i)$. We refer to the projection as the local update. Note that $\widetilde{Z}$ does not impact the optimization of $q(\boldsymbol{f})$ or the GP hyper-parameters $\boldsymbol{\theta}$, so we omit the update formula for $\widetilde{Z}$. We summarize EP in algorithm 1. In section 4 we propose a new approximation approach which is similar to EP but employs the L_2 WD rather than the KL divergence.

3.3. Wasserstein Distance

We denote by $\mathcal{M}_+^1(\Omega)$ the set of all probability measures on Ω. We consider probability measures on the d-dimensional real space $\mathbb{R}^d$. The WD between two probability distributions $\mu, \nu \in \mathcal{M}_+^1(\mathbb{R}^d)$ can be intuitively defined as the cost of transporting the probability mass from one distribution to the other. We are particularly interested in the subclass of L_p WD, formally defined as follows.

Definition 1 (L_p WD)**.** *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 WD 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.

Like the KL divergence, the L_p WD it has a minimum of zero, achieved when the distributions are equivalent, but unlike the KL, it has the appealing property of symmetry. Another property we exploit for computational efficiency is:

Proposition 1. *(Peyré et al., 2019, Remark 2.30) The L_p WD between 1-d distribution functions μ and $\nu \in \mathcal{M}_+^1(\mathbb{R})$*

equals the L_p distance between the quantile functions:

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

where $F_z:\mathbb{R}\to[0,1]$ is the cumulative distribution function (CDF) of z, defined as $F_z(x)=\int_{-\infty}^x dz$, and F_z^{-1} is the pseudoinverse or quantile function, defined as $F_z^{-1}(y)=\min_x\{x\in\mathbb{R}\cup\{-\infty\}:F_z(x)\geq y\}$.

We expoloit the following translation property of the L_2 WD in our proof of the locality property.

Proposition 2. *(Peyré et al., 2019, Remark 2.19) Consider the L_2 WD 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 $W_2^2(\mu_{\boldsymbol{\tau}},\nu_{\boldsymbol{\tau}'})=W_2^2(\mu,\nu)-2(\boldsymbol{\tau}-\boldsymbol{\tau}')^\mathsf{T}(\boldsymbol{m}_\mu-\boldsymbol{m}_\nu)+\|\boldsymbol{\tau}-\boldsymbol{\tau}'\|_2^2$, 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 $W_2^2(\mu_{\boldsymbol{\tau}},\nu_{\boldsymbol{\tau}'})=W_2^2(\mu,\nu)-\|\boldsymbol{m}_\mu-\boldsymbol{m}_\nu\|_2^2$.*

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
    compute $q(\boldsymbol{f})\propto p(\boldsymbol{f})\prod_i t_i(f_i)$                 by (1)
    repeat
        for $i=1$ to $N$ do
            compute $q^{\setminus i}(f_i)\propto q(f_i)/t_i(f_i)$               cavity
            compute $\widetilde{q}(f_i)\propto q^{\setminus i}(f_i)p(y_i|f_i)$            tilted
            if EP then
                $t_i(f_i)\propto \mathrm{proj}_{\mathrm{KL}}[\widetilde{q}(f_i)]/q^{\setminus i}(f_i)$      by (3)(4)
            else if QP then
                $t_i(f_i)\propto \mathrm{proj}_{\mathrm{W}}[\widetilde{q}(f_i)]/q^{\setminus i}(f_i)$       by (5)(4)
            end if
            update $q(\boldsymbol{f})\propto p(\boldsymbol{f})\prod_i t_i(f_i)$             by (1)
        end for
    until convergence
    $\boldsymbol{\theta}=\mathrm{argmax}_{\boldsymbol{\theta}}\log q(\mathcal{D})$                    by (2)
until convergence
return $q(\boldsymbol{f})$
```

4. Quantile Propagation

We now propose our new approximation algorithm which, as summarized in Algorithm 1, employs an L_2 WD based projection rather than the forward KL divergence projection of EP. Although QP employs a more complex divergence, it has the same computational complexity as EP.

As stated in 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 local updates for the site functions and analyzing their relationships with those of EP. Later in section 5, we show the locality property of QP, meaning that the site function $t(f)$ has a univariate parameterization and so the local update can be efficiently performed using marginals only.

4.1. Convexity of L_p Wasserstein Distance

We first show $\mathrm{W}_p^p(\widetilde{q}(f),\mathcal{N}(f|\mu,\sigma^2))$ to be strictly convex about μ and σ. Given convexity, we can employ either elegant closed form expressions (if available) or simply numerical gradient-based methods) to optimize μ and σ. Formally, we have the following theorem:

Theorem 1. *Given two probability measures in $\mathcal{M}_+^1(\mathbb{R})$: a Gaussian $\mathcal{N}(\mu,\sigma^2)$ with mean μ and standard deviation $\sigma>0$, and an arbitrary measure $\widetilde{q}$, the L_p WD $W_p^p(\widetilde{q},\mathcal{N})$ is strictly convex about μ and σ.*

Proof. Let $F_{\widetilde{q}}^{-1}(y)$ and $F_{\mathcal{N}}^{-1}(y)=\mu+\sqrt{2}\sigma\mathrm{erf}^{-1}(2y-1)$, $y\in[0,1]$, be the quantile functions of $\widetilde{q}$ and the Gaussian $\mathcal{N}$, where erf is the error function. Then, we consider two distinct Gaussian measures $\mathcal{N}(\mu_1,\sigma_1^2)$ and $\mathcal{N}(\mu_2,\sigma_2^2)$ 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 $a_1+a_2=1$. Given the above, we further define $\varepsilon_k(y)=F_{\widetilde{q}}^{-1}(y)-\mu_k-\sigma_k\sqrt{2}\mathrm{erf}^{-1}(2y-1)$, $k=1,2$, for notational simplification, and derive the convexity as below:

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

where steps (a), (b) and (c) are obtained by applying Proposition 1, non-negativity of the absolute value, and the convexity of $f(x)=x^p$, $p\geq1$, over $\mathbb{R}_+$ respectively. The equality at (b) holds iff $\varepsilon_k(y)\geq0, k=1,2,\forall y\in[0,1]$, and (c)'s equality holds iff $|\varepsilon_1(y)|=|\varepsilon_2(y)|$, $\forall y\in[0,1]$. These two conditions for equality cannot be attained simultaneously as otherwise it would contradict that $\mathcal{N}(\mu_1,\sigma_1^2)$ is different from $\mathcal{N}(\mu_2,\sigma_2^2)$. Therefore, $\mathrm{W}_p^p(\widetilde{q},\mathcal{N})$, $p\geq1$, is strictly convex about μ and σ. □

4.2. Minimization of L_2 WD

An advantage of using the L_p WD with $p=2$, rather than other choices of p, is the computational efficiency it admits in the local updates. As we show in this section, optimizing the L_2 WD yields neat analytical updates of the optimal $\mu^\star$ and $\sigma^\star$ that require only univariate integration and the CDF of $\widetilde{q}(f)$. In contrast, other L_p WDs lack analytical expressions and require resorting to gradient based methods. Nonetheless, other L_p WDs may have useful properties, and we leave extension to those cases to future work.

The optimal parameters $\mu^\star$ and $\sigma^\star$ corresponding to the L_2 WD $\mathrm{W}_2^2(\widetilde{q}, \mathcal{N}(\mu, \sigma^2))$ can be obtained using Proposition 1. Specifically, we employ the quantile function reformulation of $\mathrm{W}_2^2(\widetilde{q}, \mathcal{N}(\mu, \sigma^2))$, and zero its derivatives w.r.t. μ and σ. The results provided below are derived in Appendix A:

$$\begin{aligned}
\mu^\star &= \mu_{\widetilde{q}}, \\
\sigma^\star &= \sqrt{2}\int_0^1 F_{\widetilde{q}}^{-1}(y)\mathrm{erf}^{-1}(2y-1)\,\mathrm{d}y, && (5) \\
&= \sqrt{2}\int_{-\infty}^{\infty} f\mathrm{erf}^{-1}(2F_{\widetilde{q}}(f)-1)\widetilde{q}(f)\,\mathrm{d}f, && (6)
\end{aligned}$$

Interestingly, the update for μ matches that of EP, namely the expectation under $\widetilde{q}$. However, for the standard deviation we have the difficulty of deriving the CDF $F_{\widetilde{q}}$. If a closed form expression is available, we can apply numerical integration to compute the optimal standard deviation; otherwise, we may use sampling based methods to approximate it. In our experiments we employ the former.

4.3. Properties of the Variance Update

Given the update equations in the previous section, here we show that the standard deviation estimate of QP, denoted as σ_{QP}, is less or equal to that of EP, denoted as σ_{EP}, when projecting the same tilted distribution to the Gaussian space.

Theorem 2. *The variance of the Gaussian approximation to a univariate tilted distribution $\widetilde{q}(f)$ as estimated by QP and EP satisfy $\sigma^2_{QP} \le \sigma^2_{EP}$.*

Proof. Let $\mathcal{N}(\mu_{\mathrm{QP}}, \sigma^2_{\mathrm{QP}})$ be the optimal Gaussian in QP. As per Proposition 1, we reformulate the L_2 WD based projection $\mathrm{W}_2^2(\widetilde{q}, \mathcal{N}(\mu_{\mathrm{QP}}, \sigma^2_{\mathrm{QP}}))$ w.r.t. quantile functions:

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

where for (A), we used $\int \mu_{\mathrm{QP}}\sigma_{\mathrm{QP}}\mathrm{erf}^{-1}(2y-1)\,\mathrm{d}y = 0$ and the remaining factor can be easily shown to be equal to $2\sigma^2_{\mathrm{QP}}$. Furthermore, due to the non-negativity of the WD, we have $\sigma^2_{\mathrm{EP}} \ge \sigma^2_{\mathrm{QP}}$, and the equality holds iff $\widetilde{q}$ is Gaussian. □

Corollary 2.1. *The variances of the site functions updated by EP and QP satisfy: $\widetilde{\sigma}^2_{QP} \le \widetilde{\sigma}^2_{EP}$, and the variances of the approximate posterior marginals satisfy $\sigma^2_{q,QP} \le \sigma^2_{q,EP}$.*

Proof. Since the cavity distribution is unchanged, we can calculate the variance of the site function as per Equation (4) and conclude that the variance of the site function also satisfies $\widetilde{\sigma}^2_{\mathrm{QP}} \le \widetilde{\sigma}^2_{\mathrm{EP}}$. Moreover as per the definition of the cavity distribution in subsection 3.2, the approximate marginal distribution is proportional to the product of the cavity distribution and the site function $q(f_i) \propto q^{\backslash i}(f_i)t(f_i)$, which are two Gaussian distributions. By the product of Gaussians formula Equation (4), we know the variance of $q(f_i)$ estimated by EP as $\sigma^2_{q,\mathrm{EP}} = (\widetilde{\sigma}^{-2}_{\mathrm{EP}} + \sigma^{-2}_{\backslash i})^{-1} = \sigma^2_{\mathrm{EP}}$ and similarly $\sigma^2_{q,\mathrm{QP}} = \sigma^2_{\mathrm{QP}}$, where σ^2_{EP} and σ^2_{QP} are defined in Theorem 2. Thus, there is $\sigma^2_{q,\mathrm{QP}} \le \sigma^2_{q,\mathrm{EP}}$. □

Corollary 2.2. *The predictive variances of latent functions at $\boldsymbol{x}_*$ by EP and QP satisfy: $\sigma^2_{QP}(f(\boldsymbol{x}_*)) \le \sigma^2_{EP}(f(\boldsymbol{x}_*))$.*

Proof. The predictive variance of the latent function was analyzed in (Rasmussen & Williams, 2005, Equation (3.61)):

$$\sigma^2(f_*) = k_* - \boldsymbol{k}_*^\top(K + \widetilde{\Sigma})^{-1}\boldsymbol{k}_*,$$

where we define $f_* = f(\boldsymbol{x}_*)$ and $k_* = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$, and let $\boldsymbol{k}_* = (k(\boldsymbol{x}_*, \boldsymbol{x}_i))_{i=1}^N$ be the (column) covariance vector between the test data $\boldsymbol{x}_*$ and the training data $\{\boldsymbol{x}_i\}_{i=1}^N$. After updating parameters of the site function $t_i(f_i)$, the predictive variance can be written as (details in Appendix G):

$$\sigma^2_{\mathrm{new}}(f_*) = k_* - \boldsymbol{k}_*^\top A\boldsymbol{k}_* + \frac{\boldsymbol{k}_*^\top \boldsymbol{s}_i\boldsymbol{s}_i^\top \boldsymbol{k}_*}{(\widetilde{\sigma}^2_{i,\mathrm{new}} - \widetilde{\sigma}^2_i)^{-1} + A_{ii}},$$

where $\widetilde{\sigma}^2_{i,\mathrm{new}}$ is the site variance updated by EP or QP, $A = (K + \widetilde{\Sigma})^{-1}$ and $\boldsymbol{s}_i$ is the i's column of A. Since $\widetilde{\sigma}^2_{i,\mathrm{QP}} \le \widetilde{\sigma}^2_{i,\mathrm{EP}}$, we have $\sigma^2_{\mathrm{QP}}(f_*) \le \sigma^2_{\mathrm{EP}}(f_*)$. □

Remark. *We compared variance estimates of EP and QP assuming the same cavity distribution. Proving analagous statements for the fixed points of the EP and QP algorithms is more challenging, however, and we leave this to future work, while providing empirical support for these analogous statements in Figure 1a. and Figure 1b.*

5. Locality Property

In this section we detail the central result on which our QP algorithm is based upon, which we refer to as the *locality property*. That is, the optimal site function t_i is defined only in terms of the single corresponding latent variable f_i, and thereby and similarly to EP, it admits a simple and efficient sequential update of each individual site approximation.

5.1. Review: Locality Property of EP

We provide a brief review of the locality property of EP for GP models; for more details see Seeger (2005). We begin by defining the general site function $t_i(\boldsymbol{f})$ in terms of all of the latent variables, and the cavity and the tilted distributions as $q^{\backslash i}(\boldsymbol{f}) \propto p(\boldsymbol{f}) \prod_{j \neq i} \widetilde{t}_j(\boldsymbol{f})$ and $\widetilde{q}(\boldsymbol{f}) \propto q^{\backslash i}(\boldsymbol{f}) p(y_i|f_i)$, respectively. 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 further rewritten as (see details in Appendix D.1):

$$\mathrm{KL}\big(\widetilde{q}\|\mathcal{N}\big) = \mathrm{KL}\big(\widetilde{q}_i\|\mathcal{N}_i\big) + \mathbb{E}_{\widetilde{q}_i}\Big[\mathrm{KL}\big(q^{\backslash i}_{\backslash i|i}\|\mathcal{N}_{\backslash i|i}\big)\Big], \quad (7)$$

where and hereinafter, $\backslash i|i$ denotes the conditional distribution of $\boldsymbol{f}_{\backslash i}$ (taking f_i out of $\boldsymbol{f}$) given f_i, namely, $q^{\backslash i}_{\backslash i|i} = q^{\backslash i}(\boldsymbol{f}_{\backslash i}|f_i)$ and $\mathcal{N}_{\backslash i|i} = \mathcal{N}(\boldsymbol{f}_{\backslash i}|f_i)$. Note that $q^{\backslash i}_{\backslash i|i}$ and $\mathcal{N}_{\backslash i|i}$ in the second term in Equation (7) are both Gaussian, and so setting them equal to one another causes that term to vanish. Furthermore, it is well known that the term $\mathrm{KL}\big(\widetilde{q}_i\|\mathcal{N}_i\big)$ is minimized w.r.t. the parameters of $\mathcal{N}_i$ by matching the first and second moments of $\widetilde{q}_i$ and $\mathcal{N}_i$. Finally, according to the usual EP logic, we recover the site function $t_i(\boldsymbol{f})$ by dividing the optimal Gaussian $\mathcal{N}(\boldsymbol{f})$ by the cavity distribution $q^{\backslash i}(\boldsymbol{f})$:

$$\begin{aligned} t_i(\boldsymbol{f}) &\propto \mathcal{N}(\boldsymbol{f})/q^{\backslash i}(\boldsymbol{f}) \\ &= \cancel{\mathcal{N}(\boldsymbol{f}_{\backslash i}|f_i)}\mathcal{N}(f_i)/(\cancel{q^{\backslash i}(\boldsymbol{f}_{\backslash i}|f_i)}q^{\backslash i}(f_i)) \\ &= \mathcal{N}(f_i)/q^{\backslash i}(f_i). \end{aligned} \quad (8)$$

Here we can see the optimal site function $t_i(f_i)$ relies solely on the local latent variable f_i, so it is sufficient to assume a univariate expression for site functions. Besides, the site function can be efficiently updated by using the marginals $\widetilde{q}(f_i)$ and $\mathcal{N}(f_i)$ only, namely, $t_i(f_i) \propto \big(\min_{\mathcal{N}_i} \mathrm{KL}(\widetilde{q}_i\|\mathcal{N}_i))/q^{\backslash i}(f_i)\big)$.

5.2. Locality Property of QP

This section proves the locality property of QP, which turns out to be rather more involved to show than is the case for EP. We first prove the following theorem, and then follow the same procedure as for EP (Equation (8)).

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

Proof. We first apply the decomposition of the L_2 norm to rewriting the $\mathrm{W}_2^2(\widetilde{q}(\boldsymbol{f}),\mathcal{N}(\boldsymbol{f}))$ as below (see detailed derivations in Appendix D.2):

$$\mathrm{W}_2^2(\widetilde{q},\mathcal{N}) = \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i - f_i'\|_2^2 + \mathrm{W}_2^2(q^{\backslash i}_{\backslash i|i},\mathcal{N}_{\backslash i|i})\Big], \quad (9)$$

where the prime indicates that the variable is from the Gaussian $\mathcal{N}$, and for simplification, we use the notation π_i for the joint distribution $\pi(f_i, f_i')$ which belongs to a set of measures $U(\widetilde{q}_i, \mathcal{N}_i)$. Since $q^{\backslash i}(\boldsymbol{f})$ is known to be Gaussian, we define it in a partitioned form:

$$q^{\backslash i}(\boldsymbol{f}) \equiv \mathcal{N}\left(\begin{bmatrix}\boldsymbol{f}_{\backslash i}\\ f_i\end{bmatrix} \middle| \begin{bmatrix}\boldsymbol{m}_{\backslash i}\\ m_i\end{bmatrix}, \begin{bmatrix}\boldsymbol{S}_{\backslash i} & \boldsymbol{S}_{\backslash ii}\\ \boldsymbol{S}_{\backslash ii}^{\mathsf{T}} & S_i\end{bmatrix}\right), \quad (10)$$

and the conditional $q^{\backslash i}(\boldsymbol{f}_{\backslash i}|f_i)$ is expressed as:

$$\begin{aligned} &q^{\backslash i}(\boldsymbol{f}_{\backslash i}|f_i) = \mathcal{N}(\boldsymbol{f}_{\backslash i}|\boldsymbol{m}_{\backslash i|i}, \boldsymbol{S}_{\backslash i|i}), \\ &\boldsymbol{m}_{\backslash i|i} = \boldsymbol{m}_{\backslash i} + \boldsymbol{S}_{\backslash ii} S_i^{-1}(f_i - m_i) \equiv \boldsymbol{a} f_i + \boldsymbol{b}, \quad (11) \\ &\boldsymbol{S}_{\backslash i|i} = \boldsymbol{S}_{\backslash i} - \boldsymbol{S}_{\backslash ii} S_i^{-1} \boldsymbol{S}_{\backslash ii}^{\mathsf{T}}. \end{aligned}$$

We define a similar partioned expression for the Gaussian $\mathcal{N}(\boldsymbol{f}')$ by adding primes to variables and parameters on the r.h.s. of Equation (10), and as a result, the conditional $\mathcal{N}(\boldsymbol{f}'_{\backslash i}|f_i')$ is similarly written as:

$$\begin{aligned} &\mathcal{N}(\boldsymbol{f}'_{\backslash i}|f_i') = \mathcal{N}(\boldsymbol{m}'_{\backslash i|i}, \boldsymbol{S}'_{\backslash i|i}), \\ &\boldsymbol{m}'_{\backslash i|i} = \boldsymbol{m}'_{\backslash i} + \boldsymbol{S}'_{\backslash ii} S_i'^{-1}(f_i' - m_i') \equiv \boldsymbol{a}' f_i' + \boldsymbol{b}', \quad (12) \\ &\boldsymbol{S}'_{\backslash i|i} = \boldsymbol{S}'_{\backslash i} - \boldsymbol{S}'_{\backslash ii} S_i'^{-1} \boldsymbol{S}'^{\mathsf{T}}_{\backslash ii}. \quad (13) \end{aligned}$$

Given the above definitions, we exploit Proposition 2 to take the means out of the L_2 WD on the r.h.s. of Equation (9):

$$\begin{aligned} \mathrm{W}_2^2\left(\widetilde{q},\mathcal{N}\right) = \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i - f_i'\|_2^2 + \|\boldsymbol{m}_{\backslash i|i} - \boldsymbol{m}'_{\backslash i|i}\|_2^2\Big] \\ + \underbrace{\mathrm{W}_2^2\Big(\mathcal{N}(\boldsymbol{0}, \boldsymbol{S}_{\backslash i|i}), \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}'_{\backslash i|i})\Big)}_{(\mathrm{A})}. \quad (14) \end{aligned}$$

Minimizing this function requires optimizing m_i', $\boldsymbol{m}'_{\backslash i}$, S_i', $\boldsymbol{S}'_{\backslash i}$ and $\boldsymbol{S}'_{\backslash ii}$. As $\boldsymbol{S}'_{\backslash i}$ is only contained in $\boldsymbol{S}_{\backslash i|i}$ and isolated into the term (A), it can be optimized by simply setting

$$\boldsymbol{S}'_{\backslash i|i} = \boldsymbol{S}_{\backslash i|i} \overset{\text{Eqn. (13)}}{\Longrightarrow} \boldsymbol{S}^{(n)*}_{\backslash i} = \boldsymbol{S}_{\backslash i|i} + \boldsymbol{S}'_{\backslash ii} S_i'^{-1} \boldsymbol{S}'^{\mathsf{T}}_{\backslash ii}. \quad (15)$$

As a result, (A) is minimized to zero.

Next, we plug in expressions of $\boldsymbol{m}_{\backslash i|i}$ and $\boldsymbol{m}'_{\backslash i|i}$ (Equation (11) and Equation (12)) into optimized Equation (14):

$$\min_{\boldsymbol{S}'_{\backslash i}}(14) = \inf_{\pi_i} \mathbb{E}_{\pi_i}\left[\|f_i - f_i'\|_2^2 + \|\boldsymbol{a} f_i - \boldsymbol{a}' f_i' + \boldsymbol{b} - \boldsymbol{b}'\|_2^2\right], \quad (16)$$

where $\boldsymbol{m}'_{\backslash i}$ is only contained by $\boldsymbol{b}'$. Thus, we can optimize it by zeroing the derivative of the above function about $\boldsymbol{m}'_{\backslash i}$, which results in:

$$\begin{aligned} &\boldsymbol{b}' = \boldsymbol{b} + \boldsymbol{a}\mu_{\widetilde{q}_i} - \boldsymbol{a}' m_i' \overset{\text{Eqn. (12)}}{\Longrightarrow} \quad (17) \\ &\boldsymbol{m}^{(n)*}_{\backslash i} = \boldsymbol{S}'_{\backslash ii} S_i'^{-1} m_i' + \boldsymbol{b} + \boldsymbol{a}\mu_{\widetilde{q}_i} - \boldsymbol{a}' m_i', \end{aligned}$$

where $\mu_{\widetilde{q}_i}$ is the mean of $\widetilde{q}(f_i)$. The minimum value of Equation (16) thereby is (see details in subsection D.3):

$$\begin{aligned} \min_{\boldsymbol{m}'_{\backslash i}}(16) = (1 + \boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}')\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2 \sigma^2_{\widetilde{q}_i} + \\ \|\boldsymbol{a}'\|_2^2 S_i' - \boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}'\Big[\sigma^2_{\widetilde{q}_i} + S_i' + (\mu_{\widetilde{q}_i} - m_i')^2\Big], \quad (18) \end{aligned}$$

where $\sigma^2_{\widetilde{q}_i}$ is the variance of $\widetilde{q}(f_i)$. This function can be further simplified using the quantile based reformulation of

$\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i)$ (see details in Appendix D.4) which results in:

$$(18) = \mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2 \sigma_{\widetilde{q}_i}^2 \underbrace{-2^{\frac{3}{2}} \boldsymbol{a}^\top \boldsymbol{a}' c_{\widetilde{q}_i} S_i'^{\frac{1}{2}} + \|\boldsymbol{a}'\|_2^2 S_i'}_{\text{(B)}}. \tag{19}$$

Now, we are left with optimizing m_i', S_i' and $\boldsymbol{S}'_{\setminus ii}$. To optimize $\boldsymbol{S}'_{\setminus ii}$, which only exists in the above term (B), we zero the derivative of (B) w.r.t. $\boldsymbol{S}'_{\setminus ii}$ and this yields:

$$\boldsymbol{a}'^* = 2^{\frac{1}{2}} (S_i')^{-\frac{1}{2}} c_{\widetilde{q}_i} \boldsymbol{a} \overset{\text{Eqn. (12)}}{\Longrightarrow} \boldsymbol{S}'^*_{\setminus ii} = (2S_i')^{\frac{1}{2}} c_{\widetilde{q}_i} \boldsymbol{a}, \tag{20}$$

and the minimum value of Equation (19) is

$$\min_{\boldsymbol{S}'_{\setminus ii}} (19) = \mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2 (\sigma_{\widetilde{q}_i}^2 - 2c_{\widetilde{q}_i}^2). \tag{21}$$

The results of optimizing m_i' and S_i' in the above equation have already been provided in Equation (5): $m_i'^* = \mu_{\widetilde{q}_i}$ and $S_i'^* = 2c_{\widetilde{q}_i}^2$. By plugging them into Equation (20) and Equation (17), we have $\boldsymbol{a}'^* = \boldsymbol{a}$ and $\boldsymbol{b}'^* = \boldsymbol{b}$. Finally, using Equation (15), we obtain $q^{\setminus i}(\boldsymbol{f}_{\setminus i}|f_i) = \mathcal{N}(\boldsymbol{f}_{\setminus i}|\boldsymbol{a}f_i + \boldsymbol{b}, \boldsymbol{S}_{\setminus i|i}) = \mathcal{N}(\boldsymbol{f}_{\setminus i}|\boldsymbol{a}'f_i + \boldsymbol{b}', \boldsymbol{S}'_{\setminus i|i}) = \mathcal{N}(\boldsymbol{f}_{\setminus i}|f_i)$, which concludes the proof. □

Theorem 4 (Locality Property of QP). *For GP models with factorized likelihoods, QP requires only univariate site functions, and so yields efficient updates using only marginal distributions.*

Proof. We apply the same steps as in Equation (8) for the EP case to QP and we conclude that the site function $t_i(f_i) \propto \mathcal{N}(f_i)/q^{\setminus i}(f_i)$ relies solely on the local latent variable f_i. And as per Equation (21), $\mathcal{N}(f_i)$ is estimated by $\min_{\mathcal{N}_i} \mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i)$, so the local update only uses marginals and can perform efficiently. □

Benefits of the Locality Property. The locality property admits an analytically economic form for the site function $t_i(f_i)$, requiring a parameterization that depends on a single latent variable. In addition, this also yields a significant reduction in the computational complexity, as only marginals are involved in each local update. In contrast, if QP (or EP) had no such a locality property, estimating the mean and the variance would involve integrals w.r.t. high-dimensional distributions, with a significantly higher computational cost should closed form expressions be unavailable.

6. Experiments

In this section, we compare QP, EP and variational Bayes (VB, Opper & Archambeau, 2009) algorithms for binary classification and Poisson regression. The experiments employ eight real world datasets and aim to compare relative accuracy of the three methods, rather than optimizing the absolute performance. The implementations of EP and VB in Python are publicly available (GPy, since 2012), and our implementation of QP is based on that of EP. Our code will be publicly available upon publication. For both EP and QP, we stop local updates, *i.e.*, the inner loop in Algorithm 1, when the root mean squared change in parameters is less than 10^{-6}. In the outer loop, the GP hyper-parameters are optimized by L-BFGS-B (Byrd et al., 1995) with a maximum of 10^3 iterations and a relative tolerance of 10^{-9} for the function value. VB is also optimized by L-BFGS-B with the same configuration. Parameters shared by the three methods are initialized to be the same.

6.1. Binary Classification

Benchmark Data. We perform binary classification experiments on the five real world datasets employed by Kuss & Rasmussen (2005): Ionosphere, Wisconsin Breast Cancer, Sonar (Dua & Graff, 2017), Leptograpsus Crabs and Pima Indians Diabetes (Ripley, 1996). We use two additional UCI datasets as further evidence: Glass and Wine (Dua & Graff, 2017). As the Wine dataset has three classes, we conduct binary classification experiments on all pairs of classes. We summarize the dataset size and data dimensions in Table 1.

Prediction. We predict the test labels using models optimized by EP, QP and VB on the training data. For a test input $\boldsymbol{x}_*$ with a binary target y_*, the approximate predictive distribution is written as: $q(y_*|\boldsymbol{x}_*) = \int_{-\infty}^{\infty} p(y_*|f_*) q(f_*) \, \mathrm{d}f_*$ where $f_* = f(\boldsymbol{x}_*)$ is the value of the latent function at $\boldsymbol{x}_*$. We use the probit likelihood for the binary classification task, which admits an analytical expression for the predictive distribution. Correspondingly, the predicted label $\hat{y}_*$ is determined by thresholding the predictive probability at $1/2$.

Performance Evaluation. To evaluate the performance, we employ two measures: the test error (TE) and the negative test log-likelihood (NTLL). The TE and the NTLL quantify the prediction accuracy and uncertainty, respectively. Specifically, they are defined as $(\sum_{i=1}^m |y_{*,i} - \hat{y}_{*,i}|/2)/m$ and $-(\sum_{i=1}^m \log q(y_{*,i}|\boldsymbol{x}_{*,i}))/m$, respectively, for a set of test inputs $\{\boldsymbol{x}_{*,i}\}_{i=1}^m$, test labels $\{y_{*,i}\}_{i=1}^m$, and the corresponding predicted labels $\{\hat{y}_{*,i}\}_{i=1}^m$. Lower values indicate better performance for both measures.

Experiment Settings. In the experiments, we randomly split each dataset into 10 folds, each time using 1 fold for testing and the other 9 folds for training, with features standardized to zero mean and unit standard deviation. We repeat this 100 times for a random seed ranging 0 through 99. As a result, there are a total of 1,000 experiments for each dataset. We report the average and the standard deviation of the above metrics over the 100 rounds.

Results. The evaluation results are summarized in Table 1. The top section presents the results on the datasets employed

Table 1: Results on benchmark datasets. The first three columns give dataset names, the number of instances m and the number of features n. The table records the test errors (TEs) and the negative test log-likelihoods (NTLLs). The top section is on the benchmark datasets employed by Kuss & Rasmussen (2005) and the middle section uses additional datasets. * indicates that QP outperforms EP in more than 90% of experiments.

			TE ($\times 10^{-2}$)			NTLL($\times 10^{-3}$)		
Data	m	n	EP	QP	VB	EP	QP	VB
`Ionosphere`	351	34	$\mathbf{7.9_{\pm 0.5}}$	$\mathbf{7.9_{\pm 0.5}}$	$18.9_{\pm 6.9}$	$\mathbf{215.9_{\pm 8.4}}$	$\mathbf{215.9_{\pm 8.5}}$	$337.4_{\pm 70.8}$
`Cancer`	683	9	$3.2_{\pm 0.2}$	$3.2_{\pm 0.2}$	$\mathbf{3.1_{\pm 0.2}}$	$\mathbf{88.2_{\pm 3.1}}$	$\mathbf{88.2^{*}_{\pm 3.1}}$	$88.9_{\pm 19.1}$
`Pima`	732	7	$\mathbf{20.3_{\pm 1.0}}$	$\mathbf{20.3_{\pm 1.0}}$	$21.9_{\pm 0.4}$	$424.7_{\pm 13.0}$	$\mathbf{424.0^{*}_{\pm 13.2}}$	$450.3_{\pm 2.6}$
`Crabs`	200	7	$\mathbf{2.7_{\pm 0.5}}$	$\mathbf{2.7_{\pm 0.5}}$	$3.7_{\pm 0.7}$	$64.4_{\pm 8.2}$	$\mathbf{64.3_{\pm 8.4}}$	$164.7_{\pm 7.5}$
`Sonar`	208	60	$\mathbf{14.0_{\pm 1.1}}$	$\mathbf{14.0_{\pm 1.1}}$	$25.7_{\pm 3.9}$	$306.7_{\pm 10.8}$	$\mathbf{306.2^{*}_{\pm 10.9}}$	$693.1_{\pm 0.0}$
`Glass`	214	10	$1.1_{\pm 0.4}$	$\mathbf{1.0_{\pm 0.4}}$	$2.6_{\pm 0.5}$	$29.5_{\pm 5.4}$	$\mathbf{29.0^{*}_{\pm 5.5}}$	$79.5_{\pm 6.3}$
`Wine1`	130	13	$\mathbf{1.5_{\pm 0.5}}$	$\mathbf{1.5_{\pm 0.5}}$	$1.7_{\pm 0.6}$	$48.0_{\pm 3.4}$	$\mathbf{47.4^{*}_{\pm 3.4}}$	$83.9_{\pm 5.2}$
`Wine2`	107	13	$\mathbf{0.0_{\pm 0.0}}$	$\mathbf{0.0_{\pm 0.0}}$	$\mathbf{0.0_{\pm 0.0}}$	$18.0_{\pm 1.2}$	$\mathbf{17.8^{*}_{\pm 1.2}}$	$26.7_{\pm 1.9}$
`Wine3`	119	13	$2.0_{\pm 1.0}$	$2.0_{\pm 1.0}$	$\mathbf{1.2_{\pm 0.7}}$	$52.1_{\pm 5.6}$	$\mathbf{51.8^{*}_{\pm 5.6}}$	$69.4_{\pm 5.0}$
`Coal Mining`	112	1	$\mathbf{118.6_{\pm 27.0}}$	$\mathbf{118.6_{\pm 27.0}}$	$170.3_{\pm 15.9}$	$1606.8_{\pm 116.3}$	$\mathbf{1606.5_{\pm 116.3}}$	$2007.3_{\pm 119.8}$

Note: `Wine1`: Class 1 vs. 2. `Wine2`: Class 1 vs. 3. `Wine3`: Class 2 vs. 3.

by Kuss & Rasmussen (2005), whose reported TEs match ours as expected. While QP and EP exhibit similar TEs on these datasets, QP is superior to EP in terms of the NTLL. VB under-performs both EP and QP on all datasets except `Cancer`. The middle section of Table 1 expands to results to additional datasets. The TEs are again similar for EP and QP, while QP has lower NTLLs. Again, VB performs worst among the three methods. To emphasize the difference between NTLLs of EP and QP, we mark with an asterisk those results in which QP surpasses EP in more than 90% of the experiments. Furthermore, we visualise the predictive variances of QP in comparison with those of EP in Figure 1a., which shows that the variances of QP are always less than or equal to those of EP, thereby providing empirical evidence of QP alleviating the over-estimation of predictive variances associated with the EP algorithm.

6.2. Poisson Regression

Data and Settings. We perform a Poisson regression experiment to further evaluate the performance of our method. The experiment employs the coal-mining disaster dataset (Jarrett, 1979) which has 190 data points indicating the time of fatal coal mining accidents in the United Kingdom from 1851 to 1962. To generate training and test sequences, we randomly assign every point of the original sequence to either a training or test sequence with equal probability, and this is repeated 200 times (random seeds $0, \cdots, 199$), resulting in 200 pairs of training and test sequences. We use the TE and the NTLL to evaluate the fitting performance of the model on the test dataset. The NTLL has the same expression as that of the Binary classification experiment, but with a different predictive distribution $q(y_*|\boldsymbol{x}_*)$. The TE is defined slightly differently as $(\sum_{i=1}^{m} |y_{*,i} - \hat{y}_{*,i}|)/m$. To make the rate parameter of the Poisson likelihood non-negative, we use the square link function (Flaxman et al., 2017; Walder & Bishop, 2017), and as a result, the likelihood becomes $p(y|f^2)$. We use this link function because it is more mathematically convenient than the exponential function: the EP and QP update formulas, and the predictive distribution $q(y_*|\boldsymbol{x}_*)$ are available in Appendices C.2 and F, respectively.

Results. The means and the standard deviations of the evaluation results are reported in the last row of Table 1. Compared with EP, QP yields lower NTLL, which implies a better fitting performance of QP to the test sequences. We also provide the predictive variances in Figure 1b., in the variance of QP is once again seen to be less than or equal to that of EP. This experiment further supports our claim that QP alleviates the problem with EP of over-estimation of the predictive variance. Finally, once again we find that both EP and QP outperform VB.

7. Conclusions

We have proposed QP as an approximate Bayesian inference method for Gaussian process models with factorized likelihoods. Algorithmically, QP is similar to EP but uses the L_2 WD instead of the forward KL divergence for estimation of the site functions. When the likelihood factors are approximated by a Gaussian form we show that QP matches quantile functions rather than moments as in EP. Furthermore, we show that QP has the same mean update but a smaller variance than that of EP, which in turn alleviates the over-estimation by EP of the posterior variance in practice. Crucially, QP has the same favorable locality property as EP, and thereby admits efficient updates. Our experiments on binary classification and Poisson regression have shown that QP can outperform both EP and variational Bayes.

References

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

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

Benamou, J.-D. and Brenier, Y. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, Jan 2000. ISSN 0945-3245. doi: 10.1007/s002110050002.

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

Bui, T. D., Yan, J., and Turner, R. E. A unifying framework for gaussian process pseudo-point approximations using power expectation propagation. *J. Mach. Learn. Res.*, 18 (1):3649–3720, January 2017. ISSN 1532-4435.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995.

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

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

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

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

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. Variational inference via \chi upper bound minimization. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2732–2741. Curran Associates, Inc., 2017.

Dowson, D. and Landau, B. The frchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450 – 455, 1982. ISSN 0047-259X. doi: https://doi.org/10.1016/0047-259X(82)90077-X.

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

El Moselhy, T. A. and Marzouk, Y. M. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

Flaxman, S., Teh, Y. W., Sejdinovic, D., et al. Poisson intensity estimation with reproducing kernels. *Electronic Journal of Statistics*, 11(2):5081–5104, 2017.

Gelbrich, M. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990. doi: 10.1002/mana.19901470121.

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

GPy. GPy: A gaussian process framework in python. `http://github.com/SheffieldML/GPy`, since 2012.

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

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

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

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

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

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

Jarrett, R. A note on the intervals between coal-mining disasters. *Biometrika*, 66(1):191–193, 1979.

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

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

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

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

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

Mallasto, A. and Feragen, A. 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. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5660–5670. Curran Associates, Inc., 2017.

Matheron, G. Principles of geostatistics. *Economic Geology*, 58(8):1246–1266, 12 1963.

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

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

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

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

Minka, T. P. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001c.

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

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

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

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

Opper, M. and Archambeau, C. The variational gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

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

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

Owen, D. B. Tables for computing bivariate normal probabilities. *Ann. Math. Statist.*, 27(4):1075–1090, 12 1956. doi: 10.1214/aoms/1177728074.

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

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

Ripley, B. D. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996. doi: 10.1017/CBO9780511812651.

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

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

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

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

Takatsu, A. Wasserstein geometry of gaussian measures. *Osaka J. Math.*, 48(4):1005–1026, 12 2011.

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

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Walder, C. J. and Bishop, A. N. Fast bayesian intensity estimation for the permanental process. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3579–3588. JMLR. org, 2017.

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

Williams, C. K. I. and Barber, D. Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, Dec 1998. ISSN 1939-3539. doi: 10.1109/34.735807.

Wolfram, R. I. Mathematica, Version 12.0, 2019. Champaign, IL, 2019.

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

Zhang, R., Walder, C., Rizoiu, M.-A., and Xie, L. Efficient Non-parametric Bayesian Hawkes Processes. In *International Joint Conference on Artificial Intelligence*, pp. 4299–4305, 2019.

Zhang, R., Walder, C., and Rizoiu, M. A. Variational inference for sparse gaussian process modulated hawkes process. *the 34th AAAI Conference on Artificial Intelligence (AAAI-20)*, 2020.

Zou, G. A modified poisson regression approach to prospective studies with binary data. *American journal of epidemiology*, 159(7):702–706, 2004.

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

A. Minimization of L_2 WD between Univariate Gaussian and Non-Gaussian Distributions

In this section, we derive the formulas of the optimal μ^* and σ^* for the L_2 WD, *i.e.*, Eqn. (5). Recall the optimization problem: we use a univariate Gaussian distribution $\mathcal{N}(f|\mu,\sigma^2)$ to approximate a univariate non-Gaussian distribution $q(f)$ by minimizing the L_2 WD 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 defined in Proposition 1.

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:

$$\begin{aligned}\mu^* &= \int_0^1 F_q^{-1}(y) - \sqrt{2}\sigma \mathrm{erf}^{-1}(2y-1)\,\mathrm{d}y\\ &= \int_{-\infty}^{\infty} xq(x)\,\mathrm{d}x - \frac{\sqrt{2}}{2}\sigma \int_{-1}^{1} \mathrm{erf}^{-1}(y)\,\mathrm{d}y\\ &= \mu_q - \sqrt{2}\sigma \int_{-\infty}^{\infty} x\mathcal{N}(x|0,1/2)\,\mathrm{d}x\\ &= \mu_q,\\ \sigma^* &= \sqrt{2}\int_0^1 (F_q^{-1}(y)-\mu)\mathrm{erf}^{-1}(2y-1)\,\mathrm{d}y \Big/ \int_0^1 2(\mathrm{erf}^{-1})^2(2y-1)\,\mathrm{d}y\\ &= \sqrt{2}\int_0^1 F_q^{-1}(y)\mathrm{erf}^{-1}(2y-1)\,\mathrm{d}y \Big/ \underbrace{\int_{-\infty}^{\infty} 2x^2\mathcal{N}(x|0,1/2)\,\mathrm{d}x}_{=1}\\ &= \sqrt{2}\int_0^1 F_q^{-1}(y)\mathrm{erf}^{-1}(2y-1)\,\mathrm{d}y. \qquad (22)\end{aligned}$$

B. Minimization of L_p WD between Univariate Gaussian and Non-Gaussian Distributions

In this section, we describe a gradient descent approach to minimizing an L_p WD, for $p \neq 2$, in order to handle cases with no analytical expressions for the optimal parameters. Our goal is to use a univariate Gaussian distribution $\mathcal{N}(f|\mu,\sigma^2)$ to approximate a univariate non-Gaussian distribution $q(f)$. Specifically, we seek the minimiser in μ and σ of $\mathrm{W}_p^p(q,\mathcal{N})$; the derivatives of the objective function about μ and σ are:

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

where for simplification, we define $\varepsilon(y) = F_q^{-1}(y) - \mu - \sqrt{2}\sigma\mathrm{erf}^{-1}(2y-1)$ and $\eta(x) = x - \mu - \sqrt{2}\sigma\mathrm{erf}^{-1}(2F_q(x)-1)$, with F_q and F_q^{-1} being the CDF and the quantile function of q. Note the derivatives have no analytical expressions. However, if the CDF F_q is available, we can use the standard numerical integration routines; otherwise, we resort to Monte Carlo sampling. In the framework of EP or QP, $q(x) \propto q^{\backslash i}(x)p(y_i|x)$ and $q^{\backslash i}$ is Gaussian, so we may draw samples from a Gaussian proposal distribution to obtain a simple Monte Carlo method.

C. Computations for Different Likelihoods

Given the likelihood $p(y|f)$ and the cavity distribution $q^{\backslash i}(f) = \mathcal{N}(f|\mu, \sigma^2)$, a stable way to compute the mean and the variance of the tilted distribution $\widetilde{q}(f) = p(y|f)q^{\backslash i}(f)/Z$ where the normalizer $Z = \int_{-\infty}^{\infty} p(y|f)q^{\backslash i}(f)\,\mathrm{d}f$, can be found in the software manual of Rasmussen & Williams (2005). We present the key formulae below, for use in subsequent derivations:

$$
\begin{aligned}
\partial_\mu Z &= \int_{-\infty}^{\infty} \frac{f-\mu}{\sigma^2} p(y|f)\mathcal{N}(f|\mu,\sigma^2)\,\mathrm{d}f \\
\frac{\partial_\mu Z}{Z} &= \frac{1}{\sigma^2}\int_{-\infty}^{\infty} f \frac{p(y|f)\mathcal{N}(f|\mu,\sigma^2)}{Z}\,\mathrm{d}f - \frac{\mu}{\sigma^2}\int_{-\infty}^{\infty} \frac{p(y|f)\mathcal{N}(f|\mu,\sigma^2)}{Z}\,\mathrm{d}y \\
\frac{\partial_\mu Z}{Z} &= \frac{1}{\sigma^2}\mu_{\widetilde{q}} - \frac{\mu}{\sigma^2} \\
\Longrightarrow \mu_{\widetilde{q}} &= \frac{\sigma^2 \partial_\mu Z}{Z} + \mu = \sigma^2 \partial_\mu \log Z + \mu, \\
\partial_\mu^2 Z &= \int_{-\infty}^{\infty} -\frac{1}{\sigma^2} p(y|f)\mathcal{N}(f|\mu,\sigma^2) + \left(\frac{f-\mu}{\sigma^2}\right)^2 p(y|f)\mathcal{N}(f|\mu,\sigma^2)\,\mathrm{d}f \\
\frac{\partial_\mu^2 Z}{Z} &= \int_{-\infty}^{\infty} \left(-\frac{1}{\sigma^2} + \frac{\mu^2}{\sigma^4} + \frac{f^2}{\sigma^4} - \frac{2\mu f}{\sigma^4}\right) \frac{p(y|f)\mathcal{N}(f|\mu,\sigma^2)}{Z}\,\mathrm{d}f \\
\frac{\partial_\mu^2 Z}{Z} &= -\frac{1}{\sigma^2} + \frac{\mu^2}{\sigma^4} + \frac{1}{\sigma^4}(\sigma_{\widetilde{q}}^2 + \mu_{\widetilde{q}}^2) - \frac{2\mu}{\sigma^4}\mu_{\widetilde{q}} \\
\frac{\partial_\mu^2 Z}{Z} &= -\frac{1}{\sigma^2} + \frac{\sigma_{\widetilde{q}}^2}{\sigma^4} + \frac{(\mu - \mu_{\widetilde{q}})^2}{\sigma^4} = -\frac{1}{\sigma^2} + \frac{\sigma_{\widetilde{q}}^2}{\sigma^4} + \left(\frac{\partial_\mu Z}{Z}\right)^2 \\
\Longrightarrow \sigma_{\widetilde{q}}^2 &= \sigma^4 \left[\frac{\partial_\mu^2 Z}{Z} - \left(\frac{\partial_\mu Z}{Z}\right)^2\right] + \sigma^2 = \sigma^4 \partial_\mu^2 \log Z + \sigma^2.
\end{aligned}
$$

C.1. Probit Likelihood for Binary Classification

For the binary classification with labels $y \in \{-1, 1\}$, the PDF of the tilted distribution $\widetilde{q}(f)$ with the probit likelihood is provided by Rasmussen & Williams (2005):

$$
\widetilde{q}(f) = Z^{-1}\Phi(fy)\mathcal{N}(f|\mu,\sigma^2), \;\; Z = \Phi(z), \;\; z = \frac{\mu}{y\sqrt{1+\sigma^2}},
$$

and the mean estimate also has a closed form expression:

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

As per Eqn. (6), the computation of the optimal $\sigma^\star$ requires the CDF of $\widetilde{q}$, denoted as $F_{\widetilde{q}}$. For positive $y > 0$, the CDF is derived as

$$
\begin{aligned}
F_{\widetilde{q},y>0}(x) &= Z^{-1}\int_{-\infty}^{x} \Phi\left(fy\right)\mathcal{N}\left(f|\mu,\sigma^2\right)\,\mathrm{d}f \\
&= \frac{Z^{-1}}{2\pi\sigma y}\int_{-\infty}^{\mu}\int_{-\infty}^{x-\mu} \exp\left(-\frac{1}{2}\begin{bmatrix} w \\ f \end{bmatrix}^T \begin{bmatrix} v^{-2}+\sigma^{-2} & v^{-2} \\ v^{-2} & v^{-2} \end{bmatrix} \begin{bmatrix} w \\ f \end{bmatrix}\right)\mathrm{d}w\,\mathrm{d}f \\
&= Z^{-1}\int_{-\infty}^{k}\int_{-\infty}^{h} \mathcal{N}\left(\begin{bmatrix} w \\ f \end{bmatrix}\middle|\mathbf{0}, \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}\right)\mathrm{d}w\,\mathrm{d}f \\
&\overset{\text{(a)}}{=} Z^{-1}\left[\frac{1}{2}\Phi(h) - T\left(h, \frac{k+\rho h}{h\sqrt{1-\rho^2}}\right) + \frac{1}{2}\Phi(k) - T\left(k, \frac{h+\rho k}{k\sqrt{1-\rho^2}}\right) + \eta\right] \\
k &= \frac{\mu}{\sqrt{\sigma^2+1}}, \;\; h = \frac{x-\mu}{\sigma}, \;\; \rho = \frac{1}{\sqrt{1+1/\sigma^2}}
\end{aligned}
$$

where the step (a) is obtained by exploiting the work of Owen (1956) and $T(\cdot,\cdot)$ 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

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

Similarly, for $y<0$, the CDF is

$$F_{\widetilde{q},y<0}(x)=Z^{-1}\left[\frac{1}{2}\Phi(h)+T\left(h,\frac{k+\rho h}{h\sqrt{1-\rho^2}}\right)-\frac{1}{2}\Phi(k)+T\left(k,\frac{h+\rho k}{k\sqrt{1-\rho^2}}\right)-\eta\right].$$

Summarizing the two cases, we get the closed form expression of $F_{\widetilde{q}}$:

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

Provided the above, the optimal $\sigma^\star$ can be computed by numerical integration of Eqn (22).

C.2. Square Link Function for Poisson Regression

Consider Poisson regression, which uses the Poisson likelihood $p(y|g)=g^y\exp(-g)/y!$ to model count data $y\in\mathbb{N}$, with the square link function $g(f)=f^2$ (Walder & Bishop, 2017; Flaxman et al., 2017). We use the square link function because it is more mathematically convenient than the exponential function. Given the cavity distribution $q^{\backslash i}(f)=\mathcal{N}(f|\mu,\sigma^2)$, we want the tilted distribution $\widetilde{q}(f)=q^{\backslash i}(f)p(y|g(f))/Z$ where the normalizer Z is derived as:

$$\begin{aligned}
Z&=\int_{-\infty}^{\infty} q^{\backslash i}(f)p(y|g)\,\mathrm{d}f\\
&=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(f-\mu)^2}{2\sigma^2}\right)f^{2y}\exp(-f^2)/y!\,\mathrm{d}f\\
&\overset{\underline{(a)}}{=}\frac{1}{\sqrt{2\pi\sigma^2}y!\exp(\mu^2/(1+2\sigma^2))}\int_{-\infty}^{\infty}f^{2y}\exp\left(-\frac{(f-\mu/(1+2\sigma^2))^2}{2\sigma^2/(1+2\sigma^2)}\right)\,\mathrm{d}f\\
&\overset{\underline{(b)}}{=}\frac{\left(\frac{2\sigma^2}{1+2\sigma^2}\right)^{y+\frac{1}{2}}}{\sqrt{2\pi\sigma^2}y!\exp(\mu^2/(1+2\sigma^2))}\Gamma\left(y+\frac{1}{2}\right)\,{}_1F_1\left(-y;\frac{1}{2};-\frac{\mu^2}{2\sigma^2(1+2\sigma^2)}\right)\\
&=\frac{\alpha^{y+\frac{1}{2}}}{\sqrt{2\pi\sigma^2}y!\exp(h)}\Gamma\left(y+\frac{1}{2}\right)\,{}_1F_1\left(-y;\frac{1}{2};-\frac{h}{2\sigma^2}\right),\\
\alpha&=\frac{2\sigma^2}{1+2\sigma^2},\quad h=\frac{\mu^2}{1+2\sigma^2}
\end{aligned}\tag{23}$$

where the step (a) rewrites the product of two exponential functions into the form of the Gaussian distribution, (b) is achieved through Mathematica (Wolfram, 2019), $\Gamma(\cdot)$ is the Gamma function and ${}_1F_1\left(-y;\frac{1}{2};-\frac{h^2}{2\sigma^2}\right)$ is the confluent hypergeometric function of the first kind. Furthermore, we compute the first derivative of $\log Z$ w.r.t. μ and then the mean of the tilted distribution:

$$\begin{aligned}
\partial_\mu\log Z&=\left(\frac{y\,{}_1F_1\left(-y+1;\frac{3}{2};-\frac{h}{2\sigma^2}\right)}{\sigma^2\,{}_1F_1\left(-y;\frac{1}{2};-\frac{h}{2\sigma^2}\right)}-1\right)\frac{2\mu}{1+2\sigma^2}\\
\Longrightarrow \mu_{\widetilde{q}}&=\sigma^2\partial_\mu\log Z+\mu.\\
\partial_\mu^2\log Z&=\left(\frac{y\,{}_1F_1\left(-y+1;\frac{3}{2};-\frac{h}{2\sigma^2}\right)}{\sigma^2\,{}_1F_1\left(-y;\frac{1}{2};-\frac{h}{2\sigma^2}\right)}-1\right)\frac{2}{1+2\sigma^2}-\\
&\quad\left(\frac{2(1-y)\,{}_1F_1\left(-y+2;\frac{5}{2};-\frac{h}{2\sigma^2}\right)}{3\,{}_1F_1\left(-y;\frac{1}{2};-\frac{h}{2\sigma^2}\right)}+\frac{2y\,{}_1F_1\left(-y+1;\frac{3}{2};-\frac{h}{2\sigma^2}\right)^2}{{}_1F_1\left(-y;\frac{1}{2};-\frac{h}{2\sigma^2}\right)^2}\right)\frac{2\mu^2y}{\sigma^4(1+2\sigma^2)^2}
\end{aligned}$$

$$\implies \sigma_{\widetilde{q}}^2 = \sigma^4 \partial_\mu^2 \log Z + \sigma^2$$

Finally, we derive the CDF of the tilted distribution $\widetilde{q}$ by using the binomial theorem:

$$\begin{aligned}
F_{\widetilde{q}}(x) &= Z^{-1} \int_{-\infty}^{x} p(y|g)\mathcal{N}(f|\mu, \sigma^2)\, \mathrm{d}f \\
&\overset{\text{(a)}}{=} A \int_{-\infty}^{x} f^{2y} \exp\left(-\frac{(f-\mu/(1+2\sigma^2))^2}{2\sigma^2/(1+2\sigma^2)}\right)\, \mathrm{d}f \\
&= A \int_{-\infty}^{x-\frac{\mu}{1+2\sigma^2}} \left(f + \frac{\mu}{1+2\sigma^2}\right)^{2y} \exp\left(-\frac{f^2}{2\sigma^2/(1+2\sigma^2)}\right)\, \mathrm{d}f \\
&\overset{\text{(b)}}{=} A \int_{-\infty}^{x-\beta} \left[\sum_{k=0}^{2y} \binom{2y}{k} f^k \beta^{2y-k}\right] \exp\left(-\frac{f^2}{\alpha}\right)\, \mathrm{d}f \\
&= A \sum_{k=0}^{2y} \binom{2y}{k} \beta^{2y-k} \left[\int_{-\infty}^{0} f^k \exp\left(-\frac{f^2}{\alpha}\right)\, \mathrm{d}f + \int_{0}^{x-\beta} f^k \exp\left(-\frac{f^2}{\alpha}\right)\, \mathrm{d}f\right] \\
&\overset{\text{(c)}}{=} \frac{A}{2} \sum_{k=0}^{2y} \binom{2y}{k} \beta^{2y-k} \alpha^{\frac{k+1}{2}} \left[(-1)^k \Gamma\left(\frac{k+1}{2}\right) + \mathrm{sgn}(x-\beta)^{k+1} \left(\Gamma\left(\frac{k+1}{2}\right) - \Gamma\left(\frac{k+1}{2}, \frac{(x-\beta)^2}{\alpha}\right)\right)\right]
\end{aligned}$$

$$A = \frac{Z^{-1}}{\sqrt{2\pi\sigma^2}y!\exp(\mu^2/(1+2\sigma^2))} = \left[\alpha^{y+\frac{1}{2}}\Gamma\left(y+\frac{1}{2}\right)\, {}_1F_1\left(-y; \frac{1}{2}; -\frac{h}{2\sigma^2}\right)\right]^{-1}, \;\; \beta = \frac{\mu}{1+2\sigma^2},$$

where the step (a) has been derived in (a) of Eqn. (23), (b) applies the binomial theorem and (c) is obtained through Mathematica (Wolfram, 2019). And, the function $\Gamma(a, z) = \int_z^\infty t^{a-1}e^{-t}\, \mathrm{d}t$ is the upper incomplete gamma function and $\mathrm{sgn}(x)$ is the sign function, equaling 1 when $x > 0$, 0 when $x = 0$ and -1 when $x < 0$.

D. Details of Proof of Locality Property

D.1. Details of Eqn. (7)

$$\begin{aligned}
\mathrm{KL}(\widetilde{q}(\boldsymbol{f})\|\mathcal{N}(\boldsymbol{f})) &= \int \widetilde{q}(\boldsymbol{f}) \log \frac{\widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i)\widetilde{q}(f_i)}{\mathcal{N}(\boldsymbol{f}_{\setminus i}|f_i)\mathcal{N}(f_i)}\, \mathrm{d}\boldsymbol{f} \\
&= \int \widetilde{q}(f_i) \log \frac{\widetilde{q}(f_i)}{\mathcal{N}(f_i)}\, \mathrm{d}f_i + \int \widetilde{q}(f_i) \int \widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i) \log \frac{\widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i)}{\mathcal{N}(\boldsymbol{f}_{\setminus i}|f_i)}\, \mathrm{d}\boldsymbol{f}_{\setminus i}\, \mathrm{d}f_i \\
&= \mathrm{KL}\big(\widetilde{q}(f_i)\|\mathcal{N}(f_i)\big) + \mathbb{E}_{\widetilde{q}(f_i)}\Big[\mathrm{KL}\big(\widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i)\|\mathcal{N}(\boldsymbol{f}_{\setminus i}|f_i)\big)\Big] \\
\widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i) &= \frac{\widetilde{q}(\boldsymbol{f})}{\widetilde{q}(f_i)} \propto \frac{p(\boldsymbol{f})\cancel{p(y_i|f_i)}\prod_{j\neq i} t_j(\boldsymbol{f})}{q^{\setminus i}(f_i)\cancel{p(y_i|f_i)}} \\
&= q^{\setminus i}(\boldsymbol{f}_{\setminus i}|f_i).
\end{aligned} \tag{24}$$

D.2. Details of Eqn. (9)

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

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

where the superscript prime indicates that the variable is from the Gaussian $\mathcal{N}$. In (a), $\pi_i = \pi(f_i, f_i')$ and $\pi_{\setminus i|i} = \pi(\boldsymbol{f}_{\setminus i}, \boldsymbol{f}'_{\setminus i}|f_i, f_i')$. 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 $q^{\setminus i}(\boldsymbol{f}_{\setminus i}|f_i)$ (refer to Eqn. (24)).

D.3. Details of Eqn. (18)

$$
\begin{aligned}
&\min_{\boldsymbol{m}'_{\setminus i}} \text{Eqn. (16)} \\
&= \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i - f_i'\|_2^2 + \|\boldsymbol{a}(f_i - \mu_{\widetilde{q}_i}) - \boldsymbol{a}'(f_i' - m_i')\|_2^2\Big] \\
&= \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i - f_i'\|_2^2\Big] + \|\boldsymbol{a}\|_2^2\sigma^2_{\widetilde{q}_i} + \|\boldsymbol{a}'\|_2^2 S_i' - 2\boldsymbol{a}^\mathsf{T}\boldsymbol{a}'\mathbb{E}_{\pi_i}\Big(f_i f_i' - \mu_{\widetilde{q}_i} m_i'\Big) \\
&= \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i - f_i'\|_2^2\Big] + \|\boldsymbol{a}\|_2^2\sigma^2_{\widetilde{q}_i} + \|\boldsymbol{a}'\|_2^2 S_i' + \boldsymbol{a}^\mathsf{T}\boldsymbol{a}'\mathbb{E}_{\pi_i}\Big(\|f_i - f_i'\|_2^2 - f_i^2 - (f_i')^2 + 2\mu_{\widetilde{q}_i} m_i'\Big) \\
&= \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i - f_i'\|_2^2\Big] + \|\boldsymbol{a}\|_2^2\sigma^2_{\widetilde{q}_i} + \|\boldsymbol{a}'\|_2^2 S_i' + \boldsymbol{a}^\mathsf{T}\boldsymbol{a}'\mathbb{E}_{\pi_i}\Big(\|f_i - f_i'\|_2^2 - (f_i - \mu_{\widetilde{q}_i})^2 - \\
&\quad 2 f_i \mu_{\widetilde{q}_i} + \mu^2_{\widetilde{q}_i} - (f_i' - m_i')^2 - 2 f_i' m_i' + (m_i')^2 + 2\mu_{\widetilde{q}_i} m_i'\Big) \\
&= (1 + \boldsymbol{a}^\mathsf{T}\boldsymbol{a}')\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2\sigma^2_{\widetilde{q}_i} + \|\boldsymbol{a}'\|_2^2 S_i' - \boldsymbol{a}^\mathsf{T}\boldsymbol{a}'\Big(\sigma^2_{\widetilde{q}_i} + \mu^2_{\widetilde{q}_i} + S_i' + (m_i')^2 - 2\mu_{\widetilde{q}_i} m_i'\Big) \\
&= (1 + \boldsymbol{a}^\mathsf{T}\boldsymbol{a}')\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2\sigma^2_{\widetilde{q}_i} + \|\boldsymbol{a}'\|_2^2 S_i' - \boldsymbol{a}^\mathsf{T}\boldsymbol{a}'\Big[\sigma^2_{\widetilde{q}_i} + S_i' + (\mu_{\widetilde{q}_i} - m_i')^2\Big]
\end{aligned}
$$

D.4. Details of Eqn. (18)

We first use Proposition 1 to reformulate the L_2 WD $\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i)$ as:

$$
\begin{aligned}
\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) &= \int_0^1 \left(F^{-1}_{\widetilde{q}_i}(y) - m_i' - \sqrt{2S_i'}\mathrm{erf}^{-1}(2y-1)\right)^2 \mathrm{d}y, \\
&= \int_0^1 (F^{-1}_{\widetilde{q}_i}(y) - m_i')^2 + 2S_i'\mathrm{erf}^{-1}(2y-1)^2 - 2\sqrt{2S_i'}\mathrm{erf}^{-1}(2y-1)(F^{-1}_{\widetilde{q}_i}(y) - m_i')\,\mathrm{d}y, \\
&= \int_0^1 (F^{-1}_{\widetilde{q}_i}(y) - \mu_{\widetilde{q}_i} + \mu_{\widetilde{q}_i} - m_i')^2\,\mathrm{d}y + S_i' - 2\sqrt{2S_i'}c_{\widetilde{q}_i}, \\
&= \sigma^2_{\widetilde{q}_i} + (\mu_{\widetilde{q}_i} - m_i')^2 + S_i' - 2c_{\widetilde{q}_i}\sqrt{2S_i'},
\end{aligned}
$$

where $F^{-1}_{\widetilde{q}_i}(y)$ is the quantile function of $\widetilde{q}(f_i)$ and $c_{\widetilde{q}_i} \equiv \int_0^1 F^{-1}_{\widetilde{q}_i}(y)\mathrm{erf}^{-1}(2y-1)\,\mathrm{d}y$. Next, we plug this reformulation into Eqn. (18):

$$
\begin{aligned}
\text{Eqn. (18)} &= \mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \boldsymbol{a}^\mathsf{T}\boldsymbol{a}'\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2\sigma^2_{\widetilde{q}_i} + \|\boldsymbol{a}'\|_2^2 S_i' - \boldsymbol{a}^\mathsf{T}\boldsymbol{a}'\Big[\sigma^2_{\widetilde{q}_i} + S_i' + (\mu_{\widetilde{q}_i} - m_i')^2\Big] \\
&= \mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \boldsymbol{a}^\mathsf{T}\boldsymbol{a}'\,\Big[\cancel{\sigma^2_{\widetilde{q}_i} + (\mu_{\widetilde{q}_i} - m_i')^2 + S_i'} - 2c_{\widetilde{q}_i}\sqrt{2S_i'}\Big] + \|\boldsymbol{a}\|_2^2\sigma^2_{\widetilde{q}_i} + \|\boldsymbol{a}'\|_2^2 S_i' \\
&\quad - \boldsymbol{a}^\mathsf{T}\boldsymbol{a}'\Big[\cancel{\sigma^2_{\widetilde{q}_i} + S_i' + (\mu_{\widetilde{q}_i} - m_i')^2}\Big] \\
&= \mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) - 2c_{\widetilde{q}_i}\sqrt{2S_i'}\boldsymbol{a}^\mathsf{T}\boldsymbol{a}' + \|\boldsymbol{a}\|_2^2\sigma^2_{\widetilde{q}_i} + \|\boldsymbol{a}'\|_2^2 S_i'
\end{aligned}
$$

E. More Details of EP

We use the expressions $\widetilde{q}(\boldsymbol{f}) = q^{\setminus i}(\boldsymbol{f})p(y_i|f_i)/Z_{\widetilde{q}}$ and $q^{\setminus i}(\boldsymbol{f}) = q(\boldsymbol{f})/(t_i(f_i)Z_{q^{\setminus i}})$, and the derivation of $\mathrm{KL}(\widetilde{q}(\boldsymbol{f})\|q(\boldsymbol{f})) = \mathrm{KL}(\widetilde{q}(f_i)\|q(f_i))$ is shown as below:

$$
\begin{aligned}
\mathrm{KL}(\widetilde{q}(\boldsymbol{f})\|q(\boldsymbol{f})) &= \int \widetilde{q}(\boldsymbol{f}) \log \frac{q^{\setminus i}(\boldsymbol{f})p(y_i|f_i)}{Z_{\widetilde{q}}q(\boldsymbol{f})}\, \mathrm{d}\boldsymbol{f} \\
&= \int \widetilde{q}(\boldsymbol{f}) \log \frac{\cancel{q(\boldsymbol{f})}p(y_i|f_i)}{Z_{q^{\setminus i}}Z_{\widetilde{q}}\cancel{q(\boldsymbol{f})}t_i(f_i)}\, \mathrm{d}\boldsymbol{f} \\
&= \int \widetilde{q}(f_i) \log \frac{p(y_i|f_i)}{Z_{q^{\setminus i}}Z_{\widetilde{q}}t_i(f_i)}\, \mathrm{d}f_i \\
&= \int \widetilde{q}(f_i) \log \frac{q^{\setminus i}(f_i)p(y_i|f_i)}{Z_{q^{\setminus i}}Z_{\widetilde{q}}q^{\setminus i}(f_i)t_i(f_i)}\, \mathrm{d}f_i \\
&= \int \widetilde{q}(f_i) \log \frac{\widetilde{q}(f_i)}{q(f_i)}\, \mathrm{d}f_i \\
&= \mathrm{KL}(\widetilde{q}(f_i)\|q(f_i))
\end{aligned}
$$

F. Predictive Distributions of Poisson Regression

Given the approximate predictive distribution $f(\boldsymbol{x}_*) = \mathcal{N}(\mu_*, \sigma_*^2)$ and the relation $g(f) = f^2$, it is straightforward to derive the corresponding $g(\boldsymbol{x}_*) \sim \mathrm{Gamma}(k_*, c_*)$[1] where the shape k_* and the scale c_* are expressed as (Walder & Bishop, 2017; Zhang et al., 2020):

$$
k_* = \frac{(\mu_*^2 + \sigma_*^2)^2}{2\sigma_*^2(2\mu_*^2 + \sigma_*^2)}, \;\; c_* = \frac{2\sigma_*^2(2\mu_*^2 + \sigma_*^2)}{\mu_*^2 + \sigma_*^2}.
$$

Furthermore, the predictive distribution of the count value $y \in \mathbb{N}$ can also be derived straightforwardly:

$$
\begin{aligned}
p(y) &= \int_0^\infty p(g_*)p(y|g_*)\, \mathrm{d}g_* \\
&= \int \mathrm{Gamma}(g_*|k_*, c_*)\mathrm{Poisson}(y|g_*)\, \mathrm{d}g_* \\
&= \frac{c_*^y(c_*+1)^{-k_*-y}\Gamma(k_*+y)}{y!\Gamma(k_*)} = \mathrm{NB}(y|k_*, c_*/(1+c_*)),
\end{aligned}
$$

where $g_* = g(\boldsymbol{x}_*)$ and NB denotes the negative binomial distribution. The mode is obtained as $\lfloor c_*(k_*-1) \rfloor$ if $k_* > 1$ else 0.

G. Proof of Corollary 2.2

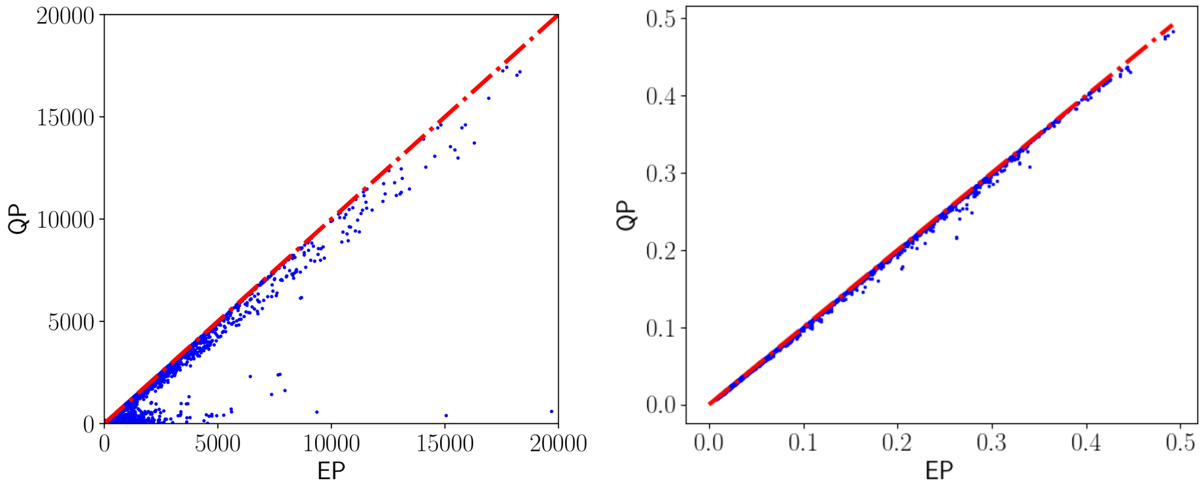
Since the site approximations of both EP and QP are Gaussian, we may analyse the predictive variances using results from the regression with Gaussian likelihood function case, namely the well known Equation (3.61) in (Rasmussen & Williams, 2005):

$$
\sigma^2(f_*) = k(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^\mathsf{T}(K + \widetilde{\Sigma})^{-1}\boldsymbol{k}_*, \tag{25}
$$

where $f_* = f(\boldsymbol{x}_*)$ is the evaluation of the latent function at $\boldsymbol{x}_*$ and $\boldsymbol{k}_* = [k(\boldsymbol{x}_*, \boldsymbol{x}_1), \cdots, k(\boldsymbol{x}_*, \boldsymbol{x}_N)]^\mathsf{T}$ is the covariance vector between the test data $\boldsymbol{x}_*$ and the training data $\{\boldsymbol{x}_i\}_{i=1}^N$, K is the prior covariance matrix and $\widetilde{\Sigma}$ is the diagonal matrix with elements of site variances $\widetilde{\sigma}_i^2$.

After updating the parameters of a site function $t_i(f_i)$, the term $(K + \widetilde{\Sigma})^{-1}$ is updated to $(K + \widetilde{\Sigma} + (\widetilde{\sigma}_{i,\text{new}}^2 - \widetilde{\sigma}_i^2)\boldsymbol{e}_i\boldsymbol{e}_i^\mathsf{T})^{-1}$ where $\widetilde{\sigma}_{i,\text{new}}$ is the site variance estimated by EP or QP and $\boldsymbol{e}_i$ is a unit vector in direction i. Using the Woodbury, Sherman

[1] $\mathrm{Gamma}(x|k, c) = \frac{1}{\Gamma(k)c^k}x^{k-1}e^{-x/c}$.



(a) Binary Classification

(b) Poisson Regression

Figure 1: A scatter plot of the predictive variances of latent functions on test data, for EP and QP. The diagonal dash line represents equivalence. We see that the predictive variance of EP is always less than or equal to that of EP.

& Morrison formula (Rasmussen & Williams, 2005, A.9), we rewrite $(K + \widetilde{\Sigma} + (\widetilde{\sigma}^2_{i,\text{new}} - \widetilde{\sigma}^2_i)\boldsymbol{e}_i\boldsymbol{e}_i^\top)^{-1}$ as

$$\begin{aligned}
&(K + \widetilde{\Sigma} + (\widetilde{\sigma}^2_{i,\text{new}} - \widetilde{\sigma}^2_i)\boldsymbol{e}_i\boldsymbol{e}_i^\top)^{-1} \\
&\equiv (A^{-1} + (\widetilde{\sigma}^2_{i,\text{new}} - \widetilde{\sigma}^2_i)\boldsymbol{e}_i\boldsymbol{e}_i^\top)^{-1} \\
&\quad = A - A\boldsymbol{e}_i[(\widetilde{\sigma}^2_{i,\text{new}} - \widetilde{\sigma}^2_i)^{-1} + \boldsymbol{e}_i^\top A\boldsymbol{e}_i]^{-1}\boldsymbol{e}_i^\top A \\
&\quad \equiv A - \boldsymbol{s}_i[(\widetilde{\sigma}^2_{i,\text{new}} - \widetilde{\sigma}^2_i)^{-1} + A_{ii}]^{-1}\boldsymbol{s}_i^\top \\
&\quad\quad = A - \frac{1}{(\widetilde{\sigma}^2_{i,\text{new}} - \widetilde{\sigma}^2_i)^{-1} + A_{ii}}\boldsymbol{s}_i\boldsymbol{s}_i^\top
\end{aligned}$$

where $A = (K + \widetilde{\Sigma})^{-1}$ and $\boldsymbol{s}_i$ is the i'th column of A. Putting the above expression into Equation (25), we have that the predictive variance is updated according to:

$$\sigma^2_{\text{new}}(f_*) = k(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^\top A\boldsymbol{k}_* + \frac{1}{(\widetilde{\sigma}^2_{i,\text{new}} - \widetilde{\sigma}^2_i)^{-1} + A_{ii}}\boldsymbol{k}_*^\top\boldsymbol{s}_i\boldsymbol{s}_i^\top\boldsymbol{k}_*.$$

In EP and QP, the first two terms on the r.h.s. of the above equation are equivalent. As the site variance provided by QP is less or equal to that by EP, *i.e.*, $\widetilde{\sigma}^2_{i,\text{QP}} \leq \widetilde{\sigma}^2_{i,\text{EP}}$, the third term on the r.h.s. for QP is less or equal to that for EP. Therefore, the predictive variance of QP is less or equal to that of EP: $\sigma^2_{\text{QP}}(f_*) \leq \sigma^2_{\text{EP}}(f_*)$.