All your loss are belong to Bayes

Christian Walder Richard Nock
Data61 & the Australian National University
first.last@data61.csiro.au

Abstract

Loss functions are a cornerstone of machine learning and the starting point of most algorithms. Statistics and Bayesian decision theory have contributed, via *properness*, to elicit over the past decades a wide set of admissible losses in supervised learning, to which most popular choices belong (logistic, square, Matsushita, *etc.*). Rather than making a potentially biased *ad hoc* choice of the loss, there has recently been a boost in efforts to *fit* the loss to the domain at hand while training the model itself. The key approaches fit a canonical link, a function which monotonically relates the closed unit interval to $\mathbb{R}$ and can provide a proper loss via integration.

In this paper, we rely on a broader view of proper *composite* losses and a recent construct from information geometry, *source* functions, whose fitting alleviates constraints faced by canonical links. We introduce a trick on squared Gaussian Processes to obtain a random process whose paths are compliant source functions with many desirable properties in the context of link estimation. Experimental results demonstrate substantial improvements over the state of the art.

1 Introduction

The loss function is a cornerstone of supervised learning. A rich literature on admissible losses has been developed from the early seventies in statistical decision theory [Sav71], and still earlier in foundational philosophical work [dF49]. The essential criterion for admissibility is *properness*, the fact that Bayes' rule is optimal for the loss at hand. Properness is *intensional*: it does not provide a particular set of functions to choose a loss from. Over the past decades, a significant body of work has focused on eliciting the set of admissible losses (see [NM20] and references therein), yet in comparison with the vivid breakthroughs on models that has flourished during the past decade in machine learning, the decades-long picture of the loss resembles a still life — more often than not, it is fixed from the start, *e.g.* by assuming the popular logistic or square loss, or by assuming a restricted parametric form [Cza97, CM00, NDF00, CR02]. More recent work has aimed to provide machine learning with greater flexibility in the loss [HT92, KS09, KKSK11, NM20] — yet these works face significant technical challenges arising from (i) the joint estimation non-parametric loss function along with the remainder of the model, and (ii) the specific part of a proper loss which is learned, called a link function, which relates class probability estimation to real-valued prediction [RW10].

In a nutshell, the present work departs from these chiefly frequentist approaches and introduces a new Bayesian standpoint to the problem. We exploit a finer characterisation of loss functions and a trick on the Gaussian Process (GP) for efficient inference while retaining guarantees on the losses. Experiments show that our approach tends to significantly beat the state of the art [KS09, KKSK11, NM20], and records better results than baselines informed with specific losses or links, which to our knowledge is a first among approaches learning a loss or link.

More specifically, we first sidestep the impediments and constraints of fitting a link by learning a *source* function, which need only be monotonic, and which allows to reconstruct a loss via a given link — yielding a proper *composite* loss [RW10]. The entire construct exploits a fine-grained information-geometric characterisation of Bregman divergences pioneered by Zhang and Amari [Ama12, Ama13, Zha04]. This is our **first contribution**. Our **second contribution** exploits a Bayesian standpoint on the problem itself: since properness does not specify *a* loss to minimise, we do not estimate nor model *a* loss based on data — instead we perform Bayesian inference on the losses themselves, and more specifically on the source function. From the losses standpoint, our technique brings the formal advantage of a fine-tuned control of the key parameters of a loss which, in addition to properness, control consistency, robustness, calibration and rates.

We perform Bayesian inference within this class of losses by a new and simple trick addressing the requirement of priors over functions which are simultaneously non-parametric, differentiable and guaranteed monotonic. We introduce for that purpose the Integrated Squared Gaussian Process (ISGP), which extends by integration (which yields monotonicity) the *squared* Gaussian Process — which has itself seen a recent burst of attention as a model for merely *non-negative* functions [ST03, MM06, LGOR15, WB17, FTS17, LH19]. The ISGP contrasts with previous approaches to learning monotonic functions which are either non-probabilistic [ABE$^+$55, YW09, APS10, LR14, Bac18, Lim18], resort to a discretisation [HHLK19, KEC19, UKEC19], or are only approximately monotonic due to the use of only a finite set of monotonicity promoting constraints [RV10, GBCC15, SPV16, LRG$^+$17, YLR$^+$18, ANARL19].

▷ **Organisation.** In Section 2 we introduce properness and *source functions*. We define the ISGP model in Section 3, and provide relevant Bayesian inference procedures in Section 4. Numerical experiments are provided in Section 5 before concluding with Section 6. Technical details and an extensive suite of illustrations are provided in the supplementary appendices.

2 Definitions and key properties for losses

Our notations follow [NM20, RW10]. In the context of statistical decision theory as discussed more than seventy years ago [dF49] and later theorised [Sav71], the key properties of a supervised loss function start with ensuring that Bayes' rule is optimal for the loss, a property known as *properness*.

▷ **Proper (composite) losses**: let $\mathcal{Y} \doteq \{-1, 1\}$ be a set of labels. A loss for binary class probability estimation [BSS05] is a function $\ell : \mathcal{Y} \times [0, 1] \to \overline{\mathbb{R}}$ (closure of $\mathbb{R}$). Its *(conditional) Bayes risk* function is the best achievable loss when labels are drawn with a particular positive

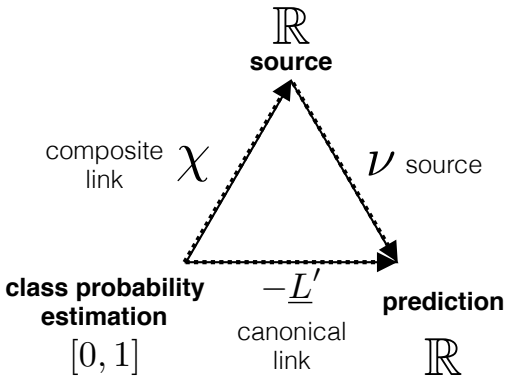


Figure 1: Correspondence between source, canonical link and composite link. Domains and names are given in the context of supervised learning.

base-rate,

$$\underline{L}(\pi) \doteq \inf_u \mathbb{E}_{\mathsf{Y}\sim\text{Bernoulli}(\pi)}\ell(\mathsf{Y}, u),$$

where $\Pr[\mathsf{Y}=1] = \pi$. A loss for class probability estimation ℓ is ***proper*** iff Bayes prediction locally achieves the minimum everywhere: $\underline{L}(\pi) = \mathbb{E}_{\mathsf{Y}}\ell(\mathsf{Y}, \pi), \forall \pi \in [0,1]$, and strictly proper if Bayes is the unique minimum. Fitting a classifier's prediction $h(\boldsymbol{z}) \in \mathbb{R}$ ($\boldsymbol{z}$ being an observation) into some $u \in [0,1]$ is done via an invertible *link function* $\chi : [0,1] \to \mathbb{R}$ connecting real valued prediction and class probability estimation. A proper loss augmented with a link, $\ell(y, \chi^{-1}(x))$ with $x \in \mathbb{R}$ [RW10] is called *proper composite*. There exists a particular link uniquely defined (up to multiplication or addition by a scalar [BSS05]) for any proper loss, the *canonical link*, as: $\chi \doteq -\underline{L}'$ [RW10, Section 6.1]: for example, the logistic loss yields the canonical link $\chi(u) = (-\underline{L}')(u) = \log(u/(1-u))$, with inverse the well-known sigmoid $\chi^{-1}(x) = (-\underline{L}')^{-1}(x) = (1+\exp(-x))^{-1}$. A proper loss augmented with its canonical link instead of any composite link is called *proper canonical*. We remind that the Bayes risk of a proper loss is concave [RW10]. There exists a particularly useful characterisation of proper composite losses [NM20, Theorem 1]: a loss is proper composite iff

$$\ell(y, \chi^{-1}(x)) = D_{-\underline{L}}\left(y^* \| \chi^{-1}(x)\right) = D_{(-\underline{L})^\star}\left(-\underline{L}' \circ \chi^{-1}(x) \| -\underline{L}'(y^*)\right),$$

where $y^* \doteq (y+1)/2 \in \{0,1\}$, $D_{-\underline{L}}$ is the Bregman divergence with generator $-\underline{L}$ and $(-\underline{L})^\star$ is the Legendre conjugate of $-\underline{L}$. Let G be convex differentiable. Then we have $D_G(x\|x') \doteq G(x) - G(x') - (x-x')G'(x')$ [Bre67] and $G^\star(x) \doteq \sup_{x' \in \text{dom}(G)}\{xx' - G(x')\}$ [BV04]. We assume in this paper that losses are twice differentiable for readability (differentiability assumption(s) can be alleviated [RW10, Footnote 6]) and strictly proper for the canonical link to be invertible, both properties being satisfied by popular choices (log, square, Matsushita losses, *etc.*).

▷ **Margin losses**: an important class of losses for real-valued prediction are functions of the kind $G(yh)$ where $G : \mathbb{R} \to \mathbb{R}$, $y \in \{-1, 1\}$ and yh is called a margin in [RW10]. There exists a rich theory on the connections between margin losses and proper losses: If $G(x) \doteq (-\underline{L})^\star(-x)$ for a proper loss ℓ whose Bayes risk is symmetric around $1/2$ (equivalent to having no

class-conditional misclassification costs), then minimising the margin loss is equivalent to minimising the proper canonical loss [NN08, Lemma 1]. In the more general case, given a couple (G, χ), there exists a proper loss ℓ such that $G(y\chi(u)) = \ell(y, u)$ $(\forall y, u)$ iff [RW10, Corollary 14]:

$$\chi^{-1}(x) = \frac{G'(-x)}{G'(-x) + G'(x)}, \forall x.$$

When this holds, we say that the couple (G, χ) is *representable* as a proper loss. When properness is too tight a constraint, one can rely on *classification calibration* [BJM06], which relaxes the optimality condition on just predicting labels. Such a notion is particularly interesting for margin losses. We assume in this paper that margin losses are twice differentiable, a property satisfied by most popular choices (logistic, square, Matsushita, *etc.*), but notably not satisfied by the hinge loss.

▷ **Key quantities for a good loss**: there are two crucial quantities that indicate the capability of a function to be a good choice for a loss, its Lipschitz constant and weight function. For any function $G : \mathbb{R} \to \mathbb{R}$, the Lipschitz constant of G is $\mathrm{lip}_G \doteq \sup_{x,x'} |G(x) - G(x')|/|x - x'|$ and weight function $w_G(x) \doteq G''(x)$. Controlling the Lipschitz constant of a loss is important for robustness [CBG+17] and mandatory for consistency [Tel13]. The weight function, on the other hand, defines properness [Sch89] but also controls optimisation rates [KM99, NN08] and defines the geometry of the loss [AN00, Section 3]. More than the desirable properness, in fitting or tuning a loss, one ideally needs to make sure that levers are easily accessible to control these two quantities.

▷ (u, v)**-geometric structure**: a proper loss defines a canonical link. A proper composite loss therefore makes use of *two* link functions. Apart from flexibility, there is an information geometric justification to such a more general formulation, linked to a finer characterisation of the information geometry of Bregman divergences introduced in [Ama12, Ama13, Zha04] and more recently analysed, the (u, v)-geometric structure [NNA16]. u, v are two differentiable invertible functions and the gradient of the generator of the divergence is given by $u \circ v^{-1}$. This way, the two dual forms of a Bregman divergence (in (2)) can be represented using the same coordinate system known as the *source*, which specifies the dual coordinate system and Riemannian metric induced by the divergence. We consider the (ν, χ)-geometric structure of the divergence $D_{-\underline{L}}$, which therefore gives for ν:

$$\nu = -\underline{L}' \circ \chi^{-1},$$

and yields a local composite estimate of Bayes' rule for a corresponding $x \in \mathbb{R}$ as in [NM20, Eq. 4]:

$$\hat{p}(y = 1|x; \nu, \underline{L}) \doteq \chi^{-1}(x) = (-\underline{L}')^{-1} \circ \nu(x).$$

Figure 1 depicts ν, which we call "source" as well. The properties of ν in (2) are summarised below.

Definition 1. *A* ***source*** *(function) ν is a differentiable, strictly increasing function.*

Any loss completed with a source gives a proper composite loss. There are two expressions of interest for a loss involving source ν, which follow from (2) and margin loss $G(.)$, with $y \in \{0,1\}$ and $x \in \mathbb{R}, u \in [0,1]$ (with the correspondence $x \doteq \chi(u)$ and $y^* \doteq (y+1)/2$):

$$\ell_\nu(y,u) \doteq D_{(-\underline{L})^\star}\left(\nu(x) \| -\underline{L}'(y^*)\right)$$
$$G_\nu(y\chi(u)) \doteq G(y \cdot (\nu^{-1} \circ (-\underline{L}'))(u)).$$

When it is clear from the context, we shall remove the index "ν" for readability. Instead of learning the canonical link of a loss, which poses significant challenges due to its closed domain [KKSK11, NM20], we choose to infer the source given a loss. As we now see, this can be done efficiently using Gaussian Process inference and with strong guarantees on the losses involved.

3 The Integrated Squared Gaussian Process

▷ **Model Definition.** An Integrated Squared Gaussian Processes (ISGP) is a non-parametric random process whose sample paths turn out to be source functions.

Definition 2. *An ISGP is defined by the following model:*

$$\nu_0 \sim \mathcal{N}(\mu, \gamma^{-1}); \qquad f \sim \mathrm{GP}(k(\cdot,\cdot)) \quad \Leftrightarrow \quad \begin{cases} f(\cdot) = \boldsymbol{w}^\top \boldsymbol{\varphi}(\cdot) \\ \boldsymbol{w} \sim \mathcal{N}(0, \Lambda) \end{cases}$$

and $\nu(x) \doteq \nu_0 + \int_0^x f^2(z)\,\mathrm{d}z$.

Here, $\boldsymbol{\varphi}(x) = (\varphi_1(x), \varphi_2(x), \ldots, \varphi_M(x))^\top$ represents M basis functions, $\Lambda \in \mathbb{R}^{M\times M}$ is a diagonal matrix with elements $\lambda_m > 0$, and $\mathrm{GP}(k(\cdot,\cdot))$ is the zero-mean Gaussian process with kernel $k : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$. Hence the above equivalence (denoted $\Leftrightarrow$) implies the Mercer expansion

$$k(x,z) = \boldsymbol{\varphi}(x)^\top \Lambda \boldsymbol{\varphi}(z).$$

The ISGP is Gaussian in the parameters $\boldsymbol{w}$ and ν_0, and admits a simple form for ν:

$$\nu(x) = \nu_0 + \int_0^x f^2(z)\,\mathrm{d}z = \nu_0 + \int_0^x \left(\boldsymbol{w}^\top \boldsymbol{\varphi}(z)\right)^2 \mathrm{d}z = \nu_0 + \boldsymbol{w}^\top \boldsymbol{\psi}(x) \boldsymbol{w},$$

with the positive semi-definite matrix (to be discussed later),

$$\boldsymbol{\psi}(x) \doteq \int_0^x \boldsymbol{\varphi}(z)\boldsymbol{\varphi}(z)^\top \,\mathrm{d}z \in \mathbb{R}^{M\times M}.$$

It is our choice of *squaring* of f — previously exploited to model *non-negativity* [ST03, MM06, LGOR15, WB17, FTS17, LH19] — which leads to this simple form for the *monotonic* ν.

▷ **Sanity checks of the ISGP prior for loss inference.** We now formally analyse the goodness of fit of an ISGP in the context of loss inference. The following Lemma is immediate and links the ISGP to inference on functions expressed as Bregman divergences, and therefore to proper losses (Section 2).

Lemma 1. *$\nu(x)$ as in Definition 2 is a source function with probability one.*

Remark 3.1. Bregman divergences are also the analytical form of losses in other areas of machine learning (*e.g.* unsupervised learning [BGW05]), so ISGPs may *via* Lemma 1 be of broader interest.

The following Theorem safeguards inference on loss functions with an ISGP.

Theorem 2. *For any convex G and any ISGP, the following holds with probability one: (i) if G is decreasing, there exists a canonical link $-\underline{L}'$ such that the couple (G, χ) where χ is obtained from (2) is representable as a proper loss; (ii) if G is classification calibrated, the margin loss $G_\nu(x) \doteq G \circ \nu^{-1}(x)$ is classification calibrated.*

(Proof in Appendix C.) To summarise, the whole support of an ISGP is fully representable as proper losses, but the proof of point (i) is not necessarily constructive as it involves a technical invertibility condition. On the other hand, part (ii) relaxes properness for classification calibration but the invertibility condition is weakened to that of ν for the margin loss involved.

Theorem 2 begs for a converse on the condition(s) on the ISGP to ensure that *any* proper composite loss can be represented in its support. Fortunately, a sufficient condition is already known [MXZ06].

Theorem 3. *Suppose kernel k in Definition 2 is universal. Then for any proper loss ℓ and any composite link χ, there exists ν as in Definition 2 such that the proper composite loss $\ell(y, \chi^{-1}(x))$ can be represented as in (2): $\ell(y, \chi^{-1}(x)) = \ell_\nu(y, u), \forall u \in [0, 1], y \in \{-1, 1\}$ and $x \doteq \chi(u)$.*

To avoid a paper laden paper with notation, we refer to [MXZ06] for the exact definition of universality which, informally, postulates that the span of the kernel can approximate any function to arbitrary precision. Interestingly, a substantial number of kernels are universal, including some of particular interest in our setting (see below). We now investigate a desirable property of the first-order moment of a loss computed from an ISGP. We say that the ISGP is *unbiased* iff $\mathbb{E}_\nu [\nu(x)] = x, \forall x$.

Theorem 4. *Letting $\{(x, y)\} \subset \mathbb{R} \times \mathcal{Y}$ be a sample of labelled training values. For any unbiased ISGP and any proper canonical loss ℓ, the proper composite loss ℓ formed by ℓ and composite link $\chi \doteq \nu^{-1} \circ -\underline{L}'$ (2) satisfies: $\mathbb{E}_{(x,y)} [\ell(y, (-\underline{L}')^{-1}(x))] \leq \mathbb{E}_{(x,y),\nu} [\ell_\nu(y, \chi^{-1}(x))]$.*

(Proof in Appendix C.) In the context of Bayesian inference this means that the prior uncertainty in ν induces a prior on losses whose expected loss upper bounds that of the canonical link. We exploit this property to *initialise* our Bayesian inference scheme with fixed canonical link (see Section 4.2). Of course, this property follows from Theorem 4 *if* the ISGP is unbiased, which we now show is trivial to guarantee. We say that a kernel k is translation invariant iff $k(x, x') = k(0, x - x'), \forall x, x'$.

Theorem 5. *For any translation invariant k, the ISGP ν satisfies $\mathbb{E}_\nu [\nu(x)] = \mu + k(0, 0) \cdot x$.*

Proof. By linearity of expectation this mean function equals $\mathbb{E}[\nu_0] = \mu$ plus the simple term

$$\mathbb{E}_{f\sim\mathrm{GP}(k(\cdot,\cdot))}\left[\int_0^x f^2(z)\,\mathrm{d}z\right] = \int_0^x \mathbb{E}\left[f^2(z)\right]\mathrm{d}z = \int_0^x k(z,z)\,\mathrm{d}z = x\cdot k(0,0),$$

yielding the statement of the Theorem. □

Hence, if $\mu = 0$ and $k(0,0) = 1$ in Definition 2, the ISGP is unbiased. Interestingly, some translation invariant kernels are universal [MXZ06, Section 4]. We now dig into how an ISGP allows to control the properties of supervised losses that govern robustness and statistical consistency, among others [CBG+17, Tel13]. Denote $\|\boldsymbol{\varphi}\|^2_{\max} = \sup_x \|\boldsymbol{\varphi}(x)\|_2^2$, which is simple to upper-bound for the $\boldsymbol{\varphi}$ we will adopt later (see (4.3)).

Lemma 6. *For ISGP ν the Lipschitz constant of ℓ_ν in* (2) *satisfies* $\mathrm{lip}_{\ell_\nu} \leq \|\boldsymbol{\varphi}\|^2_{\max}\cdot\|\boldsymbol{w}\|_2^2\cdot\mathrm{lip}_\ell$.

Proof. By Cauchy-Schwartz $\nu'(x) = f^2(x) = (\boldsymbol{w}^\top\boldsymbol{\varphi}(x))^2 \leq \sup_x(\boldsymbol{w}^\top\boldsymbol{\varphi}(x))^2 \leq \sup_x\|\boldsymbol{\varphi}(x)\|_2^2\cdot\|\boldsymbol{w}\|_2^2 \leq \|\boldsymbol{\varphi}\|^2_{\max}\cdot\|\boldsymbol{w}\|_2^2$. The Lemma can then follows from the chain rule on ℓ_ν in (2). □

Similarly, we have $\chi = \nu^{-1}\circ(-\underline{L}')$ from (2) and so $\chi'(u) = w_\ell(u)\cdot(\boldsymbol{w}^\top(\boldsymbol{\varphi}\boldsymbol{\varphi}^\top)(\chi(u))\boldsymbol{w})^{-1}$, yielding this time a Lipschitz constant for G_ν of (2) which is proportional to the weight of ℓ and inversely proportional to f^2. This shows that the prior's uncertainty still allows for a probabilistic handle on the Lipschitz constant, Λ — the eigenspectrum in the Mercer expansion.

4 Inference with Integrated Squared Gaussian Processes

In Section 4.1 we give an approximate inference method for the ISGP with simple i.i.d. likelihood functions — this is analogous to the basic GP regression and classification in *e.g.* [RW05]. The subsequent Section 4.2 builds on this to provide an inference method for proper losses for the generalised linear model — this is a Bayesian analog of the loss fitting methods in [KKSK11, NM20].

4.1 Univariate Case

It is straight-forward to construct a Laplace approximation to the posterior in the parameters $\Gamma = \{\boldsymbol{w}, \nu_0\}$, provided that *e.g.* the likelihood function for the data $\mathcal{D} = \{x_n, y_n\}_{n=1}^N$ factorises in the usual way as $\log p(\mathcal{D}|\nu,\Theta) = \sum_{n=1}^N \log p(y_n|\nu(x_n),\Theta)$. Here Θ represents the hyper-parameters — which are γ, μ, any parameters of the kernel $k(\cdot,\cdot)$, and any parameters of the likelihood function. We use automatic differentiation and numerical optimisation to obtain the posterior mode $\widehat{\Gamma} = \mathrm{argmax}_\Gamma \log p(\mathcal{D},\Gamma)$. We then use automatic differentiation once more to compute the Hessian

$$H = \left.\frac{\partial^2}{\partial\Gamma\partial\Gamma^\top}\right|_{\Gamma=\widehat{\Gamma}} - \log p(\mathcal{D},\Gamma).$$

Letting $\widehat{\Sigma}_\Gamma = H^{-1}$, the approximate posterior is then $p(\Gamma|\mathcal{D}) \approx \mathcal{N}(\Gamma|\widehat{\Gamma},\widehat{\Sigma}_\Gamma) \doteq q(\Gamma|\mathcal{D})$.
▷ **Predictive distribution:** posterior samples of the monotonic function ν are obtained

from samples $\Gamma^{(s)} = \{\boldsymbol{w}^{(s)}, \nu_0^{(s)}\}$ from q via the deterministic relation (3). For the predictive mean function, we exploit the property that the predictive distribution is the sum of the Gaussian ν_0 plus the well studied quadratic function of a multivariate normal. If we let $\widehat{\nu_0}$ be the element of $\widehat{\Gamma}$ corresponding to ν_0, and let $\widehat{\Sigma}_{\boldsymbol{w}}$ be the sub-matrix of $\widehat{\Sigma}_\Gamma$ corresponding to $\boldsymbol{w}$, *etc.*, then the following holds.

Lemma 7. *Using notations just defined, the mean function may be written as*

$$\mathbb{E}_{q(\Gamma|\mathcal{D})}\left[\nu(x)\right] = \widehat{\nu_0} + \operatorname{tr}(\boldsymbol{\psi}(x)\widehat{\Sigma}_{\boldsymbol{w}}) + \widehat{\boldsymbol{w}}^\top \widehat{\Sigma}_{\boldsymbol{w}} \widehat{\boldsymbol{w}}.$$

Proof. We have $\mathbb{E}_{q(\Gamma|\mathcal{D})}\left[\nu(x)\right] = \widehat{\nu_0} + \mathbb{E}_{\mathrm{N}(\boldsymbol{w}|\widehat{\boldsymbol{w}},\widehat{\Sigma}_{\boldsymbol{w}})}\left[\boldsymbol{w}^\top \boldsymbol{\psi}(x) \boldsymbol{w}\right]$ and, for $\boldsymbol{w} \sim \mathrm{N}(\boldsymbol{w}|\widehat{\boldsymbol{w}}, \widehat{\Sigma}_{\boldsymbol{w}})$,

$$\mathbb{E}\left[\boldsymbol{w}^\top \boldsymbol{\psi}(x) \boldsymbol{w}\right] = \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{\psi}(x) \boldsymbol{w}\boldsymbol{w}^\top\right)\right] = \operatorname{tr}\left(\boldsymbol{\psi}(x) \mathbb{E}\left[\boldsymbol{w}\boldsymbol{w}^\top\right]\right) = \operatorname{tr}\left(\boldsymbol{\psi}(x)(\widehat{\Sigma}_{\boldsymbol{w}} + \widehat{\boldsymbol{w}}\widehat{\boldsymbol{w}}^\top)\right),$$

which leads to the required result. □

This mean function differs from that of a simple GP, which is linear w.r.t. $\widehat{\boldsymbol{w}}$. Closed form expressions for the higher moments of $\boldsymbol{w}^\top \boldsymbol{\psi}(x) \boldsymbol{w}$ are provided in [MP92, §3.2b: *Moments of a Quadratic Form*].

We can easily replicate the use of Jensen's inequality in the proof of Theorem 4 to obtain the following guarantee for the posterior mean function. We recall that $\underline{L}$ is the Bayes risk of a proper loss ℓ and $\chi^{-1} \doteq (-\underline{L}')^{-1} \circ \nu$ is the construct of the (inverse) composite link using source ν (of (2)).

Theorem 8. *For any sample of labelled data points $\{(x, y)\} \subset \mathbb{R} \times \mathcal{Y}$, any proper loss ℓ and source ν sampled from approximate posterior $q(\Gamma|\mathcal{D})$, we have:*

$$\mathbb{E}_{(x,y),\nu}\left[\ell_\nu(y, \chi^{-1}(x))\right] \leq \mathbb{E}_{(x,y)}\left[D_{(-\underline{L})^\star}\left(\widehat{\nu_0} + \operatorname{tr}(\boldsymbol{\psi}(x)\widehat{\Sigma}_{\boldsymbol{w}}) + \widehat{\boldsymbol{w}}^\top \widehat{\Sigma}_{\boldsymbol{w}} \widehat{\boldsymbol{w}} \| - \underline{L}'(y^*)\right)\right].$$

This clarifies the importance of Bayesian modelling of ν, as the resulting expected loss is merely upper bounded by that for the fixed mean function. Alternatively the result suggests a cheap approximation to marginalising (2) w.r.t. ν for prediction, namely putting the mean function for ν into that equation.

▷ **Marginal likelihood approximation and optimisation:** we follow [SF06] and use automatic differentiation to achieve the same result as — while avoiding the manual gradient calculations of — the usual approach in the GP literature [RW05, §3.4]. See Section D for further details on the marginal likelihood, computational complexity, and likelihood functions for the univariate case.

4.2 Using the ISGP to fit Bayes estimates

At this stage, inferring the loss in a machine learning model boils down to simply placing an ISGP prior on the source function and doing Bayesian inference. In accordance with part i) of Theorem 2, this induces a posterior over sources which in turn implies a distribution over proper composite losses. We now present an effective inference procedure for the generalised linear model, which has been the standard model for closely related isotonic regression

approaches to the problem [KS09, KKSK11, NM20]. We first choose a proper loss' link $-\underline{L}'$ to get the proper composite estimate (2). Here we consider the log loss and therefore the inverse sigmoid for $-\underline{L}'$.

▷ **Model:** to summarise, given a dataset $\mathcal{D} = \{\boldsymbol{z}_n, y_n\}_{n=1,\dots,N} \subset \mathbb{R}^D \times \mathcal{Y}$, we model $\hat{p}(x) \doteq (-\underline{L}')^{-1} \circ \nu(x)$ from (2) with $(-\underline{L}')^{-1}(x) = (1 + \exp(-x))^{-1}$, to obtain the classification model

$$y_n | \boldsymbol{z}_n \sim \text{Bernoulli}(\hat{p}(x_n)) \qquad ; \qquad x_n = \boldsymbol{\beta}^\top \boldsymbol{z}_n,$$

where $\nu \sim \mathcal{F}$ and $\mathcal{F}$ is a prior over functions such as the GP or ISGP. We leave the prior on the linear model weight vector $\boldsymbol{\beta} \in \mathbb{R}^D$ unspecified as we perform maximum (marginal) likelihood on that parameter. This model generalises logistic regression, which we recover for $\nu(x) = x$.

▷ **Inference:** we use Expectation Maximisation (EM) to perform maximum likelihood in $\boldsymbol{\beta}$, marginal of ν. In accordance with Theorem 4 we initialise $\boldsymbol{\beta}^{(\text{old})}$ using traditional fixed link $(-\underline{L}')^{-1}$. In the *E-step*, $x_n = \boldsymbol{\beta}^{(\text{old})\top} \boldsymbol{z}_n$ are fixed and we use the Laplace approximate inference scheme of Section 4.1 to compute the posterior in ν (or equivalently $\Gamma = \{\boldsymbol{w}, \nu\}$), taking the likelihood function to be that of $y_n | x_n$ from (4.2), above. Let that approximate posterior — previously denoted by $q(\Gamma|\mathcal{D})$ — be denoted here by $q(\nu|\boldsymbol{\beta}^{(\text{old})})$. The *M-step* then updates $\boldsymbol{\beta}$ by (letting $\boldsymbol{y} = \{y_1, y_2, \dots, y_N\}$ and $\boldsymbol{x} = \{x_1, x_2, \dots, x_N\}$),

$$\boldsymbol{\beta}^{(\text{new})} = \operatorname*{argmax}_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(\text{old})}) \qquad ; \qquad Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(\text{old})}) = \mathbb{E}_{q(\nu|\boldsymbol{\beta}^{(\text{old})})} \left[\log p(\boldsymbol{y}|\boldsymbol{x}, \nu)\right].$$

Although the expectation for Q is analytically intractable, the univariate setting suggests that Monte Carlo will be effective. We draw S samples $\nu^{(\text{s})}$, $s = 1, \dots, S$ from $q(\nu|\boldsymbol{\beta}^{(\text{old})})$, and approximate

$$\mathbb{E}_{q(\nu|\boldsymbol{\beta}^{(\text{old})})} \left[\log p(\boldsymbol{y}|\boldsymbol{x}, \nu)\right] \approx \frac{1}{S} \sum_{s=1}^{S} \sum_{n=1}^{N} \log \text{Bernoulli}(y_n | (-\underline{L}')^{-1} \circ \nu^{(s)}(\boldsymbol{\beta}^\top \boldsymbol{z}_n)),$$

where $\nu^{(s)}$ is given by (3). The above expression may be automatically differentiated and optimised using standard numerical tools. To achieve this efficiently under the ISGP prior for $\mathcal{F}$, we implement a custom derivative function based on the relation

$$\nu'(x) = f^2(x) = \left(\boldsymbol{w}^\top \boldsymbol{\varphi}(x)\right)^2,$$

rather than requiring the automatic differentiation library to differentiate through $\boldsymbol{\psi}$ in (3).

4.3 The trigonometric kernel

We introduce a novel kernel designed to lend itself to the ISGP. See Appendix F for the requirements this entails, and a discussion of a tempting yet restricting alternative (the Nyström method). Our main insight is that we need not fix the kernel and derive the corresponding Mercer expansion — even if such an expansion is available, the integrals for $\boldsymbol{\psi}(x)$ are generally intractable. Instead we recall that $k(x, z) = \boldsymbol{\varphi}(x)^\top \Lambda \boldsymbol{\varphi}(z) = \sum_{m=1}^{M} \lambda_m \varphi_m(x) \varphi_m(z)$,

and we let the $(\lambda_m, \varphi_m(x))$ pairs (of which there are an even number, M) be given by the following union of two sets of $M/2$ pairs,

$$\left\{\big(b/a^m,\, \cos\left(\pi m c x\right)\big)\right\}_{m=1}^{M/2} \;\bigcup\; \left\{\big(b/a^m,\, \sin\left(\pi m c x\right)\big)\right\}_{m=1}^{M/2},$$

on the domain $x \in [-1/c, +1/c]$, where $a > 1$ and $b > 0$ are input and output length-scale hyper-parameters and $M \approx 100$ is chosen large enough to provide sufficient flexibility. This is related to the construction in [WB17], but has the advantage of admitting closed form expressions for k. It is easy to show that the kernel is translation invariant and that the prior variance — which is especially relevant here due to Theorem 5 — is given by

$$k(x, x) = k(0, 0) = b\,(1 - a^{-M/2})/(a - 1).$$

We give expressions for $k(x, z)$ and $\boldsymbol{\psi}(x)$ in Appendix E, and an illustration in Figure 4.

5 Experiments

We provide illustrative examples and quantitative comparisons of the ISGP prior in univariate regression/classification, and inferring proper losses. We fixed $M = 64$ in (4.3). Classification problems used $(-\underline{L}')^{-1} = \sigma$ with $\sigma(x) = 1/(1 + \exp(-x))$. Optimisation was performed with `L-BFGS` [BLNZ95]. Gradients were computed with automatic differentiation (with the notable exception of (4.2)) using `PyTorch` [PGM+19]. The experiments took roughly two CPU months.

▷ **Univariate Regression Toy.** As depicted in Figure 3, we combined the Gaussian likelihood function (D.1) with both the GP and ISGP function priors for ν. We investigate this setup further in the supplementary Figure 5 and Figure 6, which also depict the latent f and f^2, which are related by $\nu = f^2$. Finally, Figure 7 visualises the marginal likelihood surface near the optimal point found by the method of Section D. See the captions for details.

▷ **Univariate Classification Toy.** We illustrate the mechanics of the classification model with ISGP prior and the sigmoid-Bernoulli likelihood function of (D.1). We constructed a univariate binary classification problem by composing the sigmoid with a piece-wise linear function, which allows us to compare both the inferred ν and the inferred (inverse) canonical link $\chi^{-1} = (-\underline{L}')^{-1} \circ \nu$ to their respective ground truths — see Figure 8 and the caption therein for more details.

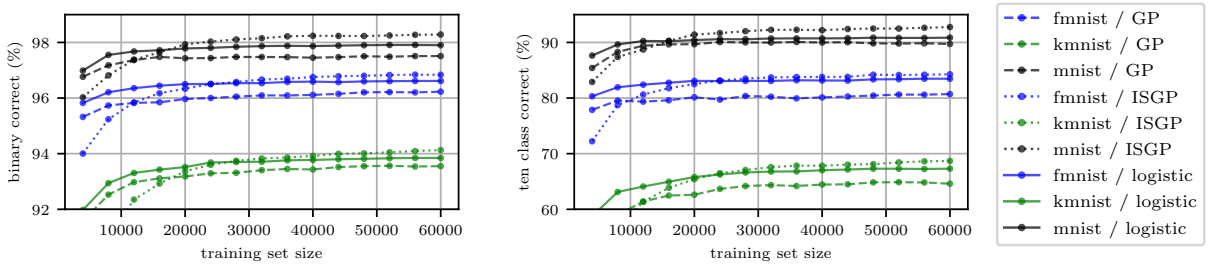
▷ **Real-world Univariate Regression.** We compared our ISGP based regression model to *a)* GP regression with the same kernel of Section 4.3, *b)* the widely used pool adjacent violators algorithm (PAVA) [Kru64] for isotonic regression, and *c)* linear least squares. For the ISGP and GP models we used maximum marginal likelihood hyper parameters. The task was to regress each of the four real-valued features of the `auto-mpg` dataset [DG17] onto the target (car efficiency in miles per gallon). We partitioned the dataset into five splits. For the `Small` problems, we trained on one split and tested on the remaining four ; for the `Large` problems we did the converse. We repeated this all $C_1^5 = C_4^5 = 5$ ways and averaged the results to obtain Table 1. As expected the Bayesian GP and ISGP models — which have the advantage of making stronger regularity assumptions — perform relatively better with less data (the `Small` case). Our ISGP is in turn superior to the GP in that case, in line with the

Table 1: Mean test set negative log likelihoods for various isotonic regression methods. See text for details.

Train	Input	GP	ISGP	PAVA	Linear
Large	accel.	2080.9	2083.4	**2055.0**	2105.9
	displ.	788.1	780.6	**760.7**	907.2
	power	763.4	763.9	**755.5**	1002.2
	weight	**735.5**	750.7	769.9	788.4
Small	accel.	**2173.3**	2216.8	2212.9	2199.6
	displ.	871.0	**849.4**	871.4	950.9
	power	827.3	**811.8**	826.4	1033.5
	weight	777.2	**770.7**	835.8	815.8

Table 2: Test AUC for generalised linear models with various link methods (ordering in decreasing average). See text for details.

	mnist	fmnist
ISGP-Linkgistic	99.9 %	99.2 %
GP-Linkgistic	99.9 %	99.1 %
Logistic regression	99.9 %	98.5%
GLMTron	99.6%	98.1%
BregmanTron	99.7%	97.9%
$\text{BregmanTron}_{\text{label}}$	99.6%	97.7%
$\text{BregmanTron}_{\text{approx}}$	99.3%	94.6%
SlIsotron	94.6%	90.7%



(a) binary classification accuracy

(b) multiclass classification accuracy

Figure 2: Test performance *v.s.* training set size for the *MNIST*, *Kuzushiji-MNIST* and *Fashion-MNIST* datasets. We compare logistic regression (logistic), GP-Linkgistic (GP) and ISGP-Linkgistic (ISGP). On the left we show the mean *one vs the rest* classification accuracy, and on the right we show the ten class classification accuracy obtained by combining the *one vs the rest* models.

appropriateness of the monotonicity assumption which it captures.

▷ **Real-world Problems of Learning the Loss.** We bench-marked our loss inference algorithm of Section 4.2. We used both the GP and ISGP function priors, and we refer to the resulting algorithms as GP-Linkgistic and ISGP-Linkgistic, respectively — only the latter of which gives rise to guaranteed *proper* losses. To ease the computational burden, we fixed the hyper-parameters throughout. We chose set $\mu = 0$ and $\gamma = 0.01$. We set the length-scale parameter $a = 1.2$ and chose b such that $k(0, 0) = 1$ (using (4.3)) — by (3) this makes $\nu(x) = x$ the most likely source *a priori*, to roughly bias towards traditional logistic regression. For the GP we tuned b for good performance on the test set (for the full 60K training examples), to give the GP an unfair advantage over the ISGP, although this advantage is small as GP-Linkgistic is rather insensitive to the choice of b.

Table 2 compares the SlIsotron and BregmanTron algorithms from [KKSK11] and [NM20], respectively, along with the other baselines from the latter work. Further details on the experimental setup are provided in [NM20]. In contrast with the SlIsotron and BregmanTron algorithms, our models successfully match or outperform the logistic regression

baseline. Moreover, the monotonic ISGP-LINKGISTIC slightly outperforms GP-LINKGISTIC, and as far as we know records the first result beating logistic regression on this problem, by a reasonable margin on `fmnist` [NM20].

We further bench-marked ISGP-LINKGISTIC against GP-LINKGISTIC and logistic regression (as the latter was the strongest practical algorithm in the experiments of [NM20]) on a broader set of tasks, namely the three MNIST-like datasets of [LC10, XRV17, CBIK$^+$18]. We found that ISGP-LINKGISTIC dominates on all three datasets, as the training set size increases — see Figure 2 and the caption therein for more details. Figure 9 depicts an example of the learned (inverse) link functions.

6 Conclusion

We have introduced a Bayesian approach to inferring a posterior distribution over loss functions for supervised learning that complies with the Bayesian notion of properness. Our contribution thereby advances the seminal work of [KKSK11] and the more recent [NM20] in terms of modelling flexibility (which we inherit from the Bayesian approach) and — as a direct consequence — practical effectiveness as evidenced by our state of the art performance. Our model is both highly general, and yet capable of out-performing even the most classic baseline, the logistic loss for binary classification. As such, this represents an interesting step toward more flexible modelling in a wider machine learning context, which typically works with a loss function which is prescribed *a priori*.
Since the tricks we use essentially rely on the loss being expressible as a Bregman divergence and since Bregman divergences are also a principled distortion measure for *un*supervised learning — such as in the context of the popular k-means and EM algorithms — an interesting avenue for future work is to investigate the potential of our approach for unsupervised learning.

References

[ABE$^+$55] Miriam Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, (4), 1955.

[Ama12] S.-I. Amari. New developments of information geometry (17): Tsallis q-entropy, escort geometry, conformal geometry. In *Mathematical Sciences (suurikagaku)*, number 592, pages 73–82. Science Company, October 2012. in japanese.

[Ama13] S.-I. Amari. New developments of information geometry (26): Information geometry of convex programming and game theory. In *Mathematical Sciences (suurikagaku)*, number 605, pages 65–74. Science Company, November 2013. in japanese.

[AN00] S.-I. Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000.

[ANARL19] Clement Abi Nader, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. Monotonic Gaussian Process for Spatio-Temporal Disease Progression Modeling in Brain Imaging Data. *NeuroImage*, 2019.

[APS10] Pankaj K. Agarwal, Jeff M. Phillips, and Bardia Sadri. Lipschitz unimodal and isotonic regression on paths and trees. In *Proc. of the 9th Latin American Symposium on Theoretical Informatics*, pages 384–396, 2010.

[Bac18] Francis Bach. Efficient algorithms for non-convex isotonic regression through submodular optimization. In *Advances in Neural Information Processing Systems 31*. 2018.

[BGW05] A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a bregman predictor. *IEEE Trans. IT*, 51:2664–2669, 2005.

[BJM06] P. Bartlett, M. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *J. of the Am. Stat. Assoc.*, 101:138–156, 2006.

[BLNZ95] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computation*, 1995.

[Bre67] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comp. Math. and Math. Phys.*, 7:200–217, 1967.

[BSS05] A. Buja, W. Stuetzle, and Y. Shen. Loss functions for binary class probability estimation ans classification: structure and applications, 2005. Technical Report, University of Pennsylvania.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[CBG+17] M. Cissé, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier. Parseval networks: improving robustness to adversarial examples. In *34th ICML*, 2017.

[CBIK+18] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature, 2018.

[CM00] Claudia Czado and Axel Munk. Noncanonical links in generalized linear models - when is the effort justified? *Journal of Statistical Planning and Inference*, 87, 2000.

[CR02] Claudia Czado and Adrian Raftery. Choosing the link function and accounting for link uncertainty in generalized linear models using bayes factors. *Statistical Papers*, 47, 2002.

[Cza97] Claudia Czado. On selecting parametric link transformation families in generalized linear models. *Journal of Statistical Planning and Inference*, 61:125–139, 05 1997.

[dF49] B. de Finetti. Rôle et domaine d'application du théorème de Bayes selon les différents points de vue sur les probabilités (*in French*). In *18th International Congress on the Philosophy of Sciences*, pages 67–82, 1949.

[DG17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[FTS17] S. Flaxman, Y.W. Teh, and D. Sejdinovic. Poisson Intensity Estimation with Reproducing Kernels. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[GBCC15] Shirin Golchi, D. Bingham, H. Chipman, and David Campbell. Monotone emulation of computer experiments. *SIAM/ASA Journal on Uncertainty Quantification*, 3:370–392, 01 2015.

[HHLK19] Pashupati Hegde, Markus Heinonen, Harri Lähdesmäki, and Samuel Kaski. Deep learning with differential gaussian process flows. In *Proceedings of Machine Learning Research*, volume 89, pages 1812–1821, 16–18 Apr 2019.

[HT92] W.K. Hardle and Berwin Turlach. Nonparametric approaches to generalized linear models. 02 1992.

[KEC19] Ieva Kazlauskaite, Carl Henrik Ek, and Neill D. F. Campbell. Gaussian process latent variable alignment learning. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, 2019.

[KKSK11] Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems 24*. 2011.

[KM99] M.J. Kearns and Y. Mansour. On the boosting ability of top-down decision tree learning algorithms. *J. Comp. Syst. Sc.*, 58:109–128, 1999.

[Kru64] J.B. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 1964.

[KS09] Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*, 2009.

[LC10] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[LGOR15] Chris Lloyd, Tom Gunter, Michael Osborne, and Stephen Roberts. Variational inference for gaussian process modulated poisson processes. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1814–1822, Lille, France, 2015. PMLR.

[LH19] Siqi Liu and Milos Hauskrecht. Nonparametric regressive point processes based on conditional gaussian processes. In *Advances in Neural Information Processing Systems 32*. 2019.

[Lim18] Cong Han Lim. An efficient pruning algorithm for robust isotonic regression. In *Advances in Neural Information Processing Systems 31*. 2018.

[LR14] Ronny Luss and Saharon Rosset. Generalized isotonic regression. *Journal of Computational and Graphical Statistics*, 23, 2014.

[LRG+17] Cheng Li, Santu Rana, Satyandra K. Gupta, Vu Nguyen, and Svetha Venkatesh. Bayesian optimization with monotonicity information. In *NIPS Workshop on Bayesian Optimization*, 2017.

[MM06] Peter McCullagh and Jesper Møller. The permanental process. *Advances in Applied Probability*, 38(4), 2006.

[MP92] Arak Mathai and Serge Provost. *Quadratic Forms in Random Variables: Theory and Applications*. Marcel Dekker, Inc., 1992.

[MXZ06] Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *J. Mach. Learn. Res.*, 7:2651–2667, 2006.

[NDF00] Ioannis Ntzoufras, Petros Dellaportas, and Jonathan Forster. Bayesian variable and link determination for generalised linear models. *Journal of Statistical Planning and Inference*, 111, 03 2000.

[NM20] Richard Nock and Aditya Krishna Menon. Supervised learning: No loss no cry. In *ICML'20*, 2020.

[NN08] Richard Nock and Frank Nielsen. On the efficient minimization of classification-calibrated surrogates. In *NIPS*21*, pages 1201–1208, 2008.

[NNA16] Richard Nock, Frank Nielsen, and Shun-ichi Amari. On conformal divergences and their population minimizers. *IEEE Trans. IT*, 62:1–12, 2016.

[Nys28] Evert Johannes Nyström. Über die praktische auflösung von linearen integralgleichungen mit anwendungen auf randwertaufgaben der potentialtheorie. *Commentationes physico-mathematicae*, 4:1–52, 1928.

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[RV10] Jaakko Riihimäki and Aki Vehtari. Gaussian processes with monotonicity information. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.

[RW05] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[RW10] Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

[Sav71] L.-J. Savage. Elicitation of personal probabilities and expectations. *J. of the Am. Stat. Assoc.*, 66:783–801, 1971.

[Sch89] M.-J. Schervish. A general method for comparing probability assessors. *Ann. of Stat.*, 17(4):1856–1879, 1989.

[SF06] Hans J. Skaug and David A. Fournier. Automatic approximation of the marginal likelihood in non-gaussian hierarchical models. *Computational Statistical Data Analysis*, 51, 2006.

[SPV16] Eero Siivola, Juho Piironen, and Aki Vehtari. Automatic monotonicity detection for gaussian processes. In *arXiv 1610.05440*, 2016.

[ST03] Tomoyuki Shirai and Yoichiro Takahashi. Random point fields associated with certain fredholm determinants ii: Fermion shifts and their ergodic and gibbs properties. *The Annals of Probability*, (3), 07 2003.

[Tel13] M. Telgarsky. Boosting with the logistic loss is consistent. In *26 th COLT*, pages 911–965, 2013.

[UKEC19] Ivan Ustyuzhaninov, Ieva Kazlauskaite, Carl Henrik Ek, and Neill D. F. Campbell. Monotonic gaussian process flow. In *arXiv 1905.12930*, 2019.

[WB17] Christian J. Walder and Adrian N. Bishop. Fast bayesian intensity estimation for the permanental process. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, 2017.

[XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[YLR+18] Ang Yang, Cheng Li, Santu Rana, Sunil Gupta, and Svetha Venkatesh. Sparse approximation for gaussian process with derivative observations. In *AI 2018: Advances in Artificial Intelligence*, 2018.

[YW09] L Yeganova and W Wilbur. Isotonic regression under lipschitz constraint. *Journal of Optimization Theory and Applications*, 2009.

[ZE02] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Knowledge Discovery and Data Mining*, 2002.

[Zha04] J. Zhang. Divergence function, duality, and convex analysis. *Neural Computation*, 16:159–195, 2004.

A Notations used

$\mathcal{Y}$	$\doteq \{-1, 1\}$, real-valued labels
$\mathcal{Y}^*$	$\doteq \{0, 1\}$, Boolean labels
ℓ, ℓ_ν	(proper) loss, loss depending on source ν
$\underline{L}$	(conditional) Bayes risk
$-\underline{L}', \chi$	canonical link, composite link
σ	sigmoid function
ν	source
G, G_ν	margin loss, margin loss depending on source ν
D_G	Bregman divergence with (convex) generator G
$G^\star$	Legendre conjugate (G differentiable)
$\nu_0; \mu, \gamma$	ISGP constant; prior mean, prior precision (of constant)
$f; \boldsymbol{w}$	GP function and Mercer weights
$k(\cdot,\cdot); \boldsymbol{\varphi}, \Lambda$	GP kernel, kernel basis, kernel eigenvalues
M	size of the Mercer expansion
$\boldsymbol{\psi}$	positive-definite integral of the kernel feature's outer product
Θ	collected ISGP *hyper*-parameters *i.e.* γ, μ, $k(\cdot,\cdot)$ + likelihood params
$\Gamma \doteq \{\boldsymbol{w}, \nu_0\}$	collected ISGP parameters
q	approximate posterior distribution
$\widehat{\Gamma}, \widehat{\Sigma}_\Gamma$	approximate posterior mean, covariance (here, of Γ)
α	univariate Gaussian likelihood precision
$\boldsymbol{\beta}$	generalised linear model weight vector

B On the Origin Problem and an Alternative Solution

Here we discuss the role of the random intercept ν_0 in definition 2 and offer an alternative formulation.

While the integration and squaring transformations guarantee monotonicity, the question remains where to integrate from. To see this, omit the random intercept ν_0, denote by r the l.h.s. lower limit of integration, and define $\nu^{(r)}(x) = \int_r^x f^2(z)\,\mathrm{d}z$. For GP distributed f, this construction induces an artefact in the distribution for $\nu^{(r)}$, which is roughly speaking the fact that $\nu^{(r)}(r) = 0$ with probability one.

In the main text, we alleviate this issue by introducing the random intercept ν_0. We offer here an alternative solution, the idea of which is to shift the starting point of integration outside the domain of interest. Assume without loss of generality that the domain of interest is the positive real line. Concretely, the alternative formulation would model $\hat{\nu}^{(r)}(x)$ as, for $r \leq 0$,

$$\hat{\nu}^{(r)}(x) = \int_r^x f^2(z)\,\mathrm{d}z - \mu(r)$$

$$f \sim \mathrm{GP}(k(\cdot,\cdot)) \Leftrightarrow \begin{cases} f(\cdot) = \boldsymbol{w}^\top \boldsymbol{\varphi}(\cdot) \\ \boldsymbol{w} \sim \mathcal{N}(0, \Lambda), \end{cases}$$

where we choose $\mu(r) \doteq \mathbb{E}\big[\int_r^0 f^2(z)\,\mathrm{d}z\big]$ in order to ensure that the prior mean at the origin is zero (*i.e.* $\mathbb{E}\big[\nu_r(0)\big] = 0$), while $r < 0$ controls the prior variance at the origin. In other words, we simply integrate f^2 from outside of the domain of interest.

Given the result (3), for stationary kernels this is equivalent to defining

$$\hat{\nu}^{(r)}(x) = \boldsymbol{w}^\top \boldsymbol{\psi}(x+r)\boldsymbol{w} - r \times k(0,0),$$

with $\boldsymbol{w}$ and k of definition 2 and (3), respectively, and $\boldsymbol{\psi}(x)$ defined by (3).

This parsimonious approach allows to dispense with the *a priori* Gaussian intercept ν_0, but is appropriate only for data which is known to lie on *e.g.* the positive real line, since $\nu_r(r) = -\mu(r)$ with probability one.

C Proof of Theorems 2 and 4

▷ Proof of Theorem 2 – (i) we rewrite (2) using $\chi = \nu^{-1} \circ g$ for a function g that we want to elicit as the canonical link of a proper loss (*i.e.* negative the derivative of its Bayes risk). From (2), g must satisfy

$$g^{-1}(x) = \frac{G'(-\nu^{-1}(x))}{G'(-\nu^{-1}(x)) + G'(\nu^{-1}(x))},$$

Since G is decreasing and convex, its derivative is negative increasing. ν^{-1} is increasing because ν is, and so the right-hand side of (C) is increasing with values in the $[0,1]$ interval, furthermore having $g^{-1}(0) = 1/2$. Its inverse g is therefore also increasing in the interval $[0,1]$ and its derivative can therefore be used as weight function to craft a proper loss ℓ following *e.g.* [RW10, Theorem 1], yielding for this loss $g = -\underline{L}'$, as claimed for point (i). The proof of (ii) is immediate: as G is convex and classification calibrated, it satisfies $G'(0) < 0$ [BJM06, Thm. 6], implying $G'_\nu(0) = \rho(0)G'(0) < 0$ since $\rho(x) \doteq 1/\nu'(\nu^{-1}(x)) > 0$, assuming wlog $|f| \ll \infty$ almost everywhere. This implies G_ν classification calibrated [BJM06, Thm. 6].

▷ Proof of Theorem 4 – Our proof uses (2) and (2). We consider a proper canonical loss $\ell(y,x)$ where x denotes a real-valued prediction. Because a Bregman divergence is always convex in its left parameter, we have

$$\begin{aligned}
\mathbb{E}_{(x,y)}\left[\ell(y,(-\underline{L}')^{-1}(x))\right] &= \mathbb{E}_{(x,y)}\left[D_{(-\underline{L})^\star}\left(x \| -\underline{L}'(y^*)\right)\right] \\
&= \mathbb{E}_{(x,y)}\left[D_{(-\underline{L})^\star}\left(\mathbb{E}_\nu\left[\nu(x)\right] \| -\underline{L}'(y^*)\right)\right] \\
&\leq \mathbb{E}_{(x,y),\nu}\left[D_{(-\underline{L})^\star}\left(\nu(x) \| -\underline{L}'(y^*)\right)\right] = \mathbb{E}_{(x,y),\nu}\left[\ell_\nu(y,\chi^{-1}(x))\right],
\end{aligned}$$

as claimed.

D Inference with Integrated Squared Gaussian Processes (Additional Details)

Recall that we denote the parameters Γ and the hyper-parameters Θ. The Laplace approximation to the marginal likelihood is then

$$\log p(\mathcal{D}|\boldsymbol{\theta}) \approx \log q(\mathcal{D}|\boldsymbol{\theta}) \doteq \log p(\mathcal{D}, \widehat{\Gamma}) - \frac{1}{2}\log\det H.$$

in terms of the Hessian of (4.1). Optimising this expression with respect to $\boldsymbol{\theta}$ is non-trivial since $\widehat{\Gamma}$ depends on $\boldsymbol{\theta}$. We employ the generic approach from [SF06], which uses automatic differentiation to achieve the same result as — while avoiding the manual gradient calculations of — the usual approach in the GP literature (see *e.g.* [RW05, §3.4]).

We first use the total derivative to decompose

$$\nabla_{\boldsymbol{\theta}} \log q(\mathcal{D}|\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log q(\mathcal{D}|\boldsymbol{\theta}) + \left(\frac{\partial}{\partial \boldsymbol{\theta}}\widehat{\Gamma}\right)^{\top} \left.\frac{\partial}{\partial \Gamma}\right|_{\Gamma=\widehat{\Gamma}} \log q(\mathcal{D}|\boldsymbol{\theta}),$$

where we may show using the implicit function theorem that

$$\frac{\partial}{\partial \boldsymbol{\theta}}\widehat{\Gamma} = -H^{-1} \left.\frac{\partial^2}{\partial \Gamma \partial \boldsymbol{\theta}^{\top}}\right|_{\Gamma=\widehat{\Gamma}} \log p(\mathcal{D}, \Gamma).$$

In line with [SF06], we employ automatic differentiation to compute both H (as a function of Γ), and then to differentiate $\log q(\mathcal{D}|\boldsymbol{\theta})$ (which is a function of $\det H$). In summary, the entire marginal likelihood maximisation procedure requires only *i)* the (trivial) implementation of the log prior and log likelihood functions, *ii)* automatic differentiation software such as [PGM+19] (to be invoked in three different ways[1]), and *iii)* , non-linear optimisation software.

▷ **Computational Complexity.** For likelihood functions of the form given at the top of Section 4.1 we may use (3) to compute $\nu(x_n)$. As a result, although inference under the ISGP prior may appear challenging due to the integral in definition 2, the log posterior can be evaluated in only $\mathcal{O}(M^2N)$ time, where M is the size of the basis and N is the number of data points. This is the usual cost for sparse GP approximations with M basis functions (or inducing points). The choice of *squared* transformation in definition 2 makes this possible.

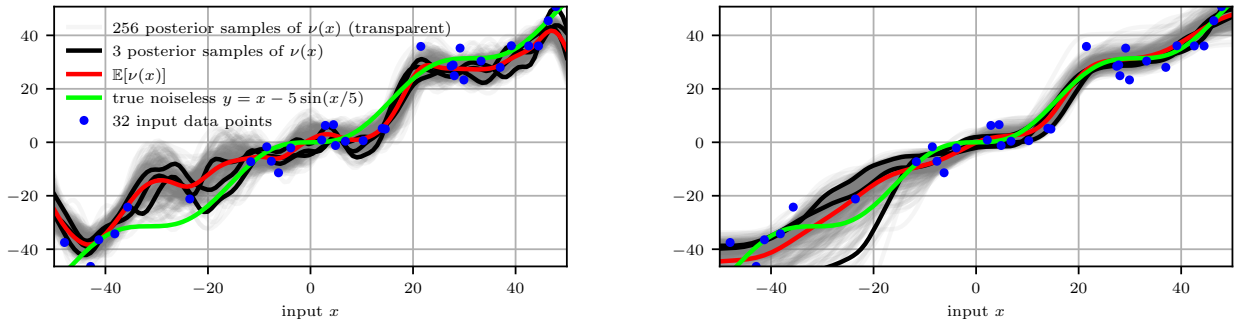
D.1 Likelihood Functions

For concreteness, and to specify the parameterisations we employ, we complete this section by introducing the two univariate likelihood functions used throughout the paper.

▷ **Gaussian Likelihood for Regression.** Here we have $y_n \in \mathbb{R},\ n = 1, 2, \ldots, N$, and

$$\begin{aligned}\log p(y_n|\nu(x_n), \Theta) &= \log \mathcal{N}(y_n|\nu(x_n), \alpha^{-1}) \\ &= \frac{1}{2}\log(\alpha) - \frac{1}{2}\log(2\pi) - \frac{\alpha}{2}\left(\nu_0 + \boldsymbol{w}^{\top}\boldsymbol{\psi}(x_n)\boldsymbol{w} - y_n\right)^2,\end{aligned}$$

[1] These include: 1) differentiating $\log p(\Gamma|\mathcal{D})$ with respect to Γ for *maximum a posteriori* optimisation, 2) computing the Hessian w.r.t. Γ, H, and 3) differentiating $\frac{1}{2}\log\det H$ w.r.t. Θ for *maximum marginal likelihood* optimisation — see [SF06] for the details.



(a) GP Prior (b) ISGP Prior

Figure 3: Regression with a prior that is *a)* a Gaussian process and *b)* our integrated squared Gaussian process of section 3. We use (Laplace approximate) maximum marginal likelihood hyper-parameters along with our novel trigonometric kernel of subsection 4.3 in both cases.

where for simplicity (and since our model is discriminative) we neglect to notate conditioning on x both above and, as appropriate, throughout. This model is illustrated on the r.h.s. of Figure 3.

▷ **Sigmoid-Bernoulli Likelihood for Classification.** Here we let $y_n \in \{0, 1\}$, and

$$p(y_n|\nu(x_n), \Theta) = \text{Bernoulli}\left(y_n|(-\underline{L}')^{-1} \circ \nu(x_n)\right),$$

where $\text{Bernoulli}(y|\rho) = \rho^y(1-\rho)^{1-y}$, and we are composing with the logistic sigmoid $(-\underline{L}')^{-1}(\nu) = 1/(1+\exp(-\nu))$. The above likelihood function may be expanded analogously to (D.1), to obtain a readily implementable form.

E Trigonometric Kernel (Additional Details)

▷ **Closed form $k(x, z)$.** Although we do not require it, the kernel is available in closed form. Letting $d = c\,|x - z|$ we have

$$k(x,z) = \frac{b\left(ae^{\frac{i\pi d(2M+3)}{2}} + ae^{\frac{i\pi d}{2}} - e^{i\pi d(M+1)} - e^{i\pi d} - 2a^M e^{\frac{i\pi d(M+2)}{2}}\left(a\cos\left(\frac{\pi d}{2}\right) - 1\right)\right)}{2a^M e^{\frac{i\pi d(M+1)}{2}}\left(ae^{i\pi d} - (a^2+1)\,e^{\frac{i\pi d}{2}} + a\right)},$$

and for $M \to \infty$,

$$k(x,z) = \frac{b}{2}\left(\frac{a}{a - \exp(\frac{i\pi d}{2})} + \frac{1}{a\exp(a - \frac{i\pi d}{2}) - 1}\right).$$

▷ **Closed form $\boldsymbol{\psi}$.** The integrals needed for $\boldsymbol{\psi}(x) \in \mathbb{R}^{M\times M}$ of (3) are, for the pairs of sine terms,

$$\int_0^x \sin\left(\pi mcz\right)\sin\left(\pi ncz\right)\mathrm{d}z = \begin{cases} \frac{x}{2} - \frac{\sin(2\pi cmx)}{4\pi cm} & m = n \\ \frac{n\sin(\pi cmx)\cos(\pi cnx) - m\cos(\pi cmx)\sin(\pi cnx)}{\pi cm^2 - \pi cn^2} & m \neq n, \end{cases}$$

for the pairs of cosine terms,

$$\int_0^x \cos\left(\pi mcz\right)\cos\left(\pi ncz\right)\mathrm{d}z = \begin{cases} \frac{x}{2} + \frac{\sin(2\pi cmx)}{4\pi cm} & m = n \\ \frac{m\sin(\pi cmx)\cos(\pi cnx) - n\cos(\pi cmx)\sin(\pi cnx)}{\pi cm^2 - \pi cn^2} & m \neq n, \end{cases}$$

and for the mixed terms,

$$\int_0^x \sin\left(\pi mcz\right)\cos\left(\pi ncz\right)\mathrm{d}z = \begin{cases} \frac{\sin^2(\pi cmx)}{2\pi cm} & m = n \\ \frac{-n\sin(\pi cmx)\sin(\pi cnx) - m\cos(\pi cmx)\cos(\pi cnx) + m}{\pi cm^2 - \pi cn^2} & m \neq n. \end{cases}$$

F Nyström Approximation

Our *trigonometric kernel* of Section 4.3 is ideally suited to inference under the ISGP model, in that it admits efficient computation of the matrix $\boldsymbol{\psi}(x)$ of (3). There is another more subtle condition which must be satisfied, however, in order for our Laplace approximate hyper-parameter optimisation procedure to be efficient. That is, only the spectrum Λ may depend on the hyper-parameters of the kernel, while the basis $\boldsymbol{\varphi}(x)$ must be fixed. Due to these requirements the tempting and popular Nyström approximation [Nys28] is generally insufficient, and our trigonometric kernel is therefore essential for tractable inference under the ISGP model.

In the remainder of this section we investigate an alternative approach based on the Nyström approximation to the kernel. In many GP inference problems, this approximation method is applicable to a rather wide range of kernels. Here however, we require the integral for $\boldsymbol{\psi}(x)$, which is not generally available. Fortunately, as we now demonstrate, these terms are available in closed form for the Gaussian kernel.

Nonetheless, a drawback of the Nyström method remains as — unlike our Trigonometric kernel — the hyper-parameters of the kernel affect the basis $\boldsymbol{\varphi}(x)$, not just the spectrum. This dependence renders the optimisation of our marginal likelihood approximation in Section 4.1 prohibitively expensive — for fixed hyper-parameters the Nyström method may be useful, however, and for completeness we derive the key expressions here.

▷ **General Setup.** The Nyström idea is to note that the φ_i, λ_i pairs are eigenfunctions of the integral operator (see [RW05] section 4.3) on the reproducing kernel Hilbert space $\mathcal{H}(k)$ induced by k,

$$\begin{aligned} T_k : \mathcal{H}(k) &\to \mathcal{H}(k) \\ f &\mapsto T_k f \doteq \int_{x\in\Omega} k(x,\cdot) f(x) p(x)\,\mathrm{d}x, \end{aligned}$$

where p may be freely chosen provided it has an appropriate support. The idea of the Nyström approximation [Nys28] to T_k is to draw M samples $X = \{x_1, x_2, \ldots, x_M\}$ from p and define the Monte Carlo approximation $T_k^{(X)} g \doteq \frac{1}{M}\sum_{x\in X} k(x,\cdot)g(x)$. Then the eigenfunctions and eigenvectors of T_k may be approximated via the eigenvectors $\boldsymbol{e}_i^{(\mathrm{mat})}$ and eigenvalues $\lambda_i^{(\mathrm{mat})}$ of $k(X,X)$, as (we abuse the notation so that $k(X,X)$ is an $M \times M$ matrix of kernel evaluations,

etc.)

$$\varphi_i^{(X)}(z) \doteq \sqrt{M}/\lambda_i^{(\mathrm{mat})} k(X,z)^\top \boldsymbol{e}_i^{(\mathrm{mat})}$$
$$\lambda_i^{(X)} \doteq \lambda_i^{(\mathrm{mat})}/M.$$

For our setting, we further require (in addition to the usual matrix eigendecomposition algorithm), the following integral

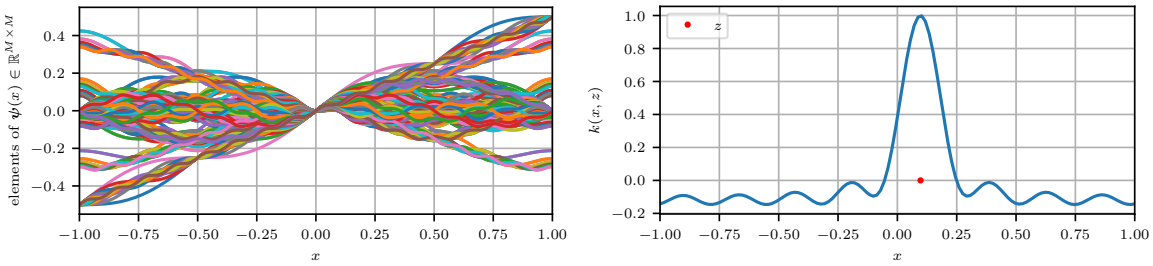
$$\begin{aligned}(\boldsymbol{\psi}(x))_{ij} &= \int_0^x \varphi_i^{(X)}(z)\varphi_j^{(X)}(z)\,\mathrm{d}z \\ &= \frac{M}{\lambda_i^{(\mathrm{mat})}\lambda_j^{(\mathrm{mat})}}\, \boldsymbol{e}_i^{(\mathrm{mat})^\top} \underbrace{\int_0^x K(z,X)^\top K(z,X)\,\mathrm{d}z}_{\doteq \boldsymbol{\psi}_{k,X}(x)\in\mathbb{R}^{M\times M}} \boldsymbol{e}_j^{(\mathrm{mat})}.\end{aligned}$$

▷ **Squared Exponential Kernel.** Although a Mercer decomposition is available in closed form (see *e.g.* [RW05]) for the popular kernel

$$k(x,z) = b\exp\big(-\frac{a}{2}(x-z)^2\big),$$

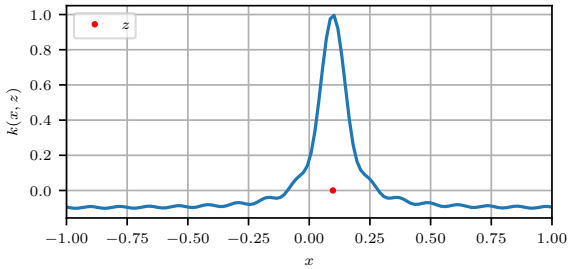
it turns out that the integrals we require for $\boldsymbol{\psi}$ are challenging for that decomposition. Fortunately the Nyström approximation is convenient, since the key term in (F) is given by

$$\begin{aligned}(\boldsymbol{\psi}_{k,X}(t))_{ij} &= \int_0^t k(x_i,z)k(x_j,z)\,\mathrm{d}z \\ &= \frac{b^2\sqrt{\pi}}{2\sqrt{a}}\exp\big(\frac{a}{4}(x_i-x_j)^2\big)\Big(\mathrm{erfi}\big(\frac{\sqrt{a}}{2}(x_i+x_j)\big) - \mathrm{erfi}\big(\frac{\sqrt{a}}{2}(x_i+x_j-2t)\big)\Big).\end{aligned}$$

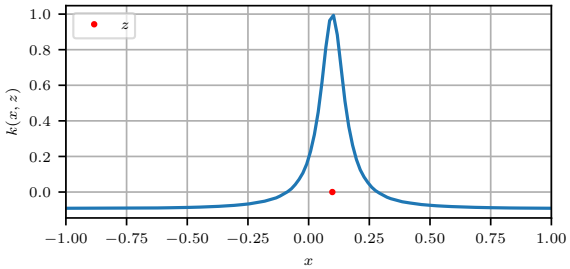


(a) elements of $\boldsymbol{\psi}(\cdot)$ (for $M = 16$)

(b) kernel (for $M = 16$)



(c) kernel (for $M = 32$)



(d) kernel (for $M = 64$)

Figure 4: Visualisation of our trigonometric kernel of Section 4.3, and the $\boldsymbol{\psi}(z)$ of (3) induced by it. M is (half) the number of basis functions. The remaining hyper-parameters are $a = 1.2$, $b = 1$ and $c = 1$. See the labels on the figures for details.

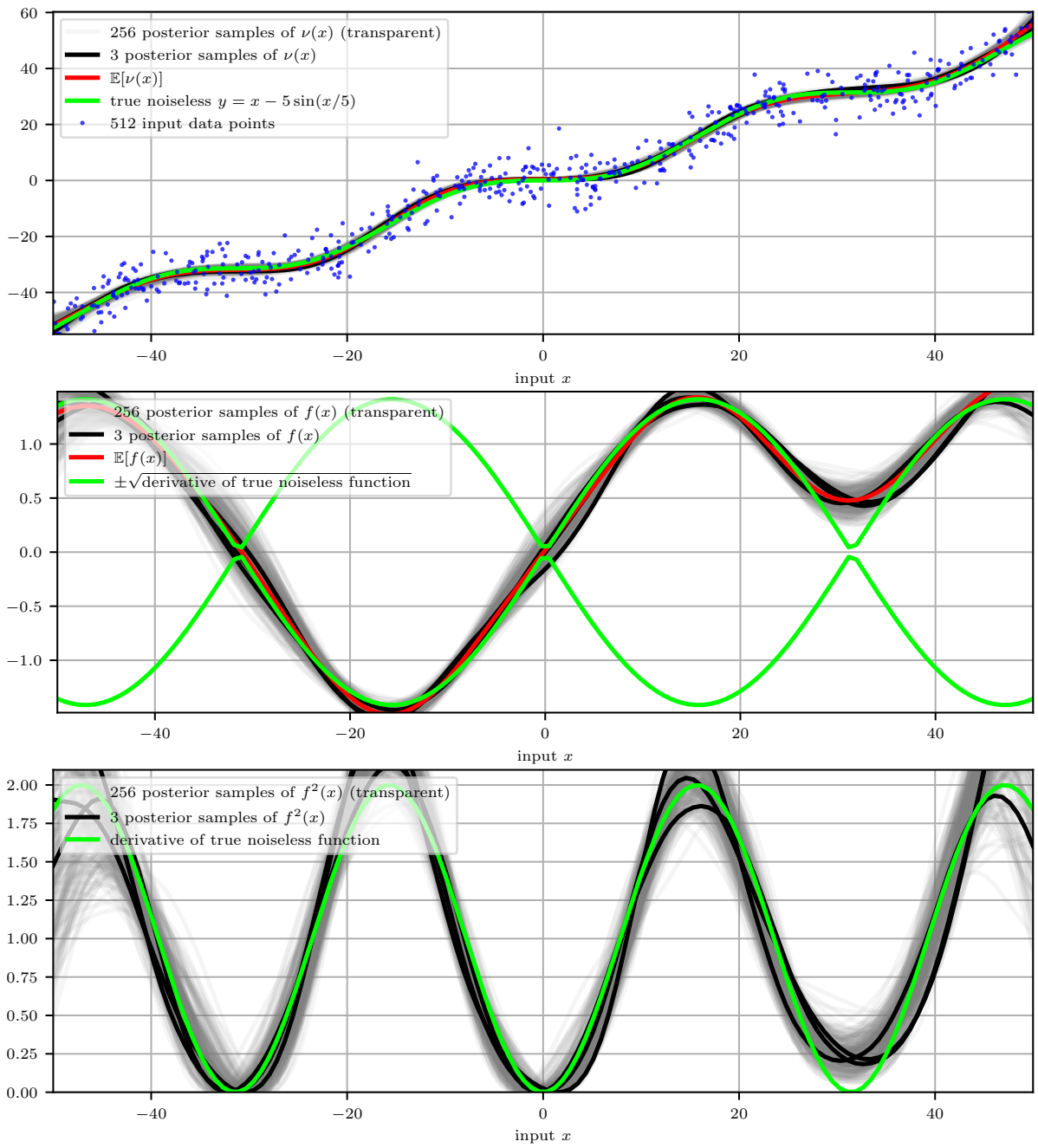


Figure 5: A toy regression problem with ISGP prior and Gaussian likelihood function. *Upper plot:* The posterior predictive distribution for ν, our inferred monotonic function, along with the ground truth function and training data points. *Middle plot:* The posterior predictive distribution for our f of definition 2, which is the square root of the derivative of ν, along with $\pm$ the square root of the derivative of the ground truth function. *Lower plot:* Similar to the middle plot but with the squared transformation included. We use maximum marginal likelihood parameters with the trigonometric kernel of subsection 4.3, and a kernel scale parameter of $c = 1/100$, so that the inferred functions are periodic with period 200.

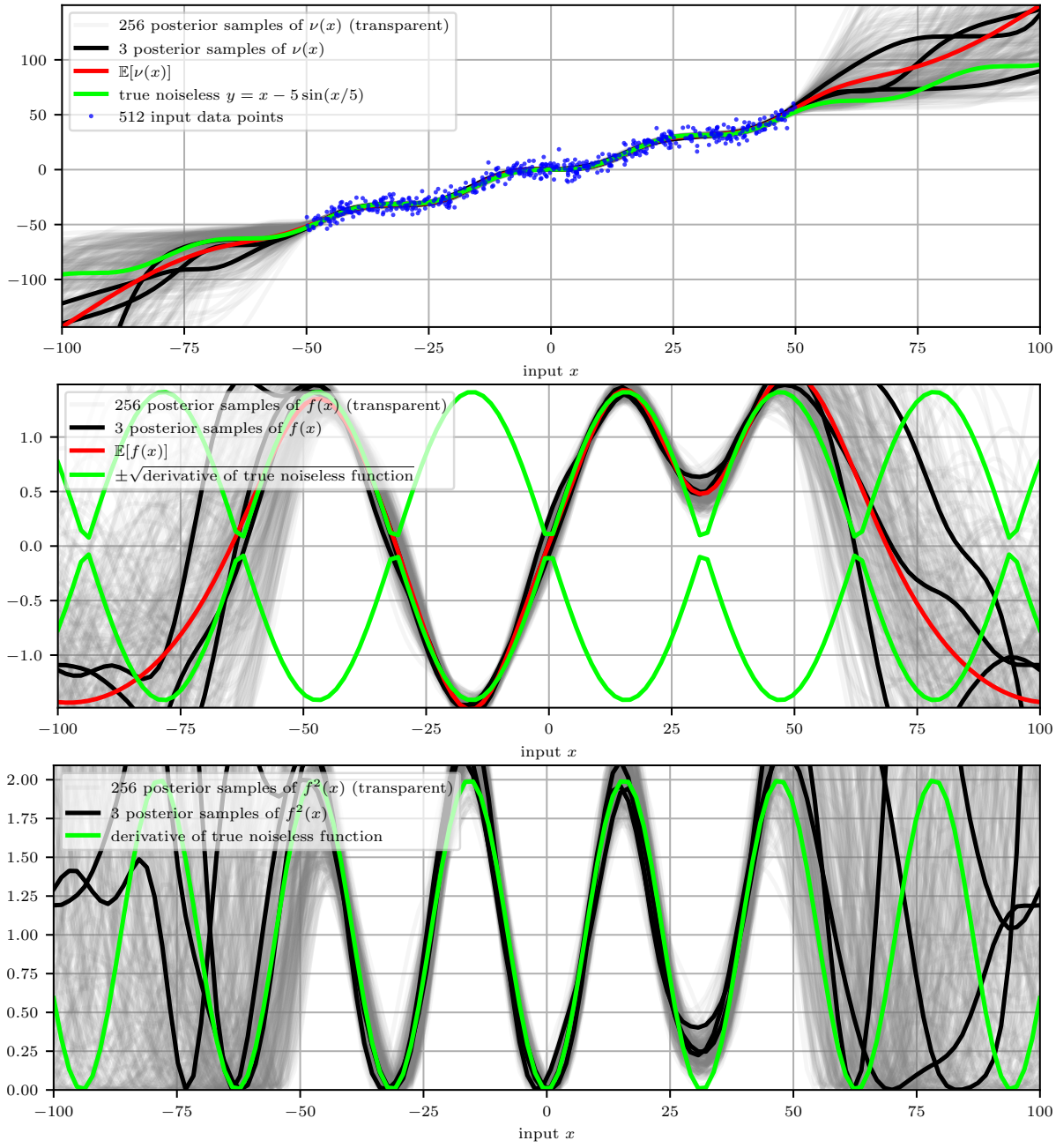


Figure 6: A zoomed out version of Figure 5.

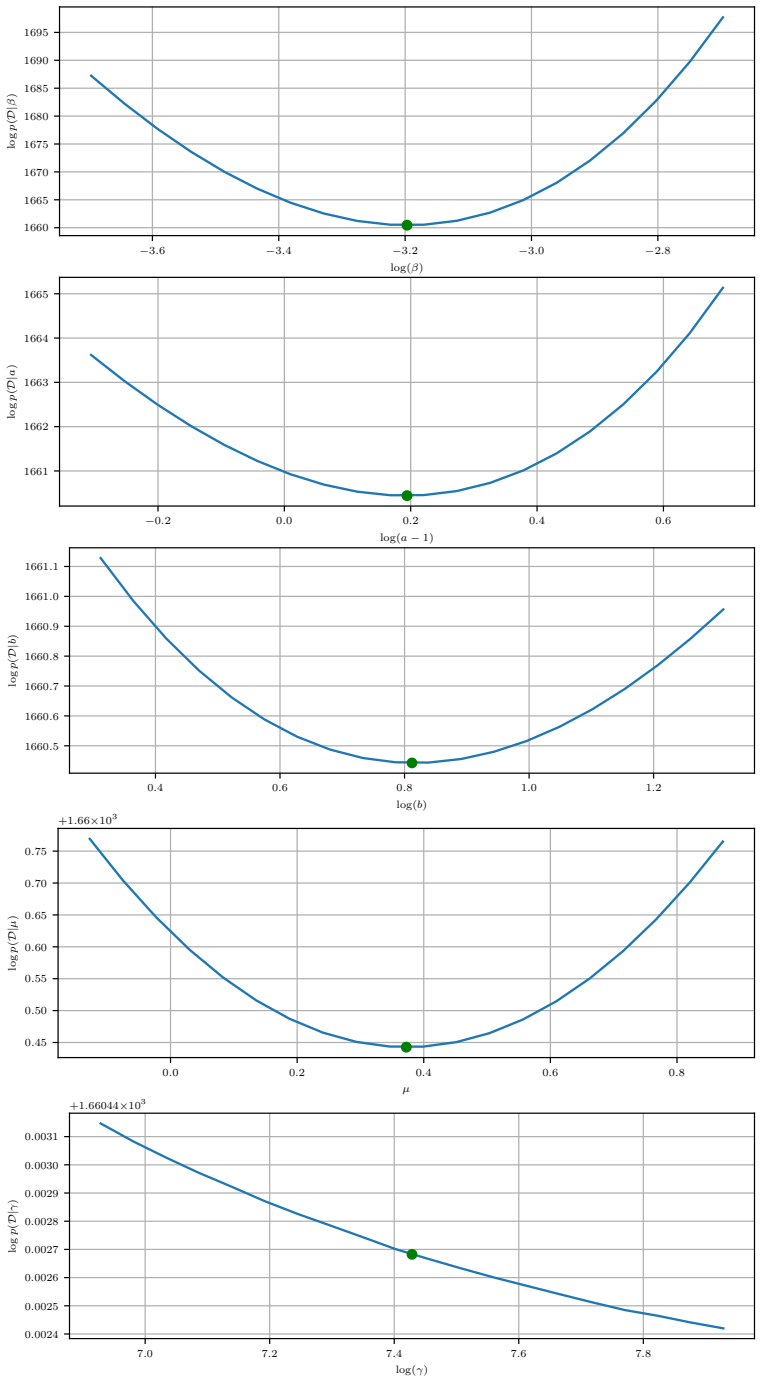
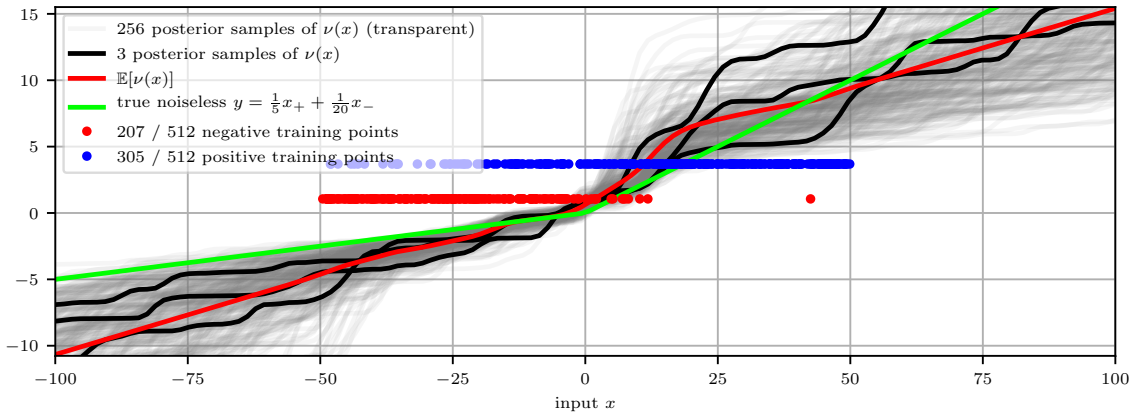
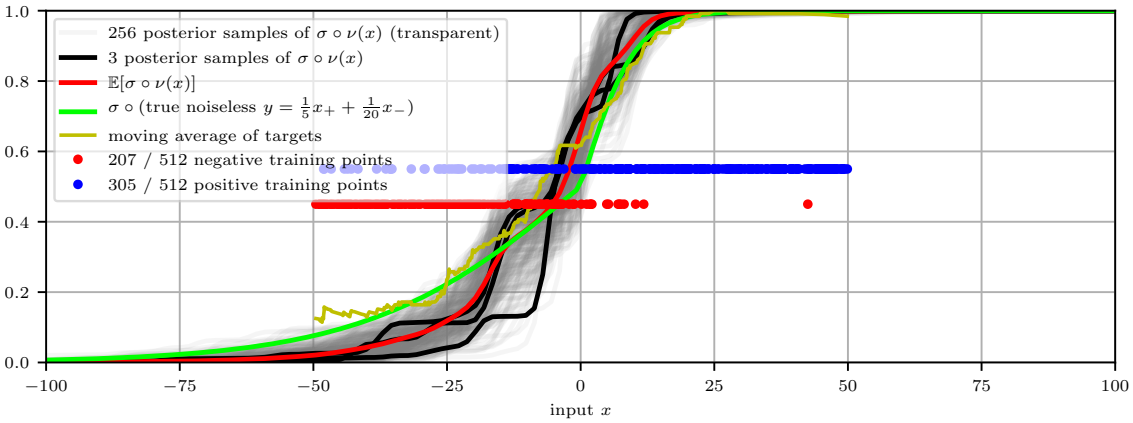


Figure 7: Visualising the log marginal likelihood for the problem of Figure 5 (and Figure 6). We computed the maximum marginal likelihood parameters using the method of subsection 4.1. Then we varied each hyper-parameter about this optimal value (represented by the greed dots), holding the others fixed. With the exception of the extremely flat lowest plot (for the prior variance γ of the intercept ν_0, which is expected to have a flat marginal posterior), the marginal likelihood optimisation finds a stable local minimum. Note that for clarity the vertical axis labels neglect to notate conditioning on certain variables.

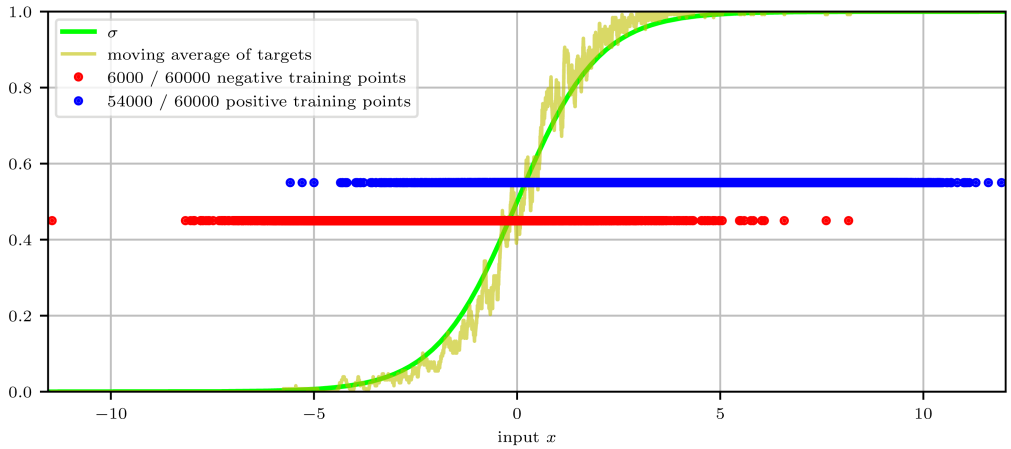


(a) posterior latent source function $\nu(\cdot)$

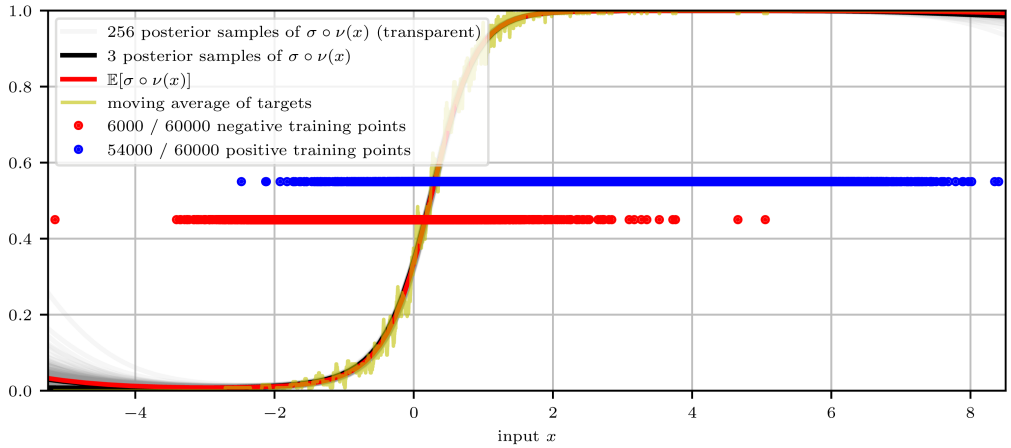


(b) composition of the source function from (a) with the sigmoid $\sigma(y) = 1/(1 + \exp(-y))$

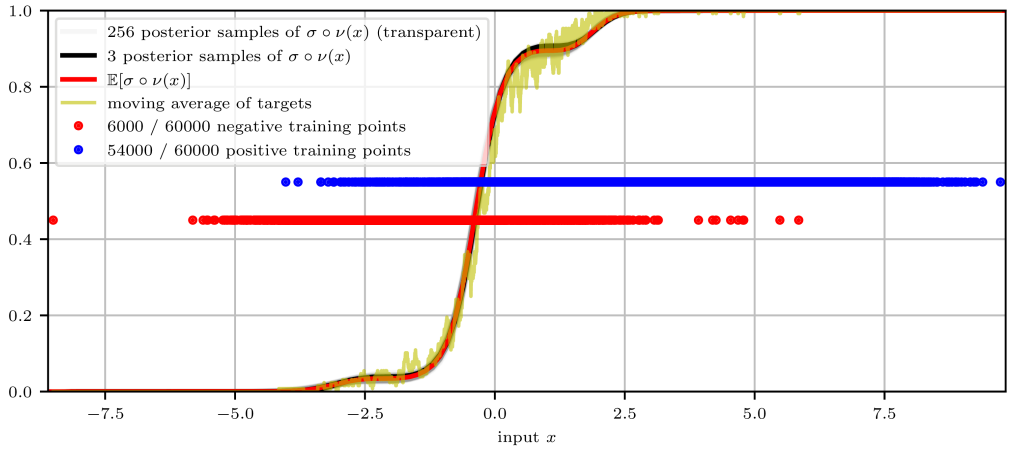
Figure 8: A toy classification problem with ISGP prior and Bernoulli likelihood function. *Upper plot:* The posterior predictive distribution for ν (our inferred monotonic function) along with the ground truth function and training data points (with binary labels represented by two distinct y-values). *Lower plot:* The posterior predictive distribution for the inverse link function $\sigma \circ \nu$, where $\sigma(\nu) = 1/(1 + \exp(-\nu))$ is a shorthand for the sigmoid function (inverse canonical link of the log loss).



(a) logistic regression



(b) GP linkgistic regression



(c) ISGP linkgistic regression

Figure 9: The inverse link function for logistic regression (upper), GP linkgistic regression (lower) and ISGP linkgistic regression (lower), on the *Fashion-MNIST* task of class 3 (*dress*) *vs.* the rest. The x-axis is the input to the (inverse) link function, which is equal to the output of the generalised linear model, *i.e.* $x_n = \boldsymbol{\beta}^\top \boldsymbol{z}_n$ as per Section 4.2.