PoRB-Nets: Poisson Process Radial Basis Function Networks

Beau Coker
Department of Biostatistics
Harvard University

Melanie F. Pradier
SEAS
Harvard University

Finale Doshi-Velez
SEAS
Harvard University

Abstract

Bayesian neural networks (BNNs) are flexible function priors well-suited to situations in which data are scarce and uncertainty must be quantified. Yet, common weight priors are able to encode little functional knowledge and can behave in undesirable ways. We present a novel prior over radial basis function networks (RBFNs) that allows for independent specification of functional amplitude variance and lengthscale (i.e., smoothness), where the inverse lengthscale corresponds to the concentration of radial basis functions. When the lengthscale is uniform over the input space, we prove consistency and approximate variance stationarity. This is in contrast to common BNN priors, which are highly nonstationary. When the input dependence of the lengthscale is unknown, we show how it can be inferred. We compare this model's behavior to standard BNNs and Gaussian processes using synthetic and real examples.

1 INTRODUCTION

Neural networks (NNs) are flexible universal function approximators that have been applied with success in many domains. Bayesian neural networks (BNNs) capture function space uncertainty in a principled manner by placing priors over network parameters (Hinton and Neal, 1995). Unfortunately, priors in parameter space often lead to unexpected behavior in function space, making it difficult to incorporate meaningful information about function space properties (Lee, 2004). Two such properties of importance are amplitude variance and lengthscale, including how they might vary over the input space.

While Gaussian processes (GPs) are function priors that can easily encode these properties via the covariance function, there are many situations in which we would prefer BNNs to GPs: BNNs may be computationally more scalable, especially at test time, and they have an explicit parametric expression for posterior samples, which is convenient when additional computation is needed on the function (e.g., finding a minima) (Hernández-Lobato et al., 2014).

Therefore, a natural question arises: can we design BNN priors that encode function space properties as in GPs while retaining the benefits of BNNs? Some approaches use sample-based methods to evaluate the discrepancy between the function space distribution and a reference distribution with desired properties (Flam-Shepherd et al., 2017; Sun et al., 2019). Pearce et al. (2019) explores different BNN architectures to recover equivalent GP kernel combinations in the infinite width limit. While promising, these approaches require challenging optimizations or rely on infinite width assumptions.

As a first step towards more expressivity for BNNs, this work focuses on a particular type of NN called a radial basis function network (RBFN). RBFNs are widely used across scientific disciplines (Dash et al., 2016) and have received renewed interest recently, both from a theoretical (Que and Belkin, 2016) and inferential perspective (Zadeh et al., 2018; Asadi et al., 2020). Importantly, each hidden unit has a center parameter corresponding to a localized activation function, which enables controlling where (over the input space) the hidden units contribute to the complexity of the function.

In this work, we introduce Poisson Process Radial Basis Function Networks (PoRB-Nets), an interpretable family of RBFNs that employ a Poisson process (PP) prior over the center parameters in an RBFN. The proposed formulation enables direct specification of functional amplitude variance and lengthscale, the latter of which can vary over the input space. We show that these properties are *decoupled*; that is, each can be specified independently

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

of the other. Intuitively, PoRB-Nets work by trading off between the concentration and scale of the radial basis functions. Consider that a higher concentration of basis functions allows for a smaller lengthscale but also a larger variance, since the basis functions add up. By making the scale of the basis functions depend inversely on their concentration, PoRB-Nets undo the impact on the variance.

PoRB-Nets have the additional benefit that the choice of the lengthscale determines the network architecture (width of the layer), since the expected number of hidden units is equal to the integral of the PP intensity over the input space. Hidden units are added or deleted from the network during inference to adjust the overall lengthscale to the data, and when the input dependence of the lengthscale is unknown, we show how it can be inferred using a sigmoidal Gaussian Cox process as a prior (Adams et al., 2009). As with GPs, and unlike networks that force a specific property (Anil et al., 2018), these properties can adjust given data. We focus on single-layer RBFNs since our interest is in theoretical properties and examining the true posterior.

Specifically, we make the following contributions: (i) we introduce a novel, intuitive prior formulation for RBFNs that encodes distributional knowledge in function space, decoupling notions of lengthscale and amplitude variance in the same way as a GP with a radial basis function (RBF) kernel; (ii) we prove important theoretical properties of consistency and amplitude stationarity; (iii) we provide an inference algorithm to learn an input dependent lengthscale and (iv) we empirically demonstrate the potential of PoRB-Nets on synthetic and real examples. The code is available at https://github.com/dtak/porbnet.

2 RELATED WORK

Early weight space priors for BNNs. Most classical NN priors aim for regularization and model selection while minimizing the amount of undesired inductive biases (Lee, 2004). MacKay (1992) proposes a hierarchical prior[1] combined with empirical Bayes. Lee (2003) proposes an improper prior for NNs, which avoids the injection of prior biases at the cost of higher sensitivity to overfitting. Robinson (2001) proposes priors to alleviate overparametrization of NN models. We build on classical weight space priors but with the goal of obtaining specific properties in function space.

[1]Hierarchical priors are convenient when there is limited parameter interpretability. The addition of upper levels to the prior reduces the influence of the choice made at the top level, making the prior at the bottom level (the original parameters) more diffuse (Lee, 2004).

Function space priors for BNNs. Some works (Flam-Shepherd et al., 2017; Sun et al., 2019) match BNN priors to specific function space priors (e.g., GPs) but rely on sampling function values at a collection of input points. These approaches do not provide guarantees outside of the sampled region, and even in that region, their enforcement of properties is approximate. Neural processes (Garnelo et al., 2018) use meta-learning to identify functional properties that may be present in new functions, but they rely on having many prior examples and do not allow the user to specify basic properties directly. In contrast, we encode functional properties via prior design, without relying on function samples.

Bayesian formulations of RBFN models. Closest to our work are Bayesian formulations of RBFNs. Barber and Schottky (1998) consider a fixed number of hidden units, fixed scale, and use a Gaussian approximation to the posterior distribution, which is available in closed form in this case. Holmes and Mallick (1998) and Andrieu et al. (2001) propose fully Bayesian formulations that employ homogeneous Poisson process priors on the center parameters, but their focus is on inferring the number of hidden units and their formulation does not decouple amplitude variance and lengthscale.

3 BACKGROUND

Bayesian neural networks (BNNs). Let $y = f(x \,|\, \boldsymbol{w}, \boldsymbol{b}) + \epsilon$, where ϵ is a noise variable and $\boldsymbol{w}$ and $\boldsymbol{b}$ refer to the weights and biases of a neural network f respectively. In the Bayesian setting, we assume a prior $\boldsymbol{w}, \boldsymbol{b} \sim p(\boldsymbol{w}, \boldsymbol{b})$. One common choice is i.i.d. normal distributions over each parameter. For better comparison to PoRB-Nets we focus on BNNs with Gaussian $\phi(z) = \exp(-z^2)$ activations. We will refer to such a model as a standard BNN (Neal, 1996).

Radial basis function networks (RBFNs). RBFNs are classical shallow neural networks that approximate arbitrary nonlinear functions through a linear combination of radial kernels (Powell, 1987). They are universal function approximators (Park and Sandberg, 1991) and are widely used across disciplines such as numerical analysis, biology, finance, and classification in spatio-temporal models (Dash et al., 2016). For an input $x \in \mathbb{R}^D$, the output of a single-hidden-layer RBFN of width K is given by:

$$f(x \,|\, \boldsymbol{\theta}) = b + \sum_{k=1}^{K} w_k \exp\left(-\frac{1}{2} s_k^2 \|x - c_k\|^2\right), \quad (1)$$

where $s_k^2 \in \mathbb{R}$ and $c_k \in \mathbb{R}^D$ are the scale and center parameters, respectively, $w_k \in \mathbb{R}$ are the hidden-to-output

weights, and $b \in \mathbb{R}$ is the bias parameter. Each k-th hidden unit can be interpreted as a local receptor centered at c_k, with radius of influence s_k and relative importance w_k (Powell, 1987).

Poisson process. A Poisson process (PP) on $\mathbb{R}^D$ is a stochastic process characterized by a positive real-valued intensity function $\lambda(c)$. For any set $\mathcal{C} \subset \mathbb{R}^D$, the number of points in $\mathcal{C}$ follows a Poisson distribution with parameter $\int_{\mathcal{C}} \lambda(c)dc$. The process is *inhomogeneous* if $\lambda(c)$ is non-constant. We use a PP as a prior on the center parameters of an RBFN.

Gaussian Cox process. A Bayesian model consisting of a Poisson process likelihood and a log Gaussian process prior $g(c)$ on the intensity function $\lambda(c)$ is called a log Gaussian Cox Process (Møller et al., 1998). Adams et al. (2009) present an extension, called the *sigmoidal Gaussian Cox process*, which passes the Gaussian process through a scaled sigmoid function. To infer an input dependent lengthscale of an RBFN, we use this process as a model for the intensity function of the PP prior on the center parameters of the RBFN.

4 MODEL

In this section we introduce Poisson Process Radial Basis Function Networks (PoRB-Nets), which achieve two essential desiderata for a functional prior. First, they enable the user to encode the fundamental basic properties of lengthscale (i.e., smoothness), amplitude variance (i.e., signal variance), and (non)stationarity. Second, PoRB-Nets adapt the complexity of the network based on the inputs. For example, if the data suggests that the function needs to be less smooth in a certain input region, then that data can override the prior. Importantly, PoRB-Nets fulfill these desiderata while retaining appealing properties of NN-based models, as discussed in Section 1.

Generative model. As in a standard BNN, we assume a Gaussian likelihood centered on the network output, and independent Gaussian priors on the weight and bias parameters. Unique to the novel PoRB-Net formulation is a Poisson process prior over the set of center parameters and a deterministic dependence of the scale parameters on the Poisson process intensity. The generative model is given by:

$$\{c_k\}_{k=1}^K \,|\, \lambda \;\sim\; \exp\left(-\int_{\mathcal{C}} \lambda(c)dc\right) \prod_{k=1}^K \lambda(c_k) \quad (2)$$

$$s_k^2 \,|\, \lambda, c_k \;=\; s_0^2 \lambda^2(c_k) \quad (3)$$

$$w_k \;\sim\; \mathcal{N}\left(0, \sigma_w^2\right) \quad (4)$$

$$b \;\sim\; \mathcal{N}\left(0, \sigma_b^2\right) \quad (5)$$

$$y_n \,|\, x_n, \boldsymbol{\theta} \;\sim\; \mathcal{N}\left(f(x_n; \boldsymbol{\theta}), \sigma^2\right), \quad (6)$$

where $f(x_n; \boldsymbol{\theta})$ is given by Eq. (1); $\lambda : \mathcal{C} \to \mathbb{R}^+$ is the (possibly non-constant) Poisson process intensity; $\boldsymbol{\theta}$ is the set of RBFN parameters, including the centers, weights, bias, and intensity; and s_0^2 is a hyperparameter that defines the scale of the radial basis function when the intensity is one. In practice, s_0^2 allows the user to control the baseline number of hidden units. For example, if computational constraints limit the number of hidden units that can be used, decreasing s_0^2 allows the user to model a smaller lengthscale without adding more units.

Different priors could be considered for the intensity function λ. One simple case is to assume a uniform intensity $\lambda(c) = \lambda$ with $\lambda^2 \sim \text{Gamma}(\alpha_\lambda, \beta_\lambda)$. Under this specific formulation, Section 5 proves that the amplitude variance is stationary as the size of the region $\mathcal{C}$ tends to infinity, and Section 6 proves that the posterior regression function is consistent as the number of observations tends to infinity; such amplitude variance only depends on the variance of the hidden-to-output weights and output bias $\mathbb{V}[f(x)] \approx \sigma_b^2 + \tilde{\sigma}_w^2$, where $\tilde{\sigma}_w^2$ is just σ_w^2 scaled by s_0. We further show that the intensity λ controls the lengthscale.

Hierarchical prior for unknown input dependence of the lengthscale. In the case when the input-dependence of the lengthscale is unknown, we further model the intensity function $\lambda(c)$ of the Poisson process by a sigmoidal Gaussian Cox process (Adams et al., 2009):

$$g \;\sim\; \mathcal{GP}(0, \Sigma(\cdot, \cdot)) \quad (7)$$

$$\lambda(c) \;=\; \lambda^* \sigma(g(c)), \quad (8)$$

where λ^* is an upper bound parameter on the intensity function and $\sigma(z) = (1 + e^{-z})^{-1}$ is the sigmoid function. In the forward pass of the network, we use the posterior mean of g to evaluate $\lambda(c)$.

Contrast to BNNs with Gaussian priors. In Sections 5 and 6, we prove that the proposed formulation has the desired properties described above. However, before doing so, we briefly emphasize that the i.i.d. Gaussian weight space prior commonly used with BNNs does not enjoy these properties. To see why, let us consider a standard feed-forward NN layer with 1-dimensional input and a Gaussian $\phi(z) = \exp(-z^2)$ activation function. We can rewrite the hidden units as $\phi(w_k x + b_k) = \phi(w_k(x - (-b_k/w_k)))$. This means that the corresponding center of the k-th hidden unit is $c_k = -b_k/w_k$ and the scale is $s_k = w_k$. If b_k and w_k have i.i.d. Gaussian priors with zero mean, as in standard BNNs, then the center parameter has a Cauchy distribution centered around zero. This is an important observation that motivates our work: A standard BNN concentrates the center of hidden units near the origin, resulting in nonstationary priors in function space.

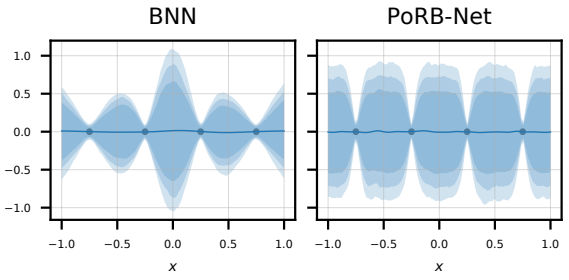


Figure 1: **PoRB-Net captures amplitude stationarity while a standard BNN does not.** Posterior predictive distributions given 4 observations.

5 VARIANCE AND LENGTHSCALE

We now return to the core desiderata: to specify a prior that separately controls a function's lengthscale and amplitude variance, as one could do using a GP with an RBF kernel. To do so, we first derive the covariance of the proposed PoRB-Net model. The full derivations supporting this section are available in Appendix A.

Neal (1996) showed that the covariance function for a single-layer BNN with a *fixed* number of hidden units $\rho(x;\theta_1),\ldots,\rho(x;\theta_K)$ and independent $\mathcal{N}(0,\sigma_w^2)$ and $\mathcal{N}(0,\sigma_b^2)$ priors on the hidden-to-output weights and output bias takes the following general form:

$$\mathrm{Cov}(f(x_1),f(x_2)) = \sigma_b^2 + \sigma_w^2 K \mathbb{E}_\theta\left[\rho(x_1;\theta)\rho(x_2;\theta)\right].$$

We show that the covariance function for a BNN with a *distribution* over the number of hidden units takes an analogous form, replacing the fixed number of hidden units K with its expectation:

$$\mathrm{Cov}(f(x_1),f(x_2)) = \sigma_b^2 + \sigma_w^2 \mathbb{E}\left[K\right] \underbrace{\mathbb{E}_\theta\left[\rho(x_1;\theta)\rho(x_2;\theta)\,|\,K\right]}_{:=U(x_1,x_2)}.$$

In the PoRB-Net model, $\theta = \{\lambda(\cdot), c_k\}$, $\rho(x;\theta_k) = \phi(\lambda(c_k)s_0\|x-c_k\|)$ where $\phi(z)=\exp(-\frac{1}{2}z^2)$, and $\mathbb{E}\left[K\right] = \int_{\mathcal{C}} \lambda(c)\,dc$. By deriving the form of $U(x_1,x_2)$ for the case of a homogeneous Poisson process, we next show that the covariance becomes increasingly stationary as the region $\mathcal{C}$ increases in size. We then illustrate how the covariance is decoupled from the lengthscale.

A homogeneous PP yields stationarity. In the case of constant intensity $\lambda(c)=\lambda$ defined over $\mathcal{C}=[C_0,C_1]$, the expression of $U(x_1,x_2)$ can be derived in closed form:

$$U(x_1,x_2) = \frac{1}{\mu(\mathcal{C})}\sqrt{\frac{\pi}{s^2}}\exp\left\{-s^2\left(\frac{x_1-x_2}{2}\right)^2\right\} \left[\Phi((C_1-x_m)\sqrt{2s^2}) - \Phi((C_0-x_m)\sqrt{2s^2}\lambda)\right], \quad (9)$$

where $s^2 = s_0^2\lambda^2$, Φ is the cumulative distribution function of a standard Gaussian, and $x_m = (x_1+x_2)/2$ is the midpoint of the inputs. As the bounded region $\mathcal{C}$ increases, the second term approaches one, and so the covariance of a PoRB-Net approaches a squared exponential kernel with inverse lengthscale $s_0^2\lambda^2$ and amplitude variance $\tilde{\sigma}_w^2 := \sqrt{\pi/s_0^2}$ (defined for convenience):

$$\mathrm{Cov}\left(f(x_1),f(x_2)\right) \approx \sigma_b^2 + \tilde{\sigma}_w^2 \exp\left\{-s_0^2\lambda^2\left(\frac{x_1-x_2}{2}\right)^2\right\}, \quad (10)$$

which is stationary since it only depends on the squared difference between x_1 and x_2. Notice that this result does not rely on an infinite width limit of the network, but only on the Poisson process region $[C_0,C_1]$ being relatively large compared to the midpoint x_m. In practice, $[C_0,C_1]$ can be set larger than the range of observed x values to achieve covariance stationarity over the input domain. Figure 2 shows that over the region $[-5,5]$ the analytical covariance from Equation (9) is fairly constant with only slight drops near the boundaries. In Appendix A we also derive the covariance when $\lambda^2 \sim \mathrm{Gamma}(\alpha_\lambda,\beta_\lambda)$, which results in a qualitatively similar shape. In contrast, the covariance function of an RBFN with a Gaussian prior on the center parameters is not approximately stationary. Specifically, for $c_k \sim \mathcal{N}(0,\sigma_c^2)$ and a fixed scale $s^2 = 1/(2\sigma_s^2)$, Williams (1997) shows that $U(x_1,x_2)$ takes the following form, which Figure 2 shows is highly nonstationary:

$$U(x_1,x_2) \propto \underbrace{\exp\left(-\frac{(x_1-x_2)^2}{2(2\sigma_s^2+\sigma_s^4/\sigma_c^2)}\right)}_{\text{Stationary}} \underbrace{\exp\left(-\frac{x_1^2+x_2^2}{2(2\sigma_c^2+\sigma_s^2)}\right)}_{\text{Nonstationary}}.$$

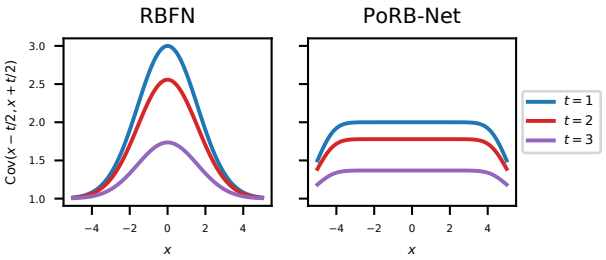


Figure 2: **PoRB-Net captures amplitude stationarity while an RBFN with a Gaussian prior on the centers does not.** The lines are $\mathrm{Cov}(x-t/2, x+t/2)$ for different t. We set all of $\sigma_w^2 = s_0^2 = s^2 = \lambda = 1$ and $\mathcal{C}=[-5,5]$.

Decoupling of variance and lengthscale. From Equation 9, notice the variance is $\mathbb{V}[f(x)] \approx \sigma_b^2 + \tilde{\sigma}_w^2$, which has no dependence on the intensity λ, freeing it to act as an inverse lengthscale. This is a point of differentiation

of PoRB-Nets. If the scale were fixed or independent of the intensity, as is the case in previous priors over RBFNs (e.g., Holmes and Mallick (1998)), the variance would be $\mathbb{V}[f(x)] \approx \sigma_b^2 + \lambda\tilde{\sigma}_w^2$. Intuitively this happens because a higher intensity implies a higher number of basis functions, which implies a higher amplitude variance as the basis functions add up. If we instead allow the scale parameters s^2 to increase as a function of the intensity, thus making the radial basis functions more narrow, we can counteract the impact of their concentration on the amplitude.

To support the hypothesis that the intensity λ controls the lengthscale, we examine the average number of upcrossings of $y = 0$ of sample functions. For a GP with an RBF kernel, the expected number of upcrossings u over the unit interval is inversely related to the lengthscale l via $u = (2\pi l)^{-1}$. Figure 3 shows a histogram of the upcrossings from functions drawn from a PoRB-Net with a stepwise intensity $\lambda(c)$ (greater above $x = 0$). Notice the lengthscale is clearly smaller above $x = 0$ but the amplitude variance $\mathbb{V}[f(x)]$ is approximately constant for all x.

An inhomogeneous PP yields non-stationarity. When the intensity is a non-constant function $\lambda(c)$, then Equation (9) does not hold. However, we find that setting the scale parameter of each hidden unit to $s_k^2 = s_0^2\lambda(c_k)^2$, where $\lambda(c_k)$ is the intensity evaluated at the center parameter c_k, allows for an input dependent lengthscale that is approximately decoupled from the variance.

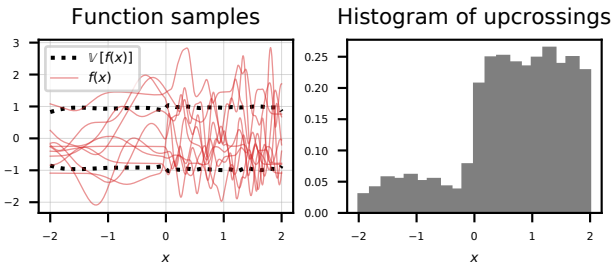


Figure 3: **PoRB-Nets decouple lengthscale (as measured by the upcrossings) and variance**.

6 CONSISTENCY

In this section, we study *consistency of predictions*. That is, as the number of observations goes to infinity, whether the posterior predictive concentrates around the true function. When dealing with priors that can produce an unbounded number of parameters, consistency is a basic but important property. To our knowledge, we are the first to provide consistency for RBFNs with a Poisson distributed number of hidden units (no consistency guarantees were derived by Andrieu et al. (2001)).

Define $r_0(x)$ to be the true regression function and $\hat{r}_n(x) = \mathbb{E}_{\hat{f}_n}[Y \mid X]$ to be the estimated regression function, where $\hat{p}_n$ is the estimated density in parameter space based on n observations. The estimator $\hat{r}_n(x)$ is said to be consistent with respect to the true regression function $r_0(x)$ if, as n tends to infinity:

$$\int (\hat{r}_n(x) - r_0(x))^2 \, dx \xrightarrow{p} 0. \tag{11}$$

Doob's theorem shows that Bayesian models are consistent as long as the prior places positive mass on the true parameter (Miller, 2018). For finite dimensional parameter spaces, one can ensure consistency by simply restricting the set of zero prior probability to have arbitrarily small or zero measure. Unfortunately, in infinite dimensional parameter spaces, this set might be very large (Freedman, 1963). In our case where functions correspond to uncountably infinite sets of parameters, we cannot restrict this set of inconsistency to have measure zero.

Instead, we aim to show a strong form of consistency called Hellinger consistency. We closely follow the approach of Lee (2000), who shows consistency for standard BNNs with normal priors on the parameters. Formally, let $(x_1, y_1), \ldots, (x_n, y_n) \sim p_0$ be the observed data drawn from the ground truth density p_0 and define the Hellinger distance between joint densities p and p_0 over (X, Y) as:

$$D_H(p, p_0) = \sqrt{\iint \left(\sqrt{p(x,y)} - \sqrt{p_0(x,y)}\right)^2 dx\, dy}.$$

The posterior is said to be consistent over Hellinger neighborhoods if for all $\epsilon > 0$,

$$p(\{f : D_H(p, p_0) \leq \epsilon\}) \xrightarrow{p} 1.$$

Lee (2000) shows that Hellinger consistency of joint density functions implies frequentist consistency as described in Equation (11). The following theorem describes an analogous result for PoRB-Nets with homogeneous intensities.

Theorem 1. *(Consistency of PoRB-Nets) A radial basis function network with a homogeneous Poisson process prior on the location of hidden units is Hellinger consistent as the number of observations goes to infinity.*

Proof. Leveraging the results and proof techniques from Lee (2000), we use bracketing entropy from empirical process theory to bound the posterior probability outside Hellinger neighborhoods. We need to check that this model satisfies two key conditions. Informally, the first

condition is that the prior probability placed on parameters larger in absolute value than a bound B_n, where B_n is allowed to grow with the data, is asymptotically bounded *above* by an exponential term $\exp(-nt)$, for some $t > 0$. The second condition is that the prior probability placed on KL neighborhoods of the ground truth density function p_0 is asymptotically bounded *below* by an exponential term $\exp(-n\nu)$, for some $\nu > 0$. The proof is in the Appendix B. □

Note that consistency of predictions does not imply concentration of the posterior in weight space, since radial basis function networks, like other deep neural models, are not identifiable.

7 INFERENCE

We infer the posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$ over the network parameters $\boldsymbol{\theta}$ with Markov-Chain Monte Carlo (MCMC) and model predictions for new observations and their associated uncertainties with the posterior predictive distribution:

$$p(y^\star|x^\star, \mathcal{D}) = \int p(y^\star|x^\star, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}.$$

The inference algorithm can be broken down into three steps. Step 1 updates the network weight, center, and bias parameters $\left(\{w_k, c_k\}_{k=1}^K, b\right)$ conditional on the network width K and intensity function with Hamiltonian Monte-Carlo (HMC) (Neal, 1996). Step 2 updates the network width K conditional on the network parameters and intensity function with birth and death Metropolis-Hastings (MH) steps. Finally, Step 3 updates the Poisson process intensity conditional on the other network parameters and network width. In the case of a homogeneous intensity with a Gamma prior, we use an MH step. In the case of a inhomogeneous intensity defined by Equations 7 and 8 we follow the inference procedure of Adams et al. (2009) for a sigmoidal Gaussian Cox process, treating the current center parameters $\{c_k\}$ as the observed events. This involves introducing three auxiliary variables: a collection of "thinned" center parameters $\{\tilde{c}_m\}$, the number of thinned center parameters M, and the latent GP evaluated at the thinned center parameters $\{\tilde{g}_m\}$. Step 3 requires updating each of these auxiliary variables, along with the latent GP values $\{g_k\}$ evaluated at the current center parameters $\{c_k\}$. For convenience we define $\mathbf{g}_{M+K}$ as vector concatenating $\{\tilde{g}_m\}_{m=1}^M$ and $\{g_k\}_{k=1}^K$ and $\mathbf{c}_{M+K}$ as the vector concatenating $\{\tilde{c}_m\}_{m=1}^M$ and $\{c_k\}_{k=1}^K$. We also define $L(\boldsymbol{\theta})$ as the likelihood of the data given all network parameters. We next describe these steps in more detail assuming a sigmoidal Gaussian Cox process prior on an inhomogeneous intensity $\lambda(c)$, but the full details of the inference procedure are available in the Appendix C.

Step 1: Update network weights, bias, and centers. The full conditional distribution of the weights, bias, and centers can be written as:

$$\begin{aligned} &p(\{w_k\}, b, \{c_k\} \mid K, \{c_m\}, \{\tilde{g}_m\}, \{\tilde{g}_k\}) \\ &\quad \propto L(\boldsymbol{\theta}) \exp\left\{-\frac{1}{2\sigma_b^2} b^2\right\} \exp\left\{-\frac{1}{2\sigma_w^2} \sum_{k=1}^K w_k^2\right\} \\ &\qquad |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2} \mathbf{g}_{M+K}^T \Sigma^{-1} \mathbf{g}_{M+K}\right\}, \end{aligned}$$

where Σ is the kernel matrix of the GP underlying the intensity evaluated at all of the center parameters. We use HMC, which requires tuning L leap-frog steps of size ϵ, to propose updates from this distribution.

Step 2: Update network width K. We adapt the network width with birth or death Metropolis-Hastings (MH) steps chosen with equal probability. For a birth step, we propose a weight w' and a center c' from their prior distributions, and we propose a GP function value g' (representing $g(c')$) from the GP conditioned on the current function values $\mathbf{g}_{M+K}$ observed at $\mathbf{c}_{M+K}$. For the death step, we propose to delete the k'th hidden unit by uniformly selecting among the existing hidden units. Therefore, we can write the hidden unit birth and death proposal densities as follows:

$$\begin{aligned} q(K \to K+1) &\propto \mathcal{N}(w'; 0, \sigma_w^2) \\ &\qquad p(g' \mid c', \mathbf{c}_{M+K}, \mathbf{g}_{M+K}) / \mu(\mathcal{C}) \\ q(K \to K-1) &= 1/K \end{aligned}$$

Note that since the GP has a zero mean function, we propose c' uniformly over $\mu(\mathcal{C})$, but for any fixed intensity we propose from the density $\lambda(c)/\Lambda$. The acceptance rates work out to:

$$\begin{aligned} a_{\text{birth}} &= \frac{L(\boldsymbol{\theta}')}{L(\boldsymbol{\theta})} \frac{\lambda^* \sigma(g') \mu(\mathcal{C})}{K+1} \\ a_{\text{death}} &= \frac{L(\boldsymbol{\theta}')}{L(\boldsymbol{\theta})} \frac{K}{\lambda^* \sigma(g_{k'}) \mu(\mathcal{C})}. \end{aligned}$$

Step 3: Update Poisson process intensity λ. We adopt an inference procedure similar to (Adams et al., 2009) with two crucial differences: the "events" $\{c_k\}$ (center parameters in our case) are unobserved and the full conditional of the function values $\mathbf{g}_{M+K}$ includes the likelihood $L(\boldsymbol{\theta})$ of the data $\mathcal{D}$, since the forward pass of the network uses the posterior mean of g to evaluate the intensity $\lambda(c) = \lambda^* \sigma(g(c))$. We proceed as follows: i) update the number M of thinned centers using birth and death

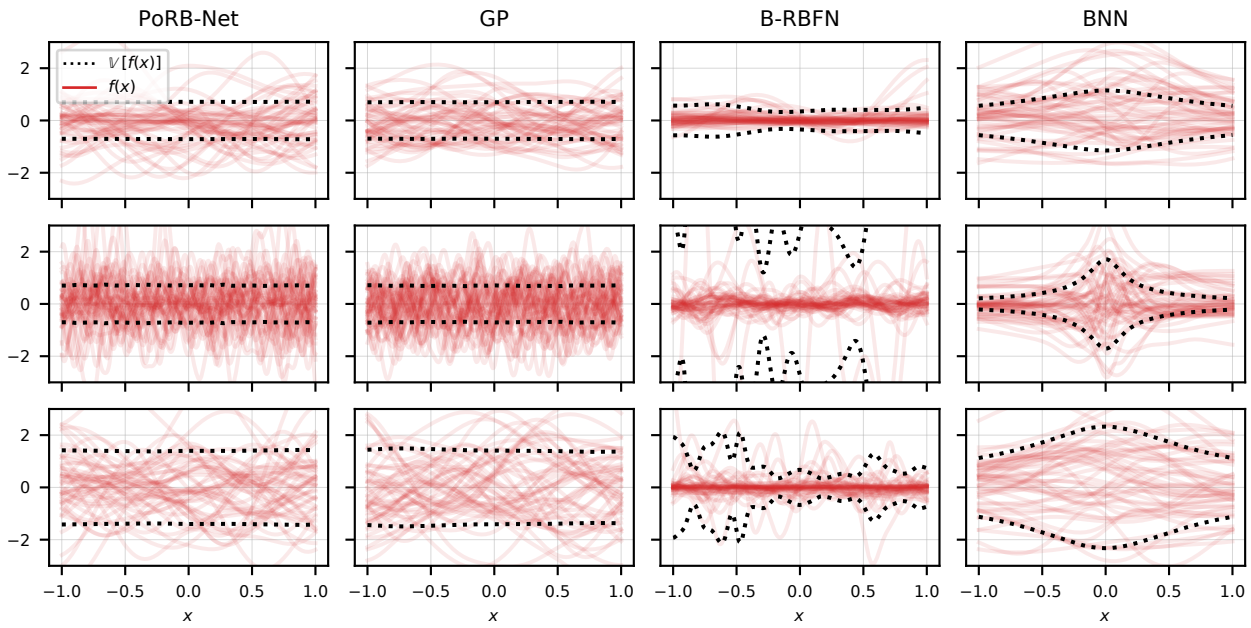


Figure 4: **PoRB-Net allows for easy specification of lengthscale and amplitude like a GP.** We show prior samples from PoRB-Net with a homogeneous intensity, a GP with RBF kernel, B-RBFN (Andrieu et al., 2001), and a BNN (Neal, 1996) with a Gaussian activation. Compared to the first row, the second row has lower lengthscale and similar amplitude, while the third row has higher amplitude and similar lengthscale.

steps, analogous to updating the number of actual centers K; ii) update the thinned center parameters $\{c_m\}_{m=1}^M$ using MH steps with perturbative proposals; iii) update the GP function values $\mathbf{g}_{M+K}$ using HMC.

8 RESULTS

Next we empirically demonstrate desirable properties of PoRB-Net. In particular, PoRB-Net allows for (a) easy specification of lengthscale and amplitude variance information (analogous to a GP), and (b) learning of an input-dependent lengthscale. We report additional empirical results on synthetic and real datasets in Appendix D.

PoRB-Net allows for easy specification of stationary lengthscale and signal variance. Figure 4 shows prior function samples from different models (columns) with different prior settings (rows). Compared to the top row, the second row has a smaller overall lengthscale and the bottom row has a higher overall variance. We plot 50 function samples (red lines) and the estimated variance based on 10,000 function samples (black, dotted line). Like a GP, the amplitude variance of PoRB-Net is constant over the input space and does not depend on the lengthscale. On the other hand, the model of Andrieu et al. (2001) (B-RBFN), which effectively assumes a homogeneous Poisson process prior on the center parameters but does not rescale the basis functions based on the intensity, has a variance that changes over the input space and *does* depend on the lengthscale. For a standard BNN (last column), the amplitude variance and lengthscale are concentrated near the origin and the variance increases as we decrease the lengthscale (from 1st to 2nd row).

PoRB-Net can recover a known, input dependent lengthscale. Figure 5 illustrates the capacity of PoRB-Net to infer an input-dependent lengthscale. Here the true function is a GP with a sinusoidal lengthscale (see kernel in the Appendix D). The right panel shows the center parameter intensity, inferred from noisy (x, y) observations, corresponds to the inverse of the true lengthscale.

PoRB-Nets exhibit competitive performance on synthetic and real datasets. We compare the performance of PoRB-Nets, GPs, and single-layer BNNs with Gaussian activations, with the first two sets of models trained with and without inferring the input dependence of the lengthscale. For the GP models, to use a constant lengthscale we use a regular GP with an RBF kernel; to infer an input dependent lengthscale we use the nonstationary GP model of Heinonen et al. (2016), which we denote by LGP.

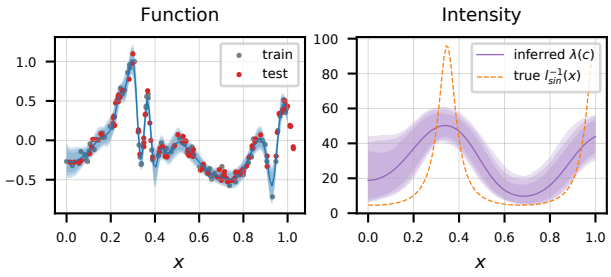


Figure 5: **PoRB-Net is able to learn input-dependent lengthscale information.** The ground truth synthetic example has been generated from a nonstationary GP with a sinusoidal lengthscale function $l_{\sin}(x)$.

At a high level, we see qualitative similarity between PoRB-Nets and GPs that infer the lengthscale, and PoRB-Nets and GPs that do not infer the lengthscale, but the BNNs look different from the rest. This is due to the nonstationarity of the prior, which has higher variability near the origin. All models except the GP are inferred using HMC (including the LGP).

We use four synthetic datasets — all drawn from GPs with known lengthscale functions $l(x)$ — and six real, nonstationary time series datasets – four from mimic (Johnson et al., 2016), the CBOE volatility index over one year starting in October 2018 ("finance"), and the motorcycle dataset (Silverman, 1985). The datasets drawn using a sinusoidal lengthscale $l_{\sin}(x)$ and an increasing lengthscale (from left to right) $l_{\text{inc}}(x)$ can be seen in Figures 5 and 6, respectively. $l_{\text{const}}(x)$ is a constant lengthscale, on which the GP with a stationary, RBF kernel not surprisingly performs best (with PoRB-Net coming in second).

To highlight differences in model behavior rather than prior specification, we first identify the variance and lengthscale parameters that optimize the log marginal likelihood of the GP. We then match the overall variance and lengthscale (as measured by the number of upcrossings mentioned in Section 5) of the BNN and PoRB-Net to the GP by a grid search over the model parameters. Note that the BNN will still have a different input dependence of variance and upcrossings over the input space (both concentrated near the origin). Since adjusting the lengthscale of PoRB-Net adjusts the prior expected number of hidden units, and during inference they can further adapt to the data, we train BNNs with 25, 50, and 100 units, roughly corresponding to the range of units used by PoRB-Net.

There are two main takeaways from these results:

- Examining the posterior predictives in Figure 6 qualitatively, both PoRB-Net and the LGP adapt the local lengthscale to the smoothness of the data, though the effect is more pronounced in the LGP. In contrast, the BNN underestimates uncertainty near $x \approx .2$ in the synthetic dataset (top row) and overestimates uncertainty near $x \approx .8$ in the real dataset (bottom row).

- The test log likelihoods in Table 1 show PoRB-Net exhibits strong performance across the datasets. In contrast, the performance of the BNN varies greatly by the number of hidden units. PoRB-Nets remove this choice by averaging over different numbers of units, fully taking advantage of the Bayesian paradigm.

Table 1: **Test Log Likelihoods.** For the BNN, we show the best(worst) performance among models of size 25, 50, and 100 units.

	PoRB-Net†	PoRB-Net	GP	LGP	BNN
sin*	0.77	**0.82**	0.73	0.81	0.79 (0.74)
inc*	-0.40	0.00	-0.23	**0.18**	-0.15 (-0.28)
inc2*	0.66	**0.75**	0.54	0.18	0.68 (0.63)
const*	0.28	0.33	**0.41**	0.24	0.01 (-0.30)
mimic1	0.89	0.95	0.83	0.90	**1.05** (0.91)
mimic2	0.53	**0.60**	0.56	0.54	0.47 (0.39)
mimic3	-0.63	**-0.57**	-0.67	-0.58	-0.59 (-0.65)
mimic4	-1.72	-1.53	-1.85	-1.44	**-0.59** (-1.38)
finance	-1.41	-0.52	-1.97	**0.03**	-0.73 (-2.63)
motor.	**0.18**	0.16	0.17	0.14	0.16 (0.12)

**synthetic dataset †infers homogeneous intensity*

Test RMSEs, posterior predictives, and inferred intensities for all datasets are available in the Appendix D. Note that HMC is a gold standard for posterior inference; the fact that the standard BNN lacks desirable properties under HMC demonstrates that its failings come from the model and not the inference.

9 CONCLUSION

This work presents a novel Bayesian prior for neural networks called PoRB-Net that allows for easy encoding and inference of two basic functional properties: amplitude variance and lengthscale. We provide a principled inference scheme and future work can address how it can be scaled.

Under standard BNN formulations, we show that it is impossible to get such properties. The essential pieces to achieve these properties were: i) a center-scale parametrization (instead of classical weight-bias), ii) an automatic adaptation of the number of hidden units, and iii) a rescaling of the radial basis functions based on their concentration.

We focused on Gaussian activations because they have

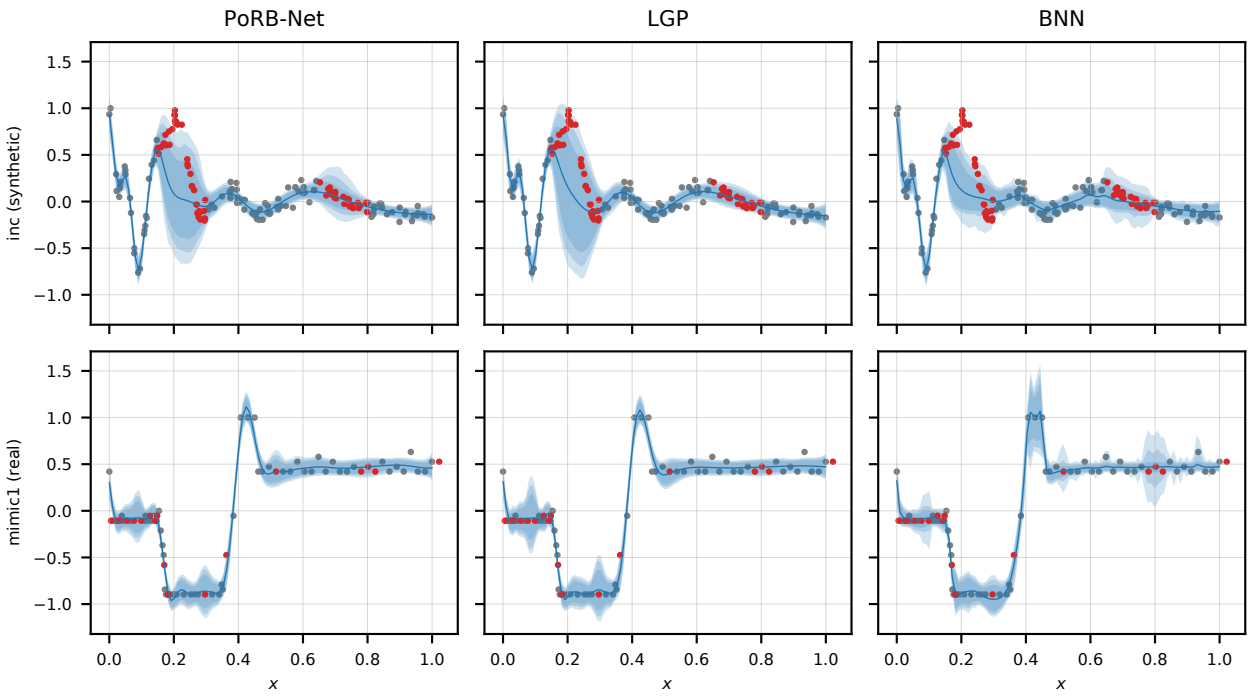


Figure 6: **PoRB-Net posterior predictive captures non-stationary patterns in real scenarios, adapting the lengthscale locally as needed.** Priors for all models have been matched to have about the same amplitude variance and lengthscale. BNNs exhibit undesired uncertainty while PoRB-Nets and LGPs adapt the local uncertainty to the data. Gray points used for training and red points used for testing.

a limited region of effect, unlike other popular activations like tanh or ReLU. Exploring how to get desirable properties for those activations seems challenging, and remains an area for future exploration. That said, we emphasize that RBFNs are commonly used in many practical applications, as surveyed in (Dash et al., 2016).

Finally, all of our work was developed in the context of single-layer networks. From a theoretical perspective this is not an overly restrictive assumption, as single layer networks are still universal function approximators (Park and Sandberg, 1991). However, deep RBFNs, where only the last layer has a radial basis function parameterization, have received renewed interest (Zadeh et al., 2018), so exploring deep PoRB-Nets is an interesting area of future work.

Given the popularity of NNs and the need for uncertainty quantification in them, understanding prior assumptions—which will govern how we will quantify uncertainty—is essential. If prior assumptions are not well understood and not properly specified, the Bayesian framework makes little sense: the posteriors that we find may not be ones that we expect or want. Though we focus on RBFNs, our work provides an important step toward specifying NN priors with desired basic functional properties.

References

Adams, R. P., Murray, I., and MacKay, D. J. C. (2009). Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*.

Andrieu, C., Freitas, N. d., and Doucet, A. (2001). Robust full Bayesian learning for radial basis networks. *Neural Computation*, 13(10):2359–2407.

Anil, C., Lucas, J., and Grosse, R. (2018). Sorting out Lipschitz function approximation. *arXiv:1811.05381 [cs.LG]*.

Asadi, K., Parr, R. E., Konidaris, G. D., and Littman, M. L. (2020). Deep RBF value functions for continuous control. *arXiv:2002.01883 [cs.LG]*.

Barber, D. and Schottky, B. (1998). Radial basis functions: a Bayesian treatment. In *Advances in Neural Information Processing Systems 10*.

Dash, C. S. K., Behera, A. K., Dehuri, S., and Cho, S.-B. (2016). Radial basis function neural networks: a topical state-of-the-art survey. *Open Computer Science*, 6(1).

Flam-Shepherd, D., Requeima, J., and Duvenaud, D. (2017). Mapping Gaussian process priors to Bayesian

neural networks. In *NIPS Bayesian Deep Learning Workshop*.

Freedman, D. A. (1963). On the aymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4):1386–1403.

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S. M. A., and Teh, Y. W. (2018). Neural processes. *arXiv:1807.01622 [cs.LG]*.

Heinonen, M., Mannerström, H., Rousu, J., Kaski, S., and Lähdesmäki, H. (2016). Non-stationary gaussian process regression with hamiltonian monte carlo. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*.

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. *arXiv:1406.2541 [stat.ML]*.

Hinton, G. E. and Neal, R. M. (1995). Bayesian learning for neural networks.

Holmes, C. C. and Mallick, B. K. (1998). Bayesian radial basis functions of variable dimension. *Neural computation*, 10(5):1217–1233.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Lee, H. (2004). *Bayesian nonparametrics via neural networks*. SIAM.

Lee, H. K. (2000). Consistency of posterior distributions for neural networks. *Neural Networks: The Official Journal of the International Neural Network Society*, 13(6):629–642.

Lee, H. K. (2003). A noninformative prior for neural networks. *Machine Learning*, 50(1-2):197–212.

MacKay, D. J. (1992). *Bayesian methods for adaptive models*. PhD Thesis, California Institute of Technology.

Miller, J. W. (2018). A detailed treatment of Doob's theorem. *arXiv:1801.03122 [math.ST]*.

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

Neal, R. M. (1996). Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer.

Park, J. and Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural Computation*, 3(2):246–257.

Pearce, T., Tsuchida, R., Zaki, M., Brintrup, A., and Neely, A. (2019). Expressive priors in Bayesian neural networks: kernel combinations and periodic functions. *arXiv:1905.06076 [stat.ML]*.

Powell, M. J. D. (1987). Algorithms for Approximation. pages 143–167. Clarendon Press.

Que, Q. and Belkin, M. (2016). Back to the future: radial basis function networks revisited. In *Artificial Intelligence and Statistics*, pages 1375–1383.

Robinson, M. (2001). *Priors for Bayesian Neural Networks*. PhD thesis, University of British Columbia.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society*, 47:1–52.

Sun, S., Zhang, G., Shi, J., and Grosse, R. (2019). Functional variational Bayesian neural networks. In *International Conference on Learning Representations*.

Williams, C. K. (1997). Computing with infinite networks. In *Advances in Neural Information Processing Systems 9*.

Zadeh, P. H., Hosseini, R., and Sra, S. (2018). DeepRBF Networks Revisited: Robust Classification with Rejection. *arXiv:1812.03190 [cs.LG]*.

PoRB-Nets: Poisson Process Radial Basis Function Networks (Appendix)

Contents

A COVARIANCE DERIVATION

In this section we derive the covariance function of a PoRB-Net for one dimensional inputs. First we show our model has a prior mean of zero. Note that b, $\{(w_k, c_k)\}_{k=1}^K$, and K are all random variables the scales s_k^2 are fixed as a function of the intensity: $s_k^2 = s_0^2 \lambda(c_k)^2$.

$$\mathbb{E}[f(x)] = \mathbb{E}\left[b + \sum_{k=1}^{K} w_k \phi(s_k(x - c_k))\right] \tag{1}$$

$$= \mathbb{E}[b] + \mathbb{E}\left[\sum_{k=1}^{K} w_k \phi(s_k(x - c_k))\right] \tag{2}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\sum_{k=1}^{K} w_h \phi(s_k(x - c_k)) \mid K = K_0\right]\right] \tag{3}$$

$$= \sum_{K_0=0}^{\infty} \Pr[K = K_0] \mathbb{E}\left[\sum_{k=1}^{K} w_k \phi(s_k(x - c_k)) \mid K = K_0\right] \tag{4}$$

$$= \sum_{K_0=0}^{\infty} \Pr[K = K_0] \sum_{k=1}^{K_0} \mathbb{E}\left[w_k \phi(s_k(x - c_k)) \mid K = K_0\right] \tag{5}$$

$$= \sum_{K_0=0}^{\infty} \Pr[K = K_0] \sum_{k=1}^{K_0} \mathbb{E}\left[w_k \phi(s_k(x - c_k))\right] \tag{6}$$

$$= \sum_{K_0=0}^{\infty} \Pr[K = K_0] K_0 \mathbb{E}\left[w_k \phi(s_k(x - c_k))\right] \tag{7}$$

$$= \mathbb{E}\left[w_k \phi(s_k(x - c_k))\right] \sum_{K_0=0}^{\infty} \Pr[K = K_0] K_0 \tag{8}$$

$$= \underbrace{\mathbb{E}[w_k]}_{0} \mathbb{E}\left[\phi(s_k(x - c_k))\right] \mathbb{E}[K_0] \tag{9}$$

$$= 0 \tag{10}$$

In Equation (6) we drop the condition $K = K_0$ since conditional on the network width K being fixed, the weights w_k are independently normally distributed and the centers are independently distributed according to the normalized intensity $\lambda(c)/\Lambda$, so they do not depend on the actual value of the network width.

Next we consider the covariance:

$$\begin{aligned}
\mathrm{Cov}\left[f(x_1), f(x_2)\right] &= \mathbb{E}\left[f(x_1) f(x_2)\right] \\
&= \mathbb{E}\left[\left(b + \sum_{k=1}^{K} w_k \phi(s_k(x_1 - c_k))\right)\left(b + \sum_{k=1}^{K} w_k \phi(s_k(x_2 - c_k))\right)\right] \qquad (11) \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(b + \sum_{k=1}^{K} w_k \phi(s_k(x_1 - c_k))\right)\left(b + \sum_{k=1}^{K} w_k \phi(s_k(x_2 - c_k)z)\right) \mid K = K_0\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[b^2 + \sum_{k_1=1}^{K}\sum_{k_2=1}^{K} w_{k_1} w_{k_2} \phi(s_{k_1}(x_1 - c_{k_1}))\phi(s_{k_2}(x_2 - c_{k_2})) \mid K = K_0\right]\right] \\
&= \sigma_b^2 + \mathbb{E}\left[\mathbb{E}\left[\sum_{k=1}^{K} w_k^2 \phi(s_k(x_1 - c_k))\phi(s_k(x_2 - c_k)) \mid K = K_0\right]\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[2\sum_{k_1=1}^{K}\sum_{k_2=k_1+1}^{K} w_{k_1} w_{k_2} \phi(s_{k_1}(x_1 - c_{k_1}))\phi(s_{k_2}(x_2 - c_{k_2})) \mid K = K_0\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&= \sigma_b^2 + \mathbb{E}\left[\sum_{k=1}^{K_0} \underbrace{\mathbb{E}\left[w_k^2\right]}_{\tilde{\sigma}_w^2} \mathbb{E}\left[\phi(s_k(x_1 - c_k))\phi(s_k(x_2 - c_k)) \mid K = K_0\right]\right] \\
&\quad + \mathbb{E}\left[2 \sum_{k_1=1}^{K} \sum_{k_2=k_1+1}^{K} \underbrace{\mathbb{E}\left[w_{k_1}\right]}_{0} \underbrace{\mathbb{E}\left[w_{k_2}\right]}_{0} \mathbb{E}\left[\phi(s_{k_1}(x_1 - c_{k_1}))\phi(s_{k_2}(x_2 - c_{k_2})) \mid K = K_0\right]\right] \\
&= \sigma_b^2 + \mathbb{E}\left[\sum_{k=1}^{K} \tilde{\sigma}_w^2 \mathbb{E}\left[\phi(s_k(x_1 - c_k))\phi(s_k(x_2 - c_k)) \mid K = K_0\right]\right] \qquad (12) \\
&= \sigma_b^2 + \mathbb{E}\left[K \tilde{\sigma}_w^2 \mathbb{E}\left[\phi(s(x_1 - c))\phi(s(x_2 - c)) \mid K = K_0\right]\right] \\
&= \sigma_b^2 + \tilde{\sigma}_w^2 \mathbb{E}\left[K_0\right] \underbrace{\mathbb{E}\left[\phi(s(x_1 - c))\phi(s(x_2 - c)) \mid K_0\right]}_{:= U(x_1, x_2)}, \qquad (13)
\end{aligned}
$$

where $\tilde{\sigma}_w^2 = \sqrt{s_0^2/\pi}\sigma_w^2$ is the prior variance for the weights. To actually evaluate the covariance we need to evaluate the $U(x_1, x_2) = \mathbb{E}\left[\phi(s(x_1 - c))\phi(s(x_2 - c))\right]$ term. We next consider two cases. Case 1 is a homogeneous Poisson process prior over c and Case 3 is an inhomogeneous Poisson process prior over c. Note that in both cases. Note that in both cases, the Poisson process prior over c is unconditional on the network width. Conditioned on the network width, as in the expectation we are trying to evaluate, Case 1 is a uniform distribution over $\mathcal{C}$ and Case 3 has PDF $\lambda(c)/\Lambda$.

A.1 CASE 1: HOMOGENEOUS POISSON PROCESS

First we consider the case where the intensity is fixed, i.e., $\lambda(c) = \lambda$, meaning the center parameters are uniformly distributed over $\mathcal{C}$. Then we have:

$$
\begin{aligned}
&U(x_1, x_2) && (14) \\
&= \int_{\mathcal{C}} \phi(s(x_1 - c))\phi(s(x_2 - c)) \frac{1}{\mu(\mathcal{C})}\, dc && (15) \\
&= \int_{\mathcal{C}} \exp\left\{-\frac{1}{2}(s(x_1 - c))^2\right\} \exp\left\{-\frac{1}{2}(s(x_1 - c))^2\right\} \frac{1}{\mu(\mathcal{C})}\, dc && (16) \\
&= \int_{\mathcal{C}} \exp\left\{-\frac{1}{2}s^2[(x_1 - c)^2 + (x_2 - c)^2]\right\} \frac{1}{\mu(\mathcal{C})}\, dc && (17) \\
&= \int_{\mathcal{C}} \exp\left\{-s^2\left[\left(\frac{x_1 - x_2}{2}\right)^2 + \left(\frac{x_1 + x_2}{2} - c\right)^2\right]\right\} \frac{1}{\mu(\mathcal{C})}\, dc && (18) \\
&= \int_{\mathcal{C}} \exp\left\{-s^2\left(\frac{x_1 - x_2}{2}\right)^2\right\} \exp\left\{-s^2\left[\left(\frac{x_1 + x_2}{2} - c\right)^2\right]\right\} \frac{1}{\mu(\mathcal{C})}\, dc && (19) \\
&= \int_{\mathcal{C}} \exp\left\{-s_0^2\lambda^2\left(\frac{x_1 - x_2}{2}\right)^2\right\} \exp\left\{-s_0^2\lambda^2\left[\left(\frac{x_1 + x_2}{2} - c\right)^2\right]\right\} \frac{1}{\mu(\mathcal{C})}\, dc && (20) \\
&= \underbrace{\exp\left\{-s_0^2\lambda^2\left(\frac{x_1 - x_2}{2}\right)^2\right\}}_{\text{SE kernel}} \underbrace{\int_{\mathcal{C}} \exp\left\{-s_0^2\lambda^2\left[\left(\frac{x_1 + x_2}{2} - c\right)^2\right]\right\} \frac{1}{\mu(\mathcal{C})}\, dc}_{\text{uniform mixture of Gaussians}} && (21)
\end{aligned}
$$

In Equation (20) we plug in $s^2 = s_0^2\lambda(c)^2 = s_0^2\lambda^2$. In equation (21) we point out we can write this term as the product of an SE kernel and a mixture of Gaussians. Considering only the uniform mixture of Gaussian term we have:

$$
\begin{aligned}
\int_{\mathcal{C}} \exp\left\{-s_0^2\lambda^2\left[\left(\frac{x_1 + x_2}{2} - c\right)^2\right]\right\} \frac{1}{\mu(\mathcal{C})}\, dc &= \frac{1}{\mu(\mathcal{C})} \int_{\mathcal{C}} \exp\left\{-s_0^2\lambda^2\left[\left(\frac{x_1 + x_2}{2} - c\right)^2\right]\right\} dc && (22) \\
&= \frac{1}{\mu(\mathcal{C})} \int_{C_0}^{C_1} \exp\left\{-\frac{1}{2\psi^2}\left[(x_m - c)^2\right]\right\} dc && (23)
\end{aligned}
$$

$$= \frac{1}{\mu(\mathcal{C})}\psi\sqrt{2\pi}\int_{(C_0-x_m)/\psi}^{(C_1-x_m)\psi}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}u^2\right\}\,du \tag{24}$$

$$= \frac{1}{\mu(\mathcal{C})}\frac{1}{\sqrt{2s_0^2\lambda^2}}\sqrt{2\pi}\left[\Phi((C_1-x_m)/\psi)-\Phi((C_1-x_m)/\psi)\right] \tag{25}$$

$$= \frac{1}{\mu(\mathcal{C})}\sqrt{\frac{\pi}{s_0^2\lambda^2}}\left[\Phi((C_1-x_m)\sqrt{2}s_0\lambda)-\Phi((C_0-x_m)\sqrt{2}s_0\lambda)\right] \tag{26}$$

where Φ is the cumulative distribution function (CDF) of a standard Gaussian. In Equation (23) we define $\psi^2 := 1/(2s_0^2\lambda^2)$ and $x_m := (x_1+x_2)/2$ as the midpoint. In Equation (24) we use the change of variables $u = (c-x_m)/\sigma$. Noting that $\mathbb{E}[K] = \lambda * \mu\mathcal{C}$ and plugging Equation (26) in Equation (21) and Equation (21) into Equation (13) we have:

$$\text{Cov}\left[f(x_1), f(x_2)\right] = \sigma_b^2 + \sigma_w^2\exp\left\{-s_0^2\lambda^2\left(\frac{x_1-x_2}{2}\right)^2\right\}\left[\Phi((C_1-x_m)\sqrt{2}s_0\lambda)-\Phi((C_0-x_m)\sqrt{2}s_0\lambda)\right] \tag{27}$$

This gives a closed form representation for the covariance (to the extent that the standard Gaussian CDF Phi is closed form). If we further assume C_1 and C_0, where $\mathcal{C} = [C_0, C_1]$ is where the Poisson process intensity is defined, are large in absolute value relative to the midpoint x_m. In other words, the Poisson Process is defined over a larger region than the data. Then the difference in error functions is approximately 1 (i.e., the integral over the tails of the Gaussian goes to zero) and the covariance becomes:

$$\text{Cov}\left[f(x_1), f(x_2)\right] \approx \sigma_b^2 + \sigma_w^2\exp\left\{-s_0^2\lambda^2\left(\frac{x_1-x_2}{2}\right)^2\right\} \tag{28}$$

Finally, notice that the variance depends only on the weight and bias variance parameters:

$$\mathbb{V}\left[f(x)\right] \approx \sigma_b^2 + \sigma_w^2 \tag{29}$$

A.2 CASE 2: HOMOGENEOUS POISSON PROCESS WITH GAMMA PRIOR

$$U(x_1, x_2) \tag{30}$$

$$= \iint \phi(s(x_1-c))\phi(s(x_2-c))p(c\mid\lambda)p(\lambda)\,d\lambda\,dc \tag{31}$$

$$= \iint \phi(s(x_1-c))\phi(s(x_2-c))p(c\mid\lambda)p(\lambda)\,d\lambda\,dc \tag{32}$$

$$= \iint \exp\left\{-\frac{1}{2}(s_0^2\lambda^2(x_1-c))^2\right\}\exp\left\{-\frac{1}{2}(s_0^2\lambda^2(x_2-c))^2\right\}\frac{1}{\mu(\mathcal{C})}\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{2(\alpha-1)}e^{-\beta\lambda^2}\,d\lambda\,dc \tag{33}$$

$$= \frac{1}{\mu(\mathcal{C})}\frac{\beta^\alpha}{\Gamma(\alpha)}\iint\lambda^{2(\alpha-1)}\exp\left\{-\lambda^2\left(\underbrace{\frac{1}{2}s_0^2(x_1-c)^2+\frac{1}{2}s_0^2(x-c)+\beta}_{:=\tilde{\beta}(c)}\right)\right\}\,d\lambda\,dc \tag{34}$$

$$= \frac{1}{\mu(\mathcal{C})}\frac{\beta^\alpha}{\Gamma(\alpha)}\iint\lambda^{2(\alpha-1)}\exp\left\{-\lambda^2\tilde{\beta}(c)\right\}\,d\lambda\,dc \tag{35}$$

$$= \frac{\beta^\alpha}{\mu(\mathcal{C})}\int\tilde{\beta}^{-\alpha}(c)\,dc \tag{36}$$

In Equation 36 we recognize the form of the Gamma probability density function to solve the inner integral. We now rewrite $\tilde{\beta}(c)$ as:

$$\tilde{\beta}(c) := \frac{1}{2}s_0^2(x_1-c)^2+\frac{1}{2}s_0^2(x-c)+\beta \tag{37}$$

$$= s_0^2\left(c^2 - 2c\left(\underbrace{\frac{x_1+x_2}{2}}_{:=x_m}\right)\right) + \frac{1}{2}\left(x_1^2 + x_2^2\right) + \beta \tag{38}$$

$$= \underbrace{s_0^2(c-x_m)^2}_{:=u^2} + \underbrace{s_0^2\left(\frac{x_1-x_2}{2}\right)^2 + \beta}_{:=r^2} \tag{39}$$

$$= u^2 + r^2 \tag{40}$$

where we define $x_m := (x_1+x_2)/2$ as the midpoint of x_1 and x_2, and $u = s_0^2(c-x_m)^2$, and $r^2 := s_0^2\left((x_1-x_2)/2\right)^2 + \beta$ to simplify the notation. Using this expression for $\tilde{\beta}(c)$, the integral in Equation 36 becomes:

$$\int \tilde{\beta}(c)^{-\alpha}\, dc = \int_{u_0}^{u_1} (u^2+r^2)^{-\alpha}\, du \tag{41}$$

$$= ur^{-2\alpha}{}_2F_1\left(\frac{1}{2}, \alpha; \frac{3}{2}; -\frac{u^2}{r^2}\right)\Bigg|_{U_0}^{U_1} \tag{42}$$

where $U_0 = s_0^2(C_0 - x_m)$, $U_1 = s_0^2(C_1 - x_m)$ and the hypergeometric function ${}_2F_1$ is defined by:

$${}_2F_1(a,b;c;z) = \sum_{n=0}^{\infty} \frac{(a)_n(b)_n}{(c)_n}\frac{z^n}{n!} \tag{43}$$

where:

$$(q)_n = \begin{cases} 1 & n = 0 \\ q(q+1)\cdots(q+n-1) & n > 0 \end{cases} \tag{44}$$

Plugging the expression for $\int b(c)\, dc$ back into Equation 36 we have:

$$V(x_1,x_2) = \frac{1}{\mu(\mathcal{C})}\left(\frac{\beta}{r^2}\right)^{-\alpha}\left[\frac{{}_2F_1\left(\frac{1}{2},\alpha;\frac{3}{2};-\frac{s_0^2(C_0-x)^2}{r^2}\right)}{(x-C_0)} + \frac{{}_2F_1\left(\frac{1}{2},\alpha;\frac{3}{2};-\frac{s_0^2(C_1-x)^2}{r^2}\right)}{(C_1-x)}\right] \tag{45}$$

Before plugging the expression for $V(x_1,x_2)$ into Equation 13, notice we can write the expected number of units $\mathbb{E}[K]$ as the product of the expected intensity $\mathbb{E}[\lambda]$ and the volume of the Poisson process region $\mu(\mathcal{C})$:

$$\mathbb{E}[K] = \mathbb{E}\left[\mathbb{E}[K \mid \lambda] \mid \lambda\right] = \mathbb{E}[\lambda\mu(\mathcal{C}) \mid \lambda] = \mu(\mathcal{C})\mathbb{E}[\lambda] \tag{46}$$

Therefore, the covariance is:

$$\text{Cov}\left[f(x_1), f(x_2)\right] = \sigma_b^2 + \tilde{\sigma}_w^2\mathbb{E}[\lambda]\left(\frac{\beta}{r^2}\right)^{-\alpha}\left[\frac{{}_2F_1\left(\frac{1}{2},\alpha;\frac{3}{2};-\frac{s_0^2(C_0-x)^2}{r^2}\right)}{(x-C_0)} + \frac{{}_2F_1\left(\frac{1}{2},\alpha;\frac{3}{2};-\frac{s_0^2(C_1-x)^2}{r^2}\right)}{(C_1-x)}\right] \tag{47}$$

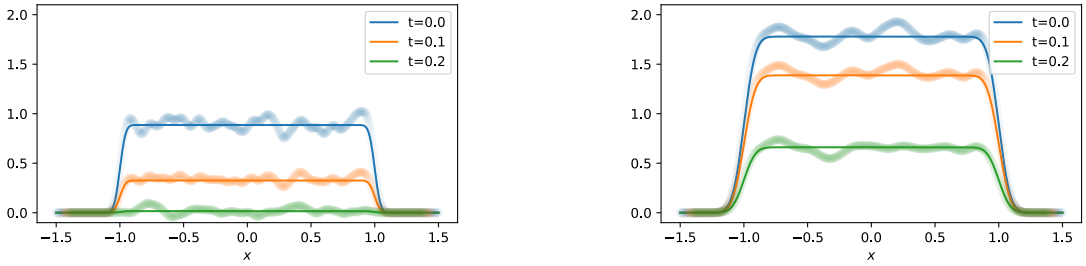
where $r^2 = s_0^2\left((x_1-x_2)/2\right)^2 + \beta$. Notice the variance simplifies to:

$$\text{Var}\left[f(x)\right] = \sigma_b^2 + \tilde{\sigma}_w^2\mathbb{E}[\lambda]\left[\frac{{}_2F_1\left(\frac{1}{2},\alpha;\frac{3}{2};-\frac{s_0^2(C_0-x)^2}{\beta}\right)}{(x-C_0)} + \frac{{}_2F_1\left(\frac{1}{2},\alpha;\frac{3}{2};-\frac{s_0^2(C_1-x)^2}{\beta}\right)}{(C_1-x)}\right] \tag{48}$$

Note that $\mathbb{E}[\lambda] \approx \sqrt{\alpha/\beta}$. This is because $\lambda^2 \sim \text{Gamma}(\alpha,\beta)$ implies $\lambda \sim \text{Nakagami}(m=\alpha, \Omega=\alpha/\beta)$. Using the approximation $\Gamma(\alpha+\frac{1}{2})/\Gamma(\alpha) \approx \sqrt{\alpha}$ (follows from Sterling's formula) we have:

$$\mathbb{E}[\lambda] = \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)}\beta^{-1/2} \approx \sqrt{\frac{\alpha}{\beta}} \tag{49}$$

Figure 1 plots the exact functional covariance $\text{Cov}(f(x - t/2), f(x + t/2))$ with and without the Gamma prior on the intensity (given by Equations 27 and 47, respectively) as a function of input x for different values of a fixed separation t (so $t = 0$ corresponds to the variance). Also shown are empirical estimates based on 1000 samples drawn from the prior. The covariance drops of sharply near the boundaries of $\mathcal{C} = [-1, 1]$ but is approximately constant within this region.



(a) Uniform intensity

(b) Uniform intensity with Gamma prior

Figure 1: True (solid line) and estimated (dots) functional covariance $\text{Cov}(f(x - t/2), f(x + t/2))$ of PoRB-Net with a uniform intensity with and without a Gamma prior. For both models we set $s_0^2 = 1$, $\sigma_w^2 = 1$, $\sigma_b^2 = 0$, $\mathcal{C} = [-1, 1]$, and $\mathbb{E}[K] = 20$.

B CONSISTENCY

We are interested in the posterior behavior of our model as the number of observations $n \to \infty$. For better comparison with existing results, we use slightly different notation in this section. We want to show that the estimated regression function $\hat{g}_n(x) := \mathbb{E}[Y \mid X = x]$ is asymptotically consistent for the true regression function $g_0(x)$, i.e.:

$$\int (\hat{g}_n(x) - g_0(x))^2 dx \xrightarrow{p} 0$$

To do this, we first show that the posterior probability assigned to all joint distribution functions $f(X, Y)$ in any Hellinger neighborhood of the true joint distribution function $f_0(X, Y)$ approaches one as $n \to \infty$. That is, if $A_\epsilon = \{f \mid D_H(f, f_0) \leq \epsilon\}$ defines a Hellinger neighborhood of the true distribution function, then $\forall \epsilon > 0$:

$$p(A_\epsilon \mid (X_1, Y_1), \ldots, (X_n, Y_n)) \xrightarrow{p} 1$$

We assume that the marginal distribution of X is uniform on $[0, 1]$ (i.e., $f(X) = 1$), so the joint distribution $f(X, Y)$ and the conditional distribution $f(Y \mid X)$ are the same, since $f(X, Y) = f(Y \mid X)f(X) = f(Y \mid X)$. The estimated regression function is defined as $\hat{g}_n(x) = \mathbb{E}_{\hat{f}_n}[Y \mid X = x]$, where $\hat{f}_n$ is given by the posterior predictive density:

$$\hat{f}_n(X, Y) = \int f(X, Y)\, dP(f \mid (X_1, Y_1)), \ldots, (X_n, Y_n)).$$

After introducing a few definitions and notation, Section B.1 discusses the necessary conditions on the prior required for any radial basis function network to achieve consistency, with many results taken or adapted from [Lee, 2000]. Section B.2 checks that these necessary conditions are met by PoRB-Net with a homogeneous Poisson process prior on the number of hidden units. We first show asymptotic consistency when the number of hidden units is allowed to grow with the data. This gives a sequence of models known as a sieve. We then extend this to the case when the number of hidden units is inferred.

Definitions and notation

We begin by specifying our notation and definitions, which differs from other sections in this paper.

- D is the input dimension.
- K is the network width.
- I, $I^{(w)}$ and $I^{(c)}$ are the number of total parameters, weight parameters, and center parameters, respectively. $I = I^{(w)} + I^{(c)} + 1$.
- $\mathcal{I}$, $\mathcal{I}^{(w)}$, $\mathcal{I}^{(c)}$, and $\mathcal{I}^{(\lambda^2)}$ are the index set of total parameters, weight parameters, center parameters, and intensity respectively (e.g., $\mathcal{I} = 1, 2, \ldots, I$). $\mathcal{I}^{(w)} \subset \mathcal{I}$, $\mathcal{I}^{(c)} \subset \mathcal{I}$, $\mathcal{I}^{(\lambda^2)} \subset \mathcal{I}$, $I = |\mathcal{I}|$, $I^{(w)} = |\mathcal{I}^{(w)}|$, $I^{(c)} = |\mathcal{I}^{(c)}|$, and $1 = |\mathcal{I}^{(\lambda^2)}|$.
- The subscript n always denotes the sample size dependence (applies to K_n, I_n, $\mathcal{I}_n$, $I_n^{(w)}$, $\mathcal{I}_n^{(w)}$, $I_n^{(c)}$, $\mathcal{I}_n^{(c)}$, C_n).
- Let θ_i denote any parameter, c_i denote a center parameter, and w_i denote a weight parameter.
- C_n is a bound on the absolute value of the parameters. For the sieves approach in we assume $C_n \leq \exp(n^{b-a})$, where $0 < a < b < 1$.
- Assume that the Poisson process intensity function $\lambda(c)$ is only defined on a bounded region $\mathcal{C}$.
- Let $f(x, y)$ denote a joint density of covariates X and label Y and let $g(x) = \mathbb{E}[Y \mid X = x]$ denote a regression function.
- Let $f_0(x, y)$ and $g_0(x)$ denote the true joint density and regression function, respectively.

- We assume $x \in \mathcal{X} = [0,1]^D$ and that the marginal density of x is uniform, i.e. $f(X) = 1$.
- Let $D_H(f_0, f)$ denote the Hellinger distance and let $A_\epsilon = \{f : D_H(f_0, f) \leq \epsilon\}$.
- Let $D_K(f_0, f)$ denote the KL divergence and let K_γ denote a KL neighborhood of the true joint density: $K_\gamma = \{f \mid D_K(f_0, f) \leq \gamma\} = \{f : D_K(f_0, f) \leq \gamma\}$.
- Let $(x_1, y_1), \ldots, (x_n, y_n)$ denote the n observations and π_n denote a prior probability distribution over the parameters of a single hidden layer PoRB-Net conditional on there being K_n nodes, where K_n increases with n. Let I_n denote the number of parameters for an RBFN network with K_n nodes.
- Let $\mathcal{F}$ denote the space of all single-layer radial basis function networks RBFN$(x; \boldsymbol{\theta}) \mapsto y$, let $\mathcal{F}_n \subset \mathcal{F}$ be its restriction to networks with parameters less than $C_n > 0$ in absolute value, where C_n also increases with n; let $\mathcal{H}_n \subset \mathcal{F}$ be its restriction to networks with K_n nodes; and let $\mathcal{G}_n = \mathcal{F}_n \cap \mathcal{G}_n$ be the intersection of both restrictions.

B.1 CONSISTENCY OF RBFNs WITH ARBITRARY PRIORS

B.1.1 Supporting results

The following theorems are used in proof of Lemma 2, which is adapted from [Lee, 2000]. Theorem 1 upper bounds the bracketing number $N_{[]}(\,)$ by the covering number $N(\,)$. Define the Hellinger bracketing entropy by $H_{[]}(\,) := \log N_{[]}(\,)$.

Theorem 1. *[van der Vaart and Wellner, 1996] Let $s, t \in \mathcal{F}_n$, i.e., s and t are realizations of the parameter vector. Let $f_t(x, y) \in \mathcal{F}^*$ be a function of x and y with parameter vector equal to t. Suppose that:*

$$|f_t(x, y) - f_s(x, y)| \leq d^*(s, t) F(x, y) \tag{50}$$

for some metric d^*, for some fixed function F, and for every s, t, and every (x, y). Then for any norm $\|\cdot\|$,

$$N_{[]}(2\epsilon \|F\|, \mathcal{F}^*, \|\cdot\|) \leq N(\epsilon, \mathcal{F}_n, d^*). \tag{51}$$

Theorem 2. *[Wong and Shen, 1995] Define the ratio of joint likelihoods between the inferred density and the true density as*

$$R_n(f) = \prod_{i=1}^{n} \frac{f(x_i, y_i)}{f_0(x_i, y_i)}. \tag{52}$$

For any $\epsilon > 0$ there exists constants a_1, a_2, a_3, a_4 such that if

$$\int_{\epsilon^2/2^8}^{\sqrt{\epsilon}} \sqrt{H_{[]}(u/a_3)}\, du \leq 2a_4\sqrt{n}\epsilon^2, \tag{53}$$

then

$$P^*\left(\sup_{f \in A_\epsilon^c \cap \mathcal{F}_n} R_n(f) \geq \exp(-a_1 n\epsilon^2)\right) \leq 4\exp(-a_2 n\epsilon^2). \tag{54}$$

Lemma 1. *(Adaptation of Lemma 1 in Lee [2000])*[1] *Suppose that $H_{[]}(u) \leq \log[(a' n^a C_n^{a''} I_n/u)^{I_n}]$, where $I_n = (D+1)K_n + 1$, $K_n \leq n^a$, $a', a'' > 0$, and $C_n \leq \exp(n^{b-a})$ for $0 < a < b < 1$. Then for any fixed constants $a''', \epsilon > 0$ and for all sufficiently large n,*

$$\int_0^{\epsilon} \sqrt{H_{[]}(u)} \leq c\sqrt{n}\epsilon^2. \tag{55}$$

Proof. Let $a_n = a' n^a C_n^{a''} I_n$, so $H_{[]}(u) \leq \log[(a_n/u)^{I_n}] = I_n \log(a_n/u)$. Taking the square root and integrating each side, we have:

$$\int_0^{\epsilon} \sqrt{H_{[]}(u)}\, du = \int_0^{\epsilon} \sqrt{I_n \log(a_n/u)}\, du \tag{56}$$

[1] This lemma differs from [Lee, 2000] because they assume $H_{[]}(u) \leq \log[(C_n^2 I_n/u)^{I_n}]$ and $I_n = (D+2)K_n + 1$.

$$= \sqrt{I_n/2} \int_0^\epsilon \sqrt{2\log(a_n/u)}\, du \tag{57}$$

$$= \sqrt{I_n/2} \int_0^\epsilon z\, du, \tag{58}$$

where we define the substitution $z := \sqrt{2\log(a_n/u)}$. Then:

$$dv = \frac{1}{2}(2\log(a_n/u))^{-1/2}(2)\frac{(-a_n/u^2)}{a_n/u}\, dz = -z^{-1}u^{-1}\, du \tag{59}$$

$$\implies du = -zu\, dz = -a_n z u/a_n\, dz = -a_n z \exp\left(-\frac{1}{2}\underbrace{2\log(a_n/u)}_{z^2}\right)\, dz = -a_n z \exp(-z^2/2)\, dz. \tag{60}$$

Thus:

$$\int_0^\epsilon \sqrt{H_{[]}(u)}\, du \leq -\sqrt{I_n/2}\int_\infty^{z_\epsilon} a_n z^2 \exp(-v^2/2)\, dz \tag{61}$$

$$= a_n\sqrt{I_n/2}\int_{z_\epsilon}^\infty z^2 \exp(-v^2/2)\, dz \tag{62}$$

where we define $z_\epsilon = \sqrt{2\log(a_n/\epsilon)}$. Next, integrate by parts (using $u = z$ and $dv = z\exp(-z^2/2)\, dz$), giving:

$$\int_0^\epsilon \sqrt{H_{[]}(u)}\, du = a_n\sqrt{I_n/2}\left[-z\, \exp(-z^2/2)\big|_{z_\epsilon}^\infty + \int_{z_\epsilon}^\infty \exp(-z^2/2)\, dz\right] \tag{63}$$

$$= a_n\sqrt{I_n/2}\left[z_\epsilon \exp(-z_\epsilon^2/2) + \sqrt{2\pi}\int_{z_\epsilon}^\infty \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)\, dz\right] \tag{64}$$

$$\leq a_n\sqrt{I_n/2}\left[z_\epsilon \exp(-z_\epsilon^2/2) + \sqrt{2\pi}\frac{\phi(z_\epsilon)}{z_\epsilon}\right] \qquad \text{Mill's Ratio} \tag{65}$$

$$= a_n\sqrt{I_n/2}z_\epsilon\left[\exp(-z_\epsilon^2/2) + \sqrt{2\pi}\frac{\frac{1}{\sqrt{2\pi}}\exp(-z_\epsilon^2/2)}{z_\epsilon^2}\right] \tag{66}$$

$$= a_n\sqrt{I_n/2}z_\epsilon \exp(-z_\epsilon^2/2)\left[1 + \frac{1}{z_\epsilon^2}\right] \tag{67}$$

$$= a_n\sqrt{I_n/2}z_\epsilon \underbrace{\exp(-z_\epsilon^2/2)}_{\epsilon/a_n}\left[1 + \frac{1}{z_\epsilon^2}\right] \tag{68}$$

$$= \epsilon\sqrt{I_n/2}z_\epsilon\left[1 + \frac{1}{z_\epsilon^2}\right]. \tag{69}$$

Since $a_n \to \infty$ as $n \to \infty$, we have $z_\epsilon^2 = 2\log(a_n/\epsilon) \to \infty$ as well, so $[1 + 1/z_\epsilon^2] \leq 2$ for large n. Continuing:

$$\int_0^\epsilon \sqrt{H_{[]}(u)}\, du \leq \epsilon\sqrt{I_n/2}z_\epsilon \tag{70}$$

$$= \epsilon\sqrt{I_n/2}\sqrt{2\log(a_n/\epsilon)} \tag{71}$$

$$= \epsilon\sqrt{I_n}\sqrt{\log(a_n/\epsilon)} \tag{72}$$

$$\leq \epsilon\sqrt{I_n}\sqrt{\log(a' n^a C_n^{a''} I_n/\epsilon)} \tag{73}$$

$$\leq \epsilon\sqrt{I_n}\sqrt{\log(a') + a\log(n) + a''\log(C_n) + \log(I_n) - \log(\epsilon)} \tag{74}$$

$$\leq \epsilon\sqrt{(D+1)n^a + 1}\sqrt{\log(a') + a\log(n) + a'' n^{b-a} + \log((D+1)n^a + 1) - \log(\epsilon)} \tag{75}$$

where we plug in $I_n = (D+1)K_n + 1 \leq (D+1)n^a + 1$ and $C_n = \exp(n^{b-a})$.

Since $0 < a < b < 1$, there exists a γ such that $a < \gamma < b$ and $b - a < 1\gamma$. This follows from the fact that since $0 < a < b < 1$, there must exist a $\delta > 0$ such that $a + \delta < b$ and $b + \delta < 1$. Now let $\gamma = a\delta$ to see that $b - a = b + \delta - (a + \delta) < 1 - (a + \delta) = 1 - \gamma$. Multiplying by $1/\sqrt{n} = \sqrt{n^{-\gamma}}\sqrt{n^{-(1-\gamma)}}$ on each side:

$$\frac{1}{\sqrt{n}} \int_0^\epsilon \sqrt{H_{[]}(u)}\, du \leq \epsilon \sqrt{n^{-\gamma}} \sqrt{(D+1)n^a + 1} \tag{76}$$

$$\sqrt{n^{-(1-\gamma)}} \sqrt{\log(a'/\epsilon) + a\log(n) + a'' n^{b-a} + \log((D+1)n^a + 1)} \tag{77}$$

$$= \epsilon \sqrt{(D+1)n^{-(\gamma-a)} + n^{-\gamma}} \tag{78}$$

$$\sqrt{n^{-(1-\gamma)}\log(a'\epsilon) + an^{-(1-\gamma)}\log(n) + a'' n^{-((1-\gamma)-(b-a))} + n^{-(1-\gamma)}\log((D+1)n^a + 1)} \tag{79}$$

$$\to \infty \text{ as } n \to \infty \tag{80}$$

since each of γ, $1 - \gamma$, $\gamma - a$, and $(1 - \gamma) - (b - a)$ are positive. Thus, for any $a''', \epsilon > 0$

$$\frac{1}{\sqrt{n}} \int_0^\epsilon \sqrt{H_{[]}(u)}\, du \leq a''' \epsilon^2 \tag{81}$$

□

Lemma 2. *(Adaptation of Lemma 2 in [Lee, 2000] (same statement but particularized for RBFNs)) Define the ratio of joint likelihoods between the inferred density and the true density as*

$$R_n(f) = \prod_{i=1}^n \frac{f(x_i, y_i)}{f_0(x_i, y_i)}. \tag{82}$$

Under the assumptions of Lemma 1,

$$\sup_{f \in A_\epsilon^c \cap \mathcal{F}_n} R_n(f) \leq 4\exp(-a_2 n\epsilon^2) \tag{83}$$

almost surely for sufficiently large n, where a_2 is the constant from Theorem 2.

Proof. Much of this proof is reproduced exactly as in Lemma 2 in [Lee, 2000], with only a few adaptations that we mention along the way. We first bound the Hellinger bracketing entropy using Theorem 1 and then use Lemma 1 to show the conditions of Theorem 2.

Since we are interested in computing the Hellinger bracketing entropy for neural networks, we need to use the L_2 norm on the square roots of the density function, f. Later, we compute the L_∞ covering number of the parameter space, so here $d^* = L_\infty$. We would like to apply Theorem 1 particularized for the L_2 norm, i.e., $|\sqrt{f_t(x,y)} - \sqrt{f_s(x,y)}| \leq d^*(s,y)F(x,y)$ for some F then $N_{[]}(2\epsilon \|F\|_2, \mathcal{F}^*, \|\cdot\|_2) \leq N(\epsilon, \mathcal{F}_n, d^*)$. To show that the condition holds true, apply the Fundamental Theorem of Integral Calculus. For particular vectors s and t, let $g(u) = \sqrt{f_{(1-u)s+ut}(x,y)}$. Let $v_i = (1-u)s_i + ut_i$ and denote the space of θ by Θ_i.

$$|\sqrt{f_t(x,y)} - \sqrt{f_s(x,y)}| = \int_0^1 \frac{g}{du}\, du \tag{84}$$

$$= \int_0^1 \sum_{i=1}^I \frac{\partial g}{\partial \theta_i} \frac{\partial \theta_i}{\partial u}\, du \tag{85}$$

$$= \sum_{i=1}^I (t_i - s_i) \int_0^1 \frac{\partial g}{\partial \theta_i}\, du \tag{86}$$

$$\leq \sum_{i=1}^I \sup_i |t_i - s_i| \int_0^1 \sup_{\theta_i \in \Theta_i} \left| \frac{\partial g}{\partial \theta_i} \right| du \tag{87}$$

$$= \sup_i |t_i - s_i| \sum_{i=1}^{I} \sup_{\theta_i \in \Theta_i} \left| \frac{\partial g}{\partial \theta_i} \right| \int_0^1 du \tag{88}$$

$$\leq \sup_i |t_i - s_i| I \sup_i \left[\sup_{\theta_i \in \Theta_i} \left| \frac{\partial g}{\partial \theta_i} \right| \right] \tag{89}$$

$$= \|t - s\|_\infty F(x, y) \tag{90}$$

where $F(x,y) = I \sup_i [\sup_{\theta_i \in \Theta_i} |\partial g / \partial \theta_i|]$. Here $\partial g / \partial \theta_i$ is the partial derivative of $\sqrt{f}$ with respect to the ith parameter. Recall that $f(x,y) = f(y \mid x) f(x)$, where $f(x) = 1$ since $X \sim U[0,1]$ and $f(y \mid x)$ is normal with mean determined by the neural network and variance 1.

So far, this proof follows Lemma 2 in [Lee, 2000] exactly. Now we make a slight modification for an RBFN model. By Lemma 3, $|\partial g / \partial \theta_i| \leq (8\pi e^2)^{-1/4} 2 n^a C_n^3 = n^a C_n^3 / 2$, where $a' := 4(8\pi e^2)^{-1/4}$. Then set $F(x,y) = a' n^a C_n^3 I / 2$, so $||F||_2 = a' n^a C_n^3 I / 2$. Applying Theorem 1 to bound the bracketing number by the covering number we have:

$$N_{[]}(u, \mathcal{F}^*, ||\cdot||_2) = N_{[]} \left(2 \left(\frac{u}{2||F||_2} \right) ||F||_2, \mathcal{F}^*, ||\cdot||_2 \right) \tag{91}$$

$$\leq N \left(\frac{u}{2||F||_2}, \mathcal{F}^*, ||\cdot||_2 \right) \tag{92}$$

Notice that the covering number of $\mathcal{F}_n$ is clearly less than $((2C_n)/(2\epsilon) + 1)^I$. So, for any $\eta > 0$, we have:

$$N(\eta, \mathcal{F}^*, L_\infty) \leq \left(\frac{2C_n}{2\eta} + 1 \right)^I = \left(\frac{C_n + \eta}{\eta} \right)^I \leq \left(\frac{C_n + 1}{\eta} \right)^I. \tag{93}$$

Therefore,

$$N_{[]}(u, \mathcal{F}^*, ||\cdot||_2) \leq \left(\frac{C_n + 1}{\frac{u}{2||F||_2}} \right)^I \tag{94}$$

$$= \left(\frac{2||F||_2 (C_n + 1)}{u} \right)^I \tag{95}$$

$$= \left(\frac{a' n^a C_n^3 I_n (C_n + 1)}{u} \right)^I \tag{96}$$

$$= \left(\frac{a' n^a \tilde{C}_n^4 I_n}{u} \right)^I \tag{97}$$

where $\tilde{C}_n = C_n + 1$. For notational convenience, we drop $\mathcal{F}^*$ and $||\cdot||_2$ going forward. Taking the logarithm:

$$H_{[]}(u) \leq \log[(a' n^a C_n^{a''} I_n / u)^I]. \tag{98}$$

The bound above holds for a fixed network size, but we can now let K_n grow such that $K_n \leq n^a$ for any $0 < a < 1$. Thus by Lemma 1, we have:

$$\frac{1}{\sqrt{n}} \int_0^\epsilon \sqrt{H_{[]}(u)}\, du \leq a''' \epsilon^2, \tag{99}$$

which shows the conditions of Lemma 1. Therefore, we have that for any $a''', \epsilon > 0$,

$$\int_0^\epsilon \sqrt{H_{[]}(u)}\, du \leq a''' \sqrt{n} \epsilon^2, \tag{100}$$

With an eye on applying Theorem 2, notice that $\int_{\epsilon^2/2^8}^{\epsilon} \sqrt{H_{[]}(u)}\, du < \int_0^\epsilon \sqrt{H_{[]}(u)}\, du$. Substituting $\sqrt{2}\epsilon$ for ϵ, we get

$$\int_{\epsilon^2/2^8}^{\sqrt{\epsilon}} \sqrt{H_{[]}(u)}\, du \leq 2 a''' \sqrt{n} \epsilon^2, \tag{101}$$

letting $a_3 = 1$ and $a_4 := 2a'''$, where a_3 and a_4 are the constants required by Theorem 2. This gives the necessary conditions for Theorem 2, which implies that

$$P^* \left(\sup_{f \in A_\epsilon^c \cap \mathcal{F}_n} R_n(f) \geq \exp(-a_1 n \epsilon^2) \right) \leq 4 \exp(-a_2 n \epsilon^2). \tag{102}$$

Now apply the first Borel-Cantelli Lemma to get the desired result. □

B.1.2 Main theorems

The following theorem is proved by Lee [2000] for single-layer feedforward networks with a logistic activation and Gaussian priors. With a few modifications to the proof as described below, it can be applied to RBFNs. Here, the number of units is allowed to grow with the number of observations but it is not inferred from the data. We call this a sieves approach.

Theorem 3. *(Consistency when width grows with data (sieves approach)) [Lee, 2000] Suppose the following conditions hold:*

(i) There exists an $r > 0$ and an $N_1 \in \mathbb{N}$ such that $\forall n \geq N_1$, $\pi_n(\mathcal{F}_n^c) < \exp(-nr)$.

(ii) For all $\gamma > 0$ and $\nu > 0$, there exists an $N_2 \in \mathbb{N}$ such that $\forall n \geq N_2$, $\pi_n(K_\gamma) \geq \exp(-n\nu)$.

Then $\forall \epsilon > 0$, the posterior is asymptotically consistent for f_0 over Hellinger neighborhoods, i.e.:

$$P(A_\epsilon \mid (x_1, y_1), \ldots, (x_n, y_n)) \xrightarrow{p} 1. \tag{103}$$

Proof. Lee [2000] proves this result for single-layer feedforward networks with a logistic activation and Gaussian priors (Theorem 1 in their paper). Their proof relies on their Lemmas 3 and 5. Their Lemma 5 needs no adaptation for RBFNs but their Lemma 3 depends on their Lemma 2, which does need adaptation for RBFNs. Above we proved their Lemma 2 for RBFNs, which we call Lemma 1. Thus their Lemma 3 holds, so their Theorem 1 holds, which gives the results of this theorem. □

Lee [2000] shows that Hellinger consistency gives asymptotic consistency.

Corollary B.1. *(Hellinger consistency gives asymptotic consistency for sieves prior) [Lee, 2000] Under the conditions of Theorem 3, $\hat{g}_n$ is asymptotically consistent for g_0, i.e.:*

$$\int (\hat{g}_n(x) - g_0(x))^2 dx \xrightarrow{p} 0. \tag{104}$$

The following is an extension of Theorem 3 to when there is a prior over the number of units. The proof in [Lee, 2000] assumes a feedforward network with a logistic activation and Gaussian priors, but these assumptions are not used beyond their use in applying Theorem 3. Since we adapt Theorem 3 to our model, the proof of the following Theorem 4 needs no additional adaptation.

Theorem 4. *(Consistency for prior on width) [Lee, 2000] Suppose the following conditions hold:*

(i) For each $i = 1, 2, \ldots$ there exists a real number $r_i > 0$ and an integer $N_i > 0$ such that $\forall n \geq N_i$, $\pi_i(\mathcal{F}_n^c) < \exp(-r_i n)$.

(ii) For all $\gamma, \nu > 0$ there exists an integer $I > 0$ such that for any $i > I$ there exists an integer $M_i > 0$ such that for all $n \geq M_i$, $\pi_i(K_\gamma) \geq \exp(-\nu n)$.

(iii) B_n is a bound that grows with n such that for all $r > 0$ there exists a real number $q > 1$ and an integer $N > 0$ such that for all $n \geq N$, $\sum_{i=B_n}^{\infty} \lambda_i < \exp(-rn^q)$.

(iv) For all i, $\lambda_i > 0$.

Then $\forall \epsilon > 0$, the posterior is asymptotically consistent for f_0 over Hellinger neighborhoods, i.e.:

$$P(A_\epsilon \mid (x_1, y_1), \ldots, (x_n, y_n)) \xrightarrow{p} 1. \tag{105}$$

Corollary B.2. *(Hellinger consistency gives asymptotic consistency for prior on width). Under the conditions of Theorem 4, $\hat{g}_n$ is asymptotically consistent for g_0, i.e.:*

$$\int (\hat{g}_n(x) - g_0(x))^2 dx \xrightarrow{p} 0. \tag{106}$$

Proof. The conditions of Theorem 4 imply the conditions of Theorem 3, so then Corollary B.1 must hold. □

B.2 CONSISTENCY OF PORB-NET

B.2.1 Supporting results

Theorem 5. *(RBFNs are universal function approximators) Park and Sandberg [1991] Define S_ϕ as the set of all functions of the form:*

$$RBFN_\phi(x; \theta) = \sum_{k=1}^{K} w_k \phi\left(\lambda(x - c_k)\right), \tag{107}$$

where $\lambda > 0$, $w_k \in \mathbb{R}$, $c_k \in \mathbb{R}^D$ and $\theta = \{\{w_k\}_{k=1}^K, \{c_k\}_{k=1}^K, \lambda\}$ is the collection of network parameters. If $\phi : \mathbb{R}^d \to \mathbb{R}$ is an integrable bounded function such that ϕ is continuous almost everywhere and $\int_{\mathbb{R}^d} \phi(z)\, dz \neq 0$, then the family S_ϕ is dense in $L_p(\mathbb{R}^d)$ for every $p \in [1, \infty)$.

In our case, $\phi(z) = \exp(-z^2)$, which clearly satisfies the conditions of Theorem 5. We will denote $\text{RBFN}(x; \theta)$ the expression in Equation (107) particularized for the squared exponential ϕ function.

Lemma 3. *(Bound on network gradients)*

$$\frac{\partial \sqrt{f(x, y; \theta)}}{\partial \theta_i} \leq (8\pi e^2)^{-1/4} \frac{\partial RBFN(x; \theta)}{\partial \theta} = (8\pi e^2)^{-1/4} 2n^a C_n^3 \tag{108}$$

Proof. Applying the chain rule we have:

$$\left| \frac{\partial \sqrt{f(x, y; \theta)}}{\partial \theta_i} \right| = \frac{1}{2} \left(f(x, y; \theta)\right)^{-1/2} \frac{\partial f(x, y; \theta)}{\partial \theta_i} \tag{109}$$

$$= \frac{1}{2}(2\pi)^{-1/4} \exp\left(-\frac{1}{4}(y - \text{RBFN}(x; \theta))^2\right) |y - \text{RBFN}(x; \theta)| \left| \frac{\partial RBFN(x; \theta_i)}{\partial \theta_i} \right| \tag{110}$$

First we show that we can bound the middle terms by:

$$\exp\left(-\frac{1}{4}(y - \text{RBFN}(x; \theta))^2\right) |y - \text{RBFN}(x; \theta)| \leq \exp(-1/2) 2^{1/2} \tag{111}$$

To see this, rewrite the left-hand-side of Equation 111 as $s(z) := \exp(-(1/4)z^2)|z|$, where $z = y - \text{RBFN}(x; \theta)$. Taking the derivative we have:

$$\frac{\partial s(z)}{\partial z} = \begin{cases} -\frac{1}{2} z^2 \exp(-\frac{1}{4} z^2) + \exp(-\frac{1}{4} z^2) & z \geq 0 \\ \frac{1}{2} z^2 \exp(-\frac{1}{4} z^2) - \exp(-\frac{1}{4} z^2) & z < 0 \end{cases} \tag{112}$$

$$= \begin{cases} \exp(-\frac{1}{4} z^2)(-\frac{1}{2} z^2 + 1) & z \geq 0 \\ \exp(-\frac{1}{4} z^2)(\frac{1}{2} z^2 - 1) & z < 0 \end{cases} \tag{113}$$

Setting to zero, we must have that $\frac{1}{2} z^2 = 1 \implies z = \sqrt{2}$. Thus, $a(z) \leq \exp(-1/2) 2^{1/2}$, as in Equation 111.

Next, consider the derivatives of the radial basis function network:

$$\left|\frac{\partial RBFN(x;\theta_i)}{\partial b}\right| = 1 \tag{114}$$

$$\left|\frac{\partial RBFN(x;\theta_i)}{\partial w_k}\right| = \exp(-\frac{1}{2}\lambda^2(x-c_k)^2) \leq 1 \tag{115}$$

$$\begin{aligned}
\left|\frac{\partial RBFN(x;\theta_i)}{\partial w_k}\right| &= |w_k|\exp(-\frac{1}{2}\lambda^2(x-c_k)^2)\lambda^2|x-c| && (116)\\
&\leq |w_k|\lambda^2(|c_k|+1) && (117)\\
&\leq C_n^2(C_n+1) && (118)\\
&\leq C_n^3 + C_n^2 && (119)\\
&\leq 2C_n^3 && (120)
\end{aligned}$$

since $C_n^2 = \exp(2n^{b-a}) < \exp(3n^{b-a}) = C_n^3$

$$\begin{aligned}
\left|\frac{\partial RBFN(x;\theta_i)}{\partial w_k}\right| &= \frac{1}{2}\left|\sum_{k=1}^{K_n} w_k \exp(-\frac{1}{2}\lambda^2(x-c_k)^2)(x-c)^2\right| && (121)\\
&= \frac{1}{2}\sum_{k=1}^{K_n}\left|w_k \exp(-\frac{1}{2}\lambda^2(x-c_k)^2)(x-c)^2\right| && (122)\\
&\leq \frac{1}{2}\sum_{k=1}^{K_n}|w_k|(|c|+1)^2 && (123)\\
&\leq \frac{1}{2}\sum_{k=1}^{n^a} C_n(C_n+1)^2 && (124)\\
&= \frac{1}{2}n^a C_n(C_n+1)^2 && (125)\\
&= \frac{1}{2}n^a C_n(C_n^2+2C_n+1) && (126)\\
&= \frac{1}{2}n^a(C_n^3+2C_n^2+C_n) && (127)\\
&\leq \frac{1}{2}n^a(C_n^3+2C_n^3+C_n^3) && (128)\\
&\leq 2n^a C_n^3 && (129)
\end{aligned}$$

Plugging everything in to Equation 110 we have the desired inequality. □

Lemma 4. *(Bounding sum of exponentially bounded terms). For two sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ suppose there exists real numbers $r_a > 0$ and $r_b > 0$ as well as integers $N_a > 0$ and $N_b > 0$ such that $a_n \leq \exp(-r_a n)$ for all $n \geq N_a$ and $b_n \leq \exp(-r_b n)$ for all $n \geq N_b$. Then there exists a real number $r > 0$ and an integer $N > 0$ such that $a_n + b_n \leq \exp(-rn)$ for all $n \geq N$.*

Proof. Set $\tilde{r} = \min\{r_a, r_b\}$ and $\tilde{N} = \max\{N_a, N_b\}$. Then we have:

$$\begin{aligned}
a_n &\leq \exp(-r_a n), \quad \forall n \geq \tilde{N} \geq N_a && (130)\\
&\leq \exp(-\tilde{r}n,) \quad \forall n \geq \tilde{N} && (131)
\end{aligned}$$

. Similarly, $b_n \leq \exp(-\tilde{r}n), \forall n \geq \tilde{N}$. Thus we have $a_n + b_n \leq 2\exp(-\tilde{r}n), \forall n \geq \tilde{N}$.

Now set $N = \max\{\lceil \frac{\log 2}{\tilde{r}} \rceil + 1, \tilde{N}\}$ and $r = \tilde{r} - \frac{\log 2}{N}$. Notice $r > 0$, since $N \geq \lceil \frac{\log 2}{r} \rceil + 1 > \frac{\log 2}{r}$ implies $r = \tilde{r} - \frac{\log 2}{N} > \tilde{r} - \log 2 \frac{\tilde{r}}{\log 2} = 0$. It follows that $2\exp(-rn) \leq \exp(-rn)$, $\forall n \geq N$, since:

$$
\begin{aligned}
& 2\exp(-\tilde{r}n) \leq \exp(-rn) && (132)\\
\iff & \log 2 - \tilde{r}n \leq -rn && (133)\\
\iff & \log 2 - \tilde{r}n \leq -\left(\tilde{r} - \frac{\log 2}{N}\right) n && (134)\\
\iff & \log 2 - \tilde{r}n \leq -\tilde{r}n + \frac{n \log 2}{N} && (135)\\
\iff & N \leq n && (136)
\end{aligned}
$$

□

Lemma 5. *(Useful equality) For all $\delta \leq 1$ and $x \in [0, 1]$, if $|\tilde{c} - c| \leq \delta$ and $|\tilde{\lambda} - \lambda| \leq \delta$, then there exists a constant ξ such that $|\xi| \leq A(|c|, \lambda)\delta$ and:*

$$\tilde{\lambda}^2(x - \tilde{c})^2 = \lambda^2(x - c)^2 + \xi, \tag{137}$$

where $A(|c|, \lambda) = 2\lambda(|c| + 1)(\lambda + |c| + 2) + (\lambda + |c| + 2)^2$

Proof. Since $|\tilde{c} - c| \leq \delta$ and $|\tilde{\lambda} - \lambda| \leq \delta$ there exists constants ξ_1 and ξ_2, where $|\xi_1| \leq \delta$ and $|\xi_2| \leq \delta$, such that $\tilde{c} = c + \xi_1$ and $\tilde{\lambda} = \lambda + \xi_2$

Plugging $\tilde{c} = c + \xi$ and $\tilde{\lambda} = \lambda + \xi_2$ into the left-hand-side of the desired inequality:

$$
\begin{aligned}
\tilde{\lambda}(x - \tilde{c}) &= (\lambda + \xi_2)(x - c - \xi_1) && (138)\\
&= \lambda(x - c) + \underbrace{(-\lambda\xi_1) + \xi_2(x - c) - \xi_1\xi_2}_{:=\xi_3} && (139)
\end{aligned}
$$

Notice:

$$
\begin{aligned}
|\xi_3| &= |(-\lambda\xi_1) + \xi_2(x - c) - \xi_1\xi_2| && (140)\\
&\leq \lambda|\xi_1| + |\xi_2||x - c| + |\xi_1||\xi_2| && (141)\\
&\leq \lambda\delta + \delta(|c| + 1) + \delta^2 && (142)\\
&\leq (\lambda + |c| + 2)\delta && (143)
\end{aligned}
$$

In Equation 142 we use $|x - c| \leq (|c| + 1)$, which follows since we assume $x \in [0, 1]$, as well as $\xi_1 \leq \delta$ and $\xi_2 \leq \delta$. In Equation 143 we use $\delta^2 \leq \delta$, which follows since we assume $\delta \leq 1$. Squaring the left-hand-side of the desired inequality:

$$
\begin{aligned}
\tilde{\lambda}^2(x - \tilde{c})^2 &= (\tilde{\lambda}(x - \tilde{c}))^2 && (144)\\
&= (\lambda(x - c) + \xi_3)^2 && (145)\\
&= \lambda^2(x - c)^2 + \underbrace{2\lambda(x - c)\xi_3 + \xi_3^2}_{:=\xi_4} && (146)
\end{aligned}
$$

Notice:

$$
\begin{aligned}
|\xi_4| &= |2\lambda(x - c)\xi_3 + \xi_3^2| && (147)\\
&\leq 2\lambda|x - c||\xi_3| + |\xi_3^2| && (148)\\
&\leq 2\lambda(|c| + 1)(\lambda + |c| + 2)\delta + (\lambda + |c| + 2)^2\delta^2 && (149)\\
&\leq \underbrace{\left(2\lambda(|c| + 1)(\lambda + |c| + 2) + (\lambda + |c| + 2)^2\right)}_{:=A(|c|,\lambda)}\delta && (150)
\end{aligned}
$$

In Equation 149 we use $|\xi_3| \leq (\lambda + |c| + 2)\delta$ and $|x - c| \leq (|c| + 1)$ again and Equation 150 we use $\delta^2 \leq \delta$. This proves the desired inequality for $\xi := \xi_4$. □

Lemma 6. *(Proximity in parameter space leads to proximity in function space). Let g be an RBFN with K nodes and parameters $(\theta_1, \dots, \theta_I)$ and let $\tilde{g}_n$ be an RBFN with $\tilde{K}_n$ nodes and parameters $(\tilde{\theta}_1, \dots, \tilde{\theta}_{\tilde{I}(n)})$, where $\tilde{K}_n$ grows with n. Define $\theta_i = 0$ for $i > \mathcal{I}$, $\tilde{\theta}_i = 0$ for $i > \tilde{i}$, and M_δ, for any $\delta > 0$, as the set of all networks $\tilde{g}$ that are close in parameter space to g:*

$$M_\delta(g) := \{\tilde{g}_n \mid |\tilde{\theta}_i - \theta_i|, i = 1, \dots\} \tag{151}$$

Then for any $\tilde{g} \in M_\delta$ and sufficiently large n,

$$\sup_{x \in \mathcal{X}} \left(\tilde{g}(x) - g(x)\right)^2 \leq \left(3\tilde{K}_n\right)^2 \delta^2 \tag{152}$$

Proof.

$$\sup_{x \in \mathcal{X}} \left(\tilde{g}(x) - g(x)\right)^2 \tag{153}$$

$$= \sup_{x \in \mathcal{X}} \left(\tilde{b} + \sum_{k=1}^{\tilde{K}_n} \tilde{w}_k \exp(-\tilde{\lambda}^2 (x - \tilde{c}_k)^2) - b - \sum_{k=1}^{K} w_k \exp(-\lambda^2 (x - c_k)^2) \right)^2 \tag{154}$$

$$= \sup_{x \in \mathcal{X}} \left((\tilde{b} - b) + \left(\sum_{k=1}^{\tilde{K}_n} \tilde{w}_k \exp(-\tilde{\lambda}^2 (x - \tilde{c}_k)^2) - \sum_{k=1}^{K} w_k \exp(-\lambda^2 (x - c_k)^2) \right) \right)^2 \tag{155}$$

$$= \sup_{x \in \mathcal{X}} \left((\tilde{b} - b) + \left(\sum_{k=1}^{\tilde{K}_n^*} \tilde{w}_k \exp(-\tilde{\lambda}^2 (x - \tilde{c}_k)^2) - w_k \exp(-\lambda^2 (x - c_k)^2) \right) \right)^2 \tag{156}$$

$$\leq \sup_{x \in \mathcal{X}} \left(|\tilde{b} - b| + \left| \sum_{k=1}^{\tilde{K}_n^*} \tilde{w}_k \exp(-\tilde{\lambda}^2 (x - \tilde{c}_k)^2) - w_k \exp(-\lambda^2 (x - c_k)^2) \right| \right)^2 \tag{157}$$

$$= \sup_{x \in \mathcal{X}} \left[|\tilde{b} - b|^2 + 2|\tilde{b} - b| \left| \sum_{k=1}^{\tilde{K}_n^*} \tilde{w}_k \exp(-\tilde{\lambda}^2 (x - \tilde{c}_k)^2) - w_k \exp(-\lambda^2 (x - c_k)^2) \right| \right. \tag{158}$$

$$\left. + \left| \sum_{k=1}^{\tilde{K}_n^*} \tilde{w}_k \exp(-\tilde{\lambda}^2 (x - \tilde{c}_k)^2) - w_k \exp(-\lambda^2 (x - c_k)^2) \right|^2 \right] \tag{159}$$

$$\leq |\tilde{b} - b|^2 + 2|\tilde{b} - b| \sup_{x \in \mathcal{X}} \left| \sum_{k=1}^{\tilde{K}_n^*} \tilde{w}_k \exp(-\tilde{\lambda}^2 (x - \tilde{c}_k)^2) - w_k \exp(-\lambda^2 (x - c_k)^2) \right| \tag{160}$$

$$+ \sup_{x \in \mathcal{X}} \left| \sum_{k=1}^{\tilde{K}_n^*} \tilde{w}_k \exp(-\tilde{\lambda}^2 (x - \tilde{c}_k)^2) - w_k \exp(-\lambda^2 (x - c_k)^2) \right|^2 \tag{161}$$

$$= |\tilde{b} - b|^2 + 2|\tilde{b} - b| \sup_{x \in \mathcal{X}} \left| \sum_{k=1}^{\tilde{K}_n^*} \tilde{w}_k \exp(-\tilde{\lambda}^2 (x - \tilde{c}_k)^2) - w_k \exp(-\lambda^2 (x - c_k)^2) \right| \tag{162}$$

$$+ \left(\sup_{x \in \mathcal{X}} \left| \sum_{k=1}^{\tilde{K}_n^*} \tilde{w}_k \exp(-\tilde{\lambda}^2 (x - \tilde{c}_k)^2) - w_k \exp(-\lambda^2 (x - c_k)^2) \right| \right)^2 \tag{163}$$

$$\leq |\tilde{b} - b|^2 + 2|\tilde{b} - b| \sup_{x \in \mathcal{X}} \sum_{k=1}^{\tilde{K}_n^*} \left| \tilde{w}_k \exp(-\tilde{\lambda}^2 (x - \tilde{c}_k)^2) - w_k \exp(-\lambda^2 (x - c_k)^2) \right| \tag{164}$$

$$+\left(\sup_{x\in\mathcal{X}}\sum_{k=1}^{\tilde{K}_n^*}\left|\tilde{w}_k\exp(-\tilde{\lambda}^2(x-\tilde{c}_k)^2)-w_k\exp(-\lambda^2(x-c_k)^2)\right|\right)^2 \tag{165}$$

$$\leq|\tilde{b}-b|^2+2|\tilde{b}-b|\sum_{k=1}^{\tilde{K}_n^*}\sup_{x\in\mathcal{X}}\left|\tilde{w}_k\exp(-\tilde{\lambda}^2(x-\tilde{c}_k)^2)-w_k\exp(-\lambda^2(x-c_k)^2)\right| \tag{166}$$

$$+\left(\sum_{k=1}^{\tilde{K}_n^*}\sup_{x\in\mathcal{X}}\left|\tilde{w}_k\exp(-\tilde{\lambda}^2(x-\tilde{c}_k)^2)-w_k\exp(-\lambda^2(x-c_k)^2)\right|\right)^2 \tag{167}$$

$$=|\tilde{b}-b|^2+2|\tilde{b}-b|\sum_{k=1}^{\tilde{K}_n^*}\Gamma_k+\left(\sum_{k=1}^{\tilde{K}_n^*}\Gamma_k\right)^2 \tag{168}$$

$$\leq\delta^2+2\delta\sum_{k=1}^{\tilde{K}_n^*}\Gamma_k+\left(\sum_{k=1}^{\tilde{K}_n^*}\Gamma_k\right)^2, \tag{169}$$

where:

$$\Gamma_k:=\sup_{x\in\mathcal{X}}\left|\tilde{w}_k\exp(-\tilde{\lambda}^2(x-\tilde{c}_k)^2)-w_k\exp(-\lambda^2(x-c_k)^2)\right| \tag{170}$$

Let $u(x)^2:=\lambda^2(x-c_k)^2$ and $\tilde{u}(x)^2=\tilde{\lambda}^2(x-\tilde{c}_k)^2$ and pick any $x\in\mathcal{X}$. By Lemma 5 there exists a constant η such that $|\eta|\leq A(|c|,\lambda)\delta$ and

$$\tilde{u}(x)^2=u(x)^2+\eta. \tag{171}$$

Now define $\xi=\sqrt{|\eta|}$ and consider two cases.

- If $\tilde{u}(x)^2\geq u(x)^2$, then Equation 171 is equivalent to $\tilde{u}(x)^2=u(x)^2+\xi^2$. Then Γ_k becomes:

$$\Gamma_k=\sup_{x\in\mathcal{X}}\left|\tilde{w}_k\exp(-\tilde{u}^2(x))-w_k\exp(-u^2(x))\right| \tag{172}$$

$$=\left|\tilde{w}_k\exp(-u^2(x)-\xi^2)-w_k\exp(-u^2(x)^2)\right| \tag{173}$$

$$=\sup_{x\in\mathcal{X}}\exp(-u(x)^2)\left|\tilde{w}_k\exp(-\xi^2)-w_k\right| \tag{174}$$

$$=\left|\tilde{w}_k\exp(-\xi^2)-w_k\right|\sup_{x\in\mathcal{X}}\exp(-u(x)^2) \tag{175}$$

$$\leq\left|\tilde{w}_k\exp(-\xi^2)-w_k\right| \tag{176}$$

  Since $|\tilde{w}_k-w_k|\leq\delta$, there exists τ, where $|\tau|\leq\delta$, such that $\tilde{w}_k=w_k+\tau$. Plugging this in:

$$\Gamma_k\leq\left|(w_k+\tau)\exp(-\xi^2)-w_k\right| \tag{177}$$

$$\leq\left|w_k(\exp(-\xi^2)-1)+\tau\right| \tag{178}$$

$$\leq|w_k||\exp(-\xi^2)-1|+|\tau| \tag{179}$$

$$\leq|w_k|\xi^2+\delta, \tag{180}$$

  where we use the result that $1-\xi^2\leq\exp(-\xi^2)$ in Equation 179.

- If $\tilde{u}(x)^2<u(x)^2$, then Equation 171 is equivalent to $u(x)^2=\tilde{u}(x)^2+\xi^2$. Then Γ_k becomes:

$$\Gamma_k=\sup_{x\in\mathcal{X}}\left|\tilde{w}_k\exp(-\tilde{u}^2(x))-w_k\exp(-u^2(x))\right| \tag{181}$$

$$=\sup_{x\in\mathcal{X}}\left|\tilde{w}_k\exp(-\tilde{u}^2(x)-w_k\exp(-\tilde{u}^2(x)-\xi^2)\right| \tag{182}$$

$$=\sup_{x\in\mathcal{X}}\exp(-\tilde{u}^2(x))\left|\tilde{w}_k-w_k\exp(-\xi^2)\right| \tag{183}$$

$$= \left|\tilde{w}_k - w_k \exp(-\xi^2)\right| \sup_{x \in \mathcal{X}} \exp(-\tilde{u}^2(x)) \tag{184}$$

$$\leq \left|\tilde{w}_k - w_k \exp(-\xi^2)\right| \tag{185}$$

Using the same τ as above:

$$\Gamma_k \leq \left|(w_k + \tau)\exp(-\xi^2) - w_k\right| \tag{186}$$

$$\leq \left|w_k(1 - \exp(-\xi^2)) + \tau\right| \tag{187}$$

$$\leq |w_k||1 - \exp(-\xi^2)| + |\tau| \tag{188}$$

$$= |w_k||\exp(-\xi^2) - 1| + |\tau| \tag{189}$$

$$\leq |w_k|\xi^2 + \delta. \tag{190}$$

In either of the two cases, we have $\Gamma_k \leq |w_k|\xi^2 + \delta$. Proceeding:

$$\Gamma_k \leq |w_k|\xi^2 + \delta \tag{191}$$

$$\leq |w_k|A(|c|, \lambda)\delta + \delta \tag{192}$$

$$= (|w_k|A(|c|, \lambda) + 1)\delta \tag{193}$$

Now consider

$$\sum_{k=1}^{\tilde{K}_n^*} \Gamma_k \leq \sum_{k=1}^{\tilde{K}_n} \Gamma_k \qquad \text{for large } n \tag{194}$$

$$\leq \delta \sum_{k=1}^{\tilde{K}_n} (|w_k|A(|c|, \lambda) + 1) \tag{195}$$

$$\leq \delta \left(\sum_{k=1}^{\tilde{K}_n} |w_k|A(|c|, \lambda) + \tilde{K}_n \right) \tag{196}$$

$$\leq \delta \left(\tilde{K}_n + \tilde{K}_n \right) \tag{197}$$

$$= 2\delta\tilde{K}_n \qquad \text{for large } n \tag{198}$$

Equation 197 follows because for $k \geq K$, $w_k = 0$ by definition, so $\sum_{k=1}^{\tilde{K}_n} |w_k|A(|c|, \lambda)$ is a constant and thus less than $\tilde{K}_n$ for large n.

Plugging Equation 198 into Equation 169:

$$\sup_{x \in \mathcal{X}} (\tilde{g}(x) - g(x))^2 \leq \delta^2 + 2(2\delta\tilde{K}_n) + (2\delta\tilde{K}_n)^2 \tag{199}$$

$$= \left(1 + 2(2\tilde{K}_n) + (2\tilde{K}_n)^2\right)\delta^2 \tag{200}$$

$$= \left(1 + 2\tilde{K}_n)\right)^2 \delta^2 \tag{201}$$

$$\leq (3\tilde{K}_n)^2\delta^2 \tag{202}$$

□

B.2.2 Main theorems for PoRB-Net

Recall the generative model for PoRB-Net in the case of a uniform intensity function with a Gamma prior on its level. For simplicity and w.l.o.g, we consider the case where the hyperparameter s_0^2 and the observation variance are fixed to 1.

We first consider the case where the width of the network is allowed to grow with the data but is fixed in the prior. We call the estimated regression function $\hat{g}_n$, with width K_n and prior π_n, where n is the number of observations. The following theorem gives consistency for this model.

Note that the following proof uses [Park and Sandberg, 1991] to show the existence of a neural network that approximates any square integrable function. We assume that the center parameters of this network are contained in the bounded region over which the Poisson process is defined, which can be made arbitrarily large.

Theorem 6. *(PoRB-Net consistency with fixed width that grows with the number of observations). If there exists a constant $a \in (0,1)$ such that $K_n \leq n^a$, and $K_n \to \infty$ as $n \to \infty$, then for any square integrable ground truth regression function g_0, $\hat{g}_n$ is asymptotically consistent for g as $n \to \infty$, i.e.*

$$\int (\hat{g}_n(x) - g_0(x))^2 dx \xrightarrow{p} 0. \tag{203}$$

Proof.

Proof outline

- Show Condition (i) of Theorem 3 is met
  - Write prior probability of large parameters as a sum of integrals over each parameter
  - Bound each set of parameters:
    - Bound weights (as in Lee [2000])
    - Bound centers (trivial since parameter space bounded)
    - Bound λ^2 with Chernoff bound
  - Bound sum using Lemma 4
- Show Condition (ii) of Theorem 3 is met.
  - Assume true regression function g_0 is L_2
  - Use Theorem 5 to find an RBFN g that approximates g_0
  - Define M_δ as RBFNs close in parameter space to g
  - Show $M_\delta \subset K_\gamma$ using Lemmas 5 and 6.
  - Show $\pi_n(M_\delta) \geq \exp(-rn)$:
    - Show you can write as a product of integrals over parameters
    - Bound each term separately:
      - Bound weights as in Lee [2000]
      - Bound centers and λ^2

Condition (i) We want to show that there exists an $r > 0$ and an $N_1 \in \mathbb{N}$ such that $\forall n \geq N_1$:

$$\pi_n(\mathcal{F}_m^c) < \exp(-nr).$$

Write prior probability of large parameters as a sum of integrals over each parameter. The prior π_n assigns zero probability to RBFNs with anything but K_n nodes, so there is no issue writing $\pi_n(\mathcal{F}_n)$ and its value is equivalent to $\pi_n(\mathcal{G}_n)$, even though $\mathcal{G}_n \subset \mathcal{F}_n$.

Notice that $\pi_n(\mathcal{G}_n^c)$ requires evaluating a multiple integral over a subset of the product space of I_n parameters. Notice $\mathcal{G}_n$ can be written as an intersection of sets:

$$\mathcal{G}_n = \bigcap_{i=1}^{I_n} \{\text{RBFN} \in \mathcal{H}_n \mid |\theta_i| \leq C_n\}.$$

Therefore we have:

$$\begin{aligned}
\pi_n\left(\mathcal{F}_n^c\right) &= \pi_n\left(\mathcal{G}_n^c\right) \\
&= \pi_n\left(\left[\bigcap_{i=1}^{I_n}\{\text{RBFN} \in \mathcal{H}_n \mid |\theta_i| \leq C_n\}\right]^c\right) \\
&= \pi_n\left(\bigcup_{i=1}^{I_n}\{\text{RBFN} \in \mathcal{H}_n \mid |\theta_i| \leq C_n\}^c\right) && \text{De Morgan} \\
&= \pi_n\left(\bigcup_{i=1}^{I_n}\{\text{RBFN} \in \mathcal{H}_n \mid |\theta_i| > C_n\}\right) \\
&\leq \sum_{i=1}^{I_n} \pi_n\left(\{\text{RBFN} \in \mathcal{H}_n \mid |\theta_i| > C_n\}\right) && \text{Union bound.}
\end{aligned} \tag{204}$$

Next, independence in the prior will allow us to write each term in Equation 204 as an integral over a single parameter. Define the following sets:

$$\begin{aligned}
\mathcal{C}_i(n) &:= \Theta_i \setminus [-C_n, C_n] \\
\mathcal{R}_i(n) &:= \Theta_1 \times \ldots \times \Theta_{i-1} \times \mathcal{C}_i(n) \times \Theta_{i+1} \times \ldots \times \Theta_{I_n}
\end{aligned}$$

where Θ_i is the parameter space corresponding to parameter θ_i (either $\mathbb{R}$ or $\mathbb{R}^+$). Notice that because $\mathcal{R}_i(n)$ is a union of two rectangular sets (one where θ_i is less than $-C_n$ and one where θ_i is greater than C_n), we can apply Fubini's theorem. Thus, each term in Equation 204 can be written as:

$$\pi_n\left(\{\text{RBFN} \in \mathcal{H}_n \mid |\theta_i| > C_n\}\right) \tag{205}$$

$$= \int \ldots \int_{\mathcal{R}_i(n)} \pi_n(\theta_1, \ldots, \theta_{I_n}) d(\theta_1, \ldots, \theta_{I_n}) \tag{206}$$

$$= \int d\theta_1 \ldots \int d\theta_{I_n}\ \pi_n(\theta_1, \ldots, \theta_{I_n}) \tag{207}$$

$$= \int d\lambda^2 \int dw_1 \ldots \int dw_{I_n^w} \int dc_1 \ldots \int dw_{I_n^c}\ \pi_n(\lambda^2) \prod_j \pi_n(c_j \mid \lambda^2) \prod_j \pi_n(w_j) \tag{208}$$

$$= \left(\int d\lambda^2\ \pi_n(\lambda^2) \int dc_1 \ldots \int dc_{I_n^c}\ \prod_j \pi_n(c_j \mid \lambda^2)\right)\left(\int dw_1 \ldots \int dw_{I_n^w}\ \prod_j \pi_n(w_j)\right) \tag{209}$$

$$= \left(\int d\lambda^2\ \pi_n(\lambda^2) \prod_j \int dc_j\ \pi_n(c_j \mid \lambda^2)\right)\left(\prod_j \int dw_j\ \pi_n(w_j)\right) \tag{210}$$

$$= \begin{cases} \int_{\mathcal{C}_n} d\lambda^2\ \pi_n(\lambda^2) & i = \mathcal{I}_n^{(\lambda^2)} \\ \int_{\mathcal{C}_n} dw\ \pi_n(w) & i \in \mathcal{I}_n^{(w)} \\ \int_{\mathcal{R}^+} d\lambda^2\ \pi_n(\lambda^2) \int_{\mathcal{C}_n} dc_i\ \pi_n(c_i \mid \lambda^2) & i \in \mathcal{I}_n^{(c)} \end{cases} \tag{211}$$

In Equation 206 we apply Fubini's theorem, which allows us to write a multiple integral as an interated integral. It is understood that the ith integral is over the restricted parameters space $[-C_n, C_n]$ while the remaining integrals are over the entire parameter space, meaning they integrate to 1. This allows us to write the result in Equation 211.

Therefore, by Equations 204 and 211 we have:

$$\pi_n\left(\mathcal{F}_n^c\right) \leq \underbrace{\int_{\mathcal{C}_n} d\lambda^2\ \pi_n(\lambda^2)}_{\lambda^2 \text{ term}} + \underbrace{\sum_{i \in \mathcal{I}_n^{(w)}} \int_{\mathcal{C}_n} dw\ \pi_n(w)}_{W \text{ term}} + \underbrace{\sum_{i \in \mathcal{I}_n^{(c)}} \int_{\mathcal{R}^+} d\lambda^2\ \pi_n(\lambda^2) \int_{\mathcal{C}_n} dc_i\ \pi_n(c_i \mid \lambda^2)}_{C \text{ term}} \tag{212}$$

Bound each term in the sum. We will deal with each of these terms separately.

- *W term.* With some minor difference for the dependence of the number of weight parameters on the network width (DK_n in our case compared to $(D+2)K_n+1$), equations 119-128 in [Lee, 2000] show for all $n \geq N_w$ for some N_w:

$$\sum_{i \in \mathcal{I}_n^{(w)}} \int_{\mathcal{C}_i(n)} \pi_n(w_i)\, dw_i \leq \exp(-nr)$$

- *C term.* Since the parameter bound $C_n \to \infty$ as $n \to \infty$ and since the prior over the center parameters is defined over a bounded region, as $n \to \infty$ the bounded region will be contained in $[-C_n, C_n]$ and thus disjoint from $\mathcal{C}_i(n) := \Theta_i \setminus [-C_n, C_n]$. Thus, for all n greater than some N_c, $\int_{\mathcal{C}_i(n)} \pi_n(c_i)\, dc_i = 0$ for all center parameters.

- *λ^2 term.*

$$\begin{aligned}
\int \pi_n(\lambda^2) d\lambda &= \int_{C_n}^{\infty} \frac{{\beta_\lambda}^{\alpha_\lambda}}{\Gamma(\alpha_\lambda)} \lambda^{2(\alpha_\lambda - 1)} \exp(-\beta_\lambda^2 \lambda) d\lambda^2 && (213)\\
&\leq \left(\frac{\beta_\lambda C_n}{\alpha_\lambda}\right)^{\alpha_\lambda} \exp(\alpha_\lambda - \beta_\lambda C_n) && \text{Chernoff Bound} \quad (214)\\
&\leq \left(\frac{\beta_\lambda e}{\alpha_\lambda}\right)^{\alpha} \exp(\alpha_\lambda n^{b-a}) \exp(-\beta_\lambda \exp(n^{b-a})) && C_n \leq n^{b-a} \quad (215)
\end{aligned}$$

Taking the negative log we have:

$$\begin{aligned}
-\log\left(\int \pi_n(\lambda^2) d\lambda^2\right) &\geq \underbrace{-\alpha \log\left(\frac{\beta e}{\alpha}\right)}_{:=A} + \beta \exp(n^{b-a}) - \alpha n^{b-a} && (216)\\
&= A + \beta\left(\sum_{j=0}^{\infty} \frac{(n^{b-a})^j}{j!}\right) - \alpha n^{b-a} && (217)\\
&= A + \beta\left(1 + n^{b-a} + \frac{1}{2} n^{2(b-a)} + \sum_{j=3}^{\infty} \frac{(n^{b-a})^j}{j!}\right) - \alpha n^{b-a} && (218)\\
&= \underbrace{(A+\beta) + (\beta - \alpha) n^{b-a} + \frac{1}{2}\beta n^{2(b-a)}}_{:=h(n)} + \beta \sum_{j=3}^{\infty} \frac{(n^{b-a})^j}{j!} && (219)\\
&= h(n) + \beta \sum_{j=3}^{\infty} \frac{(n^{b-a})^j}{j!}. && (220)
\end{aligned}$$

Now pick $k^* \in \{3, 4, \dots\}$ such that $(b-a)k^* \geq 1$, so $n^{(b-a)k^*} \geq n$, and pick any $r \in (0, \beta/(k^*!))$. Then, since every term in the sum is positive, we have:

$$\begin{aligned}
-\log\left(\int \pi_n(\lambda^2) d\lambda^2\right) &\geq h(n) + \beta \frac{n^{(b-a)k^*}}{k^*!} && (221)\\
&\geq h(n) + \frac{\beta}{k^*!} n && (222)\\
&\geq h(n) + rn && (223)\\
&\geq rn && \forall n \geq N_\lambda, \quad (224)
\end{aligned}$$

where the last inequality holds because $\beta > 0$ and $(b-a) \in (0,1)$ clearly implies there exists an $N_\lambda > 0$ such that for all $n \geq N_\lambda$, $h(n) > 0$. Negating and exponentiating each side we have:

$$\int \pi_n(\lambda^2) d\lambda^2 \leq \exp(-rn) \qquad \forall n \geq N_\lambda. \quad (225)$$

Bound sum. For any $n \geq N_c$, since the C term is zero in this case, we have:

$$\pi_n\left(\mathcal{F}_n^c\right) \leq \sum_{i \in \mathcal{I}_n^{(w)}} \int_{\mathcal{C}_i(n)} \pi_n(w_i)\, dw_i + \int_{\mathcal{C}_i(n)} \pi_n(\lambda^2)\, d\lambda^2 \tag{226}$$

$$\leq \exp(-rn) \qquad \forall n \geq N \tag{227}$$

where the last inequality follows from Lemma 4 applied to the sequences:

$$a_n := \sum_{i \in \mathcal{I}_n^{(w)}} \int_{\mathcal{C}_i(n)} \pi_n(w_i)\, dw_i \tag{228}$$

$$b_n := \int_{\mathcal{C}_i(n)} \pi_n(\lambda^2)\, d\lambda^2 \tag{229}$$

which we already showed to be exponentially bounded above for large n.

Condition (ii) Let $\gamma, \nu > 0$.

Assume true regression function. Assume $g_0 \in L_2$ is the true regression function

Find RBFN near ground truth function. Set $\epsilon = \sqrt{\gamma/2}$. By Theorem 5 there exists an RBFN g such that $\|g - g_0\|_2 \leq \epsilon$. We assume the center parameters of g are contained in the bounded region $\mathcal{C}$ over which the Poisson process is defined, which can be made arbitrarily large.

Define M_δ. Set $\delta = \epsilon/(3n^a)$ and let M_δ be defined as in Lemma 6. Then by Lemma 6, for any $\tilde{g} \in M_\delta$ we have:

$$\sup_{x \in \mathcal{X}} \left(\tilde{g}(x) - g(x)\right)^2 \leq \left(3\tilde{K}_n \delta\right)^2 = \epsilon^2 \tag{230}$$

Next we show that $M_\delta \subset K_\gamma$ for all $\gamma > 0$ and appropriately chosen δ. This means we only need to show $\pi_n\left(M_\delta\right) \geq \exp(-n\nu)$, since $M_\delta \subset K_\gamma$ implies $\pi_n\left(K_\gamma\right) \geq \pi_n\left(M_\delta\right)$.

Show M_δ contained in K_γ. Next we show that for any $\tilde{g} \in M_\delta$, $D_K(f_0, \tilde{f}) \leq \gamma$ i.e. $M_\delta \subset K_\gamma$. The following are exactly equations 129-132 and then 147-151 from Lee [2000].

$$D_K(f_0, \tilde{f}) = \int\int f_0(x,y) \log \frac{f_0(x,y)}{\tilde{f}(x,y)}\, dy\, dx \tag{231}$$

$$= \frac{1}{2} \int\int \left[(y - \tilde{g}(x))^2 - (y - g_0(x))^2\right] f_0(y \mid x) f_0(x)\, dy\, dx \tag{232}$$

$$= \frac{1}{2} \int\int \left[-2y\tilde{g}(x) + \tilde{g}(x)^2 + 2yg_0(x) - g_0(x)^2\right] f_0(y \mid x) f_0(x)\, dy\, dx \tag{233}$$

$$= \frac{1}{2} \int (\tilde{g}(x) - g_0(x))^2 f_0(x)\, dx \tag{234}$$

$$= \frac{1}{2} \int (\tilde{g}(x) - g(x) + g(x) - g_0(x))^2 f_0(x)\, dx \tag{235}$$

$$\leq \frac{1}{2} \left[\int \underbrace{\sup_{x \in \mathcal{X}} (\tilde{g}(x) - g(x))^2}_{\text{Lemma 6}} f_0(x)\, dx + \int \underbrace{(g(x) - g_0(x))^2}_{\text{Theorem 5}} f_0(x)\, dx \right. \tag{236}$$

$$+2 \underbrace{\sup_{x \in \mathcal{X}} |\tilde{g}(x) - g(x)|}_{\text{Lemma 6}} \int \underbrace{|g(x) - g_0(x)|}_{\text{Theorem 5}} f_0(x) dx \Bigg] \tag{237}$$

$$< \frac{1}{2}[\epsilon^2 + \epsilon^2 + 2\epsilon^2] \tag{238}$$

$$= 2\epsilon^2 = \gamma \tag{239}$$

Show mass on M_δ is greater than exponential

$$\begin{aligned}
\pi_n(M_\delta) &= \int_{\theta_1-\delta}^{\theta_1+\delta} \cdots \int_{\theta_{\tilde{I}_n}-\delta}^{\theta_{\tilde{I}_n}+\delta} \pi_n(\tilde{\theta}_1, \ldots, \tilde{\theta}_{\tilde{I}_n})\, d\tilde{\theta}_1 \ldots d\theta_{\tilde{I}_n} \\
&= \int_{\theta_1-\delta}^{\theta_1+\delta} \cdots \int_{\theta_{\tilde{I}_n}-\delta}^{\theta_{\tilde{I}_n}+\delta} \pi_n(\tilde{\lambda^2}) \prod_i \pi_n(\tilde{c}_i \mid \tilde{\lambda^2}) \prod_i \pi_n(w)\, d\tilde{\theta}_1 \ldots d\theta_{\tilde{I}_n} \\
&= \int_{\lambda^2-\delta}^{\lambda^2+\delta} \pi_n(\tilde{\lambda^2}) \prod_{i=1}^{\tilde{I}_n^{(c)}} \int_{c_i-\delta}^{c_i+\delta} \pi_n(\tilde{c}_i \mid \tilde{\lambda^2})\, d\tilde{c}_i\, d\tilde{\lambda^2} \times \prod_{i=1}^{\tilde{I}_n^{(w)}} \int_{w_i-\delta}^{w_i+\delta} \pi_n(\tilde{w}_i)\, d\tilde{w}_i \\
&= \int_{\lambda^2-\delta}^{\lambda^2+\delta} \pi_n(\tilde{\lambda^2}) \prod_{i=1}^{\tilde{I}_n^{(c)}} \int_{c_i-\delta}^{c_i+\delta} \frac{1}{\mu(\mathcal{C})} 1_{[\tilde{c}_i \in \mathcal{C}]}\, d\tilde{c}_i\, d\tilde{\lambda^2} \times \prod_{i=1}^{\tilde{I}_n^{(w)}} \int_{w_i-\delta}^{w_i+\delta} \pi_n(\tilde{w}_i)\, d\tilde{w}_i \\
&= \underbrace{\int_{\lambda^2-\delta}^{\lambda^2+\delta} \pi_n(\tilde{\lambda^2})\, d\tilde{\lambda^2}}_{\lambda^2 \text{ term}} \times \underbrace{\prod_{i=1}^{\tilde{I}_n^{(c)}} \int_{c_i-\delta}^{c_i+\delta} \frac{1}{\mu(\mathcal{C})} 1_{[\tilde{c}_i \in \mathcal{C}]}\, d\tilde{c}_i}_{\text{C term}} \times \underbrace{\prod_{i=1}^{\tilde{I}_n^{(w)}} \int_{w_i-\delta}^{w_i+\delta} \pi_n(\tilde{w}_i)\, d\tilde{w}_i}_{W \text{ term}}
\end{aligned}$$

- W term. The following correspond to equations 138-145 from [Lee, 2000].

$$\text{W term} = \prod_{i=1}^{\tilde{I}_m^{(w)}} \int_{w_i-\delta}^{w_i+\delta} \pi_n(\tilde{w}_i)\, d\tilde{w}_i \tag{240}$$

$$= \prod_{i=1}^{\tilde{I}_n^{(w)}} \int_{w_i-\delta}^{w_i+\delta} (2\pi\sigma_w^2)^{-1/2} \exp\left(-\frac{1}{2\sigma_w^2}\tilde{w}_i^2\right)\, d\tilde{w}_i \tag{241}$$

$$\geq \prod_{i=1}^{\tilde{I}_n^{(w)}} 2\delta \inf_{\tilde{\theta}_i \in [\theta_i-1, \theta_i+1]} \left\{(2\pi\sigma_w^2)^{-1/2} \exp\left(-\frac{1}{2\sigma_w^2}\tilde{w}_i^2\right)\right\} \tag{242}$$

$$\geq \prod_{i=1}^{\tilde{I}_n^{(w)}} \delta\sqrt{\frac{2}{\pi\sigma_w^2}} \exp\left(-\frac{1}{2\sigma_w^2}\zeta_i\right) \qquad \zeta_i := \max\{(\theta_i-1)^2, (\theta_i+1)^2\} \tag{243}$$

$$\geq \left(\delta\sqrt{\frac{2}{\pi\sigma_w^2}}\right)^{\tilde{I}_n^{(w)}} \exp\left(-\frac{1}{2\sigma_w^2}\zeta\tilde{I}_n^{(w)}\right) \qquad \zeta := \max\{\zeta_1, \ldots, \zeta_{\tilde{I}_n^{(w)}}\} \tag{244}$$

$$= \exp\left(-\tilde{I}_n^{(w)}\left[\delta^{-1}\sqrt{\frac{\pi\sigma_w^2}{2}}\right]\right) \exp\left(-\frac{1}{2\sigma_w^2}\zeta\tilde{I}_n^{(w)}\right) \tag{245}$$

$$\tag{246}$$

$$= \exp\left(-\tilde{I}_n^{(w)}\left[\frac{3n^a}{\epsilon}\sqrt{\frac{\pi\sigma_w^2}{2}}\right] - \frac{1}{2\sigma_w^2}\zeta\tilde{I}_n^{(w)}\right) \tag{247}$$

$$= \exp\left(-\tilde{I}_n^{(w)}\left[a\log n - \log\sqrt{\frac{9\pi\sigma_w^2}{2\epsilon^2}} + \frac{1}{2\sigma_w^2}\zeta\right]\right) \tag{248}$$

$$= \exp\left(-\tilde{I}_n^{(w)}\left[2a\log n + \frac{1}{2\sigma_w^2}\zeta\right]\right) \qquad \text{for large } n \tag{249}$$

$$\geq \exp\left(-Dn^a\left[2a\log n + \frac{1}{2\sigma_w^2}\zeta\right]\right) \qquad \tilde{I}_n^{(w)} \leq Dn^a \tag{250}$$

$$\geq \exp(-\nu n) \qquad \text{for large } n \tag{251}$$

Let N_w denote the integer large enough so that Equations 249 and 251 hold for $\nu/3$.

- C term.

$$\prod_{i=1}^{\tilde{I}_n^{(c)}} \int_{c_i-\delta}^{c_i+\delta} \frac{1}{\mu(\mathcal{C})} 1_{[\tilde{c}_i \in \mathcal{C}]}\, d\tilde{c}_i \geq \prod_{i=1}^{\tilde{I}_n^{(c)}} \frac{\delta}{\mu(\mathcal{C})} \tag{252}$$

$$\geq \left(\frac{\delta}{\mu(\mathcal{C})}\right)^{\tilde{I}_n^{(c)}} \tag{253}$$

$$= \exp\left(-Dn^a \log\left[\frac{\mu(\mathcal{C})}{\delta}\right]\right) \tag{254}$$

$$= \exp\left(-Dn^a \log\left[\frac{3\mu(\mathcal{C})n^a}{\epsilon}\right]\right) \tag{255}$$

$$= \exp\left(-Dn^a\left[a\log n - \log\left(\frac{3\mu(\mathcal{C})}{\epsilon}\right)\right]\right) \tag{256}$$

$$= \exp\left(-Dn^a\left[2a\log n\right]\right) \qquad \text{for large } n \tag{257}$$

$$= \exp\left(-2aDn^a \log n\right) \tag{258}$$

$$\geq \exp(-\nu n) \qquad \text{for large } n \tag{259}$$

Let N_c denote the integer large enough so that Equations 257 and 259 hold for $\nu/3$.

- λ^2 term.

$$\int_{\lambda^2-\delta}^{\lambda^2+\delta} \pi_n(\tilde{\lambda^2})\, d\tilde{\lambda^2} = \int_{[\lambda^2-\delta,\lambda^2+\delta]\cap\mathbb{R}^+} \frac{\beta^\alpha}{\Gamma(\alpha)} \tilde{\lambda^2}^{\alpha-1} \exp(-\beta\tilde{\lambda^2})\, d\tilde{\lambda^2} \tag{260}$$

$$\geq \delta\left(\inf_{\tilde{\lambda^2}\in[\lambda^2-\delta,\lambda^2+\delta]\cap\mathbb{R}^+}\left\{\frac{\beta^\alpha}{\Gamma(\alpha)}\tilde{\lambda^2}^{\alpha-1}\exp(-\beta\tilde{\lambda^2})\right\}\right) \tag{261}$$

$$\geq \underbrace{\delta\left(\inf_{\tilde{\lambda^2}\in[\lambda^2-1,\lambda^2+1]\cap\mathbb{R}^+}\left\{\frac{\beta^\alpha}{\Gamma(\alpha)}\tilde{\lambda^2}^{\alpha-1}\exp(-\beta\tilde{\lambda^2})\right\}\right)}_{:=A} \qquad \text{for large } n \tag{262}$$

$$= \delta A \tag{263}$$

$$= \frac{A\epsilon}{3n^a} \tag{264}$$

$$\geq \exp(-\nu n) \qquad \text{for large n} \tag{265}$$

In Equation 261 we note that the length of the interval $[\lambda^2-\delta, \lambda^2+\delta]\cap\mathbb{R}^+$ is at least δ, since $\lambda^2 \in \mathbb{R}^+$. In Equation 262 we note that $\delta < 1$ for large n, allowing us to define the quantity A that does not depend on n. Let N_λ denote the integer large enough so that Equations 262 and 265 hold for $\nu/3$.

Bound product Set $N_2 = \max\{N_w, N_c, N_\lambda\}$. Then for all $n \geq N_2$:

$$\begin{aligned}\pi_n\left(M_\delta\right) &\geq \exp(-n\nu/3)\exp(-n\nu/3)\exp(-n\nu/3) \\ &= \exp(-n\nu)\end{aligned}$$

This shows condition (ii). Thus, the conditions of Theorem 3 are met, so the model is Hellinger consistent. By Corollary B.1 this gives asymptotic consistency.

□

Now we consider the case where the number of hidden units K of the network is a parameter of the model. Since the center parameters follow a Poisson process prior with intensity λ over the region $\mathcal{C}$, then conditional on λ, K follows a Poisson distribution with parameter $\mu(\mathcal{C})\lambda$, where μ is the measure of $\mathcal{C}$. We again denote the estimated regression function by $\hat{g}_n$ with the understanding that the number of hidden units not fixed.

Theorem 7. *(PoRB-Net consistency for homogeneous intensity). For any square integrable ground truth regression function g_0, $\hat{g}_n$ is asymptotically consistent for g as $n \to \infty$, i.e.*

$$\int (\hat{g}_n(x) - g_0(x))^2 dx \xrightarrow{p} 0. \tag{266}$$

Proof. Since the number of hidden units follows a Poisson prior, the proof of this result is exactly as in Theorem 7 of Lee [2000]. Their result relies on their Theorem 8, but we have adapted this result in Theorem 4 to our model and the remainder of the proof requires no additional assumptions regarding the model. Asymptotic consistency follows from Corollary B.2. □

C MODEL SPECIFICATION AND MCMC ALGORITHM

C.1 HOMOGENEOUS INTENSITY

C.2 Notation

name	symbol	domain
centers	$\{c_k\}_{k=1}^K$	$c_k \in \mathbb{R}^D$
weights	$\{w_k\}_{k=1}^K$	$w_k \in \mathbb{R}$
bias	b	$\mathbb{R}$
intensity	λ	$\mathbb{R}$
number of hidden units	K	$\mathbb{K}$

Table 1: Overview of all parameters in the PoRB-Net with homogeneous intensity.

When the meaning is clear, we suppress the subscript and superscripts outside the bracket. For example, $\{c_k\}$ denotes $\{c_k\}_{k=1}^K$.

C.3 Likelihood

$$L(\boldsymbol{\theta}) := p(\{y_n\} \,|\, \{x_n\}, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}(f(x_n; \boldsymbol{\theta}), \sigma_x^2) \tag{267}$$

where $\boldsymbol{\theta} = \{\{w_k\}, b, \{c_k\}, K, \lambda^2\}$ and:

$$f(x; \boldsymbol{\theta}) = b + \sum_{k=1}^{K} w_k \exp\left(-\frac{1}{2} s_0^2 \lambda^2 (x - c_k)^T (x - c_k)\right) \tag{268}$$

is the network output.

C.4 Prior

$$w_k \,|\, K \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \tilde{\sigma}_w^2\right), \quad k = 1 \ldots, K \tag{269}$$

$$b \sim \mathcal{N}(0, \tilde{\sigma}_b^2) \tag{270}$$

$$\{c_k\}_{k=1}^K \,|\, \lambda \sim \exp(-\Lambda) \prod_{k=1}^{K} \lambda(c_k), \quad \text{where } \Lambda = \int_{\mathcal{C}} \lambda(u) du. \tag{271}$$

where $\tilde{\sigma}_w^2 = \sqrt{s_0^2/\pi}\sigma_w^2$. If the intensity function is uniform, we use a Gamma prior:

$$\lambda^2 \sim \text{Gamma}(\alpha, \beta) \tag{272}$$

A note regarding the Poisson process prior. If you do not condition on the number of centers K, the on the centers prior is:

$$p(\{c_k\}_{k=1}^K \,|\, \lambda) = \exp(-\Lambda) \prod_{k=1}^{K} \lambda(c_k), \quad \text{where } \Lambda = \int_{\mathcal{C}} \lambda(u) du. \tag{273}$$

If you do condition on K, the prior on the centers is:

$$p(\{c_k\}_{k=1}^K \,|\, K, \lambda) = K! \prod_{k=1}^{K} \frac{\lambda(c_k)}{\Lambda} \tag{274}$$

Notice you can relate this to the prior when you do not condition on K, since K has a Poisson distribution with parameter Λ.

$$
\begin{aligned}
p(\{c_k\}_{k=1}^K \mid \lambda) &= p(\{c_k\}_{k=1}^K \mid K, \lambda) p(K, \lambda) && (275)\\
&= \left(\cancel{K!} \prod_{k=1}^K \frac{\lambda(c_k)}{\cancel{\Lambda}}\right)\left(\frac{\exp(-\Lambda)\cancel{\Lambda^K}}{\cancel{K!}}\right) && (276)\\
&= \exp(-\Lambda) \prod_{k=1}^K \lambda(c_k) && (277)
\end{aligned}
$$

Joint distribution

$$
\begin{aligned}
&p(\{y_n\}, \{x_n\}, \{w_k\}, b, \{c_k\}, K, \lambda^2)\\
&\quad = p(\{y_n\} \mid \{x_n\}, \{w_k\}, b, \{c_k\}, K, \lambda^2) \times p(\{x_n\}, \{w_k\}, b, \{c_k\}, K, \lambda^2)\\
&\quad \propto p(\{y_n\} \mid \{x_n\}, \underbrace{\{w_k\}, b, K, \lambda^2}_{\boldsymbol{\theta}}) \times p(\{w_k\}, b, \{c_k\}, K, \lambda^2 \mid \cancel{\{x_n\}}) \times \underbrace{\cancel{p(\{x_n\})}}_{1}\\
&\quad \propto p(\{y_n\} \mid \{x_n\}, \boldsymbol{\theta}) \times p(\{w_k\}, b, \{c_k\}, K, \lambda^2)\\
&\quad \propto p(\{y_n\} \mid \{x_n\}, \boldsymbol{\theta}) \times p(b \mid \cancel{\{w_k\}}, \cancel{\{c_k\}}, \cancel{K}, \cancel{\lambda^2}) \times p(\{w_k\}, \{c_k\}, K, \lambda^2)\\
&\quad \propto p(\{y_n\} \mid \{x_n\}, \boldsymbol{\theta}) \times p(b) \times p(\{w_k\} \mid \cancel{\{c_k\}}, K, \cancel{\lambda^2}) \times p(\{c_k\}, K, \lambda^2)\\
&\quad \propto p(\{y_n\} \mid \{x_n\}, \boldsymbol{\theta}) \times p(b) \times p(\{w_k\} \mid K) \times p(\{c_k\}, K, \lambda^2)\\
&\quad \propto p(\{y_n\} \mid \{x_n\}, \boldsymbol{\theta}) \times p(b) \times p(\{w_k\} \mid K) \times p(\{c_k\}, K \mid \lambda^2) \times p(\lambda^2)\\
&\quad \propto \left(\prod_{n=1}^N \mathcal{N}(y_n; f(x_n; \boldsymbol{\theta}), \sigma_y^2)\right) \mathcal{N}(b; 0, \tilde{\sigma}_b^2) \left(\prod_{k=1}^K \mathcal{N}(w_k; 0, \tilde{\sigma}_w^2)\right) \left(\exp(-\Lambda) \prod_{k=1}^K \lambda(c_k)\right) \text{Gamma}(\lambda^2; \alpha, \beta)\\
&\quad \propto \left(\prod_{n=1}^N \exp\left\{-\frac{1}{2\sigma_y^2}(y_n - f(x_n; \boldsymbol{\theta}))^2\right\}\right) \exp\left\{-\frac{1}{2\sigma_b^2} b^2\right\} \left(\prod_{k=1}^K (2\pi\sigma_w^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma_w^2} w_k^2\right\}\right)\\
&\qquad \left(\exp(-\Lambda) \prod_{k=1}^K \lambda(c_k)\right) (\lambda^2)^{\alpha-1} \exp\{-\beta\lambda^2\}\\
&\quad \propto L(\boldsymbol{\theta}) \exp\left\{-\frac{1}{2\sigma_b^2} b^2\right\} (2\pi\sigma_w^2)^{-K/2} \exp\left\{-\frac{1}{2\sigma_w^2} \sum_{k=1}^K w_k^2\right\} \exp(-\Lambda) \left(\prod_{k=1}^K \lambda(c_k)\right) (\lambda^2)^{\alpha-1} \exp\{-\beta\lambda^2\}
\end{aligned}
$$

C.5 Gibbs steps

There are 3 steps:

1. Update $\{w_k\}_{k=1}^K, b, \{c_k\}_{k=1}^K$ with HMC
2. Update K with birth or death MH steps
3. Update λ^2 with an MH step (only if intensity is uniform)

Updating $\{w_k\}_{k=1}^K, b, \{c_k\}_{k=1}^K$ with HMC The full conditional distribution of the weight, bias, and center parameters is given by:

$$
\begin{aligned}
&p(\{w_k\}, b, \{c_k\} \mid \_\_\_)\\
&\qquad \propto p(\{y_n\}, \{x_n\}, \{w_k\}, b, \{c_k\}, K, \lambda^2)\\
&\qquad \propto L(\boldsymbol{\theta}) \exp\left\{-\frac{1}{2\sigma_b^2} b^2\right\} \cancel{(2\pi\sigma_w^2)^{-K/2}} \exp\left\{-\frac{1}{2\sigma_w^2} \sum_{k=1}^K w_k^2\right\} \cancel{\exp(-\Lambda)} \left(\prod_{k=1}^K \lambda(c_k)\right) \cancel{(\lambda^2)^{\alpha-1}} \cancel{\exp\{-\beta\lambda^2\}}
\end{aligned}
$$

$$\propto L(\boldsymbol{\theta}) \exp\left\{-\frac{1}{2\sigma_b^2} b^2\right\} \exp\left\{-\frac{1}{2\sigma_w^2} \sum_{k=1}^{K} w_k^2\right\} \prod_{k=1}^{K} \lambda(c_k).$$

We update the weight, bias, and center parameters using HMC with $-\log p(\{w_k\}, b, \{c_k\} \mid \_\_\_)$ as the potential energy function.

Updating K We update the network width K with birth or death MH steps of hidden units. Each iteration of the sampler, we perform either a birth or death step with equal probability. The full conditional distribution of the network width is given by:

$$\begin{aligned}
p(K \mid ___) &\propto p(\{y_n\}, \{x_n\}, \{w_k\}, b, \{c_k\}, K, \lambda^2) \\
&\propto L(\boldsymbol{\theta}) \cancel{\exp\left\{-\frac{1}{2\sigma_b^2} b^2\right\}} (2\pi\sigma_w^2)^{-K/2} \exp\left\{-\frac{1}{2\sigma_w^2} \sum_{k=1}^{K} w_k^2\right\} \cancel{\exp(-\Lambda)} \left(\prod_{k=1}^{K} \lambda(c_k)\right) \cancel{(\lambda^2)^{\alpha-1} \exp\{-\beta\lambda^2\}} \\
&\propto L(\boldsymbol{\theta}) (2\pi\sigma_w^2)^{-K/2} \exp\left\{-\frac{1}{2\sigma_w^2} \sum_{k=1}^{K} w_k^2\right\} \prod_{k=1}^{K} \lambda(c_k).
\end{aligned}$$

Birth step

A proposal for a birth consists of two parameter proposals:

- Sample $w'_{K+1} \sim \mathcal{N}(w_{K+1}; 0, \sigma_w^2)$
- Sample $c'_{K+1} \sim \lambda(c)/\Lambda$ (i.e., from the prior intensity conditioned on the number of units K).

Therefore the proposal density is:

$$q(K \to K+1) = \mathcal{N}(w_{K+1}; 0, \sigma_w^2) \frac{\lambda(c_{K+1})}{\Lambda} \tag{278}$$

A proposal for a death consists only of sampling a unit uniformly at random. The proposal density for a birth is therefore:

$$q(K \to K-1) = 1/K \tag{279}$$

The ratio of proposal densities is therefore:

$$\frac{q(K+1 \to K)}{q(K \to K+1)} = \frac{\Lambda}{(K+1)\lambda(c_{K+1})\mathcal{N}(w_{K+1}; 0, \sigma_w^2)} \tag{280}$$

To derive the acceptance ratio, we next derive the ratio of posterior probabilities, letting $\boldsymbol{\theta}' = \{\{w_k\}, b, \{c_k\}, K+1, \lambda^2\}$:

$$\frac{p(K+1 \mid \_\_\_)}{p(K \mid \_\_\_)} \tag{281}$$

$$\propto \frac{L(\boldsymbol{\theta}')(2\pi\sigma_w^2)^{-K/2} \exp\left\{-\frac{1}{2\sigma_w^2} \sum_{k=1}^{K+1} w_k^2\right\} \prod_{k=1}^{K+1} \lambda(c_k)}{L(\boldsymbol{\theta}) \cancel{(2\pi\sigma_w^2)^{-K/2}} \cancel{\exp\left\{-\frac{1}{2\sigma_w^2} \sum_{k=1}^{K} w_k^2\right\}} \cancel{\prod_{k=1}^{K} \lambda(c_k)}} \tag{282}$$

$$\propto \frac{L(\boldsymbol{\theta}')\mathcal{N}(w_{K+1}; 0, \sigma_w^2)\lambda(c_{K+1})}{L(\boldsymbol{\theta})} \tag{283}$$

The acceptance rate is then:

$$a_{\text{birth}} = \frac{p(K+1 \mid \_\_\_)}{p(K \mid \_\_\_)} \frac{q(K+1 \to K)}{q(K \to K+1)} \tag{284}$$

$$= \frac{L(\boldsymbol{\theta}') \cancel{\mathcal{N}(w_{K+1}; 0, \sigma_w^2)} \cancel{\lambda(c_{K+1})}}{L(\boldsymbol{\theta})} \frac{\Lambda}{(K+1) \cancel{\lambda(c_{K+1})} \cancel{\mathcal{N}(w_{K+1}; 0, \sigma_w^2)}} \tag{285}$$

$$= \frac{L(\boldsymbol{\theta}')}{L(\boldsymbol{\theta})} \frac{\Lambda}{K+1} \tag{286}$$

Death step

Now letting $\boldsymbol{\theta}' = \{\{w_k\}, b, \{c_k\}, K-1, \lambda^2\}$, the acceptance probability for a death step can be derived analogously to the birth step discussed above:

$$a_{\text{death}} = \frac{p(K-1 \,|\, \_\_\_)}{p(K \,|\, \_\_\_)} \frac{q(K \to K-1)}{q(K-1 \to K)} \tag{287}$$

$$= \frac{L(\boldsymbol{\theta}')}{L(\boldsymbol{\theta})\cancel{\mathcal{N}(w_K; 0, \sigma_w^2)}\cancel{\lambda(c_K)}} \frac{(K+1)\cancel{\lambda(c_K)}\cancel{\mathcal{N}(w_K; 0, \sigma_w^2)}}{\Lambda} \tag{288}$$

$$= \frac{L(\boldsymbol{\theta}')}{L(\boldsymbol{\theta})} \frac{K+1}{\Lambda} \tag{289}$$

Updating λ^2 We only update the intensity function when it is uniform (i.e., $\lambda(c) = \lambda$ for any c). Therefore, the integral of the intensity is given by $\Lambda := \int_{\mathcal{C}} \lambda(c) dc = \mu(\mathcal{C})\lambda$. The full conditional distribution is therefore:

$$\begin{aligned}
p(\lambda^2 \,|\, ___) &\propto p(\{y_n\}, \{x_n\}, \{w_k\}, b, \{c_k\}, K, \lambda^2) \\
&\propto L(\boldsymbol{\theta})\cancel{\exp\left\{-\frac{1}{2\sigma_b^2} b^2\right\}}\cancel{(2\pi\sigma_w^2)^{-K/2}}\cancel{\exp\left\{-\frac{1}{2\sigma_w^2}\sum_{k=1}^{K} w_k^2\right\}} \exp(-\Lambda) \left(\prod_{k=1}^{K} \lambda(c_k)\right) (\lambda^2)^{\alpha-1} \exp\{-\beta\lambda^2\} \\
&\propto L(\boldsymbol{\theta}) \exp(-\mu(\mathcal{C})\lambda) \lambda^K (\lambda^2)^{\alpha-1} \exp\{-\beta\lambda^2\} \\
&\propto L(\boldsymbol{\theta}) \lambda^{K+2(\alpha-1)} \exp\{-\mu(\mathcal{C})\lambda - \beta\lambda^2\}
\end{aligned}$$

To update λ^2 we use a Metropolis-Hastings step with a normal proposal distribution centered around the current value of λ^2. Since this distribution is symmetric, the proposal distributions cancel out in the acceptance ratio. Letting $(\lambda')^2$ denote the proposed value of λ^2 and $\boldsymbol{\theta}' = \{\{w_k\}, b, \{c_k\}, K, (\lambda')^2\}$, the acceptance rate is therefore:

$$a = \frac{p(\lambda^{2*} \,|\, \_\_\_)}{p((\lambda')^2 \,|\, \_\_\_)} \tag{290}$$

$$\propto \frac{L(\boldsymbol{\theta}')(\lambda')^{K+2(\alpha-1)} \exp\{-\mu(\mathcal{C})\lambda - \beta(\lambda')^2\}}{L(\boldsymbol{\theta})\lambda^{K+2(\alpha-1)} \exp\{-\mu(\mathcal{C})\lambda - \beta\lambda^2\}} \tag{291}$$

$$\propto \frac{L(\boldsymbol{\theta}')}{L(\boldsymbol{\theta})} \exp\{-\mu(\mathcal{C})(\lambda' - \lambda)\} \exp\{-\beta((\lambda')^2 - \lambda^2)\} \left(\frac{\lambda'}{\lambda}\right)^{K+2(\alpha-1)} \tag{292}$$

C.6 INHOMOGENEOUS INTENSITY

C.7 Notation

name	symbol	domain
centers	$\{c_k\}_{k=1}^K$	$c_k \in \mathbb{R}^D$
weights	$\{w_k\}_{k=1}^K$	$w_k \in \mathbb{R}$
bias	b	$\mathbb{R}$
thinned centers	$\{\tilde{c}_m\}_{m=1}^M$	$\tilde{c}_m \in \mathbb{R}^D$
GP function values	$\{g_k\}_{k=1}^K$	$\tilde{g}_m \in \mathbb{R}$
thinned GP function values	$\{\tilde{g}_k\}_{m=1}^M$	$\tilde{g}_m \in \mathbb{R}$
intensity upper bound	λ^*	$\mathbb{R}$
number of hidden units	K	$\mathbb{K}$
number of thinned centers	M	$\mathbb{N}$

Table 2: Overview of all parameters in PoRB-Net when an input dependent intensity is inferred.

When the meaning is clear, we suppress the subscript and superscripts outside the bracket. For example, $\{c_k\}$ denotes $\{c_k\}_{k=1}^K$. For convenience we also use the following notation (also applied analogously to the centers):

symbol	meaning
$\tilde{\mathbf{g}}_M$	vector of $\{\tilde{g}_m\}_{k=1}^M$
$\mathbf{g}_K$	vector of $\{g_k\}_{k=1}^K$
$\mathbf{g}_{M+K}$	vector of $\{g_m\}_{k=1}^M$ and $\{g_k\}_{k=1}^K$
$\mathbf{g}_{M+K+1}$	$\mathbf{g}_{M+K}$ with one additional component
$\mathbf{g}_{M+K-i}$	$\mathbf{g}_{M+K}$ without component i

Table 3: Alternative notation for convenience. The same subscripts are also applied to the centers.

C.8 Likelihood

$$L(\boldsymbol{\theta}) := p(\{y_n\} \,|\, \{x_n\}, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}(f(x_n; \boldsymbol{\theta}), \sigma_x^2) \tag{293}$$

where $\boldsymbol{\theta} = \{\{w_k\}, b, \{c_k\}, \{g_k\}, \lambda^*\}$ and:

$$f(x; \boldsymbol{\theta}) = b + \sum_{k=1}^{K} w_k \exp(-s_k^2 (x - c_k)^T (x - c_k)) \tag{294}$$

$$s_k^2 = (s_0 \lambda(c_k))^2 = (s_0 \lambda^* \sigma(h(c_k)))^2 \tag{295}$$

is the network, where $\sigma(\cdot)$ is the sigmoid (logistic) function and h is a GP.

C.9 Prior

$$w_k \,|\, K \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \tilde{\sigma}_w^2\right), \quad k = 1 \ldots, K \tag{296}$$

$$b \sim \mathcal{N}(0, \tilde{\sigma}_b^2) \tag{297}$$

$$\lambda^* \sim \text{Gamma}(\alpha, \beta) \tag{298}$$

$$p(\{c_k\},\{\tilde{c}_m\}, \{g_k\}, \{\tilde{g}_m\}, K, M \,|\, \lambda^*) \tag{299}$$

$$\propto (\lambda^*)^{K+M} \exp\left(-\lambda^* \mu(\mathcal{C})\right) \prod_{k=1}^{K} \sigma(g_k) \prod_{m=1}^{M} \sigma(-\tilde{g}_m) \times \mathcal{GP}(\{g_k\}, \{\tilde{g}_m\} \,|\, \{c_k\}, \{\tilde{c}_m\}) \tag{300}$$

where $\mu(\mathcal{C})$ is the measure of $\mathcal{C}$ and $\tilde{\sigma}_w^2 = \sqrt{s_0^2/\pi}\sigma_w^2$.

Note can write the GP prior in a few ways (just different notation):

$$\mathcal{GP}(\{g_k\}, \{\tilde{g}_m\} \,|\, \{c_k\}, \{\tilde{c}_m\}) = p(\mathbf{g}_{M+K} \,|\, \mathbf{c}_{M+K}) \tag{301}$$

$$= (2\pi)^{-(M+K)/2} \,|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2} \mathbf{g}_{M+K}^T \Sigma^{-1} \mathbf{g}_{M+K}\right\} \tag{302}$$

where $\Sigma = \text{kernel}(\mathbf{c}_{M+K}, \mathbf{c}_{M+K})$ is the $M + K \times M + K$ kernel matrix evaluated on all of the center parameters $\mathbf{c}_{M+K}$.

Joint distribution

$$p(\{y_n\}, \{x_n\}, \{w_k\}, b, \{c_k\}, \{\tilde{c}_m\}, \{g_k\}, \{\tilde{g}_m\}, K, M, \lambda^*)$$
$$= p(\{y_n\} \,|\, \{x_n\}, \{w_k\}, b, \{c_k\}, \cancel{\{\tilde{c}_m\}}, \{g_k\}, \cancel{\{\tilde{g}_m\}}, K, \cancel{M}, \lambda^*) \times p(\{x_n\}, \{w_k\}, b, \{c_k\}, \{\tilde{c}_m\}, \{g_k\}, \{\tilde{g}_m\}, K, M, \lambda^*)$$

$$
\begin{aligned}
&\propto p(\{y_n\} \mid \{x_n\}, \underbrace{\{w_k\}, b, \{c_k\}, \{g_k\}, K, \lambda^*}_{\boldsymbol{\theta}}) \times p(\{w_k\}, b, \{c_k\}, \{\tilde{c}_m\}, \{g_k\}, \{\tilde{g}_m\}, K, M, \lambda^* \mid \cancel{\{x_n\}}) \times \underbrace{\cancel{p(\{x_n\})}}_{1} \\
&\propto p(\{y_n\} \mid \{x_n\}, \boldsymbol{\theta}) \times p(\{w_k\}, b, \{c_k\}, \{\tilde{c}_m\}, \{g_k\}, \{\tilde{g}_m\}, K, M, \lambda^*) \\
&\propto p(\{y_n\} \mid \{x_n\}, \boldsymbol{\theta}) \times p(b \mid \cancel{\{w_k\}}, \cancel{\{c_k\}}, \cancel{\{\tilde{c}_m\}}, \cancel{\{g_k\}}, \cancel{\{\tilde{g}_m\}}, \cancel{K}, \cancel{M}, \cancel{\lambda^*}) \times p(\{w_k\}, \{c_k\}, \{\tilde{c}_m\}, \{g_k\}, \{\tilde{g}_m\}, K, M, \lambda^*) \\
&\propto p(\{y_n\} \mid \{x_n\}, \boldsymbol{\theta}) \times p(b) \times p(\{w_k\} \mid \cancel{\{c_k\}}, \cancel{\{\tilde{c}_m\}}, \cancel{\{g_k\}}, \cancel{\{\tilde{g}_m\}}, K, \cancel{M}, \cancel{\lambda^*}) \times p(\{c_k\}, \{\tilde{c}_m\}, \{g_k\}, \{\tilde{g}_m\}, K, M, \lambda^*) \\
&\propto p(\{y_n\} \mid \{x_n\}, \boldsymbol{\theta}) \times p(b) \times p(\{w_k\} \mid K) \times p(\{c_k\}, \{\tilde{c}_m\}, \{g_k\}, \{\tilde{g}_m\}, K, M, \lambda^*) \\
&\propto p(\{y_n\} \mid \{x_n\}, \boldsymbol{\theta}) \times p(b) \times p(\{w_k\} \mid K) \times p(\{c_k\}, \{\tilde{c}_m\}, \{g_k\}, \{\tilde{g}_m\}, K, M \mid \lambda^*) \times p(\lambda^*) \\
&\propto \left(\prod_{n=1}^{N} \mathcal{N}(y_n; f(x_n; \boldsymbol{\theta}), \sigma_y^2) \right) \mathcal{N}(b; 0, \tilde{\sigma}_b^2) \left(\prod_{k=1}^{K} \mathcal{N}(w_k; 0, \tilde{\sigma}_w^2) \right) \\
&\quad \left((\lambda^*)^{K+M} \exp\left(-\lambda^* \mu(\mathcal{C})\right) \prod_{k=1}^{K} \sigma(g_k) \prod_{m=1}^{M} \sigma(-\tilde{g}_m) \times \mathcal{GP}(\{g_k\}, \{\tilde{g}_m\} \mid \{c_k\}, \{\tilde{c}_m\}) \right) \text{Gamma}(\lambda^*; \alpha, \beta) \\
&\propto \left(\prod_{n=1}^{N} \exp\left\{ -\frac{1}{2\sigma_y^2} (y_n - f(x_n; \boldsymbol{\theta}))^2 \right\} \right) \exp\left\{ -\frac{1}{2\sigma_b^2} b^2 \right\} \left(\prod_{k=1}^{K} (2\pi\sigma_w^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma_w^2} w_k^2 \right\} \right) \\
&\quad \left((\lambda^*)^{K+M} \exp\left(-\lambda^* \mu(\mathcal{C})\right) \prod_{k=1}^{K} \sigma(g_k) \prod_{m=1}^{M} \sigma(-\tilde{g}_m)\, \mathcal{GP}(\{g_k\}, \{\tilde{g}_m\} \mid \{c_k\}, \{\tilde{c}_m\}) \right) (\lambda^*)^{\alpha-1} \exp\{-\beta\lambda^*\} \\
&\propto L(\boldsymbol{\theta}) \exp\left\{ -\frac{1}{2\sigma_b^2} b^2 \right\} (2\pi\sigma_w^2)^{-K/2} \exp\left\{ -\frac{1}{2\sigma_w^2} \sum_{k=1}^{K} w_k^2 \right\} (\lambda^*)^{\alpha+K+M-1} \exp\left(-\lambda^*(\beta + \mu(\mathcal{C}))\right) \prod_{k=1}^{K} \sigma(g_k) \\
&\quad \prod_{m=1}^{M} \sigma(-\tilde{g}_m)\, \mathcal{GP}(\{g_k\}, \{\tilde{g}_m\} \mid \{c_k\}, \{\tilde{c}_m\})
\end{aligned}
$$

C.10 Gibbs steps

There are 6 steps:

1. Update $\{w_k\}_{k=1}^K, b, \{c_k\}_{k=1}^K$ with HMC
2. Update K with birth or death MH steps
3. Update λ^* with an MH step (optional)
4. Update M with birth or death MH steps
5. Update $\{\tilde{c}_m\}_{m=1}^M$ with a MH step
6. Update $\{g_k\}_{k=1}^K$ and $\{\tilde{g}_m\}_{m=1}^M$ with HMC

Updating $\{w_k\}_{k=1}^K, b, \{c_k\}_{k=1}^K$ with HMC The full conditional distribution of the weight, center

$$
\begin{aligned}
p(\{w_k\}, b, \{c_k\} \mid \_\_\_) &\propto p(\{y_n\}, \{x_n\}, \{w_k\}, b, \{c_k\}, \{\tilde{c}_m\}, \{g_k\}, \{\tilde{g}_m\}, K, M, \lambda^*) \\
&\propto L(\boldsymbol{\theta}) \exp\left\{ -\frac{1}{2\sigma_b^2} b^2 \right\} \cancel{(2\pi\sigma_w^2)^{-K/2}} \exp\left\{ -\frac{1}{2\sigma_w^2} \sum_{k=1}^{K} w_k^2 \right\} \\
&\quad \cancel{(\lambda^*)^{\alpha+K+M-1}} \cancel{\exp\left(-\lambda^*(\beta + \mu(\mathcal{C}))\right)} \prod_{k=1}^{K} \sigma(g_k) \cancel{\prod_{m=1}^{M} \sigma(-\tilde{g}_m)} \times p(\mathbf{g}_{M+K} \mid \mathbf{c}_{M+K}) \\
&\propto L(\boldsymbol{\theta}) \exp\left\{ -\frac{1}{2\sigma_b^2} b^2 \right\} \exp\left\{ -\frac{1}{2\sigma_w^2} \sum_{k=1}^{K} w_k^2 \right\}
\end{aligned}
$$

$$\cancel{\prod_{k=1}^{K}\sigma(g_k)\times(2\pi)^{-(M+K)/2}}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{g}_{M+K}^T\Sigma^{-1}\mathbf{g}_{M+K}\right\}$$

$$\propto L(\boldsymbol{\theta})\exp\left\{-\frac{1}{2\sigma_b^2}b^2\right\}\exp\left\{-\frac{1}{2\sigma_w^2}\sum_{k=1}^{K}w_k^2\right\}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{g}_{M+K}^T\Sigma^{-1}\mathbf{g}_{M+K}\right\}$$

Updating K

$$\begin{aligned}
p(K\,|\,___) &\propto p(\{y_n\},\{x_n\},\{w_k\},b,\{c_k\},\{\tilde{c}_m\},\{g_k\},\{\tilde{g}_m\},K,M,\lambda^*)\\
&\propto L(\boldsymbol{\theta})\cancel{\exp\left\{-\frac{1}{2\sigma_b^2}b^2\right\}}(2\pi\sigma_w^2)^{-K/2}\exp\left\{-\frac{1}{2\sigma_w^2}\sum_{k=1}^{K}w_k^2\right\}\\
&\quad(\lambda^*)^{\cancel{\alpha}+K+\cancel{M-1}}\cancel{\exp\left(-\lambda^*(\beta+\mu(\mathcal{C}))\right)}\prod_{k=1}^{K}\sigma(g_k)\cancel{\prod_{m=1}^{M}\sigma(-\tilde{g}_m)}\times p(\mathbf{g}_{M+K}\,|\,\mathbf{c}_{M+K})\\
&\propto L(\boldsymbol{\theta})(2\pi\sigma_w^2)^{-K/2}\exp\left\{-\frac{1}{2\sigma_w^2}\sum_{k=1}^{K}w_k^2\right\}(\lambda^*)^K\prod_{k=1}^{K}\sigma(g_k)\times p(\mathbf{g}_{M+K}\,|\,\mathbf{c}_{M+K})
\end{aligned}$$

Birth step

A proposal for a birth consists of three steps:

- Sample $w'_{K+1}\sim\mathcal{N}(0,\sigma_w^2)$
- Sample $c'_{K+1}\sim 1/\mu(\mathcal{C})$ uniformly
- Sample $g'_{K+1}\sim p(g_{K+1}\mid c'_{K+1},\mathbf{c}_{M+K},\mathbf{g}_{M+K})$

Therefore the proposal density is:

$$q(K\to K+1)\propto\mathcal{N}(w_{K+1};0,\sigma_w^2)p(g'_{K+1}\mid c'_{K+1},\mathbf{c}_{M+K},\mathbf{g}_{M+K})/\mu(\mathcal{C}) \tag{303}$$

A proposal for a death consists only of sampling a unit uniformly at random, so the proposal density for a death is:

$$q(K\to K-1)=\frac{1}{K} \tag{304}$$

The ratio of proposal densities is therefore:

$$\frac{q(K+1\to K)}{q(K\to K+1)}=\frac{\mu(\mathcal{C})}{(K+1)\mathcal{N}(w_{K+1};0,\sigma_w^2)p(g'_{K+1}\mid c'_{K+1},\mathbf{c}_{M+K},\mathbf{g}_{M+K})} \tag{305}$$

Ratio of posterior probabilities:

$$\frac{p(K+1\,|\,\_\_\_)}{p(K\,|\,\_\_\_)} \tag{306}$$

$$=\frac{L(\boldsymbol{\theta}')(2\pi\sigma_w^2)^{-(K+1)/2}\exp\left\{-\frac{1}{2\sigma_w^2}\sum_{k=1}^{K+1}w_k^2\right\}(\lambda^*)^{K+1}\prod_{k=1}^{K+1}\sigma(g_k)p(\mathbf{g}_{M+K+1}\,|\,\mathbf{c}_{M+K+1})}{L(\boldsymbol{\theta})\cancel{(2\pi\sigma_w^2)^{-K/2}}\cancel{\exp\left\{-\frac{1}{2\sigma_w^2}\sum_{k=1}^{K}w_k^2\right\}}\cancel{(\lambda^*)^K}\cancel{\prod_{k=1}^{K}\sigma(g_k)}p(\mathbf{g}_{M+K}\,|\,\mathbf{c}_{M+K})} \tag{307}$$

$$=\frac{L(\boldsymbol{\theta}')(2\pi\sigma_w^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma_w^2}w_{K+1}^2\right\}\lambda^*\sigma(g'_{K+1})p(g'_{K+1}\,|\,c'_{K+1},\mathbf{c}_{M+K},\mathbf{g}_{M+K})p(\mathbf{g}_{M+K}\,|\,\cancel{c'_{K+1}},\mathbf{c}_{M+K})}{L(\boldsymbol{\theta})p(\mathbf{g}_{M+K}\,|\,\mathbf{c}_{M+K})} \tag{308}$$

$$=\frac{L(\boldsymbol{\theta}')(2\pi\sigma_w^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma_w^2}w_{K+1}^2\right\}\lambda^*\sigma(g'_{K+1})p(g'_{K+1}\,|\,c'_{K+1},\mathbf{c}_{M+K},\mathbf{g}_{M+K})\cancel{p(\mathbf{g}_{M+K}\,|\,\mathbf{c}_{M+K})}}{L(\boldsymbol{\theta})\cancel{p(\mathbf{g}_{M+K}\,|\,\mathbf{c}_{M+K})}} \tag{309}$$

$$= \frac{L(\boldsymbol{\theta}')\mathcal{N}(w_{K+1}; 0, \sigma_w^2)\lambda^* \sigma(g'_{K+1}) p(g'_{K+1} \mid c'_{K+1}, \mathbf{c}_{M+K}, \mathbf{g}_{M+K})}{L(\boldsymbol{\theta})} \tag{310}$$

The acceptance rate:

$$a = \frac{p(K+1 \mid \_\_\_)}{p(K \mid \_\_\_)} \frac{q(K+1 \to K)}{q(K \to K+1)} \tag{311}$$

$$= \frac{L(\boldsymbol{\theta}')\cancel{\mathcal{N}(w_{K+1}; 0, \sigma_w^2)}\lambda^* \sigma(g'_{K+1}) p(g'_{K+1} \mid \cancel{c'_{K+1}, \mathbf{c}_{M+K}, \mathbf{g}_{M+K}})\mu(\mathcal{C})}{L(\boldsymbol{\theta})(K+1)\cancel{\mathcal{N}(w_{K+1}; 0, \sigma_w^2)} \cancel{p(g'_{K+1} \mid c'_{K+1}, \mathbf{c}_{M+K}, \mathbf{g}_{M+K})}} \tag{312}$$

$$= \frac{L(\boldsymbol{\theta}')}{L(\boldsymbol{\theta})} \frac{\lambda^* \sigma(g'_{K+1})\mu(\mathcal{C})}{K+1} \tag{313}$$

Death step

The acceptance rate:

$$a = \frac{p(K-1 \mid \_\_\_)}{p(K \mid \_\_\_)} \frac{q(K-1 \to K)}{q(K \to K-1)} \tag{314}$$

$$= \frac{L(\boldsymbol{\theta})(K)\cancel{\mathcal{N}(w_K; 0, \sigma_w^2)} p(g_K \mid c_K, \mathbf{c}_{M+K}, \mathbf{g}_{M+K})}{L(\boldsymbol{\theta}')\cancel{\mathcal{N}(w_K; 0, \sigma_w^2)}\lambda^* \sigma(g_K) \cancel{p(g_K \mid c_K, \mathbf{c}_{M+K}, \mathbf{g}_{M+K})}\mu(\mathcal{C})} \tag{315}$$

$$= \frac{L(\boldsymbol{\theta}')}{L(\boldsymbol{\theta})} \frac{K}{\lambda^* \sigma(g_K)\mu(\mathcal{C})} \tag{316}$$

Updating λ^* We use MH. Here is the full conditional:

$$p(\lambda^* \mid \_\_\_) \propto p(\{y_n\}, \{x_n\}, \{w_k\}, b, \{c_k\}, \{\tilde{c}_m\}, \{g_k\}, \{\tilde{g}_m\}, K, M)$$

$$\propto \cancel{L(\boldsymbol{\theta})\exp\left\{-\frac{1}{2\sigma_b^2}b^2\right\}(2\pi\sigma_w^2)^{-K/2}\exp\left\{-\frac{1}{2\sigma_w^2}\sum_{k=1}^{K} w_k^2\right\}}$$

$$(\lambda^*)^{\alpha+K+M-1}\exp\left(-\lambda^*(\beta+\mu(\mathcal{C}))\right)\cancel{\prod_{k=1}^{K}\sigma(g_k)\prod_{m=1}^{M}\sigma(-\tilde{g}_m)} \times \cancel{\mathcal{GP}(\{g_k\}, \{\tilde{g}_m\} \mid \{c_k\}, \{\tilde{c}_m\})}$$

$$\propto L(\boldsymbol{\theta})(\lambda^*)^{\alpha+K+M-1}\exp\left(-\lambda^*(\beta+\mu(\mathcal{C}))\right)$$

To propose a new intensity upper bound $\lambda^{*'}$ we use a normal distribution centered on the current value of λ^*. Since this distribution is symmetric, the proposal distributions cancel out in the acceptance ratio. The acceptance rate is therefore:

$$a = \frac{p(\lambda^{*'} \mid \_\_\_)}{p(\lambda^* \mid \_\_\_)} \tag{317}$$

$$= \frac{L(\boldsymbol{\theta}')(\lambda^{*'})^{\alpha+K+M-1}\exp\left(-\lambda^{*'}(\beta+\mu(\mathcal{C}))\right)}{L(\boldsymbol{\theta})(\lambda^*)^{\alpha+K+M-1}\exp\left(-\lambda^*(\beta+\mu(\mathcal{C}))\right)} \tag{318}$$

$$= \frac{L(\boldsymbol{\theta}')}{L(\boldsymbol{\theta})}\left(\frac{\lambda^{*'}}{\lambda^*}\right)^{\alpha+M+K-1}\exp\left(-(\lambda^{*'}-\lambda^*)(\beta+\mu(\mathcal{C}))\right) \tag{319}$$

Updating M

$$p(M \mid \_\_\_) \propto p(\{y_n\}, \{x_n\}, \{w_k\}, b, \{c_k\}, \{\tilde{c}_m\}, \{g_k\}, \{\tilde{g}_m\}, K, M, \lambda^*)$$

$$\propto \cancel{L(\boldsymbol{\theta})\exp\left\{-\frac{1}{2\sigma_b^2}b^2\right\}(2\pi\sigma_w^2)^{-K/2}\exp\left\{-\frac{1}{2\sigma_w^2}\sum_{k=1}^{K} w_k^2\right\}}$$

$$(\lambda^*)^{\alpha+K+M-1}\cancel{\exp\left(-\lambda^*(\beta+\mu(\mathcal{C}))\right)}\cancel{\prod_{k=1}^{K}\sigma(g_k)}\prod_{m=1}^{M}\sigma(-\tilde{g}_m)\times p(\mathbf{g}_{M+K}\,|\,\mathbf{c}_{M+K})$$

$$\propto L(\boldsymbol{\theta})(\lambda^*)^M\prod_{m=1}^{M}\sigma(-\tilde{g}_m)\times p(\mathbf{g}_{M+K}\,|\,\mathbf{c}_{M+K})$$

Birth step A proposal for a birth consists of two parameter proposals:

- Sample $\tilde{c}'_{M+1}\sim 1/\mu(\mathcal{C})$ uniformly over $\mathcal{C}$.
- Sample $\tilde{g}'_{M+1}\sim p(\tilde{g}_{M+1}\,|\,\tilde{c}_{M+1},\mathbf{c}_{M+K},\mathbf{g}_{M+K})$

The proposal probability for a birth is therefore:

$$q(M\to M+1)=\frac{p(\tilde{g}'_{M+1}\,|\,\tilde{c}_{M+1},\mathbf{c}'_{M+K},\mathbf{g}_{M+K})}{\mu(\mathcal{C})} \tag{320}$$

A proposal for a death consists only of sampling a hidden unit uniformly at random. The proposal proability for a death is therefore:

$$q(M\to M-1)=\frac{1}{M} \tag{321}$$

The ratio of proposal densities is:

$$\frac{q(M+1\to M)}{q(M\to M+1)}=\frac{\mu(\mathcal{C})}{(M+1)p(\tilde{g}'_{M+1}\,|\,\tilde{c}_{M+1},\mathbf{c}'_{M+K},\mathbf{g}_{M+K})} \tag{322}$$

The ratio of posterior probabilities is:

$$\frac{p(M+1\,|\,\_\_\_)}{p(M\,|\,\_\_\_)}=\frac{L(\boldsymbol{\theta}')(\lambda^*)^{M+1}\prod_{m=1}^{M+1}\sigma(-\tilde{g}_m)\times p(\mathbf{g}_{M+K+1}\,|\,\mathbf{c}_{M+K+1})}{L(\boldsymbol{\theta})\cancel{(\lambda^*)^M}\cancel{\prod_{m=1}^{M}\sigma(-\tilde{g}_m)}\times p(\mathbf{g}_{M+K}\,|\,\mathbf{c}_{M+K})} \tag{323}$$

$$=\frac{L(\boldsymbol{\theta}')\lambda^*\sigma(-\tilde{g}'_{M+1})\times p(\tilde{g}'_{M+1}\,|\,\tilde{c}'_{M+1},\mathbf{c}_{M+K},\mathbf{g}_{M+K})\times p(\mathbf{g}_{M+K}\,|\,\cancel{\tilde{c}_{M+1}},\mathbf{c}_{M+K})}{L(\boldsymbol{\theta})p(\mathbf{g}_{M+K}\,|\,\mathbf{c}_{M+K})} \tag{324}$$

$$=\frac{L(\boldsymbol{\theta}')\lambda^*\sigma(-\tilde{g}'_{M+1})\times p(\tilde{g}'_{M+1}\,|\,\tilde{c}'_{M+1},\mathbf{c}_{M+K},\mathbf{g}_{M+K})\times \cancel{p(\mathbf{g}_{M+K}\,|\,\mathbf{c}_{M+K})}}{L(\boldsymbol{\theta})\cancel{p(\mathbf{g}_{M+K}\,|\,\mathbf{c}_{M+K})}} \tag{325}$$

$$=\frac{L(\boldsymbol{\theta}')\lambda^*\sigma(-\tilde{g}'_{M+1})\times p(\tilde{g}'_{M+1}\,|\,\tilde{c}'_{M+1},\mathbf{c}_{M+K},\mathbf{g}_{M+K})}{L(\boldsymbol{\theta})} \tag{326}$$

The acceptance rate is therefore:

$$a=\frac{p(M+1\,|\,\_\_\_)}{p(M\,|\,\_\_\_)}\frac{q(M+1\to M)}{q(M\to M+1)} \tag{327}$$

$$=\frac{L(\boldsymbol{\theta}')\lambda^*\sigma(-\tilde{g}'_{M+1})\times \cancel{p(\tilde{g}'_{M+1}\,|\,\tilde{c}'_{M+1},\mathbf{c}_{M+K},\mathbf{g}_{M+K})}}{\cancel{L(\boldsymbol{\theta})}}\frac{\mu(\mathcal{C})}{(M+1)\cancel{p(\tilde{g}'_{M+1}\,|\,\tilde{c}'_{M+1},\mathbf{c}_{M+K},\mathbf{g}_{M+K})}} \tag{328}$$

$$=\frac{L(\boldsymbol{\theta}')\lambda^*\mu(\mathcal{C})}{L(\boldsymbol{\theta})(M+1)(1+\exp(\tilde{g}'_{M+1}))} \tag{329}$$

Death step

We sample a thinned center to be deleted uniformely at random. For notational simplicity, assume element M is deleted. The acceptance rate for deleting this unit follows analogously to the birth step above.

$$a=\frac{p(M-1\,|\,\_\_\_)}{p(M\,|\,\_\_\_)}\frac{q(M-1\to M)}{q(M\to M-1)} \tag{330}$$

$$= \frac{L(\boldsymbol{\theta}')\lambda^*\sigma(-\tilde{g}'_{M+1}) \times p(\tilde{g}'_M \mid \tilde{c}'_M, \mathbf{c}_{M+K}, \mathbf{g}_{M+K})}{L(\boldsymbol{\theta})} \frac{\mu(\mathcal{C})}{(M)p(\tilde{g}_M \mid \tilde{c}_M, \mathbf{c}_{M+K}, \mathbf{g}_{M+K})} \tag{331}$$

$$= \frac{L(\boldsymbol{\theta}')\lambda^*\mu(\mathcal{C})}{L(\boldsymbol{\theta})M(1+\exp(\tilde{g}_M))} \tag{332}$$

Updating $\{\tilde{c}_m\}_{m=1}^M$ with MH Thinned center parameters and thinned GP function values are proposed jointly and accepted based on MH. The full conditional is:

$$\begin{aligned}
&p(\tilde{c}_i, \tilde{g}_i \mid \mathbf{c}_{M+K-i}, \mathbf{g}_{M+K-i}, ___) \\
&\quad \propto p(\{y_n\}, \{x_n\}, \{w_k\}, b, \{c_k\}, \{\tilde{c}_m\}, \{g_k\}, \{\tilde{g}_m\}, K, M, \lambda^*) \\
&\quad \propto L(\boldsymbol{\theta}) \exp\left\{-\frac{1}{2\sigma_b^2}b^2\right\} (2\pi\sigma_w^2)^{-K/2} \exp\left\{-\frac{1}{2\sigma_w^2}\sum_{k=1}^K w_k^2\right\} \\
&\qquad (\lambda^*)^{\alpha+K+M-1} \exp\left(-\lambda^*(\beta+\mu(\mathcal{C}))\right) \prod_{k=1}^K \sigma(g_k) \prod_{m=1}^M \sigma(-\tilde{g}_m) \times p(\mathbf{g}_{M+K} \mid \mathbf{c}_{M+K}) \\
&\quad \propto L(\boldsymbol{\theta}) \prod_{m=1}^M \sigma(-\tilde{g}_m) \times p(\mathbf{g}_{M+K} \mid \mathbf{c}_{M+K}) \\
&\quad \propto L(\boldsymbol{\theta})\sigma(-\tilde{g}_i) \prod_{m \in \{1,\ldots,M\} \setminus i} \sigma(-\tilde{g}_m) \times p(\tilde{g}_i, \mathbf{g}_{M+K-i} \mid \tilde{c}_i, \mathbf{c}_{M+K}) \\
&\quad \propto L(\boldsymbol{\theta})\sigma(-\tilde{g}_i) \times p(\tilde{g}_i, \mathbf{g}_{M+K-i} \mid \tilde{c}_i, \mathbf{c}_{M+K}) \\
&\quad \propto L(\boldsymbol{\theta})\sigma(-\tilde{g}_i) \times p(\tilde{g}_i \mid \mathbf{g}_{M+K-i}, \tilde{c}_i, \mathbf{c}_{M+K-i}) \times p(\mathbf{g}_{M+K-i} \mid \tilde{g}_i, \mathbf{c}_{M+K-i}) \\
&\quad \propto L(\boldsymbol{\theta})\sigma(-\tilde{g}_i) \times p(\tilde{g}_i \mid \tilde{c}_i, \mathbf{g}_{M+K-i}, \mathbf{c}_{M+K-i}) \times p(\mathbf{g}_{M+K-i} \mid \mathbf{c}_{M+K-i}) \\
&\quad \propto L(\boldsymbol{\theta})\sigma(-\tilde{g}_i) \times p(\tilde{g}_i \mid \tilde{c}_i, \mathbf{g}_{M+K-i}, \mathbf{c}_{M+K-i})
\end{aligned}$$

The proposal probability ratio is then:

$$\frac{p(\tilde{c}'_i, \tilde{g}'_i \mid \mathbf{c}_{M+K-i}, \mathbf{g}_{M+K-i}, \_\_\_)}{p(\tilde{c}_i, \tilde{g}_i \mid \mathbf{c}_{M+K-i}, \mathbf{g}_{M+K-i}, \_\_\_)} = \frac{L(\boldsymbol{\theta}')\sigma(-\tilde{g}_i;) \times p(\tilde{g}'_i \mid \tilde{c}'_i, \mathbf{g}_{M+K-i}, \mathbf{c}_{M+K-i})}{L(\boldsymbol{\theta})\sigma(-\tilde{g}_i) \times p(\tilde{g}_i \mid \tilde{c}_i, \mathbf{g}_{M+K-i}, \mathbf{c}_{M+K-i})} \tag{333}$$

The proposal distribution for thinned unit i consists of two steps:

- Sample $\tilde{c}'_i \sim \mathcal{N}(\tilde{c}'_i; \tilde{c}_i, \sigma_{qc}^2)$
- Sample $\tilde{g}'_i \sim p(\tilde{g}' \mid \tilde{c}_i, \mathbf{c}_{M+K-i}, \mathbf{g}_{M+K-i})$

Therefore, the proposal probability from the current state to the proposed state is:

$$q((\tilde{c}_i, \tilde{g}_i) \to (\tilde{c}'_i, \tilde{g}'_i)) = \mathcal{N}(\tilde{c}'_i; \tilde{c}_i, \sigma_{qc}^2) \times p(\tilde{g}' \mid \tilde{c}'_i, \mathbf{c}_{M+K-i}, \mathbf{g}_{M+K-i}) \tag{334}$$

The proposal probability ratio is then:

$$\frac{q((\tilde{c}'_i, \tilde{g}'_i) \to (\tilde{c}_i, \tilde{g}_i))}{q((\tilde{c}_i, \tilde{g}_i) \to (\tilde{c}'_i, \tilde{g}'_i))} = \frac{\mathcal{N}(\tilde{c}_i; \tilde{c}'_i, \sigma_{qc}^2) \times p(\tilde{g}_i \mid \tilde{c}_i, \mathbf{c}_{M+K-i}, \mathbf{g}_{M+K-i})}{\mathcal{N}(\tilde{c}'_i; \tilde{c}_i, \sigma_{qc}^2) \times p(\tilde{g}'_i \mid \tilde{c}'_i, \mathbf{c}_{M+K-i}, \mathbf{g}_{M+K-i})} = \frac{p(\tilde{g}_i \mid \tilde{c}_i, \mathbf{c}_{M+K-i}, \mathbf{g}_{M+K-i})}{p(\tilde{g}'_i \mid \tilde{c}'_i, \mathbf{c}_{M+K-i}, \mathbf{g}_{M+K-i})} \tag{335}$$

The acceptance ratio is then:

$$a = \frac{p(\tilde{c}'_i, \tilde{g}_i \mid \mathbf{c}_{M+K-i}, \mathbf{g}_{M+K-i}, \_\_\_)}{p(\tilde{c}'_i, \tilde{g}_i \mid \mathbf{c}_{M+K-i}, \mathbf{g}_{M+K-i}, \_\_\_)} \frac{q((\tilde{c}'_i, \tilde{g}'_i) \to (\tilde{c}_i, \tilde{g}_i))}{q((\tilde{c}_i, \tilde{g}_i) \to (\tilde{c}'_i, \tilde{g}'_i))} \tag{336}$$

$$= \frac{L(\boldsymbol{\theta}')\sigma(-\tilde{g}'_i) \times p(\tilde{g}'_i \mid \tilde{c}'_i, \mathbf{g}_{M+K-i}, \mathbf{c}_{M+K-i})}{L(\boldsymbol{\theta})\sigma(-\tilde{g}_i) \times p(\tilde{g}_i \mid \tilde{c}_i, \mathbf{g}_{M+K-i}, \mathbf{c}_{M+K-i})} \frac{p(\tilde{g}_i \mid \tilde{c}_i, \mathbf{c}_{M+K-i}, \mathbf{g}_{M+K-i})}{p(\tilde{g}'_i \mid \tilde{c}'_i, \mathbf{c}_{M+K-i}, \mathbf{g}_{M+K-i})} \tag{337}$$

$$= \frac{L(\boldsymbol{\theta}')}{L(\boldsymbol{\theta})} \frac{1+\exp(\tilde{g}_i)}{1+\exp(\tilde{g}'_i)} \tag{338}$$

Updating $\{g_k\}_{k=1}^K$ and $\{\tilde{g}_m\}_{m=1}^M$ with HMC The full conditional distribution of the GP function values is given by:

$$
\begin{aligned}
p(\{g_k\}, \{\tilde{g}_m\} \,|\, \_\_\_) &\propto p(\{y_n\}, \{x_n\}, \{w_k\}, b, \{c_k\}, \{\tilde{c}_m\}, \{g_k\}, \{\tilde{g}_m\}, K, M, \lambda^*) \\
&\propto L(\boldsymbol{\theta}) \cancel{\exp\left\{-\frac{1}{2\sigma_b^2} b^2\right\}} \cancel{(2\pi\sigma_w^2)^{-K/2}} \cancel{\exp\left\{-\frac{1}{2\sigma_w^2} \sum_{k=1}^K w_k^2\right\}} \\
&\quad \cancel{(\lambda^*)^{\alpha+K+M-1}} \cancel{\exp\left(-\lambda^*(\beta + \mu(\mathcal{C}))\right)} \prod_{k=1}^K \sigma(g_k) \prod_{m=1}^M \sigma(-\tilde{g}_m) \times \mathcal{GP}(\{g_k\}, \{\tilde{g}_m\} \,|\, \{c_k\}, \{\tilde{c}_m\}) \\
&\propto L(\boldsymbol{\theta}) \prod_{k=1}^K \sigma(g_k) \prod_{m=1}^M \sigma(-\tilde{g}_m) \cancel{(2\pi)^{-(M+K)/2}} \cancel{|\Sigma|^{-1/2}} \exp\left\{-\frac{1}{2} \mathbf{g}_{M+K}^T \Sigma^{-1} \mathbf{g}_{M+K}\right\} \\
&\propto L(\boldsymbol{\theta}) \prod_{k=1}^K \sigma(g_k) \prod_{m=1}^M \sigma(-\tilde{g}_m) \exp\left\{-\frac{1}{2} \mathbf{g}_{M+K}^T \Sigma^{-1} \mathbf{g}_{M+K}\right\}
\end{aligned}
$$

To update the GP function values, we use HMC with $-\log p(\{g_k\}, \{\tilde{g}_m\} \,|\, \_\_\_)$ as the potential energy function.

D ADDITIONAL EXPERIMENTS AND DETAILS OF EXPERIMENTAL SETUP

D.1 EXPERIMENTS ILLUSTRATING PROPERTIES OF PORB-NETS

D.1.1 PoRB-Nets decouple amplitude variance and lengthscale

Here we examine the dependence of the variance and lengthscale (as measured by the upcrossings of $y = 0$) for three different models: a standard BNN (in this case a single layer neural network with a Gaussian activation) (Figure 2), an RBFN with a homogeneous Poisson process prior on the number of hidden units *but without scaling the hidden units by the intensity*, as in PoRB-Net (Figure 3), and PoRB-Net with a homogeneous intensity (Figure 4). We compute the average variance and upcrossings over 2000 function samples drawn from each over the interval $x \in [-0.5, 0.5]$. The goal is to examine how each of these two properties scales with the model parameters.

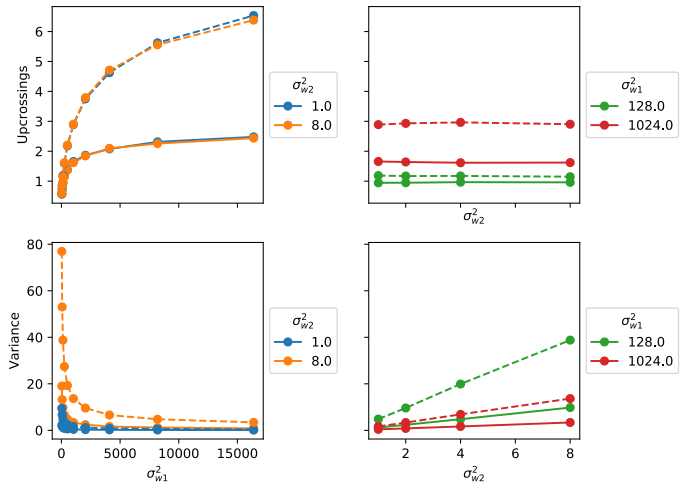


Figure 2: Average variance and upcrossings of a BNN. Dotted line: 20 hidden units, solid line: 100 hidden units. Generally the input-to-hidden weights variance $\sigma^2_{w_1}$ controls the upcrossings while the hidden-to-output weights variance $\sigma^2_{w_1}$ controls the variance, but notice how $\sigma^2_{w_1}$ impacts both properties (and in a nonlinear way). This shows that these properties cannot be controlled independently .

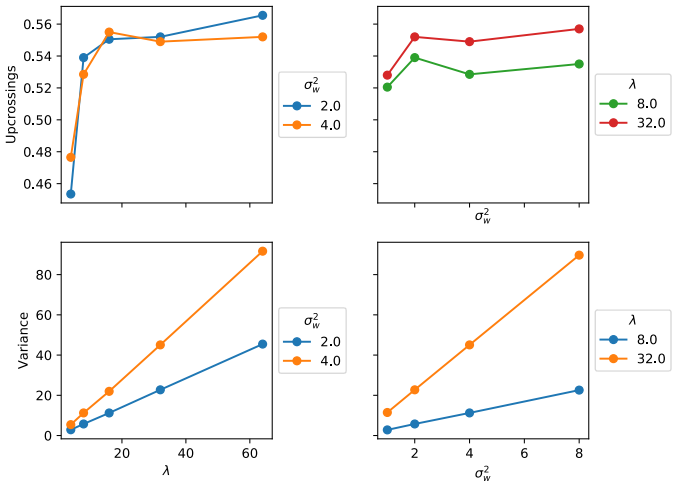


Figure 3: Average variance and upcrossings of an RBFN with homogeneous Poisson process prior on the centers but without scaling the hidden units by the intensity. Notice the intensity λ impacts both the upcrossings and the variance, since a higher intensity implies more radial basis functions, which continue to add up if their width is not scaled. The panels on the right show the hidden-to-output weights variance.

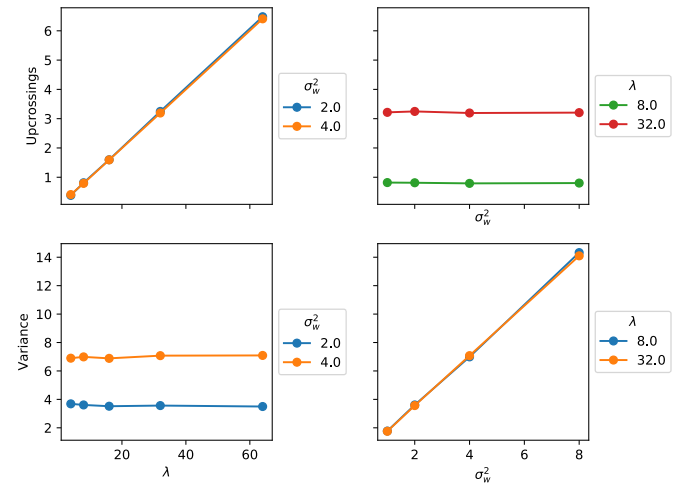
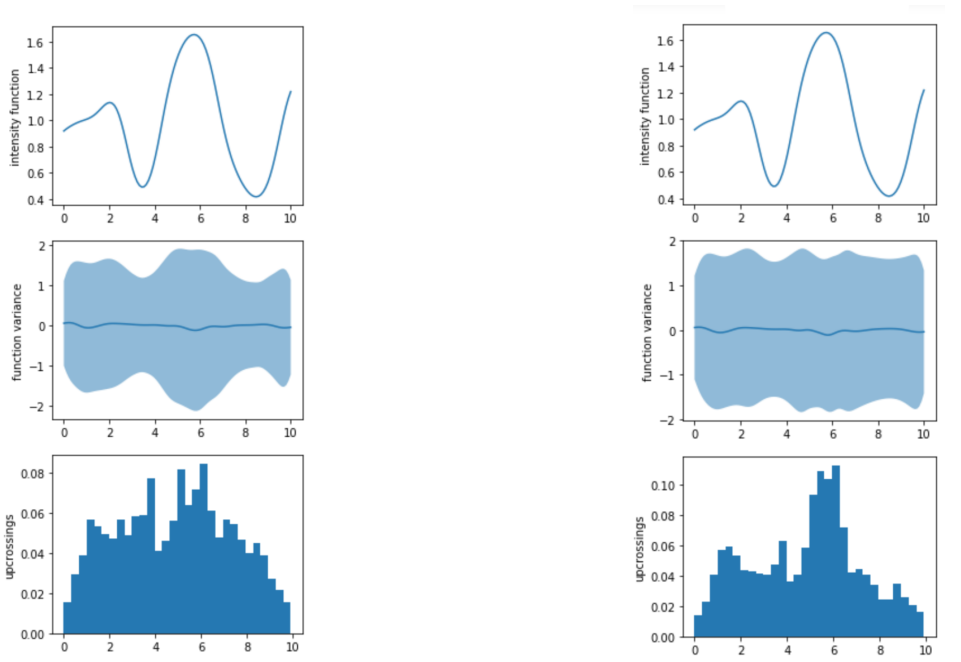


Figure 4: Average variance and upcrossings of PoRB-Net. The intensity λ and hidden-to-output weights variance σ_w^2 independently control the upcrossings and variance .

Now consider an inhomogenous Poisson process prior on the center parameters with an arbitrary intensity function $\lambda(c)$. Recall in PoRB-Net we set the scale parameters of each unit to $s_k^2 = s_0^2\lambda(c_k)$. Figure 5 shows different function samples for a single fixed intensity sampled from the prior. On the left, the s_k^2 is constant for each unit while on the right $s_k^2 = s_0^2\lambda(c_k)$. The top row shows the true intensity, the middle row shows the amplitude variance, and the bottom row show a histogram of the number of function upcrossings of $y = 0$. We see that setting the scales based on the intensity results in approximately constant function variance and increased upcrossings.



(a) Constant s_k^2

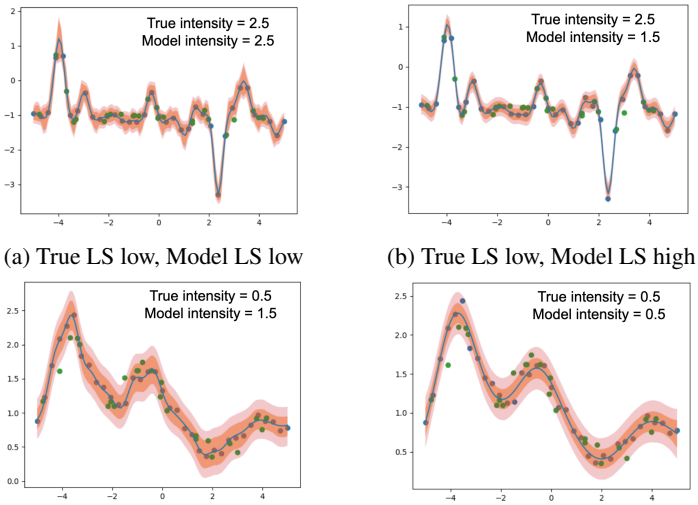
(b) $s_k^2 = s_0^2 \lambda(c_k)^2$, where for $s_0^2 = 2\pi$.

Figure 5: Setting $s_k^2 = s_0^2 \lambda(c_k)^2$ results in approximately stationary amplitude variance

.

D.1.2 PoRB-Nets can use prior information to adjust uncertainty in gaps in the training data

If prior information is available about a function's lengthscale, this can be incorporated into the PoRB-Net prior by adjusting a fixed Poisson process intensity. Figure 6 shows an example with input data x sampled uniformly while Figure 7 shows an example where there is a large gap in the x data.



(a) True LS low, Model LS low

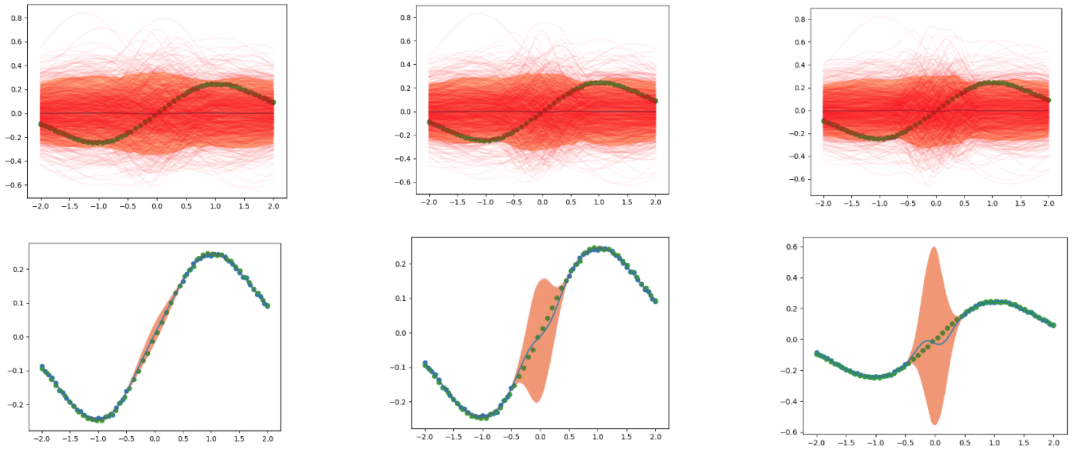
(b) True LS low, Model LS high

(c) True LS high, Model LS low

(d) True LS high, Model LS high

Figure 6: Left-to-right: increased lengthscale (LS) for the PoRB-Net model. Top-to-bottom: increased lengthscale for the true function (drawn from a PoRB-Net prior)

.

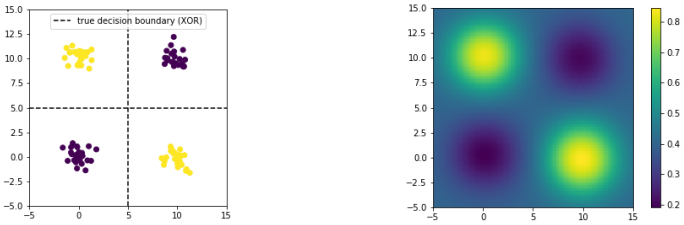


2x intensity in gap | 3x intensity in gap | 4x intensity in gap

Figure 7: By adjusting the Poisson process intensity a gap in the data (note that green points are test observations), the out of sample uncertainty can be adjusted. Higher intensity results in a smaller length scale.

D.1.3 PoRB-Nets can be used for classification

We focus on one-dimensional regression examples in this paper, since our primary interest is theoretical. However, PoRB-Nets can easily be extended to higher dimensional inputs and outputs. Figure 8 the posterior predictive distribution on a simple two dimensional XOR classification dataset.



(a) Data (b) PoRB-Net posterior predictive.

Figure 8: **PoRB-Nets can do classification.** We fit a PoRB-Net to a simple, two-dimensional XOR classification dataset.

D.2 COMPARISON WITH OTHER MODELS

D.2.1 Details on experimental setup

Data

- **sin, inc, inc2, const.** We create four synthetic datasets of 100 observations each by adding i.i.d. $\mathcal{N}(0, 0.02)$ noise to functions from Gaussian process priors with the following nonstationary kernel [Gibbs, 1997]:

$$\Sigma(x, x') = \sigma(x)\sigma(x')\sqrt{\frac{2l(x)l(x')}{l(x)^2 + l(x')^2}} \exp\left(-\frac{(x - x')^2}{l(x)^2 + l(x')^2}\right) \tag{339}$$

  Each dataset corresponds to a different lengthscale function, $l(x)$. $l_{\text{const}}(x) = 1$ is a constant function, $l_{\sin}(x) = \sin(x) + 1.1$ is a sine function shifted above zero, and $l_{\text{inc}}(x)$ is a function that increases from left to right (see plots below). Note that "inc" and "inc2" have the same lengthscale, the former just has gaps in the x data while the latter has x data sampled uniformly.

- **mimic.** Each of the four datasets from the Mimic Critical Care Database Johnson et al. [2016] shows patient heart rate over time.
- **motorcycle.** The motorcycle accident dataset [Silverman, 1985] tracks the acceleration force on the head of a motorcycle rider in the first moments after impact.
- **finance.** Chicago Board Options Exchange (CBOE) volatility index (VIX), downloaded from https://fred.stlouisfed.org/series/VIXCLS. For testing data, we create two large, artificial gaps in the data. The remaining observations are downsampled by 25%, leaving 25 training observations

  144 train 60 test

Matching priors To highlight differences between the models, we attempt to approximately match the priors in amplitude variance and lengthscale. We choose these these properties since they are the focus of this paper. For each dataset, we start by selecting the variance parameter σ^2 and lengthscale parameter l of a GP with an RBF kernel by optimizing the log marginal likelihood of the data, where an RBF kernel is given by:

$$\Sigma_{\text{rbf}}(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right). \tag{340}$$

Since some of the datasets contain gaps in x space, we constrain the lengthscale to be larger than $l_{\min} = 1/(2\pi \cdot 5) \approx 0.032$, which implies the expected number of upcrossings u of $y = 0$ over $x \in [0, 1]$ is 5, since $u = 1/(2\pi l)$ [Williams and Rasmussen, 2006]. It is also difficult to model very small lengthscales with networks of small capacity, which we needed to limit because we wished to perform full HMC inference. The observational noise variance is assumed fixed and and set to a reasonable value for each dataset (or the ground truth, if available). Once these parameters were selected for each dataset, we matched the BNN and PoRB-Net to this prior by a searching over a 25×25 grid of two model parameters. To measure lengthscale and amplitude variance of each model, we used the average upcrossings and average variance over $x \in [0, 1]$.

For the BNN (single layer), we controlled the overall lengthscale by adjusting the input-to-hidden weights variance from 10 to 15,000 and we controlled the the overall amplitude variance by adjusting the hidden-to-output weights variance from .01 to 1.0. We included the variance of the bias parameters in this search. Note that both the upcrossings and the amplitude variance are concentrated near the origin for a BNN.

For PoRB-Net, we controlled the lengthscale by adjusting the intensity λ (in the case of a homogeneous Poisson process) from 5 to 40[2] and the intensity upper bound λ^* (in the case of an inhomogeneous Poisson process) from 2×5 to 2×40 (we multiply by 2 because the mean intensity under a sigmoidal Gaussian Cox process is $1/2\lambda^*$ because the GP, assumed have zero mean, is squashed through the sigmoid function); we controlled the amplitude variance by adjusting the hidden-to-output weights variance from .01 to 1.0. Note that both the upcrossings and amplitude variance are approximately constant over $x \in [0, 1]$.

Model implementations

- **PoRB-Net.** The code is available on our GitHub: https://github.com/dtak/porbnet. For a homogeneous intensity λ (PoRB-Net† in the figures and tables), we assumed a Gamma$(\alpha_\lambda, \beta_\lambda)$ prior on the intensity. For an inhomogenous intensity we assumed used a sigmoidal Gaussian Cox process defined by intensity $\lambda(c) = \lambda^* \sigma(g(c))$, where g is a GP with an RBF kernel Σ_{rbf} of lengthscale .25 and variance 5. We set the variance of the GP to be fairly large since we did not place a prior on λ^*. We set $s_0^2 = 2$. For inference we use HMC with 5000 burn in samples and 5000 recorded samples. During the first half of the burn in, we find it is advantageous to not do any birth or death steps of (unthinned) hidden units. During all of burn in, dynamically adjust the HMC leapfrog step size ϵ to target an acceptance rate of 65%. Since the Poisson process is defined over a region $\mathcal{C}$, we implement Roll-back HMC [Yi and Doshi-Velez, 2017], which introduces a sharp sigmoid factor in the potential energy to approximate the probability drop at the boundaries of $\mathcal{C}$.
- **GP.** We use the `GPy` package, available at https://sheffieldml.github.io/GPy/. We use an RBF kernel.

[2] Technically we adjusted α_λ and set $\beta_\lambda = 1$ so that $\mathbb{E}[\lambda]$ ranged from 5 to 40.

- **LGP.** We use the Matlab code made publicly available by the authors [Heinonen et al., 2016] at https://github.com/markusheinonen/adaptivegp. We modified the code slightly to ensure that the observational noise variance was fixed that only the input dependence of the lengthscale was inferred (and not the input dependence of the amplitude variance or observational noise variance, for which this model allows). For inference we used HMC with 1000 samples, since it seems to converge faster than a BNN or PoRB-Net. We set the lengthscale parameter β_l of the GP prior on the log lengthscale function to be .25, the same as value as used in the GP defining the sigmoidal Gaussian Cox process used by PoRB-Net.
- **BNN.** We use our own implementation, also available on our GitHub. We use a Gaussian activation function $\phi(z) = \exp(-\frac{1}{2}s_0^2 z^2)$ for all experiments for better comparison to PoRB-Net (we add the scale parameter $s_0^2 = 2$ so the activation function is the same as in PoRB-Net). For inference we the same HMC implementation with the same number of samples (5000 burn in and 5000 for recorded) as for PoRB-Net.

D.2.2 Results

Numerical results Tables 4 and 5 show the test log likelihoods and root mean squared errors (RMSEs) for all datasets and all models. We evaluate the test log likelihood of the neural networks as:

$$\mathbb{E}_{p(x^\star,y^\star)}\left[\log p(y^\star|x^\star,\mathcal{D})\right] = \mathbb{E}_{p(x^\star,y^\star)}\left[\log \int p(y^\star|x^\star,\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}\right], \tag{341}$$

where $\boldsymbol{\theta}$ are the model parameters.

Table 4: **Test Log Likelihoods.** The number next to the BNN indicates the number of hidden units.

	PoRB-Net†	PoRB-Net	GP	LGP	BNN (25)	BNN (50)	BNN (100)
sin*	0.767	**0.817**	0.728	**0.814**	0.755	0.742	0.789
inc*	-0.401	0.001	-0.227	**0.183**	**-0.153**	-0.284	-0.160
inc-gap*	0.656	**0.747**	0.543	0.183	0.631	0.627	0.678
const-gap*	0.277	0.330	**0.413**	0.239	0.009	-0.295	-0.133
mimic1	0.887	0.947	0.827	0.897	**1.047**	0.952	0.912
mimic2	0.534	**0.603**	0.564	0.540	0.441	0.389	0.472
mimic3	-0.634	**-0.571**	-0.671	-0.583	-0.648	-0.626	-0.594
mimic4	-1.719	-1.531	-1.848	-1.439	**-1.383**	**-1.084**	-1.362
finance	-1.410	-0.521	-1.975	**0.033**	-2.629	-0.754	-0.734
motorcycle	**0.184**	0.155	0.167	0.141	0.159	0.125	0.127

**synthetic dataset* †*infers homogeneous intensity*

Table 5: **Test RMSEs.** The number next to the BNN indicates the number of hidden units.

	PoRB-Net†	**PoRB-Net**	**GP**	**LGP**	**BNN (25)**	**BNN (50)**	**BNN (100)**
sin*	0.075	0.068	0.089	**0.067**	0.077	0.083	0.073
inc*	0.394	0.356	0.348	0.346	**0.345**	0.399	0.349
inc-gap*	0.114	**0.099**	0.137	0.346	0.122	0.120	0.111
const-gap*	0.174	0.168	**0.159**	0.260	0.304	0.326	0.327
mimic1	0.084	0.078	0.081	0.086	**0.064**	0.071	0.078
mimic2	0.187	**0.163**	0.171	0.165	0.240	0.246	0.220
mimic3	0.204	**0.201**	0.205	0.203	0.204	0.203	0.202
mimic4	0.280	0.275	0.281	0.278	**0.262**	0.265	0.269
finance	0.586	0.403	0.628	**0.254**	0.429	0.333	0.399
motorcycle	**0.200**	0.206	0.201	0.205	0.212	0.217	0.212

**synthetic dataset* †*infers homogeneous intensity*

Posterior predictive plots

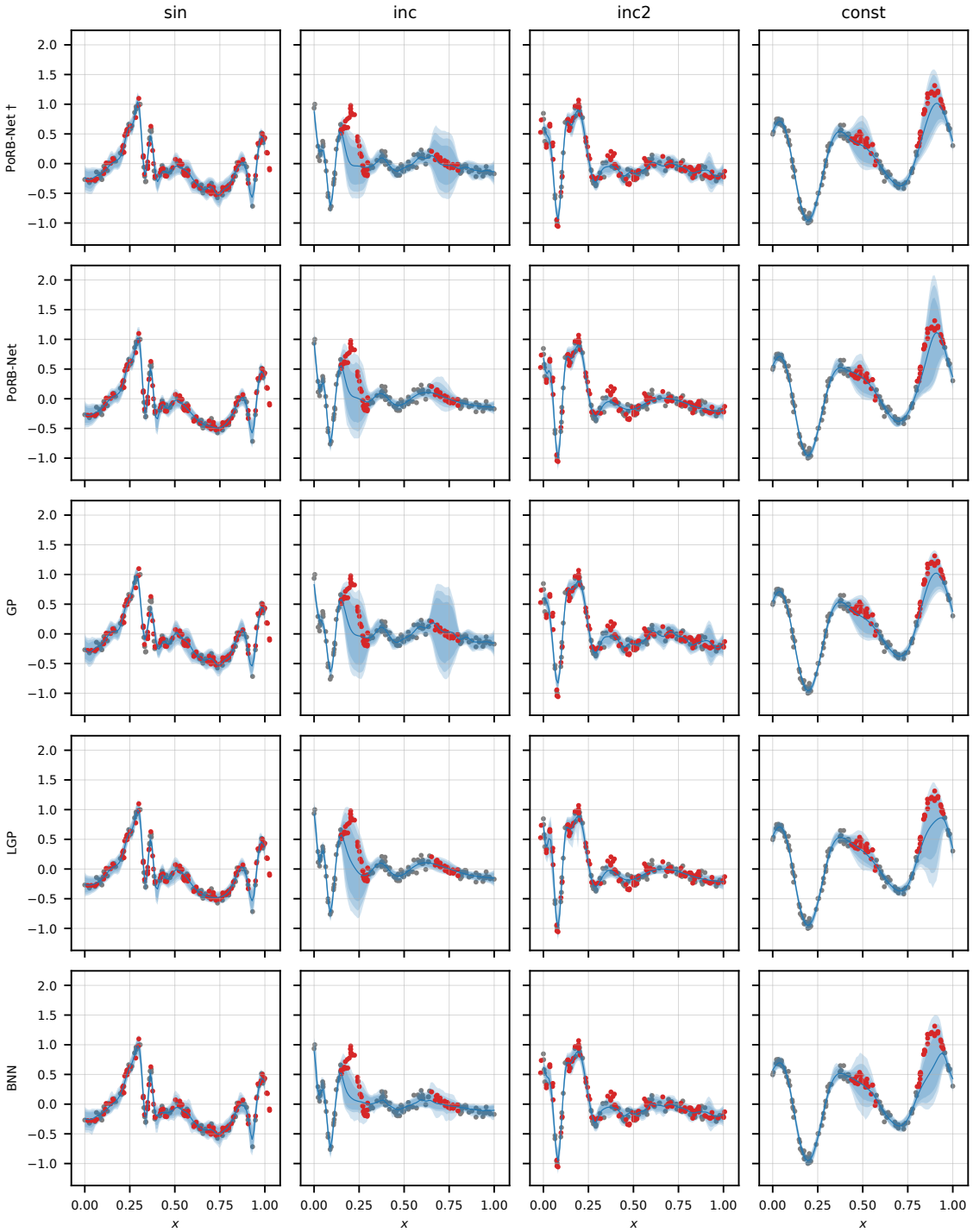


Figure 9: Each dataset is drawn from a GP with the nonstationary kernel given in Equation 339 with a different input dependent lengthscale function $l(x)$ (see Figure 12). We see qualitative similarity (especially in the gaps in the data along the x-axis) between PoRB-Net† and the GP (both stationary) and PoRB-Net and the LGP (both nonstationary). The BNN looks different from the other models (e.g., it has small uncertainty in *both* gaps in the sin dataset). Training points are gray; test points are red.

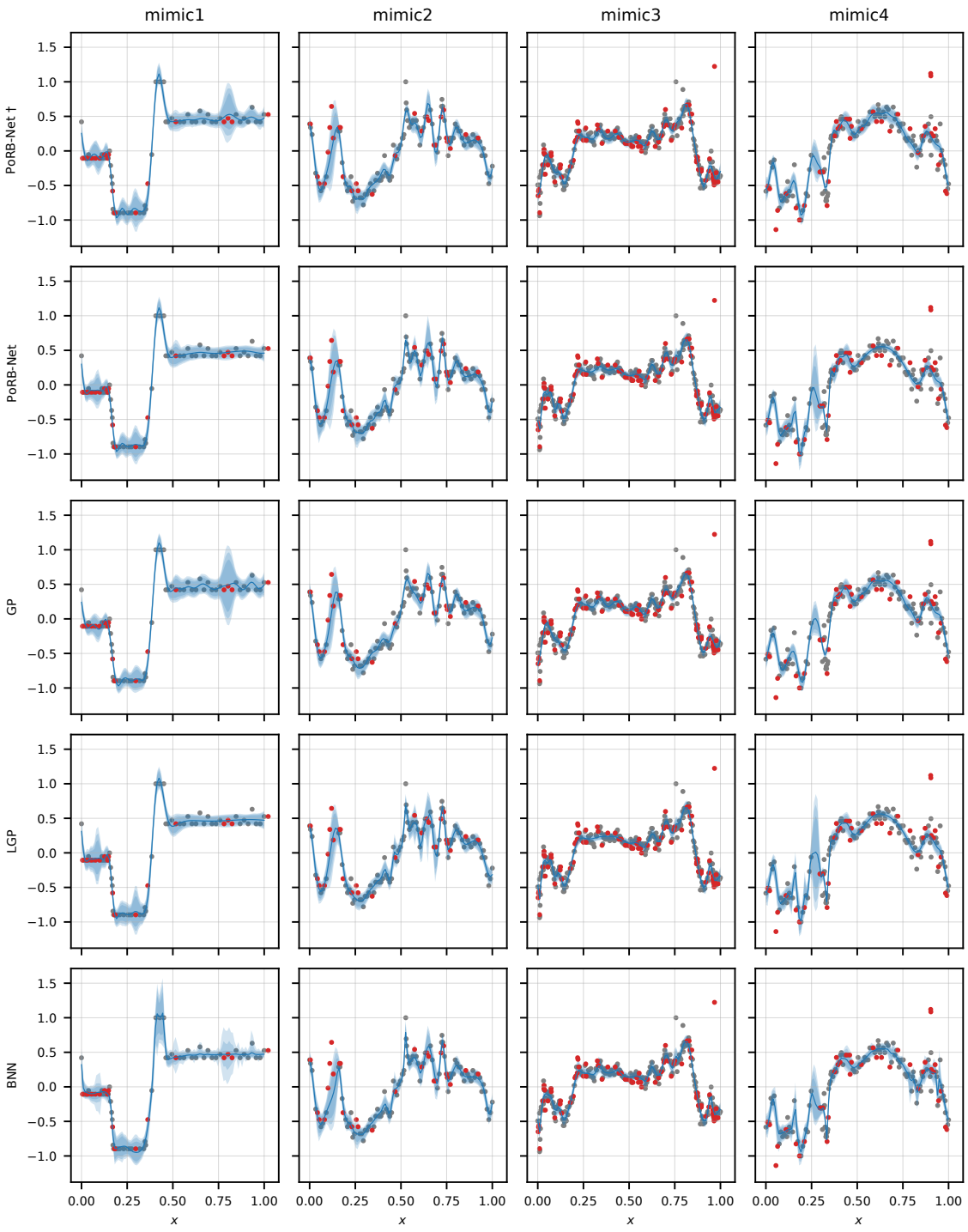


Figure 10: Posterior predictive distributions. The mimic1 dataset shows the largest qualitiative differences between models, with PoRB-Net and LGP learning a smooth function for $x > 0.5$. Meanwhile, PoRB-Net†, which has a homogeneous intensity, more closely resembles a stationary GP than a BNN, which places relatively high uncertainty near the origin and relatively less away from it. Training points are gray; test points are red.

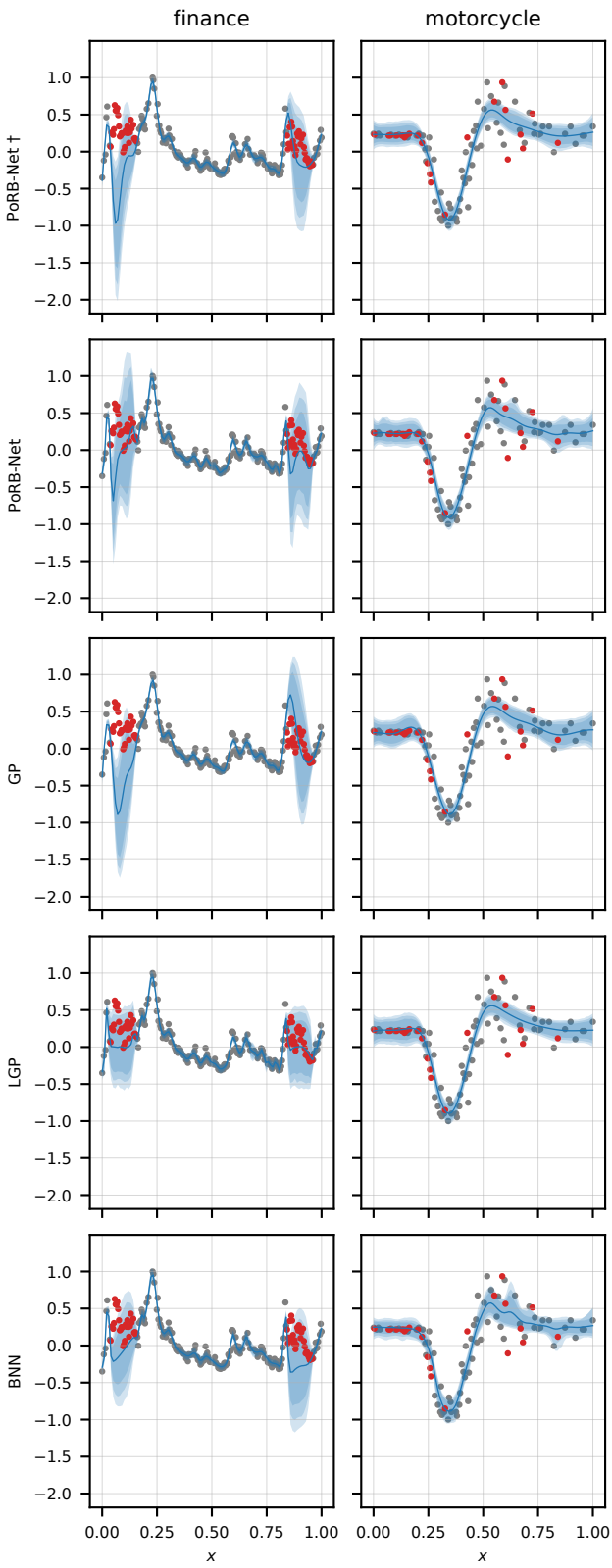


Figure 11: Posterior predictive distributions. The nonstationary models exhibit better generalization on the finance dataset. All models look fairly similar on the motorcycle dataset. Training points are gray; test points are red.

Intensity and lengthscale plots Here we compare the inhomogeneous Poisson process intensities inferred by PoRB-Net and the inverse of the input dependent lengthscales inferred by the LGP. In both cases, higher values indicate less smooth functions.

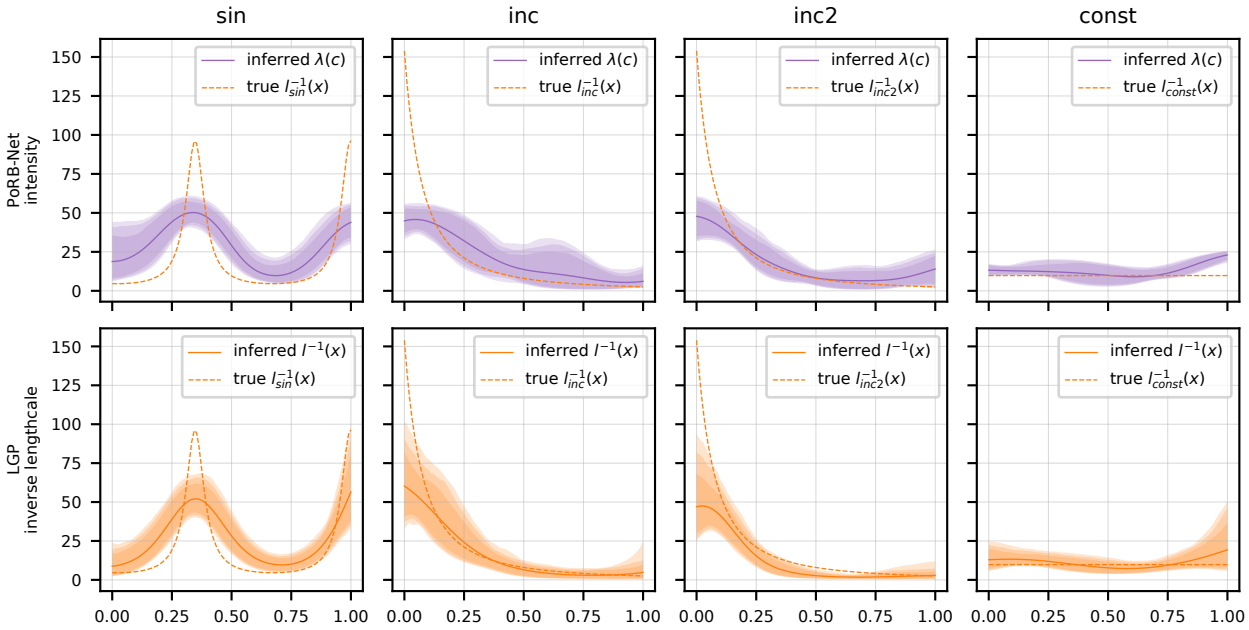


Figure 12: Each dataset is drawn from a GP with the nonstationary kernel given in Equation 339, with the inverse of the ground truth input dependent lengthscale function $l(x)$ shown in each plot. Both models pick up on the patterns in the ground truth data. Note that the LGP uses the same kernel as in the ground truth (but it infers $l(x)$ with another GP as a prior).

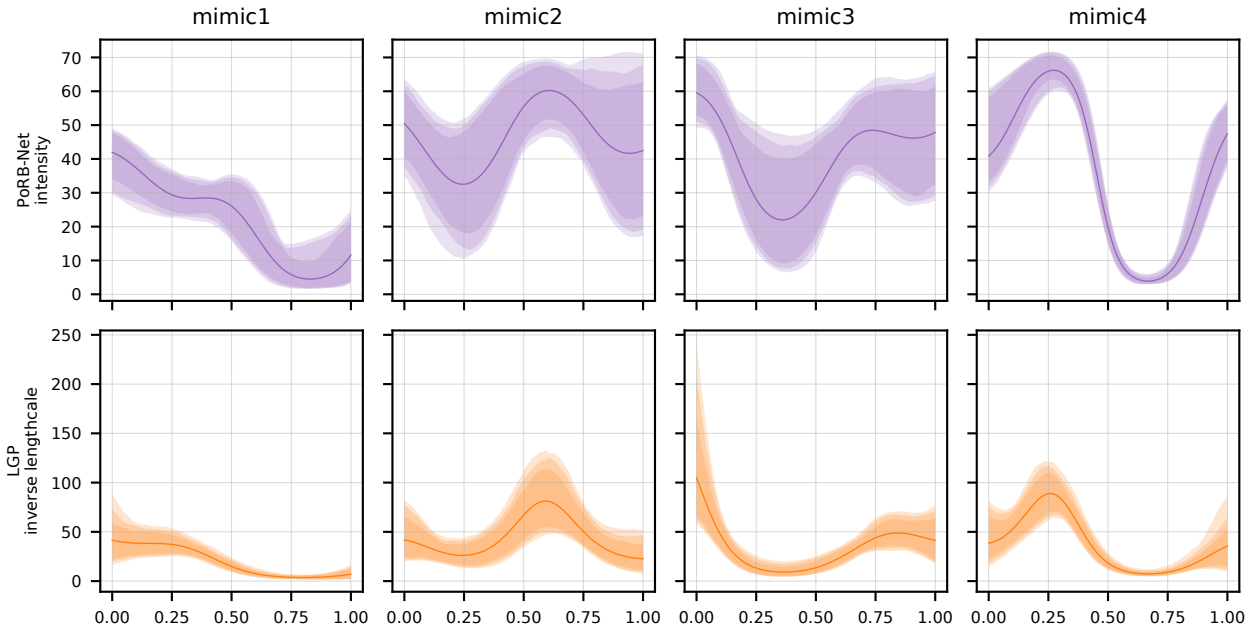


Figure 13: PoRB-Net and LGP pick up on similar lengthscale patterns in the mimic data.

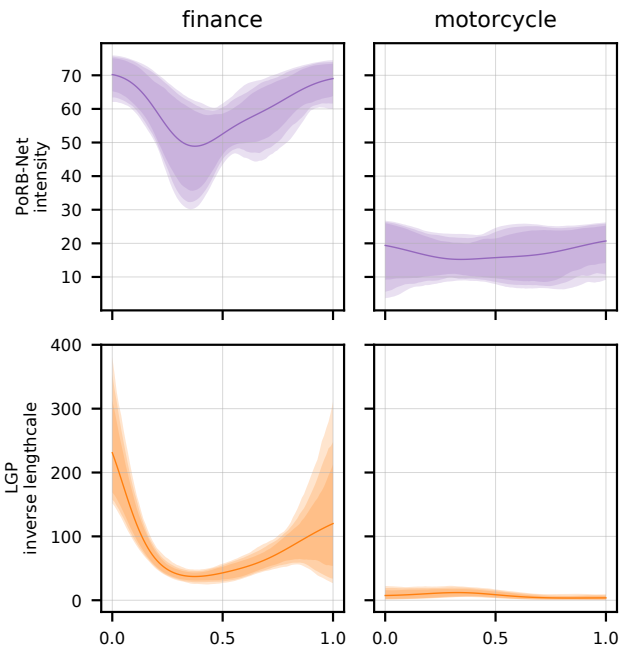


Figure 14: For the finance dataset, both models infer a smaller lengthscale near the beginning and end of the time period, where the VIX was clearly more volatile. This resulted in better uncertainty in the gaps in the data as compared to stationary models (GP and PoRB-Net with a homogeneous intensity). For the motorcycle dataset, both models infer a fairly homogeneous lengthscale, which makes sense because the motorcycle dataset is typically considered an example of input dependent amplitude variance rather than input dependent lengthscale [Heinonen et al., 2016].

References

Gibbs, M. N. (1997). *Bayesian Gaussian processes for regression and classification*. PhD Thesis, University of Cambridge.

Heinonen, M., Mannerström, H., Rousu, J., Kaski, S., and Lähdesmäki, H. (2016). Non-stationary gaussian process regression with hamiltonian monte carlo. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Lee, H. K. (2000). Consistency of posterior distributions for neural networks. *Neural Networks: The Official Journal of the International Neural Network Society*, 13(6):629–642.

Park, J. and Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural Computation*, 3(2):246–257.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society*, 47:1–52.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer. Springer New York, New York, NY.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve mles. (2).

Yi, K. and Doshi-Velez, F. (2017). Roll-back hamiltonian monte carlo. *arXiv:1709.02855 [stat.ML]*.