
NeuralSurv: Deep Survival Analysis with Bayesian Uncertainty Quantification

Mélodie Monod*
 Imperial College London
 London, United Kingdom
 melodie.monod18@imperial.ac.uk

Alessandro Micheli*
 Imperial College London
 London, United Kingdom
 a.micheli19@imperial.ac.uk

Samir Bhatt
 Imperial College London; University of Copenhagen
 London, United Kingdom; Copenhagen, Denmark
 s.bhatt@imperial.ac.uk

Abstract

We introduce *NeuralSurv*, the first deep survival model to incorporate Bayesian uncertainty quantification. Our non-parametric, architecture-agnostic framework flexibly captures time-varying covariate–risk relationships in continuous time via a novel two-stage data-augmentation scheme, for which we establish theoretical guarantees. For efficient posterior inference, we introduce a mean-field variational algorithm with coordinate-ascent updates that scale linearly in model size. By locally linearizing the Bayesian neural network, we obtain full conjugacy and derive all coordinate updates in closed form. In experiments, *NeuralSurv* delivers superior calibration compared to state-of-the-art deep survival models, while matching or exceeding their discriminative performance across both synthetic benchmarks and real-world datasets. Our results demonstrate the value of Bayesian principles in data-scarce regimes by enhancing model calibration and providing robust, well-calibrated uncertainty estimates for the survival function.

1 Introduction

Survival analysis is a branch of statistics focused on the study of time-to-event data, usually called event times. This type of data appears in a wide range of applications such as medicine [29], engineering [31], and social sciences [37]. A key objective of survival analysis is to estimate the survival function and the hazard function that govern the distribution of event times.

Traditional survival models like the Cox proportional hazards model [7] and accelerated failure time models [6] have long delivered reliable inference under strong parametric assumptions. However, such assumptions may fail to adequately capture complex and evolving baseline hazards, especially when risk relationships vary over time. To overcome these limitations, recent work has begun incorporating modern machine-learning techniques [41], and in particular deep architectures [42, 20, 28], which can learn rich, hierarchical representations directly from data. Yet most deep-survival approaches remain purely frequentist, optimizing point-estimate losses and offering no coherent uncertainty quantification. In high-stakes settings like medicine, this lack of reliable uncertainty estimates can undermine trust and impede adoption.

Bayesian statistics, by contrast, inherently quantifies uncertainty: prior beliefs are combined with observed data to yield a posterior distribution over model parameters [12]. In survival analysis, Bayesian

*Equal contribution.

methods can produce full posterior distributions for individual survival functions summarizable via credible intervals that communicate model confidence [17]. Traditional Bayesian survival tools, such as Gaussian processes (GPs) [10, 21], offer nonparametric flexibility and built-in uncertainty but often falter in high-dimensional settings due to scalability issues. To date, no method has married the representational power of deep learning with full Bayesian uncertainty quantification in a scalable survival framework. Such a synthesis would hold the potential to learn complex, high-dimensional survival dynamics while retaining principled probabilistic interpretations.

In this work, we address this critical need by introducing *NeuralSurv*, an architecture-agnostic, Bayesian deep-learning framework for survival analysis. *NeuralSurv* leverages deep neural networks to learn hierarchical representations from covariates and uses a principled variational inference framework to provide rigorous uncertainty quantification over the survival function. We develop a novel two-stage data-augmentation strategy that leverages latent marked Poisson processes and Pólya–Gamma variables. Our approach comes with theoretical guarantees and enables exact continuous-time likelihood computation. By locally linearizing the Bayesian neural network, we achieve conjugacy and derive closed-form coordinate-ascent updates that scale linearly with network size.

Through extensive experiments on synthetic and real survival datasets, in data-scarce settings, *NeuralSurv* consistently delivers superior calibration compared to state-of-the-art deep survival models, and matches or exceeds their discriminative performance. Its Bayesian formulation captures epistemic uncertainty to prevent overfitting, while informative priors induce a soft regularization that yields smooth, plausible survival functions.

2 NeuralSurv

We briefly introduce our survival analysis setup; for a comprehensive review, see Appendix A. Let (T_i, C_i) denote the survival and censoring times respectively for observations $i = 1, \dots, N$. We observe $\mathcal{D} = \{(y_i, \delta_i) : i = 1, \dots, N\}$, where $y_i = \min(T_i, C_i)$ is the observed (right-censored) event time and $\delta_i = \mathbb{1}_{\{T_i \leq C_i\}}$ is the censoring indicator. Each observation has a covariate vector $\mathbf{x}_i \in \mathbb{R}^p$, collected into $\mathbf{X} = \{\mathbf{x}_i : i = 1, \dots, N\}$. Throughout this paper, we assume that the censoring time C_i is independent of the event time T_i given \mathbf{x}_i .

Our goal is to model the hazard function λ , i.e. the instantaneous event rate at time t conditional on survival to t and covariates \mathbf{x} . We assume a sigmoidal hazard function:

$$\lambda(t \mid \mathbf{x}; \phi, g(\cdot; \boldsymbol{\theta})) = \lambda_0(t, \mathbf{x}; \phi) \sigma(g(t, \mathbf{x}; \boldsymbol{\theta})), \quad (1)$$

where the sigmoid function $\sigma(z) = 1/(1 + \exp(-z))$ maps any real input to a value in $(0, 1)$. Here, the baseline hazard $\lambda_0 : \mathbb{R}_+ \times \mathbb{R}^p \rightarrow \mathbb{R}_+$, parametrized by $\phi \in \mathbb{R}_+$, encodes our prior “best-guess” hazard profile over time and covariates. Finally, the flexible function $g : \mathbb{R}_+ \times \mathbb{R}^p \rightarrow \mathbb{R}$, parametrized by $\boldsymbol{\theta} \in \mathbb{R}^m$, provides a data-driven adjustment: once passed through the sigmoid, it multiplicatively modulates the baseline hazard, continuously scaling it between zero and λ_0 .

2.1 Likelihood Distribution

Given the hazard function in (1), the likelihood density for the observation corresponding to the i^{th} observation is given by:

$$p(y_i, \delta_i \mid \mathbf{x}_i, \phi, g(\cdot; \boldsymbol{\theta})) = \left(\lambda_0(y_i, \mathbf{x}_i; \phi) \sigma(g(y_i, \mathbf{x}_i; \boldsymbol{\theta})) \right)^{\delta_i} \exp \left(- \int_0^{y_i} \lambda_0(t, \mathbf{x}_i; \phi) \sigma(g(t, \mathbf{x}_i; \boldsymbol{\theta})) dt \right). \quad (2)$$

Assuming (y_i, δ_i) are i.i.d. conditional on $(\mathbf{x}_i, \phi, g(\cdot; \boldsymbol{\theta}))$, the full-sample likelihood is simply the product

$$p(\mathcal{D} \mid \mathbf{X}, \phi, g(\cdot; \boldsymbol{\theta})) = \prod_{i=1}^N p(y_i, \delta_i \mid \mathbf{x}_i, \phi, g(\cdot; \boldsymbol{\theta})). \quad (3)$$

2.2 Prior Distributions

Prior Distribution on θ . We assume that $g(\cdot; \theta)$ is a Bayesian Neural Network (BNN) parameterized by θ . Furthermore, denote by \mathbf{I}_m the $m \times m$ identity matrix. We place the following isotropic Gaussian prior with zero mean and identity covariance over the network weights

$$p_{\theta}(\theta) = \mathcal{N}(\theta; \mathbf{0}, \mathbf{I}_m). \quad (4)$$

This common choice [4] assumes weights are independently distributed and centered around zero, acting as an uninformative yet regularizing prior that discourages large weights and helps prevent overfitting via shrinkage.

Prior Distribution on ϕ . We further assume that the baseline hazard can be written as

$$\lambda_0(t, \mathbf{x}; \phi) = \frac{\lambda_0(t; \phi)}{Z(t, \mathbf{x})}, \quad (5)$$

for the normalization function

$$Z(t, \mathbf{x}) := \mathbb{E}_{\theta \sim p_{\theta}} [\sigma(g(t, \mathbf{x}; \theta))];$$

see Appendix D for further details on how to obtain $Z(t, \mathbf{x})$. By introducing the normalizing function $Z(t, \mathbf{x})$ in (5) we ensure that the prior mean of the sigmoidal hazard in (1) coincides with the prior mean of the baseline hazard in (5), i.e.

$$\mathbb{E}_{\phi \sim p_{\phi}, \theta \sim p_{\theta}} [\lambda(t | \mathbf{x}; \phi, g(\cdot; \theta))] = \mathbb{E}_{\phi \sim p_{\phi}} [\lambda_0(t; \phi)].$$

This approach, similar to the technique used in [10], centers the distribution around the baseline hazard $\lambda_0(t; \phi)$, favouring hazard trajectories that remain close to this prior “best-guess” profile while still permitting data-driven deviations. Notice that if $g(\cdot)$ has a fully connected architecture, then $Z(t, \mathbf{x}) \equiv \frac{1}{2}$ for all (t, \mathbf{x}) , resulting in the same normalization function value as in [10].

We adopt a Weibull-type baseline hazard

$$\lambda_0(t; \phi) = \phi t^{\rho-1}, \quad p_{\phi}(\phi) = \text{Gamma}(\alpha_0, \beta_0), \quad \rho > 0 \text{ fixed}, \quad (6)$$

where α_0 is the shape and β_0 is the rate of the Gamma distribution. The Weibull-type baseline hazard (6) is the hazard of a Weibull distribution, a common choice in survival analysis [10]. When $\rho = 1$, $\lambda_0(t; \phi)$ becomes constant and the baseline hazard reduces to the hazard of the Exponential distribution.

2.3 Posterior Distribution

Let $p(\phi, \theta | \mathcal{D}, \mathbf{X})$ denote the posterior density over the parameters ϕ and θ , defined with respect to the product measure $d\phi \times d\theta$. By Bayes’ rule, this posterior is proportional (up to normalization) to

$$p(\phi, \theta | \mathcal{D}, \mathbf{X}) \propto p(\mathcal{D} | \mathbf{X}, \phi, g(\cdot; \theta)) p_{\phi}(\phi) p_{\theta}(\theta). \quad (7)$$

The posterior in (7) is generally intractable to compute for three reasons. First, its normalizing constant is unavailable in closed form. Second, the likelihood from (2) requires evaluating N integrals, none of which admits an analytic solution. Finally, the sigmoid in (1) introduces an extra nonlinearity, rendering inference even more analytically challenging.

3 Data Augmentation Strategy

In this section, we present a data augmentation scheme that leverages the properties of Poisson processes and Pólya-Gamma random variables. Specifically, Poisson processes help overcome the challenges associated with computing the integral of the continuous-time function to evaluate the likelihood, while the Pólya-Gamma random variables allow for exact handling of the sigmoid nonlinearity without relying on analytic approximations. This combined approach allows us to efficiently perform posterior inference from the model without resorting to discretization. In Section 4.3, we develop a novel variational inference algorithm based on this augmentation scheme. Detailed reviews of Pólya-Gamma random variables and Poisson processes are provided in Appendices B and C, respectively.

3.1 Pólya-Gamma Augmentation Scheme

A primary challenge in our model arises from the sigmoid function, whose inherent nonlinearity complicates the posterior inference. To overcome this, we adopt the Pólya-Gamma data augmentation scheme introduced in [34]. The key insight of this approach is that the sigmoid function can be represented in terms of Pólya-Gamma random variables. Define the function

$$f(\omega, z) := \frac{z}{2} - \frac{z^2}{2}\omega - \log(2). \quad (8)$$

Then, the following identity holds:

$$\sigma(z) = \int_0^\infty e^{f(\omega, z)} p_{\text{PG}}(\omega \mid 1, 0) d\omega, \quad (9)$$

where $p_{\text{PG}}(\omega \mid 1, 0)$ denotes the density of a Pólya-Gamma random variable with parameters $(1, 0)$.

Since our model considers N observations, we apply this augmentation scheme to each data point. Accordingly, we introduce N independent Pólya-Gamma random variables, denoted by $\boldsymbol{\omega} = \{\omega_i\}_{i=1}^N$, each distributed according to $p_\omega(\omega_i) = p_{\text{PG}}(\omega_i \mid 1, 0)$ and with a joint density

$$p_\omega(\boldsymbol{\omega}) = \prod_{i=1}^N p_\omega(\omega_i) = \prod_{i=1}^N p_{\text{PG}}(\omega_i \mid 1, 0). \quad (10)$$

3.2 Poisson Process Augmentation Scheme

Evaluating the likelihood in (2) requires computing an integral involving a sample function drawn from the BNN prior. This integral is generally analytically intractable, due to the nonparametric and highly non-linear nature of BNN sample paths. To address this, we propose a Poisson process augmentation scheme. By substituting the sigmoid identity from (9), the intractable integral for the i^{th} data point becomes

$$\int_0^{y_i} \lambda_0(t, \mathbf{x}_i; \phi) \sigma(g(t, \mathbf{x}_i; \boldsymbol{\theta})) dt = \int_0^{y_i} \int_0^\infty \left(1 - e^{f(\omega, -g(t, \mathbf{x}_i; \boldsymbol{\theta}))}\right) \lambda_0(t, \mathbf{x}_i; \phi) p_{\text{PG}}(\omega \mid 1, 0) d\omega dt,$$

where $p_{\text{PG}}(\omega \mid 1, 0)$ is the density of a Pólya-Gamma random variable. The key insight here is that this double integral can be expressed as an expectation over a marked Poisson process.

Before proceeding further, we briefly review the concept of a marked Poisson process. A marked Poisson process extends the standard Poisson process by associating each event (or location) with an additional random variable known as a mark. In our case, each event occurs at time t and is accompanied by a positive mark ω . With this in mind, consider the space $[0, y_i] \times \mathbb{R}_+$ which consists of points (t, ω) where $t \in [0, y_i]$ and $\omega \in \mathbb{R}_+$. We then denote by Ψ_i a marked Poisson process on $[0, y_i] \times \mathbb{R}_+$ with intensity

$$\lambda_i(t, \omega; \phi) := \lambda_0(t, \mathbf{x}_i; \phi) p_{\text{PG}}(\omega \mid 1, 0), \quad (t, \omega) \in [0, y_i] \times \mathbb{R}_+. \quad (11)$$

Under suitable assumptions on the BNN $g(t, \mathbf{x}_i; \boldsymbol{\theta})$, Campbell's theorem allows us to express the integral as

$$\exp\left(-\int_0^{y_i} \int_0^\infty \left(1 - e^{f(\omega, -g(t, \mathbf{x}_i; \boldsymbol{\theta}))}\right) \lambda_i(t, \omega; \phi) d\omega dt\right) = \mathbb{E}_{\Psi_i \sim \mathbb{P}_{\Psi_i|\phi}} \left[\prod_{(t, \omega)_j \in \Psi_i} e^{f(\omega_j, -g(t_j, \mathbf{x}_i; \boldsymbol{\theta}))} \right], \quad (12)$$

where $\mathbb{P}_{\Psi_i|\phi}$ is the path measure of the process Ψ_i . In (12), we take the convention that an empty product equals 1. Equation (12) corresponds to the term with the intractable integral on the right-hand side of (2). This representation enables us to avoid time discretization, allowing an exact and efficient evaluation of the integral. Since our model involves N observations, we apply this augmentation scheme to each data point by introducing N independent marked Poisson processes, denoted by $\boldsymbol{\Psi} = \{\Psi_i\}_{i=1}^N$.

3.3 Augmented Likelihood

Leveraging both the Pólya–Gamma and the marked Poisson process augmentation schemes, we can reformulate the likelihood given in (2) in a tractable way. With these auxiliary variables, we define the *augmented likelihood* density for the i^{th} observation as

$$p(y_i, \delta_i \mid \mathbf{x}_i, \phi, g(\cdot; \boldsymbol{\theta}), \omega_i, \Psi_i) := \left(\lambda_0(y_i, \mathbf{x}_i; \phi) e^{f(\omega_i, g(y_i, \mathbf{x}_i; \boldsymbol{\theta}))} \right)^{\delta_i} \left(\prod_{(t, \omega)_{j \in \Psi_i}} e^{f(\omega_j, -g(t_j, \mathbf{x}_i; \boldsymbol{\theta}))} \right), \quad (13)$$

where $f(\omega, z)$ was defined in (8). The following proposition formalizes the data augmentation scheme.

Theorem 3.1 (Data Augmentation). *Assume for each $i = 1, \dots, N$ that the function $g(\cdot, \mathbf{x}_i; \cdot) \in C([0, y_i] \times \mathbb{R}^m)$. Let $p(y_i, \delta_i \mid \mathbf{x}_i, \phi, g(\cdot; \boldsymbol{\theta}))$ be the likelihood density given in (2). Additionally, let $p(y_i, \delta_i \mid \mathbf{x}_i, \phi, g(\cdot; \boldsymbol{\theta}), \omega_i, \Psi_i)$ be the augmented likelihood density defined in (13). Then,*

$$p(y_i, \delta_i \mid \mathbf{x}_i, \phi, g(\cdot; \boldsymbol{\theta})) = \mathbb{E}_{\omega_i \sim p_{\omega}, \Psi_i \sim \mathbb{P}_{\Psi_i \mid \phi}} [p(y_i, \delta_i \mid \mathbf{x}_i, \phi, g(\cdot; \boldsymbol{\theta}), \omega_i, \Psi_i)].$$

The proof of Theorem 3.1 is postponed to Appendix N.1. Analogous data augmentation schemes have been proposed in [9, 44, 2]. However, to the best of our knowledge, this is the first application of such an approach to survival analysis, and the first to provide a rigorous theoretical framework that establishes the validity of the methodology.

Using the assumption from Section 2.1 that (y_i, δ_i) are i.i.d. conditional on $(\mathbf{x}_i, \phi, g(\cdot; \boldsymbol{\theta}))$, and given the structure of the data augmentation, we observe that (y_i, δ_i) are conditionally independent of ω_j and Ψ_j for all $j \neq i$. As a result, the full-sample augmented likelihood factorizes as a simple product:

$$p(\mathcal{D} \mid \mathbf{X}, \phi, g(\cdot; \boldsymbol{\theta}), \boldsymbol{\omega}, \boldsymbol{\Psi}) = \prod_{i=1}^N p(y_i, \delta_i \mid \mathbf{x}_i, \phi, g(\cdot; \boldsymbol{\theta}), \omega_i, \Psi_i). \quad (14)$$

4 Variational Inference in the Augmented Space

4.1 Variational Mean–Field Approximation

Computing the posterior distribution $\mathbb{P}(\phi, \boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\Psi} \mid \mathcal{D}, \mathbf{X})$ is analytically intractable. Therefore, we consider a variational inference algorithm that aims to find an approximating variational distribution $\mathbb{Q}(\phi, \boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\Psi})$ that minimizes the KL divergence from the true posterior distribution.

To make the optimization tractable, we restrict our search to distributions that satisfy the following *mean-field* factorization:

$$\mathbb{Q}(\phi, \boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\Psi}) = \mathbb{Q}_{\phi}(\phi) \times \mathbb{Q}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \times \mathbb{Q}_{\boldsymbol{\omega}}(\boldsymbol{\omega}) \times \mathbb{Q}_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}).$$

Here, we take $\mathbb{Q}_{\phi}(\phi)$, $\mathbb{Q}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $\mathbb{Q}_{\boldsymbol{\omega}}(\boldsymbol{\omega})$ to admit densities $q_{\phi}(\phi)$, $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $q_{\boldsymbol{\omega}}(\boldsymbol{\omega})$ with respect to the Lebesgue measures $d\phi$, $d\boldsymbol{\theta}$ and $d\boldsymbol{\omega}$. The remaining factor $\mathbb{Q}_{\boldsymbol{\Psi}}(\boldsymbol{\Psi})$ is a measure on the space of marked point-process paths, which does not admit a density with respect to the Lebesgue measures (see, e.g., a similar discussion for GPs in [30]).

To handle this within the variational inference framework, we must introduce a reference measure $\mathbb{P}_{\boldsymbol{\Psi},*}$, which plays the role of a “Lebesgue-like” base measure on path space (see Definition E.1 for details). We then assume our variational law $\mathbb{Q}_{\boldsymbol{\Psi}}$ is absolutely continuous with respect to $\mathbb{P}_{\boldsymbol{\Psi},*}$, so that it admits a strictly positive Radon–Nikodym derivative $\frac{d\mathbb{Q}_{\boldsymbol{\Psi}}}{d\mathbb{P}_{\boldsymbol{\Psi},*}}$ which satisfies the normalization $\mathbb{E}_{\boldsymbol{\Psi} \sim \mathbb{P}_{\boldsymbol{\Psi},*}} \left[\frac{d\mathbb{Q}_{\boldsymbol{\Psi}}}{d\mathbb{P}_{\boldsymbol{\Psi},*}}(\boldsymbol{\Psi}) \right] = 1$. These conditions ensure that $\mathbb{Q}_{\boldsymbol{\Psi}}$ is a valid probability measure on the space of marked point-process paths (see Appendix E for further technical details).

This formulation enables us to express the KL divergence between the variational distribution and the true posterior in terms of the ELBO:

$$D_{\text{KL}}(\mathbb{Q}(\phi, \boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\Psi}) \parallel \mathbb{P}(\phi, \boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\Psi} \mid \mathcal{D}, \mathbf{X})) = -\mathcal{L}_{\text{ELBO}}(g) + \text{const}, \quad (15)$$

where the ELBO is defined as

$$\mathcal{L}_{\text{ELBO}}(g) := \mathbb{E}_{\phi \sim q_\phi, \theta \sim q_\theta, \omega \sim q_\omega, \Psi \sim \mathbb{Q}_\Psi} \left[\log \frac{p(\mathcal{D} \mid \phi, g(\cdot; \theta), \omega, \Psi) p_\phi(\phi) p_\theta(\theta) p_\omega(\omega) \frac{d\mathbb{P}_{\Psi|\phi}}{d\mathbb{P}_{\Psi,*}}(\Psi)}{q_\phi(\phi) q_\theta(\theta) q_\omega(\omega) \frac{d\mathbb{Q}_\Psi}{d\mathbb{P}_{\Psi,*}}(\Psi)} \right] \quad (16)$$

and where $\frac{d\mathbb{P}_{\Psi|\phi}}{d\mathbb{P}_{\Psi,*}}$ is the Radon-Nykodim derivative of the true conditional law $\mathbb{P}_{\Psi|\phi}$ with respect to $\mathbb{P}_{\Psi,*}$, cf. (28). From (15), it follows that minimizing the KL divergence is equivalent to maximizing the ELBO.

4.2 Local Linearization of the Bayesian Neural Network

A crucial insight is that the data augmentation strategy transforms the intractable likelihood density in (2) into a form that is *conditionally Gaussian*, as shown below:

$$p(y_i, \delta_i \mid \mathbf{x}_i, \phi, g(\cdot; \theta), \omega_i, \Psi_i) \propto \exp \left(\delta_i \frac{g(y_i, \mathbf{x}_i; \theta)}{2} - \delta_i \frac{g(y_i, \mathbf{x}_i; \theta)^2}{2} \omega_i \right) \exp \left(\sum_{(t, \omega)_j \in \Psi_i} \frac{g(t_j, \mathbf{x}_i; \theta)}{2} - \frac{g(t_j, \mathbf{x}_i; \theta)^2}{2} \omega_j \right).$$

This transformation is particularly advantageous when placing a GP prior on $g(\cdot; \theta)$, as it induces conjugacy in the model. Conjugacy is crucial for variational inference because it enables efficient computation of the ELBO (16), which involves taking expectations over the distribution of θ . However, when $g(\cdot; \theta)$ is a BNN, these expectations generally lack closed-form solutions, making exact Bayesian updates intractable. As a result, we seek to approximate $g(\cdot; \theta)$ in a way that retains the expressive power of neural networks while preserving Gaussian conjugacy to enable tractable inference.

We adopt the *local linearization* approximation introduced in [18]. This approach approximates the BNN $g(\cdot; \theta)$ using a first-order Taylor expansion around a reference point θ^* :

$$g(t, \mathbf{x}; \theta) \approx g^{\text{lin}}(t, \mathbf{x}; \theta) := g(t, \mathbf{x}; \theta^*) + \mathbf{J}_{\theta^*}(t, \mathbf{x})^\top (\theta - \theta^*), \quad (17)$$

where $[\mathbf{J}_\theta(t, \mathbf{x})]_j = \frac{\partial g(t, \mathbf{x}; \theta)}{\partial \theta_j}$ is the Jacobian of the BNN with respect to the parameters θ . Following [18], we select $\theta^* = \theta_{\text{MAP}}$ as the maximum a posteriori (MAP) estimate, which is defined as:

$$(\theta_{\text{MAP}}, \phi_{\text{MAP}}) := \arg \max_{\theta, \phi} p(\theta, \phi \mid \mathcal{D}, \mathbf{X}), \quad (18)$$

where $p(\theta, \phi \mid \mathcal{D}, \mathbf{X})$ is the posterior density defined in (7). By centering the linearization at θ_{MAP} , we ensure maximal approximation accuracy precisely where Bayesian inference is most sensitive — in the high-probability region of the posterior that dominates both parameter uncertainty quantification and predictive distributions. The procedure used to obtain the MAP estimates of (18) is detailed in Appendix F. Under the assumption of a Gaussian prior on the BNN parameters (4), the local linearization induces the GP prior

$$g^{\text{lin}} \sim \mathcal{GP}(\mu, \kappa)$$

with mean function μ and and covariance function κ given by:

$$\begin{aligned} \mu(t, \mathbf{x}) &:= g(t, \mathbf{x}; \theta_{\text{MAP}}) + \mathbf{J}_{\theta_{\text{MAP}}}(t, \mathbf{x})^\top (\mathbb{E}_{\theta \sim p_\theta}[\theta] - \theta_{\text{MAP}}) \\ \kappa((t, \mathbf{x}), (t', \mathbf{x}')) &:= \mathbf{J}_{\theta_{\text{MAP}}}(t, \mathbf{x}) \mathbf{J}_{\theta_{\text{MAP}}}(t', \mathbf{x}')^\top. \end{aligned}$$

Incorporating this approximation into our variational framework allows us to exploit Gaussian conjugacy for fast, closed-form updates, while still preserving the flexibility of neural networks. Concretely, we take a Taylor expansion of the ELBO around g^{lin} and, by truncating at the lowest order term, obtain the simple approximation

$$\mathcal{L}_{\text{ELBO}}(g) \approx \mathcal{L}_{\text{ELBO}}(g^{\text{lin}}).$$

Our approach is analogous to the method introduced in [40, Section 3.2], where the authors apply Delta Method Variational Inference by approximating the ELBO around a fixed point in parameter space. In contrast, we extend this idea by approximating the ELBO around a reference function g^{lin} , rather than a fixed point.

4.3 Coordinate Ascent Variational Inference

We adopt a Coordinate Ascent Variational Inference (CAVI) approach, allowing us to draw on standard results from variational inference (see, e.g., [3, Chapter 10.1]). In this framework, the optimal variational distributions are derived by maximizing the linearized ELBO, $\mathcal{L}_{\text{ELBO}}(g^{\text{lin}})$, with each distribution depending on the current state of the others. The algorithm proceeds by cyclically updating each variational distribution while keeping the others fixed. This iterative process progressively refines the optimal variational distributions, ultimately leading to the best possible approximation of the posterior distribution. A complete derivation of each optimal variational distribution is provided in Appendix G while the complete CAVI algorithm is presented in Appendix H.

At the k^{th} iteration the optimal variational distributions for the parameters ϕ and θ are given by

$$q_{\phi}^{(k)}(\phi) = \text{Gamma}\left(\tilde{\alpha}^{(k)}, \tilde{\beta}\right), \quad q_{\theta}^{(k)}(\theta) = \mathcal{N}\left(\tilde{\mu}^{(k)}, \tilde{\Sigma}^{(k)}\right),$$

where $(\tilde{\alpha}^{(k)}, \tilde{\beta})$ and $(\tilde{\mu}^{(k)}, \tilde{\Sigma}^{(k)})$ are given in Appendix G.3 and G.4, respectively. At the k^{th} iteration, the optimal update for the auxiliary parameters ω is given by

$$q_{\omega}^{(k)}(\omega) = \prod_{i=1}^N p_{\text{PG}}(\omega_i \mid 1, \tilde{c}_i^{(k)}),$$

where $\tilde{c}_i^{(k)}$ is given in Appendix G.1. Finally, at the k^{th} iteration, the optimal variational law $\mathbb{Q}_{\Psi}^{(k)}$ is the probability measure under which each Ψ_i ($i = 1, \dots, N$) is a marked Poisson process on $[0, y_i] \times \mathbb{R}_+$ with intensity function $\lambda_i^{\mathbb{Q},(k)}$, as given in Appendix G.2.

It is important to emphasize that we did not impose a specific form on the variational distributions — for example, we did not assume $q_{\theta}(\theta)$ to be Gaussian. Instead, we derived our results by variationally minimizing the KL divergence over the full space of distributions. This contrasts with methods that fix a parametric form and use the reparameterization trick with Monte Carlo gradient estimates.

Finally, in Appendix I, we demonstrate that, by exploiting the Woodbury matrix identity, our inference updates require only $\mathcal{O}(m)$ time complexity (m is the number of weights in the neural network architecture). This linear scaling renders our Bayesian framework feasible for contemporary large-scale deep neural architectures, which are well suited to model high-dimensional data.

5 Experiments

Details on the experimental setup, including dataset descriptions, hyperparameter tuning for the benchmark methods and evaluation metrics definitions are provided in Appendix J. Moreover, the implementation details for *NeuralSurv* are provided in Appendix K.

To comprehensively evaluate *NeuralSurv*, we compare its performance against the same set of benchmark models evaluated in *DySurv* [32]. These include: *MTLR* [43], *DeepHit* [28], *DeepSurv* [20], *Logistic Hazard* [13], *CoxTime* [25], *CoxCC* [25], *PMF* [24], *PCHazard* [24], *BCESurv* [26], and *DySurv* [32]. A detailed overview of these methods is provided in Appendix L and Table A1. Except for *DySurv*, which employs an autoencoder framework, we adopt the same neural-network architecture across all benchmark models and *NeuralSurv* to parameterize the hazard function. For *DySurv*, we use the original autoencoder architecture specified in its implementation.

We assess discriminative performance using the Antolini’s concordance index (C-index) [1], and evaluate model calibration with the inverse probability of censoring weighting (IPCW) integrated Brier score (IBS) [14]. The C-index evaluates how well a model performs by measuring the concordance between the rankings of the predicted event times and the true event times. The C-index ranges from 0 to 1, where higher values indicate better discriminative performance; a value of 0.5 corresponds to random guessing. Similar to the mean squared error, the Brier score (BS) assesses the accuracy of an estimated survival function at some time t . The IPCW are observation-specific weights that account for censoring in survival data, ensuring that the BS remains unbiased. The IPCW IBS is the integral of the IPCW BS over the observational period. The C-index and the IPCW IBS metrics are computed using the *TorchSurv* package [33].

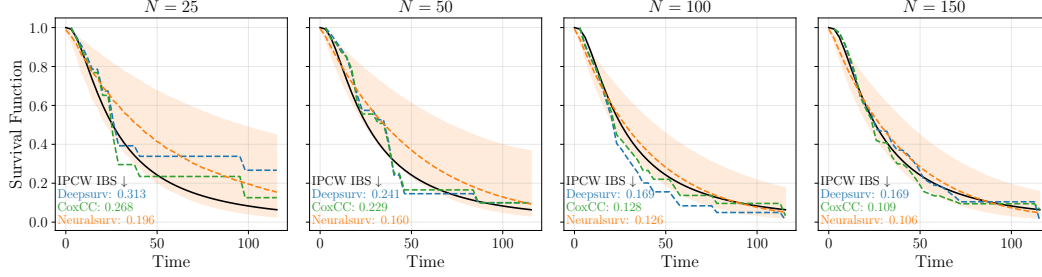


Figure 1: Comparison of the true survival function (black) with the estimated survival functions from *NeuralSurv* and the two top-performing benchmark models (colored) on synthetic data. The time axis is truncated at the maximum observed event time in the training data. Each panel represents a different training set size. The IPCW IBS score is reported for each method in each panel, with lower values indicating better predictive accuracy. *NeuralSurv* estimates the full posterior over survival functions, and the 90% credible interval is shown as a ribbon around its estimate.

5.1 Synthetic Data Experiment

In this section, we present experiments conducted on synthetic data. The experimental setup was inspired by [10] and constitutes a broadly applicable evaluation benchmark. We simulated the training sets with increasing sizes $N = 25, 50, 100$ and 150 samples where the event time was drawn from two distributions: $p_0(T) = \log \text{Normal}(3, 0.8^2)$ and $p_1(T) = \log \text{Normal}(3.5, 1^2)$. Each observation included a covariate indicating whether the event time was sampled from p_0 or p_1 , along with three additional noisy covariates generated from a standard normal distribution. The censoring times were drawn from an exponential distribution with a rate of 0.025 yielding an average censoring rate of 54% across the four synthetic datasets. The test set was generated using the same data-generating process, fixed to 100 observations, and held constant across all experiments.

Figure 1 presents the true survival function alongside the predicted functions from *NeuralSurv* and the two top-performing benchmark models, selected based on IPCW IBS. Each panel represents a different training set size. As the number of training samples increases, the predicted survival functions match more closely the true survival function. The results show that *NeuralSurv* consistently ranks as the best method according to IPCW IBS, and its predictive accuracy improves with larger sample sizes. Beyond its competitive performance, *NeuralSurv* also provides Bayesian credible intervals, offering uncertainty estimates for survival probabilities, an important feature typically absent in benchmark models. Notably, these credible intervals appropriately narrow as more data becomes available, demonstrating well-calibrated uncertainty quantification. Corresponding C-index and IPCW IBS scores for all methods are reported in Table A2.

5.2 Real Survival Data Experiments

To comprehensively evaluate *NeuralSurv*, we conduct experiments on eight real survival datasets: the chemotherapy for colon cancer (COLON), the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), the Rotterdam and German Breast Cancer Study Group (GBSG), the National Wilms’ Tumor Study (NWTCS), the Worcester Heart Attack Study (WHAS), the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatment (SUPPORT), the Veterans administration Lung Cancer trial (VLC) and the Sac 3 simulation study. Each dataset is subsampled to 125 observations to highlight the advantages of a Bayesian approach in data-scarce regimes. The data is randomly partitioned into five equally sized folds, with each fold serving as a distinct train/test split, comprising 100 training samples and 25 test samples per fold.

Table 1 presents the results on the held-out test sets for three representative datasets, while results for the remaining datasets are shown in Table A3. Across all eight datasets, *NeuralSurv* achieves the best IPCW IBS score on seven, demonstrating superior calibration performance compared to benchmark methods. This improvement can be attributed to the Bayesian framework, which naturally incorporates model uncertainty and provides better regularization in data-scarce settings.

Method	COLON		METABRIC		GBSG	
	C-index \uparrow	IPCW IBS \downarrow	C-index \uparrow	IPCW IBS \downarrow	C-index \uparrow	IPCW IBS \downarrow
MTLR [43]	0.562	0.298	0.548	0.279	0.602	0.273
DeepHit [28]	0.478	0.28	0.511	<u>0.243</u>	0.578	0.309
DeepSurv [20]	0.572	0.326	0.523	0.289	0.618	0.252
Logistic Hazard [13]	0.490	0.321	0.541	0.317	0.618	0.296
CoxTime [25]	0.578	<u>0.277</u>	0.533	0.307	0.599	0.285
CoxCC [25]	<u>0.584</u>	0.289	0.575	0.257	0.646	<u>0.240</u>
PMF [24]	0.509	0.324	0.440	0.336	<u>0.655</u>	0.250
PCHazard [24]	0.538	0.297	0.541	0.291	0.609	0.249
BCESurv [26]	0.491	0.302	0.616	0.277	0.581	0.273
DySurv [32]	0.488	0.536	0.561	0.465	0.572	0.485
NeuralSurv (Ours)	0.671	0.218	<u>0.584</u>	0.212	0.657	0.188

Table 1: Performance comparison of deep survival models over five different train/test splits of each dataset. The best results for each metric are shown in bold, and the second-best results are underlined. \uparrow indicates higher is better; \downarrow indicates lower is better.

Additional an ablation study using a larger training set of 250 observations is presented in Table A4. *NeuralSurv* continues to outperform benchmark methods under this setting in terms of calibration performance demonstrating the robustness of the method to training size. Furthermore, we also include results from traditional survival models, such as the Cox Proportional Hazards model [7], the Weibull Accelerated Failure Time model [6], the Random Survival Forest [19], and the Survival Support Vector Machine [36] in Table A5. These models, reported, often achieve strong performance in data-scarce regimes. However, they are not designed to leverage high-dimensional or complex feature representations, which limits their applicability in modern deep learning contexts. Our focus remains on evaluating deep survival methods that can scale with data complexity, but we include these classical baselines for reference and completeness.

6 Conclusion

We propose the first fully Bayesian framework for deep survival analysis with time-varying covariate-risk relationships. On both synthetic and real-world datasets, in data-scarce regimes, our method consistently achieves better calibration than state-of-the-art deep survival models and matches or surpasses their discriminative performance, while offering fully Bayesian uncertainty quantification.

Despite its strengths, *NeuralSurv* relies on three key simplifying assumptions. First, we assume a sigmoidal hazard function, a choice shared by prior work (e.g., [10, 21]), which may not capture all risk dynamics. Second, our mean-field variational inference treats parameters ϕ and θ as independent, ignoring posterior correlations. Third, we linearize the network around the MAP estimate to enforce conjugacy. In real-world settings, however, the true posterior can be multimodal and strongly correlated, so this local, factorized approximation may overlook secondary modes or misestimate joint uncertainty.

Concerning the computational efficiency of our method, the coordinate-ascent updates scale linearly with network size but still require full-dataset passes each iteration. For very large cohorts, this becomes a bottleneck. Extending the algorithm to use stochastic or mini-batch updates would preserve conjugacy benefits while improving scalability.

We believe that *NeuralSurv* has the potential to make a positive societal impact. For instance, as healthcare data becomes increasingly diverse, there is a growing need for models that can handle multimodal data within time-to-event analyses effectively. *NeuralSurv* represents an important first step toward accommodating such data within a Bayesian deep learning framework.

References

- [1] Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005.
- [2] Ifigeneia Apostolopoulou, Scott Linderman, Kyle Miller, and Artur Dubrawski. Mutually Regressive Point Processes. In *Advances in Neural Information Processing Systems*, volume 32,

2019.

- [3] Christopher Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, NY, 2016.
- [4] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Network. *Proceedings of the 32nd International Conference on Machine Learning*, 37:1613–1622, 2015.
- [5] Pierre Bremaud. *Point Processes and Queues*. Springer Series in Statistics. Springer, 1981.
- [6] Kevin J. Carroll. On the use and utility of the Weibull model in the analysis of survival data. *Controlled Clinical Trials*, 24(6):682–701, 2003.
- [7] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 34(2):187–202, 1972.
- [8] Cameron Davidson-Pilon. lifelines: survival analysis in Python. *Journal of Open Source Software*, 4(40):1317, 2019. (version 0.30.0).
- [9] Christian Donner and Manfred Opper. Efficient Bayesian Inference of Sigmoidal Gaussian Cox Processes. *Journal of Machine Learning Research*, 19(67):1–34, 2018.
- [10] Tamara Fernandez, Nicolas Rivera, and Yee Whye Teh. Gaussian Processes for Survival Analysis. *Advances in Neural Information Processing Systems*, 29, 2016.
- [11] Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. GPYtorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. In *Advances in Neural Information Processing Systems*, volume 31, 2018.
- [12] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Chapman & Hall/CRC, 3 edition, 2013.
- [13] Michael F. Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019.
- [14] Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17–18):2529–2545, 1999.
- [15] Frank E Harrell, Kerry L Lee, and Daniel B Mark. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine*, 15(4):361–387, 1996.
- [16] Nicholas J Higham. *Functions of matrices : theory and computation*. Cambridge University Press, 2008.
- [17] Joseph G Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. *Bayesian Survival Analysis*. Springer series in statistics. Springer, 2010.
- [18] Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 130:703–711, 2021.
- [19] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3), 2008.
- [20] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), 2018.
- [21] Minyoung Kim and Vladimir Pavlovic. Variational inference for gaussian process models for survival analysis. *UAI*, pages 435–445, 2018.

- [22] J. F. C. Kingman. *Poisson Processes*. Oxford Studies in Probability. Clarendon Press, 1992.
- [23] Håvard Kvamme. pycox: Survival analysis with PyTorch. <https://pypi.org/project/pycox/>, 2024. (version 0.3.0).
- [24] Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 27(4):710–736, 2021.
- [25] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-Event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.
- [26] Håvard Kvamme and Ørnulf Borgan. The Brier Score under Administrative Censoring: Problems and a Solution. *Journal of Machine Learning Research*, 24(2):1–26, 2023.
- [27] Jerald F Lawless. *Statistical models and methods for lifetime data*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2 edition, 2002.
- [28] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van der Schaar. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.
- [29] Jialiang Li and Shuangge Ma. *Survival Analysis in Medicine and Genetics*. Chapman & Hall/CRC Biostatistics Series. Chapman & Hall/CRC, 2023.
- [30] Alexander G de G Matthews, James Hensman, Richard Turner, and Zoubin Ghahramani. On sparse Variational Methods and the Kullback-Leibler Divergence between Stochastic Processes. In *Artificial Intelligence and Statistics*, pages 231–239, 2016.
- [31] J W McPherson. *Reliability Physics and Engineering: Time-To-Failure Modeling*. Springer International Publishing, 3 edition, 2019.
- [32] Munib Mesinovic, Peter Watkinson, and Tingting Zhu. DySurv: dynamic deep learning model for survival analysis with conditional variational inference. *Journal of the American Medical Informatics Association*, page ocae271, 2024.
- [33] Mélodie Monod, Peter Krusche, Qian Cao, Berkman Sahiner, Nicholas Petrick, David Ohlssen, and Thibaud Coroller. TorchSurv: A Lightweight Package for Deep Survival Analysis. *Journal of Open Source Software*, 9(104):7341, 2024. (version 0.1.4).
- [34] James G. Scott Nicholas G. Polson and Jesse Windle. Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.
- [35] Sebastian Pölsterl. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. (version 0.24.0).
- [36] Sebastian Pölsterl, Nassir Navab, and Amin Katouzian. *Fast Training of Support Vector Machines for Survival Analysis*, page 243–259. Springer International Publishing, 2015.
- [37] Alejandro Quiroz Flores. *Survival Analysis: A New Guide for Social Scientists*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press, 2022.
- [38] Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep gaussian processes. *Advances in Neural Information Processing Systems*, 30, 2017.
- [39] Terry M Therneau. survival: A package for survival analysis in R. <https://CRAN.R-project.org/package=survival>, 2024. (version 3.7.0).
- [40] Chong Wang and David M. Blei. Variational Inference in Nonconjugate Models. *Journal of Machine Learning Research*, 14(1):1005–1031, 2013.
- [41] Ping Wang, Yan Li, and Chandan K. Reddy. Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys*, 51(6):1–36, 2019.

- [42] Simon Wiegrebe, Philipp Kopper, Raphael Sonabend, Bernd Bischl, and Andreas Bender. Deep learning for survival analysis: a review. *Artificial Intelligence Review*, 57(3), 2024.
- [43] Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors. *Advances in Neural Information Processing Systems*, 24, 2011.
- [44] Feng Zhou, Zhidong Li, Xuhui Fan, Yang Wang, Arcot Sowmya, and Fang Chen. Efficient Inference for Nonparametric Hawkes Processes Using Auxiliary Latent Variables. *Journal of Machine Learning Research*, 21(241):1–31, 2020.

A Review of Survival Analysis

This appendix offers a concise summary of the survival-analysis framework on which our approach is built. For an in-depth review, the reader is referred to [27].

Survival data for each observation consist of three components:

- **Feature vector:** A covariate vector $\mathbf{x} \in \mathbb{R}^p$ capturing baseline characteristics;
- **Event time:** a nonnegative random variable T measuring the time from baseline to the occurrence of the event of interest;
- **Event indicator:** A binary variable δ , which takes the value 1 if the event is observed, and 0 if the event is not observed within the observational period. In the latter case, the observation's data is said to be right-censored, meaning that the only available information is the time of the last follow-up before the event could occur.

To handle censoring uniformly, we introduce a censoring time C and record the observed time $y = \min(T, C)$. The event indicator can then be written succinctly as $\delta = \mathbb{1}_{\{T \leq C\}}$. Throughout, we assume noninformative right-censoring, i.e. conditional on the covariates, the censoring time is independent of the event time: $C \perp T \mid \mathbf{x}$. Conditional on \mathbf{x} , we let the event time T have cumulative distribution function $F(t \mid \mathbf{x})$ and probability density function $f(t \mid \mathbf{x})$ such that

$$F(t \mid \mathbf{x}) = \mathbb{P}(T \leq t \mid \mathbf{x}) = \int_0^t f(s \mid \mathbf{x}) \, ds$$

for $t \in [0, \infty)$. The survival function gives the probability of remaining event-free beyond time t :

$$S(t \mid \mathbf{x}) := \mathbb{P}(T > t \mid \mathbf{x}) = 1 - F(t \mid \mathbf{x}) = \int_t^\infty f(s \mid \mathbf{x}) \, ds,$$

for $t \in [0, \infty)$. An important modeling quantity is the hazard function, which represents the instantaneous event rate at time t given survival up to t :

$$\lambda(t \mid \mathbf{x}) := \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}(t \leq T < t + \Delta t \mid T \geq t, \mathbf{x})}{\Delta t} = \frac{f(t \mid \mathbf{x})}{S(t \mid \mathbf{x})}.$$

Equivalently,

$$\lambda(t \mid \mathbf{x}) = -\frac{d}{dt} \log S(t \mid \mathbf{x}),$$

so that the survival function can be written in terms of the hazard:

$$S(t \mid \mathbf{x}) = \exp\left(-\int_0^t \lambda(s \mid \mathbf{x}) \, ds\right).$$

B Review of Pólya-Gamma Random Variables

We follow [34] in defining the family of Pólya-Gamma distributions and their properties.

Definition B.1 (Pólya-Gamma Distribution). A random variable ω is said to follow a *Pólya-Gamma distribution* with parameters $b > 0$ and $c \in \mathbb{R}$, denoted by $\omega \sim \text{PG}(b, c)$, if

$$\omega \stackrel{d}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{\left(k - \frac{1}{2}\right)^2 + \frac{c^2}{4\pi^2}}, \quad \text{with } g_k \stackrel{\text{i.i.d.}}{\sim} \text{Gamma}(b, 1). \quad (19)$$

The following result expresses the reciprocal of the hyperbolic cosine function raised to the power b as an infinite Gaussian mixture. This representation is central to connecting the Pólya-Gamma density with a parameter $c \neq 0$ to the case when $c = 0$.

Proposition B.2. *The reciprocal of the hyperbolic cosine raised to the power b can be represented as an infinite Gaussian mixture:*

$$\left[\cosh\left(\frac{c}{2}\right)\right]^{-b} = \int_0^\infty \exp\left(-\frac{c^2}{2}\omega\right) p_{\text{PG}}(\omega \mid b, 0) d\omega.$$

Notice that Proposition B.2 can also be read as providing a closed-form expression for the expectation $\mathbb{E}_{\omega \sim p_{\text{PG}}(\omega \mid b, 0)} \left[\exp\left(-\frac{c^2}{2}\omega\right)\right]$. Building on this representation, we can relate the density function of a Pólya-Gamma random variable with a non-zero parameter c through an exponential tilting of the Pólya-Gamma random density with $c = 0$. This connection is summarized in the next proposition.

Proposition B.3. *The Pólya-Gamma density (19) can be re-written in the form*

$$p_{\text{PG}}(\omega \mid b, c) = \exp\left(-\frac{c^2}{2}\omega\right) (\cosh(c/2))^b p_{\text{PG}}(\omega \mid b, 0). \quad (20)$$

The previous propositions not only establish key representations of the Pólya-Gamma density but also facilitate the derivation of its moment properties. In particular, one can derive the moment generating function, from which the first moment follows directly. This is captured in the next result.

Proposition B.4. *Let $p_{\text{PG}}(\omega \mid b, c)$ denote the density function of the random variable $\omega \sim \text{PG}(b, c)$, with $b > 0$ and $c \in \mathbb{R}$. Using Propositions B.2 and B.3, the moment generating function is given by*

$$\int_0^\infty e^{\xi\omega} p_{\text{PG}}(\omega \mid b, c) d\omega = \frac{\cosh^b(c/2)}{\cosh^b\left(\frac{1}{2}\sqrt{c^2 - 2\xi}\right)}. \quad (21)$$

In particular, the first moment is obtained by differentiating (21) with respect to ξ at $\xi = 0$:

$$\mathbb{E}_{\omega \sim p_{\text{PG}}(\omega \mid b, c)}[\omega] = \frac{b}{2c} \tanh\left(\frac{c}{2}\right). \quad (22)$$

Finally, the following theorem illustrates how the Pólya-Gamma distribution can be used to derive useful integral identities.

Theorem B.5. *Let $p_{\text{PG}}(\omega \mid b, 0)$ denote the density function of the random variable $\omega \sim \text{PG}(b, 0)$ with $b > 0$. Then, for all $a \in \mathbb{R}$, the following integral identity holds:*

$$\frac{e^{\psi a}}{(1 + e^\psi)^b} = 2^{-b} e^{\kappa\psi} \int_0^\infty \exp\left(-\frac{\omega\psi^2}{2}\right) p_{\text{PG}}(\omega \mid b, 0) d\omega,$$

where $\kappa = a - \frac{b}{2}$.

The following corollary is a direct application of Theorem B.5.

Corollary B.6. *Let $f(\omega, z) := \frac{z}{2} - \frac{z^2}{2}\omega - \log(2)$. Then,*

$$\sigma(z) = \frac{e^{\frac{z}{2}}}{2 \cosh(\frac{z}{2})} = \int_0^\infty e^{f(\omega, z)} p_{\text{PG}}(\omega \mid 1, 0) d\omega. \quad (23)$$

C Review of Poisson processes

This appendix briefly summarizes the properties of a Poisson process that are most relevant to our analysis. For a more comprehensive treatment, see Chapters 3 and 5 of [22].

Definition C.1 (Poisson Process). Let \mathcal{Z} be a measurable space. A random countable subset

$$\Psi = \{z \in \mathcal{Z}\}$$

is said to be a *Poisson process* on \mathcal{Z} if it satisfies the following properties:

1. **Independence:** For any sequence of disjoint subsets $\{\mathcal{Z}_k \subset \mathcal{Z}\}_{k=1}^K$, the counts

$$N(\mathcal{Z}_k) = |\Psi \cap \mathcal{Z}_k|$$

are mutually independent.

2. **Poisson Counts:** For each measurable subset $\mathcal{Z}_k \subset \mathcal{Z}$, the count $N(\mathcal{Z}_k)$ is Poisson distributed with mean

$$\int_{\mathcal{Z}_k} \lambda(z) dz,$$

where $\lambda : \mathcal{Z} \rightarrow \mathbb{R}_+$ is the intensity function.

Given a point process Ψ , we denote its path measure — that is, the probability measure induced on its sample-path space — by \mathbb{P}_Ψ . If the intensity function $\lambda(z)$ is constant, $\lambda(z) \equiv \lambda$, then Ψ is called *homogeneous*; otherwise, it is *inhomogeneous*. We now extend the concept of a Poisson process by incorporating additional random attributes, known as *marks*.

Definition C.2 (Marked Poisson Process). Let $\Psi = \{z \in \mathcal{Z}\}$ be a Poisson process on \mathcal{Z} with intensity function $\lambda : \mathcal{Z} \rightarrow \mathbb{R}_+$. Suppose that for each point z , associate a random variable ω , such that $\omega \sim p_{\omega|z}(\omega|z)$, taking values in some space \mathcal{M} . Then the collection

$$\Psi_{\mathcal{M}} = \{(z, \omega) \in \mathcal{Z} \times \mathcal{M}\}$$

defines a Poisson process on the product space $\mathcal{Z} \times \mathcal{M}$. The resulting process is known as a *marked Poisson process* with intensity

$$\lambda(z, \omega) = \lambda(z) p_{\omega|z}(\omega|z).$$

Next, we present Campbell's Theorem, which describes the law of sums taken over the points of a Poisson process (see [22, Sec. 3.2]).

Theorem C.3 (Campbell's Theorem). Let $\Psi_{\mathcal{M}}$ be a marked Poisson process on $\mathcal{Z} \times \mathcal{M}$ with intensity function $\lambda(z, \omega)$ and let $f : \mathcal{Z} \times \mathcal{M} \rightarrow \mathbb{R}$ be measurable. Then the sum

$$H(\Psi_{\mathcal{M}}) = \sum_{(z, \omega)_j \in \Psi_{\mathcal{M}}} f(z_j, \omega_j)$$

is absolutely convergent with probability one if and only if

$$\int_{\mathcal{Z} \times \mathcal{M}} \min(|f(z, \omega)|, 1) \lambda(z, \omega) dz d\omega < \infty.$$

If this condition holds, then

$$\mathbb{E}_{\Psi_{\mathcal{M}} \sim \mathbb{P}_{\Psi_{\mathcal{M}}}} \left[e^{sH(\Psi_{\mathcal{M}})} \right] = \exp \left(\int_{\mathcal{Z} \times \mathcal{M}} (e^{sf(z, \omega)} - 1) \lambda(z, \omega) dz d\omega \right)$$

for any $s \in \mathbb{C}$ for which the integral on the right converges. Moreover

$$\mathbb{E}_{\Psi_{\mathcal{M}} \sim \mathbb{P}_{\Psi_{\mathcal{M}}}} [H(\Psi_{\mathcal{M}})] = \int_{\mathcal{Z} \times \mathcal{M}} f(z, \omega) \lambda(z, \omega) dz d\omega$$

in the sense that the expectation exists if and only if the integral converges.

D Obtaining the normalizing function $Z(t, \mathbf{x})$

In this appendix we derive an efficient approximation for the normalizing constant

$$Z(t, \mathbf{x}) = \mathbb{E}_{\boldsymbol{\theta} \sim p_{\boldsymbol{\theta}}} [\sigma(g^{\text{lin}}(t, \mathbf{x}; \boldsymbol{\theta}))], \quad (24)$$

which is needed when computing the CAVI optimal updates (see Appendix G).

Recall from (4) that $\boldsymbol{\theta}$ has the following prior distribution

$$\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m),$$

where \mathbf{I}_m is the $m \times m$ identity matrix. Moreover, recall from Section 4.2 that we approximate the network output $g(t, \mathbf{x}; \boldsymbol{\theta})$ around some reference $\boldsymbol{\theta}^*$ by its first-order linearization

$$g^{\text{lin}}(t, \mathbf{x}; \boldsymbol{\theta}) := g(t, \mathbf{x}; \boldsymbol{\theta}^*) + \mathbf{J}_{\boldsymbol{\theta}^*}(t, \mathbf{x})^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*),$$

where $\mathbf{J}_{\boldsymbol{\theta}^*}(t, \mathbf{x})$ denotes the Jacobian of $g(t, \mathbf{x}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Because $\boldsymbol{\theta}$ is Gaussian, the linearized output is also Gaussian:

$$g^{\text{lin}}(t, \mathbf{x}; \boldsymbol{\theta}) \sim \mathcal{N}\left(g(t, \mathbf{x}; \boldsymbol{\theta}^*) - \mathbf{J}_{\boldsymbol{\theta}^*}(t, \mathbf{x})^\top \boldsymbol{\theta}^*, \|\mathbf{J}_{\boldsymbol{\theta}^*}(t, \mathbf{x})\|_2^2\right).$$

In order to approximate $Z(t, \mathbf{x})$ we wish to leverage a well-known asymptotic approximation. Specifically, for a normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ it holds that

$$\mathbb{E}_{X \sim \mathcal{N}(\mu, \sigma^2)} [\sigma(X)] \approx \sigma \left(\frac{\mu}{\sqrt{1 + \frac{\pi}{8} \sigma^2}} \right). \quad (25)$$

We can apply the result in (25) to the normal random variable $g^{\text{lin}}(t, \mathbf{x}; \boldsymbol{\theta})$ and approximate $Z(t, \mathbf{x})$ as

$$Z(t, \mathbf{x}) \approx \sigma \left(\frac{g(t, \mathbf{x}; \boldsymbol{\theta}^*) - \mathbf{J}_{\boldsymbol{\theta}^*}(t, \mathbf{x})^\top \boldsymbol{\theta}^*}{\sqrt{1 + \frac{\pi}{8} \|\mathbf{J}_{\boldsymbol{\theta}^*}(t, \mathbf{x})\|_2^2}} \right).$$

Since in (24) we are taking the expectation under the prior $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, it is natural to linearize around the prior mean, therefore, we set $\boldsymbol{\theta}^* = \mathbf{0}$.

E Combining Variational Inference with Poisson Processes

In this appendix, we outline how our variational-inference framework integrates marked Poisson processes — an essential part in the mean-field variational approximation of Section 4.1. For a fully rigorous, measure-theoretic treatment, the reader is referred to Brémaud’s text [5]. Our development relies in particular on Theorem T10 in Chapter VIII of that book, which shows how the law of a marked Poisson process arises via a change of measure using the appropriate Radon–Nikodym derivative.

We begin by fixing a reference measure on path space:

Definition E.1 (Reference measure $\mathbb{P}_{\Psi,*}$). Let $\Psi = (\Psi_1, \dots, \Psi_N)$ be N independent marked Poisson processes, where each Ψ_i is defined on the product space $[0, y_i] \times \mathbb{R}_+$. We define $\mathbb{P}_{\Psi,*}$ to be their joint law where each Ψ_i has intensity

$$\lambda_{*,i}(t, \omega) = t^{\rho-1} p_{\text{PG}}(\omega \mid 1, 0) \quad \text{for all } (t, \omega) \in [0, y_i] \times \mathbb{R}_+. \quad (26)$$

Next, let $\gamma_i^{\mathbb{Q}}(t)$ be a deterministic function on $[0, y_i]$ and let $h_i^{\mathbb{Q}}(t, \omega)$ be a deterministic density on $[0, y_i] \times \mathbb{R}_+$ satisfying

$$\int_0^\infty h_i^{\mathbb{Q}}(t, \omega) p_{\text{PG}}(\omega \mid 1, 0) d\omega = 1 \quad \text{and} \quad \int_0^{y_i} \gamma_i^{\mathbb{Q}}(t) t^{\rho-1} dt < \infty \quad (27)$$

for all $t \in [0, y_i]$ and $i = 1, \dots, N$. It is convenient to introduce the function

$$\lambda_i^{\mathbb{Q}}(t, \omega) := \gamma_i^{\mathbb{Q}}(t) h_i^{\mathbb{Q}}(t, \omega) \lambda_{*,i}(t, \omega) \quad \text{for all } (t, \omega) \in [0, y_i] \times \mathbb{R}_+,$$

as well as the functional

$$L(\Psi) := \prod_{i=1}^N \left(\prod_{(t, \omega)_j \in \Psi_i} \gamma_i^{\mathbb{Q}}(t_j) h_i^{\mathbb{Q}}(t_j, \omega_j) \right) \exp \left(\int_0^{y_i} \int_0^\infty (\lambda_{*,i}(t, \omega) - \lambda_i^{\mathbb{Q}}(t, \omega)) d\omega dt \right).$$

By Theorem T10.b [5, Chapter VIII], whenever $\mathbb{E}_{\Psi \sim \mathbb{P}_{\Psi,*}}[L(\Psi)] = 1$, the measure $\mathbb{Q}_{\Psi}(\Psi)$ defined by $\frac{d\mathbb{Q}_{\Psi}}{d\mathbb{P}_{\Psi,*}}(\Psi) = L(\Psi)$ is exactly the law under which each Ψ_i is a marked Poisson process on $[0, y_i] \times \mathbb{R}_+$ with intensity $\lambda_i^{\mathbb{Q}}(t, \omega)$. The above result underpins the analysis in Appendix G.2, where we show that the optimal variational measure \mathbb{Q}_{Ψ} coincides with the law of a collection of independent marked Poisson processes.

Finally, the measure $\mathbb{P}_{\Psi|\phi}$ also admits a Radon–Nykodim derivative with respect to $\mathbb{P}_{\Psi,*}$ which is given by :

$$\frac{d\mathbb{P}_{\Psi|\phi}}{d\mathbb{P}_{\Psi,*}}(\Psi) = \prod_{i=1}^N \left(\prod_{(t, \omega)_j \in \Psi_i} \frac{\phi}{Z(t_j, \mathbf{x}_i)} \right) \exp \left(\int_0^{y_i} \int_0^\infty (\lambda_{*,i}(t, \omega) - \lambda_i(t, \omega; \phi)) d\omega dt \right). \quad (28)$$

Notice that $\frac{\phi}{Z(t_j, \mathbf{x}_i)} = \frac{\lambda_i(t_j, \omega_j; \phi)}{\lambda_{*,i}(t_j, \omega_j)}$, i.e. the ratio of the intensities of $\mathbb{P}_{\Psi|\phi}$ and $\mathbb{P}_{\Psi,*}$.

F Obtaining the maximum a posteriori θ_{MAP}

We seek the maximum a posteriori (MAP) estimates

$$(\theta_{\text{MAP}}, \phi_{\text{MAP}}) = \arg \max_{\theta, \phi} \log p(\theta, \phi \mid \mathcal{D}, \mathbf{X}).$$

Applying Bayes' rule gives the following expression for the posterior density

$$\log p(\theta, \phi \mid \mathcal{D}, \mathbf{X}) \propto \log p(\mathcal{D} \mid \mathbf{X}, g(\cdot; \theta), \phi) + \log p_{\theta}(\theta) + \log p_{\phi}(\phi),$$

where the likelihood density $p(\mathcal{D} \mid \mathbf{X}, g(\cdot; \theta), \phi)$ and the prior densities $p_{\theta}(\theta)$ and $p_{\phi}(\phi)$ are specified in Equations (3), (4), and (6), respectively. Since the log likelihood distribution is intractable, direct optimization of the posterior distribution is infeasible.

F.1 Approximating the Log Likelihood distribution

Variational Mean-Field Approximation. Our aim is to approximate the log likelihood density $\log p(\mathcal{D} \mid \mathbf{X}, g(\cdot; \theta), \phi)$. In order to do so, we introduce a variational distribution $\check{\mathbb{Q}}(\omega, \Psi \mid \theta, \phi)$ to approximate the true distribution $\mathbb{P}(\omega, \Psi \mid \mathcal{D}, \mathbf{X}, g(\cdot; \theta), \phi)$. Such variational distribution differs from the one used for full-model inference in Section 4.3 because it is conditioned on the values of θ and ϕ . Hence, we adopt the notation $\check{\mathbb{Q}}$ (instead of \mathbb{Q}) to highlight this difference. We restrict our search to distributions that satisfy the following *mean-field* factorization:

$$\check{\mathbb{Q}}(\omega, \Psi \mid \theta, \phi) = \check{\mathbb{Q}}_{\omega \mid \theta, \phi}(\omega \mid \theta, \phi) \times \check{\mathbb{Q}}_{\Psi \mid \theta, \phi}(\Psi \mid \theta, \phi).$$

Here, we take $\check{\mathbb{Q}}_{\omega \mid \theta, \phi}(\omega \mid \theta, \phi)$ to admit the density $\check{q}_{\omega \mid \theta, \phi}(\omega \mid \theta, \phi)$ with respect to the Lebesgue measure $d\omega$.

For the marked point process component, we assume that the variational law $\check{\mathbb{Q}}_{\Psi \mid \theta, \phi}$ is absolutely continuous with respect to $\mathbb{P}_{\Psi, *}$, so that it admits a strictly positive Radon-Nikodym derivative $\frac{d\check{\mathbb{Q}}_{\Psi \mid \theta, \phi}}{d\mathbb{P}_{\Psi, *}}$ which satisfies the normalization $\mathbb{E}_{\Psi \sim \mathbb{P}_{\Psi, *}} \left[\frac{d\check{\mathbb{Q}}_{\Psi \mid \theta, \phi}}{d\mathbb{P}_{\Psi, *}}(\Psi) \right] = 1$. These two conditions guarantee that $\check{\mathbb{Q}}_{\Psi \mid \theta, \phi}$ is indeed a probability measure on the space of marked point-process paths.

We decompose the log-likelihood as follows:

$$\begin{aligned} \log p(\mathcal{D} \mid \mathbf{X}, g(\cdot; \theta), \phi) &= \\ D_{\text{KL}} \left(\check{\mathbb{Q}}_{\omega, \Psi \mid \theta, \phi}(\omega, \Psi \mid \theta, \phi) \parallel \mathbb{P}(\omega, \Psi \mid \mathcal{D}, \mathbf{X}, g(\cdot; \theta), \phi) \right) &+ \check{\mathcal{L}}_{\text{ELBO}}, \end{aligned} \quad (29)$$

where the ELBO is given by:

$$\check{\mathcal{L}}_{\text{ELBO}} := \mathbb{E}_{\omega \sim \check{q}_{\omega \mid \theta, \phi}, \Psi \sim \check{\mathbb{Q}}_{\Psi \mid \theta, \phi}} \left[\log \frac{p(\mathcal{D} \mid \mathbf{X}, g(\cdot; \theta), \phi, \omega, \Psi) p_{\omega}(\omega) \frac{d\mathbb{P}_{\Psi \mid \phi}}{d\mathbb{P}_{\Psi, *}}(\Psi)}{\check{q}_{\omega \mid \theta, \phi}(\omega \mid \theta, \phi) \frac{d\check{\mathbb{Q}}_{\Psi \mid \theta, \phi}}{d\mathbb{P}_{\Psi, *}}(\Psi \mid \theta, \phi)} \right], \quad (30)$$

and where $\frac{d\mathbb{P}_{\Psi \mid \phi}}{d\mathbb{P}_{\Psi, *}}$ is the Radon-Nykodim derivative of the true conditional law $\mathbb{P}_{\Psi \mid \phi}$ with respect to $\mathbb{P}_{\Psi, *}$, cf. (28).

Minimizing the KL Divergence. When the variational distribution $\check{\mathbb{Q}}(\omega, \Psi \mid \theta, \phi)$ matches the true posterior $\mathbb{P}(\omega, \Psi \mid \mathcal{D}, \mathbf{X}, g(\cdot; \theta), \phi)$, the KL divergence term in (29) vanishes. Consequently, the ELBO becomes equal to the marginal log-likelihood, and maximizing the ELBO is equivalent to maximizing log-likelihood directly. In practice, we minimize the KL divergence so that our ELBO provides the closest possible lower bound to the true log-likelihood. Therefore, in order to obtain the closest lower bound to the log-likelihood $\log p(\mathcal{D} \mid \mathbf{X}, g(\cdot; \theta), \phi)$ we must find the distribution $\check{\mathbb{Q}}(\omega, \Psi \mid \theta, \phi)$ which minimizes the KL divergence in (29).

Using standard mean-field variational inference techniques (see, e.g., Chapter 10.1 of [3]), the optimal distribution for the latent variables ω given $(\theta^{(\ell)}, \phi^{(\ell)})$ is obtained by computing the expectation of the joint log-density with respect to the other variational factors, that is

$$\begin{aligned} \log \check{q}_{\omega \mid \theta, \phi}(\omega) &= \\ \mathbb{E}_{\Psi \sim \check{\mathbb{Q}}_{\Psi \mid \theta^{(\ell)}, \phi^{(\ell)}}} \left[\log p(\mathcal{D} \mid \mathbf{X}, g(\cdot; \theta^{(\ell)}), \phi^{(\ell)}, \omega, \Psi) + \log p_{\omega}(\omega) + \log \frac{d\mathbb{P}_{\Psi \mid \phi^{(\ell)}}}{d\mathbb{P}_{\Psi, *}}(\Psi) \right] &+ \text{const.} \end{aligned}$$

A similar update applies for Ψ given $(\theta^{(\ell)}, \phi^{(\ell)})$,

$$\log \frac{d\check{\mathbb{Q}}_{\Psi|\theta,\phi}}{d\mathbb{P}_{\Psi,*}}(\Psi) = \mathbb{E}_{\omega \sim \check{q}_{\omega|\theta,\phi}} \left[\log p(\mathcal{D} | \mathbf{X}, g(\cdot; \theta^{(\ell)}), \phi^{(\ell)}, \omega, \Psi) + \log p_{\omega}(\omega) + \log \frac{d\mathbb{P}_{\Psi|\phi^{(\ell)}}}{d\mathbb{P}_{\Psi,*}}(\Psi) \right] + \text{const.}$$

Following the same derivation as in Appendix G.1, we find the optimal variational distribution of ω given $(\theta, \phi)^{(\ell)}$:

$$\check{q}_{\omega|\theta^{(\ell)}, \phi^{(\ell)}}(\omega | \theta^{(\ell)}, \phi^{(\ell)}) = \prod_{i=1}^N \check{q}_{\omega_i|\theta^{(\ell)}, \phi^{(\ell)}}(\omega_i | \theta^{(\ell)}, \phi^{(\ell)}) = \prod_{i=1}^N p_{\text{PG}}(\omega_i | 1, \check{c}_i^{(\ell)}), \quad (31)$$

where

$$\check{c}_i^{(\ell)} = \delta_i |g(y_i, \mathbf{x}_i; \theta^{(\ell)})|. \quad (32)$$

By mirroring the derivation in Appendix G.2, one shows that the optimal measure $\check{\mathbb{Q}}_{\Psi|\theta,\phi}(\Psi | \theta, \phi)$ is exactly the law under which each Ψ_i , for $i = 1, \dots, N$, is a marked Poisson process on $[0, y_i] \times \mathbb{R}_+$ with intensity

$$\lambda_i^{\check{\mathbb{Q}}}(t, \omega | \theta^{(\ell)}, \phi^{(\ell)}) = \lambda_i^{\check{\mathbb{Q}}}(t | \theta^{(\ell)}, \phi^{(\ell)}) p_{\text{PG}}(\omega | 1, |g(t, \mathbf{x}_i; \theta^{(\ell)})|),$$

where we set

$$\lambda_i^{\check{\mathbb{Q}}}(t | \theta^{(\ell)}, \phi^{(\ell)}) := \frac{t^{\rho-1}}{Z(t, \mathbf{x})} \phi^{(\ell)} \sigma(|g(t, \mathbf{x}_i; \theta^{(\ell)})|) \exp \left(-\frac{g(t, \mathbf{x}_i; \theta^{(\ell)}) + |g(t, \mathbf{x}_i; \theta^{(\ell)})|}{2} \right). \quad (33)$$

F.2 EM Algorithm for MAP Estimation

EM Algorithm. The Expectation-Maximization (EM) algorithm provides an efficient framework to iteratively maximize the Q-function. At each iteration $\ell = 0, 1, 2, \dots$, we perform the following three steps:

1. *Latent Variables Update.* Given the current estimates $(\theta, \phi)^{(\ell)}$, update $\check{q}_{\omega|\theta^{(\ell)}, \phi^{(\ell)}}(\omega | \theta^{(\ell)}, \phi^{(\ell)})$ and $\check{\mathbb{Q}}_{\Psi|\theta^{(\ell)}, \phi^{(\ell)}}(\Psi | \theta^{(\ell)}, \phi^{(\ell)})$ according to (31) and (33), respectively.
2. *E-Step.* Given current estimates $(\theta, \phi)^{(\ell)}$, compute the Q-function:

$$Q((\theta, \phi) | (\theta, \phi)^{(\ell)}) = \mathbb{E}_{\omega \sim \check{q}_{\omega|\theta^{(\ell)}, \phi^{(\ell)}}, \Psi \sim \check{\mathbb{Q}}_{\Psi|\theta^{(\ell)}, \phi^{(\ell)}}} \left[\log \left(p(\mathcal{D} | \mathbf{X}, g(\cdot; \theta), \phi, \omega, \Psi) p_{\omega}(\omega) \frac{d\mathbb{P}_{\Psi|\phi}}{d\mathbb{P}_{\Psi,*}}(\Psi) \right) \right] + \log p_{\theta}(\theta) + \log p_{\phi}(\phi). \quad (34)$$

Note that the entropy term of the ELBO (i.e., the denominator) is not included as it does not depend on the parameters (θ, ϕ) but on the current estimates $(\theta, \phi)^{(\ell)}$, hence it is irrelevant to the parameters' optimization.

3. *M-Step.* Update the parameters by maximizing the Q-function:

$$(\theta, \phi)^{(\ell+1)} = \arg \max_{\theta, \phi} Q((\theta, \phi) | (\theta, \phi)^{(\ell)}).$$

Steps 1-3 are repeated until a given convergence criterion is met. We provide an algorithmic description of our EM algorithm in Algorithm 1.

Algorithm 1 Expectation-Maximization (EM) for maximum a posteriori (MAP) Estimation

- 1: Initialize: Set initial value for $(\boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)})$.
 - 2: **Set:** iteration counter $\ell \leftarrow 0$
 - 3: **repeat**
 - 4: $\ell \leftarrow \ell + 1$
 - 5: **Latent Variables Update:**
 - 6: **Update** $\check{q}_{\omega}^{(\ell)}$:
 - 7: Update: $\left\{ \check{c}_i^{(\ell)} \right\}_{i=1}^N$ given $\boldsymbol{\theta}^{(\ell)}$ following (32).
 - 8: **Update** $\check{Q}_{\Psi}^{(\ell)}$:
 - 9: Update: $\left\{ \lambda_i^{\check{Q},(\ell)}(\cdot) \right\}_{i=1}^N$ given $\boldsymbol{\theta}^{(\ell)}$ and $\phi^{(\ell)}$ following (40).
 - 10: **E-step:** Evaluate the Q-function $Q((\boldsymbol{\theta}, \phi) \mid (\boldsymbol{\theta}, \phi)^{(\ell)})$ given $\left\{ \check{c}_i^{(\ell)}, \lambda_i^{\check{Q},(\ell)}(\cdot) \right\}_{i=1}^N$, $\boldsymbol{\theta}^{(\ell)}$ and $\phi^{(\ell)}$ following (34)
 - 11: **M-step:** Update parameters by

$$(\boldsymbol{\theta}, \phi)^{(\ell+1)} = \arg \max_{\boldsymbol{\theta}, \phi} Q((\boldsymbol{\theta}, \phi) \mid (\boldsymbol{\theta}, \phi)^{(\ell)})$$
 - 12: **until** Convergence criterion is met
 - 13: **return** $(\boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)})$
-

Computing the Q-function. The optimal distributions which minimize the KL divergence can now be plugged in the ELBO of (30) to obtain the closest lower bound to the log-likelihood. We now recast the MAP optimization problem in term of this lower bound. Specifically, define the following Q-function

$$\begin{aligned} Q((\boldsymbol{\theta}, \phi) | (\boldsymbol{\theta}, \phi)^{(\ell)}) &= \mathbb{E}_{\boldsymbol{\omega} \sim \check{q}_{\boldsymbol{\omega} | \boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)}}, \boldsymbol{\Psi} \sim \check{Q}_{\boldsymbol{\Psi} | \boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)}}} [\log p(\mathcal{D} | \mathbf{X}, g(\cdot; \boldsymbol{\theta}), \phi, \boldsymbol{\omega}, \boldsymbol{\Psi})] \\ &\quad + \mathbb{E}_{\boldsymbol{\omega} \sim \check{q}_{\boldsymbol{\omega} | \boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)}}, \boldsymbol{\Psi} \sim \check{Q}_{\boldsymbol{\Psi} | \boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)}}} \left[\log(p_{\boldsymbol{\omega}}(\boldsymbol{\omega}) + \log \left(\frac{d\mathbb{P}_{\boldsymbol{\Psi} | \phi}}{d\mathbb{P}_{\boldsymbol{\Psi}, *}}(\boldsymbol{\Psi}) \right) \right] \\ &\quad + \log p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \log p_{\phi}(\phi). \end{aligned}$$

We now wish to derive a closed-form expression for the Q-function which can be used in the MAP optimization. Specifically, using the augmented likelihood factorization in (14), we obtain

$$\begin{aligned} Q((\boldsymbol{\theta}, \phi) | (\boldsymbol{\theta}, \phi)^{(\ell)}) &= \sum_{i=1}^N \mathbb{E}_{\omega_i \sim \check{q}_{\omega_i | \boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)}}, \Psi_i \sim \check{Q}_{\Psi_i | \boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)}}} [\log p(\mathcal{D} | \mathbf{X}, g(\cdot; \boldsymbol{\theta}), \phi, \omega_i, \Psi_i)] \\ &\quad + \mathbb{E}_{\boldsymbol{\omega} \sim \check{q}_{\boldsymbol{\omega} | \boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)}}} [\log p_{\boldsymbol{\omega}}(\boldsymbol{\omega})] + \mathbb{E}_{\boldsymbol{\Psi} \sim \check{Q}_{\boldsymbol{\Psi} | \boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)}}} \left[\log \frac{d\mathbb{P}_{\boldsymbol{\Psi} | \phi}}{d\mathbb{P}_{\boldsymbol{\Psi}, *}}(\boldsymbol{\Psi}) \right] + \log p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \log p_{\phi}(\phi) + \text{const.} \end{aligned}$$

Next, by substituting the expression for the augmented likelihood in (13), for the priors $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ in (4) and $p_{\phi}(\phi)$ in (5) and for the Radon–Nikodym derivative of $\mathbb{P}_{\boldsymbol{\Psi} | \phi}$ with respect to $\mathbb{P}_{\boldsymbol{\Psi}, *}$ from (28), we obtain

$$\begin{aligned} Q((\boldsymbol{\theta}, \phi) | (\boldsymbol{\theta}, \phi)^{(\ell)}) &= \sum_{i=1}^N \left(\delta_i \left(\log \phi + \frac{g(y_i, \mathbf{x}_i; \boldsymbol{\theta})}{2} - \frac{g(y_i, \mathbf{x}_i; \boldsymbol{\theta})^2}{2} \mathbb{E}_{\omega_i \sim \check{q}_{\omega_i | \boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)}}} [\omega_i] \right) \right. \\ &\quad + \mathbb{E}_{\Psi_i \sim \check{Q}_{\Psi_i | \boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)}}} \left[\sum_{(t, \omega)_j \in \Psi_i} f(\omega_j, -g(t_j, \mathbf{x}_i; \boldsymbol{\theta})) \right] - \int_0^{y_i} \lambda_0(y_i, \mathbf{x}_i; \phi) dt \\ &\quad + \mathbb{E}_{\Psi_i \sim \check{Q}_{\Psi_i | \boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)}}} \left[\sum_{(t, \omega)_j \in \Psi_i} \log \left(\frac{\phi}{Z(t_j, \mathbf{x}_i)} \right) \right] \Bigg) \\ &\quad - \frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} + \log(\phi)(\alpha_0 - 1) - \phi \beta_0 + \text{const.} \end{aligned}$$

We apply Campbell's theorem (see Theorem C.3), we substitute the expression for the baseline hazard $\lambda_i(\cdot; \phi)$ from (11) and we substitute the expectation using the optimal variational distribution of ω_i from (31), to obtain

$$\begin{aligned} Q((\boldsymbol{\theta}, \phi) | (\boldsymbol{\theta}, \phi)^{(\ell)}) &= \sum_{i=1}^N \left[\delta_i \left(\frac{g(y_i, \mathbf{x}_i; \boldsymbol{\theta})}{2} - \frac{g(y_i, \mathbf{x}_i; \boldsymbol{\theta})^2}{4\check{c}_i^{(\ell)}} \tanh \left(\frac{\check{c}_i^{(\ell)}}{2} \right) \right) \right. \\ &\quad - \frac{1}{2} \int_0^{y_i} g(t, \mathbf{x}_i; \boldsymbol{\theta}) \lambda_i^{\check{Q}}(t | \boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)}) dt \\ &\quad - \frac{1}{4} \int_0^{y_i} \frac{g(t, \mathbf{x}_i; \boldsymbol{\theta})^2}{|g(t, \mathbf{x}_i; \boldsymbol{\theta}^{(\ell)})|} \tanh \left(\frac{|g(t, \mathbf{x}_i; \boldsymbol{\theta}^{(\ell)})|}{2} \right) \lambda_i^{\check{Q}}(t | \boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)}) dt \Bigg] \\ &\quad - \frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} + \log(\phi) \left(\alpha_0 + \sum_{i=1}^N \left(\delta_i + \int_0^{y_i} \lambda_i^{\check{Q}}(t | \boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)}) dt \right) - 1 \right) \\ &\quad - \phi \left(\beta_0 + \sum_{i=1}^N \int_0^{y_i} \frac{t^{\rho-1}}{Z(t, \mathbf{x}_i)} dt \right) + \text{const.}, \end{aligned}$$

where $\lambda_i^{\check{Q}}(t | \boldsymbol{\theta}^{(\ell)}, \phi^{(\ell)})$ is shown in (33).

G Coordinate Ascent Variational Inference Optimal Updates

In this Appendix we present a heuristic derivation of the CAVI optimal updates presented in Section 4.3. Before presenting the next results, we define here for convenience

$$\tilde{m}_i^{(k)}(t) := \mathbb{E}_{\boldsymbol{\theta} \sim q_{\boldsymbol{\theta}}^{(k)}} [g^{\text{lin}}(t, \mathbf{x}_i; \boldsymbol{\theta})], \quad \tilde{s}_i^{(k)}(t) := \sqrt{\mathbb{E}_{\boldsymbol{\theta} \sim q_{\boldsymbol{\theta}}^{(k)}} [g^{\text{lin}}(t, \mathbf{x}_i; \boldsymbol{\theta})^2]}$$

for $k \geq 0$.

G.1 Optimal Update for ω

Using standard mean-field variational inference techniques (see, e.g., Chapter 10.1 of [3]), the optimal update for the latent variables ω is obtained by computing the expectation of the joint log-density with respect to the other variational factors. In particular, we have

$$\log q_{\omega}^{(k)}(\omega) = \mathbb{E}_{\phi \sim q_{\phi}^{(k-1)}, \boldsymbol{\theta} \sim q_{\boldsymbol{\theta}}^{(k-1)}, \Psi \sim \mathbb{Q}_{\Psi}^{(k-1)}} \left[\log p(\mathcal{D} \mid \phi, g^{\text{lin}}(\cdot; \boldsymbol{\theta}), \omega, \Psi) \right] + \log p_{\omega}(\omega) + \text{const.}$$

Using the augmented likelihood factorization in (14), the expression decomposes as

$$\begin{aligned} \log q_{\omega}^{(k)}(\omega) &= \sum_{i=1}^N \mathbb{E}_{\phi \sim q_{\phi}^{(k-1)}, \boldsymbol{\theta} \sim q_{\boldsymbol{\theta}}^{(k-1)}, \Psi_i \sim \mathbb{Q}_{\Psi_i}^{(k-1)}} \left[\log p(y_i, \delta_i \mid \mathbf{x}_i, \phi, g^{\text{lin}}(\cdot; \boldsymbol{\theta}), \omega_i, \Psi_i) \right] \\ &\quad + \log p_{\omega}(\omega) + \text{const.} \end{aligned}$$

Next, by substituting the expression for the prior $p_{\omega}(\omega)$ from (10) and the augmented likelihood from (13), we obtain

$$\log q_{\omega}^{(k)}(\omega) = \sum_{i=1}^N \left(-\frac{\omega_i \delta_i}{2} \left(\tilde{s}_i^{(k-1)}(y_i) \right)^2 + \log p_{\text{PG}}(\omega_i \mid 1, 0) \right) + \text{const.}$$

Finally, by applying the identity in (20), we deduce that the optimal variational distribution factorizes as

$$q_{\omega}^{(k)}(\omega) = \prod_{i=1}^N q_{\omega_i}^{(k)}(\omega_i) = \prod_{i=1}^N p_{\text{PG}}(\omega_i \mid 1, \tilde{c}_i^{(k)}),$$

where

$$\tilde{c}_i^{(k)} = \delta_i \tilde{s}_i^{(k-1)}(y_i) \quad (35)$$

Optimal Variational Expectations for ω . From Proposition B.4, we obtain the required expectation for updating the other variational factors with

$$\mathbb{E}_{\omega_i \sim q_{\omega_i}^{(k)}} [\omega_i] = \frac{1}{2\tilde{c}_i^{(k)}} \tanh \left(\frac{\tilde{c}_i^{(k)}}{2} \right) \quad (36)$$

for $i = 1, \dots, N$. Notably, since this expectation is always multiplied by δ_i when updating other variational factors, it remains well-defined in all cases.

G.2 Optimal Update for Ψ

Using standard mean-field variational inference techniques (see, e.g., Chapter 10.1 of [3]), we obtain the optimal Radon-Nykodim derivative $\frac{d\mathbb{Q}_{\Psi}}{d\mathbb{P}_{\Psi,*}}$ by taking the expectation of the joint log-density with respect to the other variational factors. In particular, we have

$$\begin{aligned} \log \frac{d\mathbb{Q}_{\Psi}^{(k)}}{d\mathbb{P}_{\Psi,*}}(\Psi) &= \mathbb{E}_{\phi \sim q_{\phi}^{(k-1)}, \boldsymbol{\theta} \sim q_{\boldsymbol{\theta}}^{(k-1)}, \omega \sim q_{\omega}^{(k)}} \left[\log p(\mathcal{D} \mid \phi, g^{\text{lin}}(\cdot; \boldsymbol{\theta}), \omega, \Psi) \right] \\ &\quad + \mathbb{E}_{\phi \sim q_{\phi}^{(k-1)}} \left[\log \frac{d\mathbb{P}_{\Psi|\phi}}{d\mathbb{P}_{\Psi,*}}(\Psi) \right] + \text{const.}, \quad (37) \end{aligned}$$

where the constant term absorbs all terms irrelevant to the optimisation. Using the augmented likelihood factorization in (14), the expression in (37) decomposes as

$$\begin{aligned} \log \frac{d\mathbb{Q}_{\Psi}^{(k)}}{d\mathbb{P}_{\Psi,*}}(\Psi) &= \sum_{i=1}^N \mathbb{E}_{\phi \sim q_{\phi}^{(k-1)}, \theta \sim q_{\theta}^{(k-1)}, \omega_i \sim q_{\omega_i}^{(k)}} [\log p(y_i, \delta_i | \phi, g^{\text{lin}}(\cdot; \theta), \omega_i, \Psi_i)] \\ &\quad + \mathbb{E}_{\phi \sim q_{\phi}^{(k-1)}} \left[\log \frac{d\mathbb{P}_{\Psi|\phi}}{d\mathbb{P}_{\Psi,*}}(\Psi) \right] + \text{const.} \end{aligned}$$

Next, by substituting the augmented likelihood from (13) and the Radon–Nikodym derivative of $\mathbb{P}_{\Psi|\phi}$ with respect to $\mathbb{P}_{\Psi,*}$ from (28), we arrive at the unnormalised form

$$\begin{aligned} \log \frac{d\mathbb{Q}_{\Psi}^{(k)}}{d\mathbb{P}_{\Psi,*}}(\Psi) &= \sum_{i=1}^N \sum_{(t, \omega)_j \in \Psi_i} \mathbb{E}_{\theta \sim q_{\theta}^{(k-1)}} [f(\omega_j, -g^{\text{lin}}(t_j, \mathbf{x}_i; \theta))] \\ &\quad + \sum_{i=1}^N \sum_{(t, \omega)_j \in \Psi_i} \mathbb{E}_{\phi \sim q_{\phi}^{(k-1)}} \left[\log \left(\frac{\phi}{Z(t_j, \mathbf{x}_i)} \right) \right] + \text{const.} \quad (38) \end{aligned}$$

Plugging in the definition of $f(\cdot, \cdot)$ from (8) simplifies (38) to

$$\begin{aligned} \log \frac{d\mathbb{Q}_{\Psi}^{(k)}}{d\mathbb{P}_{\Psi,*}}(\Psi) &= - \sum_{i=1}^N \sum_{(t, \omega)_j \in \Psi_i} \left[\frac{\tilde{m}_i^{(k)}(t_j)}{2} + \frac{(\tilde{s}_i^{(k)}(t_j))^2}{2} \omega_j + \log(2) \right] \\ &\quad + \sum_{i=1}^N \sum_{(t, \omega)_j \in \Psi_i} \left[\mathbb{E}_{\phi \sim q_{\phi}^{(k-1)}} [\log \phi] - \log Z(t_j, \mathbf{x}_i) \right] + \text{const.} \end{aligned}$$

To express this in closed form, define for each $i = 1, \dots, N$ and $(t, \omega) \in [0, y_i] \times \mathbb{R}_+$ the functions

$$\begin{aligned} h_i^{\mathbb{Q},(k)}(t, \omega) &:= \exp \left(- \frac{(\tilde{s}_i^{(k-1)}(t))^2}{2} \omega \right) \cosh \left(\frac{\tilde{s}_i^{(k-1)}(t)}{2} \right), \\ \gamma_i^{\mathbb{Q},(k)}(t) &:= \frac{1}{Z(t, \mathbf{x}_i)} \sigma(\tilde{s}_i^{(k-1)}(t)) \exp \left(- \frac{\tilde{m}_i^{(k-1)}(t) + \tilde{s}_i^{(k-1)}(t)}{2} + \mathbb{E}_{\phi \sim q_{\phi}^{(k-1)}} [\log \phi] \right), \\ \lambda_i^{\mathbb{Q},(k)}(t, \omega) &:= \gamma_i^{\mathbb{Q},(k)}(t) h_i^{\mathbb{Q},(k)}(t, \omega) \lambda_{*,i}(t, \omega), \end{aligned}$$

where $\lambda_{*,i}(t, \omega)$ is the intensity defined in (26). Furthermore, we define for convenience,

$$\lambda_i^{\mathbb{Q},(k)}(t) := t^{\rho-1} \gamma_i^{\mathbb{Q},(k)}(t). \quad (39)$$

Notice that by using expression (20), the function $\lambda_i^{\mathbb{Q},(k)}(t, \omega)$ can be written as

$$\lambda_i^{\mathbb{Q},(k)}(t, \omega) = \lambda_i^{\mathbb{Q},(k)}(t) p_{\text{PG}} \left(\omega \mid 1, \tilde{s}_i^{(k-1)}(t) \right). \quad (40)$$

Finally, enforcing the normalisation condition

$$\mathbb{E}_{\Psi \sim \mathbb{P}_{\Psi,*}} \left[\frac{d\mathbb{Q}_{\Psi}^{(k)}}{d\mathbb{P}_{\Psi,*}}(\Psi) \right] = 1$$

together with Campbell's theorem (Theorem C.3) yields the normalized derivative

$$\begin{aligned} \frac{d\mathbb{Q}_{\Psi}^{(k)}}{d\mathbb{P}_{\Psi,*}}(\Psi) &= \\ &\prod_{i=1}^N \left(\prod_{(t, \omega)_j \in \Psi_i} \gamma_i^{\mathbb{Q},(k)}(t_j) h_i^{\mathbb{Q},(k)}(t_j, \omega_j) \right) \exp \left(\int_0^{y_i} \int_0^{\infty} (\lambda_{*,i}(t, \omega) - \lambda_i^{\mathbb{Q},(k)}(t, \omega)) d\omega dt \right). \end{aligned}$$

Notice that the products $\gamma_i^{\mathbb{Q},(k)}(t_j)h_i^{\mathbb{Q},(k)}(t_j, \omega_j)$ are all strictly positive ², hence $\frac{d\mathbb{Q}_{\Psi,*}^{(k)}}{d\mathbb{P}_{\Psi,*}}$ is also strictly positive. Under suitable regularity conditions on g , one can show that $h_i^{\mathbb{Q},(k)}(t, \omega)$ and $\gamma_i^{\mathbb{Q},(k)}(t)$ satisfy the integrability criteria of (27), so that $\mathbb{Q}_{\Psi}^{(k)}$ is the probability measure under which each Ψ_i ($i = 1, \dots, N$) is a marked Poisson Process on $[0, y_i] \times \mathbb{R}_+$ with intensity function $\lambda_i^{\mathbb{Q},(k)}(t, \omega)$.

Optimal Variational Expectations for Ψ . From Proposition B.4, we obtain the required integrals for updating the other variational factors

$$\begin{aligned} \int_{\mathbb{R}_+} \lambda_i^{\mathbb{Q},(k)}(t, \omega) d\omega &= \lambda_i^{\mathbb{Q},(k)}(t), \\ \int_{\mathbb{R}_+} \lambda_i^{\mathbb{Q},(k)}(t, \omega) \omega d\omega &= \lambda_i^{\mathbb{Q},(k)}(t) \frac{1}{2\tilde{s}_i^{(k-1)}(t)} \tanh\left(\frac{\tilde{s}_i^{(k-1)}(t)}{2}\right). \end{aligned}$$

G.3 Optimal Update for ϕ

Using standard mean-field variational inference techniques (see, e.g., Chapter 10.1 of [3]), the optimal variational factor for the parameter ϕ is obtained by computing the expectation of the joint log-density with respect to the other variational factors. In particular, we have

$$\begin{aligned} \log q_{\phi}^{(k)}(\phi) &= \mathbb{E}_{\boldsymbol{\theta} \sim q_{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\omega} \sim q_{\boldsymbol{\omega}}^{(k)}, \boldsymbol{\Psi} \sim \mathbb{Q}_{\boldsymbol{\Psi}}^{(k)}} \left[\log p(\mathcal{D} \mid \phi, g^{\text{lin}}(\cdot; \boldsymbol{\theta}), \boldsymbol{\omega}, \boldsymbol{\Psi}) + \log \frac{d\mathbb{P}_{\boldsymbol{\Psi}|\phi}}{d\mathbb{P}_{\boldsymbol{\Psi},*}}(\boldsymbol{\Psi}) \right] \\ &\quad + \log p_{\phi}(\phi) + \text{const.} \end{aligned}$$

Using the augmented likelihood factorization in (14), the expression decomposes as

$$\begin{aligned} \log q_{\phi}^{(k)}(\phi) &= \sum_{i=1}^N \mathbb{E}_{\boldsymbol{\theta} \sim q_{\boldsymbol{\theta}}^{(k-1)}, \omega_i \sim q_{\omega_i}^{(k)}, \Psi_i \sim \mathbb{Q}_{\Psi_i}^{(k)}} [\log p(y_i, \delta_i \mid \mathbf{x}_i, \phi, \boldsymbol{\theta}, \omega_i, \Psi_i)] \\ &\quad + \mathbb{E}_{\boldsymbol{\Psi} \sim \mathbb{Q}_{\boldsymbol{\Psi}}^{(k)}} \left[\log \frac{d\mathbb{P}_{\boldsymbol{\Psi}|\phi}}{d\mathbb{P}_{\boldsymbol{\Psi},*}}(\boldsymbol{\Psi}) \right] + \log p_{\phi}(\phi) + \text{const.} \end{aligned}$$

Next, by substituting the expression for the augmented likelihood from (13), the Radon–Nikodym derivative of $\mathbb{P}_{\boldsymbol{\Psi}|\phi}$ with respect to $\mathbb{P}_{\boldsymbol{\Psi},*}$ from (28), and the prior of ϕ from (5), we obtain,

$$\begin{aligned} \log q_{\phi}^{(k)}(\phi) &= \sum_{i=1}^N \left(\delta_i \log \lambda_0(y_i, \mathbf{x}_i; \phi) - \int_0^{y_i} \lambda_0(t, \mathbf{x}_i; \phi) dt \right. \\ &\quad \left. + \mathbb{E}_{\Psi_i \sim \mathbb{Q}_{\Psi_i}^{(k)}} \left[\sum_{(t, \omega)_j \in \Psi_i} \log \left(\frac{\phi}{Z(t_j, \mathbf{x}_i)} \right) \right] \right) + (\alpha_0 - 1) \log(\phi) - \beta_0 \phi + \text{const.} \end{aligned}$$

We apply Campbell’s Theorem (Theorem C.3) and substitute the expression for the baseline hazard $\lambda_0(\cdot)$ from (5), to obtain

$$\begin{aligned} &\log q_{\phi}^{(k)}(\phi) \\ &= \log(\phi) \left(\alpha_0 + \sum_{i=1}^N \left(\delta_i + \int_0^{y_i} \lambda_i^{\mathbb{Q},(k)}(t) dt \right) - 1 \right) - \phi \left(\beta_0 + \sum_{i=1}^N \int_0^{y_i} \frac{t^{\rho-1}}{Z(t, \mathbf{x}_i)} dt \right) + \text{const.}, \end{aligned}$$

where $\lambda_i^{\mathbb{Q},(k)}(t)$ is shown in (39). We deduce that

$$q_{\phi}^{(k)}(\phi) = \text{Gamma}(\tilde{\alpha}^{(k)}, \tilde{\beta}),$$

where with shape $\tilde{\alpha}^{(k)}$ and rate $\tilde{\beta}$ given by

$$\tilde{\alpha}^{(k)} = \alpha_0 + \sum_{i=1}^N \left(\delta_i + \int_0^{y_i} \lambda_i^{\mathbb{Q},(k)}(t) dt \right), \quad \tilde{\beta} = \beta_0 + \sum_{i=1}^N \int_0^{y_i} \frac{t^{\rho-1}}{Z(t, \mathbf{x}_i)} dt. \quad (41)$$

²See Lemma N.1 for a proof of the strict positivity of the normalizing constant $Z(t, \mathbf{x}_i)$.

Optimal Variational Expectation for ϕ . We obtain the required expectation for updating the other variational factors with

$$\mathbb{E}_{\phi \sim q_{\phi}^{(k)}} [\log \phi] = \psi(\tilde{\alpha}^{(k)}) - \log(\tilde{\beta}), \quad (42)$$

where $\psi(\cdot)$ is the digamma function.

G.4 Optimal Update for θ

Using standard mean-field variational inference techniques (see, e.g., Chapter 10.1 of [3]), the optimal variational factor for the parameters θ is obtained by computing the expectation of the joint log-density with respect to the other variational factors. In particular, we have

$$\log q_{\theta}^{(k)}(\theta) = \mathbb{E}_{\phi \sim q_{\phi}^{(k)}, \omega \sim q_{\omega}^{(k)}, \Psi \sim \mathbb{Q}_{\Psi}^{(k)}} [\log p(\mathcal{D} \mid \phi, g^{\text{lin}}(\cdot; \theta), \omega, \Psi)] + \log p_{\theta}(\theta) + \text{const.}$$

Using the augmented likelihood factorization in (14), we obtain

$$\begin{aligned} \log q_{\theta}^{(k)}(\theta) &= \sum_{i=1}^N \mathbb{E}_{\phi \sim q_{\phi}^{(k)}, \omega_i \sim q_{\omega_i}^{(k)}, \Psi_i \sim \mathbb{Q}_{\Psi_i}^{(k)}} [\log p(y_i, \delta_i \mid \mathbf{x}_i, \phi, g^{\text{lin}}(\cdot; \theta), \omega_i, \Psi_i)] \\ &\quad + \log p_{\theta}(\theta) + \text{const.} \end{aligned}$$

Next, by substituting the expression for the augmented likelihood (13) and for the prior for θ from (4), we obtain,

$$\begin{aligned} \log q_{\theta}^{(k)}(\theta) &= \sum_{i=1}^N \left(\frac{\delta_i}{2} \left(g^{\text{lin}}(y_i, \mathbf{x}_i; \theta) - \mathbb{E}_{\omega_i \sim q_{\omega_i}^{(k)}} [\omega_i] g^{\text{lin}}(y_i, \mathbf{x}_i; \theta)^2 \right) \right. \\ &\quad \left. + \mathbb{E}_{\Psi_i \sim \mathbb{Q}_{\Psi_i}^{(k)}} \left[\sum_{(t, \omega)_j \in \Psi_i} f(\omega_j, -g^{\text{lin}}(t_j, \mathbf{x}_i; \theta)) \right] \right) - \frac{1}{2} \theta^{\top} \theta + \text{const.} \end{aligned}$$

We apply Campbell's Theorem (Theorem C.3) to obtain,

$$\begin{aligned} \log q_{\theta}^{(k)}(\theta) &= \sum_{i=1}^N \left(\frac{\delta_i}{2} \left(g^{\text{lin}}(y_i, \mathbf{x}_i; \theta) - \mathbb{E}_{\omega_i \sim q_{\omega_i}^{(k)}} [\omega_i] g^{\text{lin}}(y_i, \mathbf{x}_i; \theta)^2 \right) \right. \\ &\quad \left. + \frac{1}{2} \int_{\mathcal{Z}_i} (-g^{\text{lin}}(t, \mathbf{x}_i; \theta) - g^{\text{lin}}(t, \mathbf{x}_i; \theta)^2 \omega) \lambda_i^{\mathbb{Q}, (k)}(t, \omega) dt d\omega \right) - \frac{1}{2} \theta^{\top} \theta + \text{const.}, \end{aligned}$$

where $\lambda_i^{\mathbb{Q}, (k)}(t, \omega)$ is shown in Equation (40). Next, we recall the expression for $g^{\text{lin}}(\cdot; \theta)$ from (17) and we notice that

$$g^{\text{lin}}(\cdot; \theta) = \theta^{\top} \mathbf{J}_{\theta_{\text{MAP}}}(\cdot) + \text{const.}$$

$$g^{\text{lin}}(\cdot; \theta)^2 = \theta^{\top} \mathbf{J}_{\theta_{\text{MAP}}}(\cdot) (2g(\cdot; \theta_{\text{MAP}}) - 2\mathbf{J}_{\theta_{\text{MAP}}}(\cdot)^{\top} \theta_{\text{MAP}}) + \theta^{\top} \mathbf{J}_{\theta_{\text{MAP}}}(\cdot) \mathbf{J}_{\theta_{\text{MAP}}}(\cdot)^{\top} \theta + \text{const.},$$

where the constant term represents terms that do not depend on θ . We substitute the expression for $g^{\text{lin}}(\cdot; \theta)$ and $g^{\text{lin}}(\cdot; \theta)^2$ and we obtain,

$$\log q_{\theta}^{(k)}(\theta) = \theta^T \mathbf{A}^{(k)} - \theta^{\top} \mathbf{B}^{(k)} \theta + \text{const.},$$

where

$$\begin{aligned} \mathbf{A}^{(k)} &= \sum_{i=1}^N \frac{1}{2} \left(\delta_i \mathbf{J}_{\theta_{\text{MAP}}}(y_i, \mathbf{x}_i) \left(1 - 2\mathbb{E}_{\omega_i \sim q_{\omega_i}^{(k)}} [\omega_i] (g(y_i, \mathbf{x}_i; \theta_{\text{MAP}}) - \mathbf{J}_{\theta_{\text{MAP}}}(y_i, \mathbf{x}_i)^{\top} \theta_{\text{MAP}}) \right) \right. \\ &\quad \left. - \left(\mathcal{I}_{1,i}^{(k)} + 2 \left(\mathcal{I}_{2,i}^{(k)} - \mathcal{I}_{3,i}^{(k)} \theta_{\text{MAP}} \right) \right) \right) \end{aligned} \quad (43)$$

$$\mathbf{B}^{(k)} = \sum_{i=1}^N \frac{1}{2} \left(\delta_i \mathbb{E}_{\omega_i \sim q_{\omega_i}^{(k)}} [\omega_i] \mathbf{J}_{\theta_{\text{MAP}}}(y_i, \mathbf{x}_i) \mathbf{J}_{\theta_{\text{MAP}}}(y_i, \mathbf{x}_i)^{\top} + \mathcal{I}_{3,i}^{(k)} \right) + \frac{1}{2} \mathbf{I}_m \quad (44)$$

and

$$\begin{aligned}\mathcal{I}_{1,i}^{(k)} &= \int_0^{y_i} \mathbf{J}_{\boldsymbol{\theta}_{\text{MAP}}}(t, \mathbf{x}_i) \lambda_i^{\mathbb{Q},(k)}(t) dt \\ \mathcal{I}_{2,i}^{(k)} &= \int_0^{y_i} \mathbf{J}_{\boldsymbol{\theta}_{\text{MAP}}}(t, \mathbf{x}_i) g(t, \mathbf{x}_i; \boldsymbol{\theta}_{\text{MAP}}) \lambda_i^{\mathbb{Q},(k)}(t) \frac{\tanh\left(\tilde{s}_i^{(k-1)}(t)/2\right)}{2\tilde{s}_i^{(k-1)}(t)} dt \\ \mathcal{I}_{3,i}^{(k)} &= \int_0^{y_i} \mathbf{J}_{\boldsymbol{\theta}_{\text{MAP}}}(t, \mathbf{x}_i) \mathbf{J}_{\boldsymbol{\theta}_{\text{MAP}}}(t, \mathbf{x}_i)^\top \lambda_i^{\mathbb{Q},(k)}(t) \frac{\tanh\left(\tilde{s}_i^{(k-1)}(t)/2\right)}{2\tilde{s}_i^{(k-1)}(t)} dt.\end{aligned}$$

\mathbf{A} , $\mathcal{I}_{1,i}$ and $\mathcal{I}_{2,i}$ are vectors of the same length of $\boldsymbol{\theta}$. \mathbf{B} and $\mathcal{I}_{3,i}$ are square matrices for which each dimension is the length of $\boldsymbol{\theta}$, and \mathbf{I}_m is the identity matrix of length of $\boldsymbol{\theta}$. We deduce that

$$q_{\boldsymbol{\theta}}^{(k)}(\boldsymbol{\theta}) = \mathcal{N}\left(\tilde{\boldsymbol{\mu}}^{(k)}, \tilde{\boldsymbol{\Sigma}}^{(k)}\right),$$

where

$$\tilde{\boldsymbol{\mu}}^{(k)} = \frac{1}{2} \left(\mathbf{B}^{(k)}\right)^{-1} \mathbf{A}^{(k)}, \quad \tilde{\boldsymbol{\Sigma}} = \frac{1}{2} \left(\mathbf{B}^{(k)}\right)^{-1}. \quad (45)$$

Optimal Variational Expectation for $\boldsymbol{\theta}$. We obtain the required expectation for updating the other variational factors,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta} \sim q_{\boldsymbol{\theta}}^{(k)}}[g^{\text{lin}}(t, \mathbf{x}_i; \boldsymbol{\theta})] &= g(t, \mathbf{x}_i; \boldsymbol{\theta}_{\text{MAP}}) + \mathbf{J}_{\boldsymbol{\theta}_{\text{MAP}}}(t, \mathbf{x}_i)^\top \left(\tilde{\boldsymbol{\mu}}^{(k)} - \boldsymbol{\theta}_{\text{MAP}}\right), \\ \mathbb{E}_{\boldsymbol{\theta} \sim q_{\boldsymbol{\theta}}^{(k)}}[g^{\text{lin}}(t, \mathbf{x}_i; \boldsymbol{\theta})^2] &= \left(g(t, \mathbf{x}_i; \boldsymbol{\theta}_{\text{MAP}}) + \mathbf{J}_{\boldsymbol{\theta}_{\text{MAP}}}(t, \mathbf{x}_i)^\top \left(\tilde{\boldsymbol{\mu}}^{(k)} - \boldsymbol{\theta}_{\text{MAP}}\right)\right)^2 \\ &\quad + \mathbf{J}_{\boldsymbol{\theta}_{\text{MAP}}}(t, \mathbf{x}_i)^\top \tilde{\boldsymbol{\Sigma}}^{(k)} \mathbf{J}_{\boldsymbol{\theta}_{\text{MAP}}}(t, \mathbf{x}_i).\end{aligned} \quad (46)$$

H Coordinate Ascent Variational Inference Algorithm

Algorithm 2 Coordinate Ascent Variational Inference (CAVI)

- 1: **Compute:** Compute $\tilde{\beta}$ following (41).
 - 2: **Initialize::** Set initial values for $\tilde{\alpha}^{(0)}$ and $(\tilde{\mu}, \tilde{\Sigma})^{(0)}$.
 - 3: **Compute:** $\mathbb{E}_{\phi \sim q_{\phi}^{(0)}} [\log \phi]$ given $(\tilde{\alpha}^{(0)}, \tilde{\beta})$ following (42).
 - 4: **Compute:** $\{(\tilde{m}_i(\cdot), \tilde{s}_i(\cdot))^{(0)}\}_{i=1}^N$ given $(\tilde{\mu}, \tilde{\Sigma})^{(0)}$ following (46).
 - 5: **Set:** iteration counter $k \leftarrow 0$
 - 6: **repeat**
 - 7: $k \leftarrow k + 1$
 - 8: **Update** $q_{\omega}^{(k)}$:
 - 9: Update: $\{\tilde{c}_i^{(k)}\}_{i=1}^N$ given $\{\tilde{s}_i(\cdot)^{(k-1)}\}_{i=1}^N$ following (35).
 - 10: Compute: $\{\mathbb{E}_{\omega_i \sim q_{\omega_i}^{(k)}} [\omega_i]\}_{i=1}^N$ given $\{\tilde{c}_i^{(k)}\}_{i=1}^N$ following (36).
 - 11: **Update** $\mathbb{Q}_{\Psi}^{(k)}$:
 - 12: Update: $\{\lambda_i^{\mathbb{Q},(k)}(\cdot)\}_{i=1}^N$ given $(\{(\tilde{m}_i(\cdot), \tilde{s}_i(\cdot))^{(k-1)}\}_{i=1}^N, \mathbb{E}_{\phi \sim q_{\phi}^{(k-1)}} [\log \phi])$ following (40).
 - 13: **Update** $q_{\phi}^{(k)}$:
 - 14: Update: $\tilde{\alpha}^{(k)}$ given $\{\lambda_i^{\mathbb{Q},(k)}(\cdot)\}_{i=1}^N$ following (41).
 - 15: Compute: $\mathbb{E}_{\phi \sim q_{\phi}^{(k)}} [\log \phi]$ given $(\tilde{\alpha}^{(k)}, \tilde{\beta})$ following (42).
 - 16: **Update** $q_{\theta}^{(k)}$:
 - 17: Update: $(\tilde{\mu}, \tilde{\Sigma})^{(k)}$ given $\{(\mathbb{E}_{\omega_i \sim q_{\omega_i}^{(k)}} [\omega_i], \lambda_i^{\mathbb{Q},(k)}(\cdot))\}_{i=1}^N$ following (45).
 - 18: Compute: $\{(\tilde{m}_i(\cdot), \tilde{s}_i(\cdot))^{(k)}\}_{i=1}^N$ given $(\tilde{\mu}, \tilde{\Sigma})^{(k)}$ following (46).
 - 19: **until** Convergence criterion is met
 - 20: **Return:** Optimized variational distributions $q_{\theta}^{(k^*)}(\theta) = \mathcal{N}(\tilde{\mu}^{(k^*)}, \tilde{\Sigma}^{(k^*)})$ and $q_{\phi}^{k^*}(\phi) = \text{Gamma}(\tilde{\alpha}^{(k^*)}, \tilde{\beta})$, where k^* is the final iteration after convergence.
-

I Computational Speed-Ups

Survival-analysis cohorts often comprise only a few hundred to a few thousand observations, yet modern deep learning models may involve millions of parameters, putting us in the $N \ll m$ regime. To exploit this disparity, we develop two complementary strategies that avoid any expensive m -dimensional inversions or factorizations by leveraging the fact that the nontrivial part of our key matrix is low-rank relative to the full parameter dimension m . We also show how heavy censoring further reduces the computational burden.

To streamline what follows, let us introduce the shorthand

$$\mathbf{J}_i := \mathbf{J}_{\theta_{\text{MAP}}}(y_i, \mathbf{x}_i) \in \mathbb{R}^{m \times 1}$$

for $i = 1, \dots, N$. With this notation (and dropping the CAVI-iteration index for clarity), the matrix $\mathbf{B} \in \mathbb{R}^{m \times m}$ defined in (43) becomes

$$\mathbf{B} = \sum_{i=1}^N \frac{1}{2} \left(\delta_i \mathbb{E}_{\omega_i \sim q_{\omega_i}} [\omega_i] \mathbf{J}_i \mathbf{J}_i^T + \mathcal{I}_{3,i} \right) + \frac{1}{2} \mathbf{I}_m.$$

Here, each $\mathcal{I}_{3,i}$ is the integral

$$\mathcal{I}_{3,i} = \int_0^{y_i} \mathbf{J}_{\theta_{\text{MAP}}}(t, \mathbf{x}_i) \mathbf{J}_{\theta_{\text{MAP}}}(t, \mathbf{x}_i)^T \lambda_i^{\mathbb{Q}}(t) \frac{\tanh(\tilde{s}_i(t)/2)}{2\tilde{s}_i(t)} dt$$

and in general admits no closed-form solution. We therefore approximate it by any standard quadrature rule (e.g. trapezoid, Simpson's, or Gauss–Legendre). In what follows, we will illustrate the argument with the trapezoid rule, though the same steps apply to any other quadrature method.

We begin by introducing a uniform grid of points along the time axis:

$$t_1, t_2, \dots, t_K,$$

where $t_1 := 0$ and $t_K := \max\{y_i\}_{i=1}^N$. We associate a set of quadrature weights $\{v_{ik}\}_{k=1}^K$ to the time grid points, tailored for each observation i . These weights correspond to the trapezoidal rule for numerical integration on the interval $[0, y_i]$, and are defined as:

$$v_{ik} = \begin{cases} \frac{t_2 - t_1}{2}, & \text{if } k = 1 \text{ and } t_1 < y_i, \\ \frac{t_{k+1} - t_{k-1}}{2}, & \text{if } 1 < k < K_i \text{ and } t_k < y_i, \\ \frac{t_{K_i} - t_{K_i-1}}{2}, & \text{if } k = K_i, \\ 0, & k > K_i, \end{cases}$$

where $K_i = \max\{k \in \{1, \dots, K\} : t_k < y_i\}$. Further we denote by \mathbf{V}_i the collection of quadrature weights for observation i , such that

$$\mathbf{V}_i := (v_{i1}, \dots, v_{iK}) \in \mathbb{R}^K.$$

We collect the Jacobian evaluations into the matrices

$$\mathbf{Q}_i := [\mathbf{J}_{\theta_{\text{MAP}}}(t_1, \mathbf{x}_i) \quad \mathbf{J}_{\theta_{\text{MAP}}}(t_2, \mathbf{x}_i) \quad \dots \quad \mathbf{J}_{\theta_{\text{MAP}}}(t_K, \mathbf{x}_i)] \in \mathbb{R}^{m \times K}.$$

With these definitions in hand, any K -point quadrature rule yields the approximation

$$\mathcal{I}_{3,i} \approx \sum_{k=1}^K v_{ik} \mathbf{J}_{\theta_{\text{MAP}}}(t_k, \mathbf{x}_i) \mathbf{J}_{\theta_{\text{MAP}}}(t_k, \mathbf{x}_i)^T = \mathbf{Q}_i \mathbf{V}_i \mathbf{Q}_i^T.$$

Likewise, each term

$$\delta_i \mathbb{E}_{\omega_i \sim q_{\omega_i}} [\omega_i] \mathbf{J}_i \mathbf{J}_i^T$$

can be written in the form $\mathbf{J}_i \mathbf{C}_i \mathbf{J}_i^T$, where the scalar $\mathbf{C}_i = \delta_i \mathbb{E}_{\omega_i \sim q_{\omega_i}} [\omega_i]$.

We collect all contributions into a single matrix $\mathbf{U} \in \mathbb{R}^{m \times R}$, where $R = N + NK$. This matrix is constructed by horizontally concatenating the vectors \mathbf{J}_i and \mathbf{Q}_i for $i = 1, \dots, N$, as follows:

$$\mathbf{U} := \begin{bmatrix} \underbrace{\mathbf{J}_1}_{(m \times 1)}, \mathbf{J}_2, \dots, \mathbf{J}_N, \underbrace{\mathbf{Q}_1}_{(m \times K)}, \mathbf{Q}_2, \dots, \mathbf{Q}_N \end{bmatrix}.$$

Further, we define the block-diagonal weight matrix

$$\mathbf{C} := \text{diag}(\underbrace{\delta_1 \mathbb{E}_{\omega_1 \sim q_{\omega_1}}[\omega_1], \dots, \delta_N \mathbb{E}_{\omega_N \sim q_{\omega_N}}[\omega_N]}_{(N)}, \underbrace{\mathbf{V}_1, \dots, \mathbf{V}_N}_{(K)}) \in \mathbb{R}^{R \times R}.$$

It is straightforward to verify that

$$\mathbf{B} = \frac{1}{2} (\mathbf{I}_m + \mathbf{U} \mathbf{C} \mathbf{U}^\top).$$

Applying the Woodbury identity (see [16, Appendix B.10]) then reduces the inversion of \mathbf{B} to that of an $R \times R$ matrix:

$$\mathbf{B}^{-1} = 2(\mathbf{I}_m + \mathbf{U} \mathbf{C} \mathbf{U}^\top)^{-1} = 2[\mathbf{I}_m - \mathbf{U}(\mathbf{C}^{-1} + \mathbf{U}^\top \mathbf{U})^{-1} \mathbf{U}^\top].$$

Forming the Gram matrix $\mathbf{U}^\top \mathbf{U}$ requires $\mathcal{O}(mR^2)$ operations (each of its R^2 entries is an inner product of two length- m vectors) while inverting the resulting dense $R \times R$ matrix costs $\mathcal{O}(R^3)$. Therefore, assembling and solving the small system costs

$$\mathcal{O}(mR^2) + \mathcal{O}(R^3) = \mathcal{O}(mR^2 + R^3)$$

instead of $\mathcal{O}(m^3)$ for a full $m \times m$ inversion. Whenever $R \ll m$, this yields a dramatic speed-up. By replacing the direct $\mathcal{O}(R^3)$ factorization with a Conjugate-Gradient (CG) solver — as is commonly done in Gaussian-process toolkits such as GPyTorch [11] — we reduce the cost to $\mathcal{O}(R^2)$.

Finally, many survival datasets exhibit censoring, i.e. $\delta_i = 0$ for a fraction of observations. Since censored observations contribute only through the integral term, we may further partition the low-rank factor \mathbf{U} into blocks for uncensored and censored cases. The effective rank becomes $R' = N_{\text{uncensored}} + NK$ where $N_{\text{uncensored}}$ is the number of uncensored observations, so that any Cholesky or CG solve scales with $(N_{\text{uncensored}} + NK)$ rather than $(N + NK)$. When $N_{\text{uncensored}} \ll N$, this yields an additional, potentially large reduction in computational cost.

J Experiment Set-Up

J.1 Real Survival Data

The real survival data used in Section 5.2 are presented below. In the central experiment, each dataset was subsampled to contain 125 observations in total. In an ablation experiment, each dataset was subsampled to contain 250 observations in total. Then, we performed 5-fold cross-validation, where the dataset was randomly divided into five equal parts. In each fold, one part (20%) was used as the test set (central experiment: 25 samples, ablation experiment: 50 samples), while the remaining four parts (80%) formed the training set (central experiment: 100 samples, ablation experiment: 200 samples). From the training set, 20% (central experiment: 20 samples, ablation experiment: 40 samples) was further attributed to the validation set.

Colon. The first successful trials of adjuvant chemotherapy for colon cancer dataset was obtained from the `survival` package [39]. The dataset contains records of 1,822 observations with 15 covariates among which 49.23% are censored. All rows with missing values were excluded from the dataset.

NWTCO. The National Wilm’s Tumor Study (NWTCO) was obtained from the `pycox` package [23]. The dataset contains records of 4,028 observations with 7 covariates among which 14.18% are censored.

GBSG. The Rotterdam and German Breast Cancer Study Group (GBSG) was obtained from the `pycox` package [23]. The dataset contains records of 2,232 observations with 7 covariates among which 43.23% are censored.

METABRIC. The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset was obtained from the `pycox` package [23]. The dataset contains records of 1,904 observations with 9 covariates among which 42.07% are censored.

WHAS. The Worcester Heart Attack Study (WHAS) dataset was obtained from the `sksurv` package [35]. The dataset contains records of 500 observations with 14 covariates among which 43.00% are censored.

SUPPORT. The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatment (SUPPORT) dataset was obtained from the `pycox` package [23]. The dataset contains records of 8,873 observations with 14 covariates among which 31.97% are censored.

VLC. The Veterans administration Lung Cancer trial (VLC) dataset was obtained from the `sksurv` package [35]. The dataset contains records of 137 observations with 8 covariates among which 6.57% are censored.

SAC 3. The Sac 3 dataset from the simulation study in [24, Appendix A.1] was obtained from the `pycox` package [23]. The dataset contains records of 100,000 observations with 45 covariates among which 37.20% are censored.

J.2 Benchmark Methods

J.2.1 Benchmark Deep Survival Methods

All deep learning methods share the same neural network architecture, which is detailed in Section K. The benchmark deep survival models were trained using the Adam optimizer with a learning rate selected via grid search. Batch normalization was applied, and a dropout rate of 0.1 was used. Training was conducted for 1,000 epochs with a batch size of 256.

MTLR. The Multi-Task Logistic Regression [43] was implemented using the `MTLR` class from the `pycox` package [23].

DeepHit. The DeepHit method [28] was implemented using the `DeepHitSingle` class from the `pycox` package [23]. The hyperparameters α and σ were set to 0.2 and 0.1, respectively. Those are the default values.

DeepSurv. The DeepSurv model [20] was implemented using the `CoxPH` class from the `pycox` package [23].

Logistic Hazard. The Logistic Hazard method [43] was implemented using the `LogisticHazard` class from the `pycox` package [23].

CoxTime. The CoxTime method [25] was implemented using the `CoxTime` class from the `pycox` package [23].

CoxCC. The CoxCC method [25] was implemented using the `CoxCC` class from the `pycox` package [23].

PMF. The PMF method [24] was implemented using the `PMF` class from the `pycox` package [23].

PCHazard. The PCHazard method [24] was implemented using the `PCHazard` class from the `pycox` package [23].

BCESurv. The BCESurv method [24] was implemented using the `BCESurv` class from the `pycox` package [23].

DySurv. The DySurv method [32] was implemented using the official code provided by the authors, available at https://github.com/munibmesinovic/DySurv/blob/main/Models/Results/Static_Benchmarks_GBSG_Example.ipynb (Accessed on May 13 2025).

J.2.2 Traditional Survival Methods

CoxPH. The Cox Proportional Hazards model [7] was implemented using the `CoxPHFitter` class from the `lifelines` package [8]. The Breslow method was used to compute the survival function.

Weibull AFT. The Weibull Accelerated Failure Time model [6] was implemented using the `WeibullAFTFitter` class from the `lifelines` package [8].

RSF. The Random Survival Forest [19] was implemented using the `RandomSurvivalForest` class from the `sksurv` package [35]. The number of trees in the forest is set to 1,000. The minimum number of samples required to split an internal node is 10, and the minimum number of samples required to be at a leaf node is 15. Those were the same hyperparameters as used in [32].

SSVM. The Survival Support Vector Machine [36] was implemented using the `FastSurvivalSVM` class from the `sksurv` package [35]. The optimal regularization hyperparameter α was selected via grid search by evaluating model performance on the training set using the C-index. This method does not allow for estimation of the survival function. Predicted ranks were used as risk scores for computing the C-index.

J.3 Evaluation metrics

C-index. Let $\hat{q}_i(t)$ be the predicted risk score of observation with covariates \mathbf{x}_i at time t . The C-index estimate [15] is given by

$$\text{C-index} = \frac{\sum_{i=1}^N \sum_{j \neq i} \delta_i \mathbb{1}_{\{y_i < y_j\}} \left(\mathbb{1}_{\{\hat{q}_i(y_i) > \hat{q}_j(y_i)\}} + \frac{1}{2} \mathbb{1}_{\{\hat{q}_i(y_i) = \hat{q}_j(y_i)\}} \right)}{\sum_{i=1}^N \sum_{j \neq i} \delta_i \mathbb{1}_{\{y_i < y_j\}}}.$$

Let $\hat{S}_i(t)$ be the predicted survival function of observation with covariates \mathbf{x}_i at time t . When the predicted risk score is taken to be the negative of the survival function, i.e., $\hat{q}_i(t) = -\hat{S}_i(t)$, the

C-index is referred to as the Antolini's C-index [1] and is found with

$$\text{C-index} = \frac{\sum_{i=1}^N \sum_{j \neq i} \delta_i \mathbb{1}_{\{y_i < y_j\}} \left(\mathbb{1}_{\{\hat{S}_i(y_i) < \hat{S}_j(y_i)\}} + \frac{1}{2} \mathbb{1}_{\{\hat{S}_i(y_i) = \hat{S}_j(y_i)\}} \right)}{\sum_{i=1}^N \sum_{j \neq i} \delta_i \mathbb{1}_{\{y_i < y_j\}}}.$$

The C-index is obtained using the `ConcordanceIndex` class from the `TorchSurv` package [33].

IPCW Integrated Brier Score Let $\hat{S}_i(t)$ be the predicted survival function of observation with covariates \mathbf{x}_i at time t . Let the inverse probability censoring weight (IPCW) at time t be defined as the inverse of the probability of being uncensored, $\xi(t) = 1/\hat{C}(t)$, where $\hat{C}(t)$ denotes the Kaplan–Meier estimate of the censoring survival function. Under right censorship, the IPCW Brier score (BS) [14] at time t is given by

$$\text{IPCW BS}(t) = \frac{1}{N} \sum_{i=1}^N \xi(y_i) \mathbb{1}_{\{y_i \leq t, \delta_i = 1\}} (0 - \hat{S}_i(t))^2 + \xi(t) \mathbb{1}_{\{y_i > t\}} (1 - \hat{S}_i(t))^2. \quad (47)$$

The IBS is the integral of the Brier Score in (47). The IPCW weights and the IPCW IBS are computed using the `get_ipcw` function and the `BrierScore` class from the `TorchSurv` package [33].

K Implementation Details

Architecture. We employed a feedforward neural network with two hidden layers, each containing 16 neurons and using ReLu activations. The input of the network for observation $i = 1, \dots, N$ is the pair (t, \mathbf{x}_i) .

Time normalization. The observation period is normalized to the interval $[0, 1]$ by dividing each time value by the maximum observed time in the training set.

EM algorithm. The parameters are initialized so that they match their prior expected values. Specifically, we set $\theta^{(0)} = \mathbf{0}$ and $\phi^{(0)} = \alpha_0/\beta_0$. The maximization step of the EM algorithm is performed using the L-BFGS-B algorithm. The EM algorithm is considered to have converged when the relative change in the Q-function between consecutive iterations falls below a tolerance threshold of 10^{-6} for two successive iterations.

CAVI algorithm. The hyperparameters are initialized so that the expected values of the model parameters match the MAP estimates. Specifically, we set $\tilde{\alpha}^{(0)} = \phi_{\text{MAP}} \times \tilde{\beta}$, and $(\tilde{\mu}, \tilde{\Sigma})^{(0)} = (\theta_{\text{MAP}}, \mathbf{I}_m)$. The CAVI algorithm is considered to have converged when the relative change between successive parameter estimates falls below a tolerance threshold of 10^{-6} .

Integral approximation. The integrals required to compute the Q-function in the EM algorithm, as well as those involved in the optimal variational updates of ϕ and θ in the CAVI algorithm, are approximated using the trapezoidal rule.

Prior and ρ . For all experiments, we fix the hyperparameters of the prior distribution over ϕ , given in (6), to be $\alpha_0, \beta_0 = 1$. Furthermore, we fix $\rho = 1$.

Machine. The experiments were conducted on NVIDIA RTX A6000 GPUs with 48GB of memory. The complete synthetic experiment, spanning four datasets, took approximately 30 minutes to run. For the real survival data, each data fold required about 20 minutes, while each fold in the ablation study took around 40 minutes. All folds and datasets were run in parallel across multiple GPUs.

Code availability. The code is available on the GitHub repository <https://github.com/MLGlobalHealth/neuralsurv> under the MIT License.

L Related Work

Survival analysis methodologies have evolved significantly over the past decades, encompassing parametric, semi-parametric, non-parametric, and more recently, deep learning-based approaches. We review these developments, focusing on their applicability to high-dimensional data and uncertainty quantification capabilities.

Parametric and Semi-parametric Traditional Models. Traditional survival models often impose parametric or semi-parametric assumptions on the hazard function. The Accelerated Failure Time (AFT) model [6] assumes a linear relationship between covariates and the logarithm of survival time, with parametric baseline distributions (e.g., Weibull). While interpretable, such models struggle with high-dimensional data and nonlinear covariate effects. The Cox Proportional Hazards (CoxPH) model [7], a semi-parametric approach, avoids specifying the baseline hazard but assumes proportional hazards. Though widely adopted, CoxPH’s linear predictor and proportionality constraints limit its flexibility in complex data regimes.

Non-parametric Traditional Models. To mitigate parametric assumptions, non-parametric methods like Random Survival Forests (RSF) [19] and Survival Support Vector Machines (SSVM) [36] emerged. RSF leverages ensemble learning for risk stratification but faces challenges in high-dimensional settings due to greedy tree induction. GP survival models [10] offer flexibility by modeling the hazard function nonparametrically, with inherent uncertainty quantification. Existing work has sought to address the cubic complexity in sample size of GPs by introducing variational inference techniques [21]. However, GPs remain fundamentally limited in scalability, particularly struggling with high-dimensional inputs and lacking the capacity to learn hierarchical representations, such as those required in image-based tasks [38].

Deep Survival Models. The advent of deep learning revolutionized survival analysis by enabling automatic feature learning from high-dimensional inputs. DeepSurv [20] extended CoxPH with neural networks, while DeepHit [28] employed multi-task learning for competing risks via discrete-time hazards. Discrete-time methods, including MTLR [43] and PCHazard [24], discretize the time axis to simplify likelihood computation, with recent advances like DySurv [32] incorporating conditional variational inference for dynamic prediction. Despite their predictive prowess, these models rely on frequentist training, yielding point estimates without uncertainty quantification — a significant shortcoming in safety-critical applications. Comprehensive reviews [42] highlight the rapid growth of deep survival methods but underscore their neglect of probabilistic uncertainty.

Bayesian and Uncertainty-Aware Approaches. Bayesian methods provide a natural framework for uncertainty quantification but have seen limited integration with deep survival models. GP-based approaches [10, 21] inherit GP limitations in scalability and high-dimensional processing. Recent works like BCESurv [26] explore bootstrap confidence intervals, yet these post-hoc approximations lack the coherence of Bayesian posteriors. Consequently, existing Bayesian survival models either sacrifice scalability for uncertainty quantification or compromise on model flexibility, leaving a critical gap in high-dimensional, uncertainty-aware survival analysis.

Summary. While parametric and semi-parametric models provide interpretability, they falter in high-dimensional, nonlinear regimes. Non-parametric methods like RSF and GP improve flexibility but face scalability challenges. Deep learning approaches excel at feature extraction yet lack principled uncertainty quantification. Bayesian methods, though theoretically sound, remain confined to traditional architectures or partial approximations. Our work bridges this divide by proposing the first scalable, deep Bayesian survival model that harmonizes neural networks with full probabilistic uncertainty, addressing a critical need in modern applications.

Method	Uncertainty (Bayesian)	Continuous Time	Deep Learning
CoxPH [7]	✓	✓	✗
AFT [6]	✓	✓	✗
RSF [19]	✗	✓	✗
SSVM [36]	✗	✓	✗
GP survival models [10, 21]	✓	✓	✗
MTLR [43]	✗	✗	✓
DeepHit [28]	✗	✗	✓
DeepSurv [20]	✓	✓	✓
Logistic Hazard [13]	✗	✗	✓
CoxTime [25]	✗	✓	✓
CoxCC [25]	✗	✓	✓
PMF [24]	✗	✗	✓
PCHazard [24]	✗	✓	✓
BCESurv [26]	✗	✗	✓
DySurv [32]	✗	✗	✓
<i>NeuralSurv (Ours)</i>	✓	✓	✓

Table A1: Summary of survival-analysis methods: uncertainty quantification, time domain, and deep-learning status.

M Further Results

M.1 Synthetic Data Experiment

Method	$N = 25$		$N = 50$		$N = 100$		$N = 150$	
	C-index \uparrow	IPCW IBS \downarrow	C-index \uparrow	IPCW IBS \downarrow	C-index \uparrow	IPCW IBS \downarrow	C-index \uparrow	IPCW IBS \downarrow
MTLR [43]	<u>0.56</u>	0.284	0.505	0.239	0.491	0.171	0.542	0.17
DeepHit [28]	0.473	0.239	0.469	<u>0.214</u>	0.502	0.171	0.574	0.114
DeepSurv [20]	0.492	0.313	0.471	0.241	0.507	0.169	0.517	0.169
Logistic Hazard [13]	0.477	0.297	0.498	0.256	0.507	0.199	0.499	0.176
CoxTime [25]	0.424	0.284	0.532	0.273	0.52	0.184	0.575	0.118
CoxCC [25]	0.421	0.268	0.497	0.229	0.526	<u>0.128</u>	0.513	<u>0.109</u>
PMF [24]	0.573	0.261	0.551	0.334	0.523	0.168	0.607	0.184
PCHazard [24]	0.477	0.337	0.501	0.249	0.467	0.174	0.486	0.193
BCESurv [26]	0.545	0.287	0.585	0.256	<u>0.558</u>	0.185	0.559	0.16
DySurv [32]	0.399	<u>0.237</u>	0.491	0.239	0.459	0.218	0.489	0.174
NeuralSurv (Ours)	0.378	0.196	<u>0.554</u>	0.16	0.589	0.126	<u>0.589</u>	0.106

Table A2: Performance comparison of survival models over synthetic data. The best results for each metric are shown in bold, and the second-best results are underlined. \uparrow indicates higher is better; \downarrow indicates lower is better.

M.2 Real Data Experiment

COLON			METABRIC		GBSG	
Method	C-index ↑	IPCW IBS ↓	C-index ↑	IPCW IBS ↓	C-index ↑	IPCW IBS ↓
MTLR [43]	0.562	0.298	0.548	0.279	0.602	0.273
DeepHit [28]	0.478	0.28	0.511	<u>0.243</u>	0.578	0.309
DeepSurv [20]	0.572	0.326	0.523	0.289	0.618	0.252
Logistic Hazard [13]	0.490	0.321	0.541	0.317	0.618	0.296
CoxTime [25]	0.578	<u>0.277</u>	0.533	0.307	0.599	0.285
CoxCC [25]	<u>0.584</u>	0.289	0.575	0.257	0.646	<u>0.240</u>
PMF [24]	0.509	0.324	0.440	0.336	<u>0.655</u>	0.250
PCHazard [24]	0.538	0.297	0.541	0.291	0.609	0.249
BCESurv [26]	0.491	0.302	0.616	0.277	0.581	0.273
DySurv [32]	0.488	0.536	0.561	0.465	0.572	0.485
NeuralSurv (Ours)	0.671	0.218	<u>0.584</u>	0.212	0.657	0.188

NWTCO			WHAS		SUPPORT	
Method	C-index ↑	IPCW IBS ↓	C-index ↑	IPCW IBS ↓	C-index ↑	IPCW IBS ↓
MTLR [43]	0.592	0.301	0.490	0.315	0.432	0.357
DeepHit [28]	0.516	0.296	0.510	0.303	0.452	0.341
DeepSurv [20]	0.527	0.248	0.654	0.281	0.505	0.354
Logistic Hazard [13]	0.512	0.298	0.545	0.315	0.536	0.378
CoxTime [25]	<u>0.550</u>	<u>0.199</u>	0.678	<u>0.250</u>	0.547	<u>0.327</u>
CoxCC [25]	0.531	0.237	<u>0.654</u>	0.281	<u>0.566</u>	0.312
PMF [24]	0.482	0.312	0.520	0.299	0.512	0.399
PCHazard [24]	0.551	0.209	0.527	0.291	0.514	0.335
BCESurv [26]	0.530	0.272	0.548	0.292	0.446	0.398
DySurv [32]	0.402	0.683	0.424	0.523	0.525	0.342
NeuralSurv (Ours)	0.712	0.166	0.602	0.233	0.599	0.333

VLC			SAC3	
Method	C-index ↑	IPCW IBS ↓	C-index ↑	IPCW IBS ↓
MTLR [43]	0.432	0.299	0.471	0.276
DeepHit [28]	0.409	0.236	0.456	0.289
DeepSurv [20]	0.642	<u>0.186</u>	0.530	0.264
Logistic Hazard [13]	0.413	0.272	0.480	0.348
CoxTime [25]	0.671	0.212	0.485	0.276
CoxCC [25]	0.645	0.169	0.533	<u>0.261</u>
PMF [24]	0.445	0.284	0.472	0.270
PCHazard [24]	0.502	0.294	0.527	0.276
BCESurv [26]	0.428	0.263	0.440	0.300
DySurv [32]	0.436	0.162	0.476	0.303
NeuralSurv (Ours)	<u>0.667</u>	0.142	<u>0.532</u>	0.204

Table A3: Performance comparison of deep survival models over five different train/test splits of each dataset. The best results for each metric are shown in bold, and the second-best results are underlined. ↑ indicates higher is better; ↓ indicates lower is better.

Method	COLON		METABRIC		GBSG	
	C-index \uparrow	IPCW IBS \downarrow	C-index \uparrow	IPCW IBS \downarrow	C-index \uparrow	IPCW IBS \downarrow
MTLR [43]	0.545	0.291	0.572	0.290	<u>0.567</u>	0.312
DeepHit [28]	0.564	0.284	0.545	0.301	0.563	0.272
DeepSurv [20]	0.600	0.295	0.605	0.265	0.531	0.277
Logistic Hazard [13]	0.501	0.289	0.553	0.252	0.562	0.287
CoxTime [25]	0.621	<u>0.259</u>	0.621	0.264	0.578	0.255
CoxCC [25]	0.640	0.277	<u>0.610</u>	0.254	0.565	<u>0.244</u>
PMF [24]	0.541	0.291	0.554	0.300	0.537	0.304
PCHazard [24]	0.549	0.280	0.561	<u>0.246</u>	0.524	0.295
BCESurv [26]	0.537	0.289	0.565	0.289	0.554	0.301
DySurv [32]	0.478	0.543	0.516	0.491	0.506	0.508
NeuralSurv (Ours)	<u>0.601</u>	0.215	0.543	0.198	0.546	0.212

Table A4: Performance comparison of deep survival models on the ablation study with 250 observations, over five different train/test splits of each dataset. The best results for each metric are shown in bold, and the second-best results are underlined. \uparrow indicates higher is better; \downarrow indicates lower is better.

Method	COLON		NWTGO		GBSG	
	C-index \uparrow	IPCW IBS \downarrow	C-index \uparrow	IPCW IBS \downarrow	C-index \uparrow	IPCW IBS \downarrow
CoxPH [7]	0.669	0.192	0.710	0.136	0.694	0.171
Weibull AFT [6]	0.681	0.198	0.697	0.134	0.673	0.179
RSF [19]	0.590	0.210	0.604	0.156	0.588	0.193
SSVM [36]	0.654	-	0.734	-	0.695	-

Method	METABRIC		WHAS		SUPPORT	
	C-index \uparrow	IPCW IBS \downarrow	C-index \uparrow	IPCW IBS \downarrow	C-index \uparrow	IPCW IBS \downarrow
CoxPH [7]	0.653	0.171	0.655	0.207	0.653	0.225
Weibull AFT [6]	0.658	0.172	0.622	0.224	0.650	0.239
RSF [19]	0.587	0.189	0.683	0.209	0.601	0.225
SSVM [36]	0.649	-	0.653	-	0.636	-

Method	VLC		SAC3	
	C-index \uparrow	IPCW IBS \downarrow	C-index \uparrow	IPCW IBS \downarrow
CoxPH [7]	0.697	0.125	0.569	0.190
Weibull AFT [6]	0.690	0.127	0.607	0.287
RSF [19]	0.687	0.139	0.487	0.182
SSVM [36]	0.698	-	0.504	-

Table A5: Performance comparison of traditional survival models over five different train/test splits of each dataset. \uparrow indicates higher is better; \downarrow indicates lower is better. The SSVM method does not provide estimates of the survival function; therefore, the predicted ranks are used for the corresponding C-index evaluations while the IPCW-IBS metric cannot be computed.

N Proofs

N.1 Proof of Theorem 3.1

Before proving Theorem 3.1 we must show some intermediate results.

Lemma N.1. *Assume that for each $i = 1, \dots, N$ the function $g(\cdot, \mathbf{x}_i; \cdot) \in C([0, y_i] \times \mathbb{R}^m)$. Then, it follows that*

$$\int_0^{y_i} \lambda_0(t, \mathbf{x}_i; \phi) dt < \infty$$

for every $i = 1, \dots, N$.

Proof. Fix an arbitrary index $i \in \{1, \dots, N\}$. From Section 2.2, recall that $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is the probability density function of a multivariate normal distribution with zero mean and identity covariance matrix \mathbf{I}_m . Our goal is to show that the normalizing constant $Z(t, \mathbf{x}_i)$ admits a strictly positive lower bound on $[0, y_i]$, from which the integrability of $\lambda_0(t, \mathbf{x}_i; \phi)$ will follow.

Step 1: Continuity of $Z(t, \mathbf{x}_i)$ on $[0, y_i]$. Fix any $t_0 \in [0, y_i]$, and let $(t_n)_{n \geq 1}$ be a sequence in $[0, y_i]$ such that $t_n \rightarrow t_0$ as $n \rightarrow \infty$. Define, for each n , the functions

$$\begin{aligned} h_n(\boldsymbol{\theta}) &:= \sigma(g(t_n, \mathbf{x}_i; \boldsymbol{\theta})) p_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \quad n \geq 1, \\ h(\boldsymbol{\theta}) &:= \sigma(g(t_0, \mathbf{x}_i; \boldsymbol{\theta})) p_{\boldsymbol{\theta}}(\boldsymbol{\theta}). \end{aligned}$$

Since $g(\cdot, \mathbf{x}_i; \cdot) \in C([0, y_i] \times \mathbb{R}^m)$ and the sigmoid $\sigma(\cdot)$ is a continuous function, it follows that

$$\lim_{n \rightarrow \infty} h_n(\boldsymbol{\theta}) = h(\boldsymbol{\theta})$$

pointwise for all $\boldsymbol{\theta} \in \mathbb{R}^m$. Furthermore, observe that

$$|h_n(\boldsymbol{\theta})| \leq p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$$

since $0 < \sigma(\cdot) < 1$. Because $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ integrates to 1 over \mathbb{R}^m , we may apply the Dominated Convergence Theorem (DCT) to conclude that:

$$\lim_{n \rightarrow \infty} Z(t_n, \mathbf{x}_i) = \lim_{n \rightarrow \infty} \int_{\mathbb{R}^m} h_n(\boldsymbol{\theta}) d\boldsymbol{\theta} \stackrel{\text{DCT}}{=} \int_{\mathbb{R}^m} h(\boldsymbol{\theta}) d\boldsymbol{\theta} = Z(t_0, \mathbf{x}_i).$$

Since t_0 was arbitrary in $[0, y_i]$, Z is continuous everywhere on that interval.

Step 2: Strict positivity of $Z(t, \mathbf{x}_i)$ on $[0, y_i]$. For each fixed $t \in [0, y_i]$, since $\sigma(g(t, \mathbf{x}_i; \boldsymbol{\theta})) > 0$ and $p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \mathbb{R}^m$, we have:

$$Z(t, \mathbf{x}_i) = \int_{\mathbb{R}^m} \sigma(g(t, \mathbf{x}_i; \boldsymbol{\theta})) p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta} > 0.$$

Since $Z(t, \mathbf{x}_i)$ is a continuous and strictly positive function on the compact interval $[0, y_i]$, the Weierstrass Extreme Value Theorem ensures that Z attains a minimum on this interval. Define:

$$z^* = \min_{t \in [0, y_i]} Z(t, \mathbf{x}_i) > 0$$

Step 3: Integrability of $\lambda_0(t, \mathbf{x}_i; \phi)$. Note that for all $t \in [0, y_i]$, we have

$$\lambda_0(t, \mathbf{x}_i; \phi) = \frac{\lambda_0(t; \phi)}{Z(t, \mathbf{x}_i)} \leq \frac{\lambda_0(t; \phi)}{z^*}.$$

It is straightforward to verify that $\lambda_0(t; \phi)$ is integrable on $[0, y_i]$, therefore it follows that

$$\int_0^{y_i} \lambda_0(t, \mathbf{x}_i; \phi) dt \leq \frac{1}{z^*} \int_0^{y_i} \lambda_0(t; \phi) dt < \infty.$$

This completes the proof. \square

Our next result verifies a condition needed for applying Campbell's Theorem in the proof of Theorem 3.1. To establish this, we will use the following Pólya–Gamma identity:

$$\mathbb{E}_{\omega \sim \text{pPG}(\omega|1,0)}[\omega] = \frac{1}{4}, \quad (48)$$

which follows by taking the limit $c \rightarrow 0$ in equation (22). Alternatively, to prove (48), one can start from the representation in equation (19), apply Tonelli's theorem to interchange expectation and infinite summation, and then invoke the series identity

$$\sum_{k=1}^{\infty} \frac{1}{(k - \frac{1}{2})^2} = \frac{\pi^2}{2}.$$

We are now ready to present our next result.

Lemma N.2. *Assume that for each $i = 1, \dots, N$ the function $g(\cdot, \mathbf{x}_i; \cdot) \in C([0, y_i] \times \mathbb{R}^m)$. Then, with probability 1 the sum*

$$H(\Psi_i) = \sum_{(t, \omega)_j \in \Psi_i} f(\omega_j, -g(t_j, \mathbf{x}_i; \boldsymbol{\theta}))$$

is absolutely convergent for every $i = 1, \dots, N$.

Proof. Fix an arbitrary index $i \in \{1, \dots, N\}$. Recall the definition of $f(\omega, z)$ from (8). From Theorem C.3, it suffices to show

$$\int_0^{y_i} \int_0^\infty \min(|f(\omega, -g(t, \mathbf{x}_i; \boldsymbol{\theta}))|, 1) \lambda_i(t, \omega; \phi) d\omega dt < \infty. \quad (49)$$

Since $\omega \in \mathbb{R}_+$, then it follows from the triangle inequality that

$$\begin{aligned} \min(|f(\omega, -g(t, \mathbf{x}_i; \boldsymbol{\theta}))|, 1) &\leq |f(\omega, -g(t, \mathbf{x}_i; \boldsymbol{\theta}))| \\ &\leq \frac{|g(t, \mathbf{x}_i; \boldsymbol{\theta})|}{2} + \frac{g(t, \mathbf{x}_i; \boldsymbol{\theta})^2}{2} \omega + \log(2). \end{aligned}$$

Hence it remains to prove finiteness of three integrals:

$$\begin{aligned} \mathcal{I}_1 &:= \int_0^{y_i} \int_0^\infty \frac{|g(t, \mathbf{x}_i; \boldsymbol{\theta})|}{2} \lambda_i(t, \omega; \phi) d\omega dt, \\ \mathcal{I}_2 &:= \int_0^{y_i} \int_0^\infty \frac{g(t, \mathbf{x}_i; \boldsymbol{\theta})^2}{2} \omega \lambda_i(t, \omega; \phi) d\omega dt, \\ \mathcal{I}_3 &:= \log(2) \int_0^{y_i} \int_0^\infty \lambda_i(t, \omega; \phi) d\omega dt. \end{aligned}$$

\mathcal{I}_1 is finite. Since $g(t, \mathbf{x}_i; \boldsymbol{\theta})$ is continuous on the compact interval $[0, y_i]$, it is bounded by some $M > 0$. Then,

$$\mathcal{I}_1 = \left(\int_0^\infty p_{\text{PG}}(\omega|1,0) d\omega \right) \int_0^{y_i} \frac{|g(t, \mathbf{x}_i; \boldsymbol{\theta})|}{2} \lambda_0(t, \mathbf{x}_i; \phi) dt \leq M \int_0^{y_i} \lambda_0(t, \mathbf{x}_i; \phi) dt < \infty,$$

where the last inequality is Lemma N.1.

\mathcal{I}_2 is finite. Likewise $g(t, \mathbf{x}_i; \boldsymbol{\theta})^2$ is bounded by some $C > 0$ over $[0, y_i]$ and $\mathbb{E}_{\omega \sim \text{pPG}(\omega|1,0)}[\omega] = \frac{1}{4}$ (see (48)), so

$$\mathcal{I}_2 = (\mathbb{E}_{\omega \sim \text{pPG}(\omega|1,0)}[\omega]) \int_0^{y_i} \frac{g(t, \mathbf{x}_i; \boldsymbol{\theta})^2}{2} \lambda_0(t, \mathbf{x}_i; \phi) dt \leq \frac{C}{8} \left(\int_0^{y_i} \lambda_0(t, \mathbf{x}_i; \phi) dt \right) < \infty,$$

where the last inequality is Lemma N.1.

\mathcal{I}_3 is finite. Finally,

$$\mathcal{I}_3 = \log(2) \int_0^{y_i} \lambda_0(t, \mathbf{x}_i; \phi) dt < \infty,$$

again by Lemma N.1.

Since $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$, are all finite, the condition in (49) is satisfied and the sum $H(\Psi_i)$ converges absolutely with probability 1. \square

The next result presents an integral identity which is key to proving the data augmentation scheme of Theorem 3.1.

Lemma N.3. *Assume that for each $i = 1, \dots, N$ the function $g(\cdot, \mathbf{x}_i; \cdot) \in C([0, y_i] \times \mathbb{R}^m)$. Then the double integral*

$$\int_0^{y_i} \int_0^\infty \left(1 - e^{f(\omega, -g(t, \mathbf{x}_i; \theta))}\right) p_{PG}(\omega|1, 0) \lambda_0(t, \mathbf{x}_i; \phi) d\omega dt$$

converges, and in fact

$$\int_0^{y_i} \int_0^\infty \left(1 - e^{f(\omega, -g(t, \mathbf{x}_i; \theta))}\right) p_{PG}(\omega|1, 0) \lambda_0(t, \mathbf{x}_i; \phi) d\omega dt = \int_0^{y_i} \lambda_0(t, \mathbf{x}_i; \phi) \sigma(g(t, \mathbf{x}_i; \theta)) dt \quad (50)$$

for every $i = 1, \dots, N$.

Proof. Fix an arbitrary index $i \in \{1, \dots, N\}$. By Lemma N.1

$$\int_0^{y_i} \lambda_0(t, \mathbf{x}_i; \phi) dt < \infty.$$

Since $0 < \sigma(\cdot) < 1$, we have

$$0 \leq \lambda_0(t, \mathbf{x}_i; \phi) \sigma(g(t, \mathbf{x}_i; \theta)) < \lambda_0(t, \mathbf{x}_i; \phi)$$

and therefore

$$\int_0^{y_i} \lambda_0(t, \mathbf{x}_i; \phi) \sigma(g(t, \mathbf{x}_i; \theta)) dt < \infty. \quad (51)$$

This shows the finiteness of the right-hand side of (50). By combining $\sigma(z) = 1 - \sigma(-z)$ with (23) we obtain that

$$\int_0^{y_i} \lambda_0(t, \mathbf{x}_i; \phi) \sigma(g(t, \mathbf{x}_i; \theta)) dt = \int_0^{y_i} \int_0^\infty \left(1 - e^{f(\omega, -g(t, \mathbf{x}_i; \theta))}\right) p_{PG}(\omega|1, 0) \lambda_0(t, \mathbf{x}_i; \phi) d\omega dt. \quad (52)$$

Putting together the finiteness from (51) with the equality of (52) completes the proof. \square

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. Fix an arbitrary index $i \in \{1, \dots, N\}$. The joint expectation factors into two independent pieces:

1. **Expectation over ω_i :** This term recovers $\lambda_0(y_i, \mathbf{x}_i; \phi)^{\delta_i} \sigma(g(y_i, \mathbf{x}_i; \theta))^{\delta_i}$;
2. **Expectation over Ψ_i :** This term recovers $\exp\left(-\int_0^{y_i} \lambda_0(t, \mathbf{x}_i; \phi) \sigma(g(t, \mathbf{x}_i; \theta)) dt\right)$.

Step (1): Expectation over ω_i . Since $\delta_i \in \{0, 1\}$,

$$\left(e^{f(\omega_i, g(y_i, \mathbf{x}_i; \boldsymbol{\theta}))}\right)^{\delta_i} = \begin{cases} e^{f(\omega_i, g(y_i, \mathbf{x}_i; \boldsymbol{\theta}))}, & \delta_i = 1, \\ 1, & \delta_i = 0. \end{cases}$$

Hence,

$$\int_0^\infty \left(e^{f(\omega_i, g(y_i, \mathbf{x}_i; \boldsymbol{\theta}))}\right)^{\delta_i} p_{\text{PG}}(\omega_i | 1, 0) d\omega_i = \left(\int_0^\infty e^{f(\omega_i, g(y_i, \mathbf{x}_i; \boldsymbol{\theta}))} p_{\text{PG}}(\omega_i | 1, 0) d\omega_i\right)^{\delta_i}.$$

By the Pólya–Gamma identity (Eq. (23)), the bracketed integral equals $\sigma(g(y_i, \mathbf{x}_i; \boldsymbol{\theta}))$. Multiplying by $\lambda_0(y_i, \mathbf{x}_i; \phi)^{\delta_i}$ gives exactly

$$\lambda_0(y_i, \mathbf{x}_i; \phi)^{\delta_i} \sigma(g(y_i, \mathbf{x}_i; \boldsymbol{\theta}))^{\delta_i}.$$

Step (2): Expectation over Ψ_i . By Lemma N.2 the random sum

$$H(\Psi_i) = \sum_{(t, \omega)_j \in \Psi_i} f(\omega_j, -g(t_j, \mathbf{x}_i; \boldsymbol{\theta}))$$

is absolutely convergent with probability 1, and by Lemma N.3 the corresponding integral converges. Therefore, we may apply Campbell’s Theorem (Theorem C.3) together with the PG-sigmoid identity from (50) to conclude

$$\mathbb{E}_{\Psi_i \sim \mathbb{P}_{\Psi_i} | \phi} \left[\prod_{(t, \omega)_j \in \Psi_i} e^{f(\omega_j, -g(t_j, \mathbf{x}_i; \boldsymbol{\theta}))} \right] = \exp \left(- \int_0^{y_i} \lambda_0(t, \mathbf{x}_i; \phi) \sigma(g(t, \mathbf{x}_i; \boldsymbol{\theta})) dt \right).$$

Putting Steps (1) and (2) together reproduces precisely the two factors of the original likelihood $p(y_i, \delta_i | \mathbf{x}_i, \phi, g)$. This completes the proof. \square