arXiv:2303.10358v2 [cs.LG] 4 Oct 2023

神经脆弱性机器：突破神经生存回归中的比例风险假设

吴若凡*†、乔佳伟*‡、吴明哲§、于文‡、郑明‡、刘腾飞†、张天翼†和王伟强†

†蚂蚁集团 ‡复旦大学 §Coupang{ruofan.wrf, aaron.ltf, zty113091, weiqiang.wwq}@antgroup.com jeremyqjw@163.com, wumingzhe.darcy@gmail.com, {wenyu, mingzheng}@fudan.edu.cn

摘要

我们提出神经脆弱性机器（NFM），这是一种功能强大且灵活的神经模型框架，用于生存回归。NFM 框架将生存分析中经典的乘法脆弱性思想作为一种有理论依据的方法，用于扩展比例风险假设，同时能够利用神经网络架构强大的近似能力来处理协变量的非线性依赖关系。在该框架下，我们推导出两个具体模型，分别扩展了神经比例风险模型和非参数风险回归模型。两个模型均可通过似然目标函数进行高效训练。理论上，对于这两个模型，我们通过刻画其收敛速率，建立了关于非参数分量的神经函数逼近统计保证。实验方面，我们提供了合成实验来验证理论结论。此外，我们在 6 个不同规模的基准数据集上进行了实验评估，结果表明，所提出的NFM模型具有与最先进的生存模型相当、甚至有时更优的预测性能。我们的代码已公开发布于 https://github.com/Rorschach1989/nfm

1 引言

对生存时间数据进行回归分析 [40] 一直是临床研究中最重要的建模工具之一，并且在公司金融 [22], 推荐系统 [38], 和计算广告 [71]等领域也日益受到关注。生存时间数据与其他类型数据的关键区别在于，它们往往无法被完整观测，其中最常见的不完整形式是右删失机制 [40]。在右删失机制下，样本对象的持续时间有时只能确定大于观察时间，而无法被精确记录。生存分析领域众所周知，即使在线性回归中，简单地丢弃删失观测也会导致统计上有偏的估计结果 [7], ，并且当删失比例较高时，还会损失样本效率。

Cox比例风险模型（CoxPH） [14] 以负偏似然的凸目标函数 [15] 为基础，是对右删失生存时间数据（为避免歧义，下文简称删失数据）建模的实际首选。该模型是半参数模型， [4] 其基准风险函数无需作参数假设。CoxPH 的原始形式

*贡献相同

第37届神经信息处理系统大会（NeurIPS 2023）。

假设模型为线性形式，因此灵活性有限，因为真实关系未必是线性的。后续研究借鉴非参数回归、 [36, 8, 9], 集成学习 [37], 和神经网络 [25, 42]的思想，将 Cox比例风险模型扩展为非线性变体。尽管这些扩展能够灵活地刻画与协变量之间的非线性依赖结构，但其学习目标仍是在比例风险（PH）假设下推导的，而研究表明，该假设在许多现实场景中并不适用 [31]。最突出的例子是无法对交叉风险现象进行建模 [61]。因此，探索能够同时刻画协变量间非线性依赖关系并放宽 PH 假设的 Cox比例风险模型扩展，具有重要意义。

脆弱性模型 [69, 21] 是现代生存分析中最重要的研究主题之一，因为它们通过引入乘性随机效应来捕捉未观测异质性，为扩展Cox比例风险模型提供了一种有理论依据的方法。在特定的脆弱性分布族选择下，由此得到的参数化包含许多有用的Cox比例风险模型变体，例如比例优势模型 [3], 。尽管脆弱性模型的理论已相当成熟 [48, 49, 53, 45], ，但其中大多数研究都集中在线性情形。近期将神经网络方法应用于生存分析的研究 [42, 46, 64, 56] 在经验预测性能方面展现出令人鼓舞的结果，但大多缺乏理论探讨。因此，采用现代深度学习技术 [29] 构建更强大的脆弱性模型，并为其提供可证明的统计保证，具有重要意义。

本文提出了一个用于扩展脆弱性模型的通用框架，称为神经脆弱性机器（NFM）。在该框架下，我们推导出两种具体的神经网络架构：第一种采用比例脆弱性假设，能够直观地解释带有乘性随机效应的神经Cox比例风险模型。第二种进一步放宽比例脆弱性假设，可视为非参数风险回归（NHR）的一种扩展 [13, 44], ；在神经生存分析 [52]领域中，它有时被称为“全神经”模型。我们的贡献总结如下。

• 我们提出神经脆弱性机器（NFM）框架，以一种有理论依据的方式将未观测异质性纳入神经生存回归模型。该框架涵盖了许多常用的生存回归模型作为特例。

• 我们基于 NFM 框架构建了两种模型架构，分别扩展了神经网络Cox比例风险模型和神经网络NHR模型。两种模型均支持随机训练，并可扩展至大型数据集。

• 从理论上讲，我们通过刻画所提出的非参数函数估计量的收敛速率，证明了这两种模型在统计上的正确性。我们的证明方法不同于以往关于神经生存分析的理论研究，并且适用于许多其他类型的神经生存模型。

• 在实证方面，我们通过在不同规模的各种基准数据集上开展广泛研究，验证了所提出框架的实证效能。根据标准性能指标，实证结果表明，两种模型的表现均具有竞争力，能够媲美甚至有时优于最先进的神经生存模型。

2 相关研究

2.1 Cox比例风险模型的非线性扩展

Cox比例风险模型的大多数非线性扩展都源于部分似然与Cox比例风险模型的半参数剖面似然等价这一性质 [50] ，由此得到的非线性变体本质上是用非线性形式替代部分似然中的线性项： [36] 使用了平滑样条，[8, 9] 使用了局部多项式回归 [24]。基于树的模型在实践中的成功，启发了后续的发展，例如 [37] ，它为梯度提升树和随机森林等基于树的模型配备了形式为负对数部分似然的损失函数。神经生存分析的早期发展 [25] 采用了类似的扩展策略，得到了部分似然的神经网络版本。后续研究 [42] 提出，应采用随机训练这一行之有效的做法；人们认为，这正是现代神经方法在实践中取得成功的关键 [34]。然而，在部分似然目标函数下进行随机训练并非易事，因为小批量版本的对数部分似然 [42] 不再是全样本对数部分似然的有效随机梯度 [64]。

2.2 生存分析中超越Cox比例风险模型

在线性生存建模中，Cox比例风险模型之外还有一些标准替代模型，例如加速失效时间模型（AFT） [7, 73], 、扩展风险回归模型 [23], 以及线性变换模型系列 [74]。尽管这些模型允许某些类型的非线性扩展，但所得的（条件）风险函数形式仍受限于特定形式。非参数风险回归（NHR）这一思路 [13, 44, 63] 通过使用样条逼近等非参数回归技术直接对条件风险函数建模，进一步提升了非参数生存分析的灵活性。近年来，人们开发了NHR的神经网络版本，例如CoxTime模型 [46]。 [56] 使用神经网络近似条件生存函数，因此也可视为NHR的另一种简单扩展。

除了NHR方面的发展， [47] 提出了一种离散时间模型，其目标函数结合了离散似然与基于秩的评分； [75] 提出了扩展风险模型的神经网络版本，将神经网络Cox比例风险模型与神经网络加速失效时间模型统一起来； [64] 采用常微分方程方法对风险函数和累积风险函数进行建模。

2.3 神经生存模型的理论依据

尽管神经生存模型层出不穷，但对其理论性质的评估仍处于起步阶段。在 [76], 中，作者使用神经网络作为函数逼近器，建立了部分线性Cox模型的极小极大理论。 [75] 为扩展风险模型下神经网络估计值的收敛提供了保证。其中的理论进展依赖于特定形式的目标函数（部分似然和核伪似然），因此并不直接适用于生存分析中常用的标准基于似然的目标函数。

3 方法

3.1 神经脆弱性机器框架

~ ~令 $T \geq 0$ 为感兴趣的事件时间，其生存函数记为 $S(t) = \mathbb{P}(T > t)$，并与特征（协变量）向量 $Z \in \mathbb{R}^d$相关。假设 T是连续随机变量，且令 $f(t)$为其密度函数。那么， $\lambda(t) = f(t)/S(t)$为风险函数，而 $\Lambda(t) = \int_0^t \lambda(s)ds$ 为累积风险函数。除协变量 Z 外，我们还使用一个正标量随机变量 $\omega \in \mathbb{R}^+$ 来表示个体对应的未观测异质性，即脆弱性。[2]。本文将通过指定其条件风险函数，假设 $\tilde{T}$遵循以下生成机制：

$$\lambda(t|Z, \omega) = \omega\tilde{\nu}(t, Z). \tag{1}$$

这里， $\tilde{\nu}$ 是一个未指定的非负函数，我们令 ω 的分布由一个一维参数 $\theta \in \mathbb{R}$参数化。 [3] 公式（1）的形式相当一般，涵盖了传统生存分析和神经生存分析中的若干重要模型：

1. 当 ω 服从参数化分布假设，且 $\tilde{\nu}(t, Z) = \lambda(t)e^{\beta^\top Z}$时，(1) 将退化为标准的比例脆弱性模型 [45]。一种特殊情况是 ω 退化，即不存在随机性，此时模型对应于经典的 Cox比例风险模型。

2. 当 ω 退化且 $\tilde{\nu}$ 任意时，该模型等价于非参数风险回归（NHR） [13, 44]。在 NHR 中，感兴趣的函数参数通常是（条件）风险函数的对数。

本文构建了对 $\tilde{\nu}$ 的对数的神经网络近似，即 $\nu(t, Z) = \log \tilde{\nu}(t, Z)$。由此得到的模型称为神经脆弱性机器（NFM）。根据对函数 ν 的先验了解，我们提出了两种函数近似方案：

[2]例如，在医学生物学中，有研究观察到，遗传上相同的动物即使生活在尽可能相似的环境中，接触环境致癌物后通常也不会表现相同 [6][3]选择一维脆弱性族主要是为了使理论推导更加简单清晰。值得注意的是，也存在多维脆弱性分布族，例如 PVF 族 [69]。将我们的理论结果推广到这类分布族需要额外的正则性条件，这将留待未来研究。

比例脆弱性（PF）方案假设 ν 对事件时间和协变量的依赖完全解耦，即，

$$\nu(t, Z) = h(t) + m(Z). \tag{2}$$

风险函数的比例形式假设已被证明是生存分析中一种有效的归纳偏置。我们将（2）中的 h 和 m 都视为函数参数，并设计两个多层感知机（MLP）分别对它们进行近似。

全神经（FN）方案不对 ν 施加任何先验假设，是 NFM 最一般的形式。显然，最常用的生存模型，如 Cox比例风险模型、加速失效时间模型、扩展风险模型或 PF 模型，都作为特例包含在所提出的模型空间中。我们将 $\nu = \nu(t, Z)$ 视为函数参数，其输入维度为 $d+1$，并使用多层感知机（MLP）作为 ν 的函数近似器。近期一些研究 [52, 56], 提出了针对风险函数的类似近似方案，并将其称为“全神经方法”，但没有给出理论刻画。

脆弱性分布族的选择 常用的脆弱性分布族有很多 [45, 21,69], ，其中最受欢迎的是伽马脆弱性，此时 ω 服从均值为 1、方差为 θ的伽马分布。我们在附录 A 中简要介绍其他几种脆弱性分布族。

3.2 删失观测下的参数学习

在生存时间建模场景中，事件时间通常会受到右删失。令 C 为右删失时间，假设其在给定 Z的条件下与事件时间~ ~T条件独立，即 $T \perp\!\!\!\perp C|Z$。在数据收集过程中，可以观测到生存时间与~ ~删失时间的较小值，也就是观测到 $T = T \wedge C$ 以及删失指标 $\delta = I(T \leqslant C)$，其中 $a \wedge b = \min(a, b)$，a 和 b 为常数，$I(\cdot)$表示指示函数。我们假设n 个独立同分布（i.i.d.）的(T, δ, Z)副本构成训练样本 (T_i, δ_i, Z_i) $i \in [n]$，其中用 $[n]$ 表示集合 $\{1, 2, \ldots, n\}$。此外，我们假设未观测到的脆弱性变量独立同分布，即 ω_i i.i.d. $\sim$ $f_\theta(\omega)$ $i \in [n]$。接下来，我们分别基于 PF 和 FN 方案下的观测对数似然（OLL）目标推导学习过程。为得到观测似然，我们首先对给定脆弱性时的条件生存函数进行积分：

$$S(t|Z) = \mathbb{E}_{\omega \sim f_\theta}\left[e^{-\omega \int_0^t e^{\nu(s,Z)} ds}\right] =: e^{-G_\theta\left(\int_0^t e^{\nu(s,Z)} ds\right)}. \tag{3}$$

这里，脆弱性变换 $G_\theta(x) = -\log\left(\mathbb{E}_{\omega \sim f_\theta}\left[e^{-\omega x}\right]\right)$ 定义为脆弱性分布的拉普拉斯变换的负对数。因此，条件累积风险函数为 $\Lambda(t|Z) = G_\theta(\int_0^t e^{\nu(s,Z)} ds)$。对于 NFM 的 PF方案，我们使用两个多层感知机 $\widehat{h} = \widehat{h}(t; \mathbf{W}^h, \mathbf{b}^h)$ 和 $\widehat{m} = \widehat{m}(Z; \mathbf{W}^m, \mathbf{b}^m)$ 作为函数逼近器，分别逼近 ν 和 m，其参数分别为 $(\mathbf{W}^h, \mathbf{b}^h)$ 和 $(\mathbf{W}^m, \mathbf{b}^m)$。[4] 根据删失数据似然的标准结果，[40], 我们将 PF方案下的学习目标写为：

$$\begin{aligned}
&\mathcal{L}(\mathbf{W}^h, \mathbf{b}^h, \mathbf{W}^m, \mathbf{b}^m, \theta) \\
=&\frac{1}{n}\left[\sum_{i \in [n]} \delta_i \log g_\theta\left(e^{\widehat{m}(Z_i)} \int_0^{T_i} e^{\widehat{h}(s)} ds\right) + \delta_i \widehat{h}(T_i) + \delta_i \widehat{m}(Z_i) - G_\theta\left(e^{\widehat{m}(Z_i)} \int_0^{T_i} e^{\widehat{h}(s)} ds\right)\right].
\end{aligned} \tag{4}$$

这里定义 $g_\theta(x) = \frac{\partial}{\partial x} G_\theta(x)$。令 $(\widehat{\mathbf{W}}_n^h, \widehat{\mathbf{b}}_n^h, \widehat{\mathbf{W}}_n^m, \widehat{\mathbf{b}}_n^m, \widehat{\theta}_n)$ 为式 (4) 的最大化解，并进一步记 $\widehat{h}_n(t) = \widehat{h}(t; \widehat{\mathbf{W}}_n^h, \widehat{\mathbf{b}}_n^h)$ 和 $\widehat{m}_n(Z) = \widehat{m}(Z; \widehat{\mathbf{W}}_n^m, \widehat{\mathbf{b}}_n^m)$。由此得到的条件累积风险和生存函数估计量为：

$$\widehat{\Lambda}_{\mathsf{PF}}(t|Z) = G_{\widehat{\theta}_n}\left(\int_0^t e^{\widehat{h}_n(s) + \widehat{m}_n(Z)} ds\right), \qquad \widehat{S}_{\mathsf{PF}}(t|Z) = e^{-\widehat{\Lambda}_{\mathsf{PF}}(t|Z)}, \tag{5}$$

[4]这里我们采用惯用记法：$\mathbf{W}$ 表示多层感知机各层权重矩阵的集合，$\mathbf{b}$ 表示各层偏置向量的集合。

对于 FN方案，我们使用 $\widehat{\nu} = \widehat{\nu}(t, Z; \mathbf{W}^{\nu}, \mathbf{b}^{\nu})$ 来近似由 $(\mathbf{W}^{\nu}, \mathbf{b}^{\nu})$ 参数化的 $\nu(t, Z)$。OLL目标函数写为:

$$\begin{aligned}&\mathcal{L}(\mathbf{W}^{\nu}, \mathbf{b}^{\nu}, \theta)\\ =&\frac{1}{n}\left[\sum_{i\in[n]}\delta_i \log g_\theta\left(\int_0^{T_i} e^{\widehat{\nu}(s,Z_i;\mathbf{W}^{\nu},\mathbf{b}^{\nu})}ds\right) + \delta_i\widehat{\nu}(T_i, Z_i; \mathbf{W}^{\nu}, \mathbf{b}^{\nu}) - G_\theta\left(\int_0^{T_i} e^{\widehat{\nu}(s,Z_i;\mathbf{W}^{\nu},\mathbf{b}^{\nu})}ds\right)\right].\end{aligned} \tag{6}$$

设 $(\widehat{\mathbf{W}}_n^{\nu}, \widehat{\mathbf{b}}_n^{\nu}, \widehat{\theta}_n)$ 为式 (6) 的最大化解，并进一步记 $\widehat{\nu}_n(t, Z) = \widehat{\nu}(t, Z; \widehat{\mathbf{W}}_n^{\nu}, \widehat{\mathbf{b}}_n^{\nu})$。因此，条件累积风险和生存函数的估计值为:

$$\widehat{\Lambda}_{\mathsf{FN}}(t|Z) = G_{\widehat{\theta}_n}\left(\int_0^t e^{\widehat{\nu}_n(s,Z)}ds\right), \qquad \widehat{S}_{\mathsf{FN}}(t|Z) = e^{-\widehat{\Lambda}_{\mathsf{FN}}(t|Z)}. \tag{7}$$

计算式 (6) 这类目标函数及其梯度，需要对经过指数变换的多层感知机函数进行定积分计算。我们不采用仅适用于特定类型激活函数和网络结构的精确计算方法，而是使用数值积分进行此类计算，具体采用克伦肖-柯蒂斯求积 [5], ; 近期将其应用于单调神经网络的研究已显示出良好的性能和效率 [68]。

备注3.1。两种方案中脆弱性项的解释有所不同。在 PF 方案中，与 Cox比例风险模型或其神经网络变体相比，引入脆弱性效应严格提升了模型的建模能力（即刻画交叉风险的能力） [45]。在 FN 方案中，可以认为，在独立同分布情形下，NHR 框架中的边际风险函数是风险函数的一种重新参数化。因此，我们将纳入脆弱性效应视为注入一种特定于定义域的归纳偏置；事实证明，这种偏置对生存分析和生存时间回归建模有帮助，我们将在第5.2节通过实证验证这一观点。此外，在处理相关数据或聚类数据时，脆弱性尤其有用，因为此时假设某些个体组共享同一个脆弱性项 [53]。将 NFM 扩展至此类场景很有价值，我们将其留待未来探索。

4 统计保证

本节将介绍 NFM 估计值的统计保证，并从非参数回归的角度 [65], 给出收敛至真实函数参数的速率（在统计学文献中，这些参数通常称为真实参数）。本节的结果表明，我们的方法在统计上是正确的。

证明策略 我们的证明技术基于筛法 [60, 59, 11] ，该方法将神经网络视为一种特殊的非线性筛 [11] ，并具有理想的近似性质[72]。我们的策略不同于以往关于神经生存模型的理论研究 [75, 76]: 那些研究的推导隐含地要求损失函数受到 L_2 损失的良好控制，因此由于脆弱性变换的选择具有灵活性，并不能直接应用于我们的模型。

由于这两种模型都会估计函数参数，我们需要指定一个合适的函数空间。这里我们选用以下赫尔德球，这也曾用于以神经网络进行非参数估计的先前研究 [57, 26, 76]

$$\mathcal{W}_M^{\beta}(\mathcal{X}) = \left\{f : \max_{\alpha:|\alpha|\le\beta}\operatorname*{esssup}_{x\in\mathcal{X}}|D^{\alpha}(f(x))| \le M\right\}, \tag{8}$$

其中，定义域 $\mathcal{X}$假设为 d 维欧几里得空间的子集。 $\alpha = (\alpha_1, \ldots, \alpha_d)$ 是一个 d 维的非负整数元组，满足 $|\alpha| = \alpha_1 + \cdots + \alpha_d$，且 $D^{\alpha}f = \frac{\partial^{|\alpha|}f}{\partial x_1^{\alpha_1}\cdots x_d^{\alpha_d}}$是 f 的弱导数。现在假设 M 是一个相当大的常数，并令 Θ为实数轴上的一个闭区间。我们对两种方案下的真实参数作出以下假设:

条件 4.1（真实参数，PF方案）。欧几里得参数 $\theta_0 \in \Theta \subset \mathbb{R}$和两个函数参数 $m_0 \in \mathcal{W}_M^{\beta}([-1, 1]^d)$ $h_0 \in \mathcal{W}_M^{\beta}([0, \tau])$，以及 $\tau > 0$是研究持续时间的结束时间，这通常用于生存分析的理论研究中 *[67]*。

条件 4.2（真实参数，FN方案）。欧几里得参数 $\theta_0 \in \Theta \subset \mathbb{R}$，以及函数参数 $\nu_0 \in \mathcal{W}_M^{\beta}([0, \tau] \times [-1, 1]^d)$,

接下来，我们通过限制多层感知机的复杂度，使其随样本量 n“缩放”，从而构造用于函数参数近似的筛空间。

条件4.3（筛空间，PF方案）。筛空间 $\mathcal{H}_n$ 构造为满足 $\widehat{h} \in \mathcal{W}_{M_h}^{\beta}([0,\tau])$ 的多层感知机集合，其深度为 $O(\log\ n)$ 阶，总参数数量为 $O(n^{\frac{1}{\beta+d}}\log n)$ 阶。筛空间 $\mathcal{M}_n$ 构造为满足 $\widehat{m} \in \mathcal{W}_{M_m}^{\beta}([-1,1]^d)$ 的多层感知机集合，其深度为 $O(\log\ n)$ 阶，总参数数量为$O(n^{\frac{d}{\beta+d}}\log n)$ 阶。这里， M_h 和 M_m 是足够大的常数，使得根据 *[72, 定理 1]*，$\mathcal{W}_{M}^{\beta}([-1,1]^d)$ 和 $\mathcal{W}_{M}^{\beta}([0,\tau])$ 中的每个函数都能由 $\mathcal{H}_n$和 $\mathcal{M}_n$中的函数准确近似。

条件 4.4（筛空间，FN方案）。筛空间 $\mathcal{V}_n$ 构造为满足 $\widehat{\nu} \in \mathcal{W}_{M_\nu}^{\beta}([0,\tau])$ 的多层感知机集合，其深度为 $O(\log\ n)$ 阶，总参数数量为$O(n^{\frac{d+1}{\beta+d+1}}\log n)$ 阶。这里， M_ν 是一个足够大的常数，使得 $\mathcal{V}_n$ 满足近似性质，与条件4.3类似。

出于技术原因，我们假设非参数函数估计量受限于相应的筛空间内，即 $h_n \in \mathcal{H}_n$，$\widehat{m}_n \in \mathcal{M}_n$ 和 $\widehat{\nu} \in \mathcal{V}_n$。如 [26]中所述，这不会影响优化程序的实现。此外，我们限制粒子滤波方案和 FN 方案中的估计值 $\theta_n \in \Theta$。

此外，函数 $G_\theta(x)$还需满足以下正则性条件：

条件 4.5。将 $G_\theta(x)$视为双变量函数 $G\colon \Theta \times \mathcal{B} \mapsto \mathbb{R}$，其中 $\mathcal{B}$是 $\mathbb{R}$上的一个紧集。函数 $G_\theta(x)$、$\frac{\partial}{\partial\theta}G_\theta(x)$、$\frac{\partial}{\partial x}G_\theta(x)$、$\log\ g_\theta(x)$、$\frac{\partial}{\partial\theta}\log\ g_\theta(x)$、 $\frac{\partial}{\partial x}\log\ g_\theta(x)$在 $\Theta \times \mathcal{B}$上有界。

我们定义两种度量来衡量参数估计的收敛：对于 PF 方案，令 $\phi_0 = (h_0, m_0, \theta_0)$为真实参数， $\phi_n = (h_n, \widehat{m}_n, \theta_n)$为估计值。我们将$\mathbb{P}_{\phi_0, Z=z}$ 简记为真实参数下给定 $Z = z$ 时(T, δ)的条件概率分布，将$\mathbb{P}_{\widehat{\phi}_n, Z=z}$ 简记为估计参数下给定 $Z = z$ 时(T, δ)的条件概率分布。定义如下度量

$$d_{\mathsf{PF}}\left(\widehat{\phi}_n, \phi_0\right) = \sqrt{\mathbb{E}_{z\sim\mathbb{P}_Z}\left[H^2(\mathbb{P}_{\widehat{\phi}_n, Z=z} \parallel \mathbb{P}_{\phi_0, Z=z})\right]}, \tag{9}$$

其中， $H^2(\mathbb{P} \parallel \mathbb{Q}) = \int\left(\sqrt{d\mathbb{P}} - \sqrt{d\mathbb{Q}}\right)^2$ 是概率分布 $\mathbb{P}$ 与 $\mathbb{Q}$之间的海林格距离的平方。

FN方案的情形类似：令 $\psi_0 = (\nu_0, \theta_0)$ 为参数， $\widehat{\nu}_n = (\widehat{\nu}_n, \theta_n)$ 为估计值。类似于上述定义，我们将给定 $Z = z$时的真实条件分布记为 $\mathbb{P}_{\psi_0, Z=z}$，将估计的条件分布记为 $\mathbb{P}_{\widehat{\psi}_n, Z=z}$。在FN方案中，我们将使用以下度量：

$$d_{\mathsf{FN}}\left(\widehat{\psi}_n, \psi_0\right) = \sqrt{\mathbb{E}_{z\sim\mathbb{P}_Z}\left[H^2(\mathbb{P}_{\widehat{\psi}_n, Z=z} \parallel \mathbb{P}_{\psi_0, Z=z})\right]}. \tag{10}$$

现在陈述我们的主要定理。我们将 $\mathbb{P}$ 记为数据生成分布，并用 O 在大O记号中隐藏多对数因子。

定理4.6（收敛速率，PF方案）。在满足条件 4.1、4.3、4.5的情况下，对于PF方案，我们有 $d_{PF}\left(\widehat{\phi}_n, \phi_0\right) = \widetilde{O}_{\mathbb{P}}\left(n^{-\frac{\beta}{2\beta+2d}}\right)$。

定理 4.7（收敛速率，FN方案）。在 FN方案中，在条件 4.2、4.4、4.5 下，我们有 $d_{FN}\left(\widehat{\psi}_n, \psi_0\right) = \widetilde{O}_{\mathbb{P}}\left(n^{-\frac{\beta}{2\beta+2d+2}}\right)$。

注释4.8。使用海林格距离衡量筛极大似然估计收敛速率的想法由 [70]提出。如果似然函数满足某些条件（例如曲率条件 [26]），则可以在更强的拓扑（如 L_2）下获得速率。然而，这类条件通常对基于似然的目标函数而言过于严格，因此我们采用要求最低的海林格收敛。因此，只需对正则性条件稍作修改，我们的证明策略便适用于许多其他依赖神经函数逼近的生存模型，例如 [56], 。关于筛理论中度量的恰当选择，另请参阅 [11, 第 2]章的讨论。

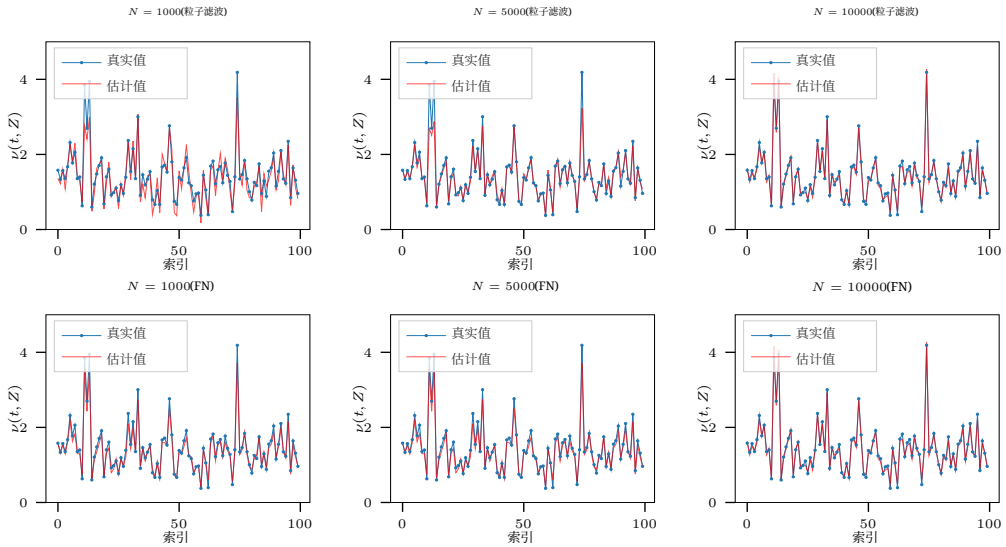


图1：NFM 框架下合成数据结果的可视化。第一行的图比较了非参数分量 $\nu(t, Z)$ 的经验估计与其在 100 个留出点上的真实值，采用的是 PF方案。第二行的图采用 FN方案绘制，其含义与第一行相同。

5 实验

本节将报告 NFM 的实证表现，并重点探讨以下两个研究问题：

研究问题1（统计正确性验证）：本节列出的结果描述了理论上的收敛结果，为准确估计所需的样本数量提供了粗略参考。尽管如此，理论收敛速率往往较为保守，因此我们希望探究适量的样本是否足以实现良好的近似。

研究问题2（实证效能评估）：尽管 NFM 在估计准确性方面具有坚实的理论依据，但我们所建立的理论并不一定能保证其作为预后方法的实证效能。因此，有必要考察 NFM 在生存分析的现实世界预测任务中有多大用处。

5.1 合成实验

为回答研究问题1，我们开展合成实验以检验经验收敛情况。具体而言，我们考察在不同样本量下对底层真实参数的经验恢复情况。

真实值 我们将真实的底层模型设定为一个具有5维特征的非线性伽马脆弱性模型。我们生成了三个规模不同的训练数据集，其中 $n \in \{1000, 5000, 10000\}$。评估将在固定的测试样本上进行，该样本包含100个留出点，均从事件时间的生成机制中独立抽取。我们采用删失机制，使每个数据集的删失比例约为40%。脆弱性模型的具体形式以及特征向量的生成分布详见附录C.2。

经验恢复结果 我们在图1中报告了针对留出测试集的非线性分量 $\nu(t, Z)$ 的经验恢复结果。从图示中可以观察到，在样本量适中 $n = 1000$ 时，NFM 对特征空间（维度相对较低）的恢复效果已令人满意，而这正是大多数公开基准数据集的普遍情况。我们还提供了其他评估结果：（i）附录D.1中使用PF方案时对 $m(Z)$ 的恢复；（ii）附录D.2中使用PF方案和FN方案时对生存函数的恢复；（iii）附录D.3中生存函数的数值恢复结果。

表 1：在四个小规模生存数据集上，以 IBS 和 INBLL 指标（%）衡量的生存预测结果。每列中，粗体分数表示最佳结果，下划线分数表示次佳结果（均为均值）。

模型	METABRIC		RotGBSG		FLCHAIN		SUPPORT	
	IBS	积分负对数似然	IBS	积分负对数似然	IBS	积分负对数似然	IBS	积分负对数似然
Cox比例风险模型	$16.46_{\pm 0.90}$	$49.57_{\pm 2.66}$	$18.25_{\pm 0.44}$	$53.76_{\pm 1.11}$	$10.05_{\pm 0.38}$	$33.18_{\pm 1.16}$	$20.54_{\pm 0.38}$	$59.58_{\pm 0.86}$
GBM	$16.61_{\pm 0.82}$	$49.87_{\pm 2.44}$	$17.83_{\pm 0.44}$	$52.78_{\pm 1.11}$	$\underline{9.98}_{\pm 0.37}$	$\underline{32.88}_{\pm 1.05}$	$19.18_{\pm 0.39}$	$56.46_{\pm 0.10}$
RSF	$16.62_{\pm 0.64}$	$49.61_{\pm 1.54}$	$17.89_{\pm 0.42}$	$52.77_{\pm 1.01}$	$\mathbf{9.96}_{\pm 0.37}$	$32.92_{\pm 1.05}$	$\underline{19.11}_{\pm 0.40}$	$\underline{56.28}_{\pm 1.00}$
DeepSurv	$16.55_{\pm 0.93}$	$49.85_{\pm 3.02}$	$17.80_{\pm 0.49}$	$52.62_{\pm 1.25}$	$10.09_{\pm 0.38}$	$33.28_{\pm 1.15}$	$19.20_{\pm 0.41}$	$56.48_{\pm 1.08}$
CoxTime	$16.54_{\pm 0.83}$	$49.67_{\pm 2.67}$	$17.80_{\pm 0.58}$	$52.56_{\pm 1.47}$	$10.28_{\pm 0.45}$	$34.18_{\pm 1.53}$	$19.17_{\pm 0.40}$	$56.45_{\pm 1.10}$
DeepHit	$17.50_{\pm 0.83}$	$52.10_{\pm 2.16}$	$19.61_{\pm 0.38}$	$56.67_{\pm 1.10}$	$11.83_{\pm 0.39}$	$37.72_{\pm 1.02}$	$20.66_{\pm 0.32}$	$60.06_{\pm 0.72}$
DeepEH	$16.56_{\pm 0.65}$	$49.42_{\pm 1.53}$	$17.62_{\pm 0.52}$	$\underline{52.08}_{\pm 1.27}$	$10.11_{\pm 0.37}$	$33.30_{\pm 1.10}$	$19.30_{\pm 0.39}$	$56.67_{\pm 0.94}$
SuMo-net	$16.49_{\pm 0.83}$	$49.74_{\pm 2.21}$	$17.77_{\pm 0.47}$	$52.62_{\pm 1.11}$	$10.07_{\pm 0.40}$	$33.20_{\pm 1.10}$	$19.40_{\pm 0.38}$	$56.87_{\pm 0.96}$
SODEN	$16.52_{\pm 0.63}$	$49.39_{\pm 1.97}$	$\mathbf{17.05}_{\pm 0.63}$	$\mathbf{50.45}_{\pm 1.97}$	$10.13_{\pm 0.24}$	$33.37_{\pm 0.57}$	$19.07_{\pm 0.50}$	$56.15_{\pm 1.35}$
SurvNode	$16.67_{\pm 1.32}$	$49.73_{\pm 3.89}$	$17.42_{\pm 0.53}$	$51.70_{\pm 1.16}$	$10.40_{\pm 0.29}$	$34.37_{\pm 1.03}$	$19.58_{\pm 0.34}$	$57.49_{\pm 0.84}$
DCM	$16.58_{\pm 0.87}$	$49.48_{\pm 2.23}$	$17.66_{\pm 0.54}$	$52.26_{\pm 1.23}$	$10.13_{\pm 0.50}$	$33.40_{\pm 1.38}$	$19.29_{\pm 0.42}$	$56.68_{\pm 1.09}$
DeSurv	$16.71_{\pm 0.75}$	$49.61_{\pm 2.15}$	$17.98_{\pm 0.46}$	$53.23_{\pm 1.15}$	$10.06_{\pm 0.62}$	$33.18_{\pm 1.93}$	$19.50_{\pm 0.40}$	$57.28_{\pm 0.89}$
NFM-PF	$\underline{16.33}_{\pm 0.75}$	$\underline{49.07}_{\pm 1.96}$	$\underline{17.60}_{\pm 0.55}$	$52.12_{\pm 1.34}$	$\mathbf{9.96}_{\pm 0.39}$	$\mathbf{32.84}_{\pm 1.15}$	$19.14_{\pm 0.39}$	$56.35_{\pm 1.00}$
NFM-FN	$\mathbf{16.11}_{\pm 0.81}$	$\mathbf{48.21}_{\pm 2.04}$	$17.66_{\pm 0.52}$	$52.41_{\pm 1.22}$	$10.05_{\pm 0.39}$	$33.11_{\pm 1.10}$	$\mathbf{18.97}_{\pm 0.60}$	$\mathbf{55.87}_{\pm 1.50}$

5.2 真实世界数据实验

为回答研究问题2，我们在6个基准数据集上进行了广泛的经验评估，其中包括5个生存数据集和1个非生存数据集。生存数据集包括国际乳腺癌分子分类联盟（METABRIC） [16], 鹿特丹肿瘤库和德国乳腺癌研究组（RotGBSG）[43], 血清游离轻链检测（FLCHAIN） [20], 了解预后、偏好、治疗结局和风险的研究（SUPPORT） [43], 以及重症监护医学信息集市（MIMIC-III）[39]。对于所有生存数据集，感兴趣事件定义为入院后的死亡。在实验中，我们将METABRIC、RotGBSG、FLCHAIN和SUPPORT视为小规模数据集，将MIMIC-III视为中等规模数据集。此外，我们还使用KKBOX 数据集 [46]进行大规模评估。在该数据集中，如果客户从KKBOX平台流失，则会观测到事件时间。我们在表3中汇总了所有数据集的基本统计信息。

基线方法 我们将 NFM 与 12 种基线方法进行比较。第一种是线性Cox比例风险模型 [14]。梯度提升机（GBM） [27, 10] 和随机生存森林（RSF） [37] 是两种基于树的非参数生存回归方法。DeepSurv [42] 和 CoxTime [46] 是两种采用部分似然神经网络变体作为目标函数的模型。SuMo-net [56] 是 NHR 的神经网络变体。此外，我们还选择了六种最新的最先进的神经生存模型：DeepHit [47], 、SurvNode [32],、DeepEH [75], 、DCM [51], 、DeSurv [19] 和 SODEN [64]。在选定的基线方法中，DeepSurv 和 SuMo-net 被视为神经网络Cox比例风险模型和神经网络NHR 的实现，因此对于实证验证脆弱性的有效性尤其重要。

评估策略 我们使用生存预测中的两种标准度量来评估模型性能：积分布里尔分数 (IBS) 和积分负二项对数似然 (INBLL)。这两种度量均由以下公式导出：

$$\mathcal{S}(\ell, t_1, t_2) = \int_{t_2}^{t_1} \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\ell(0, \widehat{S}(t|Z_i)) I(T_i \le t, \delta_i = 1)\, \widehat{S}_C}{(T_i)} + \ell(1, \widehat{S}(t|Z_i)) I(T_i > t)\, \widehat{S}_C(t) \right] dt. \tag{11}$$

其中， $S_C(t)$ 是删失变量生存函数 $S_C(t)$ 的估计值，该估计值通过对测试数据中的删失观测进行 Kaplan-Meier 估计 [41] 得到。 ℓ： $\{0,1\} \times [0,1] \mapsto \mathbb{R}^+$ 是用于二分类的某种适当损失函数 [28]。IBS 度量对应于 ℓ 为平方损失，INBLL 度量对应于 ℓ 为负二项（伯努利）对数似然 [30]。如果删失时间与生存时间相互独立，则 IBS 和 INBLL 都是适当评分规则。 [5] 我们还在附录D中报告了另一种广泛使用的度量——一致性指数（C-index）的结果。由于所有生存数据集都没有标准的训练集/测试集

否则，可以针对删失时间建立协变量依赖模型，并使用 $\widehat{S}_C(t|Z)$ 代替$\widehat{S}_C(t)$。我们采用Kaplan-Meier方法，因为它仍是评估生存预测时的主流做法。

表 2：两个较大数据集上以 IBS 和 INBLL 指标（%）衡量的生存预测结果。每列中，粗体分数表示最佳结果，下划线分数表示次佳结果（均为均值）。未报告 SODEN 和 DeepEH 这两个模型，因为我们通过实验证实，它们的计算/内存开销明显高于其他模型，而且这两个模型在这两个数据集上的表现均不理想。

模型	MIMIC-III		KKBOX	
	IBS	积分负对数似然	IBS	积分负对数似然
Cox比例风险模型	$20.40_{\pm 0.00}$	$60.02_{\pm 0.00}$	$12.60_{\pm 0.00}$	$39.40_{\pm 0.00}$
GBM	$17.70_{\pm 0.00}$	$52.30_{\pm 0.00}$	$11.81_{\pm 0.00}$	$38.15_{\pm 0.00}$
RSF	$17.79_{\pm 0.19}$	$53.34_{\pm 0.41}$	$14.46_{\pm 0.00}$	$44.39_{\pm 0.00}$
DeepSurv	$18.58_{\pm 0.92}$	$55.98_{\pm 2.43}$	$11.31_{\pm 0.05}$	$35.28_{\pm 0.15}$
CoxTime	$17.68_{\pm 1.36}$	$52.08_{\pm 3.06}$	$\underline{10.70}_{\pm 0.06}$	$\underline{33.10}_{\pm 0.21}$
DeepHit	$19.80_{\pm 1.31}$	$59.03_{\pm 4.20}$	$16.00_{\pm 0.34}$	$48.64_{\pm 1.04}$
SuMo-net	$18.62_{\pm 1.23}$	$54.51_{\pm 2.97}$	$11.58_{\pm 0.11}$	$36.61_{\pm 0.28}$
DCM	$18.02_{\pm 0.49}$	$52.83_{\pm 0.94}$	$10.71_{\pm 0.11}$	$33.24_{\pm 0.06}$
DeSurv	$18.19_{\pm 0.65}$	$54.69_{\pm 2.83}$	$10.77_{\pm 0.21}$	$33.22_{\pm 0.10}$
NFM-PF	$\mathbf{16.28}_{\pm 0.36}$	$\mathbf{49.18}_{\pm 0.92}$	$11.02_{\pm 0.11}$	$35.10_{\pm 0.22}$
NFM-FN	$\underline{17.47}_{\pm 0.45}$	$\underline{51.48}_{\pm 1.23}$	$\mathbf{10.63}_{\pm 0.08}$	$\mathbf{32.81}_{\pm 0.14}$

划分，我们沿用先前采用五折交叉验证（CV）的做法 [75]：其中 1 折用于测试，其余数据中有 20% 留作验证。在实验中，我们发现，对大多数生存数据集而言，随机将数据划分为 5 折一次并不能得到稳定的结果。因此，我们对每个生存数据集进行 10 次不同的交叉验证，并报告平均度量及其标准差。对于 KKBOX 数据集，我们使用 pycox 软件包 [46] 提供的标准训练集/验证集/测试集划分，并报告基于 10 次试验运行的结果。

实验设置 我们采用标准的预处理策略 [42, 46, 75]，将连续特征标准化为均值为零、方差为一，并对所有类别^特征进行独热编码。在所有数据集上，我们均使用带 ReLU 激活函数的多层感知机作为函数逼近器，包括 h、PF方案中的 $\hat{m}$ 以及FN方案中的 $\hat{v}$；每层的层数（深度）和隐藏单元数（宽度）均可调。我们在几种标准选项中调整脆弱性变换：伽马脆弱性、Box-Cox变换脆弱性和 IGG(α) 脆弱性，其具体形式详见附录 C.3。有关调参流程的详细说明，以及基线模型的训练配置，见附录 C.3。

结果 我们在表 1 中报告小规模数据集的实验结果，并在表 2 中报告两个较大数据集的结果。所提出的 NFM 框架取得了有竞争力的表现，与其他最先进的模型相当。具体而言，NFM 在 6 个数据集中有 5 个的平均表现最佳，并且在 MIMIC-III 数据集上，以 0.05 的经验显著性水平显著优于所有基线方法。

脆弱性的消融分析 为了更好地理解引入脆弱性建模带来的额外益处，我们计算了 NFM-PF 和 NFM-FN 相较于其非脆弱性对应模型 DeepSurv [42] 和 SuMo-net [56]的（相对）性能提升。评估采用了本文提到的全部三项度量。结果见表 6。结果表明，引入脆弱性能够带来显著改进：两种 NFM模型的相对性能提升均可超过 10%。更详细的讨论见 D.5 节。

6 讨论与结论

我们介绍了 NFM，这是一种灵活而强大的生存分析神经建模框架，理论上具有统计正确性，并且在预测任务中表现出实证有效性。尽管我们提出的框架为神经生存建模提供了一种具有理论依据的方法，但仍有一些局限性和挑战需要在未来的工作中加以解决，包括预测保证和更完善的评估方案，详见附录D.6。

7 致谢

我们感谢应志良教授和方冠华教授给予的宝贵讨论。于文的研究获得国家自然科学基金委员会项目（12071088）的资助。郑明的研究获得国家自然科学基金委员会项目（12271106）的资助。

参考文献

[1] L. Antolini、P. Boracchi 和 E. Biganzoli。生存数据的时间依赖判别指数。《医学统计学》，24(24):3927–3944，2005年。[2] P. L. Bartlett、N. Harvey、C. Liaw 和 A. Mehrabian。分段线性神经网络的近乎紧致 VC 维数与伪维数界。《机器学习研究杂志》，20(1):2285–2301，2019年。[3] S. Bennett。采用比例优势模型分析生存数据。《医学统计学》，2(2):273–277，1983年。[4] P. J. Bickel、C. A. Klaassen、Y. Ritov、J. Klaassen、J. A. Wellner 和 Y. Ritov。半参数模型的有效与自适应估计，第4卷。Springer，1993年。[5] J. P. Boyd。切比雪夫与傅里叶谱方法。Courier Corporation，2001年。[6] P. Brennan。基因—环境相互作用与癌症病因学：这意味着什么，又该如何测量？《癌变》，23(3):381–387，2002年。[7] J. Buckley 和 I. James。含删失数据的线性回归。《生物计量学》，66(3):429–436，1979年。[8] J. Cai、J. Fan、J. Jiang 和 H. Zhou。多元生存数据的部分线性风险回归。《美国统计学会会刊》，102(478):538–551，2007年。[9] J. Cai、J. Fan、J. Jiang 和 H. Zhou。多元生存数据中具有变系数的部分线性风险回归。《英国皇家统计学会会刊：B辑（统计方法学）》，70(1):141–158，2008年。[10] T. Chen 和 C. Guestrin。XGBoost：一种可扩展的树提升系统。载于第22届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集，第785–794页，2016年。[11] X. Chen。半非参数模型的大样本筛估计。《计量经济学手册》，6:5549–5632，2007年。[12] X. Chen 和 X. Shen。弱相关数据的筛极值估计。《计量经济学》，第289–314页，1998年。[13] D. D. Cox 和 F. O'Sullivan。惩罚似然及相关估计量的渐近分析。《统计学年鉴》，第1676–1695页，1990年。[14] D. R. Cox。回归模型与生命表。《英国皇家统计学会会刊：B辑（方法论）》，34(2):187–202，1972年。[15] D. R. Cox。部分似然。《生物计量学》，62(2):269–276，1975年。[16] C. Curtis、S. P. Shah、S.-F. Chin、G. Turashvili、O. M. Rueda、M. J. Dunning、D. Speed、A. G. Lynch、S. Samarajiwa、Y. Yuan 等。2000例乳腺肿瘤的基因组与转录组架构揭示了新的亚组。《自然》，486(7403):346–352，2012年。[17] J. Cuzick。秩回归。《统计学年鉴》，第1369–1389页，1988年。[18] D. M. Dabrowska 和 K. A. Doksum。含删失数据的变换模型中的部分似然。《斯堪的纳维亚统计学杂志》，第1–23页，1988年。[19] D. Danks 和 C. Yau。用于生存分析的累积分布函数基于导数的神经建模。载于人工智能与统计国际会议论文集，第7240–7256页。PMLR，2022年。[20] A. Dispenzieri、J. A. Katzmann、R. A. Kyle、D. R. Larson、T. M. Therneau、C. L. Colby、R. J. Clark、G. P. Mead、S. Kumar、L. J. Melton III 等。利用非克隆性血清免疫球蛋白游离轻链预测一般人群的总生存期。载于《梅奥诊所学报》，第87卷，第517–523页。爱思唯尔，2012年。

[21] L. Duchateau 和 P. Janssen。脆弱性模型。Springer Science & Business Media，2007年。[22] D. Duffie、A. Eckner、G. Horel 和 L. Saita。脆弱性相关违约。《金融学杂志》，64(5):2089–2123，2009年。[23] J. Etezadi-Amoli 和 A. Ciampi。针对含协变量删失生存数据的扩展风险回归：基准风险函数的样条逼近。《生物统计学》，第181–192页，1987年。[24] J. Fan 和 I. Gijbels。《局部多项式建模及其应用：统计学与应用概率专著66》，第66卷。CRC出版社，1996年。[25] D. Faraggi 和 R. Simon。生存数据的神经网络模型。《医学统计学》，14(1):73–82，1995年。[26] M. H. Farrell、T. Liang 和 S. Misra。用于估计与推断的深度神经网络。《计量经济学》，89(1):181–213，2021年。[27] J. H. Friedman。贪婪函数逼近：梯度提升机。《统计年鉴》，第1189–1232页，2001年。[28] T. Gneiting 和 A. E. Raftery。严格适当评分规则、预测与估计。《美国统计学会会刊》，102(477):359–378，2007年。[29] I. Goodfellow、Y. Bengio 和 A. Courville。深度学习。2016年。[30] E. Graf、C. Schmoor、W. Sauerbrei 和 M. Schumacher。生存数据预后分类方案的评估与比较。《医学统计学》，18(17-18):2529–2545，1999年。[31] R. J. Gray。变换线性生存模型中回归参数和风险函数的估计。《生物统计学》，56(2):571–576，2000年。[32] S. Groha、S. M. Schmon 和 A. Gusev。生存分析与多状态建模的通用框架。arXiv预印本 arXiv:2006.04893，2020年。[33] X. Han、M. Goldstein、A. Puli、T. Wies、A. Perotte 和 R. Ranganath。逆加权生存博弈。《神经信息处理系统进展》，34:2160–2172，2021年。[34] M. Hardt、B. Recht 和 Y. Singer。加快训练、提升泛化：随机梯度下降的稳定性。载于国际机器学习会议论文集，第1225–1234页。PMLR，2016年。[35] P. Hougaard。异质群体的生命表方法：描述异质性的分布。《生物计量学》，71(1):75–83，1984年。[36] J. Huang。部分线性加性Cox模型的有效估计。《统计学年鉴》，27(5):1536–1563，1999年。[37] H. Ishwaran、U. B. Kogalur、E. H. Blackstone 和 M. S. Lauer。随机生存森林。《应用统计年鉴》，2(3):841–860，2008年。[38] H. Jing 和 A. J. Smola。神经生存推荐系统。载于第十届ACM 国际网络搜索与数据挖掘会议论文集，第515–524页，2017年。[39] A. E. Johnson、T. J. Pollard、L. Shen、L.-w. H. Lehman、M. Feng、M. Ghassemi、B. Moody、P. Szolovits、L. Anthony Celi 和 R. G. Mark。MIMIC-III：一个可免费访问的重症监护数据库。《科学数据》，3(1):1–9，2016年。[40] J. D. Kalbfleisch 和 R. L. Prentice。《失效时间数据的统计分析》，第360卷。约翰·威立父子出版公司，2002年。[41] E. L. Kaplan 和 P. Meier。基于不完整观测的非参数估计。《美国统计学会会刊》，53(282):457–481，1958年。[42] J. L. Katzman、U. Shaham、A. Cloninger、J. Bates、T. Jiang 和 Y. Kluger。DeepSurv：利用Cox比例风险深度神经网络构建个性化治疗推荐系统。《BMC医学研究方法学》，18(1):1–12，2018年。[43] W. A. Knaus、F. E. Harrell、J. Lynn、L. Goldman、R. S. Phillips、A. F. Connors、N. V. Dawson、W. J. Fulkerson、R. M. Califf、N. Desbiens 等。SUPPORT预后模型：住院重症成人患者生存期的客观估计。《内科学年鉴》，122(3):191–203，1995年。

[44] C. Kooperberg、C. J. Stone 和 Y. K. Truong。风险回归。《美国统计学会会刊》，90(429):78–94，1995年。[45] M. R. Kosorok、B. L. Lee 和 J. P. Fine。单变量比例风险脆弱性回归模型的稳健推断。《统计学年鉴》，32(4):1448–1491，2004年。[46] H. Kvamme、Ø. Borgan 和 I. Scheel。利用神经网络和Cox回归预测事件发生时间。arXiv预印本 arXiv:1907.00825，2019年。[47] C. Lee、W. Zame、J. Yoon 和 M. Van Der Schaar。DeepHit：一种用于竞争风险生存分析的深度学习方法。载于《AAAI人工智能大会论文集》，第32卷，2018年。[48] S. A. Murphy。包含随机效应的比例风险模型的一致性。《统计学年鉴》，22(2):712–731，1994年。[49] S. A. Murphy。脆弱性模型的渐近理论。《统计学年鉴》，第182–198页，1995年。[50] S. A. Murphy 和 A. W. Van der Vaart。论剖面似然。《美国统计学会会刊》，95(450):449–465，2000年。[51] C. Nagpal、S. Yadlowsky、N. Rostamzadeh 和 K. Heller。用于生存回归的深度Cox混合模型。载于医疗保健机器学习会议，第674–708页。PMLR，2021年。[52] T. Omi、n. ueda 和 K. Aihara。完全基于神经网络的一般时间点过程模型。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alché-Buc、E. Fox 和 R. Gar- nett 编，《神经信息处理系统进展》，第32卷。Curran Associates, Inc.，2019年。[53] E. Parner。相关伽马脆弱性模型的渐近理论。《统计学年鉴》，26(1):183–214，1998年。[54] S. Pölsterl。scikit-survival：一个基于 scikit-learn 构建的事件发生时间分析库。机器学习研究杂志，21(212):1–6，2020年。[55] S. Purushotham、C. Meng、Z. Che 和 Y. Liu。大型医疗保健数据集上的深度学习模型基准测试。《生物医学信息学杂志》，83:112–134，2018年。[56] D. Rindt、R. Hu、D. Steinsaltz 和 D. Sejdinovic。利用适当评分规则和单调神经网络进行生存回归。机器学习研究论文集。《机器学习研究杂志》，2022年。[57] J. Schmidt-Hieber。使用 ReLU 激活函数的深度神经网络进行非参数回归。《统计学年鉴》，48(4):1875–1897，2020年。[58] S. Shalev-Shwartz 和 S. Ben-David。《理解机器学习：从理论到算法》。剑桥大学出版社，2014年。[59] X. Shen。论筛选方法与惩罚法。《统计学年鉴》，25(6):2555–2591，1997年。[60] X. Shen 和 W. H. Wong。筛选估计的收敛率。《统计学年鉴》，第580–615页，1994年。[61] D. M. Stablein 和 I. Koutrouvelis。一种对未删失数据和单删失数据中的交叉风险敏感的双样本检验。《生物统计学》，第643–652页，1985年。[62] H. Steck、B. Krishnapuram、C. Dehing-oberije、P. Lambin 和 V. C. Raykar。论生存分析中的排序：一致性指数的界。载于 J. Platt、D. Koller、Y. Singer 和 S. Roweis 编，《神经信息处理系统进展》，第20卷。Curran Associates, I nc.，2007年。[63] R. L. Strawderman 和 A. A. Tsiatis。论灵活风险估计量的渐近性质。《统计学年鉴》，24(1):41–63，1996年。[64] W. Tang、J. Ma、Q. Mei 和 J. Zhu。SODEN：通过常微分方程网络构建的可扩展连续时间生存模型。机器学习研究杂志，23(34):1–29，2022年。[65] A. Tsybakov。《非参数估计导论》。施普林格统计学丛书。施普林格纽约，2008年。

[66] A. van der Vaart、A. van der Vaart、A. van der Vaart 和 J. Wellner。《弱收敛与经验过程：及其在统计学中的应用》。施普林格统计学丛书。Springer，1996年。[67] A. W. Van der Vaart。《渐近统计学》，第3卷。剑桥大学出版社，2000年。[68] A. Wehenkel 和 G. Louppe。无约束单调神经网络。《神经信息处理系统进展》，32，2019年。[69] A. Wienke。《生存分析中的脆弱性模型》。Chapman and Hall/CRC，2010年。[70] W. H. Wong 和 X. Shen。似然比的概率不等式与筛极大似然估计的收敛率。《统计学年鉴》，第339–362页，1995年。[71] W. C.-H. Wu、M.-Y. Yeh 和 M.-S. Chen。利用删失数据预测实时竞价中的获胜价格。载于第21届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集，第1305–1314页，2015年。[72] D. Yarotsky。深度ReLU网络近似的误差界。《神经网络》，94:103–114，2017年。[73] Z. Ying。删失回归数据秩估计的大样本研究。《统计学年鉴》，第76–99页，1993年。[74] D. Zeng 和 D. Lin。计数过程半参数变换模型的有效估计。《生物计量学》，93(3):627–640，2006年。[75] Q. Zhong、J. W. Mueller 和 J.-L. Wang。用于生存分析的深度扩展风险模型。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P. Liang 和 J. W. Vaughan 编，《神经信息处理系统进展》，第34卷，第15111–15124页。Curran Associates, Inc.，2021年。[76] Q. Zhong、J. W. Mueller 和 J.-L. Wang。部分线性Cox模型的深度学习。《统计学年鉴》，2021年。

A 脆弱性设定示例

我们列出几种常用的脆弱性模型，并通过其对应的脆弱性变换 G_θ说明各自的特征：

伽马脆弱性：可以说，伽马脆弱性是应用最广泛的脆弱性模型 [48, 49, 53, 69,21], ，其

$$G_\theta(x) = \frac{1}{\theta}\log(1+\theta x), \theta \geq 0. \tag{12}$$

当 $\theta = 0$、 $G_0(x) = \lim_{\theta\to 0} G_\theta(x)$ 被定义为（逐点）极限时。伽马脆弱性设定的一个显著特点是，当满足比例脆弱性（PF）假设 (2) 时，若 $\theta = 0$，该模型退化为 Cox比例风险模型。否则，若 $\theta = 1$，该模型对应于比例优势（PO）模型 [3]。

Box-Cox 变换脆弱性：在此设定下，我们有

$$G_\theta(x) = \frac{(1+x)^\theta - 1}{\theta}, \theta \geq 0. \tag{13}$$

与伽马脆弱性的情形类似， $\theta = 0$ 的情形也以类似方式定义；在 PF 假设下，它对应于 PO 模型。当 $\theta = 1$ 时，在 PF 假设下，该模型简化为 Cox比例风险模型。

IGG(α) 脆弱性：这是对伽马脆弱性的扩展 [45] ，并包含其他类型的脆弱性设定，例如逆高斯脆弱性 [35],

$$G_\theta(x) = \frac{1-\alpha}{\alpha\theta}\left[\left(1+\frac{\theta x}{1-\alpha}\right)^\alpha - 1\right], \theta \geq 0, \alpha \in [0,1). \tag{14}$$

在一维参数范式中，假设参数 α 已知，而非可学习。当 $\alpha = 1/2$时，我们得到伽马脆弱性模型。当 $\alpha \to 0$时，其极限对应于逆高斯脆弱性。

正则性条件 4.5 的可满足性 在 [45, 命题 1], 中，作者验证了伽马脆弱性和 IGG(α) 脆弱性变量的正则性。采用类似的论证，可以直接验证 Box-Cox变换脆弱性的正则性。

B 定理证明

B.1 初步内容

补充定义 经验过程理论 [66] 将在证明中大量使用。因此，我们简要介绍一些常用记号：对于函数类 $\mathcal{F}$，定义$N(\epsilon, \mathcal{F}, \|\cdot\|)$为 $\mathcal{F}$关于范数 $\|\cdot\|$、半径为 ϵ时的覆盖数，并定义$N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$为 $\mathcal{F}$关于范数 $\|\cdot\|$、半径为 ϵ时的括号覆盖数。我们用 VC ($\mathcal{F}$) 表示 $\mathcal{F}$的 VC 维数。此外，我们用记号 $a \lesssim b$ 表示 $a \leq Cb$，其中 C为某个正的常数。

在证明定理4.6和定理4.7之前，我们先引入一些在整个证明过程中会用到的其他符号。

在PF方案中，定义

$$\begin{aligned} l(T,\delta,Z;h,m,\theta) =& \delta \log g_\theta\left(e^{m(Z)}\int_0^T e^{h(s)}ds\right) + \delta h(T) + \delta m(Z) \\ & - G_\theta\left(e^{m(Z)}\int_0^T e^{h(s)}ds\right), \end{aligned}$$

其中我们将 $g_\theta = G'(\theta)$记为.....。根据条件4.3中给出的筛空间定义，我们将参数估计值重新表述为

$$\left(\widehat{h}_n, \widehat{m}_n, \widehat{\theta}_n\right) = \underset{\widehat{h}\in\mathcal{H}_n, \widehat{m}\in\mathcal{M}_n, \theta\in\Theta}{\operatorname{argmax}} \frac{1}{n}\sum_{i\in[n]} l(T_i,\delta_i,Z_i;\widehat{h},\widehat{m},\theta).$$

类似地，在FN方案中，我们定义

$$l(T,\delta,Z;\nu,\theta)=\delta\log g_\theta\left(\int_0^T e^{\nu(s,Z)}ds\right)+\delta\nu(T,Z)-G_\theta\left(\int_0^T e^{\nu(s,Z)}ds\right)$$

根据条件 4.4 中给出的筛空间定义，我们将参数估计重述如下：

$$\left(\widehat{\nu}_n(t,z),\widehat{\theta}_n\right)=\underset{\widehat{\nu}\in\mathcal{V}_n,\theta\in\Theta}{\operatorname{argmax}}\frac{1}{n}\sum_{i\in[n]}l(T_i,\delta_i,Z_i;\widehat{\nu},\theta).$$

~我们分别用$f_{\tilde{T}|Z}(t)$和$S_{\tilde{T}|Z}(t)$表示事件时间 T 在给定 Z 时的条件密度函数和生存函数。同样，我们用 $f_{C|Z}(t)$和 $S_{C|Z}(t)$分别表示删失时间 C 在给定 Z 时的条件密度函数和生存函数。在假设$T \perp\!\!\!\perp C \mid Z$的条件下，观测时间 T和删失指标 δ 在给定Z 时的联合条件密度可表示如下：

$$\begin{aligned}p(T,\delta\mid Z) &= f_{\tilde{T}|Z}(T)^\delta S_{\tilde{T}|Z}(T)^{1-\delta}f_{C|Z}(T)^{1-\delta}S_{C|Z}(T)^\delta\\ &= \lambda_{\tilde{T}|Z}(T)^\delta S_{\tilde{T}|Z}(T)f_{C|Z}(T)^{1-\delta}S_{C|Z}(T)^\delta,\end{aligned}$$

~其中， $\lambda_{\tilde{T}|Z}(T)$是生存时间 T在给定 Z时的条件风险函数。

在 PF 方案的模型假设下， $p(T,\delta\mid Z)$可表示为

$$p(T,\delta\mid Z;h,m,\theta)=\exp\left(l(T,\delta,Z;h,m,\theta)\right)f_{C|Z}(T)^{1-\delta}S_{C|Z}(T)^\delta.$$

对于 $\phi_0=(h_0,m_0,\theta_0)$和估计量 $\widehat{\phi}=(\widehat{h},\widehat{m},\widehat{\theta})$，所定义的距离 $d_{\mathsf{PF}}\left(\widehat{\phi},\phi_0\right)$ 可明确表示为

$$d_{\mathsf{FN}}\left(\widehat{\psi},\psi_0\right)=\sqrt{\mathbb{E}_Z\left[\int\left|\sqrt{p(T,\delta\mid Z;\widehat{h},\widehat{m},\widehat{\theta})}-\sqrt{p(T,\delta\mid Z;h_0,m_0,\theta_0)}\right|^2\mu(dT\times d\delta)\right]}.$$

此处的支配测度 μ 定义为：对于任意（可测）函数 $r(T,\delta)$

$$\int r(T,\delta)\mu(dT\times d\delta)=\int_0^\tau r(T,\delta=1)dT+\int_0^\tau r(T,\delta=0)dT$$

在 FN方案的模型假设下， $p(T,\delta\mid Z)$ 可表示为

$$p(T,\delta\mid Z;\nu,\theta)=\exp\left(l(T,\delta,Z;\nu,\theta)\right)f_{C|Z}(T)^{1-\delta}S_{C|Z}(T)^\delta.$$

对于 $\psi_0=(\nu_0,\theta_0)$ 和估计量 $\widehat{\psi}=(\widehat{\nu},\widehat{\theta})$，定义的距离 $d_{\mathsf{FN}}\left(\widehat{\psi},\psi_0\right)$ 可明确表示为

$$d_{\mathsf{FN}}\left(\widehat{\psi},\psi_0\right)=\sqrt{\mathbb{E}_Z\left[\int\left|\sqrt{p(T,\delta\mid Z;\widehat{\nu},\widehat{\theta})}-\sqrt{p(T,\delta\mid Z;\nu_0,\theta_0)}\right|^2\mu(dT\times d\delta)\right]}.$$

B.2 技术引理

以下引理将用于定理4.6和4.7的证明。此后，为了记号上的^方便，我们将用 h、$\widehat{m}$ 表示条件4.3中所列相应筛空间里的任意元素，用 $\widehat{\nu}$ 表示条件 4.4 中所列筛空间里的任意元素，并用 θ 表示 Θ 中的任意元素。

引理 B.1. 在条件 4.1、4.3、4.5 下，对于 $(T,\delta,Z)\in[0,\tau]\times\{0,1\}\times[-1,1]^d$，以下各项有界：

1. $l(T,\delta,Z;\ h_0,m_0,\theta_0)$，其真实参数为 (h_0,m_0,θ_0)

2. $l(T,\delta,Z;\ h,\widehat{m},\theta)$，其参数估计值（$h,\widehat{m},\theta$）属于条件4.3中列出的任意筛空间。

引理 B.2. 在条件 4.2、4.4、4.5 下，对于（T,δ,Z）$\in[0,\tau]\times\{0,1\}\times[-1,1]^d$，以下各项有界：

1. $l(T,\delta,Z;\ \nu_0,\theta_0)$，其真实参数为（$\nu_0,\theta_0$）

2. $l(T,\delta,Z;\ \widehat{\nu},\theta)$，其参数估计值（$\widehat{\nu},\theta$）属于条件 4.4 中列出的任意筛空间。

引理 B.3. 在条件 4.1、4.3、4.5 下，设（$h,\widehat{m},\theta$）、（$h_1,\widehat{m}_1,\theta_1$）和（$h_2,\widehat{m}_2,\theta_2$）是条件4.3所定义的筛空间内任意三个参数三元组，则以下两个不等式成立。

$$\|l(T,\delta,Z;h_0,m_0,\theta_0)-l(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})\|_\infty\lesssim|\theta_0-\widehat{\theta}|+\|h_0-\widehat{h}\|_\infty+\|m_0-\widehat{m}\|_\infty$$
$$\|l(T,\delta,Z;\widehat{h}_1,\widehat{m}_1,\widehat{\theta}_1)-l(T,\delta,Z;\widehat{h}_2,\widehat{m}_2,\widehat{\theta}_2)\|_\infty\lesssim|\widehat{\theta}_1-\widehat{\theta}_2|+\|\widehat{h}_1-\widehat{h}_2\|_\infty+\|\widehat{m}_1-\widehat{m}_2\|_\infty.$$

引理 B.4。在条件 4.2、4.4、4.5 下，设 $(\widehat{\nu},\theta)$、$(\widehat{\nu}_1,\theta_1)$ 和 $(\widehat{\nu}_2,\theta_2)$ 是条件 4.4 中定义的筛空间内任意三个参数组，则以下不等式成立。

$$\|l(T,\delta,Z;\nu_0,\theta_0)-l(T,\delta,Z;\widehat{\nu},\widehat{\theta})\|_\infty\lesssim|\theta_0-\widehat{\theta}|+\|\nu_0-\widehat{\nu}\|_\infty$$
$$\|l(T,\delta,Z;\widehat{\nu}_1,\widehat{\theta}_1)-l(T,\delta,Z;\widehat{\nu}_2,\widehat{\theta}_2)\|_\infty\lesssim|\widehat{\theta}_1-\widehat{\theta}_2|+\|\widehat{\nu}_1-\widehat{\nu}_2\|_\infty.$$

引理 B.5（PF方案的逼近误差）。在 PF方案中，对于任意 n，在相应的筛空间 $\pi_n\phi_0=(\pi_nh_0,\pi_nm_0,\pi_n\theta_0)$ 中存在一个元素，使得 $d_{PF}(\pi_n\phi_0,\phi_0)=O\left(n^{-\frac{\beta}{\beta+d}}\right)$。

引理 B.6（FN方案的逼近误差）。在 FN方案中，对于任意 n，在相应的筛空间 $\pi_n\psi=(\pi_n\nu_0,\pi_n\theta_0)$ 中存在一个元素，使得 $d_{FN}(\pi_n\psi_0,\psi_0)=O\left(n^{-\frac{\beta}{\beta+d+1}}\right)$。

引理 B.7。假设 $\mathcal{F}$是一个函数类，满足对于 $\forall\varepsilon>0$，有 $N(\varepsilon,\mathcal{F},\|\cdot\|)<\infty$。我们将 $N(\varepsilon,\mathcal{F},\|\cdot\|)$ 定义为覆盖 $\mathcal{F}$所需的 ε-球 $B(f,\varepsilon)=\{g$的最小数量：$\|g-f\|<\varepsilon\}$，其半径为 ε，并进一步要求满足 $f\in\mathcal{F}$。那么我们有

$$N(\varepsilon,\mathcal{F},\|\cdot\|)\le\widetilde{N}(\varepsilon,\mathcal{F},\|\cdot\|)\le N(\frac{\varepsilon}{2},\mathcal{F},\|\cdot\|).$$

引理 B.8。假设 $\mathcal{F}$是一个函数类，满足对于 $\forall\varepsilon>0$，有 $N_{[]}(\varepsilon,\mathcal{F},\|\cdot\|_\infty)<\infty$。我们将 $N_{[]}(\varepsilon,\mathcal{F},\|\cdot\|_\infty)$定义为覆盖 $\mathcal{F}$所需的括号 $[l,u]$ 的最小数量，并进一步要求满足$\|l-u\|_\infty<\varepsilon$、$f\in\mathcal{F}$和 $l=f-\frac{\varepsilon}{2}$。那么我们有

$$N_{[]}(\varepsilon,\mathcal{F},\|\cdot\|_\infty)\le\widetilde{N}_{[]}(\varepsilon,\mathcal{F},\|\cdot\|_\infty)\le N_{[]}(\frac{\varepsilon}{2},\mathcal{F},\|\cdot\|_\infty)$$

此外，我们有 $N_{[]}(\varepsilon,\mathcal{F},\|\cdot\|_\infty)=N(\frac{\varepsilon}{2},\mathcal{F},\|\cdot\|_\infty)$。

引理 B.9（PF方案的模型容量）。设 $\mathcal{F}_n=\{l(T,\delta,Z;\ h,\widehat{m},\theta):\ h\in\mathcal{H}_n,\widehat{m}\in\mathcal{M}_n,\widehat{\theta}\in\Theta\}$。在条件 4.5 下，取 $s_h=\frac{2\beta}{2\beta+1}$和 $s_m=\frac{2\beta}{2\beta+d}$，存在常数 c_h和 $c_m>0$，使得

$$N_{[]}(\varepsilon,\mathcal{F}_n,\|\cdot\|_\infty)\lesssim\frac{1}{\varepsilon}N(c_h\varepsilon^{1/s_h},\mathcal{H}_n,\|\cdot\|_2)\times N(c_m\varepsilon^{1/s_m},\mathcal{M}_n,\|\cdot\|_2).$$

引理 B.10（FN方案的模型容量）。设 $\mathcal{G}_n=\{l(T,\delta,Z;\widehat{\nu},\theta):\ \widehat{\nu}\in\mathcal{V}_n,\theta\in\Theta\}$。在条件 4.5 下，取 $s_\nu=\frac{2\beta}{2\beta+d+1}$，存在常数 $c_\nu>0$，使得

$$N_{[]}(\varepsilon,\mathcal{G}_n,\|\cdot\|_\infty)\lesssim\frac{1}{\varepsilon}N(c_\nu\varepsilon^{1/s_\nu},\mathcal{V}_n,\|\cdot\|_2).$$

B.3 定理4.6和4.7的证明

定理4.6的证明。证明分为四个步骤。

步骤1 我们记 $\phi_0 = (h_0, m_0, \theta_0)$和 $\phi = (h, \widehat{m}, \theta)$，其中 $h \in \mathcal{H}_n$、$\widehat{m} \in \mathcal{M}_n$ 和 $\theta \in \Theta$。对于任意小的 $\varepsilon > 0$，我们有

$$\begin{aligned}
&\inf_{d_{\mathsf{PF}}(\widehat{\phi},\phi_0)\geq\varepsilon} \mathbb{E}\left[l(T,\delta,Z;h_0,m_0,\theta_0) - l(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})\right] \\
&= \inf_{d_{\mathsf{PF}}(\widehat{\phi},\phi_0)\geq\varepsilon} \mathbb{E}_Z\left[\mathbb{E}_{T,\delta|Z}\left[\log p(T,\delta \mid Z;h_0,m_0,\theta_0) - \log p(T,\delta \mid Z;\widehat{h},\widehat{m},\widehat{\theta})\right]\right] \\
&= \inf_{d_{\mathsf{PF}}(\widehat{\phi},\phi_0)\geq\varepsilon} \mathbb{E}_Z\left[\mathrm{D}_{\mathrm{KL}}\left(\mathbb{P}_{\widehat{\phi},Z} \parallel \mathbb{P}_{\phi_0,Z}\right)\right]
\end{aligned}$$

利用 $\mathrm{D}_{\mathrm{KL}}\left(\mathbb{P}_{\widehat{\phi},Z} \parallel \mathbb{P}_{\phi_0,Z}\right) \geq 2H^2(\mathbb{P}_{\widehat{\phi},\ Z} \parallel \mathbb{P}_{\phi_0,\ Z})$ 这一事实。因此，我们进一步得到

$$\begin{aligned}
&\inf_{d_{\mathsf{PF}}(\widehat{\phi},\phi_0)\geq\varepsilon} \mathbb{E}\left[l(T,\delta,Z;h_0,m_0,\theta_0) - l(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})\right] \\
&\geq \inf_{d_{\mathsf{PF}}(\widehat{\phi},\phi_0)\geq\varepsilon} 2\mathbb{E}_Z\left[H^2(\mathbb{P}_{\widehat{\phi},Z} \parallel \mathbb{P}_{\phi_0,Z})\right] \\
&= 2\inf_{d_{\mathsf{PF}}(\widehat{\phi},\phi_0)\geq\varepsilon} d^2_{\mathsf{PF}}\left(\widehat{\phi},\phi_0\right) \\
&\geq 2\varepsilon^2.
\end{aligned}$$

步骤 2 考虑以下推导。

$$\begin{aligned}
&\sup_{d_{\mathsf{PF}}(\widehat{\phi},\phi_0)\leq\varepsilon} \mathrm{Var}\left[l(T,\delta,Z;h_0,m_0,\theta_0) - l(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})\right] \\
&\leq \sup_{d_{\mathsf{PF}}(\widehat{\phi},\phi_0)\leq\varepsilon} \mathbb{E}\left[\left(l(T,\delta,Z;h_0,m_0,\theta_0) - l(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})\right)^2\right] \\
&= \sup_{d_{\mathsf{PF}}(\widehat{\phi},\phi_0)\leq\varepsilon} \mathbb{E}_Z\mathbb{E}_{T,\delta|Z}\left[\log p(T,\delta,Z;h_0,m_0,\theta_0) - \log p(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})\right]^2 \\
&= 4\sup_{d_{\mathsf{PF}}(\widehat{\phi},\phi_0)\leq\varepsilon} \mathbb{E}_Z\left[\int\left(p(T,\delta,Z;h_0,m_0,\theta_0)\left(\log\sqrt{\frac{p(T,\delta,Z;h_0,m_0,\theta_0)}{p(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})}}\right)^2\right)\mu(dT\times d\delta)\right]
\end{aligned}$$

对 $\log\ x$ 进行泰勒展开，可知存在位于 $p^{\frac{1}{2}}(T,\delta,Z;\ h_0,m_0,\theta_0)$ 与$p^{\frac{1}{2}}(T,\delta,Z;\ \widehat{h},\widehat{m},\widehat{\theta})$ 之间的 $\xi(T,\delta,Z)$，使其逐点满足

$$\begin{aligned}
&p(T,\delta,Z;h_0,m_0,\theta_0)\left(\log\sqrt{\frac{p(T,\delta,Z;h_0,m_0,\theta_0)}{p(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})}}\right)^2 \\
&= p(T,\delta,Z;h_0,m_0,\theta_0)\left(\log\sqrt{p(T,\delta,Z;h_0,m_0,\theta_0)} - \log\sqrt{p(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})}\right)^2 \\
&= \frac{p(T,\delta,Z;h_0,m_0,\theta_0)}{\xi(T,\delta,Z)^2}\left(\sqrt{p(T,\delta,Z;h_0,m_0,\theta_0)} - \sqrt{p(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})}\right)^2
\end{aligned}$$

由于

$$\frac{p(T,\delta,Z;h_0,m_0,\theta_0)}{p(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})} = e^{l(T,\delta,Z;h_0,m_0,\theta_0)-l(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})}$$

根据引理 B.1，$l(T,\delta,Z;\ h_0,m_0,\theta_0)$ 和 $l(T,\delta,Z;\ h,\widehat{m},\theta)$ 在$[0,\tau]\times\{0,1\}\times[-1,1]^d$ 上对所有 $\widehat{\phi} = (\widehat{h},\widehat{m},\widehat{\theta})$ 一致有界。因此，存在常数 C_1 和 C_2 使得$0 < C_1 \leq p(T,\delta,Z;\ h_0,m_0,\theta_0)/p(T,\delta,Z;\ h,\widehat{m},\theta) \leq C_2$。由此可知

$p(T,\delta,Z;\ h_0,m_0,\theta_0)$ $\frac{1}{\xi\ T\ \delta\ (\cdot,\cdot,Z)^2}$ 有界。我们进一步得到

$$
\begin{aligned}
& p(T,\delta,Z;h_0,m_0,\theta_0)\left(\log\sqrt{p(T,\delta,Z;h_0,m_0,\theta_0)}-\log\sqrt{p(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})}\right)^2 \\
\lesssim & \left|\sqrt{p(T,\delta,Z;h_0,m_0,\theta_0)}-\sqrt{p(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})}\right|^2 .
\end{aligned}
$$

因此，我们有

$$
\begin{aligned}
& \sup_{d_{\mathsf{PF}}\left[\widehat{\phi},\phi_0\right]\leq\varepsilon}\ \mathrm{Var}(l(T,\delta,Z;h_0,m_0,\theta_0)-l(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})) \\
\lesssim & \sup_{d_{\mathsf{PF}}\left(\widehat{\phi},\phi_0\right)\leq\varepsilon}\mathbb{E}_Z\left[\int\left|\sqrt{p(T,\delta,Z;h_0,m_0,\theta_0)}-\sqrt{p(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})}\right|^2\mu(dT\times d\delta)\right] \\
= & \sup_{d_{\mathsf{PF}}\left(\widehat{\phi},\phi_0\right)\leq\varepsilon}d^2_{\mathsf{PF}}\left(\widehat{\phi},\phi_0\right) \\
\leq & \varepsilon^2 .
\end{aligned}
$$

~ ^ ^ ^步骤 3 我们定义 $\mathcal{F}_n=\{l(T,\delta,Z;\ h,\widehat{m},\theta)$ $-\ l(T,\delta,Z;\ \pi_n h_0,\pi_n m_0,\pi_n\theta_0):\ h\in\mathcal{H}_n,\widehat{m}\in\mathcal{M}_n,\theta\in\Theta\}$。这里的 $(\pi_n h_0,\pi_n m_0,\pi_n\theta_0)$ 已在引理 B.5 中定义。显然，我们有~log $N_{[]}(\varepsilon,\mathcal{F}_n,\|\cdot\|_\infty)$ =log $N_{[]}(\varepsilon,\mathcal{F}_n,\|\cdot\|_\infty)$，其中 $\mathcal{F}$ 在引理 B.9 中定义。根据引理 B.9，我们进一步有

$$
\log N_{[]}(\varepsilon,\mathcal{F}_n,\|\cdot\|_\infty)\lesssim\log\frac{1}{\varepsilon}+\log N(c_h\varepsilon^{1/s_h},\mathcal{H}_n,\|\cdot\|_2)+\log N(c_m\varepsilon^{1/s_m},\mathcal{M}_n,\|\cdot\|_2).
$$

根据条件4.3下的 [2, 定理 7], , 我们有， $\mathcal{H}_n$ 和 $\mathcal{M}_n$的 VC 维数满足 VC $(\mathcal{H}_n)$ $\lesssim n^{\frac{1}{\beta+d}}\log^3 n$ 且 VC $(\mathcal{M}_n)$ $\lesssim n^{\frac{d}{\beta+d}}\log^3 n$。因此，我们得到

$$
\log N(c_h\varepsilon^{1/s_h},\mathcal{H}_n,\|\cdot\|_2)\lesssim\frac{\mathrm{VC}\,(\mathcal{H}_n)}{s_h}\log\frac{1}{\varepsilon}\lesssim n^{\frac{1}{\beta+d}}\log^3 n\log\frac{1}{\varepsilon},
$$

and

$$
\log N(c_m\varepsilon^{1/s_m},\mathcal{M}_n,\|\cdot\|_2)\lesssim\frac{\mathrm{VC}\,(\mathcal{M}_n)}{s_\nu}\log\frac{1}{\varepsilon}\lesssim n^{\frac{d}{\beta+d}}\log^3 n\log\frac{1}{\varepsilon}.
$$

因此，我们得到 log $N_{[]}(\varepsilon,\widetilde{\mathcal{F}}_n,\|\cdot\|_\infty)$ $\lesssim n^{\frac{d}{\beta+d}}\log^3 n\log$ $\frac{1}{\varepsilon}$。

步骤 4 根据柯西-施瓦茨不等式，我们有

$$
\begin{aligned}
& \sqrt{\mathbb{E}\left[l(T,\delta,Z;\pi_n h_0,\pi_n m_0,\pi_n\theta_0)-l(T,\delta,Z;h_0,m_0,\theta_0)\right]} \\
\leq & \left[\mathbb{E}(l(T,\delta,Z;\pi_n h_0,\pi_n m_0,\pi_n\theta_0)-l(T,\delta,Z;h_0,m_0,\theta_0))^2\right]^{\frac{1}{4}} .
\end{aligned}
$$

与第二部分类似，并根据引理 B.5，我们进一步得到

$\sqrt{\mathbb{E}\left[l(T,\delta,Z;\ \pi_n h_0,\pi_n m_0,\pi_n\theta_0)\ -l(T,\delta,Z;\ h_0,m_0,\theta_0)\right]}\lesssim\sqrt{d_{\mathsf{PF}}\left(\pi_n\phi_0,\phi_0\right)}\lesssim n^{-\frac{\beta}{2\beta+2d}}$.

现在令

$$
\tau=\frac{\beta}{2\beta+2d}-2\frac{\log\log n}{\log n}
$$

根据步骤 1、2、3 和 [60, 定理 1], , 我们有

$$
\begin{aligned}
d_{\mathsf{PF}}\left(\widehat{\phi}_n,\phi_0\right)=\max\big(& n^{-\tau},d_{\mathsf{PF}}\left(\pi_n\phi_0,\phi_0\right), \\
& \sqrt{\mathbb{E}\left[l(T,\delta,Z;\pi_n h_0,\pi_n m_0,\pi_n\theta_0)-l(T,\delta,Z;h_0,m_0,\theta_0)\right]}\big)
\end{aligned}
$$

根据引理 B.5， $d_{\mathsf{PF}}\left(\pi_n\phi_0,\phi_0\right)=O(n^{-\frac{\beta}{\beta+d}})$。

根据步骤 4, $\sqrt{\mathbb{E}\left[l(T,\delta,Z;\ \pi_n h_0,\pi_n m_0,\pi_n\theta_0)\ -l(T,\delta,Z;\ h_0,m_0,\theta_0)\right]}=O\left(n^{-\frac{\beta}{2\beta+2d}}\right)$。因此，我们有 $d_{\mathsf{PF}}\left(\widehat{\phi}_n,\phi_0\right)=O(n^{-\frac{\beta}{2\beta+2d}}\log^2 n)$ $=\widetilde{O}(n^{-\frac{\beta}{2\beta+2d}})$。

□

定理 4.7 的证明。证明分为四个步骤。

$\widehat{\quad}$ $\widehat{\quad}$ $\widehat{\quad}$步骤 1 我们记 $\psi_0 = (\nu_0, \theta_0)$ 和 $\psi = (\widehat{\nu}, \theta)$，其中 $\widehat{\nu} \in \mathcal{V}_n$ 和 $\theta \in \Theta$。对于任意$0 < \varepsilon \leq 1$，我们有

$$\begin{aligned}
&\inf_{d_{\mathsf{FN}}(\widehat{\psi},\psi_0)\geq\varepsilon} \mathbb{E}\left[l(T,\delta,Z;\nu_0,\theta_0) - l(T,\delta,Z;\widehat{\nu},\widehat{\theta})\right] \\
&= \inf_{d_{\mathsf{FN}}(\widehat{\psi},\psi_0)\geq\varepsilon} \mathbb{E}_Z\left[\mathbb{E}_{T,\delta|Z}\left[\log p(T,\delta \mid Z;\nu_0,\theta_0) - \log p(T,\delta \mid Z;\widehat{\nu},\widehat{\theta})\right]\right] \\
&= \inf_{d_{\mathsf{FN}}(\widehat{\psi},\psi_0)\geq\varepsilon} \mathbb{E}_Z\left[\mathrm{D}_{\mathrm{KL}}\left(\mathbb{P}_{\widehat{\psi},Z} \,\|\|\, \mathbb{P}_{\psi_0,Z}\right)\right]
\end{aligned}$$

利用这样一个事实：$KL(\mathbb{P}_{\widehat{\psi},Z} \,\|\, \mathbb{P}_{\psi_0,Z}) \geq 2H^2(\mathbb{P}_{\widehat{\psi},Z} \,\|\, \mathbb{P}_{\psi_0,Z})$。因此，我们进一步得到 t

$$\begin{aligned}
&\inf_{d_{\mathsf{FN}}(\widehat{\psi},\psi_0)\geq\varepsilon} \mathbb{E}\left[l(T,\delta,Z;\nu_0,\theta_0) - l(T,\delta,Z;\widehat{\nu},\widehat{\theta})\right] \\
&\geq \inf_{d_{\mathsf{FN}}(\widehat{\psi},\psi_0)\geq\varepsilon} 2\mathbb{E}_Z\left[H^2(\mathbb{P}_{\widehat{\psi},Z} \,\|\, \mathbb{P}_{\psi_0,Z})\right] \\
&= 2 \inf_{d_{\mathsf{FN}}(\widehat{\psi},\psi_0)\geq\varepsilon} d^2_{\mathsf{FN}}\left(\widehat{\psi},\psi_0\right) \\
&\geq 2\varepsilon^2.
\end{aligned}$$

步骤 2 我们考虑以下推导。

$$\begin{aligned}
&\sup_{d_{\mathsf{FN}}(\widehat{\psi},\psi_0)\leq\varepsilon} \mathrm{Var}\left[l(T,\delta,Z;\nu_0,\theta_0) - l(T,\delta,Z;\widehat{\nu},\widehat{\theta})\right] \\
&\leq \sup_{d_{\mathsf{FN}}(\widehat{\psi},\psi_0)\leq\varepsilon} \mathbb{E}\left[\left(l(T,\delta,Z;\nu_0,\theta_0) - l(T,\delta,Z;\widehat{\nu},\widehat{\theta})\right)^2\right] \\
&= \sup_{d_{\mathsf{FN}}(\widehat{\psi},\psi_0)\leq\varepsilon} \mathbb{E}_Z\left[\mathbb{E}_{T,\delta|Z}\left[\left(\log p(T,\delta,Z;\nu_0,\theta_0) - \log p(T,\delta,Z;\widehat{\nu},\widehat{\theta})\right)^2\right]\right] \\
&= 4 \sup_{d_{\mathsf{FN}}(\widehat{\psi},\psi_0)\leq\varepsilon} \mathbb{E}_Z\left[\int\left(p(T,\delta,Z;\nu_0,\theta_0)(\log\sqrt{\frac{p(T,\delta,Z;\nu_0,\theta_0)}{p(T,\delta,Z;\widehat{\nu},\widehat{\theta})}})^2\right)\mu(dT\times d\delta)\right]
\end{aligned}$$

对 $\log x$ 进行泰勒展开后，存在一个介于 $\sqrt{p(T,\delta,Z;\nu_0,\theta_0)}$ 和 $\sqrt{p(T,\delta,Z;\widehat{\nu},\widehat{\theta})}$ 之间的 $\eta(T,\delta,Z)$，使得该关系逐点成立：

$$\begin{aligned}
&p(T,\delta,Z;\nu_0,\theta_0)(\log\sqrt{\frac{p(T,\delta,Z;\nu_0,\theta_0)}{p(T,\delta,Z;\widehat{\nu},\widehat{\theta})}})^2 \\
&= p(T,\delta,Z;\nu_0,\theta_0)\left(\log\sqrt{p(T,\delta,Z;\nu_0,\theta_0)} - \log\sqrt{p(T,\delta,Z;\widehat{\nu},\widehat{\theta})}\right)^2 \\
&= \frac{p(T,\delta,Z;\nu_0,\theta_0)}{\eta(T,\delta,Z)^2}\left(\sqrt{p(T,\delta,Z;\nu_0,\theta_0)} - \sqrt{p(T,\delta,Z;\widehat{\nu},\widehat{\theta})}\right)^2
\end{aligned}$$

由于 $p(T,\delta,Z;\nu_0,\theta_0)/p(T,\delta,Z;\widehat{\nu},\widehat{\theta}) = e^{l(T,\delta,Z;\nu_0,\theta_0)-l(T,\delta,Z;\widehat{\nu},\widehat{\theta})}$，根据引理 B.2，$l(T,\delta,Z;\nu_0,\theta_0)$ 和 $l(T,\delta,Z;\widehat{\nu},\widehat{\theta})$ 在 $[0,\tau]\times\{0,1\}\times[-1,1]^d$ 上一致有界$\widehat{\quad}$ $\widehat{\quad}$，对所有 $\psi = (\widehat{\nu},\theta)$ 均成立。因此，存在常数 C_3 和 C_4，使得 $0 < C_3 \leq p(T,\delta,Z;\nu_0,\theta_0)/p(T,\delta,Z;\widehat{\nu},\widehat{\theta}) \leq C_4$。由此可知，$p(T,\delta,Z;\nu_0,\theta_0)\ \eta\ T\ \frac{1}{\delta}(\ ,\ ,Z)^2$ 有界。我们进一步有

$$\begin{aligned}
&p(T,\delta,Z;\nu_0,\theta_0)\left(\log\sqrt{p(T,\delta,Z;\nu_0,\theta_0)} - \log\sqrt{p(T,\delta,Z;\widehat{\nu},\widehat{\theta})}\right)^2 \\
&\lesssim \left|\sqrt{p(T,\delta,Z;\nu_0,\theta_0)} - \sqrt{p(T,\delta,Z;\widehat{\nu},\widehat{\theta})}\right|^2.
\end{aligned}$$

因此，我们有

$$\begin{aligned}
&\sup_{d_{\mathsf{FN}}\left(\widehat{\psi},\psi_0\right)\leq\varepsilon}\operatorname{Var}\left[l(T,\delta,Z;\nu_0,\theta_0)-l(T,\delta,Z;\widehat{\nu},\widehat{\theta})\right]\\
&\lesssim\sup_{d_{\mathsf{FN}}\left(\widehat{\psi},\psi_0\right)\leq\varepsilon}\mathbb{E}_Z\left[\int\left|\sqrt{p(T,\delta,Z;\nu_0,\theta_0)}-\sqrt{p(T,\delta,Z;\widehat{\nu},\widehat{\theta})}\right|^2\mu(dT\times d\delta)\right]\\
&=\sup_{d_{\mathsf{FN}}\left(\widehat{\psi},\psi_0\right)\leq\varepsilon}d_{\mathsf{FN}}^2\left(\widehat{\psi},\psi_0\right)\\
&\leq\varepsilon^2.
\end{aligned}$$

步骤 3 我们定义 $\mathcal{G}_n=\{l(T,\delta,Z;\widehat{\nu},\theta)-l(T,\delta,Z;\pi_n\nu_0,\pi_n\theta_0):\widehat{\nu}\in\mathcal{V}_n,\theta\in\Theta\}$。这里，$(\pi_n\nu_0,\pi_n\theta_0)$ 已在引理 B.6 中定义。显然，我们有 $\log N_{[]}(\varepsilon,\mathcal{G}_n,\|\cdot\|_\infty)=\log N_{[]}(\varepsilon,\mathcal{G}_n,\|\cdot\|_\infty)$，其中 $\mathcal{G}$ 在引理 B.10 中定义。由引理 B.10，我们进一步得到

$$\log N_{[]}(\varepsilon,\mathcal{G}_n,\|\cdot\|_\infty)\lesssim\log\frac{1}{\varepsilon}+\log N(c_\nu\varepsilon^{1/s_\nu},\mathcal{V}_n,\|\cdot\|_2).$$

根据条件 4.4 下的 [2, 定理 7], ，我们有 $\mathcal{V}_n$ 的 VC 维数满足 $\mathrm{VC}(\mathcal{V}_n)\lesssim n^{\frac{d+1}{\beta+d+1}}\log^3 n$。因此，我们得到

$$\log N(c_h\varepsilon^{1/s_\nu},\mathcal{V}_n,\|\cdot\|_2)\lesssim\frac{\mathrm{VC}(\mathcal{V}_n)}{s_\nu}\log\frac{1}{\varepsilon}\lesssim n^{\frac{d+1}{\beta+d+1}}\log^3 n\log\frac{1}{\varepsilon}.$$

此外，我们得到 $\log N_{[]}(\varepsilon,\widetilde{\mathcal{G}}_n,\|\cdot\|_\infty)\lesssim n^{\frac{d+1}{\beta+d+1}}\log^3 n\log\frac{1}{\varepsilon}$。

步骤 4 根据柯西-施瓦茨不等式，我们有

$$\sqrt{\mathbb{E}[l(T,\delta,Z;\pi_n\nu_0,\pi_n\theta_0)-l(T,\delta,Z;\nu_0,\theta_0)]}\leq\left[\mathbb{E}\left(l(T,\delta,Z;\pi_n\nu_0,\pi_n\theta_0)-l(T,\delta,Z;\nu_0,\theta_0)\right)^2\right]^{\frac{1}{4}}.$$

与第二部分类似，并根据引理 B.6，我们进一步得到

$$\sqrt{\mathbb{E}[l(T,\delta,Z;\pi_n\nu_0,\pi_n\theta_0)-l(T,\delta,Z;\nu_0,\theta_0)]}\lesssim\sqrt{d_{\mathsf{FN}}\left(\pi_n\psi_0,\psi_0\right)}\lesssim n^{-\frac{\beta}{2\beta+2d+2}}$$

现在令

$$\tau=\frac{\beta}{2\beta+2d+2}-2\frac{\log\log n}{\log n}.$$

根据步骤 1、2、3以及步骤 1、2、3和 [60, 定理 1],

$$\begin{aligned}
d_{\mathsf{FN}}\left(\widehat{\psi}_n,\psi_0\right)=\max\big(&n^{-\tau},d_{\mathsf{FN}}\left(\pi_n\psi_0,\psi_0\right),\\
&\sqrt{\mathbb{E}[l(T,\delta,Z;\pi_n\nu_0,\pi_n\theta_0)-l(T,\delta,Z;\nu_0,\theta_0)]}\big)
\end{aligned}$$

根据引理 B.6，$d_{\mathsf{FN}}\left(\pi_n\psi_0,\psi_0\right)=O(n^{-\frac{\beta}{\beta+d+1}})$

根据步骤 4，$\sqrt{\mathbb{E}[l(T,\delta,Z;\pi_n\nu_0,\pi_n\theta_0)-l(T,\delta,Z;\nu_0,\theta_0)]}=O(n^{-\frac{\beta}{2\beta+2d+2}})$。因此，我们有$d_{\mathsf{FN}}\left(\widehat{\psi}_n,\psi_0\right)=O(n^{-\frac{\beta}{2\beta+2d+2}}\log^2 n)=\widetilde{O}(n^{-\frac{\beta}{2\beta+2d+2}})$。 □

B.4 技术引理的证明

引理 B.1 的证明。由于 $h_0(T)\in\mathcal{W}_M^\beta([0,\tau])$ 且 $m_0(Z)\in\mathcal{W}_M^\beta([-1,1]^d)$，我们有 $h_0(T)\leq M, m_0(Z)\leq M$ 且 $e^{m_0(Z)}\int_0^T h_0(s)ds\leq\tau e^{2M}$。设 $\mathcal{B}=[0,\tau e^{2M}]$，则有

$$\begin{aligned}
&|l(T,\delta,Z;h_0,m_0,\theta_0)|\\
&\leq\left|\log g_{\theta_0}\left(e^{m_0(Z)}\int_0^T e^{h_0(s)}ds\right)\right|+|h_0(T)|+|m_0(Z)|+\left|G_{\theta_0}\left(e^{m_0(Z)}\int_0^T e^{h_0(s)}ds\right)\right|\\
&\leq 2M+\sup_{x\in\mathcal{B}}|\log g_{\theta_0}(x)|+\sup_{x\in\mathcal{B}}|G_{\theta_0}(x)|
\end{aligned}$$

根据条件 5，对于 $(T,\delta,Z)\in[0,\tau]\times\{0,1\}\times[-1,1]^d$，$l(T,\delta,Z;h_0,m_0,\theta_0)$ 有界。证明 $l(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})$ 有界的方法类似。 □

引理 B.2 的证明。由于 $\nu_0(T,Z)\in \mathcal{W}_M^{\beta}([0,\tau]\times[-1,1]^d)$，我们有 $\nu_0(T,Z) \leq M$and $\int_0^T e^{\nu(s,Z)}ds \leq \tau e^M$。令 $\mathcal{B}=[0,\tau e^M]$，则有

$$
\begin{aligned}
&|l(T,\delta,Z;\nu_0,\theta_0)| \\
&\leq \left|\log G'_{\theta_0}\left(\int_0^T e^{\nu_0(s,Z)}ds\right)\right| + |\nu_0(T,Z)| + \left|G_{\theta_0}\left(\int_0^T e^{\nu_0(s,Z)}ds\right)\right| \\
&\leq M + \sup_{x\in\mathcal{B}}\left|\log G'_{\theta_0}(x)\right| + \sup_{x\in\mathcal{B}}\left|G_{\theta_0}(x)\right|.
\end{aligned}
$$

根据条件 4.5，我们有 $l(T,\delta,Z;\nu_0,\theta_0)$ 在 $(T,\delta,Z)\in[0,\tau]\times\{0,1\}\times[-1,1]^d$ 中有界。$l(T,\delta,Z;\widehat{\nu},\widehat{\theta})$ 有界性的证明类似。 □

引理 B.3 的证明。根据定义，我们有

$$
\begin{aligned}
&|l(T,\delta,Z;h_0,m_0,\theta_0) - l(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})| \\
&\leq \left|\log g_{\theta_0}\left(e^{m_0(Z)}\int_0^T e^{h_0(s)}ds\right) - \log g_{\widehat{\theta}}\left(e^{\widehat{m}(Z)}\int_0^T e^{\widehat{h}(s)}ds\right)\right| + \left|h_0(T)-\widehat{h}(T)\right| \\
&+|m_0(Z)-\widehat{m}(Z)| + \left|G_{\theta_0}\left(e^{m_0(Z)}\int_0^T e^{h_0(s)}ds\right) - G_{\widehat{\theta}}\left(e^{\widehat{m}(Z)}\int_0^T e^{\widehat{h}(s)}ds\right)\right|.
\end{aligned}
$$

令 $\mathcal{B}=[0,\tau\max(e^{2M},e^{M_h+M_m})]$。根据泰勒展开，我们还可以证明

$$
\begin{aligned}
&|l(T,\delta,Z;h_0,m_0,\theta_0) - l(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})| \\
&\leq \sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial \log g_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{\theta}}}\right|\cdot\left|\theta_0-\widehat{\theta}\right| \\
&+\sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial \log g_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{x}}}\right|\cdot\left|e^{m_0(Z)}\int_0^T e^{h_0(s)}ds - e^{\widehat{m}(Z)}\int_0^T e^{\widehat{h}(s)}ds\right| \\
&+|h_0(T)-\widehat{h}(T)| + |m_0(Z)-\widehat{m}(Z)| + \sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial G_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{\theta}}}\right|\cdot\left|\theta_0-\widehat{\theta}\right| \\
&+\sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial G_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{x}}}\right|\cdot\left|e^{m_0(Z)}\int_0^T e^{h_0(s)}ds - e^{\widehat{m}(Z)}\int_0^T e^{\widehat{h}(s)}ds\right|.
\end{aligned}
$$

同样，根据泰勒展开，我们有

$$
\begin{aligned}
&\left|e^{m_0(Z)}\int_0^T e^{h_0(s)}ds - e^{\widehat{m}(Z)}\int_0^T e^{\widehat{h}(s)}ds\right| \\
&\leq \left|e^{m_0(Z)}\int_0^T (e^{h_0(s)}-e^{\widehat{h}(s)})ds\right| + \left|(e^{m_0(Z)}-e^{\widehat{m}(Z)})\int_0^T e^{\widehat{h}(s)}ds\right| \\
&\leq e^M\cdot\tau e^{\max(M,M_h)}\left\|h_0-\widehat{h}\right\|_\infty + \tau e^{M_h}\cdot e^{\max(M,M_m)}\|m_0-\widehat{m}\|_\infty.
\end{aligned}
$$

最终，我们得到

$$
\begin{aligned}
&|l(T,\delta,Z;h_0,m_0,\theta_0) - l(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})| \\
&\leq \sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial \log g_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{x}}}\right|\cdot\left[e^M\cdot\tau e^{\max(M,M_h)}\|h_0-\widehat{h}\|_\infty + \tau e^{M_h}\cdot e^{\max(M,M_m)}\|m_0-\widehat{m}\|_\infty\right] \\
&+\sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial \log g_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{\theta}}}\right|\cdot\left|\theta_0-\widehat{\theta}\right| + \left|h_0(T)-\widehat{h}(T)\right| + |m_0(Z)-\widehat{m}(Z)| \\
&+\sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial G_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{\theta}}}\right|\cdot\left|\theta_0-\widehat{\theta}\right| \\
&+\sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial G_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{x}}}\right|\cdot\left[e^M\cdot\tau e^{\max(M,M_h)}\|h_0-\widehat{h}\|_\infty + \tau e^{M_h}\cdot e^{\max(M,M_m)}\|m_0-\widehat{m}\|_\infty\right].
\end{aligned}
$$

对两边取上确界，可得

$$\|l(T,\delta,Z;h_0,m_0,\theta_0)-l(T,\delta,Z;\widehat{h},\widehat{m},\widehat{\theta})\|_\infty \lesssim |\theta_0-\widehat{\theta}|+\|h_0-\widehat{h}\|_\infty+\|m_0-\widehat{m}\|_\infty.$$

第二个不等式的证明类似。 □

引理 B.4 的证明。根据定义，我们有

$$\begin{aligned}
&|l(T,\delta,Z;\nu_0,\theta_0)-l(T,\delta,Z;\widehat{\nu},\widehat{\theta})|\\
&\le \left|\log g_{\theta_0}\left(\int_0^T e^{\nu_0(s,Z)}ds\right)-\log g_{\widehat{\theta}}\left(\int_0^T e^{\widehat{\nu}(s,Z)}ds\right)\right|+|\nu_0(T,Z)-\widehat{\nu}(T,Z)|\\
&+\left|G_{\theta_0}\left(\int_0^T e^{\nu_0(s,Z)}ds\right)-G_{\widehat{\theta}}\left(\int_0^T e^{\widehat{\nu}(s,Z)}ds\right)\right|.
\end{aligned}$$

设 $\mathcal{B}=[0,\tau\max(e^M,e^{M_\nu})]$。通过泰勒展开，我们还可以证明

$$\begin{aligned}
&|l(T,\delta,Z;\nu_0,\theta_0)-l(T,\delta,Z;\widehat{\nu},\widehat{\theta})|\\
&\le \sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial\log g_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{\theta}}}\right|\cdot\left|\theta_0-\widehat{\theta}\right|+\sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial\log g_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{x}}}\right|\cdot\left|\int_0^T e^{\nu_0(s,Z)}ds-\int_0^T e^{\widehat{\nu}(s,Z)}ds\right|\\
&+|\nu_0(T,Z)-\widehat{\nu}(T,Z)|+\sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial G_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{\theta}}}\right|\cdot|\theta_0-\widehat{\theta}|\\
&+\sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial G_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{x}}}\right|\cdot\left|\int_0^T e^{\nu_0(s,Z)}ds-\int_0^T e^{\widehat{\nu}(s,Z)}ds\right|.
\end{aligned}$$

再次通过泰勒展开，

$$\left|\int_0^T e^{\nu_0(s,Z)}ds-\int_0^T e^{\widehat{\nu}(s,Z)}ds\right|\le\tau e^{\max(M,M_\nu)}\|\nu_0-\widehat{\nu}\|_\infty,$$

最后，我们得到

$$\begin{aligned}
&\left|l(T,\delta,Z;\nu_0,\theta_0)-l(T,\delta,Z;\widehat{\nu},\widehat{\theta})\right|\\
&\le \sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial\log g_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{\theta}}}\right|\cdot\left|\theta_0-\widehat{\theta}\right|+\sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial\log g_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{x}}}\right|\cdot\tau e^{\max(M,M_\nu)}\left\|\nu_0-\widehat{\nu}\right\|_\infty\\
&+|\nu_0(T,Z)-\widehat{\nu}(T,Z)|+\sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial G_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{\theta}}}\right|\cdot\left|\theta_0-\widehat{\theta}\right|\\
&+\sup_{\tilde{\theta}\in\Theta,\tilde{x}\in\mathcal{B}}\left|\frac{\partial G_{\tilde{\theta}}(\tilde{x})}{\partial_{\tilde{x}}}\right|\cdot\tau e^{\max(M,M_\nu)}\left\|\nu_0-\widehat{\nu}\right\|_\infty.
\end{aligned}$$

对两边取上确界，可得

$$\|l(T,\delta,Z;\nu_0,\theta_0)-l(T,\delta,Z;\widehat{\nu},\widehat{\theta})\|_\infty\lesssim|\theta_0-\widehat{\theta}|+\|\nu_0-\widehat{\nu}\|_\infty,$$

第二个不等式的证明类似。 □

引理 B.5 的证明。根据 [72, 定理 1], ，存在逼近函数 h^* 和$\widehat{m}^*$，使得 $\|\widehat{h}^*-h_0\|_\infty=O\left(n^{-\frac{\beta}{\beta+d}}\right)$ 和 $\|\widehat{m}^*-m_0\|_\infty=O\left(n^{-\frac{\beta}{\beta+d}}\right)$。令 $\pi_n h_0=\widehat{h}^*$,

$\pi_n m_0 = \widehat{m}^*$，且 $\pi_n \theta = \theta_0$。我们有

$$
\begin{aligned}
& d_{\mathsf{PF}}\left(\pi_n \phi_0, \phi_0\right) \\
&= \sqrt{\mathbb{E}_Z\left[\int |\sqrt{p(T,\delta \mid Z; \pi_n h_0, \pi_n m_0, \pi_n \theta_0)} - \sqrt{p(T,\delta \mid Z; h_0, m_0, \theta_0)}|^2 \mu(dT \times d\delta)\right]} \\
&= \sqrt{\mathbb{E}_Z\left[\int [e^{\frac{1}{2} l(T,\delta,Z;\pi_n h_0, \pi_n m_0, \pi_n \theta_0)} - e^{\frac{1}{2} l(T,\delta,Z;h_0,m_0,\theta_0)}]^2 f_{C|Z}(T)^{1-\delta} S_{C|Z}(T)^{\delta} \mu(dT \times d\delta)\right]} \\
&\leq \left\| e^{\frac{1}{2} l(T,\delta,Z;\pi_n h_0, \pi_n m_0, \pi_n \theta_0)} - e^{\frac{1}{2} l(T,\delta,Z;h_0,m_0,\theta_0)} \right\|_{\infty} \\
&\qquad \times \sqrt{\mathbb{E}_Z\left[\int f_{C|Z}(T)^{1-\delta} S_{C|Z}(T)^{\delta} \mu(dT \times d\delta)\right]}.
\end{aligned}
$$

由引理 B.1 和 B.3 可得

$$
\begin{aligned}
& \| e^{\frac{1}{2} l(T,\delta,Z;\pi_n h_0, \pi_n m_0, \pi_n \theta_0)} - e^{\frac{1}{2} l(T,\delta,Z;h_0,m_0,\theta_0)} \|_{\infty} \\
&\lesssim \|\pi_n \theta_0 - \theta_0\| + \|\pi_n h_0 - h_0\|_{\infty} + \|\pi_n m_0 - m_0\|_{\infty} \\
&= O\left(n^{-\frac{\beta}{\beta+d}}\right).
\end{aligned}
$$

由于 $f_{C|Z}(T)^{1-\delta} \leq f_{C|Z}(T) + 1$ 且 $S_{C|Z}(T)^{\delta} \leq 1$，我们还可得

$$
\begin{aligned}
\sqrt{\mathbb{E}_Z\left[\int f_{C|Z}(T)^{1-\delta} S_{C|Z}(T)^{\delta} \mu(dT \times d\delta)\right]} &\leq \sqrt{\mathbb{E}_Z\left[\int (1 + f_{C|Z}(T)) \mu(dT \times d\delta)\right]} \\
&\leq \sqrt{2 + 2\tau}.
\end{aligned}
$$

因此，我们得到 $d_{\mathsf{PF}}\left(\pi_n \phi_0, \phi_0\right) = O\left(n^{-\frac{\beta}{\beta+d}}\right)$。

□

引理 B.6 的证明。根据 [72, 定理 1]，，存在一个逼近函数 $\widehat{\nu}^*$ 使得 $\|\widehat{\nu}^* - \nu_0\|_{\infty} = O\left(n^{-\frac{\beta}{\beta+d+1}}\right)$。令 $\pi_n \nu_0 = \widehat{\nu}^*$ 和 $\pi_n \theta_0 = \theta_0$。我们可得

$$
\begin{aligned}
& d_{\mathsf{FN}}\left(\pi_n \psi_0, \psi_0\right) \\
&= \sqrt{\mathbb{E}_Z\left[\int \left|\sqrt{p(T,\delta \mid Z; \pi_n \nu_0, \pi_n \theta_0)} - \sqrt{p(T,\delta \mid Z; \nu_0, \theta_0)}\right|^2 \mu(dT \times d\delta)\right]} \\
&= \sqrt{\mathbb{E}_Z\left[\int \left[e^{\frac{1}{2} l(T,\delta,Z;\pi_n \nu_0, \pi_n \theta_0)} - e^{\frac{1}{2} l(T,\delta,Z;\nu_0,\theta_0)}\right]^2 f_{C|Z}(T)^{1-\delta} S_{C|Z}(T)^{\delta} \mu(dT \times d\delta)\right]} \\
&\leq \left\| \frac{1}{2} e^{l(T,\delta,Z;\pi_n \nu_0, \pi_n \theta_0)} - \frac{1}{2} e^{l(T,\delta,Z;\nu_0,\theta_0)} \right\|_{\infty} \sqrt{\mathbb{E}_Z\left[\int f_{C|Z}(T)^{1-\delta} S_{C|Z}(T)^{\delta} \mu(dT \times d\delta)\right]}.
\end{aligned}
$$

由引理 B.2 和 B.4 可得

$$
\begin{aligned}
\left\| e^{\frac{1}{2} l(T,\delta,Z;\pi_n \nu_0, \pi_n \theta_0)} - e^{\frac{1}{2} l(T,\delta,Z;\nu_0,\theta_0)} \right\|_{\infty} &\lesssim \|\pi_n \theta_0 - \theta_0\| + \|\pi_n \nu_0 - \nu_0\|_{\infty} \\
&= O\left(n^{-\frac{\beta}{\beta+d+1}}\right).
\end{aligned}
$$

由于 $f_{C|Z}(T)^{1-\delta} \leq f_{C|Z}(T) + 1$ 且 $S_{C|Z}(T)^{\delta} \leq 1$，我们还可得

$$
\begin{aligned}
\sqrt{\mathbb{E}_Z\left[\int f_{C|Z}(T)^{1-\delta} S_{C|Z}(T)^{\delta} \mu(dT \times d\delta)\right]} &\leq \sqrt{\mathbb{E}_Z\left[\int (1 + f_{C|Z}(T)) \mu(dT \times d\delta)\right]} \\
&\leq \sqrt{2 + 2\tau}.
\end{aligned}
$$

因此，我们得到 $d_{\mathsf{FN}}\left(\pi_n \psi_0, \psi_0\right) = O\left(n^{-\frac{\beta}{\beta+d+1}}\right)$。

□

引理 B.7 的证明。根据覆盖数的定义，左侧不等式显然成立。我们需要证明右侧不等式成立。

假设我们有 $\{B(g_i, \frac{\varepsilon}{2})\}, i = 1 \ldots, N$，其中 $N = N(\frac{\varepsilon}{2}, \mathcal{F}, \|\cdot\|)$ 是覆盖 $\mathcal{F}$ 所需的最少 $\frac{\varepsilon}{2}$-球数量。那么至少存在一个 $f_i \in \mathcal{F}$，使得 $f_i \in B(g_i, \varepsilon)$。考虑以下 $\varepsilon - balls$ $\{B(f_i, \varepsilon)\}, i = 1 \ldots, N$。对于任意的 $f \in \mathcal{F} \cap B(g_i, \frac{\varepsilon}{2})$，我们有 $\|f - f_i\| \leq \|f - g_i\| + \|f_i - g_i\| \leq \varepsilon$。因此，$\{B(f_i, \varepsilon)\}$, $i = 1 \ldots, N$ 构成了 $\mathcal{F}$ 的一个 ε-覆盖。根据 $\widetilde{}$ 的定义，我们有 $N(\varepsilon, \mathcal{F}, \|\cdot\|) \leq N(\frac{\varepsilon}{2}, \mathcal{F}, \|\cdot\|)$。□

引理 B.8 的证明。前两个不等式的证明步骤与引理 B.7 完全相同。这里只需说明其余部分，即 $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) = N(\frac{\varepsilon}{2}, \mathcal{F}, \|\cdot\|_\infty)$。我们首先选取一组 $\frac{\varepsilon}{2}$-覆盖球 $\{B(f_i, \frac{\varepsilon}{2})\}, i = 1, \ldots, N_1$，其中 $N_1 = N(\frac{\varepsilon}{2}, \mathcal{F}, \|\cdot\|_\infty)$。接下来构造一组括号 $\{[l_i, u_i]\}, i = 1 \ldots, N_1$，其中 $l_i = f_i - \frac{\varepsilon}{2}$ 且 $u_i = f_i + \frac{\varepsilon}{2}$。注意到括号 $\{[l_i, u_i]\}$ 与 $B(f_i, \frac{\varepsilon}{2})$ 完全相同，因此集合 $\{[l_i, u_i]\}, i = 1, \ldots, N_1$ 覆盖了 $\mathcal{F}$，从而得到 $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \leq N(\frac{\varepsilon}{2}, \mathcal{F}, \|\cdot\|_\infty)$。同理，我们有 $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \geq N(\frac{\varepsilon}{2}, \mathcal{F}, \|\cdot\|_\infty)$。因此，我们有 $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) = N(\frac{\varepsilon}{2}, \mathcal{F}, \|\cdot\|_\infty)$。□

引理 B.9 的证明。根据引理 B.8，首先有 $N_{[]}(\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty) \leq N_{[]}(\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty)$。根据引理 B.3，存在一个常数 $c_1 > 0$，使得对于任意的 $h_1, h_2 \in \mathcal{H}_n, \widehat{m}_1, \widehat{m}_2 \in \mathcal{M}_n$ 和 $\theta_1, \theta_2 \in \Theta$，有

$$\|l(T, \delta, Z;\ h_1, \widehat{m}_1, \theta_1) - l(T, \delta, Z;\ h_2, \widehat{m}_2, \theta_2)\|_\infty \leq c_1[|\theta_1 - \theta_2| + \|h_1 - h_2\|_\infty + \|\widehat{m}_1 - \widehat{m}_2\|_\infty],$$

这表明，只要 $|\theta_1 - \theta_2| \leq \frac{\varepsilon}{3c_1}, \|h_1 - h_2\|_\infty \leq \frac{\varepsilon}{3c_1}$ 且 $\|\widehat{m}_1 - \widehat{m}_2\|_\infty \leq \frac{\varepsilon}{3c_1}$，我们就有 $\|l(T, \delta, Z;\ h_1, \widehat{m}_1, \theta_1) - l(T, \delta, Z;\ h_2, \widehat{m}_2, \theta_2)\|_\infty \leq \varepsilon$。因此，我们有

$$\widetilde{N}_{[]}(\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty) \leq \widetilde{N}_{[]}(\frac{\varepsilon}{3c_1}, \Theta, \|\cdot\|_\infty) \times \widetilde{N}_{[]}(\frac{\varepsilon}{3c_1}, \mathcal{H}_n, \|\cdot\|_\infty) \times \widetilde{N}_{[]}(\frac{\varepsilon}{3c_1}, \mathcal{M}_n, \|\cdot\|_\infty).$$

由于 Θ 是 $\mathbb{R}$上的紧集，根据引理 B.8和传统体积论证，我们有$\widetilde{N}_{[]}(\frac{\varepsilon}{3c_1}, \Theta, \|\cdot\|_\infty) \leq N_{[]}(\frac{\varepsilon}{6c_1}, \Theta, \|\cdot\|_\infty) \lesssim \frac{1}{\varepsilon}$。

对于 $N_{[]}(\frac{\varepsilon}{3c_1}, \mathcal{H}_n, \|\cdot\|_\infty)$，根据引理 B.8，我们有 $N_{[]}(\frac{\varepsilon}{3c_1}, \mathcal{H}_n, \|\cdot\|_\infty) = N(\frac{\varepsilon}{3c_1}, \mathcal{H}_n, \|\cdot\|_\infty)$。根据 [12, 引理 2], , 存在一个常数 $c_2 > 0$ 使得 $\|h_1 - h_2\|_\infty \leq c_2\|h_1 - h_2\|_2^{s_h}$，由此得到 $\widetilde{N}(\frac{\varepsilon}{3c_1}, \mathcal{H}_n, \|\cdot\|_\infty) \leq \widetilde{N}(\frac{\varepsilon^{1/s_h}}{(3c_1c_2)^{1/s_h}}, \mathcal{H}_n, \|\cdot\|_2)$。根据引理 B.7，我们进一步有$\widetilde{N}(\frac{\varepsilon^{1/s_h}}{(3c_1c_2)^{1/s_h}}, \mathcal{H}_n, \|\cdot\|_2) \leq N(\frac{\varepsilon^{1/s_h}}{2\ (3c_1c_2)^{1/s_h}}, \mathcal{H}_n, \|\cdot\|_2)$。令 $c_h = \frac{1}{2\ (3c_1c_2)^{1/s_h}}$。我们有$\widetilde{N}_{[]}(\frac{\varepsilon}{3c_1}, \mathcal{H}_n, \|\cdot\|_\infty) \leq N(c_h\varepsilon^{1/s_h}, \mathcal{H}_n, \|\cdot\|_2)$。

类似地，存在一个常数 $c_m > 0$使得 $\widetilde{N}_{[]}(\frac{\varepsilon}{3c_1}, \mathcal{M}_n, \|\cdot\|_\infty) \leq N(c_m\varepsilon^{1/s_m}, \mathcal{M}_n, \|\cdot\|_2)$。

因此，最终我们可以得到

$$N_{[]}(\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty) \lesssim \frac{1}{\varepsilon} N(c_h\varepsilon^{1/s_h}, \mathcal{H}_n, \|\cdot\|_2) \times N(c_m\varepsilon^{1/s_m}, \mathcal{M}_n, \|\cdot\|_2).$$

□

引理 B.10 的证明。根据引理 B.8，首先有 $N_{[]}(\varepsilon, \mathcal{G}_n, \|\cdot\|_\infty) \leq N_{[]}(\varepsilon, \mathcal{G}_n, \|\cdot\|_\infty)$。根据引理 B.4，存在一个常数 $c_3 > 0$，使得对于任意的 $\widehat{\nu}_1, \widehat{\nu}_2 \in \mathcal{V}_n$ 和 $\theta_1, \theta_2 \in \Theta$，我们有

$$\|l(T, \delta, Z; \widehat{\nu}_1, \widehat{\theta}_1) - l(T, \delta, Z; \widehat{\nu}_2, \widehat{\theta}_2)\|_\infty \leq c_3[|\widehat{\theta}_1 - \widehat{\theta}_2| + \|\widehat{\nu}_1 - \widehat{\nu}_2\|_\infty],$$

这表明，只要 $|\theta_1 - \theta_2| \leq \frac{\varepsilon}{2c_3}$ 且 $\|\widehat{\nu}_1 - \widehat{\nu}_2\|_\infty \leq \frac{\varepsilon}{2c_3}$，就有 $\|l(T, \delta, Z;\ \widehat{\nu}_1, \theta_1) - l(T, \delta, Z;\ \widehat{\nu}_2, \theta_2)\|_\infty \leq \varepsilon$。因此，我们有:

$$\widetilde{N}_{[]}(\varepsilon, \mathcal{G}_n, \|\cdot\|_\infty) \leq \widetilde{N}_{[]}(\frac{\varepsilon}{2c_3}, \Theta, \|\cdot\|_\infty) \times \widetilde{N}_{[]}(\frac{\varepsilon}{2c_3}, \mathcal{V}_n, \|\cdot\|_\infty).$$

表 3：基准数据集的描述性统计

	METABRIC	RotGBSG	FLCHAIN	SUPPORT	MIMIC-III	KKBOX
Size	1904	2232	6524	8873	35953	2646746
删失率	0.423	0.432	0.699	0.320	0.901	0.280
特征	9	7	8	14	26	15

由于 Θ 是 $\mathbb{R}$上的紧集，根据引理 B.8和传统体积论证，我们有$\widetilde{N}_{[]}(\frac{\varepsilon}{2c_3}, \Theta, \|\cdot\|_\infty) \leq N_{[]}(\frac{\varepsilon}{4c_3}, \Theta, \|\cdot\|_\infty) \lesssim \frac{1}{\varepsilon}$。

~ ~ ~对于 $N_{[]}(\frac{\varepsilon}{2c_3}, \mathcal{V}_n, \|\cdot\|_\infty)$，根据引理 B.8，我们有 $N_{[]}(\frac{\varepsilon}{2c_3}, \mathcal{V}_n, \|\cdot\|_\infty) = N(\frac{\varepsilon}{2c_3}, \mathcal{V}_n, \|\cdot\|_\infty)$。根据[12, 引理 2]，，存在一个常数 $c_4 > 0$使得$\|\widehat{\nu}_1 - \widehat{\nu}_2\|_\infty \leq c_4\|\widehat{\nu}_1 - \widehat{\nu}_2\|_2^{s_h}$，由此得到$\widetilde{N}(\frac{\varepsilon}{2c_3}, \mathcal{V}_n, \|\cdot\|_\infty) \leq \widetilde{N}(\varepsilon^{1/s_\nu}(2c_3c_4)^{1/s_\nu}, \mathcal{V}_n, \|\cdot\|_2)$。根据引理 B.7，我们进一步有 $\widetilde{N}(\varepsilon^{1/s_\nu}(2c_3c_4)^{1/s_\nu}, \mathcal{V}_n, \|\|_2) \leq N(\frac{1}{2}\varepsilon^{1/s_\nu}(2c_3c_4)^{1/s_\nu}, \mathcal{V}_n, \|\cdot\|_2)$。令 $c_\nu = \frac{1}{2}(2c_3c_4)^{1/s_\nu}$，则有 $\widetilde{N}_{[]}(\frac{\varepsilon}{2c_3}, \mathcal{V}_n, \|\cdot\|_\infty) \leq N(c_\nu\varepsilon^{1/s_\nu}, \mathcal{V}_n, \|\cdot\|_2)$。

因此，最终我们可以得到

$$N_{[]}(\varepsilon, \mathcal{G}_n, \|\cdot\|_\infty) \lesssim \frac{1}{\varepsilon} N(c_\nu\varepsilon^{1/s_\nu}, \mathcal{V}_n, \|\cdot\|_2).$$

□

C 实验细节

C.1 数据集摘要

我们在表3中报告了第5.2节所用6个基准数据集的描述性统计摘要。

C.2 合成实验细节

由于假设真实模型采用 PF 形式，我们根据以下变换回归模型生成事件时间 [18]:

$$\log H(\tilde{T}) = -m(Z) + \epsilon, \tag{15}$$

其中 $H(t) = \int_0^t e^{h(s)}ds$， h 定义见（2）。误差项 ϵ 的生成方式使得 e^ϵ 具有累积风险函数 G_θ。公式（15）等价于（2） [18, 17, 45]。在我们的实验中，协变量的维度为 5，且从 $[0, 1]$上的均匀分布中独立采样。我们设定 $h(t) = t$，因此 $H(t) = e^t$。函数 $m(Z)$ 的形式设为$m(Z) = \sin(\langle Z, \beta\rangle) + \langle \sin(Z)\ \beta\rangle$，其中 $\beta =$ (0.1, 0.2, 0.3, 0.4, 0.5)。随后，删失时间 C按照以下方式生成

$$\log H(C) = -m(Z) + \epsilon_C, \tag{16}$$

该方法复用协变量 Z，并独立抽取噪声向量 ϵ_C，使删失比例控制在约 40%。我们分别使用 $n \in \{1000, 5000, 10000\}$ 生成三个数据集。

超参数配置：下面列出所有任务中使用的网络架构和优化配置：

^PF方案：对于 $\widehat{m}$ 和 h，我们在 $n = 1000$中使用64个隐藏单元，在 $n =$5000中使用128个隐藏单元，并在 $n = 10000$中使用256个隐藏单元。每个模型训练100轮，批大小为128，使用Adam进行优化，学习率为0.0001，且不使用权重衰减。

FN方案：对于这两个 $\widehat{\nu}$，我们为 $n = 1000$使用64个隐藏单元，为 $n = 5000$ 使用128个隐藏单元，为 $n = 10000$使用256个隐藏单元。每个模型训练100个训练轮数，批大小为128，使用 Adam 优化，学习率为0.0001，且不使用权重衰减。

C.3 公共数据实验的详细信息

数据集预处理 对于 METABRIC、RotGBSG、FLCHAIN、SUPPORT 和 KKBOX 数据集，我们采用 pycox 软件包中提供的版本 [46]。我们将连续特征标准化为均值为零、方差为一，并对所有分类特征进行独热编码。对于 MIMIC-III 数据集，我们遵循 [55] 中的预处理流程，该流程提取了26个特征。关注事件定义为入院后的死亡，删失时间定义为患者最后一次出院的时间。该定义与 [64]中的定义类似。不过，由于该数据集未公开发布，根据我们的实现，得到的数据集删失率要高得多（90.2%，而 SODEN 论文 [64]报告为61.0%）。由于本文的主要目的是提出 NFM 框架，我们使用自己处理的数据集版本，进一步验证 NFM 的预测性能。

超参数配置 我们遵循通用训练模板，使用多层感知机作为所有$\widehat{}$非参数函数逼近器（即，PF方案中的 $\widehat{m}$ 和 h ，以及FN方案中的 $\widehat{\nu}$），并在所有数据集上使用Adam作为优化器训练100轮。可调参数及其各自的调优范围如下：

层数（网络深度） 我们调整网络深度 $L \in \{2, 3, 4\}$。通常，两层多层感知机的性能已足够令人满意。

每层的隐藏单元数（网络宽度） 我们调整网络宽度 $W \in \{2^k, 5 \leq k \leq 10\}$。

可选的丢弃 我们可以选择以 $p \in \{0.1, 0.2, 0.3, 0.5, 0.7\}$的概率应用丢弃。
批大小 我们在 $\{128, 256, 512\}$范围内调整批大小；对于KKBOX 数据集，我们还测试了更大的批大小 $\{1024\}$。
学习率和权重衰减 我们在范围 $\{0.01, 0.001, 0.0001\}$ 内调整 Adam 的学习率和权重衰减系数。

脆弱性设定 我们测试了伽马脆弱性、Box-Cox变换脆弱性和 IGG(α) 脆弱性，使用 $\alpha \in \{0, 0.25, 0.75\}$。这里需要注意，IGG(0.5) 等价于伽马脆弱性。我们还通过实验证明，将 α 设为可学习参数并未显著提升所用评估数据集上的性能。

C.4 实现

We u使用 PyTorch 实现 NFM。源代码见补充材料。
对于基线模型：

• 我们使用 sksurv 软件包中的 Cox比例风险模型、梯度提升机和随机生存森林实现 [54], 对于 KKBOX 数据集，我们使用 XGBoost 库 [10] 实现梯度提升机和随机生存森林，这可能会导致性能下降。

• 我们使用 pycox 软件包实现 DeepSurv、CoxTime 和 DeepHit 模型。
• 我们使用 SODEN 论文中提供的官方代码 [64] 实现 SODEN。
• 我们根据自行重新实现的版本获得 SuMo 和 DeepEH 的结果。

D 附加实验

D.1 PF方案中 $m(Z)$的恢复评估

We 绘制图 2 中以 (2) 式的 m 函数为目标的经验恢复结果。结果表明 s
sati在数据量适中时，即 $n \geq 1000$，恢复效果令人满意。

D.2 生存函数的恢复评估

为评估 NFM 对生存函数的恢复效果，我们考虑如下设置：在与 C.2节 相同的数据生成框架下，我们将测试$\bar{}$ $\widehat{}$ $\bar{}$特征 Z 定义为全部 100 个留出测试点的样本均值。并绘制 $S(t|Z)$ 与真实

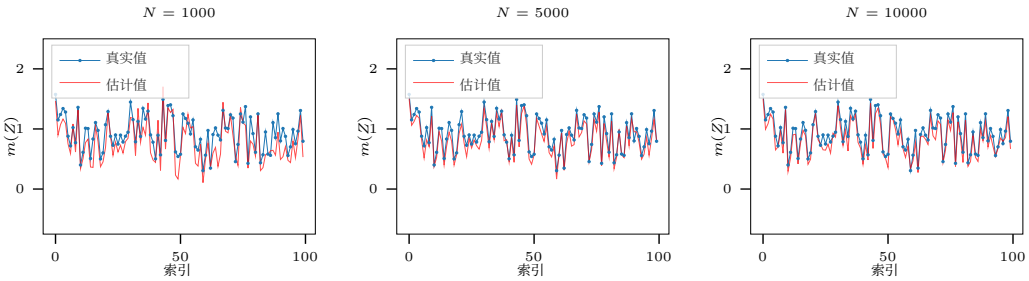


图 2：NFM 框架的 PF方案下合成数据结果的可视化，展示了式 (2) 中 m 函数的经验恢复

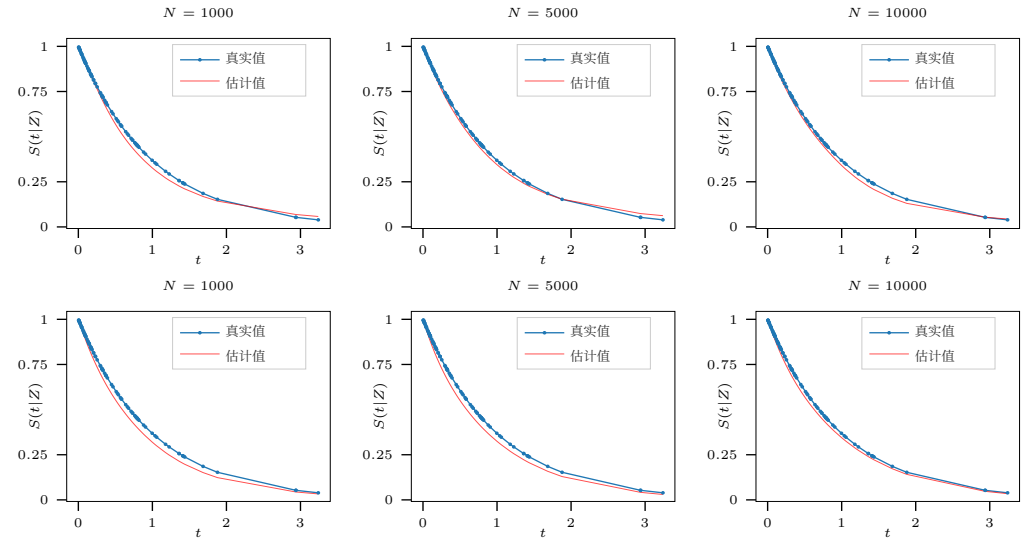


图 3：NFM 框架下合成数据结果的可视化。第一¯ ¯行中的图比较了粒子滤波方案下生存函数 $S(t|Z)$ 的经验估计与其真实值，其中 Z为 100 个留出点的特征均值。第二行中的图采用 FN方案得出，其含义与第一行中的图类似。

¯值 $S(t|Z)$ 的对比，分别针对粒子滤波和 FN方案。结果如图 3 所示。结果表明，当样本量足够大时，这两种方案都能准确估计生存函数。

D.3 合成实验的数值结果

接下来， [75], 我们报告估计生存函数相对于真实值的相对积分均方误差（RISE），并将结果列于表中。结果表明，样本量越大，拟合效果越好。此外，由于模拟中的真实模型是以 PF 模型生成的，我们发现 PF 的表现略优于 FN；鉴于 PF 在此设置下具有更准确的归纳偏置，这一结果是合理的。

表 4：合成实验中估计值生存函数的 RISE

	$N = 1000$	$N = 5000$	$N = 10000$
NFM-PF	0.0473	0.0145	0.0137
NFM-FN	0.0430	0.0184	0.0165

表5：在全部6个基准数据集上的生存预测结果，以一致性指数（%）衡量。每列中，粗体分数表示最佳结果，下划线分数表示次佳结果。最右列报告了各模型的平均排名。我们未能在两个较大的数据集 MIMIC-III 和 KKBOX 上为 DeepEH 和 SODEN 获得合理的结果，因此将它们在这些数据集上的排名设为最差。

模型	METABRIC	RotGBSG	FLCHAIN	SUPPORT	MIMIC-III	KKBOX	平均排名
Cox比例风险模型	$63.42_{\pm 1.81}$	$66.14_{\pm 1.46}$	$79.09_{\pm 1.11}$	$56.89_{\pm 0.91}$	$74.91_{\pm 0.00}$	$83.01_{\pm 0.00}$	11.33
GBM	$64.02_{\pm 1.79}$	$67.35_{\pm 1.16}$	$\mathbf{79.47}_{\pm 1.08}$	$61.46_{\pm 0.80}$	$75.20_{\pm 0.00}$	$85.84_{\pm 0.00}$	7.17
RSF	$64.47_{\pm 1.82}$	$67.33_{\pm 1.34}$	$78.75_{\pm 1.07}$	$61.63_{\pm 0.84}$	$75.47_{\pm 0.17}$	$85.79_{\pm 0.00}$	8.00
DeepSurv	$63.95_{\pm 2.12}$	$67.20_{\pm 1.22}$	$79.04_{\pm 1.14}$	$60.91_{\pm 0.85}$	$80.08_{\pm 0.44}$	$85.59_{\pm 0.08}$	8.50
CoxTime	$66.22_{\pm 1.69}$	$67.41_{\pm 1.35}$	$78.95_{\pm 1.01}$	$61.54_{\pm 0.87}$	$78.78_{\pm 0.62}$	$\mathbf{87.31}_{\pm 0.24}$	5.00
DeepHit	$66.33_{\pm 1.61}$	$66.38_{\pm 1.07}$	$78.48_{\pm 1.09}$	$\mathbf{63.20}_{\pm 0.85}$	$79.16_{\pm 0.59}$	$86.12_{\pm 0.26}$	6.50
DeepEH	$66.59_{\pm 2.00}$	$\mathbf{67.93}_{\pm 1.28}$	$78.71_{\pm 1.44}$	$61.51_{\pm 1.04}$	–	–	6.33
SuMo-net	$64.82_{\pm 1.80}$	$67.20_{\pm 1.31}$	$79.28_{\pm 1.02}$	$62.18_{\pm 0.78}$	$76.23_{\pm 1.06}$	$84.77_{\pm 0.02}$	7.00
SODEN	$64.82_{\pm 1.05}$	$66.97_{\pm 0.75}$	$79.00_{\pm 0.96}$	$61.10_{\pm 0.59}$	–	–	10.17
SurvNode	$64.64_{\pm 4.91}$	$67.30_{\pm 1.65}$	$76.11_{\pm 0.98}$	$55.37_{\pm 0.77}$	–	–	11.50
DCM	$65.76_{\pm 1.25}$	$66.75_{\pm 1.35}$	$78.61_{\pm 0.79}$	$62.19_{\pm 0.95}$	$76.45_{\pm 0.34}$	$83.48_{\pm 0.07}$	8.33
DeSurv	$65.88_{\pm 2.02}$	$67.30_{\pm 1.45}$	$78.97_{\pm 1.64}$	$61.47_{\pm 0.97}$	$\mathbf{80.97}_{\pm 0.30}$	$86.11_{\pm 0.05}$	5.67
NFM-PF	$64.98_{\pm 1.87}$	$\underline{67.77}_{\pm 1.35}$	$\underline{79.45}_{\pm 1.03}$	$61.33_{\pm 0.83}$	$79.56_{\pm 0.15}$	$86.23_{\pm 0.01}$	$\underline{4.67}$
NFM-FN	$\mathbf{66.63}_{\pm 1.82}$	$67.73_{\pm 1.29}$	$79.29_{\pm 0.93}$	$\underline{62.21}_{\pm 0.41}$	$\underline{80.18}_{\pm 0.20}$	$\underline{86.61}_{\pm 0.01}$	**2.16**

表 6：NFM模型相较于对应非脆弱性模型的相对改进，分别以积分布里尔分数、积分负对数似然和一致性指数衡量。

数据集	NFM-PF 与 DeepSurv			NFM-FN 与 SuMo-net		
	IBS	积分负对数似然	一致性指数	IBS	积分负对数似然	一致性指数
METABRIC	+1.33%	+1.56%	+1.61%	+2.30%	+3.08%	+2.79%
RotGBSG	+1.11%	+0.95%	+0.84%	+0.62%	+0.40%	+0.79%
FLCHAIN	+1.29%	+1.32%	+0.52%	+0.20%	+0.27%	+0.01%
SUPPORT	+0.31%	+0.23%	+0.69%	+2.22%	+1.76%	+0.05%
MIMIC-III	+12.38%	+12.15%	−0.64%	+6.18%	+5.56%	+5.18%
KKBOX	+2.56%	+0.51%	+0.75%	+8.20%	+10.38%	+2.17%

D.4 基于一致性指数（C-index）的性能评估

一致性指数（C-index） [1] 是生存分析中常用的另一种评估指标。一致性指数用于估计：对于随机选取的一对个体，两者预测生存时间的排序与真实生存时间的排序一致的概率。形式上，一致性指数定义为

$$\mathcal{C} = \mathbb{P}\left[\widehat{S}(T_i \mid Z_i) < \widehat{S}(T_j \mid Z_j) \mid T_i < T_j, \delta_i = 1\right]. \tag{17}$$

我们在表5中报告了6个基准数据集上的一致性指数性能评估。从表5可以看出，在所选的6个数据集上，一致性指数指标并未显示出明确的优胜者。我们推测，这一现象与一致性指数和基于似然的学习目标之间相关性较弱密切相关，正如 [56] 中观察到的那样。因此，我们计算了各模型的平均排名，以综合评估其性能，结果如表5最后一列所示。结果表明，两种 NFM模型的平均表现更好。

D.5 脆弱性的优势

我们根据表1、表2和表5中的结果，计算了 NFM-PF 和 NFM-FN 相对于其非脆弱性对应模型 DeepSurv [42] 和 SuMo-net [56] 的（相对）性能提升。结果如表6所示。结果表明，纳入脆弱性能够带来显著改进，尤其是在积分布里尔分数和积分负对数似然度量方面，两个 NFM模型的相对性能提升均超过 10%。在积分布里尔分数和积分负对数似然度量方面，所有数据集上的性能改进都保持一致。唯一出现性能下降的情况，是在 MIMIC-III 数据集上使用一致性指数进行评估时。这一现象也不难理解：由于 DeepSurv 模型在模型训练中采用了一种

部分似然（PL）的变体，正如先前研究 [62] 所指出的，PL 类型的目标函数与排序问题密切相关。由于一致性指数可视为一种特定类型的排序度量，因此，使用尺度敏感似然目标进行训练的 NFM 型模型相比之下，DeepSurv 可能具有更好的排序性能。

D.6 局限性

本节从理论和实证两个角度讨论本文的局限性。

理论局限性 尽管我们已为 NFM 的估计性质建立了形式化的统计保证，但这些保证并不一定意味着 NFM 在 IBS 和 INBLL 等度量下能够在预测任务中表现良好。按照经典学习理论的思路，要为预测提供保证，理想的做法是直接优化底层度量或其替代指标；但在生存问题中，这很难实现，因为这些度量同时涉及事件时间模型和删失时间工作模型。据我们所知，目前唯一旨在解决这一问题的尝试是逆加权生存博弈方法 [33]。然而，[33] 中的作者并未提供严格的学习理论论断，这为未来研究指明了一个有前景的方向。

实证局限性 尽管实证结果表明，NFM 是一种具有竞争力的预后生存模型，但我们从实验结果中发现，几乎无法获得相对于基线方法具有统计显著性的提升；这也是以往神经生存回归研究中常见的问题 [75, 56]。我们推测，这一现象主要由两个原因造成：首先，缺乏公开的大规模生存数据集，无法对神经生存模型进行可扩展的评估。其次，对于目前大多数可用数据集，都没有权威的训练-测试划分。因此，大多数实验只能在中等规模的数据集上进行交叉验证，导致建模算法的结果变异性相对较大。为此，我们建议生存分析领域以隐私保护方式发布更多配有标准训练-测试划分的大规模脱敏数据集；我们相信这将极大推动神经生存建模的发展。