TEMPORAL RESOLUTION AND EFFECT FLEXIBILITY: THE $K \times M$ GEOMETRY OF SURVIVAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Discrete-time survival models can be highly sensitive to interval count K, usually selected empirically without revealing the underlying mechanism. We trace this sensitivity to a larger model-optimisation problem over temporal resolution K and effect flexibility M, which governs how freely covariate effects vary across intervals. We make M explicit through a unified temporal basis model family, recovering the effect structures of two existing boundaries: time-invariant models such as proportional hazards at $M = 1$, and interval-specific models such as Nnet-survival and DeepHit at $M = K$. Under baseline and effect regularities r and s, our theory explains K-sensitivity through separate approximation–estimation balances $K^{-2r} + K/n$ for temporal resolution and $M^{-2s} + Md/n$ for effect flexibility, yielding distinct oracle scales $K^\star \asymp n^{1/(2r+1)}$ and $M^\star \asymp (n/d)^{1/(2s+1)}$, where n is the sample size and d the representation dimension. Since these scales generally differ, tuning existing $M = 1$ or $M = K$ models explores only fixed paths and cannot guarantee the joint optimum. We prove that held-out joint validation over (K, M) attains the oracle rate without knowing r or s, making optimisation over the full geometry practical. Controlled studies confirm the predicted geometry. Across six real-world datasets and four deep-survival heads, joint selection lowers mean test loss on held-out outer folds relative to $M = K$ in 20 of 24 dataset–head strata and favours $M < K$ in 353 of 360 evaluations. Thus, K-sensitivity fundamentally reflects a larger joint approximation–estimation geometry of temporal resolution and effect flexibility: existing models trace only fixed paths through it, while reaching the joint optimum requires selecting (K, M), not K alone.

1 INTRODUCTION

Survival prediction learns a conditional event-time distribution from right-censored data. Discrete-time models, widely used in deep survival learning, offer flexible and computationally convenient prediction without strong global parametric assumptions (Gensheimer & Narasimhan, 2019; Lee et al., 2018). They discretise the (continuous) follow-up horizon into K disjoint intervals over which the survival function is then estimated, making K a consequential modelling choice. Prior work recognises that finer grids increase both temporal resolution and model size, but primarily addresses this trade-off through empirical selection of K (Kvamme & Borgan, 2021; Suresh et al., 2022). This leaves a fundamental question unresolved: what statistical mechanism drives K-sensitivity?

To uncover this mechanism, we first rethink discrete-time survival models as the two-stage pipeline illustrated in Figure 1. The representation-to-interval mapping transforms covariates or a learned representation $Z \in \mathbb{R}^d$ into interval predictors $\xi_1, \ldots, \xi_K$, from which the interval-to-survival mapping constructs the survival function. The latter determines how these predictors define a survival function, while the former governs how covariate effects are parameterised and shared across intervals. Within the first mapping, time-invariant models such as the grouped proportional hazards model (Prentice & Gloeckler, 1978) share one covariate-effect vector $w_j \equiv w$ across all intervals, whereas interval-specific models such as DeepHit permit a distinct w_j in each interval (Lee et al., 2018). Increasing K therefore refines temporal resolution and adds interval intercepts α_j in both designs, but introduces independently parameterised w_j only in the latter. Prior empirical work likewise reports greater bin-count sensitivity for an interval-specific model than for a time-invariant model (Sloma et al., 2021). Thus, K-sensitivity may reflect not only temporal resolution but also how freely covariate effects vary across intervals, a property we call *effect flexibility*.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

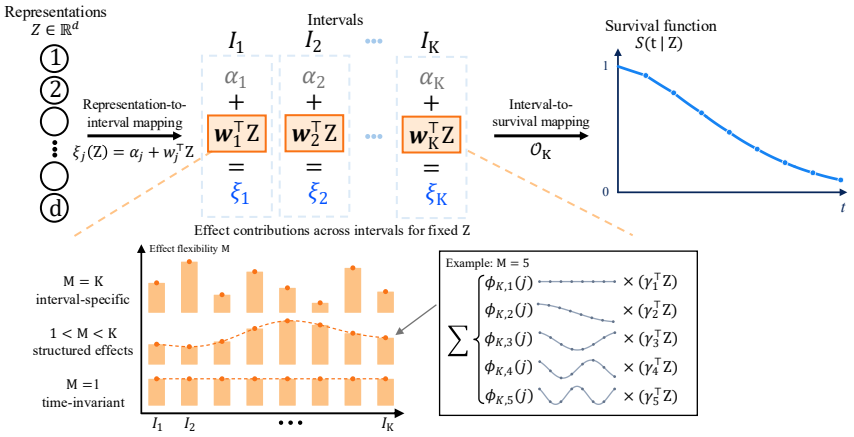

Figure 1: **Representation-to-interval mapping and temporal effect flexibility.**

We parameterise effect flexibility through a unified temporal-basis family indexed by M. The interval effect vectors $w_1, \ldots, w_K$ form a temporal sequence, which we expand in a discrete-cosine basis:

$$\xi_j(Z) = \alpha_j + \sum_{m=1}^{M} \phi_{K,m}(j)\big(\gamma_m^\top Z\big), \qquad 1 \le M \le K. \tag{1}$$

Each $\phi_{K,m}$ is a discrete-cosine basis function, ordered from constant to increasingly fine variation across intervals, and γ_m is a vector of weights for Z. Because the intercepts α_j remain unrestricted, M controls only the effect flexibility. At $M = 1$, the constant basis enforces time-invariant effects, while $M = K$ permits unrestricted interval-specific effects, with nested models between the two boundaries. Thus, existing time-invariant and interval-specific models tune K under the constraints $M = 1$ and $M = K$, respectively. Allowing M to vary exposes the larger (K, M) geometry and turns empirical K-selection into joint optimisation over temporal resolution and effect flexibility.

To characterise this geometry, we analyse the approximation–estimation cost of increasing K and M. A finer temporal grid better approximates the continuous event-time distribution but requires estimating more interval intercepts α_j, while greater effect flexibility better approximates time-varying effects but requires estimating more effect parameters. Under baseline and effect regularities r and s, an exact prediction-error decomposition has leading terms:

$$\underbrace{K^{-2r} + \frac{K}{n}}_{\text{temporal resolution}} + \underbrace{M^{-2s} + \frac{Md}{n}}_{\text{effect flexibility}},$$

where n is sample size and d is representation dimension. These balances yield generally distinct oracle scales $K^\star \asymp n^{1/(2r+1)}$ and $M^\star \asymp (n/d)^{1/(2s+1)}$, explaining why the existing $M = 1$ and $M = K$ constraints need not attain the joint optimum. In theory, we establish minimax optimality and show that held-out joint validation attains the oracle rate without knowing r or s.

Beyond the global oracle scales, we derive a local information rule for effect flexibility: an additional temporal basis function is supported only when its event-weighted signal exceeds its d/n estimation cost. We then extend discrete selection of M by hard truncation to continuous temporal smoothing through effective dimension $q(S)$, which recovers $q(S) = M$ under truncation. Finally, we relax the equal-width interval assumption and show that regular nonuniform partitions preserve the K- and M-rate exponents. The resolution–flexibility geometry therefore persists when effect complexity is controlled continuously and intervals are placed unequally.

Extensive controlled studies recover the predicted decomposition, scaling, and information threshold, with the local rule identifying the prediction-error-minimising effect dimension in 44 of 48 regimes. Across six real-world datasets and four deep-survival heads, nested selection chooses $M < K$ in 353 of 360 outer-test runs and lowers mean test loss relative to $M = K$ in 20 of 24 dataset–head strata. These results show that, at a given K, the effect flexibility favoured by the data can differ from that imposed by the architecture. Joint (K, M) selection therefore replaces tuning along an existing boundary with optimisation over the larger geometry, providing a route towards the joint optimum that neither boundary can guarantee.

Our contributions are fourfold:

- We introduce a unified temporal-basis family that exposes effect flexibility M and recovers time-invariant ($M = 1$) and interval-specific ($M = K$) effect structures.
- We develop the first statistical theory of K-sensitivity, revealing the model-dependent paths through the joint $K \times M$ prediction-error geometry.
- We establish distinct oracle scales, the matching minimax rate, adaptive joint validation, a local information rule, and an effective-dimension extension to temporal regularisation.
- We validate the geometry through controlled studies and nested evaluation across six datasets and four deep-survival heads, demonstrating the practical value of joint (K, M) selection.

2 Background and Related Work

2.1 Interval outputs under a common continuous-time target

We call a survival model interval-output when its prediction head produces K interval outputs. This class includes discrete-time survival models and continuous-time interval-parameterised models, such as PC-Hazard (Kvamme & Borgan, 2021). The K-sensitivity studied here concerns this broader class, with discrete-time models providing its most familiar examples. Given interval predictors $\boldsymbol{\xi}(Z)$, we write $S(\cdot \mid Z) = \mathcal{O}_K\{\boldsymbol{\xi}(Z)\}$ for the model-specific interval-to-survival map.

Let T be an event time, C a censoring time, X observed covariates, and $Z = f_\theta(X) \in \mathbb{R}^d$ a fixed or learned representation. Over a fixed horizon $[0, \tau_{\max}]$, we observe $Y = \min(T, C, \tau_{\max})$ and $\Delta = \mathbf{1}\{T \le C, T \le \tau_{\max}\}$. The theory conditions on Z and assumes $T \perp C \mid Z$. For a candidate hazard function $\lambda(\cdot|Z)$, the continuous right-censored negative log-likelihood (RC-NLL) is

$$\ell_{\mathrm{RC}}(\lambda) = \int_0^Y \lambda(t \mid Z)\, dt - \Delta \log \lambda(Y \mid Z), \tag{2}$$

which is a proper score for the conditional event-time distribution (Rindt et al., 2022; Yanagisawa, 2023). For a grid $\pi_K = (\tau_0, \ldots, \tau_K)$, where $0 = \tau_0 < \cdots < \tau_K = \tau_{\max}$, define interval $I_j = (\tau_{j-1}, \tau_j]$ with width $|I_j| = \tau_j - \tau_{j-1}$. A piecewise-exponential reconstruction sets $\lambda(t \mid Z) = \lambda_j(Z)$ on I_j (Holford, 1980; Laird & Olivier, 1981). If $Y \in I_j$, then

$$\ell_{\mathrm{RC}} = \sum_{l<j} \lambda_l(Z)|I_l| + \lambda_j(Z)(Y - \tau_{j-1}) - \Delta \log \lambda_j(Z). \tag{3}$$

This reconstruction retains the exact exposure within the terminal interval. Following Kvamme & Borgan (2021), the interval-to-survival maps for discrete-hazard and PMF outputs use piecewise-constant hazards that preserve survival at interval boundaries. Consequently, models with different K, M, and $\mathcal{O}_K$ can be evaluated against the same continuous-time target.

Continuous-time models optimise censored likelihood without an output grid through monotone survival networks (Rindt et al., 2022), neural ordinary differential equations (Tang et al., 2022), tractable mixture densities (Han et al., 2022), or importance-sampled density models (Ketenci et al., 2023). They replace K with architectural, distributional, or numerical choices carrying different approximation–estimation trade-offs. Our analysis instead examines how interval-output models change with K through temporal resolution and effect flexibility. We use equal-width grids to isolate the $K \times M$ geometry, while Section 4.6 shows that regular partitions preserve its rate exponents.

2.2 Temporal resolution and effect flexibility

In interval-output survival models, K sets both the temporal resolution and output-head dimension. Empirical studies show that performance can vary substantially with the interval counts, motivating its treatment as a hyperparameter. Both coarse and fine grids can be suboptimal (Kvamme & Borgan, 2021; Sloma et al., 2021; Suresh et al., 2022; Ketenci et al., 2023). Related work estimates the number or placement of intervals through Bayesian grid selection (Demarqui et al., 2008), event-quantile discretisation (Kvamme & Borgan, 2021), adaptive cutpoints (Hickey et al., 2024), or change-point detection (Cooney & White, 2021). These approaches establish the practical importance of grid

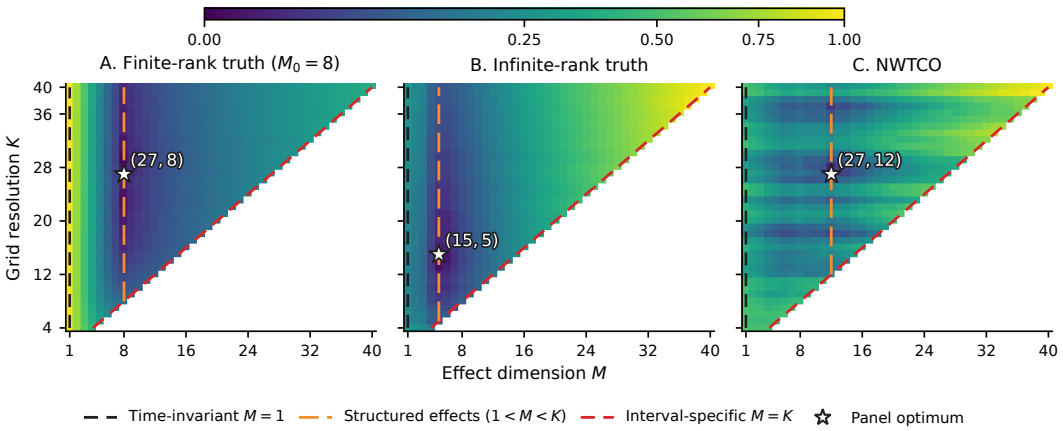

Figure 2: **Empirical $K \times M$ prediction-error geometry.** RC-NLL is min–max normalised within each panel and shown on a square-root colour scale with minima marked by stars.

design, but generally vary K or its cutpoints while holding the effect structure fixed. They therefore do not separate temporal resolution from the effect complexity accompanying grid refinement. Related work also constrains the interval intercepts α_j through parsimonious temporal specifications (Willett & Singer, 2003, Chapter 12). We leave $\boldsymbol{\alpha}$ unrestricted and focus on cross-interval covariate effects.

Effect flexibility has been studied from the complementary direction. Proportional-hazards models share one effect vector across time (Cox, 1972). Varying-coefficient and reduced-rank models represent temporal effects through lower-dimensional functions (Hastie & Tibshirani, 1993; Tian et al., 2005; Kauermann, 2005; Perperoglou, 2013), while multi-task and piecewise-additive models control interval-to-interval variation through sharing, regularisation, or smooth components (Yu et al., 2011; Fotso, 2018; Bender et al., 2018; Kopper et al., 2022). Recent approaches further allow different covariates to exhibit different degrees of temporal variation (Knaus et al., 2026; Hagemann et al., 2025). Grid design and effect modelling therefore expose two distinct axes, but existing work does not characterise their joint prediction-error geometry or select them together. We make effect flexibility explicit through M, then analyse and select both coordinates under a common continuous-time target.

3 The (K, M) Temporal-Basis Model Family

We now formalise the temporal basis construction introduced in Eq. (1). For the equal-width grid, let $t_j = (j - \frac{1}{2})/K$ denote the normalised midpoint of interval j. We use the ordered Discrete Cosine Transform type-II (DCT-II) basis

$$\Phi_{j1} = 1, \qquad \Phi_{jm} = \sqrt{2}\cos\{\pi(m-1)t_j\}, \quad m \geq 2,$$

normalised so that $K^{-1}\boldsymbol{\Phi}_{K,M}^\top \boldsymbol{\Phi}_{K,M} = I_M$. Stacking the interval predictors gives

$$\boldsymbol{\xi}_{K,M}(Z) = \boldsymbol{\alpha} + \boldsymbol{\Phi}_{K,M}\boldsymbol{\Gamma} Z, \tag{4}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^K$, $\boldsymbol{\Gamma} \in \mathbb{R}^{M \times d}$, and $Z \in \mathbb{R}^d$ excludes a constant feature. Equivalently, $W = \boldsymbol{\Phi}_{K,M}\boldsymbol{\Gamma}$ restricts each feature's interval-effect trajectory to the first M basis functions, reducing effect parameters from Kd to Md while preserving K interval intercepts. As the DCT basis progresses from constant to higher temporal frequencies, increasing M adds finer effect variation without changing the grid. These boundary identities require only a nested basis with a constant first function and an invertible complete basis. The DCT is a particular implementation whose frequency ordering makes M interpretable and, under s-regular effects, yields the $O(M^{-2s})$ effect-approximation rate.

Proposition 1 (Nested effect structures)**.** *Let $\mathcal{F}_{K,M}$ denote the model class in Eq. (4). For fixed K,*

$$\mathcal{F}_{K,1} \subseteq \cdots \subseteq \mathcal{F}_{K,K}, \quad \mathcal{F}_{K,1} = \{\boldsymbol{\alpha} + \mathbf{1}_K w^\top Z\}, \quad \mathcal{F}_{K,K} = \{\boldsymbol{\alpha} + WZ : W \in \mathbb{R}^{K \times d}\}.$$

Thus, $M = 1$ and $M = K$ recover time-invariant and interval-specific effect structures, respectively.

Proof. See Appendix B.1.

Proposition 1 concerns the representation-to-interval map, independently of the model-specific interval-to-survival map $\mathcal{O}_K$. Its $M = 1$ boundary recovers the effect structures of CoxPH (Cox, 1972) and DeepSurv (Katzman et al., 2018), while its $M = K$ boundary recovers those of Nnet-survival (Gensheimer & Narasimhan, 2019), PC-Hazard (Kvamme & Borgan, 2021), PMF/DeepHit (Kvamme & Borgan, 2021; Lee et al., 2018), and, up to an invertible cumulative reparameterisation, MTLR/N-MTLR (Yu et al., 2011; Fotso, 2018). These correspondences are exact at the effect-structure level. $\mathcal{O}_K$, training objectives, regularisers, and censoring implementations remain method-specific. Appendix B.2 details the corresponding conditions for recovering these effect structures.

For each fixed K, Proposition 1 gives a nested sequence of models as M increases. Collecting these families across K yields the triangular (K, M) design space shown in Figure 2, with time-invariant models on the $M = 1$ edge and interval-specific models on the $M = K$ diagonal. The observed interior minima show that selecting K along either boundary need not reach the joint optimum.

To study the approximation–estimation trade-offs across (K, M), we follow Kvamme & Borgan (2021) and specialise $\mathcal{O}_K$ to a piecewise-constant-hazard map, defining the continuous log-hazard as $\eta_{K,M}(t, Z) = \xi_j(Z)$ for $t \in I_j$. The resulting piecewise-exponential reconstruction places all (K, M) models against the same continuous-time prediction target. We estimate $(\boldsymbol{\alpha}, \boldsymbol{\Gamma})$ by minimising the continuous RC-NLL in Eq. (3). Section 4 analyses the resulting prediction-error geometry, while Section 5 applies the temporal-basis constraint to multiple deep-survival heads.

4 Statistical Geometry

This section develops the statistical theory of the $K \times M$ geometry. We first show how error arises from discretising time, restricting temporal effect variation, and learning from finite data. We then derive how the optimal K and M depend on sample size n, representation dimension d, and baseline and effect smoothness r and s, prove the resulting rate is minimax optimal, and show that validation adapts without knowing r or s. Finally, we extend the analysis from hard truncation in M to smooth regularisation and from equal-width to nonequal grids.

4.1 Exact censored-risk decomposition

We rescale the analysis horizon $[0, \tau_{\max}]$ to $[0, 1]$ and condition on the fixed representation Z. Assume the ground-truth log-hazard is $\eta_0(t, Z) = a_0(t) + b_0(t)^\top Z$, with baseline a_0 and time-varying effects b_0. Let $\mathcal{R}(\eta) = \mathbb{E}[\ell_{\mathrm{RC}}(e^\eta)]$ denote population RC-NLL and $\mathcal{E}(\eta) = \mathcal{R}(\eta) - \mathcal{R}(\eta_0)$ its excess risk. Since $\mathcal{R}(\eta_0)$ is constant across models, minimising $\mathcal{E}$ is equivalent to minimising population RC-NLL. All results below therefore concern this excess risk.

Let η^*_K and $\eta^*_{K,M}$ be the best population approximations within $\mathcal{F}_{K,K}$ and $\mathcal{F}_{K,M}$, respectively, and let $\widehat{\eta}_{K,M}$ be the empirical RC-NLL minimiser over $\mathcal{F}_{K,M}$. Intuitively, the path $\eta_0 \to \eta^*_K \to \eta^*_{K,M} \to \widehat{\eta}_{K,M}$ separates error from representing the continuous log-hazard on K intervals, restricting effects to M modes, and estimating from finite data. Let D_0 denote the censored-risk Bregman divergence induced by $\mathcal{R}$ (Bregman, 1967). The nesting in Proposition 1 yields the exact decomposition below.

Theorem 2 (Exact censored-risk decomposition)**.** *Under $T \perp C \mid Z$ and the regularity conditions in Appendix C,*

$$\mathcal{E}(\widehat{\eta}_{K,M}) = \underbrace{D_0(\eta^*_K, \eta_0)}_{A_{\mathrm{grid}}(K)} + \underbrace{D_0(\eta^*_{K,M}, \eta^*_K)}_{A_{\mathrm{effect}}(K,M)} + \underbrace{D_0(\widehat{\eta}_{K,M}, \eta^*_{K,M})}_{E_{\mathrm{est}}(K,M)}. \tag{5}$$

Proof. See Appendix D.1.

Affine Bregman projections make the successive residuals orthogonal, so the three terms add without remainder. All components are nonnegative, with $A_{\mathrm{effect}}(K, K) = 0$. At fixed K, grid approximation is constant across M, leaving the trade-off between effect approximation and estimation. Thus, zero effect approximation at $M = K$ does not guarantee minimum excess risk.

4.2 Rectangular approximation and estimation

Let $r \in (0, 1]$ and $s \geq r$ denote the temporal Sobolev smoothness of a_0 and b_0. Larger values permit faster approximation by K intervals and M modes, respectively. This gives the following bound:

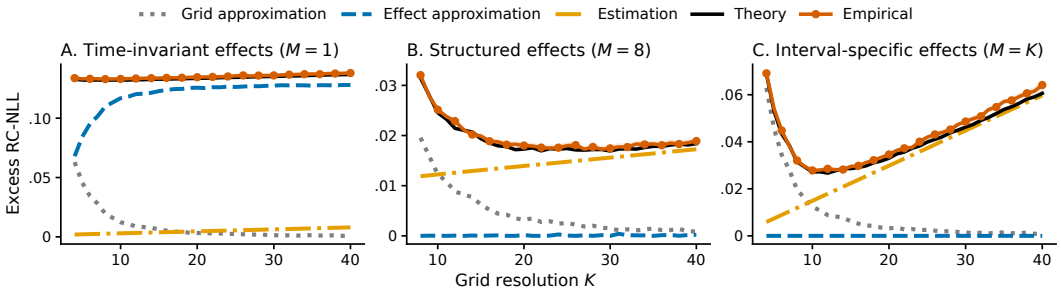

Figure 3: **Prediction-error decomposition along three effect paths.** The black curve is the theoretical sum $A_{\text{grid}} + A_{\text{effect}} + E_{\text{est}}$, and the orange curve is empirical excess RC-NLL.

Theorem 3 (Rectangular prediction-error bound)**.** *Under Appendix C conditions, for any $\delta \in (0,1)$, whenever $(K + Md + \log(1/\delta))/n$ is sufficiently small, with probability at least $1-\delta$,*

$$\mathcal{E}(\widehat{\eta}_{K,M}) \le C\left\{K^{-2r} + M^{-2s} + \frac{K + Md + \log(1/\delta)}{n}\right\}. \tag{6}$$

Proof. See Appendix D.3.

The leading bound on the exact decomposition is rectangular: K trades grid approximation K^{-2r} against baseline estimation K/n, while M trades effect approximation M^{-2s} against effect estimation Md/n. Here M^{-2s} is an approximation rate, not an identity for $A_{\text{effect}}(K,M)$, whose exact value may vary with K. Under local correct specification, estimation sharpens to $(K+Md)/(2n)+o(n^{-1})$, giving costs $1/(2n)$ per interval and $d/(2n)$ per effect mode. Piecewise-constant reconstruction limits grid approximation to order K^{-2} even for smoother truths.

Figure 3 plots the exact components along three M-paths. Grid approximation decreases with K along every path. For $M = 1$, effect approximation rises towards a nonzero misspecification gap, offsetting the grid improvement and leaving prediction error high and nearly flat. For fixed $M = 8$, effect approximation is negligible and estimation increases slowly, so prediction error decreases and then stabilises after a shallow minimum. For $M = K$, effect approximation vanishes but estimation increases steeply, producing a pronounced U-shaped error curve. The empirical errors closely follow the theoretical sums, confirming that error behaviour depends on how M changes with K.

4.3 Information-supported temporal modes

We next resolve the global geometry mode by mode by comparing the information carried by an additional temporal effect mode with its estimation price.

Proposition 4 (Local information-supported mode rule)**.** *Under local correct specification and the regularity conditions in Appendix C, let $\mathcal{R}_m$ be the local risk with m fitted effect modes after profiling the unrestricted baseline. If E_m is the information energy of the next d-dimensional mode block, then in local information-orthonormal coordinates,*

$$\mathcal{R}_{m+1} - \mathcal{R}_m = -\frac{E_m}{2} + \frac{d}{2n} + o(n^{-1}). \tag{7}$$

The additional mode is therefore supported to first order when $E_m > d/n$.

Proof. See Appendix D.4.

The rule makes effect flexibility information-limited: a mode is supported only when its information energy E_m exceeds the estimation threshold d/n. Higher representation dimension raises this threshold, while censoring and weakly informative directions reduce E_m. Under misspecification, d is replaced by the corresponding sandwich effective-dimension increment. Thus, M should increase only while successive temporal directions carry sufficient predictive information.

4.4 Optimal scales and adaptive selection

The bound in Eq. (6) separates into two balances. Temporal resolution balances grid approximation K^{-2r} against baseline estimation K/n, yielding $K \asymp n^{1/(2r+1)}$. Effect flexibility balances effect

approximation M^{-2s} against effect estimation Md/n, yielding $M \asymp (n/d)^{1/(2s+1)}$. Because these balances involve different regularities and estimation prices, their oracle scales generally differ.

Corollary 5 (Anisotropic scales and adaptation)**.** *Over the corresponding bounded survival class, when $d = o(n)$,*

$$\inf_{\widetilde{\eta}} \sup_{\eta_0} \mathbb{E}\mathcal{E}(\widetilde{\eta}) \asymp n^{-2r/(2r+1)} + (d/n)^{2s/(2s+1)}, \tag{8}$$

attained at

$$K \asymp n^{1/(2r+1)}, \qquad M \asymp (n/d)^{1/(2s+1)}. \tag{9}$$

Held-out joint selection over the admissible $K \times M$ triangle attains the same risk order up to a lower-order logarithmic term without knowing r or s.

Proof. See Appendices D.7 and D.8.

The separate scales show how sample size, temporal regularity, and representation dimension shape the $K \times M$ geometry. More data supports finer grids and richer temporal effects, while smoother components require fewer degrees of freedom and higher representation dimension increases the cost of each effect mode. Joint validation adapts to these unknown conditions over a finite candidate triangle $\mathcal{I}_n$ with only $O\{\log|\mathcal{I}_n|/n\}$ selection cost, providing a practical route to the oracle risk without coupling M to K.

4.5 Effective dimension and forced coupling

We extend the hard mode selection in Eq. (4) to spectral smoothing by weighting all K modes: $\sum_{m=1}^{K} s_m(\rho)\boldsymbol{\phi}_{K,m}\big(\boldsymbol{\gamma}_m^\top Z\big)$, where $\boldsymbol{\phi}_{K,m}$ is temporal mode m and $\boldsymbol{\gamma}_m^\top$ its full-basis coefficient row. Hard truncation uses $s_m = \mathbf{1}\{m \le M\}$, recovering Eq. (4), whereas smooth shrinkage uses fractional weights $s_m(\rho) \in [0,1]$. For the corresponding linear smoother S, define $q(S) = \|S\|_{\mathrm{HS}}^2 = \sum_{m=1}^{K} s_m(\rho)^2$. In the Gaussian experiment of Appendix D.5, the effect balance $M^{-2s} + Md/n$ in Eq. (6) has the exact smoother analogue

$$\frac{1}{2}\left\{B(S) + \frac{d}{n}q(S)\right\}, \tag{10}$$

where $B(S)$ is the worst-case squared effect bias. Hard truncation recovers $B(S) \asymp M^{-2s}$ and $q(S) = M$, while the unrestricted head has $B(S) = 0$ and $q(S) = K$. For s-regular effects, balancing smoother bias and estimation yields $q^* \asymp (n/d)^{1/(2s+1)}$ and risk $(d/n)^{2s/(2s+1)}$.

Derivation. See Appendix D.5.

This identity makes explicit M-selection and smooth shrinkage two implementations of the same effect-flexibility coordinate. The $M = K$ boundary retains every mode, forcing $q(S) = K$ and making each grid refinement increase effect-estimation cost. We next quantify this forced coupling.

Proposition 6 (Canonical cost of forced coupling)**.** *In the canonical orthogonal benchmark, suppose baseline approximation and estimation scale as K^{-2r} and K/n. Any path satisfying $q(S_K) \ge cK$ has optimised risk of order at least*

$$\left(\frac{d+1}{n}\right)^{2r/(2r+1)}. \tag{11}$$

When $s > r$, $d \to \infty$, and $d = o(n)$, its ratio to the rectangular minimax rate diverges.

Proof. See Appendix D.6.

Together with the local mode rule, this result covers both boundaries. The $M = 1$ boundary is suboptimal when $E_1 > d/n$, while the unrestricted boundary can be suboptimal when effective flexibility grows with K. Risk therefore depends jointly on temporal resolution and effective effect flexibility, so K alone cannot determine the risk-optimal model.

4.6 Robustness to interval placement

We relax the equal-width assumption on the grid π_K. Let $h(\pi_K) = \max_j |I_j|$ and construct $\boldsymbol{\Phi}_{\pi_K,M}$ by averaging the first M continuous cosine modes over the intervals I_j.

Proposition 7 (Grid-robust $K \times M$ geometry)**.** *For any positive-width grid π_K whose averaged basis has full rank, the exact decomposition in Theorem 2 remains valid with the grid-dependent quantities indexed by π_K. Under the conditions of Theorem 3, suppose additionally that $c/K \le |I_j| \le C/K$ and the grid-specific basis and information Gram matrices are uniformly conditioned. Then, with probability at least $1-\delta$,*

$$\mathcal{E}(\widehat{\eta}_{\pi_K,M}) \le C' \left\{ h(\pi_K)^{2r} + M^{-2s} + \frac{K + Md + \log(1/\delta)}{n} \right\}. \tag{12}$$

Regular grids therefore retain the same rate exponents and oracle scales as equal-width grids.

Proof. See Appendix D.9.

This shows that the $K \times M$ geometry is not an artefact of equal-width discretisation. Within regular grids, cutpoint placement may shift the optimum without collapsing the two axes. Risk therefore remains jointly governed by both coordinates and cannot be understood from K alone.

5 Experiments

5.1 Experimental setup

Controlled studies use complementary finite-rank and localised truths: the former has exactly $M_0 = 8$ nonzero temporal effect modes, whereas the latter has an infinite DCT expansion and no exact finite-M representation. The principal design fixes $d = 8$ and $n = 3000$, evaluates all 814 pairs with $4 \le K \le 40$ and $1 \le M \le K$, and runs 50 paired repetitions against a shared independent test population of 100,000 to estimate population risk with low Monte Carlo error. Targeted studies vary n, d, censoring, and baseline and effect regularities r and s, while comparing hard dimension M with smooth effective $q(S)$. We optimise the convex likelihood by IRLS and accept only converged fits with maximum absolute mean-loss gradient below 10^{-6}. Appendix E gives complete protocols.

Across six real-world datasets (Table 1), the fixed-representation study uses standardised covariates as Z and evaluates the same 814-cell $K \times M$ triangle by L-BFGS-B over ten repeated fivefold cross-validations. Learned-representation experiments pair a shared three-layer encoder ($d = 32$) with Nnet-survival, PMF (DeepHit-NLL), N-MTLR, and PC-Hazard. For each head, we evaluate all 232 pairs with $K \in \{8, 16, 24, 32, 40, 48, 64\}$ and $1 \le M \le K$ across five outer folds and three seeds, totalling 83,520 fits. Training uses each head's native likelihood, Adam, and early stopping. Inner validation selects models, and untouched outer folds evaluate continuous RC-NLL. Code, configurations, and result summaries are provided in the supplement.

5.2 Controlled checks of the theory

The two controlled surfaces support the proposed geometry for different reasons. Under finite-rank effects, the optimum $(K, M) = (27, 8)$ has excess RC-NLL 0.017, a 38% reduction from the best interval-specific model $(10, 10)$ at 0.028. The theoretical decomposition correlates above 0.999 with empirical prediction error. Under localised effects, which have no finite DCT rank, the optimum is $(15, 5)$; at $K = 16$, $M = 5$ gives 0.014 versus 0.026 at $M = K$. Intermediate M therefore reflects information-supported regularisation, not merely recovery of a true rank. Figures 2 and 3 show the corresponding surfaces and prediction-error components, with full surfaces in Appendix E. The agreement extends beyond the location of the optimum. After removing population approximation, the empirical estimation component increases at rates within 7% of $1/(2n)$ per baseline interval and $d/(2n)$ per effect mode.

Figure 4 tests the theory at local and global scales. Panel A compares adjacent-mode risk changes with the local prediction $-E_m/2 + d/(2n)$. Across 684 regular comparisons, the normalised empirical changes have slope 0.99 and correlation 0.97 with the prediction, and only six fall on the wrong side of the threshold $E_m = d/n$. Panels B and C test the oracle scaling of the two axes. As effect regularity varies over $s \in \{0.5, 1, 2\}$, the selected M scales with n/d at empirical exponents 0.45, 0.34, and 0.17, compared with predictions 0.50, 0.33, and 0.20. Varying baseline regularity over $r \in \{0.25, 0.5, 1\}$ gives empirical K-exponents 0.70, 0.48, and 0.39, compared with 0.67, 0.50, and 0.33. All 39 regular settings with the same d/n select the same effect dimension. Together, these results support the local information price d/n and the regularity-dependent scaling of K and M.

Table 1: **Dataset characteristics and real-data joint versus $M = K$ selection.** Loss differences are scaled by 10^3 and computed as joint selection minus $M = K$ selection.

	Characteristics			Fixed representation		Learned representation			
Dataset	n	Features	Cens. (%)	$(\widehat{K}, \widehat{M})$	$10^3\Delta_{\mathrm{CV}}$	Nnet-surv	PMF-DH	N-MTLR	PC-Haz
FLCHAIN	6,524	6	75.4	$(13, 1)$	-6.5	-10.7	-11.2	$+3.2$	-10.2
GBSG	2,232	7	48.6	$(22, 2)$	-15.1	-16.9	-18.7	$+2.6$	-15.3
METABRIC	1,981	63	63.9	$(9, 1)$	-57.4	-64.5	-65.7	-6.3	-61.4
NWTCO	4,028	5	86.0	$(27, 12)$	-4.8	-11.5	-8.7	$+0.1$	-11.1
Rotterdam	2,982	6	60.4	$(22, 3)$	-8.7	-13.4	-13.2	$+0.2$	-12.3
SUPPORT	8,873	14	44.6	$(26, 4)$	-16.8	-29.0	-26.0	-1.3	-29.0

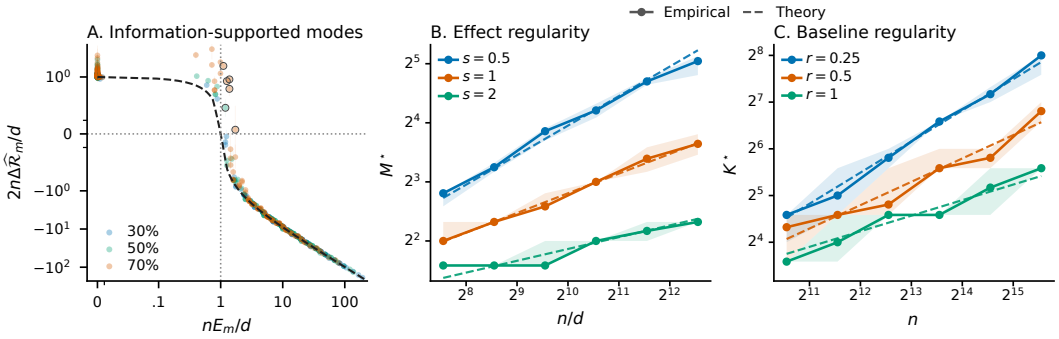

Figure 4: **Controlled checks of information and regularity scaling.** Panel A compares normalised adjacent-mode risk changes with the prediction $1 - nE_m/d$, with ringed circles marking sign disagreements. Panels B–C compare empirical optima with the predicted $M^\star$ and $K^\star$ scaling laws.

At $K = 32$, validation-selected hard truncation and smooth shrinkage outperform the unrestricted boundary in at least 99% of repetitions and select $q(S) \ll K$ (Appendix Figure 6). Truncation performs best under the finite-rank truth, whereas shrinkage performs best under the localised truth by distributing flexibility across weak modes. This supports $q(S)$ as a measure of effect-estimation cost, while the shrinkage profile determines approximation bias.

5.3 Real-data validation

Using standardised covariates as Z, the *fixed-representation study* evaluates the complete $K \times M$ triangle over ten repetitions of fivefold cross-validation, isolating temporal effect geometry from representation learning. Every dataset favours $M < K$ (range 1–12) and improves over $M = K$ (Table 1), while full-basis models with temporal smoothing prefer nonzero smoothing on all six datasets (Appendix Figure 8).

The *learned-representation study* constrains four deep-survival heads sharing a three-layer MLP ($d = 32$), trained with native NLL. PMF-DH denotes the DeepHit PMF head without ranking loss. Nested selection and outer-test evaluation use continuous RC-NLL. Joint selection chooses $M < K$ in 353 of 360 evaluations and lowers mean test loss relative to $M = K$ in 20 of 24 dataset–head strata. All 18 Nnet-survival, PMF-DH, and PC-Hazard means improve, while N-MTLR is nearly neutral. Secondary metrics like C-index and IBS broadly show the same pattern (Appendix Figure 10).

6 Conclusion

Discrete-time survival models can be highly sensitive to interval count K. Our temporal-basis family exposes a joint geometry in which K controls temporal resolution and M effect flexibility, with time-invariant and interval-specific structures as boundaries. Finer grids reduce temporal approximation error at the cost of more baseline parameters, while increasing M better approximates time-varying effects at the cost of more effect parameters. Existing model classes trace different paths through these trade-offs as K changes, while the data-supported balance need not lie on either boundary. As a modular parameterisation of cross-interval effects, our construction extends naturally to existing interval-output models. Controlled and real-world experiments demonstrate this modularity and show that joint (K, M) selection can reach optima missed by fixed model paths.

REFERENCES

Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pp. 1100–1120, 1982.

Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

Andreas Bender, Andreas Groll, and Fabian Scheipl. A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, 18(3-4):299–321, 2018.

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

Norman E Breslow and Nilanjan Chatterjee. Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):457–468, 1999.

Fabienne Comte, Stéphane Gaïffas, and Agathe Guilloux. Adaptive estimation of the conditional intensity of marker-dependent counting processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 47(4):1171–1196, 2011. doi: 10.1214/10-AIHP386.

Philip Cooney and Arthur White. Change-point detection for piecewise exponential models. *arXiv preprint arXiv:2112.03962*, 2021.

David R Cox. Regression models and life-tables. *Journal of the royal statistical society: Series B (methodological)*, 34(2):187–202, 1972.

Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403): 346–352, 2012.

Fabio N Demarqui, Rosangela H Loschi, and Enrico A Colosimo. Estimating the grid of time-points for the piecewise exponential model. *Lifetime Data Analysis*, 14(3):333–356, 2008.

Angela Dispenzieri, Jerry A Katzmann, Robert A Kyle, Dirk R Larson, Terry M Therneau, Colin L Colby, Raynell J Clark, Graham P Mead, Shaji Kumar, L Joseph Melton III, et al. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*, volume 87, pp. 517–523. Elsevier, 2012.

John A Foekens, Harry A Peters, Maxime P Look, Henk Portengen, Manfred Schmitt, Michael D Kramer, Nils Brunner, Fritz Jänicke, Marion E Meijer-van Gelder, Sonja C Henzen-Logmans, et al. The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. *Cancer research*, 60(3):636–643, 2000.

Stephane Fotso. Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512*, 2018.

Michael F Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019.

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.

Niklas Hagemann, Thomas Kneib, and Kathrin Möllenhoff. Capturing heterogeneous time-variation in covariate effects in non-proportional hazard regression models. *arXiv preprint arXiv:2501.13525*, 2025.

Xintian Han, Mark Goldstein, and Rajesh Ranganath. Survival mixture density networks. In *Machine Learning for Healthcare Conference*, pp. 224–248. PMLR, 2022.

Frank E Harrell Jr, 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.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 55(4):757–779, 1993.

Jimmy Hickey, Ricardo Henao, Daniel Wojdyla, Michael Pencina, and Matthew Engelhard. Adaptive discretization for event prediction (adept). In *International Conference on Artificial Intelligence and Statistics*, pp. 1351–1359. PMLR, 2024.

Theodore R Holford. The analysis of rates and of survivorship using log-linear models. *Biometrics*, pp. 299–305, 1980.

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):24, 2018.

Göran Kauermann. Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational statistics & data analysis*, 49(1):169–186, 2005.

Mert Ketenci, Shreyas Bhave, Noemie Elhadad, and Adler Perotte. Maximum likelihood estimation of flexible survival densities with importance sampling. In *Machine Learning for Healthcare Conference*, pp. 360–380. PMLR, 2023.

Peter Knaus, Daniel Winkler, Sebastian F Schoppmann, and Gerd Jomrich. Flexible yet sparse bayesian survival models with time-varying coefficients and unobserved heterogeneity. *Statistics in Medicine*, 45(8-9):e70458, 2026.

William A Knaus, Frank E Harrell, Joanne Lynn, Lee Goldman, Russell S Phillips, Alfred F Connors, Neal V Dawson, William J Fulkerson, Robert M Califf, Norman Desbiens, et al. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3):191–203, 1995.

Philipp Kopper, Simon Wiegrebe, Bernd Bischl, Andreas Bender, and David Rügamer. Deeppamm: Deep piecewise exponential additive mixed models for complex hazard structures in survival analysis. In *Pacific-Asia conference on knowledge discovery and data mining*, pp. 249–261. Springer, 2022.

Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime data analysis*, 27(4):710–736, 2021.

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.

Nan Laird and Donald Olivier. Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76(374):231–240, 1981.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Aris Perperoglou. Reduced rank hazard regression with fixed and time-varying effects of the covariates. *Biometrical journal*, 55(1):38–51, 2013.

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.

Ross L Prentice and Lynn A Gloeckler. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, pp. 57–67, 1978.

Shi-ang Qi, Weijie Sun, and Russell Greiner. Survivaleval: A comprehensive open-source python package for evaluating individual survival distributions. In *Proceedings of the AAAI Symposium Series*, volume 2, pp. 453–457, 2023.

David Rindt, Robert Hu, David Steinsaltz, and Dino Sejdinovic. Survival regression with proper scoring rules and monotonic neural networks. In *International conference on artificial intelligence and statistics*, pp. 1190–1205. PMLR, 2022.

Patrick Royston and Douglas G Altman. External validation of a cox prognostic model: principles and methods. *BMC medical research methodology*, 13(1):33, 2013.

Patrick Royston and Mahesh KB Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in medicine*, 21(15):2175–2197, 2002.

Martin Schumacher, Guenter Bastert, Henryk Bojar, Karl Hübner, Manfred Olschewski, Willi Sauerbrei, Cornelia Schmoor, Christa Beyerle, Rolf L Neumann, and Heinz F Rauschecker. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *Journal of Clinical Oncology*, 12(10):2086–2093, 1994.

Begoña B Sierra, Colin McLean, Peter S Hall, and Catalina A Vallejos. The c-index multiverse. *arXiv preprint arXiv:2508.14821*, 2025.

Michael Sloma, Fayeq Syed, Mohammedreza Nemati, and Kevin S Xu. Empirical comparison of continuous and discrete-time representations for survival prediction. In *Survival prediction-algorithms, challenges and applications*, pp. 118–131. PMLR, 2021.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pp. 1040–1053, 1982.

Krithika Suresh, Cameron Severn, and Debashis Ghosh. Survival prediction models: an introduction to discrete-time modeling. *BMC medical research methodology*, 22(1):207, 2022.

Weijing Tang, Jiaqi Ma, Qiaozhu Mei, and Ji Zhu. Soden: A scalable continuous-time survival model through ordinary differential equation networks. *Journal of Machine Learning Research*, 23(34): 1–29, 2022.

Lu Tian, David Zucker, and LJ Wei. On the cox model with time-varying regression coefficients. *Journal of the American statistical Association*, 100(469):172–183, 2005.

Hajime Uno, Tianxi Cai, Lu Tian, and Lee-Jen Wei. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478): 527–537, 2007.

Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and Lee-Jen Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

Sara Van de Geer. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *The Annals of Statistics*, pp. 1779–1801, 1995.

John B Willett and Judith D Singer. *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press New York, NY, USA:, 2003.

Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve mles. *The Annals of Statistics*, pp. 339–362, 1995.

Hiroki Yanagisawa. Proper scoring rules for survival analysis. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39165–39182. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/yanagisawa23a.html`.

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.

AI USE DISCLOSURE

Generative AI tools were used to assist with literature discovery, experimental design, coding and debugging, polishing the manuscript, and checking and refining mathematical claims and proofs. The authors checked all referenced literature, reviewed and tested AI-assisted code, independently verified the mathematical arguments and empirical results, and take full responsibility for the final manuscript and implementation.

A DISCUSSION, SCOPE, AND LIMITATIONS

A.1 INFORMATION-SUPPORTED COMPLEXITY

Exposing M changes ordinary tuning rather than merely adding another model variant. Searching over K along $M = K$ cannot distinguish a grid that is too coarse from covariate effects that are too flexible. Searching the full triangle can retain temporal resolution while controlling only temporal effect variation. At fixed K, selecting M is a one-dimensional nested problem. When both axes vary, the validation price grows only logarithmically with the number of candidate pairs.

The selected M is not an intrinsic rank of a dataset. It depends on the representation, basis ordering, prediction horizon, scoring rule, censoring pattern, and sample size. The localised controlled truth demonstrates that an intermediate M can be useful even when the true effect has no finite DCT rank. The scaling experiment shows how d/n and censoring move the supported boundary, while the smoothing comparison replaces integer M with a continuous effective dimension. These results favour an information-supported interpretation over a latent-rank interpretation. They also do not prescribe an interior solution. The time-invariant and interval-specific boundaries remain available whenever the data support them.

A.2 CENSORING AND THE COMMON RISK

Censoring affects both sides of the approximation–estimation balance. It changes approximation through the observable-information weight $S_T G_C \lambda_0$, and it changes estimation through the score and curvature contributed by subjects who remain observable and at risk. A parameter count alone cannot capture this temporal reweighting. The continuous RC-NLL is therefore more than a convenient evaluation metric. Together with piecewise-exponential reconstruction, it makes models with different K, M, and head maps refer to the same continuous prediction target.

The information threshold inherits this dependence on observable event information. Heavy censoring can weaken curvature, reduce effective sample information in late intervals, and destabilise empirical mode selection. The local rule is consequently a first-order guide rather than an exact finite-sample decision rule in nonregular regimes. The controlled scaling study reports these difficult regimes explicitly instead of treating all fitted cells as equally regular.

A.3 SCOPE OF THE THEORY AND EXPERIMENTS

The formal results concern interval-indexed log-hazard heads conditional on a fixed representation Z. Joint encoder training can change the representation dimension, geometry, and stochastic complexity, so it is not covered by the current bounds. The risk identities also require conditionally independent censoring. Under representation misspecification, an additional structural approximation term precedes the temporal decomposition. Under censoring misspecification, the displayed observable-information geometry need not identify the target event-time risk.

The learned-representation experiments examine Nnet-survival, PMF, PC-Hazard, and N-MTLR under a shared reconstruction and evaluation protocol. They test whether the effect-dimension principle transfers beyond the canonical log-hazard family, but they do not extend the formal theory to every link or jointly learned encoder. Continuous-time likelihood models are also outside this theory. Such models replace K with architectural, distributional, or numerical choices that carry different approximation–estimation or computational trade-offs.

The DCT is one transparent nested implementation rather than a uniquely optimal basis. It can be inefficient when effects contain abrupt or strongly localised changes. Other ordered bases, reduced-

rank maps, or temporal smoothers may be preferable. A nominally full head can also have low effective dimension when regularisation or early stopping controls its temporal variation. The smoother theory captures this case through $q(S)$ and prevents nominal output width from being confused with statistical flexibility.

A.4 Interval placement and extensions

The placement result is uniform only over regular partitions with controlled cell widths, information mass, and basis conditioning. Arbitrary cutpoint configurations can leave large unresolved cells or create nearly uninformative intervals. Continuously learned cutpoints also introduce selection complexity not covered by the finite-candidate validation argument. Regular placement preserves the two rate exponents, but it can change approximation constants, local information energies, and the selected pair (K, M).

Natural extensions include joint analysis of learned representations, continuous cutpoint optimisation, covariate-specific effect dimensions, and adaptive temporal bases. Another direction is to transfer the common-risk decomposition to continuous-time neural architectures by identifying their corresponding effective temporal complexity. These extensions would broaden the principle developed here without treating output resolution and effect flexibility as the same statistical resource.

B Boundary Effect Structures and Existing Survival Models

B.1 Proof of Proposition 1

Proof. The first DCT column is $\mathbf{1}_K$, so $M = 1$ gives $\boldsymbol{\xi}_{K,1}(Z) = \boldsymbol{\alpha} + \mathbf{1}_K w^\top Z$. Zero-padding the rows of $\boldsymbol{\Gamma}$ proves $\mathcal{F}_{K,M} \subseteq \mathcal{F}_{K,M+1}$. At $M = K$, orthogonality gives $\boldsymbol{\Phi}_{K,K}^{-1} = K^{-1}\boldsymbol{\Phi}_{K,K}^\top$. Hence every $W \in \mathbb{R}^{K\times d}$ is recovered by taking $\boldsymbol{\Gamma} = K^{-1}\boldsymbol{\Phi}_{K,K}^\top W$. This proves nesting and both boundary identities. □

B.2 Correspondence with existing survival models

Consider a survival head that applies a fixed output transformation to affine interval predictors $\boldsymbol{\xi}(Z) = \boldsymbol{\alpha} + WZ$. Any nonlinear feature extractor is absorbed into Z. Proposition 1 identifies its time-invariant effect structure with the $M = 1$ boundary when $W = \mathbf{1}_K w^\top$, and its unrestricted interval-specific effect structure with $M = K$. These correspondences concern only the representation-to-interval parameterisation; the interval-to-survival map, native likelihood, regularisation, and censoring implementation remain method specific.

For MTLR, let $\xi_1(Z), \ldots, \xi_K(Z)$ be its interval-task scores. Its pre-softmax category scores satisfy

$$u_j(Z) = \sum_{\ell=j}^{K} \xi_\ell(Z), \qquad u_{K+1}(Z) = 0, \qquad p_j(Z) = \frac{e^{u_j(Z)}}{\sum_{\ell=1}^{K+1} e^{u_\ell(Z)}}. \tag{13}$$

The reverse cumulative map is triangular and invertible because $\xi_j = u_j - u_{j+1}$. Consequently, at $M = K$, MTLR's representation-to-interval scores span the same unrestricted affine PMF-logit class while retaining its distinct parameterisation and regularisation.

These effect-structure correspondences establish a common interval parameterisation rather than a common probability model. The formal theory specialises this family to interval log-hazards, while Appendix F.3 specifies the native transformations and common continuous-time evaluation used in the learned-representation experiments.

Table 2: Existing survival-model effect structures represented at the two boundaries.

Model	Native construction	Boundary	Correspondence scope
Piecewise-exponential PH (Holford, 1980; Laird & Olivier, 1981)	$\log \lambda_j(Z) = \alpha_j + w^\top Z$	$M = 1$	Time-invariant effect structure with interval-specific baseline log-hazards.
CoxPH and DeepSurv (Cox, 1972; Katzman et al., 2018)	$\lambda(t \mid Z) = \lambda_0(t) e^{w^\top Z}$	$M = 1$	Effect structure only. Baseline and partial likelihood remain method specific.
Grouped-time PH (Cox, 1972)	$\log\{-\log(1 - h_j)\} = \alpha_j + w^\top Z$	$M = 1$	Time-invariant effect structure under the complementary log-log link.
Proportional odds (Royston & Parmar, 2002)	$\operatorname{logit} \Pr(T \le \tau_j \mid Z) = \alpha_j + w^\top Z$	$M = 1$	Time-invariant effect structure under the cumulative-logit link and ordered thresholds.
Nnet-surv (Gensheimer & Narasimhan, 2019)	$h_j(Z) = \operatorname{sigmoid}\{\xi_j(Z)\}$	$M = K$	Unrestricted interval-hazard-logit effect structure.
PC-Hazard (Kvamme & Borgan, 2021)	$\lambda_j(Z) = \operatorname{softplus}\{\xi_j(Z)\}/\lvert I_j \rvert$	$M = K$	Unrestricted interval-rate-predictor effect structure.
PMF and single-event DeepHit (Kvamme & Borgan, 2021; Lee et al., 2018)	$\boldsymbol{p}(Z) = \operatorname{softmax}\{\boldsymbol{\xi}(Z)\}$	$M = K$	Unrestricted PMF-logit effect structure. Ranking and competing-risk terms are excluded.
MTLR and N-MTLR (Yu et al., 2011; Fotso, 2018)	Reverse cumulative interval scores followed by softmax	$M = K$	Unrestricted effect structure up to invertible cumulative reparameterisation. Penalties remain method specific.
BCESurv-style heads (Kvamme & Borgan, 2021)	$S_j(Z) = \operatorname{sigmoid}\{\xi_j(Z)\}$	$M = K$	Unrestricted time-specific survival-logit effect structure. Monotonicity is separate.

C Theoretical Setup and Regularity Conditions

This section states the common notation and regularity conditions. Appendix D contains the population geometry, finite-sample, local-information, minimax, adaptation, effective-dimension, forced-coupling, and placement proofs. The remaining appendices specify the controlled and real-data protocols, cross-head reconstruction, and nested selection. All numerical statements are taken from the audited result archive; no experiment is recomputed in preparing this manuscript.

The horizon is normalised to $[0, 1]$. We assume $T \perp C \mid Z$ and condition on a fixed representation $Z \in \mathbb{R}^d$. Throughout the theory, $\mathbb{E} = \mathbb{E}_0$ denotes expectation under the true observed-data law induced by $\lambda_0 = e^{\eta_0}$ and the censoring mechanism. The truth has representation-level form

$$\eta_0(t, Z) = a_0(t) + b_0(t)^\top Z. \tag{14}$$

If this form is misspecified, a representation/structural approximation term precedes the temporal decomposition. For the finite-sample results, let

$$\mathcal{F}^H_{K,M} = \{\eta \in \mathcal{F}_{K,M} : \|\eta\|_\infty \le H\}$$

be the bounded sieve. The empirical minimiser is taken over this set; we suppress the superscript H in the theorem statements. The population projections lie in its relative interior, and the explicit approximants used below belong to the bounded sieve. The affine projection identities therefore continue to hold for every fitted candidate. The following constants are independent of n, K, M, d:

1. The true log hazard and every candidate in the finite-sample sieve satisfy $|\eta(t, Z)| \le H$ on the support of (t, Z).

2. Observable event information has common support, $0 < \underline{\mu} \le \mu_0(t, Z) = S_T(t \mid Z)G_C(t \mid Z)e^{\eta_0(t,Z)} \le \overline{\mu} < \infty$.
3. The representation is centred and $\mathbb{E}(ZZ^\top) \preceq \sigma_Z^2 I_d$. No dimension-independent bound on $\|Z\|_2$ is imposed.
4. The baseline and effects belong to spectral temporal Sobolev balls of orders $r \in (0, 1]$ and $s \ge r$. Writing $a_0(t) = \sum_{m\ge0} \beta_m \phi_m(t)$ and $b_0(t) = \sum_{m\ge0} \theta_m \phi_m(t)$ in the continuous cosine basis,
$$\sum_{m\ge1} (\pi m)^{2r} \beta_m^2 \le R_a^2, \qquad \sum_{m\ge1} (\pi m)^{2s} \|\theta_m\|_2^2 \le R_b^2. \tag{15}$$
5. With $\|h\|_{\mathcal{I}}^2 = \mathbb{E}\int_0^1 \mu_0 h^2 dt$, the event-information geometry is uniformly conditioned on represented directions: for $h(t, Z) = a(t) + b(t)^\top Z$ in any sieve difference space,
$$c_{\mathcal{I}}\{\|a\|_2^2 + \|b\|_{L_2}^2\} \le \|h\|_{\mathcal{I}}^2 \le C_{\mathcal{I}}\{\|a\|_2^2 + \|b\|_{L_2}^2\}. \tag{16}$$
The cell-average/cosine approximant constructed below lies in the fixed predictor envelope.
6. The population minimisers over $\mathcal{F}_{K,M}$ and $\mathcal{F}_{K,K}$ exist and are interior relative to their affine log-hazard spans.

These conditions are uniform along any sequence (n, d, K, M) under consideration. In particular, $d = d_n$ may grow: the finite-sample requirement is that $(K + Md)/n$ tends to zero while the displayed constants remain fixed. The feasibility condition avoids deriving a uniform envelope from an L_2 Sobolev bound and a covariance bound. Relative interiority makes first-order optimality orthogonal to every direction in the corresponding affine span, which yields the exact Pythagorean identities in Theorem 2.

These assumptions play distinct roles. Conditionally independent censoring identifies the event likelihood at the representation level. The predictor envelope supplies uniform curvature and bounded likelihood increments. Common support requires positive at-risk probability through the evaluation horizon, while (16) prevents represented directions from degenerating as their dimension changes. Sobolev regularity controls temporal approximation, while the covariance and feasibility conditions translate coefficient approximation into prediction approximation. The restriction $s \ge r$ is the anisotropic regime of primary interest: effects are at least as smooth as the grid-resolved component. It also ensures the effect approximant has uniformly bounded H^r norm when it is cell averaged.

If the representation-level varying-coefficient form is misspecified, let η^*_{rep} be its population projection. The complete excess risk then begins with $D_{F_0}(\eta^*_{\mathrm{rep}}, \eta_0)$, after which every decomposition below applies relative to η^*_{rep}. The theory does not claim that temporal-head selection repairs representation error or informative censoring after compression.

D PROOFS

D.1 PROOF OF THEOREM 2

Proof.

Setup. Normalize the prediction horizon to $[0, 1]$. We observe
$$Y = \min(T, C, 1), \qquad \Delta = \mathbf{1}\{T \le C,\ T \le 1\},$$
and assume $T \perp C \mid Z$. Write the true and candidate log hazards as
$$\eta_0(t, Z) = \log \lambda_0(t \mid Z), \qquad \eta(t, Z) = \log \lambda(t \mid Z),$$
and define
$$W_0(t, Z) = S_T(t \mid Z)G_C(t \mid Z), \qquad \mu_0(t, Z) = W_0(t, Z)e^{\eta_0(t,Z)}.$$
The population continuous RC-NLL is
$$\mathcal{R}(\eta) = \mathbb{E}_0\left[\int_0^Y e^{\eta(t,Z)} dt - \Delta\eta(Y, Z)\right]. \tag{17}$$
We rewrite its accumulated-hazard and observed-event terms separately.

Step 1: Rewrite the accumulated-hazard term. Using an at-risk indicator,

$$\int_0^Y e^{\eta(t,Z)}dt = \int_0^1 \mathbf{1}\{Y \geq t\}e^{\eta(t,Z)}dt.$$

Taking expectations and applying Fubini's theorem gives

$$\mathbb{E}\int_0^Y e^{\eta(t,Z)}dt = \mathbb{E}\int_0^1 \mathbb{E}[\mathbf{1}\{Y \geq t\} \mid Z]\, e^{\eta(t,Z)}dt. \tag{18}$$

The event $Y \geq t$ means that both $T \geq t$ and $C \geq t$. Conditionally independent censoring therefore gives

$$\Pr(Y \geq t \mid Z) = S_T(t \mid Z)G_C(t \mid Z) = W_0(t, Z).$$

Consequently,

$$\mathbb{E}\int_0^Y e^{\eta(t,Z)}dt = \mathbb{E}\int_0^1 W_0(t, Z)e^{\eta(t,Z)}dt. \tag{19}$$

Thus W_0 records the probability that the event hazard remains observable at time t.

Step 2: Rewrite the observed-event term. For any integrable function f, the expectation $\mathbb{E}\{\Delta f(Y, Z)\}$ integrates f against the density of an observed, uncensored event. Conditional on Z, this density is

$$f_T(t \mid Z)G_C(t \mid Z).$$

Because

$$f_T(t \mid Z) = S_T(t \mid Z)\lambda_0(t \mid Z) = S_T(t \mid Z)e^{\eta_0(t,Z)},$$

we have

$$f_T(t \mid Z)G_C(t \mid Z) = S_T(t \mid Z)G_C(t \mid Z)e^{\eta_0(t,Z)} = \mu_0(t, Z).$$

Hence

$$\mathbb{E}\{\Delta f(Y, Z)\} = \mathbb{E}\int_0^1 \mu_0(t, Z)f(t, Z)dt. \tag{20}$$

Taking $f = \eta$ and combining (19)–(20) yields

$$\mathcal{R}(\eta) = \mathbb{E}\int_0^1 W_0 e^{\eta}dt - \mathbb{E}\int_0^1 \mu_0\eta\, dt. \tag{21}$$

Step 3: Identify excess risk as a Bregman divergence. Define the convex functional

$$F_0(\eta) = \mathbb{E}\int_0^1 W_0(t, Z)e^{\eta(t,Z)}dt.$$

For brevity in the main text, write $D_0 = D_{F_0}$ for its Bregman divergence. Its directional derivative at η_0 in direction h is

$$\langle \nabla F_0(\eta_0), h\rangle = \mathbb{E}\int_0^1 W_0 e^{\eta_0} h\, dt = \mathbb{E}\int_0^1 \mu_0 h\, dt.$$

Subtracting (21) at η_0 from the same expression at η therefore gives

$$\begin{aligned}\mathcal{R}(\eta) - \mathcal{R}(\eta_0) &= F_0(\eta) - F_0(\eta_0) - \langle \nabla F_0(\eta_0), \eta - \eta_0\rangle \\ &= D_{F_0}(\eta, \eta_0).\end{aligned} \tag{22}$$

Writing $u = \eta - \eta_0$ and using $W_0 e^{\eta_0} = \mu_0$ gives the explicit form

$$D_{F_0}(\eta, \eta_0) = \mathbb{E}\int_0^1 \mu_0(t, Z)\{e^{u(t,Z)} - 1 - u(t, Z)\}dt. \tag{23}$$

Equivalently, returning to the hazard scale,

$$\mathcal{R}(\lambda) - \mathcal{R}(\lambda_0) = \mathbb{E}_Z \int_0^1 S_T(t \mid Z)G_C(t \mid Z)\,\{\lambda - \lambda_0 - \lambda_0 \log(\lambda/\lambda_0)\}\, dt.$$

Since $e^u - 1 - u \geq 0$, with equality only at $u = 0$, excess RC-NLL is nonnegative and vanishes only when the candidate equals the truth on directions with positive observable event information.

Step 4: Establish the affine projection identity. Let $\mathcal{S}$ be an affine log-hazard space with interior population minimiser

$$\eta^*_{\mathcal{S}} = \arg\min_{\eta \in \mathcal{S}} \mathcal{R}(\eta).$$

For completeness, the three-point identity follows directly from the definition. For any differentiable F and points x, y, z,

$$\begin{aligned} &D_F(x,z) - D_F(x,y) - D_F(y,z) \\ &= -\langle \nabla F(z), x - z\rangle + \langle \nabla F(y), x - y\rangle + \langle \nabla F(z), y - z\rangle \\ &= \langle \nabla F(y) - \nabla F(z), x - y\rangle. \end{aligned} \tag{24}$$

Taking $(x, y, z) = (\eta, \eta^*_{\mathcal{S}}, \eta_0)$ and rearranging gives, for every $\eta \in \mathcal{S}$,

$$\begin{aligned} D_{F_0}(\eta, \eta_0) = D_{F_0}(\eta, \eta^*_{\mathcal{S}}) + D_{F_0}(\eta^*_{\mathcal{S}}, \eta_0) \\ + \langle \nabla F_0(\eta^*_{\mathcal{S}}) - \nabla F_0(\eta_0), \eta - \eta^*_{\mathcal{S}}\rangle . \end{aligned} \tag{25}$$

Differentiating the risk representation in (21) with respect to the candidate η gives

$$\nabla \mathcal{R}(\eta) = \nabla F_0(\eta) - \nabla F_0(\eta_0).$$

For any direction v parallel to $\mathcal{S}$, define the one-dimensional restriction

$$g_v(\epsilon) = \mathcal{R}(\eta^*_{\mathcal{S}} + \epsilon v).$$

Because $\mathcal{S}$ is affine and the minimiser is interior, both signs of sufficiently small ϵ are feasible. First-order optimality therefore gives

$$\begin{aligned} 0 = g'_v(0) &= \langle \nabla \mathcal{R}(\eta^*_{\mathcal{S}}), v\rangle \\ &= \langle \nabla F_0(\eta^*_{\mathcal{S}}) - \nabla F_0(\eta_0), v\rangle . \end{aligned} \tag{26}$$

Since both η and $\eta^*_{\mathcal{S}}$ belong to $\mathcal{S}$, $v = \eta - \eta^*_{\mathcal{S}}$ is such a direction. The final term in (25) is therefore exactly zero, giving

$$D_{F_0}(\eta, \eta_0) = D_{F_0}(\eta, \eta^*_{\mathcal{S}}) + D_{F_0}(\eta^*_{\mathcal{S}}, \eta_0). \tag{27}$$

This is the Bregman analogue of the Pythagorean identity for an orthogonal Euclidean projection.

Step 5: Apply the identity to the $K \times M$ family. Let $\mathcal{F}_{K,M} \subseteq \mathcal{F}_{K,K}$ denote the rectangular and unrestricted-grid log-hazard spaces, with population projections $\eta^*_{K,M}$ and η^*_K. Applying (27) to $\widehat{\eta}_{K,M} \in \mathcal{F}_{K,M}$ gives

$$D_{F_0}(\widehat{\eta}_{K,M}, \eta_0) = D_{F_0}(\widehat{\eta}_{K,M}, \eta^*_{K,M}) + D_{F_0}(\eta^*_{K,M}, \eta_0).$$

Because $\eta^*_{K,M} \in \mathcal{F}_{K,K}$, applying the identity through the larger space gives

$$D_{F_0}(\eta^*_{K,M}, \eta_0) = D_{F_0}(\eta^*_{K,M}, \eta^*_K) + D_{F_0}(\eta^*_K, \eta_0).$$

Combining the two equalities proves (5). Every component is a nonnegative divergence. When $M = K$, the two spaces coincide, so $\eta^*_{K,M} = \eta^*_K$ and the effect-approximation term is exactly zero. □

Remark (quadratic equivalence). The exact divergence also connects to the weighted quadratic geometry used below. For $|\eta - \nu| \le B$, Taylor's theorem applied to $e^u - 1 - u$ gives

$$\frac{e^{-B}}{2}\|\eta - \nu\|^2_\nu \le D_{F_0}(\eta, \nu) \le \frac{e^B}{2}\|\eta - \nu\|^2_\nu, \qquad \|h\|^2_\nu = \mathbb{E}\int_0^1 W_0 e^\nu h^2 dt. \tag{28}$$

Thus all three components of (5) can be analysed through the same observable-information-weighted prediction norm.

Remark (misspecification and scope). The result does more than telescope three risk differences: it shows that all components are projections under the same censored-risk divergence and that no cross term remains. If the varying-coefficient representation is misspecified, let η^*_{rep} be its population projection. The complete decomposition begins with the additional structural term $D_{F_0}(\eta^*_{\mathrm{rep}}, \eta_0)$, after which the same temporal decomposition applies relative to η^*_{rep}. If censoring is not conditionally independent given Z, Steps 1 and 2 no longer identify the displayed risk geometry.

D.2 Localized concentration for the censored likelihood

For one observation $O = (Y, \Delta, Z)$, write

$$\ell_\eta(O) = \int_0^Y e^{\eta(t,Z)} dt - \Delta\eta(Y, Z).$$

The following lemma supplies the finite-sample step used throughout the appendix.

Lemma 8 (Localized censored-likelihood bound)**.** *Let $\mathcal{A}$ be an affine predictor space of dimension at most p, let $\mathcal{A}^H = \{\eta \in \mathcal{A} : \|\eta\|_\infty \le H\}$, and let η^* be an interior population-risk minimiser over $\mathcal{A}$. Under the uniform envelope, common-support, and information conditions in Appendix C, there is a constant C depending only on those fixed conditions such that, whenever $(p + \log(1/\delta))/n$ is sufficiently small, with probability at least $1 - \delta$,*

$$|(P_n - P)(\ell_\eta - \ell_{\eta^*})| \le \frac{1}{8} P(\ell_\eta - \ell_{\eta^*}) + C\frac{p + \log(1/\delta)}{n} \tag{29}$$

simultaneously for all $\eta \in \mathcal{A}^H$.

Proof. Put $h = \eta - \eta^*$ and $g_\eta = \ell_\eta - \ell_{\eta^*}$.

Step 1: Curvature. Interiority and the affine projection identity give

$$Pg_\eta = D_{F_0}(\eta, \eta^*) = \mathbb{E}\int_0^1 W_0 e^{\eta^*}\{e^h - 1 - h\} dt. \tag{30}$$

For $|\eta|, |\eta^*| \le H$, Taylor's theorem and the boundedness of η_0 imply

$$c_H \|h\|_{\mathcal{I}}^2 \le Pg_\eta \le C_H \|h\|_{\mathcal{I}}^2, \qquad \|h\|_{\mathcal{I}}^2 = \mathbb{E}\int_0^1 \mu_0 h^2 dt. \tag{31}$$

All constants are dimension free because the comparison is between bounded predictors rather than coefficient vectors.

Step 2: Bernstein variance and the loss metric. The loss difference can be written as

$$g_\eta(O) = \int_0^Y (e^\eta - e^{\eta^*}) dt - \Delta h(Y, Z).$$

The predictor envelope gives $|g_\eta| \le 2e^H + 2H$. Moreover,

$$\mathbb{E}\{\Delta h(Y, Z)^2\} = \mathbb{E}\int_0^1 \mu_0 h^2 dt = \|h\|_{\mathcal{I}}^2, \tag{32}$$

$$\mathbb{E}\left\{\int_0^Y (e^\eta - e^{\eta^*}) dt\right\}^2 \le e^{2H}\mathbb{E}\int_0^1 \mathbf{1}\{Y \ge t\} h(t, Z)^2 dt \le C_H \|h\|_{\mathcal{I}}^2. \tag{33}$$

The first equality uses (20); the second display uses Cauchy–Schwarz on the unit horizon, followed by (18) and $\mu_0 = W_0 e^{\eta_0}$. Consequently,

$$Pg_\eta^2 \le C_H \|h\|_{\mathcal{I}}^2 \le C'_H Pg_\eta. \tag{34}$$

The same calculation with two predictors η, ν gives the Lipschitz metric comparison

$$\|g_\eta - g_\nu\|_{L_2(P)} \le C_H \|\eta - \nu\|_{\mathcal{I}}. \tag{35}$$

Step 3: Local entropy. For $u > 0$, let

$$\mathcal{G}(u) = \{g_\eta : \eta \in \mathcal{A}^H,\ Pg_\eta \le u\}.$$

By (31), its predictor differences lie in an information-norm ball of radius $C\sqrt{u}$ inside a linear space of dimension at most p. A volumetric covering of that ball, followed by (35), yields

$$\log N\{\epsilon, \mathcal{G}(u), L_2(P)\} \le p \log\left(1 + \frac{C\sqrt{u}}{\epsilon}\right). \tag{36}$$

In particular, its entropy integral satisfies

$$\int_0^{C\sqrt{u}} \sqrt{\log N\{\epsilon, \mathcal{G}(u), L_2(P)\}}\, d\epsilon \le C\sqrt{pu}. \tag{37}$$

Step 4: Local maximal inequality and peeling. Symmetrisation and chaining using (37), followed by Bernstein concentration using (34) and the fixed envelope, give, for every $x > 0$, with probability at least $1 - 2e^{-x}$,

$$\sup_{g \in \mathcal{G}(u)} |(P_n - P)g| \le C \left\{ \sqrt{\frac{u(p+x)}{n}} + \frac{p+x}{n} \right\}. \tag{38}$$

This is the standard finite-dimensional specialization of localized counting-process and likelihood-sieve inequalities (Van de Geer, 1995; Wong & Shen, 1995).

For completeness, set $u_0 = A\{p + x\}/n$ for a sufficiently large fixed A and divide the range $Pg_\eta > u_0$ into shells $2^j u_0 < Pg_\eta \le 2^{j+1} u_0$. Apply (38) on shell j with $x_j = x + j + 1$. Since $(p + x_j)/2^j(p + x)$ is uniformly bounded, increasing A makes the right-hand side at most $Pg_\eta/8$ on every shell. On $Pg_\eta \le u_0$, the same inequality is $O(u_0)$. Summing the shell probabilities and replacing x by $\log(C/\delta)$ gives (29), after changing the universal constant. □

D.3 Proof of Theorem 3

Proof.

Setup and roadmap. Let P_K denote averaging over the K equal-width cells. We construct an explicit member $\widetilde{\eta}_{K,M} \in \mathcal{F}_{K,M}$, bound its approximation error, and then control the distance between the empirical minimiser and the population projection. The proof has three corresponding ingredients: cell approximation, cosine truncation, and finite-dimensional likelihood concentration.

Step 1: Control approximation by the grid. For every $f \in H^r([0,1])$ and $r \in (0,1]$, the fractional Poincaré inequality gives

$$\|f - P_K f\|_2^2 \le C_r K^{-2r} |f|_{H^r}^2. \tag{39}$$

When $r = 1$, this is the usual cellwise Poincaré inequality. Fractional interpolation gives intermediate r. The order cannot improve beyond K^{-2} for a piecewise-constant reconstruction, even if the underlying function is smoother.

Step 2: Match the continuous cosine approximation to the implemented DCT space. For the continuous cosine mode $\phi_m(t) = \sqrt{2}\cos(\pi m t)$ and midpoint $t_{Kj} = (j - 1/2)/K$,

$$K \int_{(j-1)/K}^{j/K} \phi_m(t)dt = \phi_m(t_{Kj}) \operatorname{sinc}\Big(\frac{\pi m}{2K}\Big). \tag{40}$$

To verify (40), direct integration and the identity $\sin a - \sin b = 2\cos\{(a+b)/2\}\sin\{(a-b)/2\}$ give

$$\begin{aligned} K \int_{(j-1)/K}^{j/K} \phi_m(t)dt &= \frac{\sqrt{2}K}{\pi m} \left[\sin\left(\frac{\pi m j}{K}\right) - \sin\left(\frac{\pi m (j-1)}{K}\right) \right] \\ &= \phi_m(t_{Kj}) \frac{\sin\{\pi m/(2K)\}}{\pi m/(2K)}. \end{aligned} \tag{41}$$

The sinc factor is nonzero for $m < K$. Therefore, for $M \le K$, cell averages of the first M continuous cosine modes span exactly the first M midpoint-DCT directions. The mode-specific sinc factors change only coefficient scaling and can be absorbed into Γ. We may therefore use the continuous cosine expansion for the approximation analysis while retaining exact membership in the implemented $K \times M$ model family.

Step 3: Construct and bound a two-stage rectangular approximant. We mirror the nested path from the continuous truth to the unrestricted K-grid model and then to the M-mode effect model:

$$\eta_0 \longrightarrow \widetilde{\eta}_K \longrightarrow \widetilde{\eta}_{K,M}.$$

First define the unrestricted-grid approximant

$$\widetilde{\eta}_K(t, Z) = (P_K a_0)(t) + (P_K b_0)(t)^\top Z. \tag{42}$$

Because the full DCT spans all K cellwise-constant directions, $\widetilde{\eta}_K \in \mathcal{F}_{K,K}$. This transition changes only temporal resolution. The covariance bound on Z and the Poincaré inequality give

$$\begin{aligned}\mathbb{E}\int_0^1 \{\widetilde{\eta}_K - \eta_0\}^2 dt &\lesssim \|P_K a_0 - a_0\|_2^2 + \|P_K b_0 - b_0\|_{L_2}^2 \\ &\lesssim K^{-2r},\end{aligned} \tag{43}$$

where $s \geq r$ ensures that b_0 also has the regularity required for the K^{-2r} grid rate.

Next truncate the true effect function to its first M cosine modes,

$$b_M(t) = \sum_{m<M} \theta_m \phi_m(t),$$

and define

$$\widetilde{\eta}_{K,M}(t, Z) = (P_K a_0)(t) + (P_K b_M)(t)^\top Z. \tag{44}$$

By Step 2, $\widetilde{\eta}_{K,M} \in \mathcal{F}_{K,M}$. This transition changes only effect flexibility:

$$\widetilde{\eta}_{K,M} - \widetilde{\eta}_K = \{P_K(b_M - b_0)\}^\top Z.$$

The covariance bound on Z and the fact that cell averaging is an L_2 contraction give

$$\mathbb{E}\int_0^1 \{\widetilde{\eta}_{K,M} - \widetilde{\eta}_K\}^2 dt \lesssim \|P_K(b_M - b_0)\|_{L_2}^2 \leq \|b_M - b_0\|_{L_2}^2 .$$

The contraction follows directly from Jensen's inequality on each cell. Moreover,

$$b_0(t) = \sum_{m\geq 0} \theta_m \phi_m(t), \qquad b_0(t) - b_M(t) = \sum_{m\geq M} \theta_m \phi_m(t).$$

Orthonormality and the Sobolev ellipsoid in (15) give

$$\begin{aligned}\|b_0 - b_M\|_{L_2}^2 &= \sum_{m\geq M} \|\theta_m\|_2^2 \\ &= \sum_{m\geq M} (\pi m)^{-2s} (\pi m)^{2s} \|\theta_m\|_2^2 \\ &\leq (\pi M)^{-2s} \sum_{m\geq M} (\pi m)^{2s} \|\theta_m\|_2^2 \leq R_b^2 (\pi M)^{-2s} = O(M^{-2s}).\end{aligned} \tag{45}$$

Combining the two transitions with $(x+y)^2 \leq 2x^2 + 2y^2$ gives

$$\mathbb{E}\int_0^1 \{\widetilde{\eta}_{K,M} - \eta_0\}^2 dt \lesssim K^{-2r} + M^{-2s}.$$

Quadratic equivalence in (28) therefore yields

$$D_{F_0}(\widetilde{\eta}_{K,M}, \eta_0) \leq C\{K^{-2r} + M^{-2s}\}. \tag{46}$$

Finally, $\eta^*_{K,M}$ minimizes population risk over $\mathcal{F}_{K,M}$. The projection identity in Theorem 2 therefore gives

$$\begin{aligned}A_{\text{grid}}(K) + A_{\text{effect}}(K, M) &= D_{F_0}(\eta^*_{K,M}, \eta_0) \\ &\leq D_{F_0}(\widetilde{\eta}_{K,M}, \eta_0) \lesssim K^{-2r} + M^{-2s}.\end{aligned} \tag{47}$$

This makes explicit that the construction bounds the two approximation components jointly while assigning their rates to separate transitions.

Step 4: Derive the sharp grid constant. For differentiable truths, write $x = (1, Z^\top)^\top$, $\vartheta_0(t) = (a_0(t), b_0(t)^\top)^\top$, and

$$H(t) = \mathbb{E}\{\mu_0(t, Z)xx^\top\}.$$

On a cell I of width $h = K^{-1}$ centred at t_I, let c_I minimise the local quadratic loss. Its normal equation is

$$\left\{\int_I H(t)dt\right\} c_I = \int_I H(t)\vartheta_0(t)dt.$$

Continuity and symmetry of the cell imply $c_I = \vartheta_0(t_I) + o(h)$. Using $\vartheta_0(t) = \vartheta_0(t_I) + (t - t_I)\vartheta_0'(t_I) + o(h)$ and $H(t) = H(t_I) + o(1)$ gives

$$\frac{1}{2}\int_I \{c_I - \vartheta_0(t)\}^\top H(t)\{c_I - \vartheta_0(t)\}dt = \frac{h^3}{24}\vartheta_0'(t_I)^\top H(t_I)\vartheta_0'(t_I) + o(h^3).$$

The factor $1/24$ is the product of the likelihood-curvature factor $1/2$ and the centred cell moment $\int_{-h/2}^{h/2} u^2 du = h^3/12$. Summing over the K cells yields

$$A_{\text{grid}}(K) = \frac{1}{24K^2}\int_0^1 \vartheta_0'(t)^\top H(t)\vartheta_0'(t)dt + o(K^{-2}). \tag{48}$$

Step 5: Control estimation in the rectangular sieve. Fix $\delta \in (0, 1)$. This step controls the final transition

$$\eta^*_{K,M} \longrightarrow \widehat{\eta}_{K,M}.$$

Let $p = K + Md$, write $\eta^* = \eta^*_{K,M}$, and define $g_\eta = \ell_\eta - \ell_{\eta^*}$. Here $Pf = \mathbb{E}f(O)$ denotes population expectation and $P_n f = n^{-1}\sum_{i=1}^n f(O_i)$ the training average. Lemma 8, applied to the p-dimensional rectangular predictor space, gives

$$|(P_n - P)g_\eta| \le \frac{1}{8}Pg_\eta + C\frac{p + \log(1/\delta)}{n} \tag{49}$$

simultaneously over the bounded rectangular sieve. Its constant is independent of d, K, M because the lemma uses the common predictor envelope and information norm directly.

Finally, empirical optimality gives

$$P_n g_{\widehat{\eta}_{K,M}} = P_n \ell_{\widehat{\eta}_{K,M}} - P_n \ell_{\eta^*} \le 0.$$

Therefore, on the event in (49),

$$\begin{aligned} Pg_{\widehat{\eta}_{K,M}} &= (P - P_n)g_{\widehat{\eta}_{K,M}} + P_n g_{\widehat{\eta}_{K,M}} \\ &\le \frac{1}{8}Pg_{\widehat{\eta}_{K,M}} + C\frac{p + \log(1/\delta)}{n}. \end{aligned} \tag{50}$$

Rearranging and substituting $p = K + Md$ gives

$$D_{F_0}(\widehat{\eta}_{K,M}, \eta^*_{K,M}) \le C\frac{K + Md + \log(1/\delta)}{n}. \tag{51}$$

Step 6: Combine approximation and estimation. The exact decomposition in (5), the approximation bound (46), and (51) give

$$D_{F_0}(\widehat{\eta}_{K,M}, \eta_0) \le C\left\{K^{-2r} + M^{-2s} + \frac{K + Md + \log(1/\delta)}{n}\right\},$$

which proves Theorem 3. □

Remark (sharp local estimation price). In a fixed-dimensional local expansion, let J be population curvature and I score covariance. Then

$$\mathbb{E}E_{\text{est}}(K, M) = \frac{1}{2n}\operatorname{tr}(J^{-1}I) + o(n^{-1}).$$

Under correct specification, $I = J$, so the expression reduces to $(K + Md)/(2n) + o(n^{-1})$. This sharp formula explains the empirical slope of risk against baseline intervals and effect modes. The high-probability theorem is stronger in a different direction: it allows the sieve dimension to grow, but it suppresses sharp constants.

D.4 Derivation of the information-supported mode rule

Setup and roadmap. We work in the local asymptotic experiment around a reference population target. The argument first converts censored likelihood into a quadratic information problem, then compares two nested effect spaces that differ by one temporal mode, and finally profiles out the unrestricted baseline.

Step 1: Obtain the local quadratic experiment. Fix a reference log hazard η_0 and consider local alternatives $\eta_{n,u} = \eta_0 + u/\sqrt{n}$. For subject i, let $N_i(t) = \mathbf{1}\{Y_i \le t, \Delta_i = 1\}$ be the observed-event counting process and $R_i(t) = \mathbf{1}\{Y_i \ge t\}$ the at-risk process. Under the reference model,

$$dM_i(t) = dN_i(t) - R_i(t)e^{\eta_0(t,Z_i)}dt$$

is the event martingale increment. The log-likelihood ratio between $\eta_{n,u}$ and η_0 has the standard counting-process LAN expansion (Andersen & Gill, 1982):

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 u(t,Z_i)dM_i(t) - \frac{1}{2}\|u\|_{\mathcal{I}}^2 + o_P(1), \qquad \|u\|_{\mathcal{I}}^2 = \mathbb{E}\int_0^1 \mu_0 u^2 dt. \tag{52}$$

In information-orthonormal coordinates, the Gaussian limit experiment observes $u + W$, where W is standard white noise. Let $\Pi_{K,M}$ be orthogonal projection onto the sieve in the information norm and put $p = K + Md$. The projected estimator of the root-n local coordinate is

$$\widehat{u}_{K,M} = \Pi_{K,M}u + \Pi_{K,M}W.$$

Returning to the original log-hazard scale, define $h_n = u/\sqrt{n}$ and $\widehat{h}_{K,M} = \widehat{u}_{K,M}/\sqrt{n}$. Then

$$\widehat{h}_{K,M} - h_n = -(I - \Pi_{K,M})h_n + n^{-1/2}\Pi_{K,M}W.$$

The two terms lie in orthogonal subspaces, so their cross product is zero. Moreover, $\mathbb{E}\,\|\Pi_{K,M}W\|_{\mathcal{I}}^2 = \operatorname{rank}(\Pi_{K,M}) = p$. Hence the local quadratic risk is

$$\mathbb{E}_{n,u}\frac{1}{2}\left\|\widehat{h}_{K,M} - h_n\right\|_{\mathcal{I}}^2 = \frac{1}{2}\|(I - \Pi_{K,M})h_n\|_{\mathcal{I}}^2 + \frac{K + Md}{2n} + o(n^{-1}). \tag{53}$$

The two leading terms are exact in the Gaussian limit experiment; the remainder records the local quadratic approximation to the original censored likelihood. The first leading term is signal left outside the fitted space, and the second is estimation variance inside it.

Step 2: Add one temporal effect mode. Because the two sieve spaces are nested, write $\Pi_{m+1} = \Pi_m + Q_m$, where Q_m projects onto the new d-dimensional effect block. Define

$$E_{m,n} = \|Q_m h_n\|_{\mathcal{I}}^2.$$

Orthogonal Pythagoras gives $\|(I - \Pi_{m+1})h_n\|_{\mathcal{I}}^2 = \|(I - \Pi_m)h_n\|_{\mathcal{I}}^2 - E_{m,n}$, while the projection rank increases by d. Thus the larger model reduces squared bias by $E_{m,n}$ and increases variance by d/n. Taking the one-half factor in (53) into account,

$$\mathcal{R}_{m+1} - \mathcal{R}_m = -\frac{E_{m,n}}{2} + \frac{d}{2n} + o(n^{-1}).$$

Thus, to first order, mode m lowers local risk when

$$E_{m,n} > \frac{d}{n}.$$

Dropping the triangular-array subscript from the original-scale information energy gives (7). Equivalently, in the root-n coordinate the condition is $\|Q_m u\|_{\mathcal{I}}^2 > d$. This explicit change of scale is essential: the LAN coordinate has order-one noise, while the original log-hazard coordinate has order-$n^{-1/2}$ noise.

Step 3: Profile the unrestricted baseline. The baseline and covariate effects are not generally information orthogonal. Partition the event-information matrix as

$$H(t) = \mathbb{E}\left\{\mu_0(t,Z)\begin{pmatrix}1\\ Z\end{pmatrix}\begin{pmatrix}1 & Z^\top\end{pmatrix}\right\} = \begin{pmatrix}h_{00} & h_{0Z}^\top\\ h_{0Z} & H_{ZZ}\end{pmatrix}. \tag{54}$$

For a baseline perturbation a and effect perturbation b, the local information form is

$$\begin{pmatrix}a\\ b\end{pmatrix}^\top H \begin{pmatrix}a\\ b\end{pmatrix} = h_{00}\left(a + \frac{h_{0Z}^\top b}{h_{00}}\right)^2 + b^\top\left(H_{ZZ} - \frac{h_{0Z}h_{0Z}^\top}{h_{00}}\right)b. \tag{55}$$

Minimizing over the unrestricted baseline perturbation a removes the first square. The remaining quadratic form is therefore the information for an effect perturbation after profiling out the baseline, namely

$$V_\mu(t) = H_{ZZ}(t) - \frac{h_{0Z}(t)h_{0Z}(t)^\top}{h_{00}(t)}. \tag{56}$$

The Schur complement removes temporal variation that can be absorbed by the baseline. Therefore, E_m measures only covariate-dependent variation that remains identifiable after baseline profiling.

Remark (role of censoring and representation width). Censoring changes E_m through the factor G_C in μ_0, even though the censoring tangent space is orthogonal to the event tangent space under conditional independence. The factor d is equally structural: each new temporal direction introduces one coefficient per representation coordinate. If profiling or regularisation gives the new block effective rank smaller than d, its variance price becomes that effective-rank increment divided by $2n$.

D.5 Derivation of effective dimension beyond hard truncation

Setup and roadmap. Let L be a positive temporal roughness operator on a Hilbert space and let $T = L^{-1/2}$ have singular values $\tau_j(T) \asymp j^{-s}$. In the whitened Gaussian experiment, observe

$$X_l = \theta_l + n^{-1/2}W_l, \qquad l = 1,\dots,d,$$

under the roughness constraint

$$\sum_{l=1}^{d}\left\|L^{1/2}\theta_l\right\|^2 \le R_b^2.$$

For a common linear smoother S, define

$$q(S) = \|S\|_{\mathrm{HS}}^2, \qquad \beta_L(S) = \left\|(I-S)L^{-1/2}\right\|_{\mathrm{op}}^2.$$

We derive its exact bias-variance risk, lower-bound the bias attainable at a given $q(S)$, and then balance the two terms.

Step 1: Derive the exact smoother risk. For each coordinate,

$$SX_l - \theta_l = -(I-S)\theta_l + n^{-1/2}SW_l.$$

The noise is centred and independent of the signal, so

$$\mathbb{E}\left\|SX_l - \theta_l\right\|^2 = \left\|(I-S)\theta_l\right\|^2 + \frac{1}{n}\|S\|_{\mathrm{HS}}^2.$$

Summing over l gives variance $dq(S)/n$. To evaluate the worst-case bias, write $\theta_l = Tu_l$, so the roughness constraint becomes $\sum_l \|u_l\|^2 \le R_b^2$. Then

$$\sum_{l=1}^{d}\left\|(I-S)\theta_l\right\|^2 = \sum_{l=1}^{d}\left\|(I-S)Tu_l\right\|^2 \le \beta_L(S)\sum_{l=1}^{d}\|u_l\|^2 \le R_b^2\beta_L(S).$$

Equality is approached by placing the available roughness energy in a right-singular direction that nearly maximizes $\|(I-S)T\|_{\mathrm{op}}$. Hence

$$\sup_\theta \frac{1}{2}\sum_{l=1}^{d}\mathbb{E}\left\|SX_l - \theta_l\right\|^2 = \frac{1}{2}\left\{R_b^2\beta_L(S) + \frac{d}{n}q(S)\right\}. \tag{57}$$

Step 2: Lower-bound bias at fixed effective dimension. Put $q = \|S\|_{\mathrm{HS}}^2$ and $N = \lfloor 4q \rfloor + 2$. Suppose, for contradiction, that

$$\|(I - S)T\|_{\mathrm{op}} < \frac{1}{2}\tau_N(T).$$

Let $Tv_j = \tau_j(T) f_j$ be a singular system. For $j \le N$, the assumed inequality gives

$$\tau_j(T)\,\|(I - S)f_j\| = \|(I - S)Tv_j\| < \frac{1}{2}\tau_N(T) \le \frac{1}{2}\tau_j(T).$$

Thus $\|(I - S)f_j\| < 1/2$, and the triangle inequality gives $\|Sf_j\| \ge 1 - \|(I - S)f_j\| > 1/2$. Consequently,

$$q = \|S\|_{\mathrm{HS}}^2 \ge \sum_{j=1}^{N} \|Sf_j\|^2 > \frac{N}{4} > q,$$

a contradiction. Therefore,

$$\beta_L(S) = \|(I - S)T\|_{\mathrm{op}}^2 \gtrsim \tau_N(T)^2 \asymp \{q(S) + 1\}^{-2s}.$$

This width argument depends only on the roughness class, not on the cosine basis.

Step 3: Show attainability and optimise. Projection onto the first m singular directions has $q(S) = m$ and bias modulus of order m^{-2s}, so it attains the reverse bound. The optimal frontier therefore balances

$$R_b^2 q^{-2s} \quad \text{and} \quad \frac{d}{n} q.$$

Equating these terms gives $q^{2s+1} \asymp R_b^2 n/d$. Therefore

$$q_n^* \asymp \left(\frac{R_b^2 n}{d}\right)^{1/(2s+1)}$$

and effect risk of order

$$(d/n)^{2s/(2s+1)}$$

up to constants depending on R_b. Hard truncation, spectral shrinkage, and non-diagonal linear smoothers all lie on the same order frontier.

Step 4: Account for a finite output grid. If the output space contains only K temporal directions, the oracle becomes

$$q_{n,K}^* \asymp \min\{K, q_n^*\}.$$

The identity smoother has $q(S) = K$ and pays effect variance $dK/(2n)$. When $K > q_n^*$, its variance exceeds the oracle variance by order K/q_n^*. A nominally full head with tuned shrinkage can instead keep $q(S)$ near q_n^*.

Remark. The invariant is effective flexibility, not nominal rank. Hard truncation makes the axis visible because $q(S) = M$, but smooth shrinkage permits fractional effective dimension while retaining a nominally full output head.

D.6 Proof of Proposition 6

Proof.

Setup. Augment the local effect experiment with an orthogonal baseline coordinate whose K-cell approximation and estimation risks are bounded below by $c_A K^{-2r}$ and $c_B K/n$. Suppose the model path forces

$$q(S_K) \ge c_q K.$$

The effect variance in (57) is then at least $c_E dK/n$ for a constant $c_E > 0$.

Step 1: Form the coupled lower bound. Combining baseline approximation, baseline estimation, and effect estimation gives

$$R_{\text{coupled}}(K) \geq c_A K^{-2r} + c_E \frac{K(d+1)}{n}. \tag{58}$$

Step 2: Optimize over resolution. The two terms balance when

$$K^{-2r} \asymp \frac{K(d+1)}{n},$$

or equivalently

$$K \asymp \left(\frac{n}{d+1}\right)^{1/(2r+1)}.$$

Substitution into (58) yields

$$\inf_K R_{\text{coupled}}(K) \asymp \left(\frac{d+1}{n}\right)^{2r/(2r+1)}. \tag{59}$$

Step 3: Compare with the rectangular rate. Let

$$a = \frac{2r}{2r+1}, \qquad b = \frac{2s}{2s+1}.$$

Consider any sequence $d = d_n \to \infty$, $d_n = o(n)$, along which the uniform conditions of Appendix C hold. The coupled benchmark is of order $(d/n)^a$. Its ratio to the rectangular grid term n^{-a} is d^a. Its ratio to the rectangular effect term $(d/n)^b$ is $(d/n)^{a-b}$. When $s > r$, we have $a < b$, so both ratios diverge. This proves Proposition 6. □

Remark (scope). This is a canonical orthogonal local benchmark, not a uniform lower bound for every nominally full neural head. It applies when effective effect flexibility genuinely grows with K. A full head whose regularisation keeps $q(S_K)$ near q_n^* is not forced coupled.

D.7 Proof of the minimax statement in Corollary 5

Proof. The construction follows classical Sobolev minimax arguments (Stone, 1982), with separate baseline and effect hypercubes. Related approximation–estimation and adaptive model-selection arguments for covariate-dependent counting-process intensities appear in Comte et al. (2011).

Upper bound. Set

$$K \asymp n^{1/(2r+1)}, \qquad M \asymp (n/d)^{1/(2s+1)}$$

in Theorem 3. When $d = d_n = o(n)$,

$$\frac{K}{n} \asymp n^{-2r/(2r+1)} \longrightarrow 0, \qquad \frac{Md}{n} \asymp (d/n)^{2s/(2s+1)} \longrightarrow 0,$$

so the finite-sample condition holds along the sequence. Integrating the exponential tail in Lemma 8 and using bounded risk outside its local range gives the corresponding expectation bound. Balancing K^{-2r} with K/n and M^{-2s} with Md/n therefore gives

$$\sup_{\eta_0} \mathbb{E} D_{F_0}(\widehat{\eta}_{K,M}, \eta_0) \lesssim n^{-2r/(2r+1)} + (d/n)^{2s/(2s+1)}.$$

Lower-bound construction. It suffices to work under administrative censoring and a regular covariate design

$$Z = \varepsilon \sqrt{d}\, e_R,$$

where R is uniform on $\{1, \ldots, d\}$ and ε is Rademacher. Then $\mathbb{E} Z Z^\top = I_d$, while only one effect coordinate is active for each observation. Although $\|Z\|_2 = \sqrt{d}$, the design is uniformly isotropic. On the bounded log-hazard class under administrative censoring,

$$\mathbb{E} \int_0^1 \mu_\omega(t, Z)\{a(t) + b(t)^\top Z\}^2 dt \asymp \|a\|_2^2 + \|b\|_{L_2}^2$$

uniformly in d and in the hypercube vertex ω. Thus the construction satisfies the common information geometry in (16) without a dimension-independent Euclidean norm bound on Z.

Let

$$h_{m,j}(t) = \sqrt{m}\,\psi(mt - j + 1)$$

be disjoint, normalised bumps generated by a smooth function ψ supported in $(0, 1)$. For a sign vector ω, define

$$a_\omega(t) = a_* + \alpha_a \sum_{j=1}^{m_a} \omega_{0j} h_{m_a,j}(t),$$
$$b_{\omega,l}(t) = \alpha_b \sum_{j=1}^{m_b} \omega_{lj} h_{m_b,j}(t), \qquad l = 1, \dots, d, \tag{60}$$

with

$$\alpha_a^2 = \alpha_b^2 = \frac{c_0}{n}, \qquad m_a \asymp n^{1/(2r+1)}, \qquad m_b \asymp (n/d)^{1/(2s+1)}.$$

Step 1: Verify membership in the smoothness class. Disjoint support makes Sobolev energy additive. The baseline and effect energies satisfy

$$\alpha_a^2 m_a^{2r+1} \lesssim 1, \qquad d\alpha_b^2 m_b^{2s+1} \lesssim 1.$$

Choose a_* strictly inside the common log-hazard envelope. The baseline perturbation has sup-norm

$$\alpha_a \sqrt{m_a} = n^{-r/(2r+1)},$$

and the active effect perturbation has order

$$\sqrt{d}\,\alpha_b \sqrt{m_b} = \{d m_b / n\}^{1/2} = (d/n)^{s/(2s+1)} = o(1)$$

when $d = o(n)$. Thus every vertex lies in the same bounded survival class.

Step 2: Control testing difficulty. Adjacent vertices differ in one baseline or effect bit, so their log hazards differ by $2\alpha_a h_{m_a,j}$ or $2\alpha_b h_{m_b,j} Z_l$. On the common bounded class, a second-order likelihood expansion gives

$$\mathrm{KL}(P_\omega, P_{\omega'}) \le C\mathbb{E} \int_0^1 \{\eta_\omega(t, Z) - \eta_{\omega'}(t, Z)\}^2 dt.$$

Because each bump has unit L_2 norm, the right-hand side is at most $C\alpha_a^2$ for a baseline edge. For an effect edge, $\mathbb{E}Z_l^2 = 1$: equivalently, the probability $1/d$ of selecting coordinate l cancels its nonzero magnitude $Z_l^2 = d$. The one-observation divergence is therefore at most $C\alpha_b^2$. For product experiments, Kullback–Leibler divergences add, so every adjacent n-sample pair has divergence at most $Cn\alpha_a^2$ or $Cn\alpha_b^2$, both equal to Cc_0. Choosing c_0 sufficiently small keeps all adjacent tests uniformly nontrivial.

Step 3: Convert Hamming separation to risk. If two vertices differ in H_0 baseline bits and H_l bits of effect coordinate l, disjoint support and $\mathbb{E}ZZ^\top = I_d$ give

$$\mathbb{E} \int_0^1 \{\eta_\omega(t, Z) - \eta_{\omega'}(t, Z)\}^2 dt = 4\alpha_a^2 H_0 + 4\alpha_b^2 \sum_{l=1}^{d} H_l.$$

Thus an incorrectly recovered baseline bit costs order α_a^2, while an incorrectly recovered effect bit costs order α_b^2. Quadratic equivalence of Bregman and L_2 loss, the uniformly nontrivial adjacent tests from Step 2, and Assouad's lemma then give

$$\inf_{\widetilde{\eta}} \sup_{\omega} \mathbb{E}_\omega D_{F_0}(\widetilde{\eta}, \eta_\omega) \gtrsim m_a \alpha_a^2 + d m_b \alpha_b^2$$
$$\asymp n^{-2r/(2r+1)} + (d/n)^{2s/(2s+1)}. \tag{61}$$

This matches the upper bound and proves the minimax statement. Compactly supported wavelets extend the construction to noninteger smoothness. □

Remark. The lower bound contains separate baseline and effect hypercubes. This is why the minimax risk is a sum of two anisotropic rates rather than a single rate based on total parameter count.

D.8 Proof of the adaptive-selection statement in Corollary 5

Proof.

Setup. Split the observations into independent training and validation samples of sizes n_t and n_v. On the training sample, fit every bounded-sieve candidate in

$$\mathcal{I}_n = \{(K, M) : 1 \le M \le K,\ K + Md \le c_0 n_t\}. \tag{62}$$

Choose the candidate with the smallest validation continuous RC-NLL. For $i = (K, M) \in \mathcal{I}_n$, abbreviate the fitted candidate by $\widehat{\eta}_i$, set $g_i = \ell_{\widehat{\eta}_i} - \ell_{\eta_0}$, and write $L_i = Pg_i = D_{F_0}(\widehat{\eta}_i, \eta_0)$. Let P_v denote the validation empirical average. If $\widehat{i} = (\widehat{K}, \widehat{M})$ is selected, then

$$P_v g_{\widehat{i}} \le P_v g_i \qquad \text{for every } i \in \mathcal{I}_n.$$

Step 1: Condition on the training sample. Conditional on training, the fitted candidates are fixed functions. Bounded likelihood differences satisfy

$$\mathrm{Var}(\ell_\eta - \ell_{\eta_0}) \le V_H\{\mathcal{R}(\eta) - \mathcal{R}(\eta_0)\}.$$

Scalar Bernstein concentration, a union bound over $\mathcal{I}_n$, and $\sqrt{uv} \le \epsilon' u/2 + v/(2\epsilon')$ imply that, simultaneously for all $i \in \mathcal{I}_n$,

$$|(P_v - P)g_i| \le \epsilon' L_i + C_{\epsilon'} \frac{\log(2|\mathcal{I}_n|/\delta)}{n_v}. \tag{63}$$

For any fixed i, the selection inequality and (63) give

$$\begin{aligned} L_{\widehat{i}} &\le P_v g_{\widehat{i}} + \epsilon' L_{\widehat{i}} + c_v \le P_v g_i + \epsilon' L_{\widehat{i}} + c_v \\ &\le L_i + \epsilon' L_i + \epsilon' L_{\widehat{i}} + 2c_v, \qquad c_v = C_{\epsilon'} \frac{\log(2|\mathcal{I}_n|/\delta)}{n_v}. \end{aligned} \tag{64}$$

Rearranging gives $(1 - \epsilon')L_{\widehat{i}} \le (1 + \epsilon')L_i + 2c_v$. Taking the minimum over i and choosing $\epsilon' = \epsilon/(2 + \epsilon)$ yields

$$\begin{aligned} D_{F_0}(\widehat{\eta}_{\widehat{K},\widehat{M}}, \eta_0) \le (1 + \epsilon) \min_{(K,M)\in\mathcal{I}_n} D_{F_0}(\widehat{\eta}_{K,M}, \eta_0) \\ + C_\epsilon \frac{\log(2|\mathcal{I}_n|/\delta)}{n_v}. \end{aligned} \tag{65}$$

Step 2: Control the training fits. Apply Theorem 3 simultaneously to all candidates fitted on the training sample. A union bound adds $O\{\log|\mathcal{I}_n|/n_t\}$ to their estimation terms. Combining this result with (65) gives the rectangular oracle rate plus

$$O\left\{\frac{\log|\mathcal{I}_n|}{n_t} + \frac{\log|\mathcal{I}_n|}{n_v}\right\}.$$

Step 3: Evaluate the search price. The triangular candidate set has $|\mathcal{I}_n| = O(n^2)$. If $n_t \asymp n_v \asymp n$, the total selection price is $O(\log n/n)$, which is lower order than both nonparametric components. Joint validation therefore attains the minimax rate without knowing r or s; the logarithmic search price parallels classical finite-family model-selection bounds (Barron et al., 1999).

At prescribed K, the same argument gives

$$K^{-2r} + \frac{K}{n} + \inf_{1\le M\le K}\left\{M^{-2s} + \frac{Md}{n}\right\} + O\{\log(K)/n\},$$

so

$$M^*_{n,K} \asymp \min\left\{K, (n/d)^{1/(2s+1)}\right\}.$$

This proves the adaptive-selection statement. □

Remark. The oracle inequality separates training uncertainty from validation uncertainty. It justifies searching the $K \times M$ triangle statistically, but it does not prescribe a particular computational strategy for fitting every candidate.

D.9 Proof of Proposition 7

Proof. The proof separates the structural statement, which requires only a fixed nested partition model, from the rate statement, which additionally requires a regular and information-stable partition.

Step 1: Construct the partition-adapted nested family. For $t \in I_j$, define the cell-average operator

$$(P_{\pi_K} f)(t) = \overline{f}_j = \frac{1}{|I_j|} \int_{I_j} f(u)\, du.$$

Let $\phi_0, \phi_1, \ldots$ be the continuous cosine basis used in the proof of Theorem 3, and let the mth partition-adapted direction be the vector of cell averages $P_{\pi_K}\phi_m$. Weighted orthogonalization may be used numerically, but it does not change the nested spans

$$\mathcal{V}_{\pi_K,M} = P_{\pi_K} \operatorname{span}\{\phi_0, \ldots, \phi_{M-1}\}.$$

By assumption, the first K averaged directions have full rank. Hence

$$\mathcal{V}_{\pi_K,1} \subseteq \cdots \subseteq \mathcal{V}_{\pi_K,K} = \mathbb{R}^K,$$

and the associated affine log-hazard families satisfy

$$\mathcal{F}_{\pi_K,1} \subseteq \cdots \subseteq \mathcal{F}_{\pi_K,K}.$$

For equal-width cells and $m < K$,

$$\frac{1}{|I_j|} \int_{I_j} \phi_m(t)\, dt = c_{m,K}\phi_m(t_j), \qquad c_{m,K} = \frac{\sin\{\pi m/(2K)\}}{\pi m/(2K)},$$

with $c_{0,K} = 1$. Every $c_{m,K}$ is nonzero, so these averaged directions span exactly the DCT-II directions in Section 3.

Step 2: Recover the exact decomposition for a fixed partition. Let $\eta^*_{\pi_K}$ and $\eta^*_{\pi_K,M}$ minimise population risk over $\mathcal{F}_{\pi_K,K}$ and $\mathcal{F}_{\pi_K,M}$. The proof of Theorem 2 uses only affine nesting and first-order optimality, not equal cell widths. Applying the same three-point Bregman identity therefore gives

$$\mathcal{E}(\widehat{\eta}_{\pi_K,M}) = D_0(\eta^*_{\pi_K}, \eta_0) + D_0(\eta^*_{\pi_K,M}, \eta^*_{\pi_K}) + D_0(\widehat{\eta}_{\pi_K,M}, \eta^*_{\pi_K,M}).$$

These are $A_{\mathrm{grid}}(\pi_K)$, $A_{\mathrm{effect}}(\pi_K, M)$, and $E_{\mathrm{est}}(\pi_K, M)$. Thus placement changes the projection spaces but not the exact accounting identity. No quasi-uniformity is required for this step.

Step 3: Control approximation on an irregular partition. Consider the partition-adapted candidate

$$\widetilde{\eta}_{\pi_K,M}(t, Z) = (P_{\pi_K} a_0)(t) + (P_{\pi_K} b_M)(t)^\top Z, \qquad b_M = \sum_{m<M} \theta_m \phi_m.$$

It belongs to $\mathcal{F}_{\pi_K,M}$ by construction. For $f \in H^r[0,1]$, $0 < r \le 1$, the cellwise fractional Poincaré inequality gives

$$\|f - P_{\pi_K} f\|^2_{L_2} \le C h(\pi_K)^{2r} |f|^2_{H^r}.$$

Because $s \ge r$, this applies to both a_0 and the represented effect functions. Cell averaging is also an L_2 contraction. Indeed, Jensen's inequality gives

$$\|P_{\pi_K} f\|^2_{L_2} = \sum_{j=1}^K |I_j| |\overline{f}_j|^2 \le \sum_{j=1}^K \int_{I_j} |f(t)|^2 dt = \|f\|^2_{L_2}.$$

Consequently, the Sobolev ellipsoid in (15) implies

$$\|P_{\pi_K}(b_0 - b_M)\|^2_{L_2} \le \|b_0 - b_M\|^2_{L_2} = \sum_{m \ge M} \|\theta_m\|_2^2 \le C M^{-2s}.$$

The candidate error follows the explicit path

$$\eta_0 - \widetilde{\eta}_{\pi_K,M} = (a_0 - P_{\pi_K} a_0) + (b_0 - P_{\pi_K} b_0)^\top Z + \{P_{\pi_K}(b_0 - b_M)\}^\top Z.$$

Combining these cell-approximation and truncation errors, using the covariance bound for Z and quadratic equivalence of D_0 on the bounded sieve, yields

$$D_0(\widetilde{\eta}_{\pi_K,M}, \eta_0) \le C\{h(\pi_K)^{2r} + M^{-2s}\}.$$

Since $\eta^*_{\pi_K,M}$ is the population minimiser, its approximation risk is no larger.

Step 4: Control estimation uniformly over regular partitions. The partition-specific model has $p = K + Md$ free parameters. Under the stated uniform conditioning of its basis and information Gram matrices, the conditions of Lemma 8 hold with constants independent of π_K. Applying the lemma therefore gives, with probability at least $1 - \delta$,

$$D_0(\widehat{\eta}_{\pi_K,M}, \eta^*_{\pi_K,M}) \le C\frac{K + Md + \log(1/\delta)}{n}.$$

Adding this estimate to the approximation bound proves (12).

Step 5: Recover the original rate exponents. For a regular partition,

$$\frac{c}{K} \le |I_j| \le \frac{C}{K} \qquad \Longrightarrow \qquad h(\pi_K)^{2r} \le C^{2r} K^{-2r}.$$

Substitution into (12) recovers

$$\mathcal{E}(\widehat{\eta}_{\pi_K,M}) \le C' \left\{ K^{-2r} + M^{-2s} + \frac{K + Md + \log(1/\delta)}{n} \right\}.$$

Thus regular placement changes constants and information energies, but preserves the two approximation–estimation balances and their oracle exponents. □

Why regularity is necessary. The exact decomposition remains valid for any fixed admissible partition, but the rate statement cannot hold uniformly over arbitrary placements. If $K - 1$ cutpoints are packed into one half of the horizon, then $h(\pi_K) \ge 1/2$, and a smooth function varying on the remaining cell has nonvanishing grid error. Extremely narrow or low-exposure cells can likewise make the information matrix nearly singular. These failures affect approximation or conditioning; they do not remove the separate effect dimension M.

Finite placement selection. If an independent validation sample selects from a finite set of regular partitions together with (K, M), apply the oracle inequality in Appendix D.8 to the enlarged candidate index. The only additional selection price is the logarithm of the number of candidate partitions divided by the validation size.

E Controlled Experiments

E.1 Data-generating processes

All controlled experiments use the horizon $[0, 1]$ and $Z \sim N(0, I_d)$. Given log-hazard $\eta_0(t, Z)$, draw $E \sim \mathrm{Exp}(1)$ and define

$$T = \inf\left\{ t : \int_0^t \exp\{\eta_0(u, Z)\}\, du \ge E \right\}, \qquad Y = \min(T, C, 1), \qquad \Delta = \mathbf{1}\{T \le C,\ T \le 1\}. \tag{66}$$

Cumulative hazards are evaluated on 4,096 equal-width numerical intervals by trapezoidal integration, with the event time solved within the crossing interval. Write

$$\phi_0(t) = 1, \qquad \phi_m(t) = \sqrt{2}\cos(\pi m t), \quad m \ge 1, \qquad g(t; c, w) = \exp\left\{ -\frac{(t - c)^2}{2w^2} \right\}. \tag{67}$$

Finite-rank effects. The finite-rank truth is

$$\eta_{\mathrm{FR}}(t, Z) = \log(0.5 + t) + \sum_{m=0}^{7} v_{m+1}\phi_m(t)Z_{m+1}, \tag{68}$$

where $v = (0.35, 0.32, 0.29, 0.26, 0.23, 0.20, 0.17, 0.14)$. It has exactly $M_0 = 8$ temporal effect modes and only administrative censoring at $t = 1$.

Localised effects. The localised truth is

$$\begin{aligned}\eta_{\mathrm{LOC}}(t, Z) = {} & \log\kappa + \log(0.5 + t) + 0.25g(t; 0.30, 0.07) + 0.20g(t; 0.72, 0.08) \\ & + \sum_{\ell=1}^{8} v_\ell g(t; c_\ell, w_\ell)Z_\ell,\end{aligned} \tag{69}$$

where

$$\begin{aligned}v &= (0.35, 0.32, -0.29, 0.26, 0.23, -0.20, 0.17, -0.14), \\ c &= (0.25, 0.25, 0.35, 0.35, 0.68, 0.68, 0.78, 0.78), \\ w &= (0.06, 0.10, 0.06, 0.11, 0.06, 0.10, 0.05, 0.09).\end{aligned} \tag{70}$$

Independent censoring follows $C \sim \mathrm{Exp}(\nu)$. We use $\kappa = 2.087$ and $\nu = 0.6538$, calibrated to $\Pr(T \le 1) = 0.90$ and $\Pr(T \le C, T \le 1) = 0.70$. Gaussian bumps have nonzero coefficients at every DCT frequency, so no finite M recovers this effect exactly.

Regularity families. The effect-regularity experiment uses

$$b_s(t) = c_s \sum_{m=1}^{L} m^{-(s+1/2)} u_m \sqrt{2}\cos(\pi m t), \qquad L = 256, \tag{71}$$

where $u_m \in \mathbb{R}^8$ are fixed random unit vectors and c_s normalises $\int_0^1 \|b_s(t)\|_2^2\, dt$ to 0.5204. Its baseline is $\log(0.5 + t)$. The baseline-regularity experiment instead uses

$$a_r(t) = a_\star + c_r \sum_{m=1}^{L} m^{-(r+1/2)} \cos(2\pi m t + \omega_m), \tag{72}$$

where the phases ω_m are fixed by the DGP seed, c_r gives integrated squared baseline variation 0.16, and $a_\star$ gives event probability 0.70 by the horizon. Its time-invariant effect is

$$b_0 = (0.20, 0.18, 0.16, 0.14, 0.12, 0.10, 0.08, 0.06)^\top.$$

Four DGP seeds determine the directions and phases. The labels r and s describe the imposed spectral decay over the tested frequency range. Repeating the relevant experiments with $L = 512$ checks sensitivity to the finite expansion.

E.2 Geometry and information-threshold experiments

The principal geometry experiments use $d = 8$, $n = 3000$, 50 paired repetitions, and a common test population of 100,000. Every pair satisfying $K = 4, \ldots, 40$ and $1 \le M \le K$ is evaluated, giving 814 candidates per repetition. Excess risk is measured against the known continuous-time truth rather than the fitted grid.

For the finite-rank truth, theoretical approximation plus sandwich estimation correlates 0.9997 with empirical excess risk across the complete triangle. The fitted grid-approximation slope is -1.91 over $K \ge 4$ and -1.97 over $K \ge 16$. After subtracting population approximation, the empirical estimation increments are 0.000156 per baseline interval and 0.001402 per effect mode, compared with $1/(2n) = 0.000167$ and $d/(2n) = 0.001333$.

The information-threshold experiment fixes $K = 32$, searches $M = 1, \ldots, 16$, and varies $n \in \{3000, 6000, 12000, 24000\}$ and $d \in \{8, 16, 32, 64\}$ under target censoring rates of 30%, 50%, and

Table 3: Selected controlled comparisons in excess continuous RC-NLL. Entries are means and 95% intervals across 50 repetitions.

Truth	Comparison	(K, M)	Mean [95% interval]
Finite rank	global empirical optimum	$(27, 8)$	0.01727 [0.01665, 0.01789]
Finite rank	best unrestricted boundary	$(10, 10)$	0.02782 [0.02727, 0.02836]
Finite rank	fine unrestricted endpoint	$(40, 40)$	0.06413 [0.06263, 0.06562]
Localised	global empirical optimum	$(15, 5)$	0.01345 [0.01294, 0.01395]
Localised	structured model at $K = 16$	$(16, 5)$	0.01375 [0.01327, 0.01424]
Localised	unrestricted model at $K = 16$	$(16, 16)$	0.02659 [0.02578, 0.02740]

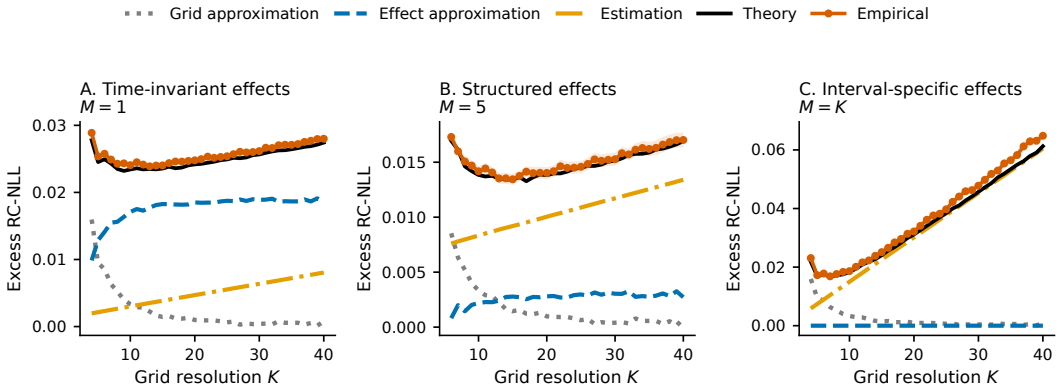

Figure 5: Risk decomposition under localised non-finite-rank effects. The structured $M = 5$ path balances effect approximation against estimation, while the unrestricted path becomes estimation dominated.

70%. Each regime targets 100 paired repetitions and uses a common test population of 200,000. Modes 2–8 are active, while modes 9–16 satisfy the independent population null-mode check.

Across 684 regular adjacent-mode comparisons, the empirical risk increment has zero-intercept slope 0.986 and correlation 0.973 with the local prediction $-E_m/2 + d/(2n)$. The ideal and sandwich-adjusted rules identify the empirical optimum in 44 of 48 regimes. All regular settings with identical d/n select the same empirical effect dimension.

E.3 Regularity and smoothing extensions

E.3.1 Regularity scaling

The effect branch fixes $d = 8$ and $K = 64$, varies $s \in \{0.5, 1, 2\}$ and $n \in \{1500, 3000, 6000, 12000, 24000, 48000\}$, and searches $M = 1, \ldots, 32$. The two largest-sample $s = 0.5$ settings extend the search to $M = 56$. The baseline branch fixes $M = 1$, varies $r \in \{0.25, 0.5, 1\}$ over the same sample sizes, and searches K up to 320 where required. Each setting uses four DGP seeds, 30 paired repetitions per seed, and a fixed test population of 100,000. No selected optimum remains at the upper candidate boundary.

The fitted log–log exponents are computed across sample sizes within each regularity and then summarised over the four DGP seeds. They test the predicted scaling exponents rather than equality of finite-sample constants.

E.3.2 Hard truncation and smooth shrinkage

The smoothing comparison fixes $K = 32$ and uses the finite-rank and localised DGPs above. For each truth, 100 paired repetitions use $n_{\text{train}} = n_{\text{val}} = 3000$, followed by refitting on the combined $n = 6000$ observations and evaluation on a common test population of 100,000. Hard truncation

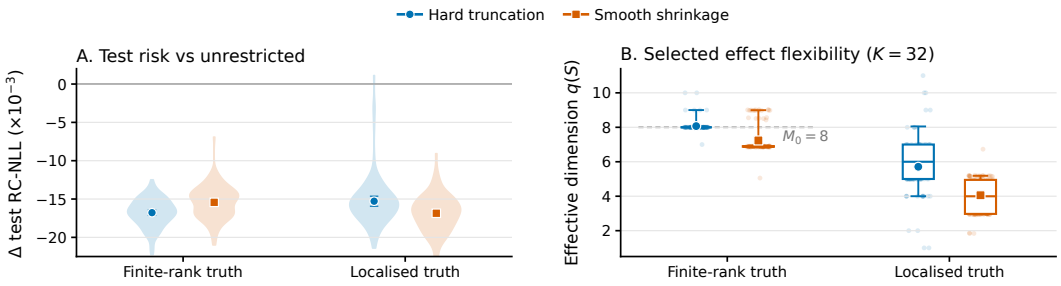

Figure 6: **Hard truncation and smooth shrinkage.** At $K = 32$, both validation-selected procedures improve test RC-NLL over the unrestricted boundary and select $q(S) \ll K$.

searches $M = 1, \ldots, 32$. Smooth shrinkage retains all K modes and adds

$$\rho \sum_{m=1}^{K-1} (\pi m)^{2p} \, \|\gamma_m\|_2^2, \qquad p \in \{1, 2\}, \tag{73}$$

leaving the constant mode unpenalised. Validation searches $\rho = 10^{-12}, 10^{-11.5}, \ldots, 10^{4}$, together with $\rho = 0$.

To connect the fitted penalty with Section 4.5, let H_{eff} denote the effect Hessian after profiling the K interval-specific intercepts, and let ν_j be the generalised eigenvalues of the penalty matrix relative to H_{eff}. The local shrinkage factors and reported variance-effective temporal dimension are

$$s_j(\rho) = \frac{1}{1 + 2\rho\nu_j}, \qquad q_{\text{eff}}(\rho) = \frac{1}{d} \sum_{j=1}^{Kd} s_j(\rho)^2. \tag{74}$$

This quantity is reported as $q(S)$ in Figure 6. It equals M for hard truncation and K when $\rho = 0$.

Under the finite-rank truth, truncation, shrinkage, and the unrestricted boundary have excess risks 0.00982, 0.01117, and 0.02659. Under the localised truth, the corresponding risks are 0.00987, 0.00830, and 0.02516. Both structured procedures reduce estimation relative to the unrestricted boundary. Their reversed ordering reflects approximation: a hard cutoff matches finite-rank effects, while smooth shrinkage distributes its effective dimension across the many weak modes of a localised effect.

E.4 Optimisation and completeness

The controlled geometry, threshold, and regularity experiments minimise the convex person-period likelihood. Fits are accepted when the largest absolute mean-loss gradient component is below 10^{-6}. Representative cells independently optimised by L-BFGS-B agree in objective value within 10^{-6}.

All geometry and regularity fits pass the acceptance checks. In the threshold experiment, 178 of 76,800 requested fits fail the regularity diagnostic. These failures occur only under 70% censoring in high-d/n, high-M regimes. The local-law analysis uses the regular subset defined independently of agreement with the theoretical prediction.

F Real-Data Experiments

F.1 Datasets and preprocessing

Experiments span diverse clinical settings. FLCHAIN is a population-based study of serum free light chains and overall survival (Dispenzieri et al., 2012); METABRIC and Rotterdam are breast-cancer cohorts (Curtis et al., 2012; Foekens et al., 2000; Royston & Altman, 2013); NWTCO contains paediatric Wilms-tumour records (Breslow & Chatterjee, 1999); and SUPPORT follows seriously ill hospitalised adults (Knaus et al., 1995). We obtain FLCHAIN, NWTCO, SUPPORT, and the processed GBSG benchmark through `pycox` (Kvamme et al., 2019). GBSG combines the Rotterdam Tumour Bank and German Breast Cancer Study Group cohorts, totalling 2,232 patients (Foekens et al., 2000; Schumacher et al., 1994).

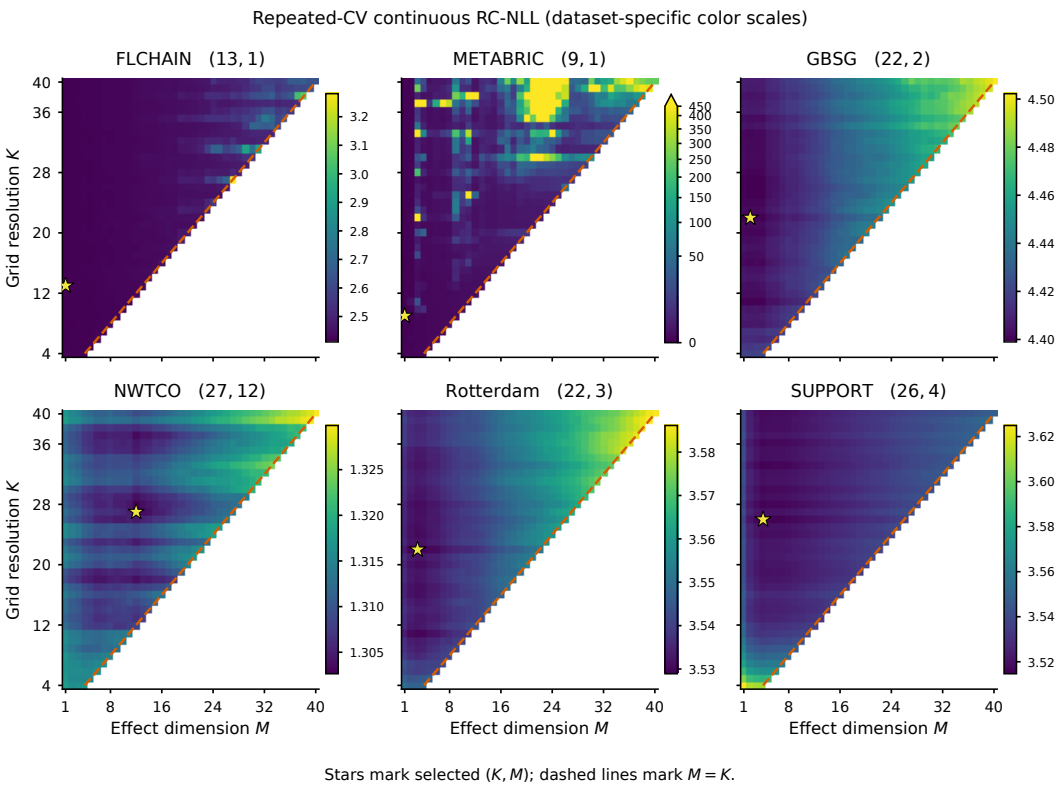

Figure 7: **Fixed-representation RC-NLL surfaces.** Stars mark selected (K, M) and dashed lines mark $M = K$. Dataset-specific colour scales cap values above the 95th percentile for display only. All cells remain eligible for selection.

Within each training fold, missing values are median-imputed, features are standardised, and constant or linearly dependent features are removed. The fitted transformations are applied unchanged to validation and test data. Three FLCHAIN deaths recorded at day zero are assigned day 0.5.

F.2 Fixed-representation experiment

Raw standardised covariates serve as Z. Each dataset evaluates the complete triangle $K = 4, \dots, 40$, $M \le K$ using ten repetitions of stratified fivefold cross-validation with shared candidate splits.

Models use L-BFGS-B with baseline ridge 10^{-8} and effect penalty

$$\frac{0.01}{K} \left\| \Phi_{K,M} \Gamma \right\|_{\mathrm{F}}^2 ,$$

which treats full-DCT and direct interval-specific parameterisations equivalently. Selection uses pooled subject-level held-out RC-NLL without the penalty. Uncertainty is calculated across the ten repetitions.

Smooth-effect comparator. To test whether the value of controlling effect flexibility is specific to hard truncation, we repeat the fixed-representation study using all K DCT modes with the second-order spectral penalty in Eq. (73). For each $K = 4, \dots, 40$, validation searches $\rho \in \{0, 10^{-12}, 10^{-11.5}, \dots, 10^{4}\}$ using the same repeated cross-validation splits and continuous RC-NLL criterion. Here $\rho = 0$ is the unrestricted interval-specific boundary, while $\rho > 0$ controls a fractional effective dimension $q(S)$.

Figure 8 shows that validation selects positive, interior smoothing levels on all six datasets, although their magnitudes vary substantially. Thus, the empirical value of controlling effect flexibility is not specific to a hard cutoff M: a full head with tuned smooth shrinkage exhibits the same joint resolution–flexibility geometry. This supports effective flexibility $q(S)$, rather than nominal output width alone, as the relevant complexity measure.

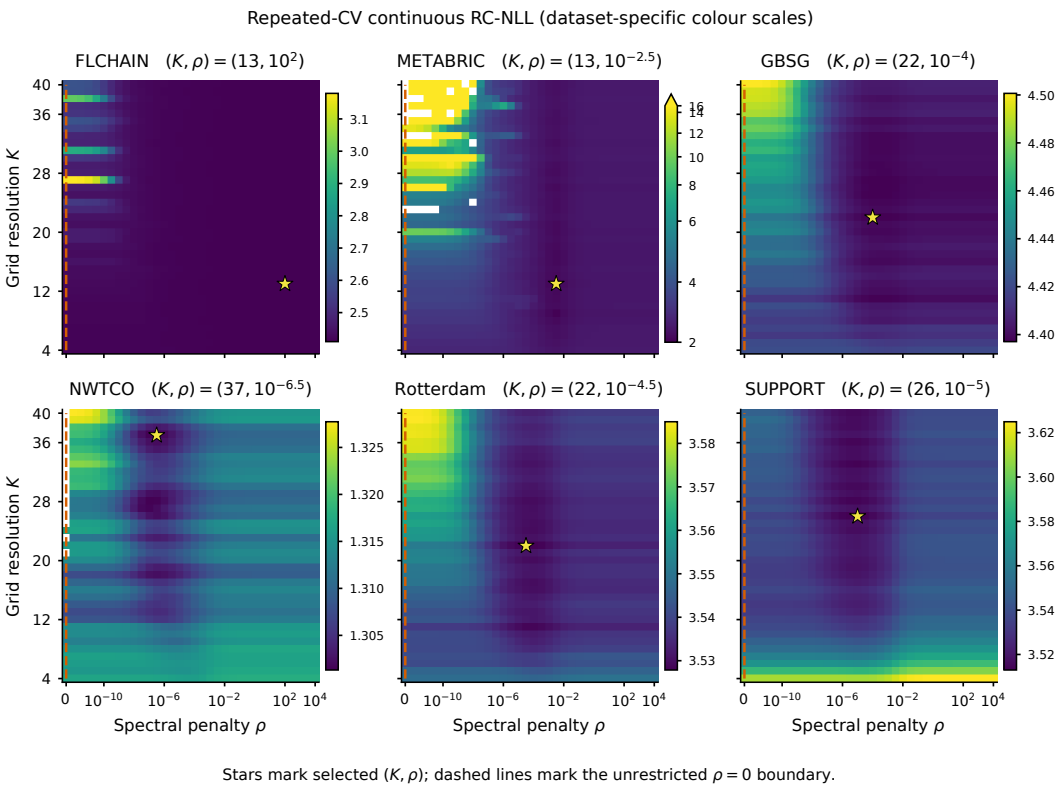

Figure 8: **Fixed-representation $K \times \rho$ RC-NLL surfaces.** Full K-mode heads are fitted with a second-order spectral roughness penalty. Stars mark the cross-validated (K, ρ), and dashed lines mark the unrestricted $\rho = 0$ boundary. Dataset-specific colour scales are used for visibility.

F.3 Learned-representation reconstruction and nested selection

For every head, the pre-transformation predictor is

$$\boldsymbol{\xi}(Z) = \boldsymbol{\alpha} + \boldsymbol{\Phi}_{K,M}\boldsymbol{\Gamma}Z. \tag{75}$$

If a head produces interval failure probability h_{ij} conditional on survival to interval j, its continuous piecewise-exponential rate is

$$\lambda_{ij} = -\frac{\log(1 - h_{ij})}{|I_j|}. \tag{76}$$

For a PMF-style head producing mass p_{ij},

$$S_{i,j-1} = 1 - \sum_{\ell<j} p_{i\ell}, \qquad h_{ij} = \frac{p_{ij}}{S_{i,j-1}}. \tag{77}$$

Nnet-survival uses the first conversion. PMF and N-MTLR use the second after their native probability transformations. PC-Hazard uses

$$\lambda_{ij} = \frac{\mathrm{softplus}(\xi_{ij})}{|I_j|}.$$

These transformations preserve survival at every interval boundary. For an observation in interval j, all heads are evaluated using

$$\ell_i = \sum_{\ell<j} \lambda_{i\ell}|I_\ell| + \lambda_{ij}(Y_i - \tau_{j-1}) - \Delta_i \log \lambda_{ij}. \tag{78}$$

The common score is used for validation and outer-test evaluation, while each head retains its native training likelihood. At $M = K$, full rank of $\Phi_{K,K}$ makes the pre-transformation covariate-dependent output matrix unrestricted. Candidates use $K \in \{8, 16, 24, 32, 40, 48, 64\}$ and $M = 1, \ldots, K$.

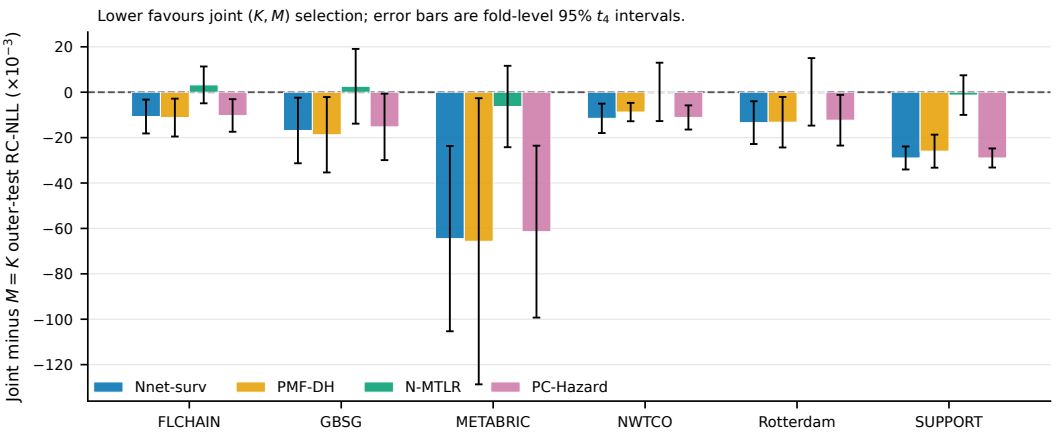

Figure 9: **Learned-representation outer-test RC-NLL.** Differences are joint (K, M) selection minus selection restricted to $M = K$; lower values favour joint selection. Error bars are 95% t_4 intervals.

Each dataset uses five stratified outer folds, with 20% of each outer-training fold reserved for inner validation and three paired training seeds.

The shared encoder has three hidden layers of width 64, ReLU activation, dropout 0.1, and a 32-dimensional representation followed by non-affine layer normalisation. Training uses Adam with learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 128, cosine learning-rate decay, gradient clipping at 5, and at most 200 epochs. Early stopping uses patience 20 and minimum improvement 10^{-5} in validation continuous RC-NLL.

Within each outer fold and seed, inner RC-NLL selects jointly over (K, M) or, for the comparator, over K with $M = K$. Both are evaluated on the paired untouched outer fold. Descriptive intervals average paired seed differences within each fold and apply a t_4 multiplier across the five fold means. This yields 83,520 fits and 360 paired outer-test comparisons.

Secondary outer-test metrics are computed from reconstructed survival curves using SurvivalEVAL 0.8.0 (Qi et al., 2023) for Harrell's C-index (Harrell Jr et al., 1996), Uno's IPCW C-index (Uno et al., 2011), and IPCW integrated Brier score (IBS) (Graf et al., 1999), and scikit-survival 0.24.1 (Pölsterl, 2020) for integrated cumulative/dynamic AUC (Uno et al., 2007). For both C-indices, linearly interpolated curves are summarised by negative RMST over the common evaluation horizon, $R_i = -\int_0^\tau \widehat{S}_i(t)\,dt$, rather than risk at a single time point (Sierra et al., 2025). Integrated AUC instead uses the time-dependent risk $1 - \widehat{S}_i(t)$. AUC and IBS are integrated over the same 100-point grid within each outer fold, with IPCW quantities estimated from the training fold. Higher values are better except for IBS. Figure 10 reports paired mean differences, reversing the IBS sign so that positive values uniformly favour joint selection. Joint selection improves Harrell's C-index in 18 of 24 dataset–head strata, Uno's C-index in 17, integrated AUC in 16, and IBS in 18. Gains are most consistent for Nnet-surv and PC-Hazard, while N-MTLR changes less consistently.

Figure 11 illustrates the architecture dependence on Rotterdam. Interior M stabilises Nnet-survival and PMF-DH as K increases, whereas N-MTLR changes little across effect paths, consistent with its native cross-time sharing.

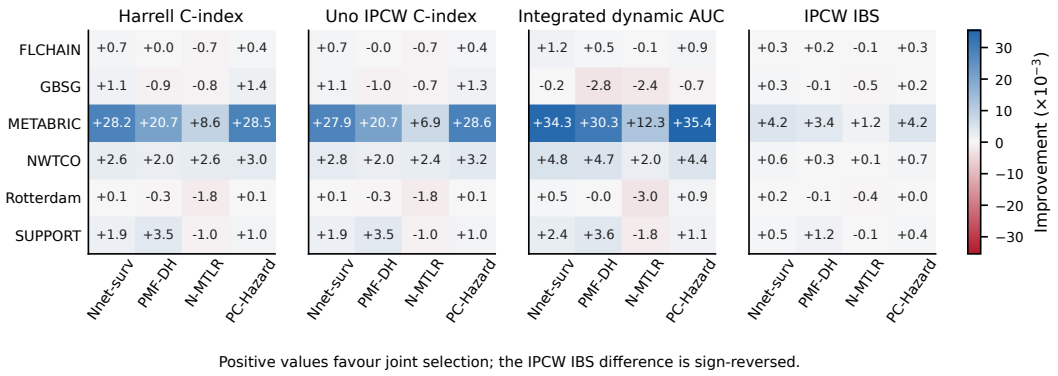

Figure 10: **Secondary predictive metrics.** Cells show 10^3 times the paired outer-test improvement under joint (K, M) selection; the IBS difference is sign reversed.

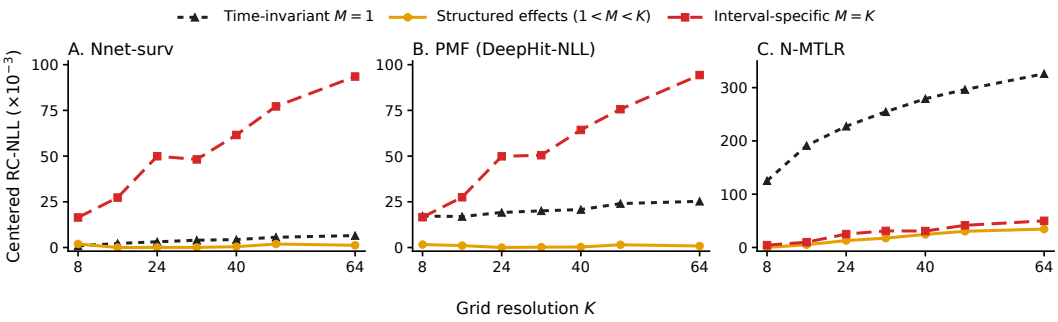

Figure 11: **Learned-representation effect-structure paths on Rotterdam.** Descriptive outer-test RC-NLL along the time-invariant ($M = 1$), best structured ($1 < M < K$), and interval-specific ($M = K$) paths.