

Mortality Prediction as Boundary Value Problem

Victor Zhorin

February 2026

Abstract

We present a mathematical framework for mortality prediction from discrete medical event sequences, formulating the problem as a boundary value problem on causal path space. Discrete event sequences lift to rough path space; signatures provide coordinate-free trajectory analysis. The transformer architecture emerges as computing weighted projections of path signatures, with Time2Vec temporal encoding providing spectral parameterization of the rough path lift. We establish connections to tropical geometry: ReLU networks compute tropical rational functions, and the decision boundary is a tropical hypersurface interpretable as the viscosity solution to a Hamilton-Jacobi equation on trajectory space. Mortality acts as absorbing boundary condition, with gradients propagating retroactively to reshape early event representations, the computational realization of “boundary conditions determine interior.” The framework unifies perspectives from causal set theory [1], rough path theory [15], tropical geometry [16], and viscosity solutions [5]. The architecture’s parameter efficiency is explained by alignment with the problem’s intrinsic mathematical structure: tropical decision boundaries, Lie algebraic dimension reduction, and binary boundary factorization.

1 Why This Mathematics: Discontinuity as Ontological Fact

1.1 The Clinical Reality

Consider a 28-year-old female. Her baseline annual mortality probability, drawn from actuarial life tables, is approximately 3×10^{-4} . A diagnosis of secondary malignant neoplasm of bone (ICD-10 C79.5) elevates her mortality probability to roughly 0.3: three orders of magnitude above baseline. This is not a steep gradient. It is a shock: a discontinuity in the mortality function on trajectory space, induced by a single discrete event. The jump occurs not in physical time but in information space, the diagnosis does not change her physiological state at the instant of coding, but it changes everything the value function must encode about her future.

Such shocks are not pathological edge cases. They are the generic structure of mortality as a function on medical event sequences. A previously healthy 45-year-old male receiving a diagnosis of glioblastoma (C71.9) undergoes a comparable discontinuity. So does a 60-year-old upon diagnosis of pancreatic cancer (C25.9), or a 35-year-old with ALS (G12.21). Every catastrophic diagnosis is a shock front propagating through the value function.

1.2 The Inadequacy of Smooth Approximation

Standard approaches to mortality prediction uniformly assume, explicitly or implicitly, that the mapping from patient state to mortality risk is smooth or at least continuous.

- Cox proportional hazards [4] models a smooth baseline hazard modulated by multiplicative covariate effects: a log-linear structure that cannot represent discontinuous jumps without infinite coefficient values.
- Logistic regression and its regularized variants (LASSO, elastic net) approximate the discontinuity by steepening a sigmoid, but the function remains C^∞ everywhere. The approximation error concentrates precisely at the shock, where clinical accuracy matters most.
- Standard neural networks with smooth activations (tanh, sigmoid, GELU) inherit the same limitation: they are universal approximators of continuous functions, and converge to discontinuities only in the limit of infinite width.
- Stochastic mortality surface models [2, 13] achieve parsimony through parametric structure but impose continuity on the mortality intensity process: precisely the assumption that catastrophic diagnoses violate.
- Gradient boosted trees (XGBoost, LightGBM) achieve piecewise-constant approximation but without the variational structure that selects the physically meaningful solution among many possible piecewise approximations.

All of these fit smooth or ad hoc piecewise functions to something that is not smooth. The approximation may achieve acceptable aggregate metrics, but it is ontologically wrong: it treats the discontinuity as a nuisance to be smoothed away rather than the fundamental structure to be represented.

1.3 Viscosity Solutions: Mathematics Built for Shocks

The theory of viscosity solutions, developed by Crandall and Lions [5], exists precisely because Hamilton-Jacobi equations generically develop gradient discontinuities. Classical (differentiable) solutions cease to exist after finite time. The viscosity framework provides [6]:

1. **Existence:** Solutions persist through and beyond shock formation.
2. **Uniqueness:** The comparison principle selects the unique physically meaningful weak solution, the one obtained as the vanishing-viscosity limit $\varepsilon \rightarrow 0^+$ of regularized (diffusive) equations.
3. **Stability:** Viscosity solutions are stable under perturbation, including the discretization implicit in numerical (and neural network) computation.

No competing mathematical framework provides all three simultaneously for the class of problems mortality prediction presents:

- Hyperbolic conservation laws (Rankine-Hugoniot conditions) describe shock propagation in physical space, but the relevant object here is a value function on trajectory space, not a conserved quantity.
- Lévy processes and jump-diffusion models handle stochastic jumps but impose parametric assumptions on jump-size distributions: inapplicable when the magnitude and meaning of each discontinuity depend on the full trajectory context.
- Distributional and Colombeau solutions extend classical PDE theory to handle singularities but sacrifice uniqueness, precisely what a prediction system cannot afford.
- Measure-valued solutions generalize further but again lose the selection principle that determines which weak solution corresponds to reality.

1.4 ReLU Networks as Native Viscosity Solvers

The connection closes through tropical geometry (Section 6). A feedforward ReLU network computes a tropical rational function: a difference of piecewise-linear convex functions [20]. Such functions are exactly the class that arises as viscosity solutions to Hamilton-Jacobi equations with piecewise-linear initial data [7].

The decision boundary of the mortality classifier is a tropical hypersurface: a piecewise-linear complex where competing linear regimes meet. At each face of this complex, the mortality value function is non-differentiable, the gradient jumps. These are the learned shock fronts.

Critically, the architecture does not approximate the discontinuity by steepening a smooth function. It represents it exactly, in finitely many parameters, through the piecewise-linear structure that ReLU activation provides natively. The shocks are not smoothed; they are resolved.

Remark 1.1 (Parameter Efficiency from Structural Alignment). *This structural alignment explains the architecture’s parameter efficiency. Representing a genuine discontinuity by smooth approximation requires parameters scaling with the reciprocal of the desired approximation error at the shock. Representing it as a tropical hypersurface requires only the specification of the linear regions and their boundaries: $O(k)$ parameters for k pieces, independent of the magnitude of the jump. A three-order-of-magnitude mortality shock and a factor-of-two risk elevation cost the same number of parameters. The mathematics does not penalize the severity of the discontinuity.*

2 Discrete Causal Structure

2.1 Event Space and Causal Ordering

Definition 2.1 (Medical Event Space). *Let \mathcal{E} be a finite vocabulary of medical event types (ICD-10 codes, procedure codes, drug codes). A timestamped event is a pair $(e, t) \in \mathcal{E} \times \mathbb{R}^+$ where t denotes occurrence time.*

Definition 2.2 (Patient Trajectory). *A trajectory is a finite sequence $\gamma = ((e_1, t_1), \dots, (e_n, t_n))$ with $t_1 \leq t_2 \leq \dots \leq t_n$. The space of all trajectories is the Kleene closure:*

$$\Gamma = \bigcup_{n=0}^{\infty} (\mathcal{E} \times \mathbb{R}^+)^n$$

equipped with the causal partial order $i \prec j \iff t_i < t_j$. See Section 7 for the embedding into rough path space.

Remark 2.3 (Kleene Closure as Discrete Path Space). *The Kleene closure Γ is the computer scientist’s path space. It embeds into the analyst’s rough path space $\Omega_p(\mathbb{R}^d)$ via the piecewise-constant lift: each discrete trajectory $\gamma = ((e_1, t_1), \dots, (e_n, t_n))$ maps to a càdlàg path $\tilde{\gamma} : [0, T] \rightarrow \mathbb{R}^d$ with $\tilde{\gamma}(t) = \phi(e_i)$ for $t \in [t_i, t_{i+1})$. This piecewise-constant lift provides the bridge to analytic path space (see Section 7 for the signature formalism).*

This embedding is measure-theoretically natural: the empirical distribution on Γ (observed trajectories) pushes forward to a measure on path space that rough path theory can analyze [15].

Remark 2.4 (Connection to Causal Set Theory). *In quantum gravity, spacetime is hypothesized to be fundamentally discrete, with causal ordering as primitive structure [1, 18]. The patient trajectory γ is analogous to a causal chain, a totally ordered subset of the causal set. Mortality prediction becomes: given a partial causal chain, estimate probability of intersection with a boundary region $\partial\Omega$.*

2.2 Embedding Map

Definition 2.5 (Event Embedding). *The embedding map $\phi : \mathcal{E} \rightarrow \mathbb{R}^d$ assigns each event type a learned vector representation.*

The embedding space $H = \mathbb{R}^d$ carries implicit geometric structure learned from co-occurrence patterns. Events that predict similar outcomes cluster; causal precursors align with their consequences.

3 Temporal Encoding as Spectral Lifting

3.1 The Problem of Irregular Time

Medical events occur at irregular intervals. Standard positional encodings assume uniform spacing. The timestamp sequence (t_1, \dots, t_n) contains critical information: rapid event acceleration often precedes mortality.

3.2 Time2Vec: Spectral Decomposition of Time

Definition 3.1 (Temporal Lifting [11]). *The Time2Vec map $\tau : \mathbb{R}^+ \rightarrow \mathbb{R}^{2K+1}$ is defined by:*

$$\tau(t) = (\omega_0 t, \sin(\nu_1 t + \phi_1), \cos(\nu_1 t + \phi_1), \dots, \sin(\nu_K t + \phi_K), \cos(\nu_K t + \phi_K))$$

where $\omega_0 \in \mathbb{R}$ (linear trend), and $\{(\nu_k, \phi_k)\}_{k=1}^K$ are learnable frequencies and phases.

Proposition 3.2 (Spectral Universality). *Any continuous periodic function $f : \mathbb{R} \rightarrow \mathbb{R}$ can be approximated arbitrarily well by linear combinations of the Time2Vec basis functions (Fourier’s theorem). With learnable frequencies, the representation extends to quasi-periodic and multi-scale patterns.*

Remark 3.3 (What the Frequencies Learn). *Empirically, learned frequencies ν_k span scales from sub-daily ($\nu \sim 2\pi/\text{day}$) to multi-year ($\nu \sim 2\pi/(5 \text{ years})$). They capture: circadian patterns in acute events; weekly/monthly medication cycles; seasonal disease patterns; and multi-year chronic disease progression. Critically, frequencies are learned globally across the population, not per-patient. Time2Vec discovers collective periodicities, the shared temporal grammar of medical event sequences.*

3.3 Combined Event-Time Representation

The full representation of a timestamped event (e_i, t_i) is:

$$\psi(e_i, t_i) = \phi(e_i) \oplus \tau(t_i) \oplus \xi(e_i, t_i, \gamma)$$

where \oplus denotes concatenation and ξ captures engineered features.

4 Attention as Learned Causal Influence

4.1 The Attention Mechanism

Given sequence representations $\{\psi_i\}_{i=1}^n$, multi-head attention [19] computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$$

where $Q = W_Q \Psi$, $K = W_K \Psi$, $V = W_V \Psi$ are learned projections.

4.2 Causal Interpretation of Attention Weights

Let $A_{ij} = [\text{softmax}(QK^\top/\sqrt{d_k})]_{ij}$ denote attention from position j to position i .

Proposition 4.1 (Attention as Soft Causal Graph). *Under causal masking ($A_{ij} = 0$ for $j > i$), the attention matrix A defines a weighted DAG over events, with edge weights learned to optimize prediction.*

Remark 4.2 (Second-Order Causal Effects). *Single attention layers capture direct influence. Stacked layers compute compositions:*

$$(A^{(2)} \cdot A^{(1)})_{ik} = \sum_j A_{ij}^{(2)} A_{jk}^{(1)}$$

This computes indirect influence: event k affects event i through intermediate event j . The connection between stacked attention and high-order causal chains relates to causal emergence in complex systems [9]. Deep attention networks learn arbitrarily high-order causal chains.

4.3 Information Aggregation

After L transformer layers, the sequence of representations $\{\psi_i^{(L)}\}$ is aggregated to a single trajectory representation:

$$h_\gamma = \text{Pool}(\psi_1^{(L)}, \dots, \psi_n^{(L)})$$

5 Mortality as Boundary Condition

5.1 The Boundary Value Formulation

Let $\partial\Omega \subset \Gamma$ denote the set of trajectories terminating in death within the prediction horizon. The prediction task is:

$$\hat{P}(\gamma \rightarrow \partial\Omega) = \sigma(w^\top h_\gamma + b)$$

where σ is the sigmoid function and (w, b) are the final classification parameters.

Remark 5.1 (Mortality as Attractor). *The vector $w \in \mathbb{R}^{d_h}$ defines a mortality direction in representation space. Training pushes trajectories of deceased patients toward high $w^\top h_\gamma$ and survivors toward low $w^\top h_\gamma$. The decision boundary $\{h : w^\top h + b = 0\}$ is a hyperplane separating life from death in representation space. This geometric structure relates to the HJB value function characterization developed in Fleming and Soner [7].*

5.2 Gradient Flow from Boundary

Training minimizes binary cross-entropy:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N [y_i \log \hat{P}_i + (1 - y_i) \log(1 - \hat{P}_i)]$$

where $y_i \in \{0, 1\}$ indicates death.

Proposition 5.2 (Retroactive Gradient Attribution). *Let $\gamma = (e_1, \dots, e_n)$ with label $y = 1$ (death). The gradient with respect to early event embedding:*

$$\frac{\partial \mathcal{L}}{\partial \phi(e_1)} = \frac{\partial \mathcal{L}}{\partial \hat{P}} \cdot \frac{\partial \hat{P}}{\partial h_\gamma} \cdot \frac{\partial h_\gamma}{\partial \psi_1^{(L)}} \cdot \prod_{\ell=1}^L \frac{\partial \psi_1^{(\ell)}}{\partial \psi_1^{(\ell-1)}} \cdot \frac{\partial \psi_1^{(0)}}{\partial \phi(e_1)}$$

propagates the mortality signal backward through time and through all transformer layers. Interpretation: The death event reaches backward through the causal chain, modifying how early events are represented. An early diagnosis that seemed benign is re-weighted once the model learns it preceded death. This is the computational realization of “boundary conditions shape interior.”

6 Tropical Geometry and Viscosity Solutions

This section develops the theoretical framework connecting ReLU network geometry to Hamilton-Jacobi PDEs via tropical algebra [16]. The discrete nature of the Kleene closure Γ makes tropical geometry natural: we work with piecewise-linear structures rather than smooth manifolds. The embedding into rough path space (Section 7) provides the analytic completion when needed.

6.1 Tropical Semiring and Max-Plus Algebra

Definition 6.1 (Tropical Semiring). *The tropical semiring $(\mathbb{R} \cup \{-\infty\}, \oplus, \odot)$ is defined by:*

$$\begin{aligned} a \oplus b &= \max(a, b) && \text{(tropical addition)} \\ a \odot b &= a + b && \text{(tropical multiplication)} \end{aligned}$$

with identity elements $-\infty$ for \oplus and 0 for \odot .

This algebra arises naturally as the “dequantization” limit of ordinary algebra [14]. Consider the logarithmic map $\log_{\hbar} : \mathbb{R}^+ \rightarrow \mathbb{R}$ and observe:

$$\lim_{\hbar \rightarrow 0^+} \hbar \log(e^{a/\hbar} + e^{b/\hbar}) = \max(a, b)$$

The smooth log-sum-exp becomes piecewise-linear max in the zero-temperature limit.

6.2 ReLU Networks as Tropical Rational Maps

Proposition 6.2 (ReLU-Tropical Correspondence [20]). *A feedforward network with ReLU activations computes a tropical rational function: a ratio of tropical polynomials. Specifically, for input $x \in \mathbb{R}^n$, the network output is:*

$$f(x) = \max_{i \in I^+} (A_i^+ x + b_i^+) - \max_{j \in I^-} (A_j^- x + b_j^-)$$

where the index sets I^+, I^- and affine coefficients are determined by network weights.

Sketch. $\text{ReLU}(z) = \max(0, z)$ is tropical addition with 0. Affine layers are tropical linear. Composition preserves tropical rationality. The difference of maxima has direct clinical interpretation: the first term aggregates evidence toward mortality (risk factors, disease progression signals), the second aggregates protective factors (treatment response, physiological reserve). The decision boundary $\{f = 0\}$ is where these competing influences balance. \square

Definition 6.3 (Tropical Hypersurface [16]). *The tropical hypersurface of a tropical polynomial $p(x) = \bigoplus_i (c_i \odot x^{\alpha_i}) = \max_i (c_i + \langle \alpha_i, x \rangle)$ is:*

$$\mathcal{T}(p) = \{x \in \mathbb{R}^n : \text{the maximum is achieved by at least two terms}\}$$

This is a piecewise-linear complex of codimension 1.

Conjecture 6.4 (Tropical Decision Boundary). *The decision boundary of the mortality classifier, pulled back through the network to input space, is a tropical hypersurface in trajectory representation space. Its combinatorial structure encodes the learned disease taxonomy, with each face of the complex corresponding to a transition between competing clinical risk regimes.*

6.3 Hamilton-Jacobi Equations and Viscosity Solutions

The connection to PDEs emerges through the following classical result:

Theorem 6.5 (Hopf-Lax Formula [7]). *Consider the Hamilton-Jacobi equation:*

$$\partial_t V + H(\nabla_x V) = 0, \quad V(x, 0) = V_0(x)$$

With convex H , the viscosity solution admits the variational representation:

$$V(x, t) = \inf_y \left\{ V_0(y) + t \cdot L\left(\frac{x - y}{t}\right) \right\}$$

where L is the Legendre transform of H .

Remark 6.6 (Convexity and Disease Dynamics). *Convexity of H is sufficient, not necessary. Viscosity solutions exist for non-convex Hamiltonians via Perron’s method [6], but lack closed-form representation. Whether the effective Hamiltonian governing mortality dynamics is convex remains open; non-convexity would correspond to multiple locally optimal disease progression pathways, consistent with multi-morbidity phenomenology.*

The viscosity solution is the unique weak solution satisfying a maximum principle, selected by adding vanishing diffusion $\varepsilon \Delta V$ and taking $\varepsilon \rightarrow 0^+$. This regularization procedure mirrors how neural network training (with noise, dropout, finite precision) selects among possible piecewise-linear decision boundaries [10].

6.4 Mortality Prediction as Hamilton-Jacobi Problem

Definition 6.7 (Value Function). *Define $V : \Gamma \times \mathbb{R}^+ \rightarrow \mathbb{R}$ as the mortality value function:*

$$V(\gamma, T) = \text{“cost-to-go from trajectory state } \gamma \text{ to mortality boundary within horizon } T\text{”}$$

Low V indicates high mortality risk; $V = 0$ on the boundary $\partial\Omega$.

Conjecture 6.8 (HJ Structure of Mortality Prediction). *The value function V satisfies, in a weak sense, a Hamilton-Jacobi-Bellman equation:*

$$\partial_T V + H(\gamma, \nabla_\gamma V) = 0$$

where the “gradient” $\nabla_\gamma V$ is understood via Fréchet derivative on path space or finite-dimensional projection via signatures [8]; the Hamiltonian $H(\gamma, p)$ encodes transition costs between medical states; and the boundary condition is $V(\gamma, T) = 0$ for $\gamma \in \partial\Omega$. The trained classifier approximates V evaluated at $T =$ prediction horizon. The piecewise-linear structure arising from ReLU activation is exactly the tropical-viscosity solution structure, and the decision boundary $\{V = c\}$ for threshold c propagates as a wavefront in trajectory space.

Remark 6.9 (Characteristics and Disease Trajectories). *Classical HJ theory solves via the method of characteristics, curves along which information propagates. In the mortality context, these correspond to canonical disease progression pathways determined by the (unknown) Hamiltonian. The attention mechanism may be performing an analogous computation: learning prototypical progression patterns and classifying patients by proximity to those patterns.*

6.5 Maslov Dequantization and the Classical Limit

The tropical-viscosity connection is a special case of Maslov dequantization [14]:

Quantum/Smooth ($\hbar > 0$)	Classical/Tropical ($\hbar \rightarrow 0^+$)
Schrödinger equation	Hamilton-Jacobi equation
Wave function ψ	Action $S = -i\hbar \log \psi$
Linear superposition	Max-plus superposition
Diffusion/spreading	Wavefront propagation
Soft decisions	Hard decisions

The ReLU network operates in the “classical limit”, making hard decisions at each neuron. Softmax smooths the tropical structure.

7 Rough Paths and Signature Methods

We connect the temporal structure to rough path theory [15], providing a coordinate-free characterization of trajectory shape. This section makes rigorous the embedding of the Kleene closure Γ into analytic path space.

7.1 Path Signatures

Definition 7.1 (Signature of a Path [3]). *Let $\gamma : [0, T] \rightarrow \mathbb{R}^d$ be a path of bounded variation. The signature of γ is the sequence of iterated integrals:*

$$S(\gamma) = (1, S(\gamma)^1, S(\gamma)^2, \dots)$$

where the k -th level is:

$$S(\gamma)_{i_1, \dots, i_k}^k = \int_{0 < t_1 < \dots < t_k < T} d\gamma_{t_1}^{i_1} \otimes \dots \otimes d\gamma_{t_k}^{i_k}$$

This lives in the tensor algebra $T((\mathbb{R}^d)) = \bigoplus_{k=0}^{\infty} (\mathbb{R}^d)^{\otimes k}$.

Theorem 7.2 (Chen [1958]; Hambly-Lyons [2010]). *The signature $S(\gamma)$ uniquely determines the path γ up to tree-like equivalence. For generic paths, the signature is a complete invariant.*

7.2 Truncated Signatures and Computational Tractability

In practice, we truncate at level M :

$$S^{\leq M}(\gamma) = (1, S(\gamma)^1, \dots, S(\gamma)^M) \in \bigoplus_{k=0}^M (\mathbb{R}^d)^{\otimes k}$$

Remark 7.3 (Kernel Trick for Signatures). *The signature kernel $k(\gamma, \gamma') = \langle S(\gamma), S(\gamma') \rangle$ can be computed without explicitly constructing signatures, via a Goursat PDE [17]. This enables signature methods to scale to the trajectory lengths typical of longitudinal medical records.*

7.3 Attention as Learned Signature Projection

Conjecture 7.4 (Attention-Signature Correspondence). *Let $\gamma = (e_1, \dots, e_n)$ be a trajectory with embeddings (ψ_1, \dots, ψ_n) . The attention mechanism computes:*

$$h_\gamma = \sum_i \alpha_i \psi_i + \sum_{i < j} \beta_{ij} (\psi_i \otimes \psi_j) + \dots$$

where the weights $\alpha_i, \beta_{ij}, \dots$ are attention-derived. This is a weighted projection of the signature onto a learned subspace [12].

Supporting evidence: the level-1 signature $S^1 = \sum_i \Delta\psi_i$ is captured by first/last pooling; the level-2 signature $S^2_{ij} = \int \int_{s < t} d\psi_s^i d\psi_t^j$ captures signed area related to inter-position attention; and multi-head attention with H heads can represent H independent signature projections.

Remark 7.5 (Time2Vec as Signature Augmentation). *Appending Time2Vec features $\tau(t_i)$ to event embeddings $\phi(e_i)$ before computing signatures incorporates temporal information into the path [11]. The frequencies ν_k select which temporal scales contribute to the signature. In rough path terms, the “rough path lift” of the trajectory includes both event content and timing; Time2Vec provides a specific parameterization of the temporal component.*

7.4 Log-Signatures and Efficient Representation

Definition 7.6 (Log-Signature [15]). *The log-signature is defined via the Baker-Campbell-Hausdorff formula:*

$$\text{LogSig}(\gamma) = \log(S(\gamma))$$

where the logarithm is in the tensor algebra (using the BCH series). The log-signature lies in the free Lie algebra, a much smaller space than the full tensor algebra.

For a path in \mathbb{R}^d , the level- M log-signature has dimension $\sim d^M/M$ (versus d^M for the truncated signature). This compression identifies the free Lie algebra as the natural target for efficient trajectory representation, with attention implicitly constrained to this subspace by the combinatorial structure of its weight-sharing scheme.

8 Synthesis: The Mathematical Framework

8.1 Three-Level Architecture

The architecture admits characterization through three mathematical lenses:

1. **Tropical/Viscosity Level:** The network computes a viscosity solution to an HJ equation on trajectory space [5, 16]. The decision boundary is a tropical hypersurface. ReLU nonlinearity implements the max-plus algebra of classical mechanics.
2. **Signature Level:** The attention mechanism extracts signature features: coordinate-free path invariants that capture trajectory shape [8, 15]. Time2Vec frequencies parameterize the temporal component of the rough path lift.
3. **Boundary Value Level:** Mortality acts as an absorbing boundary condition. Training propagates gradients backward from the boundary, shaping how interior trajectories are represented. The decision hyperplane in representation space is the image of $\partial\Omega$ under the learned embedding.

8.2 Unified Action Functional

Combining these perspectives, the action minimized during training is:

$$S[\gamma; \theta] = \underbrace{-\log P_\theta(\gamma)}_{\text{signature likelihood}} + \lambda \cdot \underbrace{d_{\text{trop}}(\gamma, \partial\Omega)}_{\text{tropical distance to boundary}}$$

The operational correspondence is:

signature likelihood \longleftrightarrow cross-entropy on trajectory representation
 tropical distance \longleftrightarrow logistic loss measuring boundary distance
 manifold projection \longleftrightarrow implicit representation regularization

8.3 Parameter Efficiency from Structural Alignment

Three structural factors explain the architecture’s parameter efficiency:

1. **Tropical efficiency [20]:** Piecewise-linear decision boundaries require $O(k)$ parameters to specify k linear regions. Smooth approximations require $O(k^2)$ or worse for equivalent expressivity at discontinuities.
2. **Signature compression [8]:** By targeting signature-like features (coordinate-free path invariants), attention avoids redundant representations. The free Lie algebra is exponentially smaller than the full tensor algebra.
3. **Boundary factorization:** Binary mortality classification factors the problem: learn a single mortality direction w rather than modeling the full outcome distribution. All information flows through the one-dimensional boundary distance [7].

References

- [1] Bombelli, L., Lee, J., Meyer, D., and Sorkin, R. D. (1987). Space-time as a causal set. *Physical Review Letters*, 59(5):521–524.
- [2] Cairns, A. J. G., Blake, D., and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *Journal of Risk and Insurance*, 73(4):687–718.
- [3] Chen, K.-T. (1958). Integration of paths: a faithful representation of paths by noncommutative formal power series. *Transactions of the American Mathematical Society*, 89(2):395–407.
- [4] Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34(2):187–220.
- [5] Crandall, M. G. and Lions, P.-L. (1983). Viscosity solutions of Hamilton-Jacobi equations. *Transactions of the American Mathematical Society*, 277(1):1–42.
- [6] Crandall, M. G., Ishii, H., and Lions, P.-L. (1992). User’s guide to viscosity solutions of second order partial differential equations. *Bulletin of the American Mathematical Society*, 27(1):1–67.
- [7] Fleming, W. H. and Soner, H. M. (2006). *Controlled Markov Processes and Viscosity Solutions*, 2nd edition. Springer, New York.

- [8] Hambly, B. and Lyons, T. (2010). Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics*, 171(1):109–167.
- [9] Hoel, E. P., Albantakis, L., and Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795.
- [10] Jensen, R. (1988). The maximum principle for viscosity solutions of fully nonlinear second order partial differential equations. *Archive for Rational Mechanics and Analysis*, 101(1):1–27.
- [11] Kazemi, S. M., Goel, R., Jain, S., Kobzyev, I., Sethi, A., Forsyth, P., and Poupart, P. (2019). Time2Vec: Learning a general purpose representation of time. *arXiv preprint*, arXiv:1907.05321.
- [12] Kidger, P. and Lyons, T. (2021). Neural controlled differential equations for irregular time series. *Advances in Neural Information Processing Systems*, 33.
- [13] Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87(419):659–671.
- [14] Litvinov, G. L. (2007). The Maslov dequantization, idempotent and tropical mathematics: a brief introduction. *Journal of Mathematical Sciences*, 140(3):426–444.
- [15] Lyons, T. J. (1998). Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2):215–310.
- [16] Maclagan, D. and Sturmfels, B. (2015). *Introduction to Tropical Geometry*. Graduate Studies in Mathematics, Vol. 161. American Mathematical Society, Providence, RI.
- [17] Salvi, C., Cass, T., Foster, J., Lyons, T., and Yang, W. (2021). The signature kernel is the solution of a Goursat PDE. *SIAM Journal on Mathematics of Data Science*, 3(3):873–899.
- [18] Sorkin, R. D. (2003). Causal sets: discrete gravity. In *Lectures on Quantum Gravity*, pages 305–327. Springer, New York.
- [19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- [20] Zhang, L., Naitzat, G., and Lim, L.-H. (2018). Tropical geometry of deep neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5824–5832. PMLR.