Mathematical Foundations of DYNAMO: Dynamic Causal-Temporal Adaptation

Karthik Srikumar

June 1, 2025

1 Problem Formulation

Given a language model $\mathcal{M}$ with parameters Θ, trained on static dataset $\mathcal{D}_{\text{static}} = \{(x_i, y_i)\}_{i=1}^{N}$, we address two limitations:

1. **Temporal Bias**: Knowledge cutoff at time T_{train} 2. **Causal Fragility**: Learned causal relations $\mathcal{C}$ become invalid when $P(Y|X)$ changes over time

Define the dynamic world state at time t:

$$\mathcal{W}_t = (\mathcal{K}_t, \mathcal{C}_t)$$

where $\mathcal{K}_t$ is factual knowledge and $\mathcal{C}_t = \{(X \to Y|\tau)\}$ are time-dependent causal relations.

2 Core Mathematical Framework

2.1 Temporal Embeddings

We extend Time2Vec (Kazemi et al., 2019) to d-dimensional temporal embeddings:

$$\phi(t)[k] = \begin{cases} \omega_k t + \varphi_k & k = 0 \\ \sin(\omega_k t + \varphi_k) & 1 \leq k \leq \lfloor d/2 \rfloor \\ \cos(\omega_k t + \varphi_k) & \lfloor d/2 \rfloor < k \leq d \end{cases}$$

2.2 Causal Graph Representation

Causal knowledge at time t is represented as a directed graph:

$$\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t), \quad \mathcal{E}_t \subseteq \mathcal{V} \times \mathcal{V} \times \mathbb{R}^+$$

with edge weights $w_{ij}^t = P(X_j|\text{do}(X_i))$ estimated from temporal data.

2.3 Adapter Architecture

For each transformer layer ℓ with hidden states $\mathbf{H}^\ell$, we apply:

$$\mathbf{H}^\ell_{\text{out}} = \mathbf{H}^\ell + \Delta\mathbf{H}^\ell(t, \mathcal{G}_t)$$

where the adapter function is:

$$\Delta\mathbf{H}^\ell = \mathbf{W}_o \cdot \sigma\Big(\mathbf{W}_t\phi(t) + \mathbf{W}_g f_g(\mathcal{G}_t) + \mathbf{b}\Big)$$

with:

$$f_g(\mathcal{G}_t) = \text{GNN}(\mathbf{A}_t)$$
$$\mathbf{A}_t[i,j] = w^t_{ij} \cdot \mathbb{I}[t \in \tau_{ij}]$$

3 Key Algorithms

3.1 Temporal Adapter Training

Algorithm 1 Causal-Temporal Adapter Optimization

1: **Input:** Base model $\mathcal{M}_\Theta$, Stream $\mathcal{S} = \{(x_i, y_i, t_i, \mathcal{G}_{t_i})\}_{i=1}^M$
2: Initialize adapter params $\Psi = \{\mathbf{W}_t, \mathbf{W}_g, \mathbf{W}_o\}$
3: **while** not converged **do**
4: Sample batch $\mathcal{B} \sim \mathcal{S}$
5: Compute loss:

$$\mathcal{L} = \underbrace{\text{CE}(\mathcal{M}_{\Theta,\Psi}(x, t, \mathcal{G}_t), y)}_{\text{factual}} + \lambda \underbrace{D_{\text{KL}}(p_\Psi \| p_{\mathcal{G}_t})}_{\text{causal reg.}}$$

6: Update $\Psi \leftarrow \Psi - \eta\nabla_\Psi\mathcal{L}$
7: **end while**

3.2 Causal Graph Update

4 Information-Theoretic Foundations

4.1 Temporal Generalization Bound

For time-shifted distributions $\mathcal{P}_t$, $\mathcal{P}_{t+\Delta}$, the generalization error is bounded by:

$$\epsilon_{\text{gen}} \leq \underbrace{\frac{1}{2}d_{\mathcal{H}}(\mathcal{P}_t, \mathcal{P}_{t+\Delta})}_{\text{temporal shift}} + \underbrace{\sqrt{\frac{\log(1/\delta)}{2m}}}_{\text{sample size}}$$

Our adapters minimize the divergence term:

$$\min_\Psi \mathbb{E}_t[d_{\mathcal{H}}(\mathcal{P}_t, \mathcal{P}_{t+\Delta|\Psi})]$$

Algorithm 2 Dynamic Causal Graph Maintenance

procedure UPDATEGRAPH($\mathcal{G}_t, \mathcal{D}_{\text{new}}$)
  for (X, Y) in causal pairs **do**
    Compute temporal mutual information:

$$\hat{w}_{XY}^{t+\Delta} = I_t(X;Y) - \alpha \frac{\partial I}{\partial t}$$

    Update edge:

$$w_{XY}^{t+\Delta} = (1-\beta) w_{XY}^{t} + \beta \hat{w}_{XY}^{t+\Delta}$$

    Update validity window:

$$\tau_{XY} \leftarrow \tau_{XY} \cup [t, t+\Delta]$$

  end for
end procedure

4.2 Causal Invariance Learning

We enforce invariant prediction via:

$$P_\Psi(Y|\text{do}(X), t_1) = P_\Psi(Y|\text{do}(X), t_2) \quad \forall t_1, t_2$$

achieved through the regularization:

$$R(\Psi) = \sum_{t_i \neq t_j} D_{\text{JS}}(P_\Psi^{(t_i)}(Y|do(X)) \| P_\Psi^{(t_j)}(Y|do(X)))$$

5 Experimental Framework

5.1 Metrics

$$\text{Temporal Accuracy} = \frac{1}{|\mathcal{Q}|} \sum_{(q,t)} \mathbb{I}[\text{Correct}(q,t)]$$

$$\text{Causal F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{where Precision} = \frac{|\mathcal{C}_{\text{pred}} \cap \mathcal{C}_{\text{true}}|}{|\mathcal{C}_{\text{pred}}|}$$

$$\text{Recall} = \frac{|\mathcal{C}_{\text{pred}} \cap \mathcal{C}_{\text{true}}|}{|\mathcal{C}_{\text{true}}|}$$

5.2 Efficiency Analysis

Adapter parameter efficiency:

$$\rho = \frac{\|\Psi\|_0}{\|\Theta\|_0} \approx 10^{-3}$$

Training complexity:

$$\mathcal{O}(|\Psi| \cdot B \cdot T) \ll \mathcal{O}(|\Theta| \cdot N)$$

where B = batch size, T = iterations, N = full dataset size.

6 Theoretical Guarantees

6.1 Stability Theorem

Theorem 1: Under Lipschitz continuity (L-smoothness) of adapter parameters, the model's output drift is bounded:

$$\|\mathcal{M}(x,t) - \mathcal{M}(x,t+\Delta)\|_2 \leq L \cdot \kappa \cdot |\Delta| \cdot \|\phi'(t)\|_2$$

where κ is the condition number of $\mathbf{W}_o$.

Proof. By Grönwall's inequality applied to the dynamical system: $\frac{\partial \mathbf{H}}{\partial t} = \mathbf{J}_\Psi \cdot \frac{\partial \phi}{\partial t}$ □

6.2 Causal Consistency

Lemma: The causal regularization term ensures:

$$\lim_{\lambda \to \infty} D_{\mathrm{KL}}(p_\Psi \| p_{\mathcal{G}_t}) = 0$$

guaranteeing alignment with ground-truth causal structures.

7 Conclusion

DYNAMO provides a mathematically rigorous framework for temporal and causal adaptation:

- Time2Vec embeddings capture continuous temporal dynamics
- GNN-based causal graph integration maintains structural relationships
- Information-theoretic regularization ensures stability
- Parameter-efficient design enables real-time deployment