Implicit Flow Attention on Cyclic Graphs: Convergent Conservation-Normalized Message Passing with Stability Bounds

Liz Lemma Future Detective

January 18, 2026

Abstract

Flow-attentional GNNs modify standard graph attention by normalizing attention scores across outgoing neighbors, aligning message passing with conservation laws in resource-flow graphs. Prior work established expressivity gains and strong empirical performance for flow attention on undirected flow graphs and on DAGs via FlowDAGNN, but many real infrastructure networks (power transmission, traffic, water) contain directed cycles that break sequential DAG processing. We introduce an implicit (fixed-point) flow-attentive layer for general directed graphs. The layer defines outgoing-normalized attention weights as a routing kernel and computes node representations as the unique fixed point of a damped conservation-aware message-passing operator. Under explicit norm/temperature conditions we prove (i) existence and uniqueness of the fixed point, (ii) linear convergence of simple iterations, and (iii) a perturbation stability bound with respect to node/edge feature changes. We also show that without contractivity such fixed-point computation is PPAD-hard, motivating the restricted design. Implementations and experiments on cyclic flow benchmarks (e.g., AC-like power grid variants, traffic loops) would validate the practical benefits and quantify the trade-off between expressivity and guaranteed convergence.

Table of Contents

3. 3. Problem Formulation: learning on cyclic flow graphs; conservative routing kernels; desiderata (permutation invariance, conservation-consistent weights, convergence/stability).

4. 4. Implicit Flow-Attentive Layer: define scoring, outgoing normalization, damped operator F_θ; handling sources/sinks and optional injection terms; implementation notes (unrolling vs implicit differentiation).

5. 5. Convergence Theory: explicit Lipschitz upper bounds for Φ_θ and F_θ; contraction conditions; linear convergence rate; iteration complexity to reach ε-accuracy.

6. 6. Stability Theory: perturbation bounds for fixed points under feature/edge perturbations; sensitivity to graph rewiring approximations; interpretation as robustness certificate.

7. 7. Lower Bounds / Necessity: PPAD-hardness of general fixed points; discussion of why contraction (or monotone operator structure) is needed for provable guarantees; minimality of reading all edges.

8. 8. Extensions: (i) capacity-aware normalizations, (ii) multiple sweeps / block-implicit variants, (iii) hybrid global-local architectures, (iv) node/edge-level tasks.

9. 9. Experimental Plan (flagged as strengthening): cyclic power/traffic benchmarks, synthetic controlled cycles, compute-vs-accuracy trade-offs, ablations on temperature/damping/iterations, comparison to unrolled recurrent GNNs and transformer hybrids.

10. 10. Discussion and Limitations: expressivity vs contraction; conditions being sufficient not necessary; failure modes; deployment considerations for real-time infrastructure.

1 Introduction

Many graph learning problems tacitly conflate two distinct semantics: *informational graphs*, in which an edge merely indicates that a feature may be used to predict another feature, and *flow graphs*, in which edges represent a directed routing of conserved mass, probability, traffic, or influence. In an informational graph, the primary requirement is typically permutation invariance and locality of computation; the update rule may freely amplify or attenuate signals as it aggregates neighbors. In a flow graph, by contrast, edges encode a constraint—often implicit rather than explicitly supervised—that outgoing influence from a node should be distributed among its outgoing edges. This distinction becomes operational when one seeks embeddings that are interpretable as steady-state quantities (e.g. equilibrium occupancies, conserved routing policies, or stabilized influence scores) rather than merely transient outputs of a fixed-depth computation.

Cycles are the rule rather than the exception in flow graphs. Feedback appears in transportation and logistics networks (re-routing and congestion), in citation and information propagation graphs (mutual reinforcement), in economic input–output networks, and in dynamical interaction graphs in biology and control. Even when the original data are acyclic, modeling choices frequently introduce effective cycles: bidirectional edges for undirected relations, reverse edges for information backflow, or residual pathways across layers. Consequently, approaches restricted to directed acyclic graphs (DAGs)—or, more generally, approaches whose correctness or stability relies on a topological ordering—are structurally mismatched to the phenomena we wish to represent. A DAG-only method can compute a well-defined forward pass, but it cannot naturally encode equilibria determined by mutual dependence, since those equilibria are fixed points of a coupled system rather than values produced by a one-way recursion.

Standard message passing neural networks and attention-based graph transformers accommodate cycles syntactically, yet their computation is typically *depth-limited*: information propagates for a prescribed number of layers, and the output depends on that truncation. This design choice is often computationally convenient, but it entangles representation with an arbitrary iteration budget and makes it difficult to interpret the result as a stable graph-derived quantity. Moreover, unnormalized attention can behave unlike any conservative routing rule: a node may simultaneously assign large weights to many outgoing edges, yielding amplification effects that are at odds with flow semantics and that complicate stability analyses. In applications where one desires a routing kernel—a probability distribution over outgoing neighbors for each node—it is natural to impose outgoing normalization, so that each node distributes a fixed unit of influence across its outgoing edges. This simple architectural constraint aligns the model with flow conservation and yields an attention mechanism that is interpretable as

local routing.

Once we adopt a conservative, outgoing-normalized attention, the remaining issue is how to reconcile cycles with well-defined computation. If the attention weights depend on hidden states, then in a cyclic graph the hidden states and the induced routing kernel are mutually dependent. In such settings, the appropriate object is not a depth-T representation but an *equilibrium* representation: a hidden state assignment consistent with the update rule everywhere on the graph. This perspective leads naturally to implicit graph layers defined by fixed-point equations. Rather than unrolling a deep stack with potentially unstable dynamics, we compute (or approximate) the fixed point of a single contractive operator. The resulting representation is independent of an arbitrary depth parameter and is, by construction, compatible with cyclic feedback.

The present work develops a flow-attentive implicit layer whose attention is normalized over outgoing neighborhoods, thereby enforcing exact conservation at every evaluation of the attention kernel. The update is defined by a damped operator that interpolates between the current iterate and an undamped message passing map; the damping plays the same conceptual role as a step size in a dynamical system. Our goal is not merely to propose this architecture, but to provide explicit conditions under which the resulting implicit layer is well posed and efficiently solvable on directed graphs with cycles.

Our technical contributions are organized around three desiderata: (i) *existence and uniqueness* of the equilibrium embedding, (ii) *efficient computation* to a prescribed accuracy, and (iii) *stability* with respect to perturbations of graph data and parameters. To this end we work with a block maximum norm and track Lipschitz constants of the constituent maps. The scoring function of attention is assumed to be Lipschitz in the hidden states, and the message transform is controlled via an explicit spectral norm bound. The attention temperature and maximum out-degree enter the analysis through the sensitivity of the outgoing softmax normalization. These ingredients yield a concrete Lipschitz bound for the undamped map, hence a contraction criterion for the damped operator.

At a high level, the guarantees we establish are as follows.

- *Contractivity and unique fixed point.* Under explicit bounds on the message transform, attention sensitivity, temperature, and graph out-degree, the damped operator becomes a contraction. By the Banach fixed-point theorem, it admits a unique fixed point, which we interpret as the equilibrium node embedding induced by flow-attentive routing on a cyclic directed graph.

- *Linear-time-per-iteration computation and logarithmic iteration complexity.* The equilibrium can be approximated by Picard iteration

(damped fixed-point iteration). Each iteration reduces to sparse edge-local computations—scoring edges, normalizing per sender, and aggregating incoming messages—and therefore costs time linear in the number of edges up to the dimensionality factors of the chosen scoring mechanism. Under contraction, the number of iterations required to reach accuracy ε grows only logarithmically in ε^{-1}.

- *Exact conservation of the attention kernel.* Because normalization is performed over each node's outgoing neighborhood, the induced attention weights form a row-stochastic routing kernel wherever out-degree is nonzero. This property holds at every iterate and therefore at the fixed point, yielding a principled notion of conserved flow in the learned representation.

- *Perturbation stability.* When the operator is contractive, the fixed point depends Lipschitz-continuously on perturbations of node features, edge features, and model parameters. This yields a clean bound relating representation drift to data drift, with a degradation factor determined by the contraction modulus. Such a statement is unavailable for general implicit layers without structural control.

- *Necessity of restrictions.* Finally, we justify why one should not expect unconditional convergence guarantees for unrestricted implicit graph layers: without contraction (or comparable monotonicity structure), approximating fixed points of continuous graph-defined operators is intractable in the worst case (PPAD-hard), via standard connections to Brouwer fixed points.

Conceptually, our analysis treats the implicit layer as a controlled dynamical system on a high-dimensional product space indexed by nodes. The outgoing-normalized attention acts as a learned, state-dependent routing rule, and the damping enforces a quantitative form of stability. The resulting model therefore sits between two classical perspectives: on the one hand, attention-based message passing as a flexible statistical learner; on the other, contractive fixed-point iteration as a principled computational mechanism for equilibria on cyclic networks. In the sequel we formalize the graph-theoretic notation, clarify the relationship between standard attention and flow-attention, and introduce the norms and Lipschitz estimates that make the above guarantees precise.

2 Preliminaries

Directed graphs and neighborhoods. We work with a finite directed graph $G = (V, E)$ with $|V| = n$ and $|E| = m$, allowing cycles and self-loops.

For a node $i \in V$ we denote its incoming and outgoing neighborhoods by

$$N_{\text{in}}(i) := \{j \in V : (j,i) \in E\}, \qquad N_{\text{out}}(i) := \{k \in V : (i,k) \in E\}.$$

We write $\Delta_{\max} := \max_{j \in V} |N_{\text{out}}(j)|$ for the maximum out-degree. Node features are $x_i \in \mathbb{R}^{d_x}$, and each directed edge $(i,j) \in E$ may carry features $a_{ij} \in \mathbb{R}^{d_e}$. When convenient, we regard node-wise hidden states as a block vector $h = (h_i)_{i \in V} \in \mathbb{R}^{n \times d}$.

The algorithmic model we have in mind is sparse adjacency access: at each iteration we may enumerate $N_{\text{out}}(j)$ for each sender j and $N_{\text{in}}(i)$ for each receiver i, and we may compute edge-local quantities for each $(j,i) \in E$. Nodes with $N_{\text{out}}(j) = \varnothing$ require a convention for normalization; throughout we interpret the outgoing-normalized attention weights only on nodes with nonzero out-degree (equivalently, we may add self-loops to eliminate sinks, or stipulate that such nodes emit no messages). This choice affects implementation but not the principal contractivity mechanism, which is driven by explicit Lipschitz bounds.

Flow conservation and routing kernels. A central object in our development is a *conservative* attention kernel that assigns, for each sender j with $|N_{\text{out}}(j)| > 0$, a probability distribution over its outgoing edges. Concretely, an array $\beta(h) = (\beta_{ij}(h))_{(j,i) \in E}$ is called outgoing-normalized if

$$\sum_{i \in N_{\text{out}}(j)} \beta_{ij}(h) = 1 \qquad \text{for all } j \text{ with } |N_{\text{out}}(j)| > 0,$$

and $\beta_{ij}(h) \geq 0$ for all edges. Under this condition, $\beta(h)$ acts as a row-stochastic routing rule indexed by senders: one may interpret node j as distributing a unit of mass (or influence) across its outgoing edges according to the weights $\beta_{ij}(h)$. This is the formal sense in which the attention mechanism respects flow conservation at the level of the learned routing kernel.

We emphasize that outgoing normalization is distinct from the normalization commonly used in standard attention-based message passing, in which weights are normalized over a receiver's incoming neighborhood. Incoming normalization ensures $\sum_{j \in N_{\text{in}}(i)} \alpha_{ij} = 1$ for fixed receiver i, which is appropriate when attention is viewed as a convex combination of candidate *inputs* to a node. Outgoing normalization instead enforces conservation at the *sender*, aligning the weights with flow semantics on directed graphs.

Standard attention versus flow-attention. Let e_θ be an attention scoring function. Given hidden states h and edge features, we form edge scores for each directed edge $(j,i) \in E$ by

$$s_{ij}(h) := \frac{1}{\tau}\, e_\theta\big(h_i, h_j, a_{ji}\big),$$

where $\tau > 0$ is a temperature parameter. (Larger τ produces smoother, less sensitive softmax weights.) Standard graph attention typically computes receiver-normalized coefficients

$$\alpha_{ij}(h) = \frac{\exp(s_{ij}(h))}{\sum_{\ell \in N_{\text{in}}(i)} \exp(s_{i\ell}(h))}, \qquad j \in N_{\text{in}}(i),$$

so that the aggregation at i is a convex mixture of transformed neighbor states. In contrast, our *flow-attention* coefficients are outgoing-normalized:

$$\beta_{ij}(h) = \frac{\exp(s_{ij}(h))}{\sum_{k \in N_{\text{out}}(j)} \exp(s_{kj}(h))}, \qquad i \in N_{\text{out}}(j).$$

Thus, for fixed sender j, the normalization is taken over the scores associated with edges leaving j. This is the minimal structural modification needed to ensure that each node distributes a fixed unit of influence across its outgoing edges. In particular, if the graph represents a transportation or routing system, $\beta_{ij}(h)$ can be interpreted as the learned policy assigning probability of routing from j to i.

The dependence of $\beta_{ij}(h)$ on *both* h_j and the receiver states $\{h_k : k \in N_{\text{out}}(j)\}$ is essential: the denominator couples all outgoing edges of a sender. On cyclic graphs, this coupling propagates globally through the fixed-point equation and is the reason we adopt an implicit (equilibrium) viewpoint rather than a finite-depth recursion.

State space, norms, and operator bounds. Our convergence and stability guarantees are stated using the block maximum norm

$$\|h\|_{\infty,2} := \max_{i \in V} \|h_i\|_2, \qquad h \in \mathbb{R}^{n \times d}.$$

This choice separates node-wise effects from graph size and allows degree-dependent factors to appear explicitly. For linear maps acting on a single node state we use the spectral norm $\|\cdot\|_2$; for example, $\|W_m h_j\|_2 \le \|W_m\|_2 \, \|h_j\|_2$.

For a map $F : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d}$ we write $\mathrm{Lip}(F)$ for the smallest L such that

$$\|F(h) - F(\tilde{h})\|_{\infty,2} \le L \, \|h - \tilde{h}\|_{\infty,2} \qquad \text{for all } h, \tilde{h}.$$

We repeatedly use standard composition rules: if g and f are Lipschitz, then $\mathrm{Lip}(g \circ f) \le \mathrm{Lip}(g)\,\mathrm{Lip}(f)$, and if $F = (1-\alpha)I + \alpha G$, then $\mathrm{Lip}(F) \le (1-\alpha) + \alpha\,\mathrm{Lip}(G)$.

The activation $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ is assumed to be applied coordinatewise, bounded, and 1-Lipschitz with respect to $\|\cdot\|_2$. Boundedness implies that iterates remain uniformly bounded once the update has the form $h \leftarrow \sigma(\cdot)$, while 1-Lipschitzness ensures that σ does not inflate perturbations introduced by attention or message aggregation.

Lipschitz assumptions for attention and data dependence. We assume that the scoring function e_θ is Lipschitz in its hidden-state arguments uniformly over edge features: there exists $L_e \geq 0$ such that for all $u, u', v, v' \in \mathbb{R}^d$ and all admissible a,

$$\big|e_\theta(u, v, a) - e_\theta(u', v', a)\big| \leq L_e\big(\|u - u'\|_2 + \|v - v'\|_2\big).$$

This condition is satisfied, for example, by common dot-product or bilinear scorers under spectral norm control, and by multilayer perceptrons with bounded operator norms.

The softmax normalization contributes an additional sensitivity factor. At a fixed sender j, the map from the score vector $(s_{ij})_{i \in N_{\text{out}}(j)}$ to the probability vector $(\beta_{ij})_{i \in N_{\text{out}}(j)}$ is Lipschitz, with modulus scaling like $1/\tau$ and depending on the neighborhood size $|N_{\text{out}}(j)|$ under the norms used to compare score and probability vectors. In our later bounds this dependence is summarized by an explicit factor $C(\Delta_{\max}) = O(\Delta_{\max})$ under $\|\cdot\|_{\infty,2}$, reflecting that a single-node perturbation can influence multiple outgoing weights through the shared normalizer.

Finally, we separate two kinds of Lipschitz control: (i) dependence on hidden states, quantified by L_e and the ensuing L_Φ for the undamped message passing map, and (ii) dependence on *data* (x, a), quantified by a data-Lipschitz constant L_{data} used in perturbation arguments. The latter captures how changes in features alter the operator even when the hidden state is held fixed.

Implicit layer maps and Picard iteration. The undamped flow-attentive update map Φ_θ is defined nodewise by

$$\Phi_\theta(h)_i := \sigma\Big(W_x x_i + \sum_{j \in N_{\text{in}}(i)} \beta_{ij}(h)\, W_m h_j + b\Big),$$

and we introduce damping via

$$F_\theta(h) := (1 - \alpha)h + \alpha\, \Phi_\theta(h), \qquad \alpha \in (0, 1].$$

We approximate equilibria by Picard iteration $h^{t+1} = F_\theta(h^t)$. When Φ_θ is Lipschitz with constant L_Φ and $\alpha L_\Phi < 1$, the map F_θ is a contraction, ensuring both existence and uniqueness of a fixed point $h^* = F_\theta(h^*)$ and linear convergence of the iterates. The remainder of our analysis is devoted to making L_Φ explicit in terms of $\|W_m\|_2$, L_e, τ, and $\Delta_{\max}$, and to translating contractivity into computational and stability guarantees.

3 Problem Formulation

Learning on directed flow graphs with cycles. We consider supervised and self-supervised learning problems in which the input is a directed

graph $G = (V, E)$ whose edges encode admissible directions of transport, influence, or routing. The key regime of interest is *cyclic* graphs: feedback loops and recirculation are not treated as exceptional, but rather as the typical case (e.g. road networks with roundabouts, supply chains with return flows, or recurrent interaction graphs). In such settings, a finite-depth message passing stack is often best understood as an *approximate solver* of an underlying equilibrium relation. We therefore frame representation learning in terms of node states $h = (h_i)_{i \in V} \in \mathbb{R}^{n \times d}$ that are intended to satisfy a fixed-point condition induced by a flow-aware attention mechanism, and we train parameters so that the resulting equilibrium embeddings are predictive for the downstream task.

Concretely, the data consist of node features $x_i \in \mathbb{R}^{d_x}$ and edge features $a_{ij} \in \mathbb{R}^{d_e}$, and optionally additional markings such as a set of sources $S \subseteq V$ and sinks $T \subseteq V$ (or, more generally, any side information specifying boundary conditions). Labels may be provided at the node, edge, or graph level. Our output is an embedding h^* (nodewise) together with task predictions $\hat{y}$ obtained by applying an appropriate head to h^* (and, for graph-level tasks, a readout that aggregates $\{h_i^*\}_{i \in V}$).

Conservative routing kernels as structural attention. A distinguishing requirement of our setting is that attention weights should be consistent with *conservation at senders*. Rather than treating attention as a receiver-side convex combination, we interpret attention as defining, for each sender j, a distribution over its outgoing edges. Thus the attention array $\beta(h) = (\beta_{ij}(h))_{(j,i) \in E}$ is constrained to be outgoing-normalized: for every j with $|N_{\text{out}}(j)| > 0$,

$$\sum_{i \in N_{\text{out}}(j)} \beta_{ij}(h) = 1, \qquad \beta_{ij}(h) \geq 0.$$

This constraint turns $\beta(h)$ into a row-stochastic routing kernel indexed by senders. In flow-centric applications, $\beta_{ij}(h)$ is naturally interpreted as the probability (or fraction of a unit of influence) that node j sends along edge (j, i), conditional on the current state h. The dependence on h is essential: the routing policy may adapt to congestion-like signals encoded in hidden states and to local edge features a_{ji}, while still respecting conservation.

The conservative constraint is not merely semantic. It enforces a form of *mass preservation* at the level of learned routing and thereby reduces degrees of freedom that can otherwise destabilize dynamics on cyclic graphs. Moreover, because normalization couples the outgoing neighborhood of each sender, conservation introduces structured, local competition among edges leaving the same node, which is the appropriate analogue of capacity allocation in many directed systems.

Equilibrium embeddings and task prediction. Given a parameterization of the flow-attentive update map, our representation is defined implicitly as an equilibrium h^* satisfying a fixed-point equation

$$h^* = F_\theta(h^*),$$

where F_θ is the (possibly damped) flow-attentive operator specified in the sequel. From the perspective of learning, this shifts emphasis from layer depth to *solution accuracy*: we may compute an approximate equilibrium h^T by an iterative method until a fixed-point residual criterion is met, and we interpret $h^T \approx h^*$ as the embedding used by the downstream head.

Let g_ψ denote a task head with parameters ψ. For node-level prediction we set $\hat{y}_i = g_\psi(h_i^*)$; for edge-level prediction $\hat{y}_{ij} = g_\psi(h_i^*, h_j^*, a_{ij})$; and for graph-level prediction we form a permutation-invariant readout, for example

$$\bar{h}^* = \mathrm{READOUT}\big(\{h_i^* : i \in V\}\big), \qquad \hat{y} = g_\psi(\bar{h}^*),$$

where READOUT may be a sum/mean pooling or a learned invariant aggregator. The learning objective is to minimize a supervised loss $\mathcal{L}(\hat{y}, y)$, potentially augmented by regularizers encoding contractive design constraints. Writing (x, a, y) for a training instance, the canonical objective takes the form

$$\min_{\theta,\psi} \ \mathbb{E}\big[\mathcal{L}(g_\psi(h^*), y)\big] \quad \text{subject to} \quad h^* = F_\theta(h^*),$$

with the understanding that h^* is computed (approximately) during the forward pass.

Desiderata: symmetry, conservation, and well-posedness. We now state the properties we require of the encoder defined implicitly by F_θ.

(D1) Permutation equivariance and invariance. Since the node identities are arbitrary labels, the embedding map must be consistent with relabeling. Formally, for any permutation π of V acting on node-indexed tensors in the natural way and transporting edges accordingly, the node embedding should be permutation equivariant:

$$h^*(\pi \cdot x, \pi \cdot a) = \pi \cdot h^*(x, a).$$

Graph-level predictions obtained after a readout should be invariant. This desideratum is satisfied when all computations in F_θ are expressed via neighborhood aggregations and shared parameters, and when the attention normalization is performed per sender using only its outgoing neighborhood.

(D2) Conservation-consistent weights. At every iterate (and hence at the fixed point), attention weights must define a valid routing distribution per sender. In particular, the normalization must be outgoing rather than incoming, and the nonnegativity and unit-sum conditions must hold exactly

(up to floating point effects). This is both an inductive bias and a hard architectural constraint: we do not rely on training to "learn" conservation.

(D3) Convergence and uniqueness on cyclic graphs. Because cycles induce feedback, naive recurrent message passing may admit multiple equilibria or fail to converge. For representation learning this is problematic: the embedding becomes initialization-dependent, gradients become ill-defined, and the forward pass may be unstable. We therefore seek explicit conditions under which the fixed point exists and is unique, and under which simple solvers converge at a controlled rate. Our goal is to enforce these conditions through explicit parameter restrictions (e.g. bounds on $\|W_m\|_2$, temperature τ, and damping α) so that the encoder is well-posed as a map from data to embeddings.

(D4) Stability to data perturbations. Graph learning frequently encounters noisy features, missing edges, and distribution shift. An implicit layer intended to model equilibria should be stable: small perturbations in (x, a) should yield proportionally small changes in h^*. This is also a prerequisite for meaningful generalization guarantees and for numerical robustness of the solver used to approximate h^*.

Computational viewpoint and approximation accuracy. In practice we do not compute h^* exactly, but rather an approximation h^T obtained by an iterative method with a stopping rule. We thus treat the fixed-point residual

$$r^t := \|h^{t+1} - h^t\|_{\infty,2}$$

as a proxy for solution quality and terminate once $r^t \leq \varepsilon$ for a user-specified tolerance $\varepsilon > 0$. This leads to a clear separation between (i) the *model* (the operator F_θ and its equilibrium) and (ii) the *solver* (the method used to approximate the equilibrium). Our subsequent analysis provides conditions ensuring that solver complexity scales predictably with ε and that the learned representations are not artifacts of nonconvergence.

Summary. We therefore formulate learning on cyclic flow graphs as the problem of fitting parameters of a permutation-equivariant, conservative, and contractively designed flow-attentive operator whose unique fixed point defines the embedding used for prediction. The next section specifies the operator in detail, discusses conventions for sources and sinks, and records implementation choices relevant for training (unrolled iteration versus implicit differentiation).

4 Implicit Flow-Attentive Layer

Local scores and outgoing-normalized flow attention. Fix hidden states $h = (h_i)_{i \in V} \in \mathbb{R}^{n \times d}$. For each directed edge $(j, i) \in E$ we compute a

sender-conditioned compatibility score

$$s_{ij}(h) \;:=\; \frac{1}{\tau}\, e_\theta\big(h_i, h_j, a_{ji}\big),$$

where $\tau > 0$ is a temperature parameter and e_θ is a shared scoring function applied locally to incident data. The flow-attention weights are then defined by an outgoing softmax over each sender neighborhood:

$$\beta_{ij}(h) \;:=\; \frac{\exp\big(s_{ij}(h)\big)}{\sum_{k \in N_{\mathrm{out}}(j)} \exp\big(s_{kj}(h)\big)} \qquad \text{for } (j,i) \in E \text{ and } |N_{\mathrm{out}}(j)| > 0.$$

We interpret $\beta_{ij}(h)$ as the fraction of a unit of influence routed from j to i under state h. The restriction to outgoing normalization is structural: competition occurs only among edges leaving the same sender, and conservation at the sender holds identically whenever $|N_{\mathrm{out}}(j)| > 0$. If $|N_{\mathrm{out}}(j)| = 0$, node j simply emits no messages (equivalently, the attention array has no entries indexed by j).

State-dependent message aggregation and update map. Given $\beta(h)$, we form incoming aggregated messages at each receiver node i by

$$m_i(h) \;:=\; \sum_{j \in N_{\mathrm{in}}(i)} \beta_{ij}(h)\, W_m h_j,$$

and define the undamped update map $\Phi_\theta : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ coordinatewise as

$$\Phi_\theta(h)_i \;:=\; \sigma\big(W_x x_i + m_i(h) + b\big).$$

We emphasize that the attention dependence on h couples all messages leaving a fixed sender j through the shared normalization constant

$$Z_j(h) \;:=\; \sum_{k \in N_{\mathrm{out}}(j)} \exp\big(s_{kj}(h)\big),$$

so that changing any receiver-side state h_k with $k \in N_{\mathrm{out}}(j)$ perturbs all weights $\{\beta_{ij}(h) : i \in N_{\mathrm{out}}(j)\}$ simultaneously. This local coupling is the mechanism by which routing decisions adapt to the current equilibrium state while remaining conservative.

Damping and the implicit equilibrium operator. On cyclic graphs, direct iteration of Φ_θ may fail to converge, so we introduce a damped operator

$$F_\theta(h) \;:=\; (1-\alpha)h + \alpha\, \Phi_\theta(h), \qquad \alpha \in (0,1].$$

Our embedding is defined implicitly as a fixed point h^* satisfying $h^* = F_\theta(h^*)$. In subsequent sections we will choose explicit restrictions (on $\|W_m\|_2$, on the temperature τ, and on α) ensuring that F_θ is contractive under $\|\cdot\|_{\infty,2}$, which yields existence and uniqueness of h^* and linear convergence of Picard iteration.

Sources, sinks, and optional injection terms. Many flow datasets specify marked source/sink sets (S, T). We incorporate such information without breaking permutation equivariance by allowing an additional *exogenous injection* term that depends only on local markings and features. Concretely, we permit

$$\Phi_\theta(h)_i \;:=\; \sigma\big(W_x x_i + m_i(h) + b + \iota_i\big),$$

where $\iota_i = \iota_\theta(x_i, \mathbf{1}_{i\in S}, \mathbf{1}_{i\in T}) \in \mathbb{R}^d$ is any bounded, shared parametrization (e.g. a learned embedding for membership in S or T added to the pre-activation). This captures steady supply/demand effects while preserving locality. When one wishes to model sinks as absorbing states, a purely graph-level convention is also available: for $j \in T$ we may remove outgoing edges (so $N_{\mathrm{out}}(j) = \emptyset$), or insert a self-loop (j, j) so that $N_{\mathrm{out}}(j) \neq \emptyset$ and $\beta_{jj}(h) = 1$ becomes feasible. Both conventions preserve the outgoing normalization definition and keep the encoder permutation equivariant.

A complementary mechanism is to impose *soft boundary conditions* by nodewise damping. Let $\alpha_i \in (0, 1]$ be a per-node step size determined from markings (e.g. $\alpha_i = \alpha_{\mathrm{src}}$ on S, $\alpha_i = \alpha_{\mathrm{int}}$ otherwise). Then we may update

$$F_\theta(h)_i \;=\; (1 - \alpha_i)h_i + \alpha_i\, \Phi_\theta(h)_i,$$

which allows sources/sinks to adapt faster or slower than interior nodes while remaining within the same fixed-point framework. The contraction analysis extends to this case by replacing α with $\max_i \alpha_i$ in the worst case.

Numerical solver: Picard iteration and stopping. To approximate h^* we use Picard iteration,

$$h^{t+1} = F_\theta(h^t), \qquad t = 0, 1, 2, \ldots,$$

initialized for example by $h^0 = 0$ or by a feature projection $h_i^0 = \sigma(W_0 x_i)$. We monitor the fixed-point residual

$$r^t \;:=\; \|h^{t+1} - h^t\|_{\infty,2} = \max_{i\in V} \|h_i^{t+1} - h_i^t\|_2,$$

and terminate once $r^t \leq \varepsilon$ for a prescribed tolerance $\varepsilon > 0$, or after a maximum number of iterations T. When F_θ is a contraction, r^t is a reliable proxy for the distance to the fixed point, and its decay is controlled by the contraction factor; the iteration complexity to achieve ε-accuracy will be stated after we develop explicit Lipschitz bounds.

Implementation remarks: sparse computation and stable normalization. Each iteration requires computing scores on edges, normalizing per sender, aggregating incoming messages, and applying the nodewise update. In sparse adjacency-list access models this is naturally implemented

by iterating over edges to compute s_{ij} and by accumulating senderwise log-normalizers Z_j. For numerical stability, we compute the softmax with the standard max-shift per sender:

$$\beta_{ij}(h) = \frac{\exp(s_{ij}(h) - c_j)}{\sum_{k \in N_{\text{out}}(j)} \exp(s_{kj}(h) - c_j)}, \qquad c_j := \max_{k \in N_{\text{out}}(j)} s_{kj}(h),$$

which preserves conservation exactly up to floating-point rounding. Memory may be reduced by streaming β_{ij}: one may recompute weights during aggregation rather than storing all per-edge values, at the cost of an additional pass over outgoing edges.

Training: unrolled differentiation versus implicit differentiation. There are two standard approaches for gradients through the equilibrium computation. In *unrolled training* we fix an iteration budget T (or an adaptive stopping rule) and backpropagate through the computation graph of h^T. This is simple and uses only automatic differentiation, but its memory scales with T unless checkpointing or recomputation is used.

In *implicit differentiation* we treat h^* as the solution to $h = F_\theta(h)$ and differentiate the fixed-point relation. Writing the training loss as $\mathcal{L}(h^*)$, one obtains a linear system for the adjoint variable v of the form

$$\big(I - J_F(h^*)^\top\big)v \;=\; \nabla_h \mathcal{L}(h^*),$$

where $J_F(h^*)$ is the Jacobian of F_θ at the fixed point. One then computes parameter gradients via Jacobian–vector products without storing the entire forward trajectory. When F_θ is contractive, $I - J_F(h^*)$ is well-conditioned in the sense relevant to iterative solvers, and v may be computed by fixed-point iteration or Krylov methods using only products of $J_F(h^*)^\top$ with vectors. This training mode aligns with our goal of well-posedness: the same structural restrictions that guarantee convergence also yield stable differentiation through the equilibrium.

The next section provides the required Lipschitz bounds and explicit contraction conditions under $\|\cdot\|_{\infty,2}$, thereby justifying both the forward solver and the implicit gradient computation on cyclic directed graphs.

5 Convergence Theory: explicit Lipschitz bounds and linear rates

Block norm and degree parameters. We work throughout with the block maximum norm

$$\|h\|_{\infty,2} := \max_{i \in V} \|h_i\|_2,$$

and we write $\Delta_{\text{in,max}} := \max_{i \in V} |N_{\text{in}}(i)|$ in addition to $\Delta_{\max} = \max_{j \in V} |N_{\text{out}}(j)|$. In bounded-degree regimes one often has $\Delta_{\text{in,max}} \lesssim \Delta_{\max}$; when this holds we may absorb $\Delta_{\text{in,max}}$ into the same constant $C(\Delta_{\max})$. We also set

$$B_\sigma := \sup_{z \in \mathbb{R}^d} \|\sigma(z)\|_2 < \infty,$$

which is finite since σ is bounded (e.g. for coordinatewise tanh, $B_\sigma \le \sqrt{d}$).

A senderwise softmax Lipschitz bound. Fix a sender j with $|N_{\text{out}}(j)| > 0$ and consider the vector of scores $s_{\cdot j} \in \mathbb{R}^{|N_{\text{out}}(j)|}$ defined by $[s_{\cdot j}]_i = s_{ij}$ for $i \in N_{\text{out}}(j)$. Let sm denote the softmax map, so that $\beta_{\cdot j} = \text{sm}(s_{\cdot j})$. The Jacobian of sm at a probability vector p is $J(p) = \text{diag}(p) - pp^\top$, and a direct column-sum bound yields

$$\|J(p)\|_{\infty \to 1} \;=\; \max_\ell \sum_k |J(p)_{k\ell}| \;=\; \max_\ell 2p_\ell(1-p_\ell) \;\le\; \frac{1}{2}.$$

Consequently, for any two score vectors u, v of the same dimension,

$$\|\text{sm}(u) - \text{sm}(v)\|_1 \;\le\; \frac{1}{2}\|u - v\|_\infty, \qquad \|\text{sm}(u) - \text{sm}(v)\|_\infty \;\le\; \frac{1}{2}\|u - v\|_\infty. \tag{1}$$

Since our scores are scaled by $1/\tau$, the effective sensitivity with respect to the unscaled logits is proportional to $1/\tau$.

Score sensitivity under $\|\cdot\|_{\infty,2}$. By the L_e-Lipschitz property of e_θ in its hidden-state arguments, for any two hidden-state arrays $h, \tilde{h}$ and any edge $(j, i) \in E$ we have

$$|s_{ij}(h) - s_{ij}(\tilde{h})| \;=\; \frac{1}{\tau}\left|e_\theta(h_i, h_j, a_{ji}) - e_\theta(\tilde{h}_i, \tilde{h}_j, a_{ji})\right| \;\le\; \frac{L_e}{\tau}\big(\|h_i - \tilde{h}_i\|_2 + \|h_j - \tilde{h}_j\|_2\big).$$

Taking the maximum over all i, j shows

$$\max_{(j,i) \in E} |s_{ij}(h) - s_{ij}(\tilde{h})| \;\le\; \frac{2L_e}{\tau}\|h - \tilde{h}\|_{\infty,2}. \tag{2}$$

Attention perturbations. Combining (1) and (2), we obtain for each sender j with $|N_{\text{out}}(j)| > 0$,

$$\sum_{i \in N_{\text{out}}(j)} |\beta_{ij}(h) - \beta_{ij}(\tilde{h})| \;=\; \|\beta_{\cdot j}(h) - \beta_{\cdot j}(\tilde{h})\|_1 \;\le\; \frac{1}{2}\|s_{\cdot j}(h) - s_{\cdot j}(\tilde{h})\|_\infty \;\le\; \frac{L_e}{\tau}\|h - \tilde{h}\|_{\infty,2}. \tag{3}$$

This estimate makes the role of the temperature explicit: increasing τ uniformly reduces the state-dependence of routing.

Lipschitz bound for message aggregation. For each receiver i we consider

$$m_i(h) = \sum_{j \in N_{\text{in}}(i)} \beta_{ij}(h)\, W_m h_j.$$

We add and subtract $\beta_{ij}(h)\, W_m \tilde{h}_j$ to obtain

$$m_i(h) - m_i(\tilde{h}) = \sum_{j \in N_{\text{in}}(i)} \beta_{ij}(h)\, W_m (h_j - \tilde{h}_j) + \sum_{j \in N_{\text{in}}(i)} (\beta_{ij}(h) - \beta_{ij}(\tilde{h}))\, W_m \tilde{h}_j.$$

Using $\|W_m v\|_2 \le \|W_m\|_2 \|v\|_2$, $\beta_{ij}(h) \le 1$, and $\|\tilde{h}_j\|_2 \le B_\sigma$ whenever $\tilde{h}$ lies in the range of Φ_θ, we bound

$$\|m_i(h) - m_i(\tilde{h})\|_2 \;\le\; \|W_m\|_2 \sum_{j \in N_{\text{in}}(i)} \|h_j - \tilde{h}_j\|_2 + \|W_m\|_2\, B_\sigma \sum_{j \in N_{\text{in}}(i)} |\beta_{ij}(h) - \beta_{ij}(\tilde{h})|.$$

The first sum contributes at most $\|W_m\|_2\, \Delta_{\text{in,max}} \|h - \tilde{h}\|_{\infty,2}$. For the second sum we use the elementary inequality $|\beta_{ij}(h) - \beta_{ij}(\tilde{h})| \le \|\beta_{\cdot j}(h) - \beta_{\cdot j}(\tilde{h})\|_1$ and then (3), obtaining

$$\sum_{j \in N_{\text{in}}(i)} |\beta_{ij}(h) - \beta_{ij}(\tilde{h})| \;\le\; \sum_{j \in N_{\text{in}}(i)} \frac{L_e}{\tau} \|h - \tilde{h}\|_{\infty,2} \;\le\; \Delta_{\text{in,max}} \frac{L_e}{\tau} \|h - \tilde{h}\|_{\infty,2}.$$

Hence, if $\|W_m\|_2 \le \kappa_m$,

$$\|m(h) - m(\tilde{h})\|_{\infty,2} \;\le\; \kappa_m\, \Delta_{\text{in,max}} \Big(1 + \frac{B_\sigma L_e}{\tau}\Big) \|h - \tilde{h}\|_{\infty,2}. \tag{4}$$

In bounded-degree settings we may summarize the dependence by writing $\Delta_{\text{in,max}} \le C(\Delta_{\max})$ with $C(\Delta_{\max}) = O(\Delta_{\max})$.

Lipschitz constant for Φ_θ and a contraction condition for F_θ. Since σ is 1-Lipschitz and $h \mapsto W_x x + b$ is constant in h, (4) implies

$$\|\Phi_\theta(h) - \Phi_\theta(\tilde{h})\|_{\infty,2} \;\le\; L_\Phi \|h - \tilde{h}\|_{\infty,2}, \qquad L_\Phi := \kappa_m\, \Delta_{\text{in,max}} \Big(1 + \frac{B_\sigma L_e}{\tau}\Big). \tag{5}$$

For the damped map $F_\theta(h) = (1 - \alpha)h + \alpha \Phi_\theta(h)$ we then have

$$\|F_\theta(h) - F_\theta(\tilde{h})\|_{\infty,2} \;\le\; \big((1 - \alpha) + \alpha L_\Phi\big) \|h - \tilde{h}\|_{\infty,2}.$$

Thus a sufficient explicit contraction condition is

$$q \;:=\; (1 - \alpha) + \alpha L_\Phi \;<\; 1, \tag{6}$$

which holds in particular whenever $L_\Phi < 1$ (and hence under appropriate restrictions on κ_m, τ^{-1}, and the relevant degree bound).

Linear convergence and iteration complexity. Assuming (6), Banach's fixed-point theorem implies that F_θ admits a unique fixed point h^* and that the Picard iterates satisfy the linear rate

$$\|h^t - h^*\|_{\infty,2} \;\le\; q^t \, \|h^0 - h^*\|_{\infty,2}, \qquad t \ge 0.$$

Equivalently, to achieve $\|h^T - h^*\|_{\infty,2} \le \varepsilon$ it suffices that

$$T \;\ge\; \frac{\log(\varepsilon^{-1}) + \log \|h^0 - h^*\|_{\infty,2}}{\log(q^{-1})}.$$

Moreover, the fixed-point residual $r^t = \|h^{t+1} - h^t\|_{\infty,2}$ controls the error a posteriori: since $h^* = F_\theta(h^*)$ and F_θ is q-contractive,

$$\|h^t - h^*\|_{\infty,2} \;\le\; \sum_{\ell=t}^{\infty} \|h^{\ell+1} - h^\ell\|_{\infty,2} \;\le\; \sum_{\ell=t}^{\infty} q^{\ell-t}\, r^t \;=\; \frac{1}{1-q}\, r^t.$$

This justifies stopping rules based on $r^t \le \varepsilon(1-q)$, and it makes explicit the tradeoff between expressivity and solvability: stronger attention sensitivity (large L_e/τ), stronger mixing (κ_m), or larger degree bounds inflate L_Φ, thereby degrading q and increasing the number of iterations required for a prescribed tolerance.

6 Stability Theory: perturbation bounds and robustness certificates

A general perturbation bound for contractive implicit layers. Throughout this section we assume that both the nominal operator F and a perturbed operator $\tilde{F}$ are contractions on $(\mathbb{R}^{n\times d}, \|\cdot\|_{\infty,2})$ with a common contraction factor at most $q < 1$. Let h^* and $\tilde{h}^*$ denote their respective fixed points. The key estimate is the standard contraction perturbation lemma:

$$\|h^* - \tilde{h}^*\|_{\infty,2} \;\le\; \frac{1}{1-q} \sup_h \|F(h) - \tilde{F}(h)\|_{\infty,2}. \tag{7}$$

Indeed, using $h^* = F(h^*)$ and $\tilde{h}^* = \tilde{F}(\tilde{h}^*)$, we write

$$\|h^* - \tilde{h}^*\|_{\infty,2} = \|F(h^*) - \tilde{F}(\tilde{h}^*)\|_{\infty,2} \le \|F(h^*) - F(\tilde{h}^*)\|_{\infty,2} + \|F(\tilde{h}^*) - \tilde{F}(\tilde{h}^*)\|_{\infty,2},$$

and we bound $\|F(h^*) - F(\tilde{h}^*)\|_{\infty,2} \le q\|h^* - \tilde{h}^*\|_{\infty,2}$ by contractivity and $\|F(\tilde{h}^*) - \tilde{F}(\tilde{h}^*)\|_{\infty,2} \le \sup_h \|F(h) - \tilde{F}(h)\|_{\infty,2}$. Rearranging yields (7). The salient feature is the amplification factor $(1-q)^{-1}$: as $q \uparrow 1$ the fixed point becomes increasingly sensitive to any modeling or numerical perturbation.

Feature perturbations: explicit dependence on $(\Delta x, \Delta a)$. We now specialize (7) to perturbations of node and edge features, keeping the graph topology fixed. Write $\Phi = \Phi_{\theta,x,a}$ for the undamped map and $F = (1-\alpha)I + \alpha\Phi$. Let $(\tilde{x}, \tilde{a})$ be perturbed features and define $\tilde{\Phi} = \Phi_{\theta,\tilde{x},\tilde{a}}$, $\tilde{F} = (1-\alpha)I + \alpha\tilde{\Phi}$ with the same (θ, α, τ). Assume additionally that the score function is Lipschitz in edge features,

$$\left|e_\theta(h_i, h_j, a_{ji}) - e_\theta(h_i, h_j, \tilde{a}_{ji})\right| \;\le\; L_a \,\|a_{ji} - \tilde{a}_{ji}\|_2,$$

uniformly over (h_i, h_j), and set $\|W_x\|_2 \le \kappa_x$.

For any fixed h, the pre-activation difference at node i is

$$u_i(h) - \tilde{u}_i(h) = W_x(x_i - \tilde{x}_i) + \sum_{j \in N_{\text{in}}(i)} \big(\beta_{ij}(h; a) - \beta_{ij}(h; \tilde{a})\big)\, W_m h_j,$$

where we emphasize the dependence of β on edge features through the logits. Using the 1-Lipschitz property of σ and $\|W_m h_j\|_2 \le \|W_m\|_2 \|h_j\|_2$, we obtain

$$\|\Phi(h) - \tilde{\Phi}(h)\|_{\infty,2} \;\le\; \kappa_x \,\|x - \tilde{x}\|_{\infty,2} + \kappa_m \,\|h\|_{\infty,2} \cdot \max_i \sum_{j \in N_{\text{in}}(i)} |\beta_{ij}(h; a) - \beta_{ij}(h; \tilde{a})|. \tag{8}$$

To bound the attention perturbation term, we again use senderwise softmax stability: for each sender j,

$$\|\beta_{\cdot j}(h; a) - \beta_{\cdot j}(h; \tilde{a})\|_1 \;\le\; \frac{1}{2}\,\|s_{\cdot j}(h; a) - s_{\cdot j}(h; \tilde{a})\|_\infty \;\le\; \frac{L_a}{2\tau}\,\|a - \tilde{a}\|_{\infty,2},$$

where the last step follows because the logits differ only through a and are scaled by $1/\tau$. Therefore, for each receiver i,

$$\sum_{j \in N_{\text{in}}(i)} |\beta_{ij}(h; a) - \beta_{ij}(h; \tilde{a})| \;\le\; \sum_{j \in N_{\text{in}}(i)} \|\beta_{\cdot j}(h; a) - \beta_{\cdot j}(h; \tilde{a})\|_1 \;\le\; \Delta_{\text{in,max}}\, \frac{L_a}{2\tau}\, \|a - \tilde{a}\|_{\infty,2}.$$

If, in addition, h lies on the forward trajectory of the bounded map Φ (hence $\|h\|_{\infty,2} \le B_\sigma$), then (8) gives the uniform operator perturbation bound

$$\sup_h \|\Phi(h) - \tilde{\Phi}(h)\|_{\infty,2} \;\le\; \kappa_x \,\|\Delta x\|_{\infty,2} + \kappa_m \, B_\sigma \, \Delta_{\text{in,max}}\, \frac{L_a}{2\tau}\, \|\Delta a\|_{\infty,2},$$

and thus, since $F - \tilde{F} = \alpha(\Phi - \tilde{\Phi})$,

$$\|h^* - \tilde{h}^*\|_{\infty,2} \;\le\; \frac{\alpha}{1-q}\Big(\kappa_x \,\|\Delta x\|_{\infty,2} + \kappa_m \, B_\sigma \, \Delta_{\text{in,max}}\, \frac{L_a}{2\tau}\, \|\Delta a\|_{\infty,2}\Big). \tag{9}$$

This provides a direct robustness guarantee: larger temperature τ damps the impact of edge-feature perturbations on routing and, hence, on the fixed point.

Sensitivity to graph rewiring and sparsified routing. We next treat perturbations that change the effective routing, as occurs under graph rewiring, edge deletions, or approximations such as neighborhood sampling. Rather than coupling two different adjacency patterns directly, we express the perturbation at the level of the senderwise routing kernel. Let $\beta(h)$ be the nominal outgoing-normalized weights and let $\tilde{\beta}(h)$ be any approximate kernel satisfying, for all senders j,

$$\delta_j \;:=\; \sup_h \|\beta_{\cdot j}(h) - \tilde{\beta}_{\cdot j}(h)\|_1 \;<\; \infty, \qquad \delta_{\max} := \max_j \delta_j. \tag{10}$$

Define $\tilde{\Phi}$ by replacing β with $\tilde{\beta}$ in the message term, keeping (x, a, θ) fixed. Then, for any bounded h with $\|h\|_{\infty,2} \le B_\sigma$,

$$\|\Phi(h) - \tilde{\Phi}(h)\|_{\infty,2} \le \max_i \sum_{j \in N_{\text{in}}(i)} |\beta_{ij}(h) - \tilde{\beta}_{ij}(h)| \cdot \|W_m h_j\|_2 \le \kappa_m \, B_\sigma \, \Delta_{\text{in,max}} \, \delta_{\max},$$

and consequently

$$\|h^* - \tilde{h}^*\|_{\infty,2} \;\le\; \frac{\alpha}{1-q} \, \kappa_m \, B_\sigma \, \Delta_{\text{in,max}} \, \delta_{\max}. \tag{11}$$

The bound (11) is particularly convenient because $\delta_{\max}$ can be estimated for a variety of approximation schemes (e.g. truncating softmax support, quantizing scores, or approximate normalization).

A common special case is edge deletion followed by renormalization. Fix a sender j and let $R_j \subseteq N_{\text{out}}(j)$ be the removed outgoing edges, and let $\tilde{\beta}_{\cdot j}$ be the renormalized restriction of $\beta_{\cdot j}$ to $N_{\text{out}}(j) \backslash R_j$. Writing $\rho_j := \sum_{i \in R_j} \beta_{ij}(h)$ for the removed mass (which may depend on h), a direct computation yields

$$\|\beta_{\cdot j}(h) - \tilde{\beta}_{\cdot j}(h)\|_1 \;=\; 2\rho_j,$$

and hence $\delta_{\max} \le 2 \sup_{j,h} \rho_j$. Thus deletions are harmless precisely when they remove only negligible attention mass; in particular, smoother attention (larger τ) tends to spread mass and makes ρ_j small for moderate-sized removals.

Interpretation as a robustness certificate. Taken together, (9) and (11) certify that the implicit embedding map $(x, a, \beta) \mapsto h^*$ is Lipschitz on any parameter regime where $q < 1$. In applications, we may report the quantity $(1-q)^{-1}$ as a condition number for the layer: it converts any bound on operator mismatch (from feature noise, approximate routing, or numerical error) into a bound on the change in the fixed point. Moreover, when the fixed point is computed approximately by Picard iteration, the residual-based a posteriori estimate

$$\|h^t - h^*\|_{\infty,2} \le \frac{1}{1-q} \, \|h^{t+1} - h^t\|_{\infty,2}$$

provides a complementary certificate of solve accuracy that is directly computable at inference time.

Limitations and transition to lower bounds. All bounds above use contractivity twice: first to ensure uniqueness of h^* and second to control perturbation amplification by $(1-q)^{-1}$. In the absence of such structure, even deciding whether a stable fixed point exists (let alone bounding its dependence on data or graph edits) becomes qualitatively more difficult; in the next section we formalize this limitation via PPAD-hardness for general continuous graph-defined maps and discuss why some restriction—contraction or a comparable monotone-operator hypothesis—is necessary for provable guarantees.

7 Lower bounds and necessity of structural restrictions

The stability estimates of §6 rely on contractivity in an essential way: uniqueness of the embedding is guaranteed only because F is a contraction, and the perturbation amplification factor $(1-q)^{-1}$ is finite only because $q < 1$. We now justify that some restriction of this type is not merely convenient but is, in a precise complexity-theoretic sense, necessary if one seeks worst-case polynomial-time guarantees for fixed-point computation and robustness.

PPAD-hardness in the absence of contractivity. Consider a general continuous operator $F : \mathcal{H} \to \mathcal{H}$ on a compact convex set $\mathcal{H} \subset \mathbb{R}^{n\times d}$. By Brouwer's theorem, F has at least one fixed point, but there is no generic polynomial-time algorithm that, given a description of F, computes (or even approximates to inverse-polynomial accuracy) a fixed point in the worst case. The canonical formalization is the complexity class PPAD, which captures total search problems guaranteed to have solutions by parity arguments; the problem of finding an ε-approximate Brouwer fixed point is PPAD-complete.

In our setting, if we drop the norm/temperature/damping restrictions that enforce $\alpha L_\Phi < 1$ and allow the local components (in particular the scoring map e_θ and any auxiliary MLPs used to compute logits or messages) to be unrestricted continuous parameterizations, then the induced graph-defined operator F_θ can emulate an arbitrary Brouwer function on an appropriate domain. Concretely, fix a dimension D and a continuous map $f : [0,1]^D \to [0,1]^D$ given, say, by an arithmetic circuit. We may encode an input point $z \in [0,1]^D$ into a subset of node states (or into the full state vector via a fixed linear embedding) and design a bounded-degree directed graph whose nodes represent circuit gates. Using sufficiently expressive local computations at each node, we can implement the circuit evaluation $z \mapsto f(z)$ by a single global application of a message passing operator; that is, we can arrange that, on a designated readout subvector $\pi(h) \in \mathbb{R}^D$,

$$\pi\big(F_\theta(h)\big) \approx f\big(\pi(h)\big), \qquad \pi(h) \in [0,1]^D,$$

while the remaining coordinates serve only to shuttle intermediate values across the circuit graph. (A bounded activation such as tanh is compatible with this emulation by rescaling and shifting signals to remain within a compact interval, and the attention mechanism can be used either as a fixed routing kernel or as a state-dependent switch so long as we do not constrain its Lipschitz constant.) Any fixed point $h = F_\theta(h)$ then induces a fixed point $z = \pi(h) \approx f(z)$, and conversely any fixed point of f can be lifted to one of F_θ by extending with consistent intermediate gate values. Since approximate fixed-point computation for f is PPAD-hard, so is approximate fixed-point computation for the induced F_θ in this unrestricted regime.

This hardness has two immediate implications. First, absent contraction (or a comparable structure), we cannot expect a worst-case convergence guarantee for Picard iteration: for general continuous F the iteration $h^{t+1} = F(h^t)$ may fail to converge, may converge to different fixed points depending on initialization, or may exhibit periodic or chaotic behavior. Second, even replacing Picard by more sophisticated black-box methods does not circumvent the obstacle: PPAD-hardness rules out a generic polynomial-time algorithm under standard complexity assumptions.

Why contraction (or monotone structure) is the right kind of restriction. The preceding reduction does not use any pathology beyond continuity and boundedness; thus, to obtain provable guarantees, we must restrict the operator class. Contractivity is a particularly clean restriction because it yields three properties simultaneously: (i) existence and uniqueness of h^*; (ii) an explicit algorithm (Picard iteration) with linear convergence rate; and (iii) perturbation stability with an explicit condition number $(1 - q)^{-1}$. Other restrictions can also suffice, but they must exclude the general Brouwer setting in a comparable way. A common alternative is monotone operator theory: if the fixed point can be recast as a root of a strongly monotone and Lipschitz map (or the solution of a strongly convex variational inequality), then projected gradient or extragradient methods admit polynomial-time rates. However, such monotonicity typically requires architectural constraints that are at least as stringent as the spectral-norm/temperature/damping controls used to enforce contraction, and in practice are less directly compatible with expressive attention mechanisms. Thus, from the standpoint of both analysis and implementability, contraction is a natural minimal hypothesis.

From a design perspective, the role of τ and $\|W_m\|_2$ is now conceptually clear. Without a lower bound on τ (preventing arbitrarily sharp softmax) and an upper bound on $\|W_m\|_2$ (preventing arbitrarily strong amplification through messages), the Lipschitz constant L_Φ can become arbitrarily large; consequently, no choice of α can ensure $\alpha L_\Phi < 1$ uniformly over parameters, and the model class again contains hard instances. The contractive

regime is therefore not merely a technical convenience: it is the regime in which fixed-point computation and robustness analysis become algorithmically well-posed.

A minimal per-iteration lower bound: one must inspect all edges. Even within the contractive regime, there are information-theoretic limits on the cost of applying F in sparse adjacency access models. We record a simple indistinguishability argument showing that any algorithm that claims to compute the exact update $h \mapsto \Phi_\theta(h)$ (or the exact attention weights $\beta(h)$) must, in the worst case, inspect every edge at least once per iteration.

Formally, consider an oracle model in which an algorithm queries adjacency lists and edge features to produce $\Phi_\theta(h)$. Suppose the algorithm does not inspect an edge $(j, i) \in E$. We can construct two instances that are identical on all inspected edges and nodes but differ on the uninspected edge feature a_{ji} in such a way that the logit $e_\theta(h_i, h_j, a_{ji})$ changes by a nonzero amount while leaving all other logits unchanged. Because the outgoing normalization couples all outgoing edges of j through the denominator, this modification changes the entire sender distribution $\beta_{\cdot j}(h)$ and hence modifies the message received by at least one neighbor of j. Therefore $\Phi_\theta(h)$ differs between the two instances, yet the algorithm would output the same value on both, contradicting correctness. In particular, under standard sparse access (adjacency lists) the per-iteration time is bounded below by $\Omega(m)$ edge inspections, matching the upper bounds in §6 up to multiplicative factors from the score computation and the hidden dimension.

This lower bound does not preclude approximation schemes (sampling, truncation, quantization), but it clarifies what must be paid for: any reduction in edge work necessarily introduces an operator mismatch $\tilde{F} \neq F$, and the stability bounds of §6 quantify precisely how such mismatch propagates to the embedding via the factor $(1 - q)^{-1}$.

Summary and transition. We have thus isolated two complementary necessities. On the algorithmic side, without contractivity (or an alternative structure excluding Brouwer-hard instances) fixed-point computation is PPAD-hard in the worst case. On the computational side, even when contraction holds, exact evaluation of a flow-attentive layer is inherently $\Omega(m)$ per iteration in sparse models. The next section discusses extensions that preserve the favorable contractive/stable regime while increasing modeling capacity, including capacity-aware normalizations, block-implicit updates, and hybrid global–local designs.

8 Extensions

We record several extensions that preserve the defining structural features of the flow-attentive layer—locality, outgoing-normalized routing, and the possibility of enforcing contractivity by explicit parameter controls—while increasing modeling capacity. In each case, the guiding requirement is that the induced operator on hidden states remains a well-posed fixed-point map on $(\mathbb{R}^{n\times d}, \|\cdot\|_{\infty,2})$, so that the guarantees of §§6–7 continue to apply after suitable modifications of the Lipschitz constants.

(i) Capacity-aware normalizations. Outgoing normalization $\sum_{i\in N_{\mathrm{out}}(j)} \beta_{ij} = 1$ encodes conservation at the sender, but in many flow-like domains (traffic, power, packet routing) one also wishes to represent receiver-side capacity constraints. A simple modification is to introduce node capacities $c_i > 0$ and replace the raw incoming aggregation by a saturating map that limits the effective inflow. For instance, letting

$$\bar{m}_i(h) \;:=\; \sum_{j\in N_{\mathrm{in}}(i)} \beta_{ij}(h)\, W_m h_j, \qquad m_i(h) \;:=\; c_i \cdot \tanh\Big(\frac{\bar{m}_i(h)}{c_i}\Big),$$

we update $h_i \leftarrow \sigma(W_x x_i + m_i(h) + b)$. Since coordinatewise tanh is 1-Lipschitz and bounded, this introduces an additional nonexpansive component and hence does not worsen the Lipschitz constant of Φ_θ beyond the factor already induced by W_m and the attention map; indeed $\|m(h) - m(\tilde{h})\|_{\infty,2} \le \|\bar{m}(h) - \bar{m}(\tilde{h})\|_{\infty,2}$. One may similarly incorporate edge capacities $u_{ji} > 0$ by tempering logits with a bounded additive bias $\log u_{ji}$, i.e.

$$\beta_{ij}(h) \;:=\; \frac{\exp\big((e_\theta(h_i, h_j, a_{ji}) + \log u_{ji})/\tau\big)}{\sum_{k\in N_{\mathrm{out}}(j)} \exp\big((e_\theta(h_k, h_j, a_{jk}) + \log u_{jk})/\tau\big)},$$

which preserves exact outgoing conservation (Proposition 4) while steering probability mass toward higher-capacity edges. When one requires approximate receiver-side conservation (e.g. $\sum_{j\in N_{\mathrm{in}}(i)} \beta_{ij} \le 1$), one can apply a second normalization or projection on the incoming weights. The mathematically clean option is a smooth projection onto a capped simplex at each receiver (implemented, for example, by an entropic regularized projection), which is Lipschitz with an explicit modulus depending on the regularization strength. The resulting operator remains amenable to the same contraction analysis, at the cost of enlarging L_Φ by the projection's Lipschitz constant.

(ii) Multiple sweeps and block-implicit variants. The Jacobi-style Picard iteration in Algorithm 1 updates all nodes simultaneously. In sparse directed graphs with strong local feedback (small directed cycles), Gauss–Seidel-type sweeps can be empirically more efficient: we update nodes in a

chosen order and immediately reuse the latest values in subsequent message computations. Formally, we partition V into blocks $V = V_1 \cup \cdots \cup V_B$ and define a block operator $\Phi_\theta^{(b)}$ that updates only coordinates in V_b while holding others fixed. A full sweep is the composition

$$\mathcal{S}_\theta \;:=\; \Phi_\theta^{(B)} \circ \cdots \circ \Phi_\theta^{(1)}, \qquad h^{t+1} \;=\; (1-\alpha)h^t + \alpha\, \mathcal{S}_\theta(h^t).$$

If the original Φ_θ is L_Φ-Lipschitz under $\|\cdot\|_{\infty,2}$, then each block map is also L_Φ-Lipschitz (as a restriction of coordinates), and hence $\mathcal{S}_\theta$ is L_Φ^B-Lipschitz in the worst case by composition. This bound is pessimistic; in practice, if blocks are chosen to reduce inter-block coupling (e.g. via a topological order on a condensation DAG, or via strongly connected components), one can obtain a much smaller effective constant. Independently of ordering, if we damp the sweep by α and enforce $\alpha\,\mathrm{Lip}(\mathcal{S}_\theta) < 1$, then Banach's theorem again yields a unique fixed point, and asynchronous iteration theory implies convergence even when blocks are updated with delays, provided the underlying map is a contraction. These block updates also admit partial implicitness: one may solve the fixed point restricted to a small subgraph (a "patch") to higher accuracy while treating the complement as fixed boundary data, thereby allocating compute adaptively without leaving the contractive regime.

(iii) Hybrid global–local architectures. Local message passing can be augmented with a global latent state $g \in \mathbb{R}^{d_g}$ (or a small set of global tokens) to represent long-range context. We consider an augmented state $\bar{h} = (h, g)$ and define

$$\Phi_\theta^{\mathrm{hyb}}(h,g) \;:=\; \big(\,\Phi_\theta^{\mathrm{loc}}(h,g),\; \Phi_\theta^{\mathrm{glob}}(h,g)\,\big), \qquad F_\theta^{\mathrm{hyb}}(\bar{h}) = (1-\alpha)\bar{h} + \alpha\,\Phi_\theta^{\mathrm{hyb}}(\bar{h}).$$

Here $\Phi_\theta^{\mathrm{loc}}$ may include an additional term Ug in the node pre-activation, while $\Phi_\theta^{\mathrm{glob}}$ aggregates node information, e.g. via a permutation-invariant readout $r(h) = \sum_i \rho(h_i)$ followed by a bounded nonlinearity. If ρ is 1-Lipschitz and $\|U\|_2$ is bounded, then $\Phi_\theta^{\mathrm{hyb}}$ is Lipschitz with an explicit constant controlled by $(L_\Phi, \|U\|_2)$ and the Lipschitz modulus of $\Phi_\theta^{\mathrm{glob}}$. Consequently, by choosing α and norm bounds to ensure $\alpha\,\mathrm{Lip}(\Phi_\theta^{\mathrm{hyb}}) < 1$, we retain a unique fixed point on the augmented space. This provides a principled route to "transformer-hybrid" behavior while maintaining the well-posedness of the implicit layer: global attention modules are admissible insofar as their operator norm is controlled, and their softmax temperature is bounded away from zero analogously to the local case.

(iv) Node- and edge-level tasks; source–sink structure. Once a fixed point h^* is defined, task-specific heads may be attached without affecting the fixed-point existence/uniqueness, since prediction occurs after convergence.

For node classification or regression we apply a pointwise map $\hat{y}_i = \psi(h_i^*)$. For edge-level tasks (link prediction, edge labels) we form edge representations from endpoint states and edge features, e.g. $\hat{y}_{ij} = \psi_{\text{edge}}([h_i^*, h_j^*, a_{ij}])$, optionally including the equilibrium routing weight $\beta_{ij}(h^*)$ as an additional feature when the task is inherently flow-like. Graph-level tasks use a readout $\hat{y} = \psi_{\text{graph}}(\sum_i \rho(h_i^*))$. In domains with distinguished sources and sinks (S, T), we can incorporate this structure into the state update by adding fixed boundary injections, e.g. an additional input term p_i with $p_i > 0$ for $i \in S$ and $p_i < 0$ for $i \in T$, or by clamping certain coordinates of h_i to prescribed values and iterating only over the free coordinates; contractivity on the free subspace suffices for uniqueness of the unconstrained variables. During training, one may either unroll a finite number of iterations or use implicit differentiation at the fixed point; the preceding extensions remain compatible with either choice provided the same contractive bounds are enforced so that the linear system underlying implicit gradients is well-conditioned (with condition number controlled by $(1 - \alpha L_\Phi)^{-1}$).

Taken together, these extensions delineate a family of expressive yet analyzable implicit GNN layers: we may enrich the routing kernel to respect capacities, allocate computation via block solves and sweeps, incorporate controlled global context, and support standard node/edge/graph supervision, all while maintaining the fixed-point guarantees by explicit control of the relevant Lipschitz constants and damping.

9 Experimental Plan

We evaluate the proposed flow-attentive implicit layer along four axes: (a) predictive utility on cyclic flow domains, (b) controlled tests isolating directed feedback and sharp routing, (c) compute–accuracy tradeoffs intrinsic to fixed-point computation, and (d) ablations and architectural comparisons clarifying when implicit contractive design is advantageous relative to unrolled recurrent GNNs and global-attention hybrids.

Benchmarks with intrinsic directed cycles. We focus on domains where directed cycles are structural rather than incidental. *(Power)* We consider standard transmission-network benchmarks with meshed (cyclic) topology, using public IEEE-style test cases with bus/branch features. Typical tasks include (i) node-level regression of voltage magnitude/angle surrogates and (ii) edge-level regression/classification of line loading or constraint violations under varying injections. We treat buses as nodes with exogenous features (loads, generations, limits) and lines as directed edges with attributes (impedance surrogates, thermal limits, direction-dependent status), and we optionally include marked source/sink sets derived from net injection sign. *(Traffic)* We consider directed road networks with recurrent

loops (ring roads, grid-like downtown cores). Tasks include predicting edge-level congestion indicators (speed, density, or travel time) and node-level accumulation proxies (intersection delay), given static attributes and (when available) aggregated OD or sensor-derived features. In both domains we report standard predictive metrics (MSE/MAE for regression; AUROC/F1 for classification) and additionally record the equilibrium residual and iteration count required to reach a prescribed tolerance.

Synthetic controlled-cycle suites. To isolate the interaction between directed feedback and attention sharpness, we generate synthetic directed graphs with planted cyclic structure. Concretely, for chosen parameters $(n, \Delta_{\max}, L)$ we (i) create a base directed Erdős–Rényi graph with bounded out-degree $\Delta_{\max}$, (ii) plant disjoint directed cycles of length L (including short cycles $L \in \{2, 3, 4\}$), and (iii) assign node/edge features so that the target depends on multi-hop circulation along the planted cycles (e.g. a parity-like signal or a diffusion-with-reinjection surrogate). We vary (a) the fraction of nodes participating in cycles, (b) the strength of cycle-dependent signal relative to noise, and (c) the degree of "bottleneck" edges whose attention logits are systematically higher. This suite permits stress-testing regimes in which small τ induces near-deterministic routing and hence potentially large effective Lipschitz constants through the softmax sensitivity.

Compute–accuracy tradeoffs for equilibrium computation. Because the forward pass solves (approximately) a fixed point, we quantify the relation between solve accuracy and downstream performance. We adopt the fixed-point residual

$$r^t \;:=\; \|h^{t+1} - h^t\|_{\infty,2},$$

and we study performance as a function of (i) a fixed iteration budget T and (ii) an adaptive stopping rule $r^t \le \varepsilon$. For each dataset we produce curves of validation error versus (wall-clock) time, and error versus achieved residual. We report the empirical linear rate by regressing $\log r^t$ on t in the convergent regime, and we compare it to the predicted dependence on α and τ (holding parameter norm bounds fixed). When implicit differentiation is used, we likewise report the number of iterations required by the backward linear solve (e.g. a Neumann-series truncation or Krylov method) to reach a relative tolerance, thereby exposing the conditioning effect governed by $(1 - \alpha L_\Phi)^{-1}$.

Ablations on temperature, damping, and norm control. We ablate the parameters directly implicated by the contraction analysis. *(Temperature)* We sweep τ on a log grid from a sharp regime to a smooth regime, recording (a) accuracy, (b) convergence failures or slowdowns (as measured

by residual decay), and (c) statistics of the equilibrium routing kernel (entropy of $\beta_{ij}(h^*)$ per sender and concentration on maximum-probability edges). *(Damping)* We sweep $\alpha \in (0, 1]$ and evaluate the predicted speed–stability tradeoff: larger α reduces averaging but can violate $\alpha L_\Phi < 1$; smaller α slows convergence but may enlarge the basin of stable behavior. *(Iteration budget)* For unrolled training we compare fixed T to adaptive ε-stopping; we record generalization versus T to identify whether improvements stem from better equilibrium approximation or from effectively deeper computation. *(Spectral norm bounds)* We compare explicit control of $\|W_m\|_2$ (via spectral normalization or reparameterization) to unconstrained training, and we measure the resulting empirical Lipschitz proxy (e.g. Jacobian-vector product norms estimated at equilibrium) as well as the frequency of divergence in Picard iteration.

Comparisons to unrolled recurrent GNNs. We compare against (i) standard message passing with attention but without outgoing-normalization, (ii) recurrent GNNs that apply the same local update for a fixed number of steps (with and without residual connections), and (iii) equilibrium-style models without conservation constraints. To make comparisons meaningful, we match parameter counts and per-iteration edge operation costs. We additionally compare training regimes: unrolled backpropagation through T steps versus implicit differentiation at the equilibrium. Beyond predictive accuracy, we report stability under feature perturbations: given perturbed inputs $(x + \Delta x, a + \Delta a)$ with controlled magnitudes, we measure $\|h^*(x, a) - h^*(x + \Delta x, a + \Delta a)\|_{\infty,2}$ and compare to analogous differences in the unrolled models at their final iterate, thereby empirically probing the contraction-based stability claim.

Comparisons to transformer-hybrid architectures. We implement controlled global–local baselines by augmenting local message passing with either (i) a global token updated by attention to nodes or (ii) a graph-level readout fed back to nodes. We consider two variants: a fully unrolled hybrid with T rounds, and an implicit hybrid constrained by operator-norm control on the global-to-local map. We then test whether the implicit, contractive hybrid achieves similar benefits of global context at lower iteration counts or with improved robustness in cyclic graphs. As an additional diagnostic we examine the sensitivity of performance to long-range dependencies by artificially increasing graph diameter while preserving local cycle statistics.

Evaluation protocol and reporting. For each task we perform multiple random seeds, report mean and standard deviation, and provide convergence statistics (fraction of runs reaching ε within budget, median iterations, and residual trajectories). We separate *optimization instability* (training diver-

gence) from *fixed-point instability* (failure of Picard iteration at inference) and report both. Finally, we include a small-scale stress test in which we intentionally violate contraction controls (e.g. decreasing τ and removing norm bounds) to document qualitative failure modes and to substantiate the necessity of explicit restrictions for reliable equilibrium computation in cyclic settings.

10 Discussion and Limitations

Our design imposes explicit contractivity controls (via damping, temperature, and operator-norm bounds) in order to obtain a unique equilibrium and a predictable iterative solve. This raises an immediate tension between *expressivity* and *guaranteed convergence*. On the one hand, allowing unconstrained message amplification, sharp routing (small τ), or highly state-sensitive attention scores can represent rich, potentially multi-stable dynamics on cyclic directed graphs. On the other hand, such regimes can destroy contraction and thereby forfeit both uniqueness of the equilibrium and algorithmic reliability of Picard iteration (and, in the worst case, tractability of fixed-point computation). We view the contractive formulation not as a claim that all useful cyclic reasoning must be contractive, but rather as a disciplined subset in which equilibrium computation can be made routine, monitored, and robust.

Expressivity under contraction. A standard concern is that contraction precludes modeling sharp, long-range, or resonant interactions in graphs with feedback. We emphasize two counterpoints. First, contraction bounds are typically enforced on a *single layer operator* in a specific norm, and can be relaxed at the architecture level via composition: stacking multiple contractive implicit layers, interleaving with feed-forward (non-implicit) transformations, or using multi-head message components can increase representational capacity while keeping each equilibrium computation well-posed. Second, the relevant question is not whether the map is contractive for *all* inputs and parameters, but whether training discovers parameters that yield stable equilibria on the data distribution. In practice, the admissible set determined by α, τ, and $\|W_m\|_2$ may still contain models that route information directionally through $\beta(h^*)$ in a data-dependent way; contraction limits sensitivity, but does not force trivial uniform attention.

Sufficient versus necessary conditions. The explicit constants appearing in our contraction analysis are sufficient and generally conservative. They are derived from worst-case Lipschitz bounds for the softmax normalization over $N_{\text{out}}(j)$, the Lipschitzness of e_θ, and degree-dependent aggregation factors under $\|\cdot\|_{\infty,2}$. None of these steps is tight in general. For example, (i)

attention logits may occupy a regime where the softmax Jacobian is much smaller than its worst-case upper bound; (ii) the effective out-degree may be substantially below $\Delta_{\max}$ once routing concentrates; and (iii) the composition of maps may exhibit cancellation not captured by submultiplicative norm bounds. Consequently, models may converge reliably even when the stated inequality $\alpha L_\Phi < 1$ is violated by the bound, and conversely may exhibit slow convergence in regimes that formally satisfy a loose bound with $q = \alpha L_\Phi$ close to 1. We therefore interpret the theory as providing *design guidance* and *certificates of stability* rather than an exact characterization of all stable parameterizations.

Failure modes beyond the theory. Even when the map is contractive in principle, several practical failure modes remain.

(i) Near-critical slowing. If $q = \alpha L_\Phi$ is close to 1, convergence is linear but slow, and iteration counts needed to reach a given ε can be large. This affects both inference latency and the conditioning of backward implicit solves, which scale with $(1-q)^{-1}$. In such regimes, it may be preferable to increase τ (smoother routing), reduce $\|W_m\|_2$, or choose smaller effective hidden dimension d for the implicit block while shifting expressivity to explicit feed-forward components.

(ii) Sharp-routing instability. Small τ induces high sensitivity of $\beta_{ij}(h)$ to changes in the scores. If the scoring network is not controlled (e.g. an MLP with large Lipschitz constant), the induced operator may behave like a switching system, leading to oscillatory Picard trajectories or apparent convergence to different equilibria under small perturbations. While damping can mitigate such effects, it may simultaneously erase the benefits of sharp routing by effectively averaging across iterations. Monitoring residual trajectories and attention entropies provides a basic diagnostic.

(iii) Numerical and implementation issues. Outgoing-normalized softmax can overflow for large logits; stable implementations require the standard log-sum-exp trick per sender j. Further, graphs with $|N_{\text{out}}(j)| = 0$ require a convention (e.g. no outgoing mass and hence no contribution to any receiver) that should be consistent across iterations. Finally, stopping rules based on $r^t = \|h^{t+1} - h^t\|_{\infty,2}$ can be deceived by finite-precision plateaus; one may supplement with a maximum iteration budget and, when needed, check $\|F_\theta(h^t) - h^t\|_{\infty,2}$ directly.

Limitations of conservation constraints. Outgoing normalization yields an exact conservation law per sender and endows $\beta(h)$ with a row-stochastic interpretation. This is appropriate when attention is intended to model routing of a conserved quantity or flow-like influence. However, some domains require *amplification* or *attenuation* at nodes, or require multiple simultaneously conserved commodities. While amplification can be represented in part

by W_m and by additive exogenous terms $W_x x_i + b$, the strict row-stochastic structure may be restrictive when the true mechanism is not approximately conservative. Extensions include adding learnable per-sender gates (with explicit bounds to preserve contractivity), or representing multiple channels with separate conservation constraints.

Deployment considerations for real-time infrastructure. For power and traffic applications, the principal operational constraint is predictable latency under distribution shift. Implicit equilibrium computation is attractive insofar as it admits a monitored, anytime solve: we can stop when $r^t \leq \varepsilon$ or fall back to a fixed iteration budget T with a known bound on approximation error when the contraction factor is certified. Nevertheless, several deployment-specific issues remain.

(i) Real-time budgets and worst-case graphs. The per-iteration cost scales with m, and the required number of iterations depends on q and the desired ε. In systems with strict deadlines, one must choose (α, τ, κ_m) to keep q comfortably below 1 under expected operating conditions, or else accept a coarser ε with a quantified error–latency tradeoff.

(ii) Nonstationarity and topology changes. Infrastructure graphs change (line outages, road closures). Stability bounds suggest that small perturbations in features yield controlled perturbations in h^* when $q < 1$, but they do not address abrupt topology edits that change $\Delta_{\max}$ or invalidate calibrated norm/temperature choices. Practical systems should include automated checks for degree changes and revalidation of convergence behavior.

(iii) Safety and interpretability. Although the routing kernel $\beta(h^*)$ is a well-defined distribution, its semantics depend on the learned scoring function. For safety-critical contexts, it is not sufficient that the model converges; one must additionally ensure that learned routing respects domain constraints (e.g. forbidding attention along de-energized lines). Such constraints can be enforced by masking edges prior to normalization; the theory remains applicable provided masking is treated as a fixed graph restriction during the solve.

In summary, contraction is a deliberate restriction that yields clear convergence and stability guarantees, but it does not eliminate the need for empirical monitoring, careful numerical implementation, and domain-specific constraint handling, particularly in real-time cyclic infrastructure settings.