Synthetic Data Generation for Anomaly Detection

Johan Slettengren[1]

[1]KTH Royal Institute of Technology, Sweden (email: johansle@kth.se)

Problem Setup

- Want to apply machine learning to detect anomalies, e.g. leaks.
- Anomalies are rare so we must create **synthetic data**.
- Traditional method need one simulation per scenario—scales poorly!

Water Pipe Networks

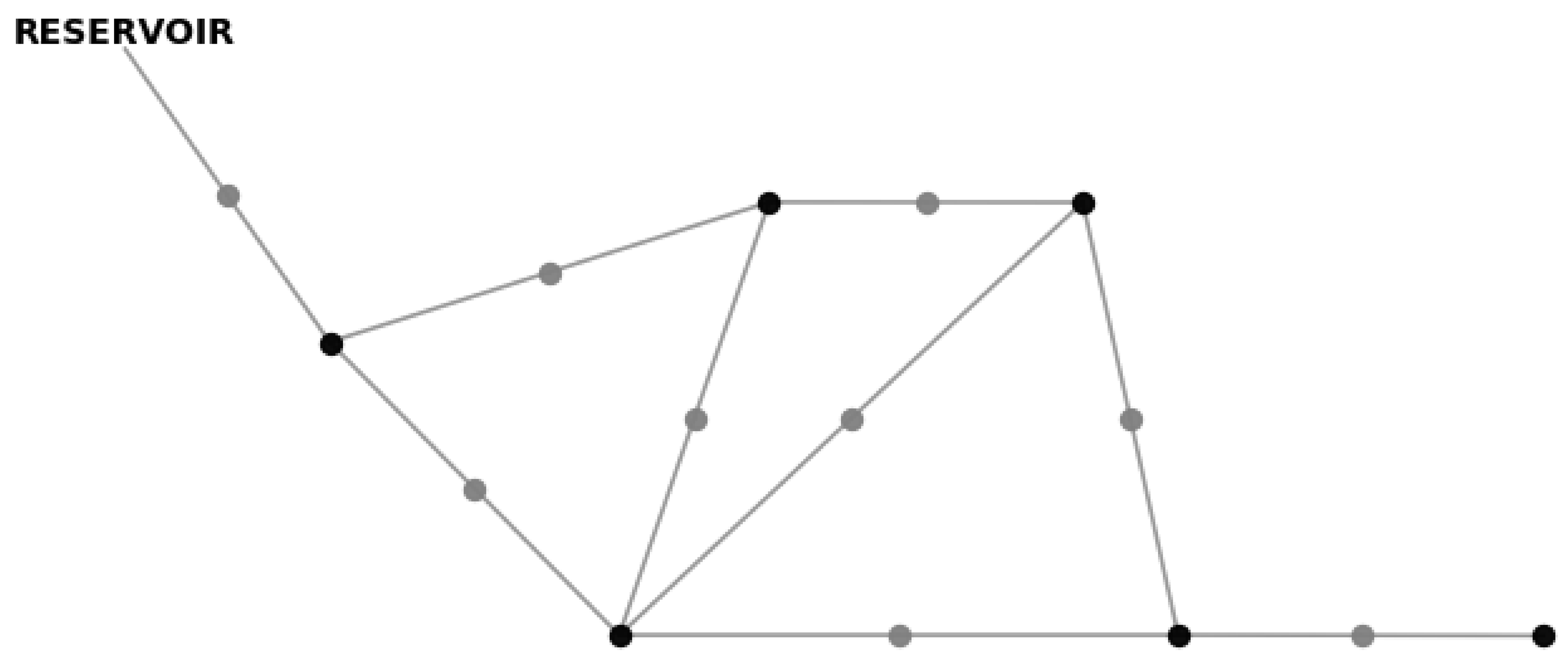


Figure: Pipe network example with junctions (•), leak nodes (•) and pipes (—)

Governing System of Equations

$$\mathcal{A}_0 \mathbf{Q} = \mathbf{D} + d_{leak}(\mathbf{H}) \quad (1a)$$

$$\mathcal{A}_0^T \mathbf{H} = \mathcal{B}\mathbf{S} - h_L(\mathbf{Q}) \quad (1b)$$

$\mathbf{Q}$: pipe flows
$\mathbf{H}$: junction hydraulic heads
$\mathcal{A}_0$: reduced incidence-matrix
$\mathbf{D}$: junction flow demands
$\mathbf{S}$: reservoirs hydraulic heads
d_{leak}: leak mechanism
h_L: friction head loss
$\mathcal{B}$: maps reservoirs to connected nodes

Scenario Parameters

- Leak node(s)
- Demand $\mathbf{D}$
- Leak location along pipe
- Leak area
- Leak discharge coefficient

Physics Informed Neural Network

- **Neural network**: parametrized "black box" function.
- **Physics informed loss**: residuals of (1).
- Scenario parameters sampled randomly—no discretization needed!

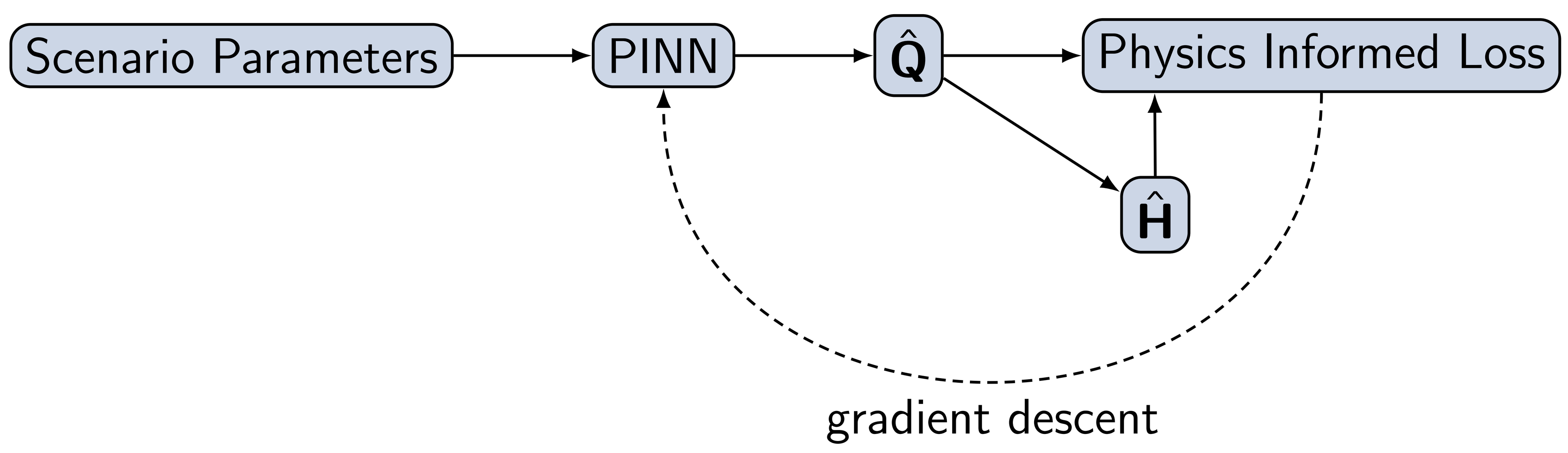


Figure: PINN training

Useful Trick

Avoid predicting $\mathbf{H}$ by setting $\hat{\mathbf{H}} = (\mathcal{A}_0^T)^+ \left(\mathcal{B}\mathbf{S} - h_L(\hat{\mathbf{Q}})\right)$.

! Must still enforce (1b) via loss.

Amortized Data Generation

- Neural network handles multiple scenarios simultaneously.
- Gives **amortized data generation**—scales well!

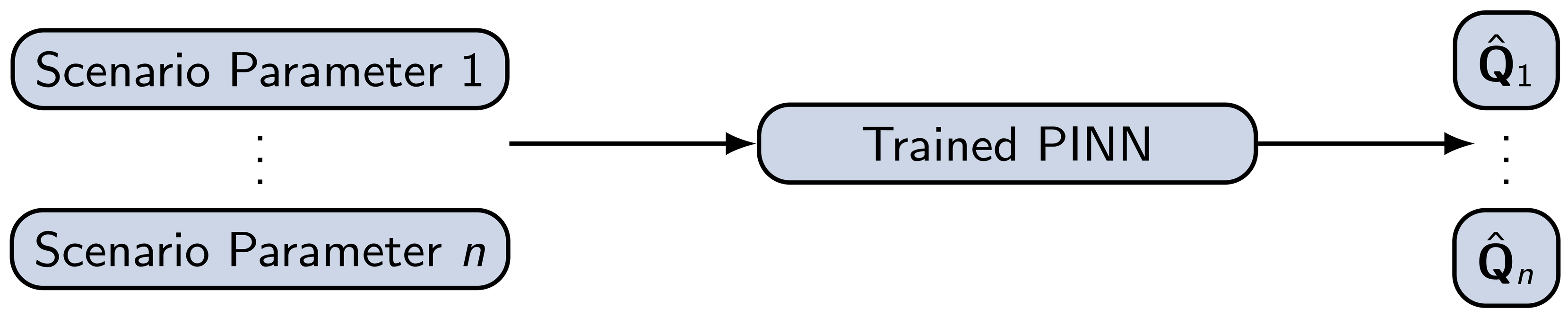


Figure: Amortized data generation via PINN

Experimental Results

- Predict flows based on leak node index and $\mathbf{D} \in [0,1]^{\#\text{ demand junctions}}$.
- Neural network trained until 90% of predictions have NRMSD $\leq$ 10 %.

$$\texttt{NRMSD} = \frac{\sqrt{\texttt{mean}(\hat{\mathbf{Q}} - \mathbf{Q})^2}}{\texttt{max}\mathbf{Q} - \texttt{min}\mathbf{Q}}$$

Experiment 1

#Points / axis	Simulator time (s)	PINN time (s)	Mean NRMSD
5	5.90	13.00	6.8%
10	40.91	8.36	7.3%
15	136.10	10.74	7.6%

Table: Simulation times for different demand grid resolutions (#demand junctions = 3).

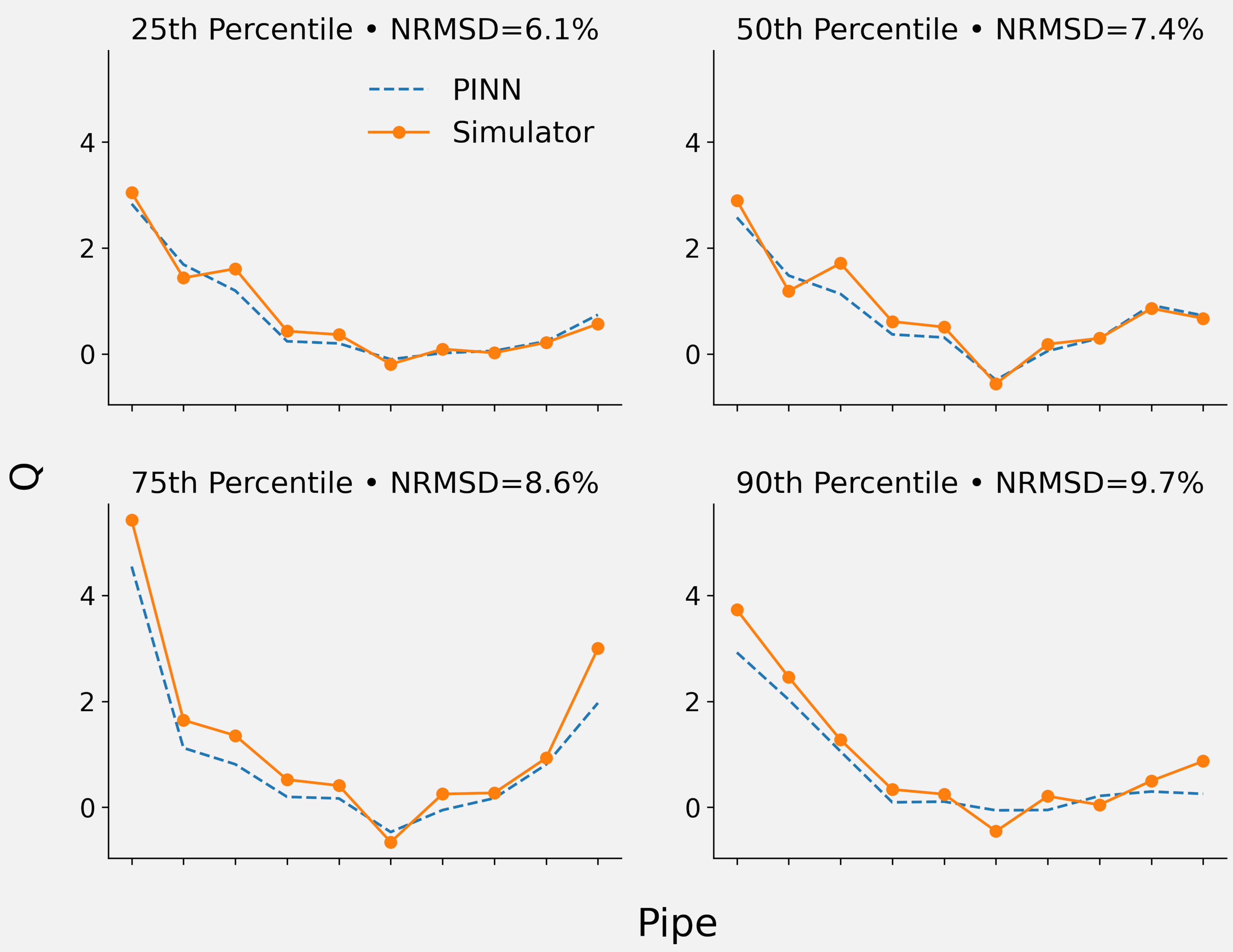
Experiment 2

#Demand junctions	Simulator time (s)	PINN time (s)	Mean NRMSD
3	5.60	12.97	6.8%
4	25.25	12.16	5.8%
5	119.19	8.90	6.4%

Table: Simulation times for different numbers of demand junctions (#points / axis = 5).

Visualized Accuracy

Figure: Predictions in different percentiles of NRMSD
(#points / axis = 15, #demand junctions = 3)



Future Work

- Increase complexity: more scenario parameters and larger pipe networks.
- Add network components, e.g. valves, tanks and pumps.
- Expand to similar systems, e.g. gas and electricity.