Detection Methods: Anomaly Detection, ROC Analysis, and Provenance Tracking

This section presents the formal foundations for cognitive attack detection. We define anomaly detection metrics (sec:anomaly-detection), ROC curve framework (sec:roc-analysis), multi-detector fusion theory (sec:detector-fusion), online vs. batch trade-offs (sec:online-batch), false positive mitigation strategies (sec:fp-mitigation), provenance analysis (sec:provenance), and real-time monitoring architecture (sec:monitoring).

> ***Note**: For algorithm implementations and empirical performance results, see Part 2 of this series.*

Anomaly Detection

Cognitive Drift Scoring

Definition (Drift Score)

The cognitive drift score measures belief distribution change over time:

$$S_{\text{drift}}(t) = D_{\text{KL}}(\mathcal{B}_i^t \| \mathcal{B}_i^{t-w}) + \lambda \cdot \max_{\phi} |\Delta \mathcal{B}_i(\phi)| \tag{1}$$