No-Regret Contextual Bandits for Cost-Sensitive Decision-Making

Public agencies like city governments face sequential, cost-sensitive choices under partial feedback, for example, deciding whether to *inspect* or *not inspect* a construction permit given categorical descriptors, spatial coordinates, and stage metadata. There are operational costs of doing inspections and thus doing all possible inspections is excessively costly. We frame this as a contextual bandit problem and ask: Can regret-minimizing online policies reduce cumulative cost without much hyperparameter tuning when conditions drift or become strategic?

We use a cost model with two parts: a fixed inspection cost and a high miss cost when a problematic case is skipped. To reflect real workloads, a backlog-aware surcharge raises effective inspection cost under heavy recent load and decays as the queue clears, making losses history-dependent and inducing endogenous non-stationarity. Evaluation combines a large administrative dataset with synthetic data generated from a calibrated Bayesian statistical model.

Methods span two families with formal guarantees. (1) EXP4 mixes experts via multiplicative weights and inverse–propensity loss estimates, with distribution-free regret $O(\sqrt{TK \log N})$ against the best expert (rounds T, actions K, experts N). (2) ILTCB / ILOVETOCONBANDITS reduces to cost-sensitive classification and, in the stochastic i.i.d. contextual-bandit setting, achieves high-probability policy regret $\tilde{O}(\sqrt{TK \log |\Pi|})$ against the best policy in a class Π (oracle-efficient reduction with a decaying exploration floor to control IPS variance). Online-Cover is the practical variant that maintains a small cover of policies to approximate ILTCB's behavior. Supervised references are cost-sensitive classifiers with randomized exploration and windowed retraining.

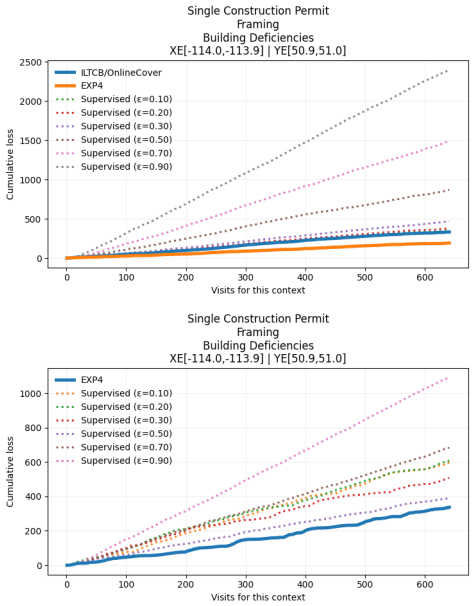

Figure 1: Cumulative loss in a context with a stochastic (top) vs. adversarial (bottom) generator.

For our experiments, we simulate both stochastic and adversarial models. Across backlog regimes and cost settings, EXP4 and ILTCB / Online-Cover achieve lower cumulative cost than supervised baselines. Bandit learners track regime changes and retain performance when a latent agent (adversarial builder) cuts corners as inspections wane. Per-context cumulative-loss trajectories show switching optimal arms and non-constant slopes, visualizing drift and strategic feedback.

EXP4 dominates across cost models and under adversaries. Adversarial robustness is achieved because multiplicative-weights concentrates mass on low-estimated-loss experts while preserving minimum action probabilities, so guarantees hold without stationarity or realizability. EXP4 depends only on normalized per-round losses; shifting the cost composition changes the scale, not the update. It is also caple of rapid regime tracking - persistent loss gaps tilt weights exponentially, moving probability mass quickly when the optimal arm switches. By mixing diverse experts, EXP4 adapts locally without brittle window/epsilon schedules.

Thus the contributions of this work is as follows: (a) A backlog-aware cost model that induces realistic, endogenous non-stationarity for civic workflows. (b) A multi-agent adversarial extension that stresses robustness. (c) A head-to-head evaluation of adversarial and oracle-efficient contextual bandits against supervised baselines under partial feedback. (d) Interpretable diagnostics using per-context loss curves that help stakeholders understand when and why bandit policies excel. For resource-constrained operations (inspections, service dispatch, audits), regret-minimizing bandits deliver robust, low-maintenance decision rules that natively handle exploration, drift, partial observability, and multi-agent feedback.

Keywords: contextual bandits; online learning; partial feedback; non-stationarity; cost-sensitive decision-making; adversarial robustness.