Minimax Optimal Estimation of Stability Under Distribution Shift

Hongseok Namkoong*, **Yuanzhe Ma**†, and Peter W. Glynn‡

*Decision, Risk, and Operations Division, Columbia University

†Department of Industrial Engineering and Operations Research, Columbia University

‡Management Science and Engineering, Stanford University

Introduction

Evaluating stability

- Distribution shift (training $\neq$ test distributions) often happens, leading to large performance degradation of machine learning models.
- Prior to deployment, it's important to evaluate the stability of machine learning models in an **interpretable** way.

Problem set-up

- Given a random cost function denoted by R, we have i.i.d. scenarios with associated costs $R_1, \ldots, R_n \overset{\text{iid}}{\sim} P$.
- For a chosen threshold $y \geq \mathbb{E}_P[R]$, the **stability** of the system is defined as the smallest distribution shift in the underlying environment that deteriorates performance above the threshold y using the Kullback–Leibler divergence:

$$I_y(P) := \inf_{Q} \left\{ D_{\text{kl}}(Q \| P) : \mathbb{E}_Q[R] \geq y \right\}.$$

- It is more intuitive for data analysts and engineers to decide on a tolerable amount of monetary loss or prediction error in the **cost scale** based on past data and domain knowledge. Our approach is similar to sensitivity analysis in the causal inference literature.

Approach

An estimator based on dual formulation of $I_y(P)$

- The above optimization problem is a challenging problem that optimizes over probability measures.
- We use a duality result by Donsker and Varadhan in 1976:

$$I_y(P) = \sup_{\lambda \in \mathbb{R}} \left\{ \lambda y - \log \mathbb{E}_P[e^{\lambda R}] \right\}.$$

- Our dual plug-in estimator is given by

$$\widehat{I}_n := \sup_{\lambda \in \mathbb{R}} \left\{ \lambda y - \log \mathbb{E}_{\widehat{P}_n}[e^{\lambda R}] \right\},$$

  where $\widehat{P}_n$ is the empirical distribution over the data $R_1, \ldots, R_n$.
- We can efficiently solve the above convex optimization problem using binary search in one dimension.
- $I_y(P)$ is also called the large deviations rate function, and our estimator can be used for estimating probabilities of rare events.

Theory

Assumptions

We study the minimax rate of convergence for estimating $I_y(P)$ over a natural class of distributions $\mathcal{P}_{\sigma,y,\alpha}$ with Gamma-like tail behavior, which contains distributions P satisfying:

1. $R \geq 0$ and $\mathbb{E}_P[e^{\sigma R}] = \infty$.
2. $\mathbb{E}_P[e^{\lambda R}] \leq \frac{\sigma}{\sigma - \lambda}$ for $0 \leq \lambda < \sigma$ and $\mathbb{E}_P[R] \leq \frac{1}{\sigma}$.
3. $\lambda^\star(P) = \operatorname{argmax}_\lambda \left\{ \lambda y - \log \mathbb{E}_P[e^{\lambda R}] \right\} \leq \sigma - \frac{\alpha}{y} =: \bar{\lambda}$.

Main results of the minimax rate of convergence

Our results characterize ($\asymp$ hides polylogarithmic factors in n and constants)

$$\inf_{\widehat{I}_n} \sup_{P \in \mathcal{P}_{\sigma,y,\alpha}} \mathbb{E}_P \left| \widehat{I}_n - I_y(P) \right| \asymp n^{-\left(\frac{1}{2} \wedge \frac{\alpha}{\sigma y}\right)},$$

where $x \wedge y := \min\{x, y\}$. In particular, our estimator $\widehat{I}_n$ achieves the rate $n^{-\left(\frac{1}{2} \wedge \frac{\alpha}{\sigma y}\right)}$, which implies it is minimax.

Proof outline

- Upper bound: use chaining techniques to uniformly bound the empirical process $\lambda \mapsto \mathbb{E}_{\widehat{P}_n}[e^{\lambda R}] - \mathbb{E}_P[e^{\lambda R}]$ over the interval $[0, \bar{\lambda}]$, which in turn bounds the estimation error $\widehat{I}_n - I$.
- Lower bound: Le Cam's method: construct $P_1, P_2 \in \mathcal{P}_{\sigma,y,\alpha}$ with small $\|P_1 - P_2\|_{\text{TV}}$ but large $|I_y(P_1) - I_y(P_2)|$.
  Two cases:
  1. Fix $\alpha \in (\frac{1}{2}, 1)$ and let $\sigma y \geq 2\alpha > 1$:

  $$f_1(x) \propto x^{\alpha + \frac{1}{\sigma x_0} - 1} e^{-\sigma x} \mathbf{1}\{x \geq 0\},$$

  $$f_2(x) \propto \begin{cases} x^{\alpha + \frac{1}{\sigma x_0} - 1} e^{-\sigma x} & \text{if } 0 \leq x \leq x_0 \\ x^{-1} e^{-\sigma x} & \text{if } x > x_0 \end{cases}$$

  with $x_0 \approx \frac{1}{\sigma} \log n$.
  2. Fix $\alpha \in (\frac{1}{2}, 1)$ and let $2\alpha > \sigma y > 1$: $f_1(x) = \sigma e^{-\sigma x} \mathbf{1}\{x \geq 0\}$ and

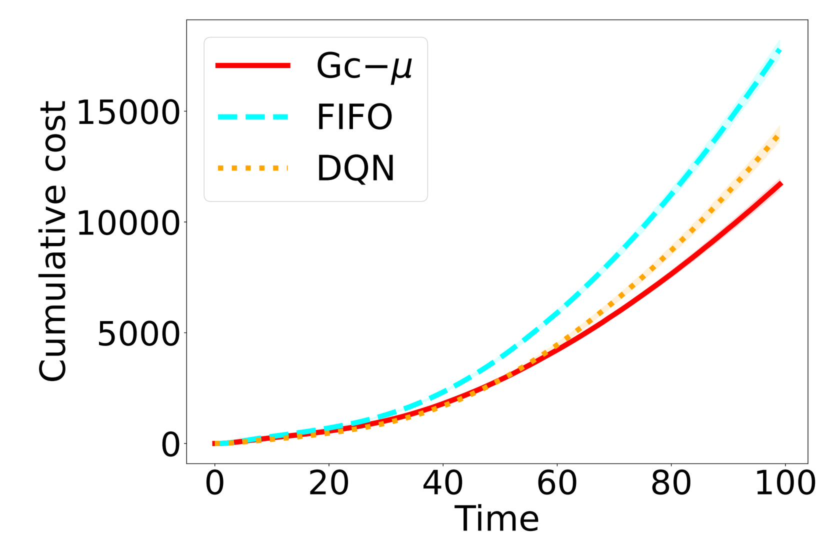
  $$f_2(x) = \begin{cases} 0 & \text{if } x < 0 \\ \sigma(1+\omega) e^{-\sigma(1+\omega)x} & \text{if } 0 \leq x \leq x_0 \\ \sigma e^{-\sigma \omega x_0} e^{-\sigma x} & \text{if } x > x_0 \end{cases}$$

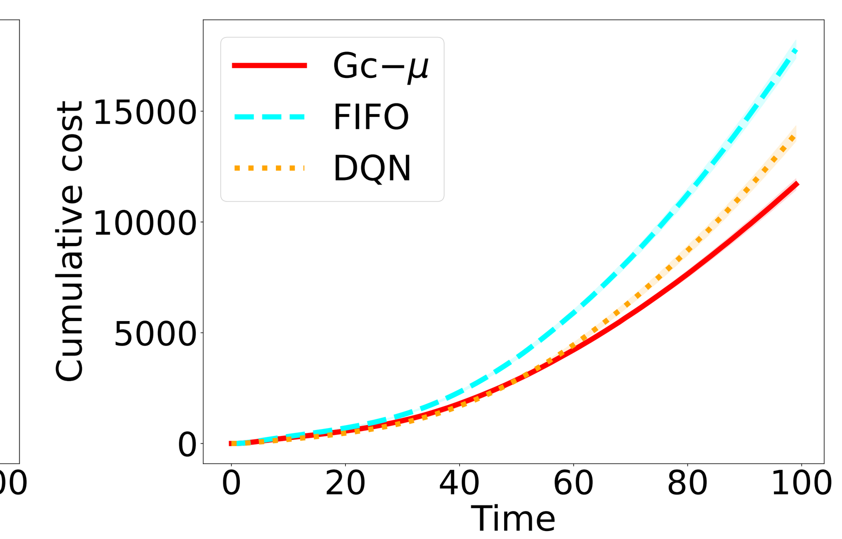
  with $x_0 = \frac{1}{\sigma} \log n$, $\omega \leq \frac{1-\alpha}{\sigma y} \wedge \frac{2 - \sigma y}{\sigma y}$.

Experiments

Sequential decision-making

- A queueing control problem with a G/G/1 queue with multi-class jobs.
- Three policies:
  1. Gc-μ: simple, index-based, optimal in a limiting regime, enjoys a natural adversarial robustness
  2. Deep Q-learning (DQN): empirically good but robustness not guaranteed
  3. FIFO: a simple benchmark
- Our stability measure suggests Gc-μ is much more robust than DQN, which is also confirmed in our simulations of concrete distribution shifts.





Supervised learning

- A problem of health utilization prediction using a supervised dataset with $X \in \mathbb{R}^{396}$ and $Y \in \{0, 1\}$.
- Take the viewpoint of an analyst who trains a model in 2015.
- According to our stability measure, the LightGBM model is significantly less stable than the other two models.
- We verify that the performance of the LightGBM model substantially degrades over time. Reason: LightGBM overly relies on a covariate that has a large distribution shift over time.