Local Glivenko-Cantelli

Doron Cohen DORONV@POST.BGU.AC.IL and **Aryeh Kontorovich** KARYEH@CS.BGU.AC.IL
Department of Computer Science
Ben-Gurion University of the Negev
Beer-Sheva, Israel

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

If μ is a distribution over the d-dimensional Boolean cube $\{0,1\}^d$, our goal is to estimate its mean $p \in [0,1]^d$ based on n iid draws from μ. Specifically, we consider the empirical mean estimator $\hat{p}_n$ and study the expected maximal deviation $\Delta_n = \mathbb{E} \max_{j \in [d]} |\hat{p}_n(j) - p(j)|$. In the classical Universal Glivenko-Cantelli setting, one seeks distribution-free (i.e., independent of μ) bounds on Δ_n. This regime is well-understood: for all μ, we have $\Delta_n \lesssim \sqrt{\log(d)/n}$ up to universal constants, and the bound is tight.

Our present work seeks to establish dimension-free (i.e., without an explicit dependence on d) estimates on Δ_n, including those that hold for $d = \infty$. As such bounds must necessarily depend on μ, we refer to this regime as *local* Glivenko-Cantelli (also known as μ-GC), and are aware of very few previous bounds of this type — which are either "abstract" or quite sub-optimal. Already the special case of product measures μ is rather non-trivial. We give necessary and sufficient conditions on μ for $\Delta_n \to 0$, and calculate sharp rates for this decay. Along the way, we discover a novel sub-gamma-type maximal inequality for shifted Bernoullis, of independent interest.

Extended abstract. Full version appears as arxiv.org/abs/2209.04054v4.

Keywords: empirical process, Bernoulli, concentration, Bernstein inequality, maximal inequality