Universal Hypothesis Testing via Mismatched Divergence

Jayakrishnan Unnikrishnan, Dayu Huang, Sean Meyn, Venu Veeravalli (Univ. of Illinois at Urbana-Champaign)

Amit Surana (UTRC)

Abstract

Optimal solution to the universal hypothesis testing problem suffers from high variance for large alphabet distributions. We propose a new approach to this problem that addresses this issue. Our solution is based on the *mismatched divergence* which is a new lower bound on Kullback-Leibler divergence (i.e., relative entropy). We present results on the asymptotic statistics of our test statistic and geometry of our mismatched test.

Universal Hypothesis Testing

- Observations $\boldsymbol{Z} = \{Z_t : t = 1, \ldots\}$ i.i.d. sequence evolving on $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$
- Objective: Determine if marginal distribution is π^0 or some other distribution
- For given alternate distribution π^1, define error-exponents

$$J_\phi^0 := \liminf_{n\to\infty} -\frac{1}{n}\log(\mathsf{P}_{\pi^0}\{\phi_n(Z_1,\ldots,Z_n) = 1\}),$$
$$J_\phi^1 := \liminf_{n\to\infty} -\frac{1}{n}\log(\mathsf{P}_{\pi^1}\{\phi_n(Z_1,\ldots,Z_n) = 0\}).$$

- Criterion for optimizing test sequence $\{\phi_n\}$:

$$\beta_\eta^* = \sup\{J_\phi^1 : \ \textit{subject to } J_\phi^0 \geq \eta\} \tag{1}$$

We seek universal solution that is oblivious of π^1

Hoeffding's Universal Solution

- Define empirical distribution (*type*): $\Gamma^n(a) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{Z_i = a\}, \quad a \in \mathcal{A}$
- Hoeffding's Test [Hoe65], [ZG91]:

$$\varphi_n^{\text{H}}(Z_1, Z_2, \ldots, Z_n) = \mathbb{I}\{D(\Gamma^n\|\pi^0) \geq \eta\}$$

- Solution unreliable for large alphabet size (N) and moderate sequence length (n)
- Suppose $Z_i \sim \pi^0$, then $D(\Gamma^n\|\pi^0) \to 0$, and,

$$\lim_{n\to\infty} \mathsf{E}[nD(\Gamma^n\|\pi^0)] = \tfrac{1}{2}(N-1) \tag{2}$$
$$\lim_{n\to\infty} \mathsf{Var}\,[nD(\Gamma^n\|\pi^0)] = \tfrac{1}{2}(N-1) \tag{3}$$
$$nD(\Gamma^n\|\pi^0) \xrightarrow[n\to\infty]{d.} \tfrac{1}{2}\chi^2_{N-1} \tag{4}$$

- High bias (2) is easily addressed by varying threshold since the bias is high even under the alternate distribution:

Suppose $Z_i \sim \pi^1$, then $D(\Gamma^n\|\pi^0) \to D(\pi^1\|\pi^0)$, and,

$$\lim_{n\to\infty} \mathsf{E}[n(D(\Gamma^n\|\pi^0) - D(\pi^1\|\pi^0))] = \tfrac{1}{2}(N-1)$$

How can we reduce the high variance?

Mismatched Divergence

Notation: For distribution π and function f,

$$\pi(f) = \sum_{i=1}^N f(a_i)\pi_i \qquad \Lambda_\pi(f) = \log(\pi(e^f))$$

Variational representation of divergence as a supremum over functions

$$D(\nu_1\|\nu_2) = \sup_f \big(\nu_1(f) - \Lambda_{\nu_2}(f)\big)$$

Supremum is achieved by the LLR function $\log\frac{\nu^1}{\nu^2}$

- Mismatched divergence is a relaxation

$$D^{\text{MM}}(\nu_1\|\nu_2) := \sup_{f\in\mathcal{F}} \big\{\nu_1(f) - \Lambda_{\nu_2}(f)\big\}$$

where $\mathcal{F}$ is a d-dimensional function class $\mathcal{F} = \{f_r : r \in \mathbb{R}^d\}$

- Mismatched test:

$$\phi_n^{\text{MM}}(Z_1, Z_2, \ldots, Z_n) = \mathbb{I}\{D^{\text{MM}}(\Gamma^n\|\pi^0) \geq \eta\}$$

If $\mathcal{F}$ contains $\{r\log\frac{\pi^1}{\pi^0} : r > 0\}$, then $\{\phi_n^{\text{MM}}\}$ achieves the supremum $J_{\phi^{\text{MM}}}^1 = \beta_\eta^*$ in (1)

Prior information regarding alternate hypotheses can be utilized in designing function class

Properties of Mismatched Divergence

- Consider parametric family of distributions

$$\mathbb{E} := \{\check{\pi}^r : r \in \mathbb{R}^d\} \qquad \text{where} \qquad \check{\pi}^r = \pi\exp(f_r - \Lambda_\pi(f_r))$$

We have a generalized Pythagorean equality:

$$D^{\text{MM}}(\nu\|\pi) = D(\nu\|\pi) - D(\nu\|\check{\pi}(\nu))$$

where $\check{\pi}(\nu)$ is the reverse I-projection of ν onto $\mathbb{E}$, i.e.,

$$\check{\pi}(\nu) = \arg\min_{\mu\in\mathbb{E}} D(\nu\|\mu)$$

For $\nu = \Gamma^n$, $\check{\pi}$ is the ML estimate of underlying distribution from $\mathbb{E}$

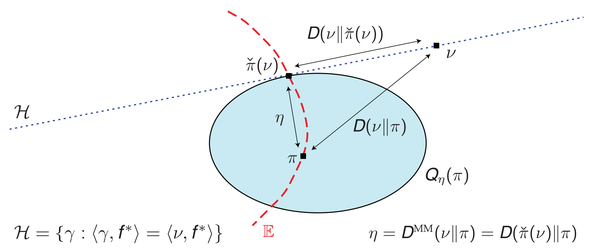
- For log-linear class mismatched divergence is a robust rate function

$$D^{\text{MM}}(\nu\|\pi) = \inf_{\mu\in\mathbb{P}} D(\nu\|\mu), \qquad \text{where } \mathbb{P} \text{ is some linear family of distributions}$$

Framework unifies ideas from statistics and information theory

Geometry of Mismatched Divergence for Linear Function Class

For linear class $\mathcal{F} = \{\sum_i r_i\psi_i : r \in \mathbb{R}^d\}$, we also have $D^{\text{MM}}(\nu\|\pi) = D(\check{\pi}(\nu)\|\pi)$



Asymptotics of Mismatched Test Statistic

- If $Z_i \sim \pi^0$ and $\mathcal{F}$ is d-dimensional, we have

$$\lim_{n\to\infty} \mathsf{E}[nD^{\text{MM}}(\Gamma^n\|\pi^0)] = \tfrac{1}{2}d$$
$$\lim_{n\to\infty} \mathsf{Var}\,[nD^{\text{MM}}(\Gamma^n\|\pi^0)] = \tfrac{1}{2}d$$
$$nD^{\text{MM}}(\Gamma^n\|\pi^0) \xrightarrow[n\to\infty]{d.} \tfrac{1}{2}\chi^2_d$$

- If $Z_i \sim \pi^1$ and $\mathcal{F}$ contains LLR function $L = \log\frac{\pi^1}{\pi^0}(.)$. Then we have, with $\sigma_1^2 := \mathsf{Cov}_{\pi^1}(L)$,

$$\lim_{n\to\infty} \mathsf{E}[n(D^{\text{MM}}(\Gamma^n\|\pi^0) - D(\pi^1\|\pi^0))] = \tfrac{1}{2}d$$
$$\lim_{n\to\infty} \mathsf{Var}\,[n^{\frac{1}{2}}D^{\text{MM}}(\Gamma^n\|\pi^0)] = \sigma_1^2$$
$$n^{\frac{1}{2}}(D^{\text{MM}}(\Gamma^n\|\pi^0) - D(\pi^1\|\pi^0)) \xrightarrow[n\to\infty]{d.} \mathcal{N}(0, \sigma_1^2).$$

- Bias and variance also computable when $L \notin \mathcal{F}$

Lower bias and variance for large alphabet size
Tradeoff between power of test and bias-variance performance

References

[Hoe65] W. Hoeffding. Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.*, 36:369–408, 1965.

[ZG91] O. Zeitouni and M. Gutman. On universal hypotheses testing via large deviations. *IEEE Trans. Inform. Theory*, 37(2):285–290, 1991.