# Beyond KL Divergence: Divergence Tests for Universal Hypothesis Testing

K V Harsha

BITS Pilani
Hyderabad Campus

Jithin Ravi

Indian Institute of Technology
Kharagpur

Tobias Koch

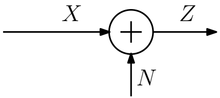
Universidad Carlos III de
Madrid

CNI Seminar Series

IISc Bangalore, 21 September 2026

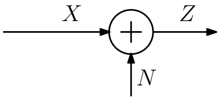
# Binary Detection

- Transmit a bit $b \in \{0, 1\}$ in the presence of Gaussian noise
- Received symbol $Z = X + N$, where $N \sim \mathcal{N}(0, 1)$

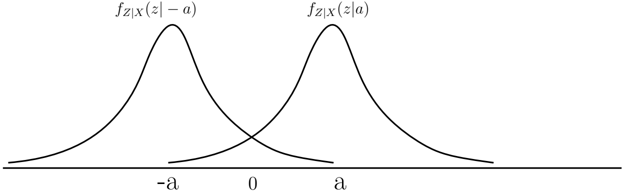


# Binary Detection

- Transmit a bit $b \in \{0, 1\}$ in the presence of Gaussian noise
- Received symbol $Z = X + N$, where $N \sim \mathcal{N}(0, 1)$

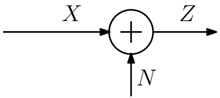



- Use BPSK signalling: $0 \to X = -a$ and $1 \to X = a$

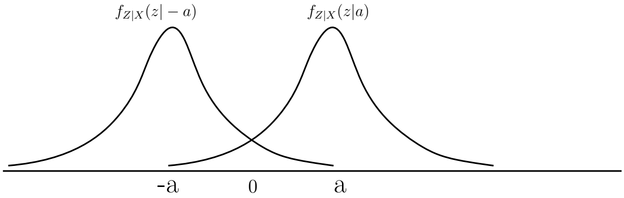


# Binary Detection

- Transmit a bit $b \in \{0, 1\}$ in the presence of Gaussian noise
- Received symbol $Z = X + N$, where $N \sim \mathcal{N}(0, 1)$



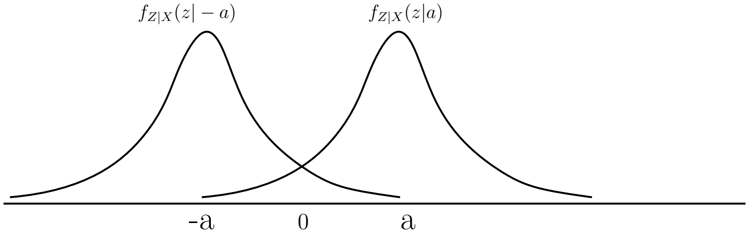

- Use BPSK signalling: $0 \to X = -a$ and $1 \to X = a$




- Under hypothesis $H_0$: $Z \sim \mathcal{N}(-a, 1)$ and under $H_1$: $Z \sim \mathcal{N}(a, 1)$

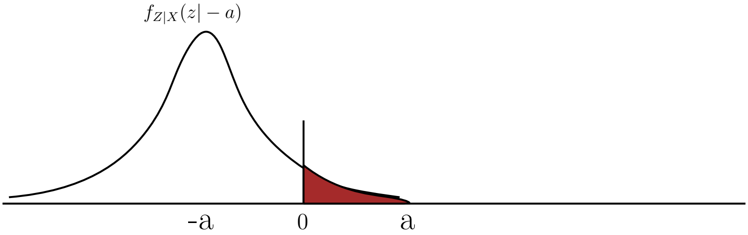
# Likelihood Ratio (LR) Test

- Likelihood ratio test (LRT): If $\frac{P(z)}{Q(z)} \geq r$ (for some threshold $r$), declare $H_0$.
- One instance: If $Z \geq 0$, declare $X = a$ ($\hat{H} = H_1$), otherwise $X = -a$ ($\hat{H} = H_0$).



# Likelihood Ratio (LR) Test

- Likelihood ratio test (LRT): If $\frac{P(z)}{Q(z)} \geq r$ (for some threshold $r$), declare $H_0$.
- One instance: If $Z \geq 0$, declare $X = a$ ($\hat{H} = H_1$), otherwise $X = -a$ ($\hat{H} = H_0$).

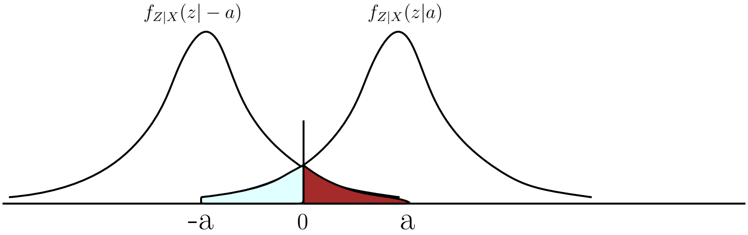



- Two types of errors for test $T$:

$$\text{Type I error:} \qquad \alpha = \Pr(\text{guess } H_1 | H = H_0)$$

# Likelihood Ratio (LR) Test

- Likelihood ratio test (LRT): If $\frac{P(z)}{Q(z)} \geq r$ (for some threshold $r$), declare $H_0$.
- One instance: If $Z \geq 0$, declare $X = a$ ($\hat{H} = H_1$), otherwise $X = -a$ ($\hat{H} = H_0$).



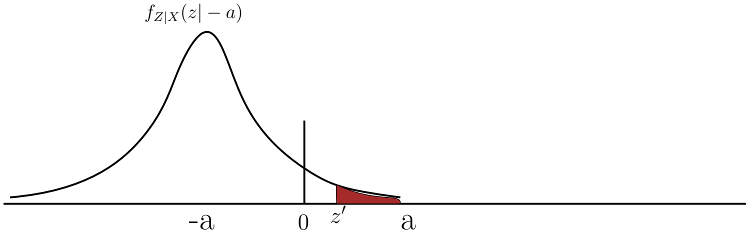

- Two types of errors for test $T$:

$$\text{Type I error:} \qquad \alpha = \Pr(\text{guess } H_1 | H = H_0)$$

$$\text{Type II error:} \qquad \beta = \Pr(\text{guess } H_0 | H = H_1)$$

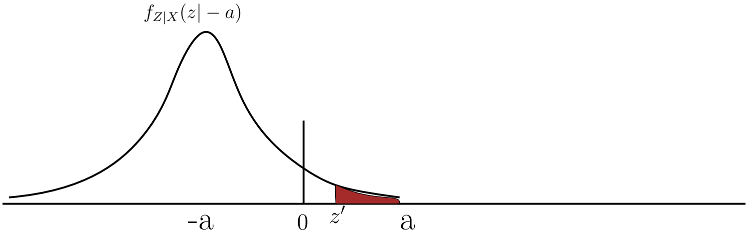
# Constraints on Type I Error

- Suppose we want type I error $\alpha \leq \epsilon$, for $\epsilon \in (0, 1)$



# Constraints on Type I Error

- Suppose we want type I error $\alpha \leq \epsilon$, for $\epsilon \in (0, 1)$




- LRT can be modified as: If $Z \leq z'$, then $H_0$; otherwise $H_1$
- Under $\alpha \leq \epsilon$, study type II error $\beta$

# LR test for i.i.d. random variables

- Also for i.i.d. copies of RV $Z$ denoted by $Z^n = (Z_1, \ldots, Z_n)$

The Likelihood Ratio Test for the setting: $H_0 : Z \sim P$ and $H_1 : Z \sim Q$.

- Declare $H_0$ if $\frac{P^n(z^n)}{Q^n(z^n)} \geq r_n$ (for some $r_n$). Declare $H_1$ if $\frac{P^n(z^n)}{Q^n(z^n)} < r_n$
- Also known as Neyman-Pearson Test[1]

[1] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 694-706, pp. 289–337, Feb. 1933

# LR test for i.i.d. random variables

- Also for i.i.d. copies of RV $Z$ denoted by $Z^n = (Z_1, \ldots, Z_n)$

The Likelihood Ratio Test for the setting: $H_0 : Z \sim P$ and $H_1 : Z \sim Q$.

- Declare $H_0$ if $\frac{P^n(z^n)}{Q^n(z^n)} \geq r_n$ (for some $r_n$). Declare $H_1$ if $\frac{P^n(z^n)}{Q^n(z^n)} < r_n$
- Also known as Neyman-Pearson Test[1]
- LR test can be done for discrete random variables, e.g., testing Bernouli distributions
  1. $H_0$: Bern$(p)$
  2. $H_1$: Bern$(q)$

[1] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 694-706, pp. 289–337, Feb. 1933

# Composite Binary Hypothesis Testing

- Observe $Z^n = (Z_1, \ldots, Z_n)$ - i.i.d. copies of RV $Z \in \mathcal{Z} = \{a_1, \ldots, a_k\}, k \geq 2$

| Different settings | Null hypothesis | vs. | Alternative hypothesis |
|---|---|---|---|
| Classical Setting | $H_0 : Z \sim P$ | | $H_1 : Z \sim Q$ |

- Classical Setting: $H_0$: Bern$(1/3)$ vs. $H_1$: Bern$(2/3)$

# Composite Binary Hypothesis Testing

- Observe $Z^n = (Z_1, \ldots, Z_n)$ - i.i.d. copies of RV $Z \in \mathcal{Z} = \{a_1, \ldots, a_k\}, k \geq 2$

| Different settings | Null hypothesis | vs. | Alternative hypothesis |
|---|---|---|---|
| Classical Setting | $H_0 : Z \sim P$ | | $H_1 : Z \sim Q$ |
| Composite setting | $H_0 : P \in \mathcal{P}$ | | $H_1 : Q \in \mathcal{Q}$ |

- Classical Setting: $H_0$: Bern$(1/3)$ vs. $H_1$: Bern$(2/3)$
- Composite Setting: $H_0$: Bern$(p)$, $p \in (\frac{1}{4}, \frac{1}{3})$ vs. $H_1$: Bern$(q)$, $q \in (\frac{2}{3}, \frac{3}{4})$

# Composite Binary Hypothesis Testing

- Observe $Z^n = (Z_1, \ldots, Z_n)$ - i.i.d. copies of RV $Z \in \mathcal{Z} = \{a_1, \ldots, a_k\}, k \geq 2$

| Different settings | Null hypothesis | vs. | Alternative hypothesis |
|---|---|---|---|
| Classical Setting | $H_0 : Z \sim P$ | | $H_1 : Z \sim Q$ |
| Composite setting | $H_0 : P \in \mathcal{P}$ | | $H_1 : Q \in \mathcal{Q}$ |
| Our focus | $H_0 : \mathcal{P} = \{P\}$ | | $H_1 : \mathcal{Q} = \mathcal{P}^c$ |

$P$ is known $Q$ can be anything except $P$ (Goodness-of-Fit test)

- Classical Setting: $H_0$: Bern$(1/3)$ vs. $H_1$: Bern$(2/3)$
- Composite Setting: $H_0$: Bern$(p)$, $p \in (\frac{1}{4}, \frac{1}{3})$ vs. $H_1$: Bern$(q)$, $q \in (\frac{2}{3}, \frac{3}{4})$
- Our focus: $H_0$: Bern$(1/3)$ vs. $H_1$: Bern$(q)$, $q \neq 1/3$

# Different Asymptotic Regimes

Consider the setting: $H_0 : Z \sim P$ and $H_1 : Z \sim Q$.

- Type I error: $\alpha_n = \Pr(\text{guess } H_1 | H = H_0)$; Type II error: $\beta_n = \Pr(\text{guess } H_0 | H = H_1)$

Chernoff regime:

- Exponential decay for $\alpha_n$ and $\beta_n$: i.e., $\alpha_n \approx e^{-\lambda n}$ and $\beta_n \approx e^{-\gamma n}$ for $\lambda > 0, \gamma > 0$.
- For a given $\lambda > 0$, if $\lim_{n\to\infty} -\frac{1}{n} \ln \alpha_n(T_n) \geq \lambda$, maximize

$$\lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n^{NP})$$

# Different Asymptotic Regimes

Consider the setting: $H_0 : Z \sim P$ and $H_1 : Z \sim Q$.

- Type I error: $\alpha_n = \Pr(\text{guess } H_1 | H = H_0)$; Type II error: $\beta_n = \Pr(\text{guess } H_0 | H = H_1)$

Chernoff regime:

- Exponential decay for $\alpha_n$ and $\beta_n$: i.e., $\alpha_n \approx e^{-\lambda n}$ and $\beta_n \approx e^{-\gamma n}$ for $\lambda > 0, \gamma > 0$.
- For a given $\lambda > 0$, if $\lim_{n\to\infty} -\frac{1}{n} \ln \alpha_n(T_n) \geq \lambda$, maximize

$$\lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n^{NP})$$

Stein's regime:

- Exponential decay only for $\beta_n$
- For $\epsilon \in (0,1)$, if $\alpha_n(T_n) \leq \epsilon$ as $n \to \infty$, then maximize

$$\lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n)$$

# Likelihood Ratio Test: Stein's Regime

The Likelihood Ratio Test for the setting: $H_0 : Z \sim P$ and $H_1 : Z \sim Q$.

- Declare $H_0$ if $\frac{P^n(z^n)}{Q^n(z^n)} \geq r_n$ (for some $r_n$). Declare $H_1$ if $\frac{P^n(z^n)}{Q^n(z^n)} < r_n$

# Likelihood Ratio Test: Stein's Regime

The Likelihood Ratio Test for the setting: $H_0 : Z \sim P$ and $H_1 : Z \sim Q$.

- Declare $H_0$ if $\frac{P^n(z^n)}{Q^n(z^n)} \geq r_n$ (for some $r_n$). Declare $H_1$ if $\frac{P^n(z^n)}{Q^n(z^n)} < r_n$
- **Chernoff-Stein lemma:** For $\epsilon \in (0,1)$, if $\alpha_n(T_n^{NP}) \leq \epsilon$ as $n \to \infty$, then

$$\lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n^{NP}) = D_{\mathrm{KL}}(P\|Q)$$

  where $D_{\mathrm{KL}}$ is the Kullback-Leibler (KL) divergence
- $\lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n^{NP})$ is also known as the error exponent; $\beta_n(T_n^{NP}) \approx e^{-nD_{\mathrm{KL}}(P\|Q)}$

# Likelihood Ratio Test: Stein's Regime

The Likelihood Ratio Test for the setting: $H_0 : Z \sim P$ and $H_1 : Z \sim Q$.

- Declare $H_0$ if $\frac{P^n(z^n)}{Q^n(z^n)} \geq r_n$ (for some $r_n$). Declare $H_1$ if $\frac{P^n(z^n)}{Q^n(z^n)} < r_n$
- **Chernoff-Stein lemma:** For $\epsilon \in (0,1)$, if $\alpha_n(T_n^{NP}) \leq \epsilon$ as $n \to \infty$, then

$$\lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n^{NP}) = D_{\mathrm{KL}}(P\|Q)$$

  where $D_{\mathrm{KL}}$ is the Kullback-Leibler (KL) divergence
- $\lim\limits_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n^{NP})$ is also known as the error exponent; $\beta_n(T_n^{NP}) \approx e^{-nD_{\mathrm{KL}}(P\|Q)}$

Our focus: $H_0 : \mathcal{P} = \{P\}$ $H_1 : \mathcal{Q} = \mathcal{P}^c$ under Stein's regime

# The Hoeffding Test

- The Hoeffding Test: Suitable for $H_0 : \mathcal{P} = \{P\}$ vs. $H_1 : \mathcal{Q} = \mathcal{P}^c$
- $T_{Z^n} = (T_{Z^n}(a_1), \ldots, T_{Z^n}(a_k))^T$ - type (empirical distribution) of $Z^n$, $T_{Z^n}(a_i) = \frac{\# a_i \text{ in } Z^n}{n}$

The Hoeffding[2] test $T_n^H(r)$:

If $D_{\mathrm{KL}}(T_{Z^n} \| P) < r_n$ for some $r_n$, then accept $H_0$; else accept $H_1$

[2] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *The Annals of Mathematical Statistics*, pp. 369–401, Apr. 1965

# The Hoeffding Test

- The Hoeffding Test: Suitable for $H_0 : \mathcal{P} = \{P\}$ vs. $H_1 : \mathcal{Q} = \mathcal{P}^c$
- $T_{Z^n} = (T_{Z^n}(a_1), \ldots, T_{Z^n}(a_k))^T$ - type (empirical distribution) of $Z^n$, $T_{Z^n}(a_i) = \frac{\# a_i \text{ in } Z^n}{n}$

The Hoeffding[2] test $T_n^H(r)$:

If $D_{\mathrm{KL}}(T_{Z^n} \| P) < r_n$ for some $r_n$, then accept $H_0$; else accept $H_1$

- Hoeffding test does not require the knowledge about $Q$
- The Hoeffding test is the Generalized Likelihood Ratio Test (GLRT) in this setting

$$H_0 : \theta_0 = P \text{ vs. } H_1 = \Theta \setminus \{\theta_0\}$$

- GLRT Test statistic $T(z)$:

$$T(z) = -2 \ln \frac{L(\theta_0|z)}{L(\hat{\theta}|z)} = D_{\mathrm{KL}}(T_{Z^n} \| P)$$

$$L(\hat{\theta}|z) = \sup_{\theta \in \Theta} L(\theta|z), \; L(\theta|z) - \text{ Likelihood function}$$

[2] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *The Annals of Mathematical Statistics*, pp. 369–401, Apr. 1965

# The Hoeffding Test: GLRT

- The Hoeffding test achieves

$$\lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n^H) = D_{\mathrm{KL}}(P\|Q)$$

- Thus, $-\ln \beta(T_n^H) = nD_{\mathrm{KL}}(P\|Q) + o(n)$. The Hoeffding test is first-order optimal.

# The Hoeffding Test: GLRT

- The Hoeffding test achieves

$$\lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n^H) = D_{\mathrm{KL}}(P\|Q)$$

- Thus, $-\ln \beta(T_n^H) = nD_{\mathrm{KL}}(P\|Q) + o(n)$. The Hoeffding test is first-order optimal.

How about finer asymptotics?

For any test $T_n$:

- First-order term: $\beta'(T_n) \triangleq \lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n)$
- Second-order term: $\beta''(T_n) \triangleq \lim_{n\to\infty} \frac{-\ln \beta_n(T_n) - n\beta'(T_n)}{\sqrt{n}}$

# The Hoeffding Test: GLRT

- The Hoeffding test achieves

$$\lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n^H) = D_{\mathrm{KL}}(P\|Q)$$

- Thus, $-\ln \beta(T_n^H) = nD_{\mathrm{KL}}(P\|Q) + o(n)$. The Hoeffding test is first-order optimal.

How about finer asymptotics?

For any test $T_n$:

- First-order term: $\beta'(T_n) \triangleq \lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n)$
- Second-order term: $\beta''(T_n) \triangleq \lim_{n\to\infty} \frac{-\ln \beta_n(T_n) - n\beta'(T_n)}{\sqrt{n}}$
- $\beta'(T_n^H) = D_{\mathrm{KL}}(P\|Q)$
- What is $\beta''(T_n^H) = ?$

# Finer Asymptotics of Type II error of LR Test

- Refined asymptotics for the LR test: For $\alpha_n(T_n^{NP}) \leq \epsilon$, type II error satisfies[3]

$$-\ln \beta_n(T_n^{NP}) = nD_{\mathrm{KL}}(P\|Q) - \sqrt{nV_{\mathrm{KL}}(P\|Q)}\mathrm{Q}^{-1}(\epsilon) + o(\sqrt{n})$$

  - $\mathrm{Q}^{-1}(\cdot)$ - the inverse of the tail probability of the standard Normal distribution
  - $V_{\mathrm{KL}}(P\|Q)$ - divergence variance:

$$V_{\mathrm{KL}}(P\|Q) \triangleq \sum_i P_i \left[\left(\ln \frac{P_i}{Q_i} - D_{\mathrm{KL}}(P\|Q)\right)^2\right]$$

  - $o(\sqrt{n})$ - terms which grows smaller than $\sqrt{n}$

For Neyman-Pearson test $T_n^{NP}$:

- $\beta'(T_n^{NP}) = D_{\mathrm{KL}}(P\|Q)$
- $\beta''(T_n^{NP}) = -\sqrt{V_{\mathrm{KL}}(P\|Q)}\mathrm{Q}^{-1}(\epsilon)$

[3] V. Y. Tan, "Asymptotic estimates in information theory with non-vanishing error probabilities," *Foundations and Trends® in Communications and Information Theory*, vol. 11, 2014

# Second-order Asymptotics of the Hoeffding Test [4]

### Theorem

*For $0 < \epsilon < 1$, there exists a sequence $r_n$ satisfying $\alpha_n(T_n^H(r_n)) \leq \epsilon$ such that, as $n \to \infty$,*

$$-\ln \beta_n(T_n^H(r_n)) = nD_{\mathrm{KL}}(P\|Q) - \sqrt{nV_{\mathrm{KL}}(P\|Q)\mathrm{Q}_{\chi^2_{k-1}}^{-1}(\epsilon)} + o(\sqrt{n}).$$

$\mathrm{Q}_{\chi^2_m}(\cdot)$: *tail probability of $\chi^2$ RV with $m$ degrees of freedom.*

[4] K. V. Harsha, J. Ravi, and T. Koch, "Second-order asymptotics of hoeffding-like hypothesis tests," in *2022 IEEE Information Theory Workshop (ITW)*, 2022, pp. 654–659

# Second-order Asymptotics of the Hoeffding Test [4]

**Theorem**

*For $0 < \epsilon < 1$, there exists a sequence $r_n$ satisfying $\alpha_n(T_n^H(r_n)) \leq \epsilon$ such that, as $n \to \infty$,*

$$-\ln \beta_n(T_n^H(r_n)) = nD_{\mathrm{KL}}(P\|Q) - \sqrt{nV_{\mathrm{KL}}(P\|Q)\mathrm{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)} + o(\sqrt{n}).$$

$\mathrm{Q}_{\chi^2_m}(\cdot)$*: tail probability of $\chi^2$ RV with $m$ degrees of freedom.*

- LR test: $-\ln \beta_n(T_n^{NP}) = nD_{\mathrm{KL}}(P\|Q) - \sqrt{nV_{\mathrm{KL}}(P\|Q)}\mathrm{Q}^{-1}(\epsilon) + o(\sqrt{n})$

[4] K. V. Harsha, J. Ravi, and T. Koch, "Second-order asymptotics of hoeffding-like hypothesis tests," in *2022 IEEE Information Theory Workshop (ITW)*, 2022, pp. 654–659

# Second-order Asymptotics of the Hoeffding Test [4]

## Theorem

*For $0 < \epsilon < 1$, there exists a sequence $r_n$ satisfying $\alpha_n(T_n^H(r_n)) \leq \epsilon$ such that, as $n \to \infty$,*

$$-\ln \beta_n(T_n^H(r_n)) = nD_{\mathrm{KL}}(P\|Q) - \sqrt{nV_{\mathrm{KL}}(P\|Q)\mathrm{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)} + o(\sqrt{n}).$$

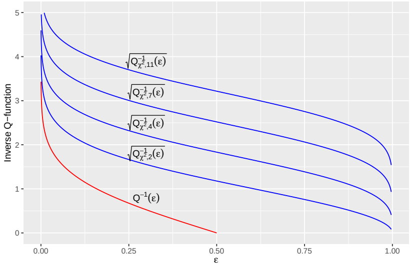
$\mathrm{Q}_{\chi^2_m}(\cdot)$: *tail probability of $\chi^2$ RV with m degrees of freedom.*

- LR test: $-\ln \beta_n(T_n^{NP}) = nD_{\mathrm{KL}}(P\|Q) - \sqrt{nV_{\mathrm{KL}}(P\|Q)}\mathrm{Q}^{-1}(\epsilon) + o(\sqrt{n})$

## Lemma

*For $0 < \epsilon < 1$ and $\eta = 1, 2, \ldots$*

$$\sqrt{\mathrm{Q}^{-1}_{\chi^2_{\eta+1}}(\epsilon)} > \sqrt{\mathrm{Q}^{-1}_{\chi^2_{\eta}}(\epsilon)} > \mathrm{Q}^{-1}(\epsilon)$$

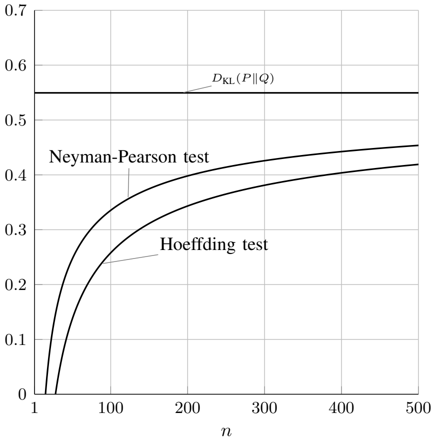



[4] K. V. Harsha, J. Ravi, and T. Koch, "Second-order asymptotics of hoeffding-like hypothesis tests," in *2022 IEEE Information Theory Workshop (ITW)*, 2022, pp. 654–659

# Performance Comparison of Second Order Terms

$$-\frac{1}{n}\ln\beta_n(T_n^H(r_n)) \approx D_{\mathsf{KL}}(P\|Q) - \left\{\sqrt{\frac{V_{\mathsf{KL}}(P\|Q)}{n}}\mathsf{Q}^{-1}(\epsilon), \sqrt{\frac{V_{\mathsf{KL}}(P\|Q)}{n}}\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)\right\}$$

- $P = (0.15, 0.6, 0.25)$
- $Q = (0.45, 0.15, 0.4)$
- $\epsilon = 0.02$



# Key tools used in the proofs

- LR test: $-\ln \beta_n(T_n^{NP}) = nD_{\mathrm{KL}}(P\|Q) - \sqrt{nV_{\mathrm{KL}}(P\|Q)}\mathrm{Q}^{-1}(\epsilon) + o(\sqrt{n})$
- Key tool: Central limit theorem - when $Z^n$ i.i.d. $P$

$$\frac{\sum_{i=1}^n \log \frac{P(Z_i)}{Q(Z_i)} - nD_{\mathrm{KL}}(P\|Q)}{\sqrt{nV_{\mathrm{KL}}(P\|Q)}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } n \to \infty$$

[5] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The annals of mathematical statistics*, vol. 9, no. 1, pp. 60–62, 1938

# Key tools used in the proofs

- LR test: $-\ln \beta_n(T_n^{NP}) = nD_{\mathrm{KL}}(P\|Q) - \sqrt{nV_{\mathrm{KL}}(P\|Q)}\mathrm{Q}^{-1}(\epsilon) + o(\sqrt{n})$
- Key tool: Central limit theorem - when $Z^n$ i.i.d. $P$

$$\frac{\sum_{i=1}^{n} \log \frac{P(Z_i)}{Q(Z_i)} - nD_{\mathrm{KL}}(P\|Q)}{\sqrt{nV_{\mathrm{KL}}(P\|Q)}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } n \to \infty$$

- Hoeffding Test: $-\ln \beta_n(T_n^{H}(r_n)) = nD_{\mathrm{KL}}(P\|Q) - \sqrt{nV_{\mathrm{KL}}(P\|Q)\mathrm{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)} + o(\sqrt{n})$
- Key tool: use Wilks's theorem[5] that when $Z^n$ i.i.d. $P$

$$2nD_{\mathrm{KL}}(T_{Z^n}\|P) \xrightarrow{d} \chi^2_{|\mathcal{Z}|-1} \quad \text{as } n \to \infty$$

- Needs Berry–Esseen type convergence results

[5] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The annals of mathematical statistics*, vol. 9, no. 1, pp. 60–62, 1938

# Key tools used in the proofs

- LR test: $-\ln \beta_n(T_n^{NP}) = nD_{\mathrm{KL}}(P\|Q) - \sqrt{nV_{\mathrm{KL}}(P\|Q)}\mathrm{Q}^{-1}(\epsilon) + o(\sqrt{n})$
- Key tool: Central limit theorem - when $Z^n$ i.i.d. $P$

$$\frac{\sum_{i=1}^n \log \frac{P(Z_i)}{Q(Z_i)} - nD_{\mathrm{KL}}(P\|Q)}{\sqrt{nV_{\mathrm{KL}}(P\|Q)}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } n \to \infty$$

- Hoeffding Test: $-\ln \beta_n(T_n^H(r_n)) = nD_{\mathrm{KL}}(P\|Q) - \sqrt{nV_{\mathrm{KL}}(P\|Q)\mathrm{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)} + o(\sqrt{n})$
- Key tool: use Wilks's theorem[5] that when $Z^n$ i.i.d. $P$

$$2nD_{\mathrm{KL}}(T_{Z^n}\|P) \xrightarrow{d} \chi^2_{|\mathcal{Z}|-1} \quad \text{as } n \to \infty$$

- Needs Berry–Esseen type convergence results

Thanks to **Shun Watanabe**, Tokyo University of Agriculture and Technology, Japan

[5] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The annals of mathematical statistics*, vol. 9, no. 1, pp. 60–62, 1938

# Extending to other divergences: Divergence Test

- Wilks's result and other properties of KL also hold for divergences like Rényi divergence

$$nD_R(T_{Z^n} \| P) \xrightarrow{d} \chi^2_{|\mathcal{Z}|-1} \quad \text{as } n \to \infty$$

- Can we extend the test to other divergences: Hoeffding-like test?

# Extending to other divergences: Divergence Test

- Wilks's result and other properties of KL also hold for divergences like Rényi divergence

$$nD_R(T_{Z^n}\|P) \xrightarrow{d} \chi^2_{|\mathcal{Z}|-1} \quad \text{as } n \to \infty$$

- Can we extend the test to other divergences: Hoeffding-like test?

**Divergence test** $T_n^D(r)$ for testing $H_0 : \mathcal{P} = \{P\}$, $H_1 : \mathcal{Q} = \mathcal{P}^c$

Upon observing $z^n$,

- if $D(T_{z^n}\|P) \leq r_n$ (for some $r_n > 0$), then accept $H_0$
- if $D(T_{z^n}\|P) > r_n$, then accept $H_1$

# Divergence

- Let $\mathcal{P}(\mathcal{Z})$ be the set of positive probability distributions.
- For any $T = (T_1, \ldots, T_k) \in \mathcal{P}(\mathcal{Z})$, let $\mathbf{T} = (T_1, \ldots, T_{k-1})$.
- We follow the definition of divergence from information geometry[6]

## Definition

A *divergence* $D : \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to [0, \infty)$ between $T$ and $R$, denoted by $D(T\|R)$ or $D(\mathbf{T}\|\mathbf{R})$, is a smooth function of $\mathbf{T}$ and $\mathbf{R}$ satisfying the following conditions:

1. $D(T\|R) \geq 0$ for every $T, R \in \mathcal{P}(\mathcal{Z})$.
2. $D(T\|R) = 0$ if, and only if, $T = R$.
3. When $\mathbf{T} = \mathbf{R} + \boldsymbol{\varepsilon}$, the Taylor expansion of $D$ satisfies

$$D(\mathbf{R} + \boldsymbol{\varepsilon}\|\mathbf{R}) = \frac{1}{2}\sum_{i,j=1}^{k-1} g_{ij}(\mathbf{R})\varepsilon_i\varepsilon_j + O(\|\boldsymbol{\varepsilon}\|_2^3), \quad \text{as } \|\boldsymbol{\varepsilon}\|_2 \to 0$$

for some positive-definite matrix $G(\mathbf{R}) = [g_{ij}(\mathbf{R})]$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{k-1})^{\mathsf{T}}$.

[6] S.-I. Amari, *Information geometry and its applications*. Springer, 2016, vol. 194

# Main Result: Second-order Asymptotics of the Divergence Test

- From the 3rd condition: $D(T\|R) = (\mathbf{T}-\mathbf{R})^T \boldsymbol{A}_{D,\mathbf{R}}(\mathbf{T}-\mathbf{R}) + O(\|\mathbf{T}-\mathbf{R}\|_2^3), \quad \text{as} \quad \|\mathbf{T}-\mathbf{R}\|_2 \to 0.$

## Theorem

*There exists a threshold value $r_n$ (which depends on $n$) satisfying $\alpha_n(\mathrm{T}_n^D(r_n)) \leq \epsilon$,*

$$-\ln \beta_n(\mathrm{T}_n^D(r_n)) = nD_{\mathrm{KL}}(P\|Q) - \sqrt{n}\sqrt{\mathbf{c}^{\mathsf{T}} \boldsymbol{A}_{D,\mathbf{P}}^{-1} \mathbf{c}}\sqrt{\mathrm{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\epsilon)} + o(\sqrt{n}).$$

# Main Result: Second-order Asymptotics of the Divergence Test

- From the 3rd condition: $D(T\|R) = (\mathbf{T}-\mathbf{R})^T \boldsymbol{A}_{D,\mathbf{R}}(\mathbf{T}-\mathbf{R}) + O(\|\mathbf{T}-\mathbf{R}\|_2^3)$, as $\|\mathbf{T}-\mathbf{R}\|_2 \to 0$.

## Theorem

*There exists a threshold value $r_n$ (which depends on $n$) satisfying $\alpha_n(\mathrm{T}_n^D(r_n)) \leq \epsilon$,*

$$-\ln \beta_n(\mathrm{T}_n^D(r_n)) = nD_{\mathrm{KL}}(P\|Q) - \sqrt{n}\sqrt{\mathbf{c}^\mathsf{T}\boldsymbol{A}_{D,\mathbf{P}}^{-1}\mathbf{c}}\sqrt{\mathrm{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\epsilon)} + o(\sqrt{n}).$$

1. $c_i \triangleq \ln\left(\frac{P_i}{Q_i}\right) - \ln\left(\frac{P_k}{Q_k}\right), \quad i = 1,\ldots,k-1$

# Main Result: Second-order Asymptotics of the Divergence Test

- From the 3rd condition: $D(T\|R) = (\mathbf{T}-\mathbf{R})^T \boldsymbol{A}_{D,\mathbf{R}}(\mathbf{T}-\mathbf{R}) + O(\|\mathbf{T}-\mathbf{R}\|_2^3)$, as $\|\mathbf{T}-\mathbf{R}\|_2 \to 0$.

## Theorem

*There exists a threshold value $r_n$ (which depends on $n$) satisfying $\alpha_n(\mathrm{T}_n^D(r_n)) \leq \epsilon$,*

$$-\ln \beta_n(\mathrm{T}_n^D(r_n)) = nD_{\mathrm{KL}}(P\|Q) - \sqrt{n}\sqrt{\mathbf{c}^{\mathrm{T}}\boldsymbol{A}_{D,\mathbf{P}}^{-1}\mathbf{c}}\sqrt{\mathrm{Q}_{\chi^2_{\boldsymbol{\lambda},k-1}}^{-1}(\epsilon)} + o(\sqrt{n}).$$

1. $c_i \triangleq \ln\left(\frac{P_i}{Q_i}\right) - \ln\left(\frac{P_k}{Q_k}\right), \quad i = 1,\ldots,k-1$
2. $\mathrm{Q}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\cdot)$: *tail probability of generalized chi-squared RV* $\psi \sim \chi^2_{\mathbf{w},m}$

   $\psi = \sum_i^m w_i\Gamma_i$, $\quad \boldsymbol{w} = (w_1,\ldots,w_m)$: *weight vector;* $\Gamma_i$ *- a* $\chi^2$ *RV with 1 degree of freedom*

# Main Result: Second-order Asymptotics of the Divergence Test

- From the 3rd condition: $D(T\|R) = (\mathbf{T}-\mathbf{R})^T \boldsymbol{A}_{D,\mathbf{R}}(\mathbf{T}-\mathbf{R}) + O(\|\mathbf{T}-\mathbf{R}\|_2^3)$, as $\|\mathbf{T}-\mathbf{R}\|_2 \to 0$.

**Theorem**

*There exists a threshold value $r_n$ (which depends on $n$) satisfying $\alpha_n(\mathrm{T}_n^D(r_n)) \le \epsilon$,*

$$-\ln \beta_n(\mathrm{T}_n^D(r_n)) = nD_{\mathrm{KL}}(P\|Q) - \sqrt{n}\sqrt{\mathbf{c}^\mathrm{T}\boldsymbol{A}_{D,\mathbf{P}}^{-1}\mathbf{c}}\sqrt{\mathrm{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\epsilon)} + o(\sqrt{n}).$$

1. $c_i \triangleq \ln\left(\frac{P_i}{Q_i}\right) - \ln\left(\frac{P_k}{Q_k}\right), \quad i = 1, \ldots, k-1$
2. $\mathrm{Q}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\cdot)$: *tail probability of generalized chi-squared RV* $\psi \sim \chi^2_{\mathbf{w},m}$

   $\psi = \sum_i^m w_i\Gamma_i$, $\quad \boldsymbol{w} = (w_1, \ldots, w_m)$: *weight vector;* $\Gamma_i$ *- a* $\chi^2$ *RV with 1 degree of freedom*
3. $\boldsymbol{\lambda}$: *vector containing eigenvalues of* $\boldsymbol{\Sigma}_\mathbf{P}^{-1/2}\boldsymbol{A}_{D,\mathbf{P}}\boldsymbol{\Sigma}_\mathbf{P}^{-1/2}$, $\boldsymbol{\Sigma}_\mathbf{P}$ *is the Fisher Information matrix*

$$\boldsymbol{\Sigma}_\mathbf{P}(i,j) = \begin{cases} \frac{1}{P_i} + \frac{1}{P_k}, & \text{for } i = j \\ \frac{1}{P_k}, & \text{for } i \neq j. \end{cases}$$

# Two Classes of Divergences

- If $\boldsymbol{A}_{D,\mathbf{P}} = \eta \boldsymbol{\Sigma}_{\mathbf{P}}$ (for some $\eta > 0$), then $\boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2} \boldsymbol{A}_{D,\mathbf{P}} \boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2}$ is a multiple of identity matrix.

# Two Classes of Divergences

- If $\boldsymbol{A}_{D,\mathbf{P}} = \eta \boldsymbol{\Sigma}_{\mathbf{P}}$ (for some $\eta > 0$), then $\boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2} \boldsymbol{A}_{D,\mathbf{P}} \boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2}$ is a multiple of identity matrix.

### Invariant Divergence (Information Geometry)

For some $\eta > 0$, if $\boldsymbol{A}_{D,\mathbf{P}} = \eta \boldsymbol{\Sigma}_{\mathbf{P}}$, where

$$\boldsymbol{\Sigma}_{\mathbf{P}}(i,j) = \begin{cases} \frac{1}{P_i} + \frac{1}{P_k}, & \text{for } i = j \\ \frac{1}{P_k}, & \text{for } i \neq j. \end{cases}$$

Examples of invariant divergence

1. f-divergence (includes KL divergence)
2. Rényi Divergence

# Two Classes of Divergences

- If $\boldsymbol{A}_{D,\mathbf{P}} = \eta \boldsymbol{\Sigma}_{\mathbf{P}}$ (for some $\eta > 0$), then $\boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2} \boldsymbol{A}_{D,\mathbf{P}} \boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2}$ is a multiple of identity matrix.

**Invariant Divergence (Information Geometry)**

For some $\eta > 0$, if $\boldsymbol{A}_{D,\mathbf{P}} = \eta \boldsymbol{\Sigma}_{\mathbf{P}}$, where

$$\boldsymbol{\Sigma}_{\mathbf{P}}(i,j) = \begin{cases} \frac{1}{P_i} + \frac{1}{P_k}, & \text{for } i = j \\ \frac{1}{P_k}, & \text{for } i \neq j. \end{cases}$$

Examples of invariant divergence

1. f-divergence (includes KL divergence)
2. Rényi Divergence

- For invariant divergence $D$: Consider $\sqrt{\mathbf{c}^{\mathsf{T}} \boldsymbol{A}_{D,\mathbf{P}}^{-1} \mathbf{c}} \sqrt{\mathrm{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\epsilon)}$
  - $\mathbf{c}^{\mathsf{T}} \boldsymbol{A}_{D,\mathbf{P}}^{-1} \mathbf{c} = \frac{1}{\eta} V_{\mathrm{KL}}(P \| Q)$
  - Eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2} \boldsymbol{A}_{D,\mathbf{P}} \boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2}$ are the same. $\mathrm{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\epsilon)$ becomes $\eta \mathrm{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)$

# Two Classes of Divergences

- If $\boldsymbol{A}_{D,\mathbf{P}} = \eta \boldsymbol{\Sigma}_{\mathbf{P}}$ (for some $\eta > 0$), then $\boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2} \boldsymbol{A}_{D,\mathbf{P}} \boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2}$ is a multiple of identity matrix.

**Invariant Divergence (Information Geometry)**

For some $\eta > 0$, if $\boldsymbol{A}_{D,\mathbf{P}} = \eta \boldsymbol{\Sigma}_{\mathbf{P}}$, where

$$\boldsymbol{\Sigma}_{\mathbf{P}}(i,j) = \begin{cases} \frac{1}{P_i} + \frac{1}{P_k}, & \text{for } i = j \\ \frac{1}{P_k}, & \text{for } i \neq j. \end{cases}$$

Examples of invariant divergence

1. f-divergence (includes KL divergence)
2. Rényi Divergence

- For invariant divergence $D$: Consider $\sqrt{\mathbf{c}^{\mathsf{T}} \boldsymbol{A}_{D,\mathbf{P}}^{-1} \mathbf{c}} \sqrt{\mathrm{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\epsilon)}$
  - $\mathbf{c}^{\mathsf{T}} \boldsymbol{A}_{D,\mathbf{P}}^{-1} \mathbf{c} = \frac{1}{\eta} V_{\mathrm{KL}}(P \| Q)$
  - Eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2} \boldsymbol{A}_{D,\mathbf{P}} \boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2}$ are the same. $\mathrm{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\epsilon)$ becomes $\eta \mathrm{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)$

If $D$ is invariant, then the second-order performance is same as that of the Hoeffding test, i.e.,

$$\sqrt{\mathbf{c}^{\mathsf{T}} \boldsymbol{A}_{D,\mathbf{P}}^{-1} \mathbf{c}} \sqrt{\mathrm{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\epsilon)} = \sqrt{V_{\mathrm{KL}}(P \| Q) \mathrm{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)}$$

# Invariant and Non-invariant Divergences

- Non-invariant Divergence: A divergence that is not invariant

# Invariant and Non-invariant Divergences

- Non-invariant Divergence: A divergence that is not invariant

Examples of non-invariant divergence:

1. Squared Mahalanobis distance:

$$D_{\mathrm{SM}}(T\|R) \triangleq (\mathbf{T}-\mathbf{R})^{\mathsf{T}}\boldsymbol{W}_{\mathbf{R}}(\mathbf{T}-\mathbf{R}), \quad \boldsymbol{W}_{\mathbf{R}} \neq \eta\boldsymbol{\Sigma}_{\mathbf{R}}$$

2. Some of the Bregman Divergences associated with $\phi$ (Itakura-Saito divergence):

$$D_{\phi}(T\|R) = \phi(\mathbf{T}) - \phi(\mathbf{R}) - (\mathbf{T}-\mathbf{R})^{\mathsf{T}}\nabla\phi(\mathbf{R})$$

   where $\nabla\phi(\mathbf{R})$ denotes the gradient of $\phi$ evaluated at $\mathbf{R}$

3. Sundaresan divergence (2007): For $\alpha > 0$,

$$\Delta_{\alpha}(T\|R) = \frac{\alpha}{1-\alpha}\ln\sum_{x} T(x)R(x)^{\alpha-1} - \frac{1}{1-\alpha}\ln\sum_{x} T(x)^{\alpha} + \ln\sum_{x} R(x)^{\alpha}$$

# Divergence Test: Extension of Wilks's theorem

- Convergence to a generalized $\chi^2$ RV

$$nD(T_{Z^n}\|P) \xrightarrow{d} \chi^2_{\boldsymbol{\lambda},|\mathcal{Z}|-1} \quad \text{as } n \to \infty$$

# Divergence Test: Extension of Wilks's theorem

- Convergence to a generalized $\chi^2$ RV

$$nD(T_{Z^n}\|P) \xrightarrow{d} \chi^2_{\boldsymbol{\lambda},|\mathcal{Z}|-1} \quad \text{as } n \to \infty$$

For more details...

K. V. Harsha, J. Ravi, and T. Koch, "On the second-order asymptotics of the Hoeffding test and other divergence tests," IEEE Trans. Inf. Theory, Oct. 2025

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 71, NO. 10, OCTOBER 2025 7459

## On the Second-Order Asymptotics of the Hoeffding Test and Other Divergence Tests

K. V. Harsha, Jithin Ravi, *Member, IEEE*, and Tobias Koch, *Senior Member, IEEE*

***Abstract*—Consider a composite hypothesis testing problem where the test has access to the null hypothesis $P$ but not to the alternative hypothesis $Q$. The generalized likelihood-ratio test (GLRT) for this problem is the Hoeffding test, which accepts $P$ if the Kullback-Leibler (KL) divergence between the empirical distribution of $Z^n$ and $P$ is below some threshold. This paper proposes a generalization of the Hoeffding test, termed divergence test, for which the KL divergence is replaced by an arbitrary divergence. For this test, the first and second-order terms of the type-II error probability for a fixed type-I error probability are characterized and compared with the error terms of the Neyman-Pearson test, which is the optimal test when both $P$ and $Q$ are known. It is demonstrated that, irrespective of the divergence, divergence tests achieve the first-order term of the**

### I. Introduction

STATISTICAL hypothesis testing is known to have applications in areas such as information theory, signal processing, and machine learning; see, e.g., [1], [2], [3], [4], [5], [6]. The most simple form of hypothesis testing is binary hypothesis testing, where the goal is to determine the distribution of a random variable $Z$ between a null hypothesis $P$ and an alternative hypothesis $Q$. There can be two types of errors in binary hypothesis testing: The type-I error is the probability of declaring the hypothesis as $Q$ when the true distribution is $P$. The type-II error is the probability of

# Observations on the Divergence Test

1. The divergence test is first-order optimal (irrespective of $D$).
2. When $D$ is invariant, it has the same second-order performance as the Hoeffding test, given by

$$\sqrt{V_{\mathsf{KL}}(P\|Q)\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)}$$

# Observations on the Divergence Test

1. The divergence test is first-order optimal (irrespective of $D$).
2. When $D$ is invariant, it has the same second-order performance as the Hoeffding test, given by

$$\sqrt{V_{\mathsf{KL}}(P\|Q)\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)}$$

3. Second-order performance is always worse than the NP test, i.e., for every $D$ and $0 < \epsilon < 1$,

$$\sqrt{\mathbf{c}^{\mathsf{T}}\boldsymbol{A}^{-1}_{D,\mathbf{P}}\mathbf{c}}\sqrt{\mathsf{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\epsilon)} > \sqrt{V_{\mathsf{KL}}(P\|Q)}\mathsf{Q}^{-1}(\epsilon)$$

# Observations on the Divergence Test

1. The divergence test is first-order optimal (irrespective of $D$).
2. When $D$ is invariant, it has the same second-order performance as the Hoeffding test, given by

$$\sqrt{V_{\mathsf{KL}}(P\|Q)\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)}$$

3. Second-order performance is always worse than the NP test, i.e., for every $D$ and $0 < \epsilon < 1$,

$$\sqrt{\mathbf{c}^{\mathsf{T}}\boldsymbol{A}^{-1}_{D,\mathbf{P}}\mathbf{c}}\sqrt{\mathsf{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\epsilon)} > \sqrt{V_{\mathsf{KL}}(P\|Q)}\mathsf{Q}^{-1}(\epsilon)$$

4. Depending on $D, P, Q, \epsilon$

$$\sqrt{\mathbf{c}^{\mathsf{T}}\boldsymbol{A}^{-1}_{D,\mathbf{P}}\mathbf{c}}\sqrt{\mathsf{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\epsilon)} \lessgtr \sqrt{V_{\mathsf{KL}}(P\|Q)\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)}$$

   - Test with a non-invariant $D$ can outperform the Hoeffding test

# Numerical Results: Setup

- Ternary observations ($|\mathcal{Z}| = 3$) and $\epsilon = 0.02$

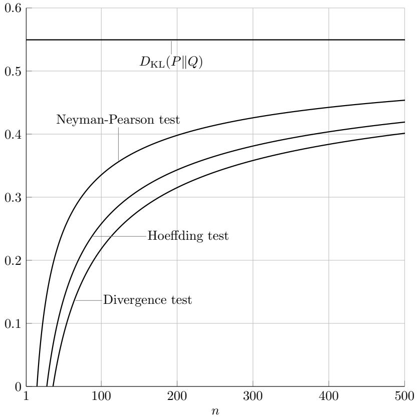
3 hypothesis tests:

1. likelihood ratio test / NP test
2. Hoeffding test ($D_{\mathrm{KL}}$)
3. $D$ is squared Mahalanobis distance

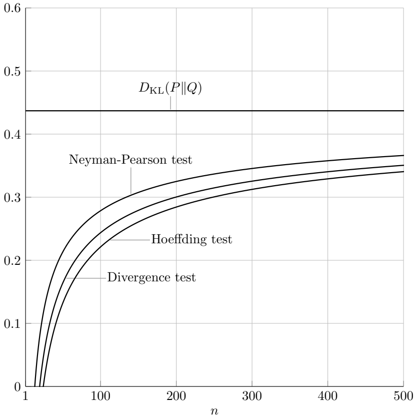
$$\boldsymbol{W}_{\mathbf{P}}(i,j) = \begin{cases} \frac{1}{2P_i^2} + \frac{1}{2P_k^2}, & i = j \\ \frac{1}{2P_k^2}, & i \neq j. \end{cases}$$

1. $-\frac{1}{n} \ln \beta_n(T_n^{NP}(r_n)) \approx D_{\mathrm{KL}}(P\|Q) - \sqrt{\frac{V_{\mathrm{KL}}(P\|Q)}{n}} \mathrm{Q}^{-1}(\epsilon)$
2. $-\frac{1}{n} \ln \beta_n(T_n^{H}(r_n)) \approx D_{\mathrm{KL}}(P\|Q) - \sqrt{\frac{V_{\mathrm{KL}}(P\|Q)}{n} \mathrm{Q}_{\chi^2_{k-1}}^{-1}(\epsilon)}$
3. $-\frac{1}{n} \ln \beta_n(T_n^{SM}(r_n)) \approx D_{\mathrm{KL}}(P\|Q) - \sqrt{\frac{\mathbf{c}^{\mathsf{T}} \boldsymbol{A}_{D,\mathbf{P}}^{-1} \mathbf{c}}{n} \mathrm{Q}_{\chi^2_{\boldsymbol{\lambda},k-1}}^{-1}(\epsilon)}$

# Numerical Results: Second-Order Performance




$Q = (0.45, 0.15, 0.4)$

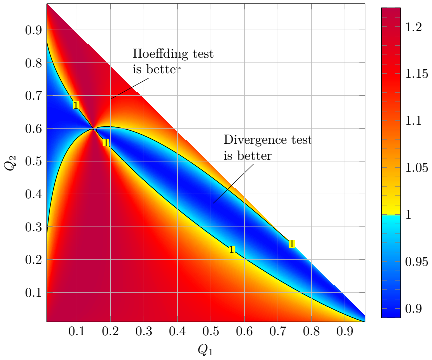



$Q = (0.6, 0.3, 0.1)$

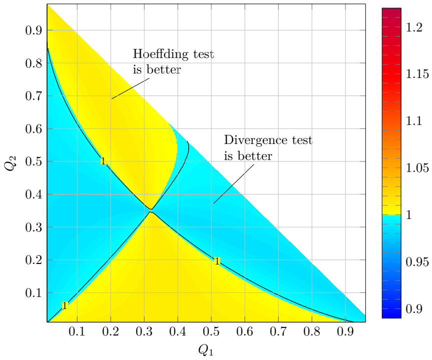
For $P = (0.15, 0.6, 0.25)$ and $\epsilon = 0.02$

# Ratio of Second-Order Terms: Changes with $P$

$$\rho(P, Q, \epsilon) = \frac{\sqrt{\mathbf{c}^{\mathsf{T}}(\boldsymbol{W}_{\mathsf{P}})^{-1}\mathbf{c}}\sqrt{\mathsf{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},2}}(\epsilon)}}{\sqrt{V_{\mathsf{KL}}(P\|Q)}\sqrt{\mathsf{Q}^{-1}_{\chi^2_2}(\epsilon)}}, \quad \text{for } \epsilon = 0.02$$



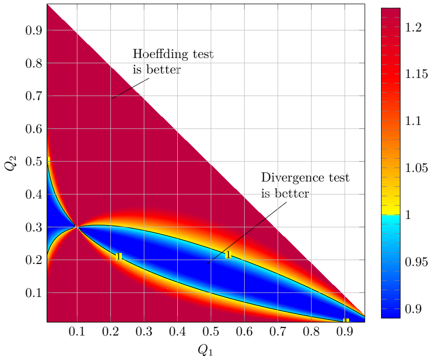

$P = (0.15, 0.6, 0.25)$



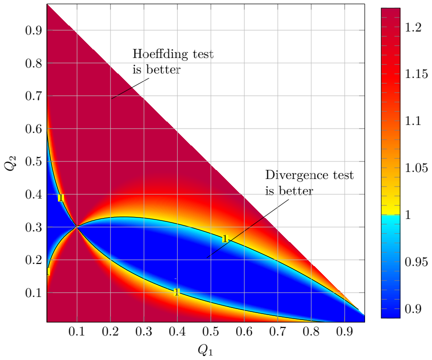

$P = (0.32, 0.35, 0.33)$

# Ratio of Second-Order Terms: Changes with $\epsilon$

$$\rho(P, Q, \epsilon) = \frac{\sqrt{\mathbf{c}^\mathsf{T}(\boldsymbol{W}_\mathsf{P})^{-1}\mathbf{c}}\sqrt{\mathsf{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},2}}(\epsilon)}}{\sqrt{V_{\mathsf{KL}}(P\|Q)}\sqrt{\mathsf{Q}^{-1}_{\chi^2_2}(\epsilon)}}, \quad \text{for } P = (0.1, 0.3, 0.6)$$




$\epsilon = 0.02$

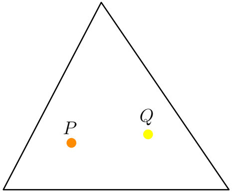



$\epsilon = 0.5$

# From One-Sample to Two-Sample Test

# Binary Hypothesis Testing on Two Samples

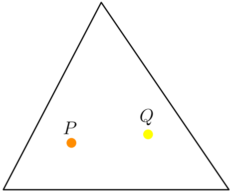
- **One-Sample Test**: Observe one i.i.d. seqeunce $Z^n = (Z_1, \ldots, Z_n)$, where $|\mathcal{Z}| = k$
- Classical Setting: $H_0 : P$ vs. $H_1 : Q$



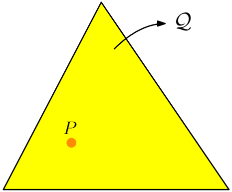

Classical Testing Problem

# Binary Hypothesis Testing on Two Samples

- **One-Sample Test**: Observe one i.i.d. seqeunce $Z^n = (Z_1, \ldots, Z_n)$, where $|\mathcal{Z}| = k$
- Classical Setting: $H_0 : P$ vs. $H_1 : Q$




Classical Testing Problem

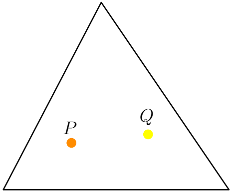



Goodness-of-Fit Testing Problem

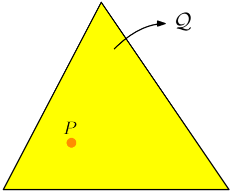
- Goodness of Fit Test: $H_0 : P$ vs. $H_1 : \mathcal{Q} = \mathcal{P} \setminus \{P\}$

# Binary Hypothesis Testing on Two Samples

- **One-Sample Test**: Observe one i.i.d. seqeunce $Z^n = (Z_1, \ldots, Z_n)$, where $|\mathcal{Z}| = k$
- Classical Setting: $H_0 : P$ vs. $H_1 : Q$




Classical Testing Problem




Goodness-of-Fit Testing Problem

- Goodness of Fit Test: $H_0 : P$ vs. $H_1 : \mathcal{Q} = \mathcal{P} \setminus \{P\}$
- **Two-Sample Test**: Observe two independent i.i.d. sequences

$$X^n = (X_1, \ldots, X_n) \text{ and } Y^n = (Y_1, \ldots, Y_n)$$

# Two-Sample Test

- **Two-Sample Test**: Observe two independent i.i.d. sequences $X^n$ and $Y^n$

Sequence $x^n$ with $n = 100$:

**0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 0**
**0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0**

Sequence $y^n$ with $n = 100$:

**1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0**
**1 0 0 1 0 0 1 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 0**

# Two-Sample Test

- **Two-Sample Test**: Observe two independent i.i.d. sequences $X^n$ and $Y^n$

Sequence $x^n$ with $n = 100$:

**0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 0**
**0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0**

Sequence $y^n$ with $n = 100$:

**1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0**
**1 0 0 1 0 0 1 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 0**

- $X^n$ i.i.d. Bern(p) and $Y^n$ i.i.d. Bern(q)
- $x^n$ contains 32 ones and 68 zeros; $y^n$ contains 26 ones and 74 zeros
- Can we conclude $p = q$? (Homogeneity Test)

# Two-Sample Test

- **Two-Sample Test**: Observe two independent i.i.d. sequences $X^n$ and $Y^n$

Sequence $x^n$ with $n = 100$:

**0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 0**
**0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0**

Sequence $y^n$ with $n = 100$:

**1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0**
**1 0 0 1 0 0 1 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 0**

- $X^n$ i.i.d. Bern(p) and $Y^n$ i.i.d. Bern(q)
- $x^n$ contains 32 ones and 68 zeros; $y^n$ contains 26 ones and 74 zeros
- Can we conclude $p = q$? (Homogeneity Test)
- Applications of the 2-sample test:
  - Safety of drinking water in two areas:
  - $H_0$: Safety levels are the same; $H_1$: they differ

# Problem Formulation: Two-Sample Test

- $X$ and $Y$ are independent and take values in the same finite alphabet $\mathcal{Z}$
- $X^n = (X_1, \ldots, X_n)$ i.i.d. according to $P_1$ and $Y^n = (Y_1, \ldots, Y_n)$ i.i.d. according to $P_2$.

Consider the setting: $H_0 : P_1 = P_2 = P$ and $H_1 : P_1 \neq P_2$.

# Problem Formulation: Two-Sample Test

- $X$ and $Y$ are independent and take values in the same finite alphabet $\mathcal{Z}$
- $X^n = (X_1, \ldots, X_n)$ i.i.d. according to $P_1$ and $Y^n = (Y_1, \ldots, Y_n)$ i.i.d. according to $P_2$.

Consider the setting: $H_0 : P_1 = P_2 = P$ and $H_1 : P_1 \neq P_2$.

- A test $T_n$ decides $\hat{H}$; $\mathrm{T}_n : \mathcal{Z}^n \times \mathcal{Z}^n \to \{H_0, H_1\}$ without any knowledge of $P_1$ and $P_2$
- $\mathcal{A}(\mathrm{T}_n)$ denote the acceptance region for $H_0$

$$\mathcal{A}(\mathrm{T}_n) = \{(x^n, y^n) : (x^n, y^n) \in \mathcal{Z}^n \times \mathcal{Z}^n, T_n(x^n, y^n) = H_0\}$$

# Problem Formulation: Two-Sample Test

- $X$ and $Y$ are independent and take values in the same finite alphabet $\mathcal{Z}$
- $X^n = (X_1, \ldots, X_n)$ i.i.d. according to $P_1$ and $Y^n = (Y_1, \ldots, Y_n)$ i.i.d. according to $P_2$.

Consider the setting: $H_0 : P_1 = P_2 = P$ and $H_1 : P_1 \neq P_2$.

- A test $T_n$ decides $\hat{H}$; $\mathrm{T}_n : \mathcal{Z}^n \times \mathcal{Z}^n \to \{H_0, H_1\}$ without any knowledge of $P_1$ and $P_2$
- $\mathcal{A}(\mathrm{T}_n)$ denote the acceptance region for $H_0$

$$\mathcal{A}(\mathrm{T}_n) = \{(x^n, y^n) : (x^n, y^n) \in \mathcal{Z}^n \times \mathcal{Z}^n, T_n(x^n, y^n) = H_0\}$$

- A test $T_n$ will result in two types of errors

$$\text{Type I Error:} \quad \alpha_n(\mathrm{T}_n) \triangleq \sum_{(x^n, y^n) \notin \mathcal{A}(\mathrm{T}_n)} P(x^n) P(y^n)$$

$$\text{Type II Error:} \quad \beta_n(\mathrm{T}_n) \triangleq \sum_{(x^n, y^n) \in \mathcal{A}(\mathrm{T}_n)} P_1(x^n) P_2(y^n)$$

# Asymptotics: Chernoff regime (Gutman Test)

Chernoff regime:

- Exponential decay for $\alpha_n$ and $\beta_n$: i.e., $\alpha_n \approx e^{-\lambda n}$ and $\beta_n \approx e^{-\gamma n}$ for $\lambda > 0, \gamma > 0$.
- For a given $\lambda > 0$, if $\lim_{n\to\infty} -\frac{1}{n} \ln \alpha_n(T_n) \geq \lambda$, maximize $\lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n)$

[7] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401–408, Mar. 1989

# Asymptotics: Chernoff regime (Gutman Test)

Chernoff regime:

- Exponential decay for $\alpha_n$ and $\beta_n$: i.e., $\alpha_n \approx e^{-\lambda n}$ and $\beta_n \approx e^{-\gamma n}$ for $\lambda > 0, \gamma > 0$.
- For a given $\lambda > 0$, if $\lim_{n\to\infty} -\frac{1}{n}\ln \alpha_n(T_n) \geq \lambda$, maximize $\lim_{n\to\infty} -\frac{1}{n}\ln \beta_n(T_n)$

Gutman Test [7]

- $P_{X^n}$ and $P_{Y^n}$ denote the types (empirical distributions) of sequences $X^n$ and $Y^n$
- If Jensen-Shannon (JS) divergence $D_{\mathrm{JS}}(P_{X^n} \| P_{Y^n}) < r$, then $H_0$ is accepted; else $H_1$ is accepted.
- The JS divergence between distributions $T$ and $R$ is defined as

$$D_{\mathrm{JS}}(T\|R) = \frac{1}{2}D_{\mathrm{KL}}\left(T \,\middle\|\, \frac{T+R}{2}\right) + \frac{1}{2}D_{\mathrm{KL}}\left(R \,\middle\|\, \frac{T+R}{2}\right)$$

- Gutman obtained maximum of $\lim_{n\to\infty} -\frac{1}{n}\ln \beta_n(T_n)$ as a function of $\lambda$.

[7] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401–408, Mar. 1989

# Two-Sample Tests in Machine Learning Literature

- Kernel-based test: Mostly through **Maximum Mean Discrepancy (MMD)** metric
- Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B. and Smola, A.,“A kernel two-sample test,” The Journal of Machine Learning Research (JMLR), 2012.

# Two-Sample Tests in Machine Learning Literature

- Kernel-based test: Mostly through **Maximum Mean Discrepancy (MMD)** metric
- Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B. and Smola, A.,"A kernel two-sample test," The Journal of Machine Learning Research (JMLR), 2012.
- Liu, Feng, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland. "Learning deep kernels for non-parametric two-sample tests," In International conference on machine learning, PMLR, 2020.
- Cheng, Xiuyuan, and Alexander Cloninger. "Classification logit two-sample testing by neural networks for differentiating near manifold densities," IEEE Transactions on Information Theory 68, 10, 2022.
- Biggs, F., Schrab, A. and Gretton, A., "MMD-FUSE: Learning and combining kernels for two-sample testing without data splitting," Advances in Neural Information Processing Systems, 36, 2023.
- Kübler, J.M., Stimper, V., Buchholz, S., Muandet, K. and Schölkopf, B., "AutoML two-sample test," Advances in Neural Information Processing Systems, 2022.
- Shekhar, S., Kim, I. and Ramdas, A., "A permutation-free kernel two-sample test." Advances in Neural Information Processing Systems, 35, pp.18168-18180, 2022.
- Podkopaev, A. and Ramdas, A., "Sequential predictive two-sample and independence testing," Advances in neural information processing systems, 36, pp.53275-53307, 2023.

$\vdots$

# Stein's Regime: Gutman Test

Stein's regime:

- Exponential decay only for $\beta_n$, $\beta_n \approx e^{-\gamma n}$
- For $\epsilon \in (0, 1)$, if $\alpha_n(T_n) \leq \epsilon$ as $n \to \infty$, then maximize $\beta' = \lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n)$

[8] L. Zhou, V. Y. F. Tan, and M. Motani, "Second-order asymptotically optimal statistical classification," *Information and Inference: A Journal of the IMA*, vol. 9, no. 1, pp. 81–111, Mar. 2020

[9] S. Zhu, B. Chen, Z. Chen, and P. Yang, "Asymptotically optimal one- and two-sample testing with kernels," *IEEE Transactions on Information Theory*, vol. 67, no. 4, pp. 2074–2092, Apr. 2021

# Stein's Regime: Gutman Test

Stein's regime:

- Exponential decay only for $\beta_n$, $\beta_n \approx e^{-\gamma n}$
- For $\epsilon \in (0,1)$, if $\alpha_n(T_n) \leq \epsilon$ as $n \to \infty$, then maximize $\beta' = \lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n)$

Stein's regime of Gutman test [8][9]

- $D_{\text{JS}}(P_{X^n} \| P_{Y^n}) < r_n$, then $H_0$ is accepted ($P_{X^n}$ and $P_{Y^n}$ are the types)
- Choose $r_n = \Theta(1/n)$; $\beta'_{\text{GT}}$ is twice the Bhattacharyya distance between $P_1$ and $P_2$, $D_B(P_1, P_2)$

$$D_B(P_1, P_2) = -\ln\Big(\sum_{z\in\mathcal{Z}} \sqrt{P_1(z)P_2(z)}\Big)$$

- $\beta'_{\text{GT}}$ is also characterized by the following minimization over all distributions $\mathcal{P}(\mathcal{Z})$ on $\mathcal{Z}$

$$\beta'_{\text{GT}} = \min_{P\in\mathcal{P}(\mathcal{Z})} \left\{D_{\text{KL}}(P\|P_1) + D_{\text{KL}}(P\|P_2)\right\} = 2D_B(P_1, P_2)$$

- Also shown that $2D_B(P_1, P_2)$ is the maximum $\beta'(T_n)$ among all tests

[8] L. Zhou, V. Y. F. Tan, and M. Motani, "Second-order asymptotically optimal statistical classification," *Information and Inference: A Journal of the IMA*, vol. 9, no. 1, pp. 81–111, Mar. 2020

[9] S. Zhu, B. Chen, Z. Chen, and P. Yang, "Asymptotically optimal one- and two-sample testing with kernels," *IEEE Transactions on Information Theory*, vol. 67, no. 4, pp. 2074–2092, Apr. 2021

# Our Work: Stein's regime of Two-Sample Test

Our work addresses two questions:

1. Why JS divergence? How about other divergences?

2. What about finer asymptotics due to the finite-sample size

$$-\ln \beta_n(T_n) = n2D_B(P_1, P_2) + o(n)$$

   Can we find finer asymptotic term $o(n)$?

# Our Work: Stein's regime of Two-Sample Test

Our work addresses two questions:

1. Why JS divergence? How about other divergences?

2. What about finer asymptotics due to the finite-sample size

$$-\ln \beta_n(T_n) = n2D_B(P_1, P_2) + o(n)$$

Can we find finer asymptotic term $o(n)$?

$$\text{Type I Error (Same distribution; declared different):} \quad \alpha_n(\mathrm{T}_n) \triangleq \sum_{(x^n, y^n) \notin \mathcal{A}(\mathrm{T}_n)} P(x^n)P(y^n)$$

$$\text{Type II Error (Different distributions; declared same):} \quad \beta_n(\mathrm{T}_n) \triangleq \sum_{(x^n, y^n) \in \mathcal{A}(\mathrm{T}_n)} P_1(x^n)P_2(y^n).$$

- Consider tests $T_n$ such that $\alpha_n(\mathrm{T}_n) \leq \epsilon$ as $n \to \infty$ and analyse $\lim_{n \to \infty} -\ln \beta_n(T_n)$

# Our Work: Stein's regime of Two-Sample Test

Our work addresses two questions:

1. Why JS divergence? How about other divergences?
2. What about finer asymptotics due to the finite-sample size
$$-\ln \beta_n(T_n) = n2D_B(P_1, P_2) + o(n)$$
Can we find finer asymptotic term $o(n)$?

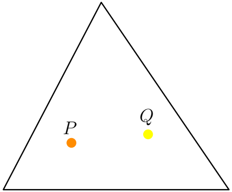
$$\text{Type I Error (Same distribution; declared different):} \quad \alpha_n(\mathrm{T}_n) \triangleq \sum_{(x^n, y^n) \notin \mathcal{A}(\mathrm{T}_n)} P(x^n)P(y^n)$$

$$\text{Type II Error (Different distributions; declared same):} \quad \beta_n(\mathrm{T}_n) \triangleq \sum_{(x^n, y^n) \in \mathcal{A}(\mathrm{T}_n)} P_1(x^n)P_2(y^n).$$

- Consider tests $T_n$ such that $\alpha_n(\mathrm{T}_n) \leq \epsilon$ as $n \to \infty$ and analyse $\lim_{n \to \infty} -\ln \beta_n(T_n)$

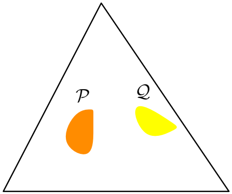
Thanks to Prof. Srikrishna Bhashyam (IIT Madras) for the initial discussions

# 2-Sample Test as a Composite Hypothesis Test



Classical Testing Problem

# 2-Sample Test as a Composite Hypothesis Test

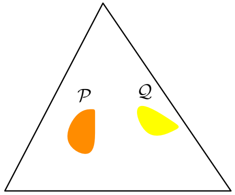


Composite Testing Problem

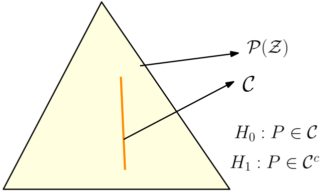
## A Case of Composite Hypothesis Testing

$$H_0 : \ \theta \in \Theta_0 \qquad \text{vs.} \qquad H_1 : \ \theta \in \Theta_1$$

# 2-Sample Test as a Composite Hypothesis Test



Composite Testing Problem



Robust Goodness of Fit (GoF) Test

## A Case of Composite Hypothesis Testing

$$H_0 : \ \theta \in \Theta_0 \qquad \text{vs.} \qquad H_1 : \ \theta \in \Theta \setminus \Theta_0$$

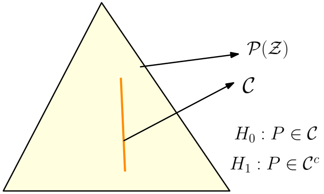
$L(\theta|z)$ - Likelihood function

$$L(\hat{\theta}_0|z) = \sup_{\theta \in \Theta_0} L(\theta|z),$$
$$L(\hat{\theta}|z) = \sup_{\theta \in \Theta} L(\theta|z).$$

Generalized Likelihood Ratio Test (GLRT) statistic:

$$T(z) = -2 \ln \frac{L(\hat{\theta}_0|z)}{L(\hat{\theta}|z)}$$

# GLRT for Robust GoF




Robust Goodness of Fit (GoF) Test

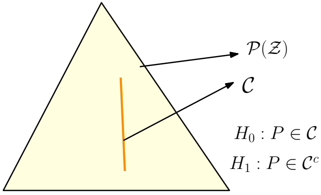
## Generalized Likelihood Ratio Test (GLRT) for Robust GoF

Test Statistics:

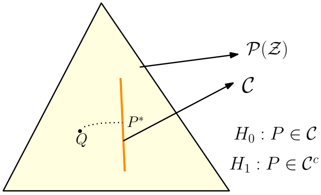
$$D_{\mathrm{KL}}^{\mathrm{ROB}}(T_{Z^n} \| \mathcal{C}) = \inf_{P \in \mathcal{C}} D_{\mathrm{KL}}(T_{Z^n} \| P)$$

Test: $D_{\mathrm{KL}}^{\mathrm{ROB}}(T_{Z^n} \| \mathcal{C}) \leq r_n$

# GLRT for Robust GoF




Robust Goodness of Fit (GoF) Test




Robust Goodness of Fit (GoF) Test

## Generalized Likelihood Ratio Test (GLRT) for Robust GoF

Test Statistics:

$$D_{\mathrm{KL}}^{\mathrm{ROB}}(T_{Z^n} \| \mathcal{C}) = \inf_{P \in \mathcal{C}} D_{\mathrm{KL}}(T_{Z^n} \| P)$$

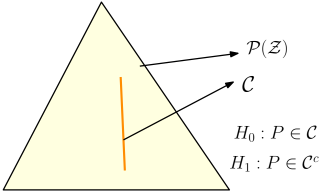
Test: $D_{\mathrm{KL}}^{\mathrm{ROB}}(T_{Z^n} \| \mathcal{C}) \le r_n$

$$\beta' = D_{\mathrm{KL}}(P^* \| Q)$$

$$P^* = \arg\min_{P \in \mathcal{C}} D_{\mathrm{KL}}(P \| Q)$$

- Project $Q$ onto the class $\mathcal{C}$. $P^*$ minimizes the KL Divergence between $Q$ and $\mathcal{C}$

# 2-Sample Test as an Extension of Robust GoF Test




Robust Goodness of Fit (GoF) Test

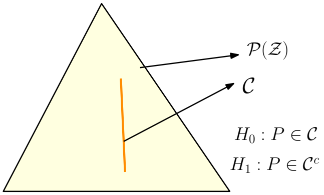
<u>GLRT for Robust GoF</u>

Test Statistics:

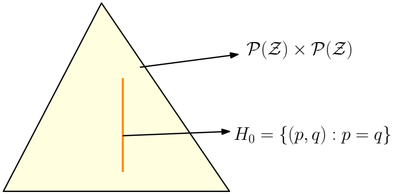
$$D^{\mathrm{ROB}}_{\mathrm{KL}}(T_{Z^n}\|\mathcal{C}) = \inf_{P\in\mathcal{C}} D_{\mathrm{KL}}(T_{Z^n}\|P)$$

Test: $D^{\mathrm{ROB}}_{\mathrm{KL}}(T_{Z^n}\|\mathcal{C}) \leq r_n$

# 2-Sample Test as an Extension of Robust GoF Test




Robust Goodness of Fit (GoF) Test




GLRT for Robust GoF

Test Statistics:

$$D_{\mathrm{KL}}^{\mathrm{ROB}}(T_{Z^n} \| \mathcal{C}) = \inf_{P \in \mathcal{C}} D_{\mathrm{KL}}(T_{Z^n} \| P)$$

Test: $D_{\mathrm{KL}}^{\mathrm{ROB}}(T_{Z^n} \| \mathcal{C}) \leq r_n$

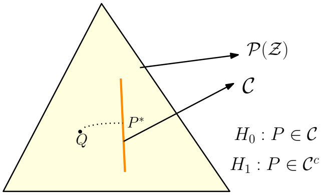
$$\Theta = \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z})$$

$$\Theta_0 = \{(P_1, P_2) \in \Theta : P_1 = P_2\}$$

**Test statistic of 2-sample test:**

$$\mathbf{T}(\mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}}) = \mathbf{4nD}_{\mathrm{JS}}(\mathbf{T}_{\mathbf{x}^{\mathbf{n}}} \| \mathbf{T}_{\mathbf{y}^{\mathbf{n}}})$$

# GLRT for 2-Sample Test

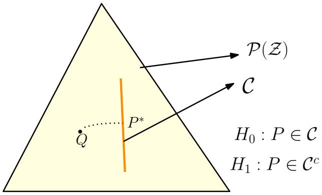



Robust Goodness of Fit (GoF) Test

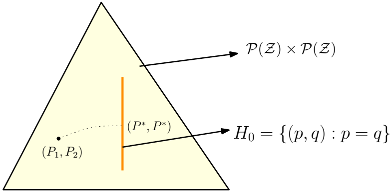
## First-order term for Robust GoF

$$\beta' = D_{\mathsf{KL}}(P^* \| Q)$$
$$P^* = \arg\min_{P \in \mathcal{C}} D_{\mathsf{KL}}(P \| Q)$$

- Project $Q$ onto the class $\mathcal{C}$. $P^*$ minimizes the KL Divergence between $Q$ and $\mathcal{C}$




Robust Goodness of Fit (GoF) Test




$$D_{\mathrm{KL}}^{\mathrm{ROB}}((T_{X^n}, T_{Y^n})\|\mathcal{C}) = \inf_{(P_1,P_2)\in\mathcal{C}} D_{\mathrm{KL}}((T_{X^n}, T_{Y^n})\|(P_1, P_2))$$
$$= 4nD_{\mathrm{JS}}(T_{X^n}\| T_{Y^n})$$
$$\beta' = D_{\mathrm{KL}}((P^*, P^*)\|(P_1, P_2))$$
$$(P^*, P^*) = \arg\min_{(P,P)\in\mathcal{C}} D_{\mathrm{KL}}((P, P)\|(P_1, P_2))$$

$$P^*(z) = \frac{\sqrt{P_1(z)P_2(z)}}{\sum_{z'\in\mathcal{Z}} \sqrt{P_1(z')P_2(z')}}$$

## First-order term for Robust GoF

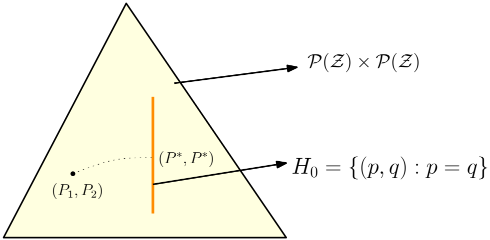
$$\beta' = D_{\mathrm{KL}}(P^*\|Q)$$
$$P^* = \arg\min_{P\in\mathcal{C}} D_{\mathrm{KL}}(P\|Q)$$

- Project $Q$ onto the class $\mathcal{C}$. $P^*$ minimizes the KL Divergence between $Q$ and $\mathcal{C}$

# Two-Sample Test: Asymptotics of GLRT

- For $\epsilon \in (0,1)$, if $\alpha_n(T_n) \leq \epsilon$ as $n \to \infty$, then maximize $\beta' = \lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n)$
- GLRT, Gutman test and JS divergence test are equivalent
- Connection to robust GoF as a bi-variate sequence

$$\begin{aligned}\beta' &= D_{\mathsf{KL}}((P^*, P^*)\|(P_1, P_2)) \\ &= D_{\mathsf{KL}}(P^*\|P_1) + D_{\mathsf{KL}}(P^*\|P_2)\end{aligned}$$

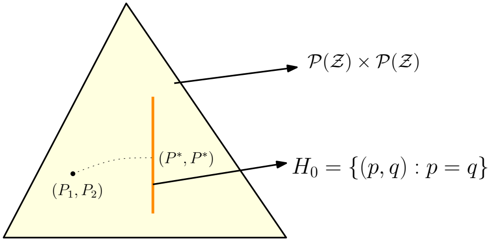


# Two-Sample Test: Asymptotics of GLRT

- For $\epsilon \in (0,1)$, if $\alpha_n(T_n) \leq \epsilon$ as $n \to \infty$, then maximize $\beta' = \lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n)$
- GLRT, Gutman test and JS divergence test are equivalent
- Connection to robust GoF as a bi-variate sequence

$$\begin{aligned}\beta' &= D_{\mathrm{KL}}((P^*, P^*)\|(P_1, P_2)) \\ &= D_{\mathrm{KL}}(P^*\|P_1) + D_{\mathrm{KL}}(P^*\|P_2)\end{aligned}$$

$$\begin{aligned}\beta' &= \min_{P\in\mathcal{P}(\mathcal{Z})} \{D_{\mathrm{KL}}(P\|P_1) + D_{\mathrm{KL}}(P\|P_2)\} \\ &= 2D_B(P_1, P_2)\end{aligned}$$

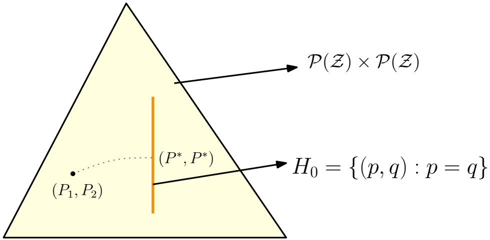


# Two-Sample Test: Asymptotics of GLRT

- For $\epsilon \in (0,1)$, if $\alpha_n(T_n) \le \epsilon$ as $n \to \infty$, then maximize $\beta' = \lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(T_n)$
- GLRT, Gutman test and JS divergence test are equivalent
- Connection to robust GoF as a bi-variate sequence

$$\begin{aligned}\beta' &= D_{\mathrm{KL}}((P^*, P^*)\|(P_1, P_2)) \\ &= D_{\mathrm{KL}}(P^*\|P_1) + D_{\mathrm{KL}}(P^*\|P_2)\end{aligned}$$

$$\begin{aligned}\beta' &= \min_{P\in\mathcal{P}(\mathcal{Z})} \left\{D_{\mathrm{KL}}(P\|P_1) + D_{\mathrm{KL}}(P\|P_2)\right\} \\ &= 2D_B(P_1, P_2)\end{aligned}$$




- GLRT (Gutman test) shows that $-\ln \beta_n(T_n) = n2D_B(P_1, P_2) + o(n)$
- Second-order term scales as $\sqrt{n}$, so we have $-\ln \beta_n(T_n) = n2D_B(P_1, P_2) + O(\sqrt{n})$

# Second-order Asymptotics of GLRT

- Since $-\ln \beta_n(T_n) = n2D_B(P_1, P_2) + O(\sqrt{n})$, we can define the second-order term as

$$\beta'' = \lim_{n\to\infty} \frac{-\ln \beta_n(\mathrm{T}_n) - n\beta'}{\sqrt{n}}, \qquad \text{where } \beta' = 2D_B(P_1, P_2)$$

# Second-order Asymptotics of GLRT

- Since $-\ln \beta_n(T_n) = n2D_B(P_1, P_2) + O(\sqrt{n})$, we can define the second-order term as

$$\beta'' = \lim_{n\to\infty} \frac{-\ln \beta_n(\mathrm{T}_n) - n\beta'}{\sqrt{n}}, \qquad \text{where } \beta' = 2D_B(P_1, P_2)$$

## Theorem

*For $0 < \epsilon < 1$, there exists a sequence $r_n$ satisfying $\alpha_n(\mathrm{T}_n^{D_{\mathrm{JS}}}(r_n)) \leq \epsilon$ such that, as $n \to \infty$,*

$$-\ln \beta_n(\mathrm{T}_n^{D_{\mathrm{JS}}}(r_n)) = n2D_B(P_1, P_2) - \sqrt{n(V_{KL}(P^*\|P_1) + V_{KL}(P^*\|P_2))}\sqrt{\mathrm{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)} + o\left(\sqrt{n}\right)$$

$\mathrm{Q}_{\chi^2_m}(\cdot)$: *tail probability of $\chi^2$ RV with $m$ degrees of freedom*

$$\beta''_{\mathrm{GT}} = -\sqrt{V_{\mathrm{KL}}(P^*\|P_1) + V_{\mathrm{KL}}(P^*\|P_2)}\sqrt{\mathrm{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)}$$

$V_{\mathrm{KL}}(P\|Q)$ - divergence variance:

$$V_{\mathrm{KL}}(P\|Q) \triangleq \sum_i P_i \left[\left(\ln \frac{P_i}{Q_i} - D_{\mathrm{KL}}(P\|Q)\right)^2\right] \qquad P^* = \arg\min_{P\in\mathcal{C}} D_{\mathrm{KL}}(P\|P_1) + D_{\mathrm{KL}}(P\|P_2)$$

# Key step: Convergence of the test static

- Convergence of $\frac{n}{2} D_{\mathrm{JS}}(P_{X^n} \| P_{Y^n})$ to a chi-square random variable with $|\mathcal{Z}| - 1$ degrees of freedom

$$\frac{n}{2} D_{\mathrm{JS}}(P_{X^n} \| P_{Y^n}) \xrightarrow{d} \chi^2_{|\mathcal{Z}|-1} \quad \text{as } n \to \infty$$

# Key step: Convergence of the test static

- Convergence of $\frac{n}{2}D_{\mathrm{JS}}(P_{X^n}\|P_{Y^n})$ to a chi-square random variable with $|\mathcal{Z}|-1$ degrees of freedom

$$\frac{n}{2}D_{\mathrm{JS}}(P_{X^n}\|P_{Y^n}) \xrightarrow{d} \chi^2_{|\mathcal{Z}|-1} \quad \text{as } n \to \infty$$

- Convergence of the test statistics helps to choose threshold $r_n = \Theta(\frac{1}{n})$ properly

# Key step: Convergence of the test static

- Convergence of $\frac{n}{2} D_{\mathrm{JS}}(P_{X^n} \| P_{Y^n})$ to a chi-square random variable with $|\mathcal{Z}| - 1$ degrees of freedom

$$\frac{n}{2} D_{\mathrm{JS}}(P_{X^n} \| P_{Y^n}) \xrightarrow{d} \chi^2_{|\mathcal{Z}|-1} \quad \text{as } n \to \infty$$

- Convergence of the test statistics helps to choose threshold $r_n = \Theta(\frac{1}{n})$ properly
- For more details... https://arxiv.org/pdf/2601.09196

## Second-Order Asymptotics of Two-Sample Tests

K V Harsha*, Jithin Ravi†, and Tobias Koch§
*Birla Institute of Technology and Science, Pilani, Hyderabad, India
†Indian Institute of Technology Kharagpur, Kharagpur, India
§Universidad Carlos III de Madrid, Leganés, Spain, and Gregorio Marañón Health Research Institute, Madrid, Spain
Emails: harsha.kv@hyderabad.bits-pilani.ac.in, jithin@ece.iitkgp.ac.in, tkoch@ing.uc3m.es

***Abstract*—In two-sampling testing, one observes two independent sequences of independent and identically distributed random variables distributed according to the distributions $P_1$ and $P_2$ and wishes to decide whether $P_1 = P_2$ (null hypothesis) or $P_1 \neq P_2$ (alternative hypothesis). The Gutman test for this problem compares the empirical distributions of the observed sequences and decides on the null hypothesis if the Jensen-Shannon (JS) divergence between these empirical distributions is below a given threshold. This paper proposes a generalization of the Gutman test, termed *divergence test*, which replaces the JS divergence by an arbitrary divergence. For this test, the exponential decay of the type-II error probability for a fixed type-I error probability is studied. First, it is shown that the divergence test achieves the optimal first-order exponent, irrespective of the choice of divergence. Second, it is demonstrated that divergence tests with invariant divergences achieve the same second-order asymptotics as the Gutman test. In addition, a connection between two-sample testing and robust goodness-of-fit testing is established.**



### I. Introduction

Two-sample testing is a fundamental problem in statistical hypothesis testing. In this problem, we observe two independent and identically distributed (i.i.d.) sequences $X^n$ and $Y^n$, generated according to the distributions $P_1$ and $P_2$,

via kernel-based methods, most notably through the Maximum Mean Discrepancy (MMD), where the MMD between the two observed sequences is compared against a threshold [3].

The asymptotic performance in Stein's regime of MMD-based two-sample testing was analyzed in [4]. Specifically, the authors studied the decay rate of the type-II error probability $\beta_n$ as $n \to \infty$ given that the type-I error $\alpha_n$ remains below a fixed $\epsilon \in (0,1)$. It was shown that the error exponent of $\beta_n$ of the MMD-based test is given by twice the Bhattacharyya distance. It was further shown that this exponent is, in fact, the largest achievable error exponent among all two-sample tests. A similar asymptotic analysis was carried out by Zhou, Tan and Motani for the Gutman test [5], who showed that the Gutman test also achieves an error exponent equal to twice the Bhattacharyya distance. It is noted that the Gutman test can be interpreted as comparing the Jensen–Shannon (JS) divergence [6, Eq. (7.8)] between the empirical distributions of $X^n$ and $Y^n$ to a threshold. They further studied the second-order asymptotics of the two-sample test under the assumption that the JS divergence between $P_1$ and $P_2$ exceeds a given $\lambda$ [5]. In particular, they characterized the first-order and second-

# Extending to other divergences: Divergence Test

- Consider using divergences such as KL divergence or Rényi divergence
- Can we extend the test to other divergences: Gutman-like test?

# Extending to other divergences: Divergence Test

- Consider using divergences such as KL divergence or Rényi divergence
- Can we extend the test to other divergences: Gutman-like test?

**Divergence test**: $T_n^D(r)$ for testing $H_0 : P_1 = P_2$, $H_1 : P_1 \neq P_2$

Upon observing $x^n$ and $y^n$:

- if $D(P_{X^n} \| P_{Y^n}) \leq r_n$, (for some $r_n > 0$), then accept $H_0$, i.e., $P_1 = P_2$
- if $D(P_{X^n} \| P_{Y^n}) > r_n$, then accept $H_1$, i.e., $P_1 \neq P_2$

# First-Order Term of Divergence Test

- First-order is the same as that of GLRT. If $\alpha_n(\mathsf{T}_n^D(r_n)) \leq \epsilon$ as $n \to \infty$, then

$$\lim_{n\to\infty} \frac{-\ln \beta_n(\mathsf{T}_n^D(r_n))}{n} = 2D_B(P_1, P_2)$$

# First-Order Term of Divergence Test

- First-order is the same as that of GLRT. If $\alpha_n(\mathsf{T}_n^D(r_n)) \leq \epsilon$ as $n \to \infty$, then

$$\lim_{n\to\infty} \frac{-\ln \beta_n(\mathsf{T}_n^D(r_n))}{n} = 2D_B(P_1, P_2)$$

- $\frac{n}{2} D(P_{X^n} \| P_{Y^n})$ converges in distribution to the generalized chi-square distribution $\chi^2_{\boldsymbol{\lambda}, k-1}$ with vector parameter $\boldsymbol{\lambda}$ and $k-1$ degrees of freedom
- Choose $r_n = \omega\left(\frac{1}{n}\right)$ and $r_n = o(1)$
- Such an $r_n$ gives that $\lim_{n\to\infty} \alpha_n\left(\mathsf{T}_n^D(r_n)\right) = 0$, and

$$\lim_{n\to\infty} -\frac{\ln \beta_n(\mathsf{T}_n^D(r_n))}{n} = 2D_B(P_1, P_2)$$

# First-Order Term of Divergence Test

- First-order is the same as that of GLRT. If $\alpha_n(\mathrm{T}_n^D(r_n)) \leq \epsilon$ as $n \to \infty$, then

$$\lim_{n\to\infty} \frac{-\ln \beta_n(\mathrm{T}_n^D(r_n))}{n} = 2D_B(P_1, P_2)$$

- $\frac{n}{2} D(P_{X^n} \| P_{Y^n})$ converges in distribution to the generalized chi-square distribution $\chi^2_{\boldsymbol{\lambda},k-1}$ with vector parameter $\boldsymbol{\lambda}$ and $k-1$ degrees of freedom

- Choose $r_n = \omega\left(\frac{1}{n}\right)$ and $r_n = o(1)$

- Such an $r_n$ gives that $\lim_{n\to\infty} \alpha_n\left(\mathrm{T}_n^D(r_n)\right) = 0$, and

$$\lim_{n\to\infty} -\frac{\ln \beta_n(\mathrm{T}_n^D(r_n))}{n} = 2D_B(P_1, P_2)$$

- To obtain the optimal second-order term, we need to choose $r_n = \Theta(\frac{1}{n})$

- We consider the second-order term for the tests with invariant divergences

# Second-order for Test with Invariant Divergence

- For an invariant divergence, $\mathbf{\Sigma}_{\mathbf{P}}^{-1/2}\boldsymbol{A}_{D,\mathbf{P}}\mathbf{\Sigma}_{\mathbf{P}}^{-1/2}$ is a multiple of identity matrix.

**Theorem**

*For invariant divergence $D$, for $0 < \epsilon < 1$, $\exists\, r_n$ satisfying $\alpha_n(\mathrm{T}_n^D(r_n)) \leq \epsilon$ such that, as $n \to \infty$,*

$$-\ln \beta_n(\mathrm{T}_n^D(r_n)) = n2D_B(P_1, P_2) - \sqrt{n(V_{KL}(P^*\|P_1) + V_{KL}(P^*\|P_2))}\sqrt{\mathrm{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)} + o\left(\sqrt{n}\right)$$

$\mathrm{Q}_{\chi^2_m}(\cdot)$: *tail probability of $\chi^2$ RV with $m$ degrees of freedom*

- Test with any invariant divergence also achieves the same second-order term as the Gutman test

$V_{\mathrm{KL}}(P\|Q)$ - divergence variance:

$$V_{\mathrm{KL}}(P\|Q) \triangleq \sum_i P_i \left[\left(\ln \frac{P_i}{Q_i} - D_{\mathrm{KL}}(P\|Q)\right)^2\right] \qquad P^* = \arg\min_{P\in\mathcal{C}} D_{\mathrm{KL}}(P\|P_1) + D_{\mathrm{KL}}(P\|P_2)$$

# On Tests using Non-Invariant Divergence

- **Invariant Divergence**: $\frac{n}{2}D(P_{X^n}\|P_{Y^n})$ converges to a chi-square random variable

$$\frac{n}{2}D(P_{X^n}\|P_{Y^n}) \xrightarrow{d} \chi^2_{|\mathcal{Z}|-1} \quad \text{as } n \to \infty$$

- **Non-Invariant Divergence**: $nD(P_{X^n}\|P_{Y^n})$ converges to a generalized chi-square random variable

$$\frac{n}{2}D(P_{X^n}\|P_{Y^n}) \xrightarrow{d} \chi^2_{\boldsymbol{\lambda},|\mathcal{Z}|-1} \quad \text{as } n \to \infty$$

# On Tests using Non-Invariant Divergence

- **Invariant Divergence**: $\frac{n}{2}D(P_{X^n}\|P_{Y^n})$ converges to a chi-square random variable

$$\frac{n}{2}D(P_{X^n}\|P_{Y^n}) \xrightarrow{d} \chi^2_{|\mathcal{Z}|-1} \quad \text{as } n \to \infty$$

- **Non-Invariant Divergence**: $nD(P_{X^n}\|P_{Y^n})$ converges to a generalized chi-square random variable

$$\frac{n}{2}D(P_{X^n}\|P_{Y^n}) \xrightarrow{d} \chi^2_{\boldsymbol{\lambda},|\mathcal{Z}|-1} \quad \text{as } n \to \infty$$

- $\boldsymbol{\lambda}$ are the eigenvalues of the matrix $\boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2}\boldsymbol{A}_{D,\mathbf{P}}\boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2}$
- Since $\boldsymbol{\lambda}$ is a function of $P$, test not knowing $P$ makes it challenging to find a suitable threshold $r_n > 0$.
- **Invariant Divergence**: Convergence of $\frac{n}{2}D(P_{X^n}\|P_{Y^n})$ under $H_0$ is independent of underlying $P$

# Conclusions

One-Sample Test: Observe $Z^n$

- GoF: Observe $H_0 : P$ vs. $H_1 : \mathcal{Q} = \mathcal{P} \setminus \{P\}$
- Divergence test: If $D(t_{Z^n} \| P) \leq r_n$, then $H_0$

Two-Sample Test: Observe $X^n \sim P_1$ and $Y^n \sim P_2$

- To test: $H_0 : P_1 = P_2$ vs. $H_1 : P_1 \neq P_2$
- Divergence test: If $D(P_{X^n} \| P_{Y^n}) \leq r_n$, then $H_0$

# Conclusions

**One-Sample Test: Observe $Z^n$**

- GoF: Observe $H_0 : P$ vs. $H_1 : \mathcal{Q} = \mathcal{P} \setminus \{P\}$
- Divergence test: If $D(t_{Z^n} \| P) \leq r_n$, then $H_0$
- Is first-order optimal: Achieves $D_{\mathrm{KL}}(P \| Q)$
- Second-Order asymptotics:
  - **Invariant** $D$: same second-order term as the Hoeffding test
  - **Non-invariant** $D$: Can outperform the Hoeffding test

**Two-Sample Test: Observe $X^n \sim P_1$ and $Y^n \sim P_2$**

- To test: $H_0 : P_1 = P_2$ vs. $H_1 : P_1 \neq P_2$
- Divergence test: If $D(P_{X^n} \| P_{Y^n}) \leq r_n$, then $H_0$
- Is first-order optimal: Achieves $2D_B(P_1, P_2)$
- Second-Order asymptotics:
  - **Invariant** $D$: same second-order term as Gutman test
  - **Non-invariant** $D$: Second-order term yet to be analyzed

# Conclusions

One-Sample Test: Observe $Z^n$

- GoF: Observe $H_0 : P$ vs. $H_1 : \mathcal{Q} = \mathcal{P} \setminus \{P\}$
- Divergence test: If $D(t_{Z^n} \| P) \leq r_n$, then $H_0$
- Is first-order optimal: Achieves $D_{\mathrm{KL}}(P\|Q)$
- Second-Order asymptotics:
  - **Invariant** $D$: same second-order term as the Hoeffding test
  - **Non-invariant** $D$: Can outperform the Hoeffding test

Two-Sample Test: Observe $X^n \sim P_1$ and $Y^n \sim P_2$

- To test: $H_0 : P_1 = P_2$ vs. $H_1 : P_1 \neq P_2$
- Divergence test: If $D(P_{X^n} \| P_{Y^n}) \leq r_n$, then $H_0$
- Is first-order optimal: Achieves $2D_B(P_1, P_2)$
- Second-Order asymptotics:
  - **Invariant** $D$: same second-order term as Gutman test
  - **Non-invariant** $D$: Second-order term yet to be analyzed

- Still Open: One and Two-Sample tests beyond finite alphabets

# Conclusions

One-Sample Test: Observe $Z^n$

- GoF: Observe $H_0 : P$ vs. $H_1 : \mathcal{Q} = \mathcal{P} \setminus \{P\}$
- Divergence test: If $D(t_{Z^n} \| P) \leq r_n$, then $H_0$
- Is first-order optimal: Achieves $D_{\mathrm{KL}}(P\|Q)$
- Second-Order asymptotics:
  - **Invariant** $D$: same second-order term as the Hoeffding test
  - **Non-invariant** $D$: Can outperform the Hoeffding test

Two-Sample Test: Observe $X^n \sim P_1$ and $Y^n \sim P_2$

- To test: $H_0 : P_1 = P_2$ vs. $H_1 : P_1 \neq P_2$
- Divergence test: If $D(P_{X^n} \| P_{Y^n}) \leq r_n$, then $H_0$
- Is first-order optimal: Achieves $2D_B(P_1, P_2)$
- Second-Order asymptotics:
  - **Invariant** $D$: same second-order term as Gutman test
  - **Non-invariant** $D$: Second-order term yet to be analyzed

- Still Open: One and Two-Sample tests beyond finite alphabets

# Thank You