Supplementary Material

A Proof of Corollary 1

We first present the two lemmas used in the proof of Theorem 3: one establishes the convergence of $d_k(P, \hat{P}_m)$ and the other describes the lower semi-continuity of the KLD.

Lemma 2. *[47, 48] Assume $0 \le k(\cdot,\cdot) \le K$. Given y^m i.i.d. $\sim P$, denote by $\hat{P}_m$ the empirical measure of y^m. It follows that*

$$\mathbf{P}_{y^m}\left(d_k(P,\hat{P}_m) > (2K/m)^{1/2} + \epsilon\right) \le \exp\left(-\frac{\epsilon^2 m}{2K}\right).$$

Lemma 3 ([50])**.** *For a fixed $Q \in \mathcal{P}$, $D(\cdot\|Q)$ is a lower semi-continuous function w.r.t. the weak topology of $\mathcal{P}$. That is, for any $\epsilon > 0$, there exists a neighborhood $U \subset \mathcal{P}$ of P such that for any $P' \in U$, $D(P'\|Q) \ge D(P\|Q) - \epsilon$ if $D(P\|Q) < \infty$, and $D(P'\|Q) \to \infty$ as P' tends to P if $D(P\|Q) = \infty$.*

Proof of Corollary 1. Since $0 \le k(\cdot,\cdot) \le K$, we have

$$\left|d_u^2(P,\hat{Q}_n) - d_k^2(P,\hat{Q}_n)\right| = \left|\frac{1}{n^2(n-1)}\sum_{i=1}^n\sum_{j\neq i} k(x_i,x_j) - \frac{1}{n^2}\sum_{i=1}^n k(x_i,x_i)\right| \le K/n.$$

It then holds that

$$\left\{x^n : d_k^2(P,\hat{Q}_n) \le \gamma_n^2\right\} \subset \left\{x^n : d_u^2(P,\hat{Q}_n) \le \gamma_n^2 + K/n\right\} \subset \left\{x^n : d_k^2(P,\hat{Q}_n) \le \gamma_n^2 + 2K/n\right\}.$$

Thus, under $H_0 : P = Q$, we have

$$P\left(d_u^2(P,\hat{Q}_n) > \gamma_n^2 + K/n\right) \le P\left(d_k^2(P,\hat{Q}_n) > \gamma_n^2\right) \le \alpha,$$

where the last inequality is from Lemma 2 and the fact that $d_k(P,\hat{Q}_n) \ge 0$. The type-II error exponent follows from

$$\begin{aligned}
&\liminf_{n\to\infty} -\frac{1}{n}\log Q\left(d_u^2(P,\hat{Q}_n) \le \gamma_n^2 + K/n\right)\\
\ge\ &\liminf_{n\to\infty} -\frac{1}{n}\log Q\left(d_k^2(P,\hat{Q}_n) \le \gamma_n^2 + 2K/n\right)\\
\ge\ &D(P\|Q).
\end{aligned}$$

The last inequality can be shown by similar argument of Eq. (1) because $\gamma_n^2 + 2K/n \to 0$ as $n \to \infty$. Applying Chernoff-Stein lemma completes the proof. □

B Proof of Theorem 4

We use a result from [25] to verify the two-sample test to be level α.

Lemma 4 ([25, Theorem 7])**.** *Let $P, Q, y^m, x^n, \hat{P}_m, \hat{Q}_n$ be defined in Theorem 4. Assume $0 \le k(\cdot,\cdot) \le K$. Then under the null hypothesis $H_0 : P = Q$,*

$$\mathbf{P}_{y^m x^n}\left(d_k(\hat{P}_m,\hat{Q}_n) > 2(K/m)^{1/2} + 2(K/n)^{1/2} + \epsilon\right) \le 2\exp\left(-\frac{\epsilon^2 mn}{2K(m+n)}\right).$$

Proof of Theorem 4. That the two-sample test is level α can be verified by the above lemma. The rest is to show the type-II error exponent being $D(P\|Q)$.

We can write the type-II error probability as

$$\mathbf{P}_{y^m x^n}\left(d_k(\hat{P}_m,\hat{Q}_n) \le \gamma_{m,n}\right) = \beta_{m,n}^u + \beta_{m,n}^l,$$

where

$$\gamma'_{m,n} = \sqrt{2K/m} + \sqrt{2KnD(P\|Q)/m}$$
$$\beta^u_{m,n} = \mathbf{P}_{y^m x^n}\left(d_k(\hat{P}_m, \hat{Q}_n) \le \gamma_{m,n}, d_k(P, \hat{P}_m) > \gamma'_{m,n}\right),$$
$$\beta^l_{m,n} = \mathbf{P}_{y^m x^n}\left(d_k(\hat{P}_m, \hat{Q}_n) \le \gamma_{m,n}, d_k(P, \hat{P}_m) \le \gamma'_{m,n}\right).$$

It suffices to show that $\max\{\beta^u_{m,n}, \beta^l_{m,n}\}$ decreases exponentially as n scales. We first have

$$\beta^u_{m,n} \le \mathbf{P}_{y^m}\left(d_k(P, \hat{P}_m) > \gamma'_{m,n}\right) \le e^{-nD(P\|Q)}, \tag{3}$$

where the last inequality is due to Lemma 2. Thus, $\beta^u_{m,n}$ vanishes at least exponentially fast with the error exponent being $D(P\|Q)$.

For $\beta^l_{m,n}$, we have

$$\begin{aligned}
\beta^l_{m,n} &= \sum_{\{\hat{P}_m : d_k(P,\hat{P}_m) \le \gamma'_{m,n}\}} P\left(\hat{P}_m\right) Q\left(d_k(\hat{P}_m, \hat{Q}_n) < \gamma_{m,n}\right) \\
&= \left(\sum_{\hat{P}_m : d_k(P,\hat{P}_m) \le \gamma'_{m,n}} P(\hat{P}_m)\right) \sup_{\{\hat{P}_m : d_k(P,\hat{P}_m) \le \gamma'_{m,n}\}} Q\left(d_k(\hat{P}_m, \hat{Q}_n) < \gamma_{m,n}\right) \\
&\le \sup_{\{\hat{P}_m : d_k(P,\hat{P}_m) \le \gamma'_{m,n}\}} Q\left(d_k(\hat{P}_m, \hat{Q}_n) < \gamma_{m,n}\right) \\
&\le Q\left(d_k(P, \hat{Q}_n) \le \gamma_{m,n} + \gamma'_{m,n}\right),
\end{aligned}$$

where the last inequality is from the triangle inequality for metric d_k. Similar to Eq. (1), we get

$$\liminf_{n\to\infty} -\frac{1}{n} \log \beta^l_{m,n} \ge D(P\|Q),$$

because $\gamma_{m,n} + \gamma'_{m,n} \to 0$ as $n \to \infty$. Together with Eq. (3), we have under $H_1 : P \ne Q$,

$$\liminf_{n\to\infty} -\frac{1}{n} \log \mathbf{P}_{y^m x^n}\left(d_k(\hat{P}_m, \hat{Q}_n) \le \gamma_{m,n}\right) \ge D(P\|Q).$$

We next show the other direction under H_1. We can write

$$\begin{aligned}
\mathbf{P}_{y^m x^n}\left(d_k(\hat{P}_m, \hat{Q}_n) \le \gamma_{m,n}\right) &\overset{(a)}{\ge} \mathbf{P}_{y^m x^n}\left(d_k(\hat{P}_m, P) \le \gamma'_m, d_k(P, \hat{Q}_n) \le \gamma'_n\right) \\
&= P\left(d_k(\hat{P}_m, P) \le \gamma'_m\right) Q\left(d_k(P, \hat{Q}_n) \le \gamma'_n\right),
\end{aligned}$$

where (a) is because d_k is a metric, and we choose $\gamma'_m = \sqrt{2K/m}\left(1 + \sqrt{-\log\alpha}\right)$ and $\gamma'_n = \sqrt{2K/n}\left(1 + \sqrt{-\log\alpha}\right)$ so that $\gamma_{m,n} > \gamma'_m + \gamma'_n$. Then Lemma 2 gives $P(d_k(P, \hat{P}_m) \le \gamma'_m) > 1 - \alpha$ and $P(d_k(P, \hat{Q}_n) \le \gamma'_n) > 1 - \alpha$, where the latter implies that $d_k(P, \hat{Q}_n) \le \gamma'_n$ is a level α test for testing $H_0 : x^n \sim P$ and $H_1 : x^n \sim Q$ with $P \ne Q$. Together with Chernoff-Stein Lemma, we get

$$\begin{aligned}
&\liminf_{n\to\infty} -\frac{1}{n} \log \mathbf{P}_{y^m x^n}\left(d_k(\hat{P}_m, \hat{Q}_n) \le \gamma_{m,n}\right) \\
\le &\liminf_{n\to\infty} -\frac{1}{n} \log\left(P\left(d_k(\hat{P}_m, P) \le \gamma'_m\right) Q\left(d_k(P, \hat{Q}_n) \le \gamma'_n\right)\right) \\
\le &\liminf_{n\to\infty} -\frac{1}{n} \log(1-\alpha) + \liminf_{n\to\infty} -\frac{1}{n} \log Q\left(d_k(P, \hat{Q}_n) \le \gamma'_n\right) \\
\le &\, D(P\|Q).
\end{aligned}$$

The proof is complete. □

C Proof of the Extended Sanov's Theorem

Our proof is inspired by [16] which proved the original Sanov's theorem w.r.t. the τ-topology. We first prove the result with a finite sample space and then extend it to the case with general Polish space. The prerequisites are two combinatorial lemmas that are standard tools in information theory.

For a positive integer t, let $\mathcal{P}_m(t)$ denote the set of probability distributions defined on $\{1,\dots,t\}$ of form $P=\left(\frac{m_1}{m},\cdots,\frac{m_t}{m}\right)$, with integers $m_1,\dots,m_t$. Stated below are the two lemmas.

Lemma 5 ([15, Theorem 11.1.1]). $|\mathcal{P}_m(t)|\le (m+1)^t$.

Lemma 6 ([15, Theorem 11.1.4]). *Assume y^m i.i.d. $\sim R$ where R is a distribution defined on $\{1,\dots,t\}$. For any $P\in\mathcal{P}_m(t)$, the probability of the empirical distribution $\hat{P}_m$ of y^m equal to P satisfies*

$$(m+1)^{-t}e^{-mD(P\|R)}\le \mathbf{P}_{y^m}(\hat{P}_m=P)\le e^{-mD(P\|R)}.$$

C.1 Finite Sample Space

Upper bound Let t denote the cardinality of $\mathcal{X}$. Without loss of generality, assume that $\inf_{(R,S)\in\operatorname{int}\Gamma} cD(R\|P)+(1-c)D(S\|Q)<\infty$. Hence, the open set $\operatorname{int}\Gamma$ is non-empty. As $0<c=\lim_{m,n\to\infty}\frac{m}{m+n}<1$, we can find m_0 and n_0 such that there exists $(P'_m,Q'_n)\in\operatorname{int}\Gamma\cap\mathcal{P}_m(t)\times\mathcal{P}_m(t)$ for all $m>m_0$ and $n>n_0$, and that $cD(P'_m\|P)+(1-c)D(Q'_n\|Q)\to\inf_{(R,S)\in\operatorname{int}\Gamma} cD(R\|P)+(1-c)D(S\|Q)$ as $m,n\to\infty$. Then we have, with $m>m_0$ and $n>n_0$,

$$\begin{aligned}
\mathbf{P}_{y^m x^n}((\hat{P}_m,\hat{Q}_n)\in\Gamma) &= \sum_{(R,S)\in\Gamma\cap\mathcal{P}_m(t)\times\mathcal{P}_m(t)} \mathbf{P}_{y^m x^n}(\hat{P}_m=R,\hat{Q}_n=S)\\
&\ge \sum_{(R,S)\in\operatorname{int}\Gamma\cap\mathcal{P}_m(t)\times\mathcal{P}_m(t)} \mathbf{P}_{y^m x^n}(\hat{P}_m=R,\hat{Q}_n=S)\\
&\ge \mathbf{P}_{y^m x^n}(\hat{P}_m=P'_m,\hat{Q}_n=Q'_n)\\
&= \mathbf{P}_{y^m}(\hat{P}_m=P'_m)\,\mathbf{P}_{x^n}(\hat{Q}_n=Q'_n)\\
&\ge (m+1)^{-t}(n+1)^{-t}e^{-mD(P'_m\|P)}e^{-nD(Q'_n\|Q)},
\end{aligned}$$

where the last inequality is from Lemma 6. It follows that

$$\begin{aligned}
&\limsup_{m,n\to\infty} -\frac{1}{m+n}\log\mathbf{P}_{y^m x^n}((\hat{P}_m,\hat{Q}_n)\in\Gamma)\\
\le&\lim_{m,n\to\infty}\frac{1}{m+n}\left(-t\log((m+1)(n+1))+mD(P'_m\|P)+nD(Q'_n\|Q)\right)\\
=&\lim_{m,n\to\infty}\frac{1}{m+n}\left(mD(P'_n\|P)+nD(Q'_n\|Q)\right)\\
=&\inf_{(R,S)\in\operatorname{int}\Gamma} cD(R\|P)+(1-c)D(S\|Q).
\end{aligned}$$

Lower bound

$$\begin{aligned}
\mathbf{P}_{y^m x^n}((\hat{P}_m,\hat{Q}_n)\in\Gamma) &= \sum_{(R,S)\in\Gamma\cap\mathcal{P}_m(t)\times\mathcal{P}_m(t)} \mathbf{P}_{y^m}(\hat{P}_m=R)\,\mathbf{P}_{x^n}(\hat{Q}_n=S)\\
&\overset{(a)}{\le} \sum_{(R,S)\in\Gamma\cap\mathcal{P}_m(t)\times\mathcal{P}_n(t)} e^{-mD(R\|P)}e^{-nD(S\|Q)}\\
&\overset{(b)}{\le} (m+1)^t(n+1)^t\sup_{(R,S)\in\Gamma} e^{-mD(R\|P)}e^{-nD(S\|Q)},
\end{aligned}\tag{4}$$

where (a) and (b) are due to Lemma 6 and Lemma 5, respectively. This gives

$$\liminf_{m,n\to\infty} -\frac{1}{m+n}\log\mathbf{P}_{y^m x^n}((\hat{P}_m,\hat{Q}_n)\in\Gamma)\ge\inf_{(R,S)\in\Gamma} cD(R\|P)+(1-c)D(S\|Q),$$

and hence the lower bound by noting that $\Gamma\in\operatorname{cl}\Gamma$. Indeed, when the right hand side is finite, the infimum over Γ equals the infimum over $\operatorname{cl}\Gamma$ as a result of the continuity of KLD for finite alphabets.

C.2 Polish Sample Space

We consider the general case with $\mathcal{X}$ being a Polish space. Now $\mathcal{P}$ is the space of probability measures on $\mathcal{X}$ endowed with the topology of weak convergence. To proceed, we introduce another topology on $\mathcal{P}$ and an equivalent definition of the KLD.

τ-topology: denote by Π the set of all partitions $\mathcal{A} = \{A_1, \ldots, A_t\}$ of $\mathcal{X}$ into a finite number of measurable sets A_i. For $P \in \mathcal{P}$, $\mathcal{A} \in \Pi$, and $\zeta > 0$, denote

$$U(P, \mathcal{A}, \zeta) = \{P' \in \mathcal{P} : |P'(A_i) - P(A_i)| < \zeta, i = 1, \ldots, t\}. \tag{5}$$

The τ-topology on $\mathcal{P}$ is the coarsest topology in which the mapping $P \to P(F)$ are continuous for every measurable set $F \subset \mathcal{X}$. A base for this topology is the collection of the sets (5). We will use $\mathcal{P}_\tau$ when we refer to $\mathcal{P}$ endowed with this τ-topology, and write the interior and closure of a set $\Gamma \in \mathcal{P}_\tau$ as $\mathrm{int}_\tau \Gamma$ and $\mathrm{cl}_\tau \Gamma$, respectively. We remark that the τ-topology is stronger than the weak topology: any open set in $\mathcal{P}$ w.r.t. weak topology is also open in $\mathcal{P}_\tau$ (see more details in [16, 19]). The product topology on $\mathcal{P}_\tau \times \mathcal{P}_\tau$ is determined by the base of the form of

$$U(P, \mathcal{A}_1, \zeta_1) \times U(Q, \mathcal{A}_2, \zeta_2),$$

for $(P, Q) \in \mathcal{P}_\tau \times \mathcal{P}_\tau$, $\mathcal{A}_1, \mathcal{A}_2 \in \Pi$, and $\zeta_1, \zeta_2 > 0$. We still use $\mathrm{int}_\tau(\Gamma)$ and $\mathrm{cl}_\tau(\Gamma)$ to denote the interior and closure of a set $\Gamma \subset \mathcal{P}_\tau \times \mathcal{P}_\tau$. As there always exists $\mathcal{A} \in \Pi$ that refines both $\mathcal{A}_1$ and $\mathcal{A}_2$, any element from the base has an open subset

$$\tilde{U}(P, Q, \mathcal{A}, \zeta) := U(P, \mathcal{A}, \zeta) \times U(Q, \mathcal{A}, \zeta) \subset \mathcal{P}_\tau \times \mathcal{P}_\tau,$$

for some $\zeta > 0$.

Another definition of the KLD: an equivalent definition of the KLD will also be used:

$$D(P\|Q) = \sup_{\mathcal{A} \in \Pi} \sum_{i=1}^{t} P(A_i) \log \frac{P(A_i)}{Q(A_i)} = \sup_{\mathcal{A} \in \Pi} D(P^{\mathcal{A}} \| Q^{\mathcal{A}}),$$

with the conventions $0 \log 0 = 0 \log \frac{0}{0} = 0$ and $a \log \frac{a}{0} = +\infty$ if $a > 0$. Here $P^{\mathcal{A}}$ denotes the discrete probability measure $(P(A_1), \ldots, P(A_t))$ obtained from probability measure P and partition $\mathcal{A}$. It is not hard to verify that for $0 < c < 1$,

$$\begin{aligned} cD(R\|P) + (1-c)D(S\|Q) &= c \sup_{\mathcal{A}_1 \in \Pi} D(R^{\mathcal{A}_1} \| P^{\mathcal{A}_1}) + (1-c) \sup_{\mathcal{A}_2 \in \Pi} D(S^{\mathcal{A}_2} \| Q^{\mathcal{A}_2}) \\ &= \sup_{\mathcal{A} \in \Pi} \left(cD\left(R^{\mathcal{A}} \| P^{\mathcal{A}}\right) + (1-c) D\left(S^{\mathcal{A}} \| Q^{\mathcal{A}}\right) \right), \end{aligned} \tag{6}$$

due to the existence of $\mathcal{A}$ that refines both $\mathcal{A}_1$ and $\mathcal{A}_2$ and the log-sum inequality [15].

We are ready to show the extended Sanov's theorem with Polish space.

Upper bound It suffices to consider only non-empty open Γ. If Γ is open in $\mathcal{P} \times \mathcal{P}$, then Γ is also open in $\mathcal{P}_\tau \times \mathcal{P}_\tau$. Therefore, for any $(R, S) \in \Gamma$, there exists a finite (measurable) partition $\mathcal{A} = \{A_1, \ldots, A_t\}$ of $\mathcal{X}$ and $\zeta > 0$ such that

$$\tilde{U}(R, S, \mathcal{A}, \zeta) = \{(R', S') : |R(A_i) - R'(A_i)| < \zeta, |S(A_i) - S'(A_i)| < \zeta, i = 1, \ldots, t\} \subset \Gamma. \tag{7}$$

Define the function $T : \mathcal{X} \to \{1, \ldots, t\}$ with $T(x) = i$ for $x \in A_i$. Then $(\hat{P}_m, \hat{Q}_n) \in \tilde{U}(R, S, \mathcal{A}, \zeta)$ with $R, S \in \Gamma$ if and only if the empirical measures $\hat{P}_m^\circ$ of $\{T(y_1), \ldots, T(y_m)\} := T(y^m)$ and $\hat{Q}_n^\circ$ of $\{T(x_1), \ldots, T(x_n)\} := T(x^n)$ lie in

$$U^\circ(R, S, \mathcal{A}, \zeta) = \{(R^\circ, S^\circ) : |R^\circ(i) - R(A_i)| < \zeta, |S^\circ(i) - S(A_i)| < \zeta, i = 1, \ldots, t\} \subset \mathbb{R}^t \times \mathbb{R}^t.$$

Thus, we have

$$\begin{aligned} \mathbf{P}_{y^m x^n}((\hat{P}_m, \hat{Q}_n) \in \Gamma) &\geq \mathbf{P}_{y^m x^n}((\hat{P}_m, \hat{Q}_n) \in \tilde{U}(R, S, \mathcal{A}, \zeta)) \\ &= \mathbf{P}_{T(y^m) T(x^n)}((\hat{P}_m^\circ, \hat{Q}_n^\circ) \in U^\circ(R, S, \mathcal{A}, \zeta)). \end{aligned}$$

As $T(x)$ and $T(y)$ takes values from a finite alphabet and $U^\circ(R,S,\mathcal{A},\zeta)$ is open, we obtain that

$$\begin{aligned}
&\limsup_{m,n\to\infty} -\frac{1}{m+n}\log \mathbf{P}_{y^m x^n}((\hat{P}_m,\hat{Q}_n)\in\Gamma)\\
\le\ &\limsup_{m,n\to\infty} -\frac{1}{m+n}\log \mathbf{P}_{T(y^m)T(x^n)}((\hat{P}^\circ_m,\hat{Q}^\circ_n)\in U^\circ(R,S,\mathcal{A},\zeta))\\
\le\ &\inf_{(R^\circ,S^\circ)\in U^\circ(R,S,\mathcal{A},\zeta)} cD(R^\circ\|P^{\mathcal{A}})+(1-c)D(S^\circ\|Q^{\mathcal{A}})\\
=\ &\inf_{(R',S')\in \tilde{U}(R,S,\mathcal{A},\zeta)} cD(R'^{\mathcal{A}}\|P^{\mathcal{A}})+(1-c)D(S'^{\mathcal{A}}\|Q^{\mathcal{A}})\\
\le\ &cD(R\|P)+(1-c)D(S\|Q),
\end{aligned}\tag{8}$$

where we have used definition of KLD in Eq. (6) and $(R,S)\in\tilde{U}(R,S,\mathcal{A},\zeta)$ in the last inequality. As (R,S) is arbitrary in Γ, the lower bound is established by taking infimum over Γ.

Lower bound With notations

$$\Gamma^{\mathcal{A}}=\{(R^{\mathcal{A}},S^{\mathcal{A}}):(R,S)\in\Gamma\},\ \Gamma(\mathcal{A})=\{(R,S):(R^{\mathcal{A}},S^{\mathcal{A}})\in\Gamma^{\mathcal{A}}\},$$

where $\mathcal{A}=\{A_1,\dots,A_t\}$ is a finite partition, it holds that

$$\begin{aligned}
&\mathbf{P}_{y^m x^n}((\hat{P}_m,\hat{Q}_n)\in\Gamma)\\
\le\ &\mathbf{P}_{y^m x^n}((\hat{P}_m,\hat{Q}_n)\in\Gamma(\mathcal{A}))\\
=\ &\mathbf{P}_{y^m x^n}((\hat{P}^{\mathcal{A}}_m,\hat{Q}^{\mathcal{A}}_n)\in\Gamma^{\mathcal{A}}\cap\mathcal{P}_n(t)\times\mathcal{P}_m(t))\\
\le\ &(n+1)^t(m+1)^t\max_{(R^\circ,S^\circ)\in\Gamma^{\mathcal{A}}\cap\mathcal{P}_n(t)\times\mathcal{P}_m(t)}\mathbf{P}_{y^m x^n}\left(\hat{P}_n=R^\circ,\hat{Q}_m=S^\circ\right)\\
\le\ &(n+1)^t(m+1)^t\exp\left(-\inf_{(R,S)\in\Gamma}\left(nD(R^{\mathcal{A}}\|P^{\mathcal{A}})+mD(S^{\mathcal{A}}\|Q^{\mathcal{A}})\right)\right),
\end{aligned}$$

where the last two inequalities are from Lemmas 5 and 6. As the above holds for any $\mathcal{A}\in\Pi$, Eq. (6) indicates

$$\begin{aligned}
&\limsup_{m,n\to\infty}\frac{1}{m+n}\log \mathbf{P}_{y^m x^n}((\hat{P}_m,\hat{Q}_n)\in\Gamma)\\
\le\ &\inf_{\mathcal{A}}\left(-\inf_{(R,S)\in\Gamma}\left(cD(R^{\mathcal{A}}\|P^{\mathcal{A}})+(1-c)D(S^{\mathcal{A}}\|Q^{\mathcal{A}})\right)\right)\\
=\ &-\sup_{\mathcal{A}}\inf_{(R,S)\in\Gamma} cD(R^{\mathcal{A}}\|P^{\mathcal{A}})+(1-c)D(S^{\mathcal{A}}\|Q^{\mathcal{A}}).
\end{aligned}$$

Then the remaining of obtaining the lower bound is to show

$$\sup_{\mathcal{A}}\inf_{(R,S)\in\Gamma} cD(R^{\mathcal{A}}\|P^{\mathcal{A}})+(1-c)D(S^{\mathcal{A}}\|Q^{\mathcal{A}})\ge\inf_{(R,S)\in\mathrm{cl}\,\Gamma} cD(R\|P)+(1-c)D(S\|Q).$$

Assuming, without loss of generality, that the left hand side is finite, we only need to show

$$\mathrm{cl}\,\Gamma\cap B(P,Q,\eta)\neq\varnothing,$$

whenever

$$\eta>\sup_{\mathcal{A}}\inf_{(R,S)\in\Gamma} cD(R^{\mathcal{A}}\|P^{\mathcal{A}})+(1-c)D(S^{\mathcal{A}}\|Q^{\mathcal{A}}).$$

Here $B(P,Q,\eta)$ is the divergence ball defined as follows

$$B(P,Q,\eta)=\{(R,S):cD(R\|P)+(1-c)D(S\|Q)\le\eta\},$$

which is compact in $\mathcal{P}\times\mathcal{P}$ w.r.t. the weak topology, due to the lower semi-continuity of $D(\cdot\|P)$ and $D(\cdot\|Q)$ as well as the fact that $0<c<1$.

To this end, we first show the following:

$$\operatorname{cl}\Gamma = \bigcap_{\mathcal{A}} \operatorname{cl}\Gamma(\mathcal{A}). \tag{9}$$

The inclusion is obvious since $\Gamma \in \Gamma(\mathcal{A})$. The reverse means that if $(R,S) \in \operatorname{cl}\Gamma(\mathcal{A})$ for each $\mathcal{A}$, then any neighborhood of (R,S) w.r.t. the weak convergence intersects Γ. To verify this, let $O(R,S)$ be a neighborhood of (R,S) w.r.t. the weak convergence, then there exists $\tilde{U}(R,S,\mathcal{B},\zeta) \in O(R,S)$ over a finite partition $\mathcal{B}$ as $O(R,S)$ is also open in $\mathcal{P}_\tau \times \mathcal{P}_\tau$. Furthermore, the partition $\mathcal{B}$ can be chosen to refine $\mathcal{A}$ so that $\operatorname{cl}\Gamma(\mathcal{B}) \subset \operatorname{cl}\Gamma(\mathcal{A})$. As τ-topology is stronger than the weak topology, a closed set in the $\mathcal{P}_\tau \times \mathcal{P}_\tau$ is closed in $\mathcal{P} \times \mathcal{P}$, and hence $\operatorname{cl}\Gamma(\mathcal{B}) \subset \operatorname{cl}_\tau \Gamma(\mathcal{B})$. That $(R,S) \in \operatorname{cl}_\tau \Gamma(\mathcal{B})$ implies that there exists $(R',S') \in \tilde{U}(R,S,\mathcal{B},\zeta) \cap \Gamma(\mathcal{B})$. By the definition of $\Gamma(\mathcal{B})$, we can also find $(\tilde{R},\tilde{S}) \in \Gamma$ such that $\tilde{R}(B_i) = R'(B_i)$ and $\tilde{S}(B_i) = S'(B_i)$ for each $B_i \in \mathcal{B}$, and hence $(\tilde{R},\tilde{S}) \in \tilde{U}(R,S,\mathcal{B},\zeta)$. In summary, we have $(\tilde{R},\tilde{S}) \in \tilde{U}(R,S,\mathcal{B},\zeta) \subset O(R,S)$ and $(\tilde{R},\tilde{S}) \in \Gamma$. Therefore, $\Gamma \cap O(R,S) \neq \varnothing$ and the claim follows.

Next we show that, for each partition $\mathcal{A}$,

$$\Gamma(\mathcal{A}) \cap B(P,Q,\eta) \neq \varnothing. \tag{10}$$

By Eq. (6), there exists $(\tilde{P},\tilde{Q})$ such that $cD(\tilde{P}^{\mathcal{A}}\|P^{\mathcal{A}}) + (1-c)D(\tilde{Q}^{\mathcal{A}}\|Q^{\mathcal{A}}) \leq \eta$. For such $(\tilde{P},\tilde{Q})$, we can construct $(P',Q') \in \Gamma(\mathcal{A})$ as

$$P'(F) = \sum_{i=1}^{t} \frac{\tilde{P}(A_i)}{P(A_i)} P(F \cap A_i),$$

$$Q'(F) = \sum_{i=1}^{t} \frac{\tilde{Q}(A_i)}{Q(A_i)} Q(F \cap A_i),$$

for any measurable subset $F \subset \mathcal{X}$. If $P(A_i) = 0$ ($Q(A_i) = 0$) and hence $\tilde{P}(A_i) = 0$ ($\tilde{Q}(A_i) = 0$), as $D(\tilde{P}^{\mathcal{A}}\|P^{\mathcal{A}}) < \infty$ ($D(\tilde{Q}^{\mathcal{A}}\|Q^{\mathcal{A}}) < \infty$), for some i, the corresponding term in the above equation is set equal to 0. Then (P',Q') belongs to $\Gamma(\mathcal{A})$ and also lies in $B(P,Q,\eta)$. The latter is because $D(P'\|P) = D(\tilde{P}^{\mathcal{A}}\|Q^{\mathcal{A}})$ and $D(Q'\|Q) = D(\tilde{Q}^{\mathcal{A}}\|Q^{\mathcal{A}})$: one can verify that any $\mathcal{B}$ that refines $\mathcal{A}$ satisfies

$$D(P'^{\mathcal{B}}\|P^{\mathcal{B}}) = D(\tilde{P}^{\mathcal{A}}\|P^{\mathcal{A}}), D(Q'^{\mathcal{B}}\|Q^{\mathcal{B}}) = D(\tilde{Q}^{\mathcal{A}}\|Q^{\mathcal{A}}).$$

For any finite collection of partitions $\mathcal{A}_i \in \Pi$ and $\mathcal{A} \in \Pi$ refining each $\mathcal{A}_i$, each $\Gamma(\mathcal{A}_i)$ contains $\Gamma(\mathcal{A})$. This implies that

$$\bigcap_{i=1}^{r} (\Gamma(\mathcal{A}_i) \cap B(p,q,\eta)) \neq \varnothing,$$

for any finite r. Finally, the set $\operatorname{cl}\Gamma(\mathcal{A}) \cap B(P,Q,\eta)$ for any $\mathcal{A}$ is compact due to the compactness of $B(P,Q,\eta)$, and any finite collection of them has non-empty intersection. It follows that all these sets is also non-empty. This completes the proof.

D Proof of Theorem 7

Proof. According to Theorem 1, d_k metrizes the weak convergence over $\mathcal{P}$. For convenience, we will write the type-I and type-II error probabilities as $\alpha_{m,n}$ and $\beta_{m,n}$, respectively; we will also use β to denote the type-II error exponent. That $\alpha_{m,n} \leq \alpha$ is clear from Lemma 4, and we only need to show that $\beta_{m,n}$ vanishes exponentially as m and n scale.

We first show $\beta \geq D^*$. With a fixed $\gamma > 0$, we have $\gamma_{m,n} \leq \gamma$ for sufficiently large n and m. Therefore,

$$\begin{aligned}
\beta &= \liminf_{m,n\to\infty} -\frac{1}{m+n} \log \mathbf{P}_{y^m x^n}(d_k(\hat{P}_m,\hat{Q}_n) \leq \gamma_{m,n}) \\
&\geq \liminf_{m,n\to\infty} -\frac{1}{m+n} \log \mathbf{P}_{y^m x^n}(d_k(\hat{P}_m,\hat{Q}_n) \leq \gamma) \\
&\geq \inf_{(R,S):d_k(R,S)\leq\gamma} cD(R\|P) + (1-c)D(S\|Q) \\
&:= D^*_\gamma,
\end{aligned} \tag{11}$$

where the last inequality is from the extended Sanov's theorem and that d_k metrizes weak convergence of $\mathcal{P}$ so that $\{(R,S): d_k(R,S)\le\gamma\}$ is closed in the product topology on $\mathcal{P}\times\mathcal{P}$. Since $\gamma>0$ can be arbitrarily small, we have

$$\beta \ge \lim_{\gamma\to 0^+} D^*_\gamma,$$

where the limit on the right hand side must exist as D^*_γ is positive, non-decreasing when γ decreases, and bounded by D^* that is assumed to be finite. Then it suffices to show

$$\lim_{\gamma\to 0^+} D^*_\gamma = D^*.$$

To this end, let (R_γ, S_γ) be such that $d_k(R_\gamma,S_\gamma)\le\gamma$ and $cD(R_\gamma\|P)+(1-c)D(S_\gamma\|Q)=D^*_\gamma$. Notice that R_γ and S_γ must lie in

$$\left\{W: D(W\|P)\le\frac{D^*}{c}, D(W\|Q)\le\frac{D^*}{1-c}\right\}:=\mathcal{W},$$

for otherwise $D^*_\gamma > D^*$. We remark that $\mathcal{W}$ is a compact set in $\mathcal{P}$ as a result of the lower semi-continuity of KLD w.r.t. the weak topology on $\mathcal{P}$ [50, 19]. Existence of such a pair can be seen from the facts that $\{(R,S): d_k(R,S)\le\gamma\}$ is closed and convex, and that both $D(\cdot\|P)$ and $D(\cdot\|Q)$ are convex functions [50].

Assume that D^* cannot be achieved. We can write

$$\lim_{\gamma\to 0^+} D^*_\gamma = D^*-\epsilon, \tag{12}$$

for some $\epsilon>0$. By the definition of lower semi-continuity, there exists a $\kappa_W>0$ for each $W\in\mathcal{W}$ such that

$$cD(R\|P)+(1-c)D(S\|Q)\ge cD(W\|P)+(1-c)D(W\|Q)-\frac{\epsilon}{2}\ge D^*-\frac{\epsilon}{2}, \tag{13}$$

whenever R and S are both from

$$\mathcal{S}_W=\{R: d_k(R,W)<\kappa_W\}.$$

Here the last inequality comes from the definition of D^* given in Theorem 7. To find a contradiction, define

$$\mathcal{S}'_W=\left\{R: d_k(R,W)<\frac{\kappa_W}{2}\right\}.$$

Since S'_W is open and $\bigcup_W \mathcal{S}'_W$ covers $\mathcal{W}$, the compactness of $\mathcal{W}$ implies that there exists finite $\mathcal{S}'_W$'s, denoted by $\mathcal{S}'_{W_1},\ldots,\mathcal{S}'_{W_N}$, covering $\mathcal{W}$. Define $\kappa^*=\min_{i=1}^N \kappa_{W_i}>0$. Now let $\gamma<\kappa^*/2$ as γ can be made arbitrarily small. Since $\bigcup_{i=1}^N \mathcal{S}'_{W_i}$ covers $\mathcal{W}$, we can find a W_i with $R_\gamma\in\mathcal{S}'_{W_i}\subset\mathcal{S}_{W_i}$. Thus, it holds that

$$d_k(S_\gamma,W_i)\le d_k(S_\gamma,R_\gamma)+d_k(R_\gamma,W_i)<\kappa_{W_i}.$$

That is, S_γ also lies in $\mathcal{S}_{W_i}$. By Eq. (13) we get

$$cD(R_\gamma\|P)+(1-c)D(S_\gamma\|Q)\ge D^*-\epsilon/2.$$

However, by our assumption in Eq. (12), it should hold that

$$cD(R_\gamma\|P)+(1-c)D(S_\gamma\|Q)\le D^*-\epsilon.$$

Therefore, $\beta\ge D^*$.

The other direction can be simply seen from the optimal type-II error exponent in Theorem 8. Alternatively, we can use Chernoff-Stein lemma in a similar manner to the proof of Theorem 3. Let P' be such that $cD(P'\|P)+(1-c)D(P'\|Q)=D^*$. Such P' exists because $0<D^*<\infty$ and $D(\cdot\|P)$ and $D(\cdot\|Q)$ are convex w.r.t. $\mathcal{P}$. That D^* is bounded implies that both $D(P'\|P)$ and $D(P'\|Q)$ are finite. We have

$$\begin{aligned}
\beta_{m,n} &= \mathbf{P}_{y^m x^n}(d_k(\hat{P}_m,\hat{Q}_n)\le\gamma_{m,n})\\
&\overset{(a)}{\ge} \mathbf{P}_{y^m x^n}(d_k(P',\hat{P}_m)+d_k(P',\hat{Q}_n)\le\gamma_{m,n})\\
&\overset{(b)}{\ge} \mathbf{P}_{y^m x^n}(d_k(P',\hat{P}_m)\le\gamma_m, d_k(P',\hat{Q}_m)\le\gamma_n)\\
&= P(d_k(P',\hat{P}_m)\le\gamma_m)\,Q(d_k(P',\hat{Q}_n)\le\gamma_n),
\end{aligned}$$

where (a) and (b) are from the triangle inequality of the metric d_k, and we pick $\gamma_n = \sqrt{2K/n}(1+\sqrt{-\log\alpha})$, and $\gamma_m = \sqrt{2K/m}(1+\sqrt{-\log\alpha})$ so that $\gamma_{m,n} > \gamma_n + \gamma_m$. Then Lemma 2 implies $P'(d_k(P', \hat{P}_m) \leq \gamma_m) > 1-\alpha$. For now assume that $D(P'\|P) > 0$ and $D(P'\|Q) > 0$. We can regard $\{y^m : d_k(P', \hat{P}_m) \leq \gamma_m\}$ as an acceptance region for testing $H_0 : y^m \sim P'$ and $H_1 : y^m \sim P$. Clearly, this test performs no better than the optimal level α test for this simple hypothesis testing in terms of the type-II error probability. Therefore, Chernoff-Stein lemma implies

$$\liminf_{m\to\infty} -\frac{1}{m}\log P(d_k(P', \hat{P}_m) \leq \gamma_m) \leq D(P'\|P). \tag{14}$$

Analogously, we have

$$\liminf_{n\to\infty} -\frac{1}{n}\log Q(d_k(P', \hat{Q}_n) \leq \gamma_n) \leq D(P'\|Q). \tag{15}$$

Now assume without loss of generality that $D(P'\|P) = 0$, i.e., $P' = P$. Then $D(P'\|Q) > 0$ under the alternative hypothesis $H_1 : P \neq Q$, and Eq. (15) still holds. Using Lemma 2, we have $P(d_k(P', \hat{P}_m) \leq \gamma_m) > 1-\alpha$, which gives zero exponent. Therefore, Eq. (14) holds with $P' = P$.

As $\lim_{m,n\to\infty} \frac{m}{m+n} = c$, we conclude that

$$\beta = \liminf_{m,n\to\infty} -\frac{1}{m+n}\log\beta_{m,n} \leq D^*.$$

The proof is complete. □

E Proof of Theorem 8

Proof. Let P' be such that $cD(P'\|P) + (1-c)D(P'\|Q) = D^*$. Consider first $D(P'\|P) \neq 0$ and $D(P'\|Q) \neq 0$. Since D^* is assumed to be finite, we have both $D(P'\|P)$ and $D(P'\|Q)$ being finite. This implies that P' is absolutely continuous w.r.t. both P and Q, so the Radon-Nikodym derivatives dP'/dP and dP'/dQ exist.

Define two sets

$$\begin{aligned} A_m &= \left\{ y^m : D(P'\|P) - \epsilon \leq \frac{1}{m}\log\frac{dP'(y^m)}{dP(y^m)} \leq D(P'\|P) + \epsilon \right\}, \\ B_n &= \left\{ x^n : D(P'\|Q) - \epsilon \leq \frac{1}{n}\log\frac{dP'(x^n)}{dQ(x^n)} \leq D(P'\|Q) + \epsilon \right\}, \end{aligned} \tag{16}$$

Recall the definition of the KLD: $D(P'\|P) = \mathbf{E}_{x\sim P'}\log(dP'(x)/dP(x))$ and $D(P'\|Q) = \mathbf{E}_{x\sim P'}\log(dP'(x)/dQ(x))$. By law of large numbers, we have for any given $\epsilon > 0$,

$$\mathbf{P}_{y^m x^n}(A_m \times B_n) \geq 1-\epsilon, \text{ for large enough } m \text{ and } n, \tag{17}$$

with y^m and x^n i.i.d. $\sim P'$.

Now consider the type-II error probability of level α tests. First, for a level α test, we have its acceptance region satisfies

$$\mathbf{P}_{y^m x^n}(\Omega_0'(m,n)) > 1-\alpha, \tag{18}$$

when y^m and x^n i.i.d. $\sim P'$, i.e., when the null hypothesis $H_0 : P = Q$ holds. Then under the alternative

hypothesis $H_1 : P \neq Q$, we have

$$\begin{aligned}
&\mathbf{P}_{y^m x^n}(\Omega_0'(m,n)) \\
&\geq \mathbf{P}_{y^m x^n}(A_m \times B_n \cap \Omega_0'(m,n)) \\
&= \int_{A_m \times B_n \cap \Omega_0'(m,n)} dP(y^m)\, dQ(x^n) \\
&\overset{(a)}{\geq} \int_{A_m \times B_n \cap \Omega_0'(m,n)} 2^{-m(D(P'\|P)+\epsilon)} 2^{-n(D(P'\|Q)+\epsilon)} dP'(y^m)\, dP'(x^n) \\
&= 2^{-mD(P'\|P)-n(D(P'\|Q)-(m+n)\epsilon} \int_{A_m \times B_n \cap \Omega_0'(m,n)} dP'(y^m)\, dP'(x^n) \\
&\overset{(b)}{\geq} 2^{-mD(P'\|P)-nD(P'\|Q)-(m+n)\epsilon}(1-\alpha-\epsilon),
\end{aligned}$$

where (a) is from Eq. (16) and (b) is due to Eqs. (17) and (18). Thus, when ϵ is small enough so that $1-\alpha-\epsilon > 0$, we get

$$\begin{aligned}
\liminf_{m,n\to\infty} -\frac{1}{m+n} \log \mathbf{P}_{y^m x^n}(\Omega_0'(m,n)) &\leq \liminf_{m,n\to\infty} -\frac{1}{m+n} \left(mD(P'\|P) + n(D(P'\|Q) + (m+n)\epsilon\right) \\
&= D^* + \epsilon. \qquad (19)
\end{aligned}$$

If a test is an asymptotic level α test, we can replace α by $\alpha + \epsilon'$ where ϵ' can be made arbitrarily small provided that m and n are large enough. Thus, Eq. (19) holds too. Finally, since ϵ can also be arbitrarily small, we conclude that

$$\liminf_{m,n\to\infty} -\frac{1}{m+n} \log \mathbf{P}_{y^m x^n}(\Omega_0'(m,n)) \leq D^*.$$

If $P' = P$, then A_m contains all $y^m \in \mathcal{X}^m$ and the above procedure gives the same result. □

F Experiments

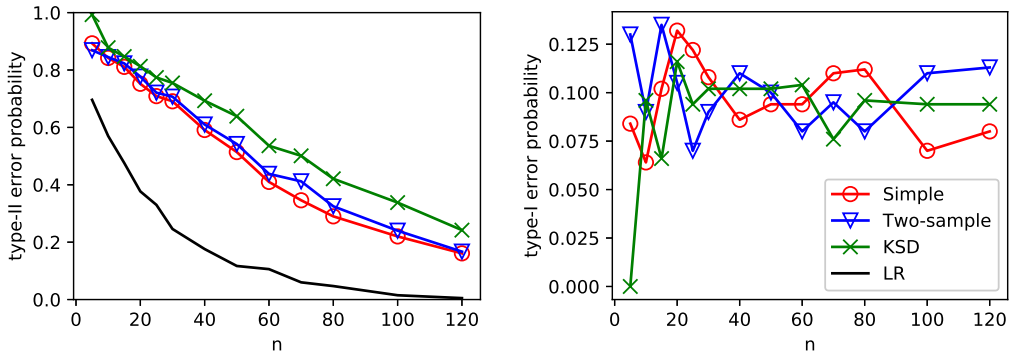
This section presents empirical results of the MMD and KSD based goodness-of-fit tests in the finite sample regime. We note that there have been extensive experiments in [12, 23, 34, 29] and the sample size m drawn from P is usually fixed for the kernel two-sample test. As such, we only consider two toy experiments and let m scale as required in Theorem 4.

We evaluate the following tests with a fixed level $\alpha = 0.1$, all using Gaussian kernel $k(x,y) = e^{-\|x-y\|_2^2/(2w)}$: 1) `Simple`: the simple kernel test $d_k(P, \hat{Q}_n)$. The acceptance threshold is estimated by drawing i.i.d. samples from P, i.e., the Monte Carlo method. The number of trials is 500. 2) `Two-sample`: the two-sample test $d_k(\hat{P}_m, \hat{Q}_n)$ with $m = n^{1.5}$. Threshold is obtained from the bootstrap method in [25], with 500 bootstrap replicates. 3) `KSD`: the KSD based test $d_S^2(P, \hat{Q}_n)$. We use wild bootstrap method from [12] with 500 replicates to estimate the α-quantile.

Gaussian vs. Laplace. We use a similar experiment setting in [29]. Consider $P : \mathcal{N}(0, 2\sqrt{2})$ and Q : Laplace$(0, 2)$, a zero-mean Laplace distribution with scale parameter 2. The parameters are chosen so that P and Q have the same mean and variance. We pick a fixed bandwidth $w = 1$ for all the kernel based tests and repeat 500 trials of each sample size n for both hypotheses. We also evaluate the likelihood ratio test `LR`, an oracle approach assuming both P and Q are known. In Figure 1a, `LR` has the lowest type-II error probabilities as expected, while `Simple` and `Two-sample` perform slightly better than `KSD`. As shown in Figure 1b, all the kernel based tests have the type-I error probabilities around the given level $\alpha = 0.1$, except for `KSD` with $n = 5$ samples.

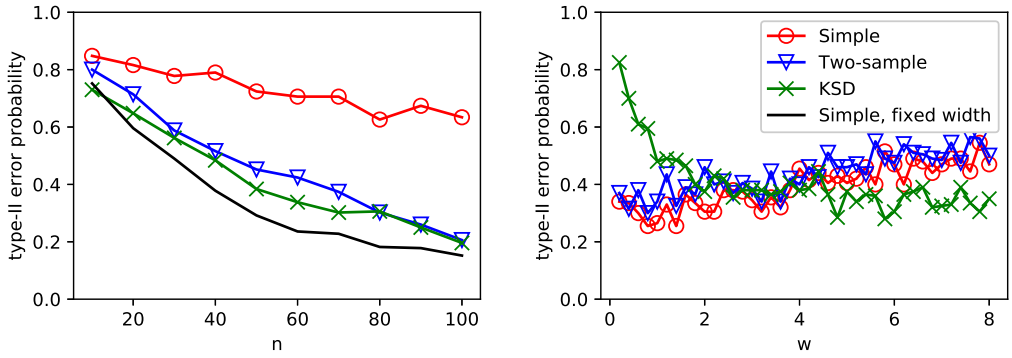
Gaussian Mixture. The next experiment is taken from [34]. The i.i.d. observations x^n are drawn from $Q : \sum_{i=1}^5 a_i \mathcal{N}(x; \mu_i, \sigma^2)$ with $a_i = 1/5$, $\sigma^2 = 1$, and μ_i randomly drawn from Uniform$[0, 10]$. We then generate P by adding standard Gaussian noise (perturbation) to μ_i. In [34], the sample number m drawn from P is fixed while the observed sample number n varies. We report the type-II error probabilities in Figure 2, averaged over 500 random trials.

With the median heuristic for bandwidth choice, `KSD` and `Two-sample` perform similarly whereas `Simple` has its type-II error probability decreasing slowly, as shown in Figure 2a. Picking a fixed bandwidth $w = 1$ for `Simple`



(a) type-II error (b) type-I error

Figure 1: Gaussian vs. Laplace.



(a) type-II error (b) type-II error

Figure 2: Gaussian mixture. (a) median bandwidth for `Simple`, `Two-sample`, and `KSD`, and a fixed bandwidth $w = 1$ for `Simple`; (b) fixing $n = 50$ samples and varying kernel bandwidths.

again results in a better performance. In light of the role of kernels, we then search over the kernel bandwidths in $[0, 8]$ for a fixed sample size $n = 50$. In Figure 2b, `Simple` and `Two-sample` tend to achieve lower type-II error probabilities when w is small, while `KSD` has a lower error probability around $w = 5$. The optimal type-II error probabilities of `Simple` and `KSD` are close and slightly lower than that of `Two-sample`. While computational issue is not the focus of this paper, we do observe that `KSD` is more efficient in this experiment, as it does not need to draw samples.

Whereas we cannot tell much statistical difference in our experiments, some experiments in the literature showed that the MMD based tests performed better than the KSD based tests and others showed the opposite [12, 23, 34, 29]. The finite sample performance depends on kernel choice as well as specific distributions. Under the universal setting, no test is known to be optimal in terms of the type-II error probability subject to a given level constraint. Statistical optimality can only be established in the large sample limit, as the one considered in the present work.