A. Terms and Notations

For the clarity of the paper, we give a summary of the commonly used notations in the main text and proof.

Symbols:

$\boldsymbol{s}_p(\boldsymbol{x})$	$\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$
$s_p^r(\boldsymbol{x})$	Projected score function $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})^T \boldsymbol{r}$
$\mathcal{X}$	A subset of $\mathbb{R}^D$
$\mathcal{K}$	A subset of $\mathbb{R}$.
k_{rg_r}	kernel function $k : \mathcal{K} \times \mathcal{K} \to \mathcal{R}$
$\mathcal{H}_{rg_r}$	Induced RKHS by the kernel k_{rg_r}.
$\|\| \cdot \|\|_{\mathcal{H}_{rg_r}}$	RKHS norm of $\mathcal{H}_{rg_r}$
$\boldsymbol{g}_r$	Input projection direction (e.g. $\boldsymbol{x}^T \boldsymbol{g}_r$) for corresponding $\boldsymbol{r}$.
$\boldsymbol{r}$	Score projection direction (e.g. $s_p^r(\boldsymbol{x}) = \boldsymbol{s}_p(\boldsymbol{x})^T \boldsymbol{r}$)
$S_{\max_{g_r}}$	*maxSSD-g* (Eq.4).
S_{g_r}	*SSD-g*, i.e. $S_{\max_{g_r}}$ (Eq.4) without $\sup_{\boldsymbol{g}_r}$. But with summation of O_r.
S_{rg_r}	*SSD-rg*, i.e. $S_{\max_{g_r}}$ (Eq.4) without $\sup_{\boldsymbol{g}_r}$ and summation of O_r. Instead, we use specific $\boldsymbol{r}$.
$SK_{\max_{g_r}}$	*maxSKSD-g*. The kernelized verison of $S_{\max_{g_r}}$
SK_{g_r}	*SKSD-g*. The kernelized verison of S_{g_r}
SK_{rg_r}	*SKSD-rg*. The kernelized verison of S_{rg_r}
PSD	Projected Stein discrepancy (Eq.9)
PSD_r	Projected Stein discrepancy (Eq.9) without summation O_r and use specific $\boldsymbol{r}$ instead.
f_r^*	Optimal test function for PSD. $f_r^*(\boldsymbol{x}) \propto s_p^r(\boldsymbol{x}) - s_q^r(\boldsymbol{x})$
$h_{rg_r}^*$	Optimal test function for S_{g_r} with specific $\boldsymbol{r}$ and $\boldsymbol{g}_r$, defined in Eq.8.
*	This indicates the optimal test function (e.g. f_r^*)
C_{sup}	Supremum of Poincaré constant defined in assumption 6.

A.1. "Sub-optimal" variants of SSD

For the ease of the analysis, we want to define the notations without the sup opeartor over the slice directions $\boldsymbol{r}$, $\boldsymbol{g}_r$. Here, we define *SSD-g* (S_{g_r}) as the *maxSSD-g* ($S_{\max_{g_r}}$ in Eq.4) without the $\sup_{\boldsymbol{g}_r}$.

$$S_{g_r} = \sum_{\boldsymbol{r} \in O_r} \sup_{h_{rg_r} \in \mathcal{F}_q} \mathbb{E}_q[s_p^r(\boldsymbol{x}) h_{rg_r}(\boldsymbol{x}^T \boldsymbol{g}_r) + \boldsymbol{r}^T \boldsymbol{g}_r \nabla_{\boldsymbol{x}^T \boldsymbol{g}_r} h_{rg_r}(\boldsymbol{x}^T \boldsymbol{g}_r)] \tag{16}$$

Similarly, we define *SSD-rg* (S_{rg_r}) as *maxSSD-rg* ($S_{\max_{rg_r}}$ in Eq.37) without $\sup_{\boldsymbol{r}, \boldsymbol{g}_r}$:

$$S_{rg_r} = \sup_{h_{rg_r} \in \mathcal{F}_q} \mathbb{E}_q[s_p^r(\boldsymbol{x}) h_{rg_r}(\boldsymbol{x}^T \boldsymbol{g}_r) + \boldsymbol{r}^T \boldsymbol{g}_r \nabla_{\boldsymbol{x}^T \boldsymbol{g}_r} h_{rg_r}(\boldsymbol{x}^T \boldsymbol{g}_r)] \tag{17}$$

As for each of the above "optimal" discrepancies, it has the corresponding kernelized version. Therefore, we need to define their "un-optimal" version as well. We define *SKSD-g* (SK_{g_r}) as *maxSKSD-g* ($SK_{\max_{g_r}}$ in Eq.6) as

$$SK_{g_r} = \sum_{\boldsymbol{r} \in O_r} ||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}}, \tag{18}$$

Similarly, we define *SKSD-rg* (SK_{rg_r}) as *maxSKSD-rg* ($SK_{\max_{rg_r}}$ in Eq.41) as

$$SK_{rg_r} = ||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}} \tag{19}$$

B. Assumptions and Definitions

Definition B.1 (Inner product in Hilbert space)**.** *We denote the algebraic space $\mathbb{R}^D$ refers to a parameter space of dimension D. The Borel sets of $\mathbb{R}^D$ is denoted as $\mathcal{B}(\mathbb{R}^D)$, and we let $\mu(x)$ be a probability measure on $\boldsymbol{x}$. We define*

$$\mathcal{H}_\mu = L^2(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D), \mu) \tag{20}$$

as the Hilbert space which contains all the measurable functions $f : \mathbb{R}^D \to \mathbb{R}$, such that $||f||_{\mathcal{H}_\mu} \leq \infty$, where we define inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\mu}$ to be

$$\langle f, g \rangle_{\mathcal{H}_\mu} = \int f(\boldsymbol{x}) g(\boldsymbol{x}) d\mu(x) \tag{21}$$

for all $f, g \in \mathcal{H}_\mu$

Definition B.2. ***(Stein Class** (Liu et al., 2016)) Assume distribution q has continuous and differentiable density $q(\boldsymbol{x})$. A function f defined on the domain $\mathcal{X} \subseteq \mathbb{R}^D$, $f : \mathcal{X} \to \mathbb{R}$ is in the **Stein class of** q if f is smooth and satisfies*

$$\int_{\mathcal{X}} \nabla_x (f(\boldsymbol{x}) q(\boldsymbol{x})) d\boldsymbol{x} = 0 \tag{22}$$

We call a function $f(\boldsymbol{x}) \in \mathcal{F}_q$ if f belongs to the Stein class of q. We say vector-valued function $\boldsymbol{f}(\boldsymbol{x}) : \mathcal{X} \subseteq \mathbb{R}^D \to \mathbb{R}^m \in \mathcal{F}_q$ if each component of $\boldsymbol{f}$ belongs to the Stein class of q.

Definition B.3 (Stein Identity)**.** *Assume q is a smooth density satisfied assumption 1 , then we have*

$$\mathbb{E}_q \left[s_q(x) f(x)^T + \nabla f(x) \right] = 0 \tag{23}$$

for any functions $f : \mathcal{X} \subseteq \mathbb{R}^D \to \mathbb{R}^D$ in Stein class of q.

We can easily see that the above holds true for $\mathcal{X} = \mathbb{R}^D$ if

$$\lim_{||\boldsymbol{x}|| \to \infty} q(\boldsymbol{x}) f(\boldsymbol{x}) = 0 \tag{24}$$

Assumption 1 (Properties of densities) Assume the two probability distributions p, q has continuous differentiable density $p(\boldsymbol{x})$, $q(\boldsymbol{x})$ supported on $\mathcal{X} \subseteq \mathbb{R}^D$, such that the induced set $\mathcal{K} = \{y \in \mathbb{R} | y = \boldsymbol{x}^T \boldsymbol{g}, ||\boldsymbol{g}||^2 = 1, \boldsymbol{x} \in \mathcal{X}\}$ is *locally compact Hausdorff* (LCH) for all possible $\boldsymbol{g} \in \mathbb{S}^{D-1}$. If $\mathcal{X} = \mathbb{R}^D$, then the density q satisfies: $\lim_{||\boldsymbol{x}|| \to \infty} q(\boldsymbol{x}) = 0$. If $\mathcal{X} \subset \mathbb{R}^D$ is compact, then $q(\boldsymbol{x}) = 0$ at boundary $\partial \mathcal{X}$.

Assumption 2 (Regularity of score functions) Denote the score function of $p(\boldsymbol{x})$ as $\boldsymbol{s}_p(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) \in \mathbb{R}^D$ and score function of $q(\boldsymbol{x})$ accordingly. Assume the score functions are bounded continuous differentiable functions and satisfying

$$\begin{aligned} &\int_{\mathcal{X}} q(\boldsymbol{x}) |(s_p(\boldsymbol{x}) - s_q(\boldsymbol{x}))^T \boldsymbol{r}| d\boldsymbol{x} < \infty \\ &\int_{\mathcal{X}} q(\boldsymbol{x}) ||(s_p(\boldsymbol{x}) - s_q(\boldsymbol{x}))^T \boldsymbol{r}||^2 d\boldsymbol{x} < \infty \end{aligned} \tag{25}$$

for all $\boldsymbol{r}$ where $\boldsymbol{r} \in \mathbb{S}^{D-1}$.

Assumption 3 (Test functions) Assume the test function $h_{rg_r} : \mathcal{K} \subseteq \mathbb{R} \to \mathbb{R}$ is smooth and belongs to the Stein class of q. Specifically, if with assumption 1, we only requires h_{rg_r} to be a bounded continuous function. Similarly, we assume this also holds for PSD (eq.9) test function $f_r(\boldsymbol{x})$.

Assumption 4 (Bounded Conditional Expectation) Define

$$h^*_{rg_r}(y_d) = \mathbb{E}_{q_{G_r}(\boldsymbol{y}_{-d}|y_d)}[(s^r_p(\boldsymbol{G}_r^{-1}\boldsymbol{y}) - s^r_q(\boldsymbol{G}_r^{-1}\boldsymbol{y}))] \tag{26}$$

as in proposition 1. We assume $h^*_{rg_r}$ is uniformly bounded for all possible $\boldsymbol{g}_r \in \mathbb{S}^{D-1}$.

Assumption 5 (universal kernel): We assume the kernel $k_{rg} : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ is bounded and c_0-universal.

Assumption 6 (Real analytic translation invariant kernel): We assume the kernel is translation invariant $k(x, y) = \phi(x - y) : \mathcal{K} \to \mathbb{R}$ and ϕ is a real analytic function. Additionally, we assume if $k(cx, cy) = k'(x, y)$ for a constant $c > 0$ where k' is also a c_0-universal kernel. For example, *radial basis kernel function* (RBF) and *inverse multiquadric* (IMQ) kernel satisfy these assumptions.

Assumption 7 (Log-concave probabilities) Assume a probability distribution q with density function such that $q(\boldsymbol{x}) = \exp(-V(\boldsymbol{x}))$, where $V(\boldsymbol{x})$ is a convex function.

Assumption 8 (Existence of supremum of Poincaré constant). For the Poincaré constant defined in lemma 5, the essential supremum exists $C_{ess,\boldsymbol{G}} = \text{ess sup}_{y_d} C_{y_d} < \infty$ and also the $C_{sup} = \sup_{\boldsymbol{G}} C_{ess,\boldsymbol{G}} < \infty$ exists over all possible orthogonal matrix $\boldsymbol{G}$.

C. Detailed Background

C.1. Stein Discrepancy

Assume we have two differentiable probability density functions $q(\boldsymbol{x})$ and $p(\boldsymbol{x})$ where $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^D$. We further define a test function $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^D$ and a suitable test function family $\mathcal{F}_q$ called *Stein's class of* q. Recall the Stein operator (Eq.1) is defined as

$$\mathcal{A}_p \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{s}_p(\boldsymbol{x})^T \boldsymbol{f}(\boldsymbol{x}) + \nabla^T_{\boldsymbol{x}}] \boldsymbol{f}(\boldsymbol{x}) \tag{27}$$

The function family $\mathcal{F}_q$ is defined as

$$\mathcal{F}_q = \left\{ \boldsymbol{f} : \mathcal{X} \to \mathbb{R}^D \mid \mathbb{E}_q[\mathcal{A}_q \boldsymbol{f}] = 0 \right\} \tag{28}$$

This function space can be quite general. For example, if $\mathcal{X} = \mathbb{R}^D$, we only require $\boldsymbol{f}$ to be differentiable and vanishing at infinity. With all the notations, *Stein discrepancy* is defined as follows:

$$D_{SD}(q, p) = \sup_{\boldsymbol{f} \in \mathcal{F}_q} \mathbb{E}_q[\mathcal{A}_p \boldsymbol{f}(\boldsymbol{x})] \tag{29}$$

which can be proved to be a valid discrepancy (Gorham & Mackey, 2017). Stein discrepancy has been shown to be closely related to *Fisher discrepancy* defined as

$$D_F(q, p) = \mathbb{E}_q ||\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})||_2^2 \tag{30}$$

Indeed, Hu et al. (2018) shows that the optimal test function for *Stein discrepancy* has the form $\boldsymbol{f}^*(\boldsymbol{x}) \propto \boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})$. By substitution, we can show *Stein discrepancy* is equivalent to *Fisher divergence* up to a multiplicative constant.

Unfortunately, the score difference $\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})$ may be intractable in practice, making SD intractable as a consequence. Thus, Liu et al. (2016); Chwialkowski et al. (2016) propose an variant of SD by restricting $\mathcal{F}_q$ to be a unit ball inside an RKHS $\mathcal{H}_k$ induced by a c_0-universal kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. By using the reproducing properties, they propose *kernelized Stein discrepancy* as

$$\begin{aligned} D^2(q, p) &= ||\mathbb{E}_q[\boldsymbol{s}_p(\boldsymbol{x}) k(\boldsymbol{x}, \cdot) + \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \cdot)]||^2_{\mathcal{H}_k} \\ &= \mathbb{E}_{\boldsymbol{x}, \boldsymbol{x}' \sim q}[u_p(\boldsymbol{x}, \boldsymbol{x}')] \end{aligned} \tag{31}$$

where $u_p(\boldsymbol{x}, \boldsymbol{x}')$ is

$$\begin{aligned} u_p(\boldsymbol{x}, \boldsymbol{x}') &= \boldsymbol{s}_p(\boldsymbol{x})^T k(\boldsymbol{x}, \boldsymbol{x}') \boldsymbol{s}_p(\boldsymbol{x}') + \boldsymbol{s}_p(\boldsymbol{x})^T \nabla_{\boldsymbol{x}'} k(\boldsymbol{x}, \boldsymbol{x}') \\ &+ \boldsymbol{s}_p(\boldsymbol{x}')^T \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{x}') + \nabla^2_{\boldsymbol{x}, \boldsymbol{x}'} k(\boldsymbol{x}, \boldsymbol{x}') \end{aligned} \tag{32}$$

and $\boldsymbol{x}$, $\boldsymbol{x}'$ are i.i.d. samples from q.

Due to its tractability, it has been extensively used in statistical test e.g. GOF test Liu et al. (2016); Chwialkowski et al. (2016); Huggins & Mackey (2018); Jitkrittum et al. (2017). However, recent work demonstrate KSD suffers from the curse-of-dimensionality problem Gong et al. (2021); Huggins & Mackey (2018); Chwialkowski et al. (2016). One potential fix is to use another variant called *sliced kernelized Stein discrepancy*.

C.2. Sliced Kernelized Stein Discrepancy

In this section, we give a more detailed introduction to sliced kernelized Stein discrepancy (SKSD). Recall the definition of Stein discrepancy:

$$D_{SD}(q,p) = \sup_{\boldsymbol{f}\in\mathcal{F}_q} \mathbb{E}_q[\boldsymbol{s}_p^T(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x}) + \nabla_{\boldsymbol{x}}^T\boldsymbol{f}(\boldsymbol{x})] \tag{33}$$

In the original paper of (Gong et al., 2021), they argue that the curse of dimensionality comes from two sources: (i) the high dimensionality of the score function $\boldsymbol{s}_p : \mathcal{X} \subseteq \mathbb{R}^D \to \mathbb{R}^D$ and (ii) the test function input $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^D$. Therefore, authors proposed two slice directions $\boldsymbol{r}$, $\boldsymbol{g}$ to project $\boldsymbol{s}_p$ and $\boldsymbol{x}$ respectively. However, this projection is equivalent to throwing away most of the information possessed by $\boldsymbol{s}_p$ and $\boldsymbol{x}$. To tackle this problem, authors proposed the first member of the SSD family by considering over all possible directions of $\boldsymbol{r}$ and $\boldsymbol{g}$ (a distribution over $\boldsymbol{r} \sim p_r, \boldsymbol{g} \sim p_g$), called *integrated sliced Stein discrepancy*:

$$S(q,p) = \mathbb{E}_{p_r,p_g}\left[\sup_{h_{rg}\in\mathcal{F}_q} \mathbb{E}_q[s_p^r(\boldsymbol{x})h_{rg}(\boldsymbol{x}^T\boldsymbol{g}) + \boldsymbol{r}^T\boldsymbol{g}\nabla_{\boldsymbol{x}^T\boldsymbol{g}}h_{rg}(\boldsymbol{x}^T\boldsymbol{g})]\right]. \tag{34}$$

where h_{rg_r} is the test function. Although it is theoretically valid (Theorem 1 in(Gong et al., 2021)), its practical useage is limited by the intractability of the integral over p_r, p_g and the optimal test function h_{rg}. Surprisingly, authors show that the integral over $\boldsymbol{r}$, $\boldsymbol{g}$ is not necessary for discrepancy validity. They achieved this in two steps.

The first step is to replace the expectation w.r.t. $\boldsymbol{r}$ by a finite summation over orthogonal basis. The author showed that this is a valid discrepancy, called *orthogonal sliced Stein discrepancy* defined as

$$S_O(q,p) = \sum_{\boldsymbol{r}\in O_r} \mathbb{E}_{p_g}\left[\sup_{h_{rg}\in\mathcal{F}_q} \mathbb{E}_q[s_p^r(\boldsymbol{x})h_{rg}(\boldsymbol{x}^T\boldsymbol{g}) + \boldsymbol{r}^T\boldsymbol{g}\nabla_{\boldsymbol{x}^T\boldsymbol{g}}h_{rg}(\boldsymbol{x}^T\boldsymbol{g})]\right]. \tag{35}$$

where O_r is an orthogonal basis (e.g. one-hot vectors). The next step is to get rid of the expectation w.r.t. $\boldsymbol{g}$ by a supremum operator. This is called *maxSSD-g*, which is defined as Eq.4 in the main text. For a quick recall, we include *maxSSD-g* in here:

$$S_{\max_{g_r}}(q,p) = \sum_{\boldsymbol{r}\in O_r} \sup_{\substack{h_{rg_r}\in\mathcal{F}_q \\ \boldsymbol{g}_r\in\mathbb{S}^{D-1}}} \mathbb{E}_q[s_p^r(\boldsymbol{x})h_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r) + \boldsymbol{r}^T\boldsymbol{g}_r\nabla_{\boldsymbol{x}^T\boldsymbol{g}_r}h_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r)] \tag{36}$$

Further, one can also use single optimal direction $\boldsymbol{r}$ to replace the summation over the orthogonal basis O_r, resulting in *maxSSD-rg*($S_{\max_{rg_r}}$):

$$S_{max_{rg_r}}(q,p) = \sup_{h_{rg}\in\mathcal{F}_q,\boldsymbol{g}_r,\boldsymbol{r}\in\mathbb{S}^{D-1}} \mathbb{E}_q\left[s_p^r(\boldsymbol{x})h_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r) + \boldsymbol{r}^T\boldsymbol{g}_r\nabla_{\boldsymbol{x}^T\boldsymbol{g}_r}h_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r)\right] \tag{37}$$

Similar to KSD, authors addressed tractability issue of the optimal h_{rg_r} by restricting the $\mathcal{F}_q$ to be a one-dimensional RKHS induced by a c_0-universal kernel $k_{rg} : \mathcal{K}\times\mathcal{K} \to \mathbb{R}$ where $\mathcal{K} \subseteq \mathbb{R}$. Thus, for each member of the above SSD family, we have a corresponding kernelized version. They are called *integrated sliced kernelized Stein discrepancy*, *orthogonal SKSD*, and *max sliced kernelized Stein discrepancy* (including *maxSKSD-g* and *maxSKSD-rg*). In practice, *maxSKSD-g* or *maxSKSD-rg* is often preferred over the others due to its computational tractability, where their optimal slices for $\boldsymbol{r}$ and $\boldsymbol{g}_r$ are obtained by gradient-based optimization.

By reproducing properties of RKHS, one can define $\xi_{p,r,g_r}(\boldsymbol{x},\cdot)$ as in Eq.5, and further define $\mu_{p,r,g_r} = \langle\xi_{p,r,g_r}(\boldsymbol{x},\cdot),\xi_{p,r,g_r}(\boldsymbol{y},\cdot)\rangle_{\mathcal{H}_{rg_r}}$

$$\begin{aligned}\mu_{p,r,g_r}(\boldsymbol{x},\boldsymbol{y}) &= s_p^r(\boldsymbol{x})k_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r,\boldsymbol{y}^T\boldsymbol{g}_r)s_p^r(\boldsymbol{y}) \\ &+ \boldsymbol{r}^T\boldsymbol{g}_r s_p^r(\boldsymbol{y})\nabla_{\boldsymbol{x}^T\boldsymbol{g}_r}k_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r,\boldsymbol{y}^T\boldsymbol{g}_r) \\ &+ \boldsymbol{r}^T\boldsymbol{g}_r s_p^r(\boldsymbol{x})\nabla_{\boldsymbol{y}^T\boldsymbol{g}_r}k_{rg}(\boldsymbol{x}^T\boldsymbol{g}_r,\boldsymbol{y}^T\boldsymbol{g}_r) \\ &+ (\boldsymbol{r}^T\boldsymbol{g}_r)^2\nabla^2_{\boldsymbol{x}^T\boldsymbol{g}_r,\boldsymbol{y}^T\boldsymbol{g}_r}k_{rg}(\boldsymbol{x}^T\boldsymbol{g}_r,\boldsymbol{y}^T\boldsymbol{g}_r).\end{aligned} \tag{38}$$

Then, by simple algebra, one can show that given $\boldsymbol{r}$, $\boldsymbol{g}_r$, the optimality w.r.t. test functions can be computed analytically:

$$\begin{aligned}&D^2_{rg_r}(q,p) \\ =&(\sup_{\substack{h_{rg_r}\in\mathcal{H}_{rg_r} \\ ||h_{rg_r}||_{\mathcal{H}_{rg_r}}\le 1}} \mathbb{E}_q[s_p^r(\boldsymbol{x})h_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r) \\ &+ \boldsymbol{r}^T\boldsymbol{g}_r\nabla_{\boldsymbol{x}^T\boldsymbol{g}_r}h_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r)])^2 \\ =&||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}} = \mathbb{E}_{q(\boldsymbol{x})q(\boldsymbol{x}')}[\mu_{p,r,g_r}(\boldsymbol{x},\boldsymbol{x}')].\end{aligned} \tag{39}$$

where $\mathcal{H}_{rg_r}$ is the RKHS induced by the kernel k_{rg_r}. Therefore, the *maxSSD-g* and *maxSSD-rg* can be computed as

$$SK_{\max_{g_r}}(q,p) = \sum_{\boldsymbol{r}\in O_r}\sup_{\boldsymbol{g}_r\in\mathbb{S}^{D-1}} D^2_{rg_r}(q,p) \tag{40}$$

and

$$SK_{\max_{rg_r}}(q,p) = \sup_{\substack{\boldsymbol{g}_r\in\mathbb{S}^{D-1} \\ \boldsymbol{r}\in\mathbb{S}^{D-1}}} D^2_{rg_r}(q,p) \tag{41}$$

D. Goodness-of-fit test

In this section, we give an introduction to the GOF test. To be general, we focus on the *SKSD-rg* ($SK_{rg_r} = ||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}}$) as other related discrepancy can be easily derived from it. Assuming we have active slices $\boldsymbol{r}$ and $\boldsymbol{g}_r$ from algorithm 1. Thus, we can estimate SK_{rg_r} using the minimum variance U-staistics (Hoeffding, 1992; Serfling, 2009):

$$\widehat{SK}_{rg_r}(q,p) = \frac{1}{N(N-1)} \sum_{1\le i\neq j\le N} \mu_{p,r,g_r}(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{42}$$

where $\mu_{\boldsymbol{x},\boldsymbol{y}}$ is defined in Eq.38 which satisfies $\mathbb{E}_{q(\boldsymbol{x})q(\boldsymbol{x}')}[\mu_{p,r,g_r}(\boldsymbol{x},\boldsymbol{x}')] = ||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}}$, and $\boldsymbol{x}$, $\boldsymbol{x}'$ are i.i.d. samples from q. With the help of the U-statistics, we characterize its asymptotic distribution.

Theorem 6. *Assume the conditions in theorem 1 are satisfied, we have the following:*

1. *If $q \neq p$, then $\widehat{SK}_{rg_r}$ is asymptotically normal. Particularly,*

$$\sqrt{N}(\widehat{SK}_{rg_r} - SK_{rg_r}) \xrightarrow{d} \mathcal{N}(0, \sigma_h^2) \tag{43}$$

where $\sigma_h^2 = var_{\boldsymbol{x}\sim q}(\mathbb{E}_{\boldsymbol{x}'\sim q}[\mu_{p,r,g_r}(\boldsymbol{x},\boldsymbol{x}')])$ and $\sigma_h \neq 0$

2. *If $q = p$, we have a degenerated U-statistics with $\sigma_h = 0$ and*

$$N\widehat{SK}_{rg_r} \xrightarrow{d} \sum_{j=1}^{\infty} c_j(Z_j^2 - 1) \tag{44}$$

where $\{Z_j\}$ are i.i.d standard Gaussian variables, and $\{c_j\}$ are the eigenvalues of the kernel $\mu_{p,r,g_r}(\boldsymbol{x},\boldsymbol{x}')$ under $q(\boldsymbol{x})$. In other words, they are the solutions of $c_j\phi_j(\boldsymbol{x}) = \int_{\boldsymbol{x}'} \mu_{p,r,g_r}(\boldsymbol{x},\boldsymbol{x}')\phi_j(\boldsymbol{x}')q(\boldsymbol{x}')d\boldsymbol{x}'$.

Proof. As the $\widehat{SK}_{rg_r}$ is the second order U-statistic of SK_{rg_r}, thus, we can directly use the results from section 5.5.1 and 5.5.2 in (Serfling, 2009). □

The above theorem indicates a well-defined asymptotic distribution for SK_{rg_r}, which allows us to use the following bootstrap method to estimate the rejection threshold (Huskova & Janssen, 1993; Arcones & Gine, 1992; Liu et al., 2016). The bootstrap samples can be computed as

$$\widehat{SK}^*_m = \sum_{1\le i\neq j\le N} (w_i^m - \frac{1}{N})(w_j^m - \frac{1}{N})\mu_{p,r,g_r}(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{45}$$

where $(w_1^m, \dots, w_N^m)_{m=1}^M$ are random weights drawn from multinomial distributions Multi$(N, \frac{1}{N}, \dots, \frac{1}{N})$. Now, we give the detailed algorithm for GOF test.

Algorithm 2 GOF test with active slices

Input: Samples $\boldsymbol{x} \sim q$, density p, kernel k_{rg_r}, active slices $\boldsymbol{r}$, $\boldsymbol{g}_r$, significance level α, and bootstrap sample size M.
Hypothesis: H_0: $p = q$ v.s. $H_1 : p \neq q$
Computing U-statistics $\widehat{SK}_{rg_r}$ using Eq.42
Generate M bootstrap samples $\{\widehat{SK}^*_m\}_{m=1}^M$ using Eq.45
Reject null hypothesis H_0 if the proportion of $\widehat{SK}^*_m > \widehat{SK}_{rg_r}$ is less than α

E. Relaxing constraints for kernelized SSD family

E.1. Validity w.r.t $\boldsymbol{r}$, $\boldsymbol{g}_r$

The key to this proof is to prove the real analyticity of SK_{g_r} (or S_{rg_r}) to slices $\boldsymbol{r}$ and $\boldsymbol{g}_r$. Therefore, let's first give a definition of multivariate real analytic function.

Definition E.1 (Real analytic function)**.** *A function $f : \mathcal{U} \to \mathbb{R}$ is real analytic if for each $\mathbf{c} \in \mathcal{U}$, there is a power series as in the form*

$$f(\boldsymbol{x}) = \sum_{\kappa\in\mathbb{N}_0^n} \alpha_\kappa(\boldsymbol{x} - \boldsymbol{c})^\kappa$$

for some choice of $(\alpha_\kappa)_{\kappa\in\mathbb{N}_0^n} \subset \mathbb{R}$ and all $\boldsymbol{x}$ in a neighbourhood of $\mathbf{c}$, and this power series converges absolutely. Namely,

$$\sum_{\kappa\in\mathbb{N}_0^n} |\alpha_\kappa||(\boldsymbol{x} - \boldsymbol{c})^\kappa| < \infty$$

where $\mathbb{N}_0 = \{0, 1, \dots\}$ denotes non-negative integers, $\kappa = (\kappa_1, \dots, \kappa_n)$ are called multiindex, and we define $\boldsymbol{x}^\kappa = x_1^{\kappa_1} \dots x_n^{\kappa_n}$.

Now, we introduce a useful lemma showing that composition of real analytic function is also real analytic.

Lemma 1 (Composition of real analytic function)**.** *Let $\mathcal{U} \subset \mathbb{R}^n$ and $\mathcal{V} \subset \mathbb{R}^m$ be open, and let $\boldsymbol{f} : \mathcal{U} \to \mathcal{V}$ and $\boldsymbol{g} : \mathcal{V} \to \mathbb{R}^p$ be real analytic. Then $\boldsymbol{g} \circ \boldsymbol{f} : \mathcal{U} \to \mathbb{R}^p$ is real analytic.*

Especially, the real analyticity is not only preserved by function composition, it is also closed under most of the simple operations: addition, multiplication, division (assuming denominator is non-zero), etc. Now we can prove the main proposition to show that the *SKSD-rg* (SK_{rg_r}) is real analytic w.r.t both $\boldsymbol{g}_r$ and $\boldsymbol{r}$. In the following, we assume the $\boldsymbol{r}, \boldsymbol{g}_r \in \mathbb{R}^D$.

Proposition 3 (*SKSD-g* is real analytic)**.** *Assume assumption 1-4 (density regularity), 5-6 (kernel richness and real analyticity) are satisfied, further we let $\boldsymbol{g}_r \in \mathbb{R}^D$, then SKSD-g ($SK_{g_r}$) is real analytic w.r.t $\boldsymbol{g}_r$ and SK_{rg_r} is real analytic to both $\boldsymbol{r} \in \mathbb{R}^D$ and $\boldsymbol{g}_r$.*

Proof. First, let's focus on the real analyticity w.r.t. $\boldsymbol{g}_r$. We re-write the *SKSD-g* as the following:

$$
\begin{aligned}
SK_{g_r} &= \sum_{\boldsymbol{r}\in O_r} ||\xi_{p,r,g_r}(\boldsymbol{x})||^2_{\mathcal{H}_{rg_r}} \\
&= \sum_{\boldsymbol{r}\in O_r} \langle \mathbb{E}_q[\underbrace{(s_p^r(\boldsymbol{x}) - s_q^r(\boldsymbol{x}))}_{f_r^*(\boldsymbol{x})} k_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r, \cdot)], \\
&\mathbb{E}_q[(s_p^r(\boldsymbol{x}) - s_q^r(\boldsymbol{x}))k_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r, \cdot)]\rangle_{\mathcal{H}_{rg_r}} \\
&= \sum_{\boldsymbol{r}\in O_r} \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[f_r^*(\boldsymbol{x})k_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r, \boldsymbol{x}'^T\boldsymbol{g}_r)f_r^*(\boldsymbol{x}')]
\end{aligned}
$$

The second equality is from the definition of RKHS norm $||\cdot||_{\mathcal{H}_{rg_r}}$ and Stein identity. We can observe that $\boldsymbol{g}_r$ appears inside the kernel k_{rg_r} in the form of $\boldsymbol{x}^T\boldsymbol{g}_r$. So in order to use the function composition lemma (lemma 1), we need to first show that for any given $\boldsymbol{x}$, $\boldsymbol{x}^T\boldsymbol{g}_r$ is real analytic. By definition of real analytic function, we need a center point $\boldsymbol{c} \in \mathbb{R}^D$, and $\boldsymbol{g}_r$ in the neighborhood of c (i.e. $|\boldsymbol{g}_r - \boldsymbol{c}| < R_c$). Then, we define the power series as

$$
h_x(\boldsymbol{g}_r) = \sum_{\kappa_1=0}^{\infty} \cdots \sum_{\kappa_D=0}^{\infty} \frac{(g_{r1} - c_1)^{\kappa_1} \dots (g_{rD} - c_D)^{\kappa_D}}{\kappa_1! \dots \kappa_D!} \alpha_{\{\kappa_i\}_i^D}
$$

with the following coefficient

$$
\begin{cases}
\alpha_{\{\kappa_i\}_i^D} = 0 & \text{if} \quad \sum_i \kappa_i > 1 \\
\alpha_{\{\kappa_i\}_i^D} = x_i & \text{if} \quad \kappa_i = 1, \sum_i \kappa_i = 1 \\
\alpha_{\{\kappa_i\}_i^D} = \boldsymbol{c}^T\boldsymbol{x} & \text{if} \quad \sum_i \kappa_i = 0
\end{cases}
$$

Then, by substitution, we have

$$
\begin{aligned}
h_x(\boldsymbol{g}) &= \sum_{d=1}^{D} (g_d - c_d)x_d + \boldsymbol{c}^T\boldsymbol{x} && (46) \\
&= \boldsymbol{x}^T\boldsymbol{g} && (47)
\end{aligned}
$$

which converges with radius of convergence $R_c = \infty$. From assumption 6, we know the kernel $k_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r, \boldsymbol{x}'^T\boldsymbol{g}_r) = \phi((\boldsymbol{x} - \boldsymbol{x}')^T\boldsymbol{g}_r)$ is translation invariant and real analytic. Thus, from lemma 1, we know $k_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r, \boldsymbol{x}'^T\boldsymbol{g}_r)$ is real analytic to $\boldsymbol{g}_r$ with radius of convergence R_k (R_k is determined by the form of the kernel function). This means we can use a power series to represents this kernel w.r.t. $\boldsymbol{g}_r$ inside some neighborhood define around center point. Specifically, for a central point $\boldsymbol{c} \in \mathbb{R}^D$ and any $\boldsymbol{g}_r$ satisfying $|\boldsymbol{g}_r - \boldsymbol{c}| < R_k$, we have

$$
k_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r, \boldsymbol{x}'^T\boldsymbol{g}_r) = \sum_{\kappa\in\mathbb{N}_0^D} \alpha_\kappa(\boldsymbol{x}, \boldsymbol{x}')(\boldsymbol{g}_r - \boldsymbol{c})^\kappa
$$

where this series converges absolutely. We substitute it into SK_{g_r}

$$
\begin{aligned}
SK_{g_r} &= \sum_{\boldsymbol{r}\in O_r} \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[f_r^*(\boldsymbol{x})k_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r, \boldsymbol{x}'^T\boldsymbol{g}_r)f_r^*(\boldsymbol{x}')] \\
&= \sum_{\boldsymbol{r}\in O_r} \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[f_r^*(\boldsymbol{x}) \sum_{\kappa\in\mathbb{N}_0^D} \alpha_\kappa(\boldsymbol{x}, \boldsymbol{x}')(\boldsymbol{g}_r - \boldsymbol{c})^\kappa f_r^*(\boldsymbol{x}')] \\
&= \sum_{\boldsymbol{r}\in O_r} \sum_{\kappa\in\mathbb{N}_0^D} \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[\alpha_\kappa(\boldsymbol{x}, \boldsymbol{x}')f_r^*(\boldsymbol{x})f_r^*(\boldsymbol{x}')](\boldsymbol{g}_r - \boldsymbol{c})^\kappa
\end{aligned}
$$

which also converges absolutely with radius of convergence R_k. The third equality is from the Fubini's theorem. The conditions of Fubini's theorem can be verified by fact that f_r^* is square integrable (assumption 2), and the power series of k_{rg_r} converges absolutely. Thus, by definition of real analytic function, *SKSD-g* is real analytic w.r.t each $\boldsymbol{g}_r$. This also implies *SKSD-rg* (SK_{rg_r}) is real analytic w.r.t. $\boldsymbol{g}_r$ (because SK_{rg_r} is just SK_{g_r} without summation over O_r).

For the real analyticity w.r.t $\boldsymbol{r}$, the proof is almost the same. The inner product $s_p^r(\boldsymbol{x}) - s_q^r(\boldsymbol{x})$ is real analytic w.r.t $\boldsymbol{r}$ obviously for given $\boldsymbol{x}$. We also use the fact that real analyticity is preserved under multiplication of two real analytic functions. In addition, note that $k_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r, \boldsymbol{x}^T\boldsymbol{g}_r)$ act as a constant w.r.t. $\boldsymbol{r}$, we can directly apply the Fubini's theorem again to form a power series w.r.t. $\boldsymbol{r}$ with absolute convergence. Thus, SK_{rg_r} is real analytic w.r.t. $\boldsymbol{r}$ for any $\boldsymbol{g}_r$. Thus, SK_{rg_r} is real analytic to both $\boldsymbol{r}$ and $\boldsymbol{g}_r$. □

Next, we introduce an important property of real analytic function:

Lemma 2 (Zero Set Theorem (Mityagin, 2015))**.** *Let $f(\boldsymbol{x})$ be a real analytic function on (a connected open domain $\mathcal{U}$ of)$\mathbb{R}^d$. If f is not identically 0, then its zero set*

$$
S(f) := \{\boldsymbol{x} \in \mathcal{U} | f(x) = 0\}
$$

has a measure 0, i.e. $mes_d S(f) = 0$

With the help from the zero-set theorem, we can prove the validity of SK_{g_r} (or SK_{rg_r}) with finite random slices $\boldsymbol{g_r}$ (and $\boldsymbol{r}$).

Proof of theorem 1

Proof. We first deal with the validity of $\boldsymbol{g}_r$ with fixed orthogonal basis O_r. It is trivial that when $p = q$, $SK_{g_r} = 0$ identically. Now, assume $p \neq q$, then, from the theorem 3 in (Gong et al., 2021), the orthogonal SKSD (Eq.48) is a valid discrepancy. Namely, we have

$$
\sum_{\boldsymbol{r}\in O_r} \int q_{g_r}(\boldsymbol{g}_r) ||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}} > 0 \qquad (48)
$$

We should note that the distribution q_{g_r} is originally defined on $\mathbb{S}^{D-1}$. But, we can easily generalize it to larger spaces.

As for $\boldsymbol{g}_r \in \mathbb{R}^D$, we can always write $\boldsymbol{g}_r = c\boldsymbol{g}'_r$, where $\boldsymbol{g}'_r \in \mathbb{S}^{D-1}$, and $c \geq 0$. As the domain for $\boldsymbol{g}_r$ is $\mathbb{R}^D$, the $\boldsymbol{g}_r$ can represents all possible directions. Thus, we can follow the same proof logic as theorem 3 in (Gong et al., 2021) to show the corresponding discrepancy is greater than 0 when $p \neq q$.

Therefore, Eq.48 represents there exists a $\boldsymbol{r} \in O_r$ such that $||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}} > 0$ for a set of $\boldsymbol{g}_r$ with non-zero measure. Namely, $||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}}$ is not 0 identically. Thus, from the propositon 3 and lemma 2, the set of $\boldsymbol{g}_r$ that make $||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}} = 0$ has a 0 measure. Then, if $\boldsymbol{g}_r$ is sampled from some distribution η_g with density supported on $\mathbb{R}^D$ (e.g. Gaussian distribution), we have

$$SK_{g_r} = \sum_{\boldsymbol{r} \in O_r} ||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}} > 0$$

almost surely.

Now, we show that SK_{rg_r} is also a valid discrepancy with $\boldsymbol{r} \sim \eta_r$. First, due to the validity of *integrated SKSD*, we have

$$\int q_r(\boldsymbol{r}) \int q_{g_r}(\boldsymbol{g}_r)||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}} d\boldsymbol{g}_r d\boldsymbol{r} > 0 \tag{49}$$

Due to the real analyticity of SK_{rg_r} ($||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}}$) w.r.t $\boldsymbol{r}$, we can easily show that

$$\int q_{g_r}(\boldsymbol{g}_r)||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}} d\boldsymbol{g}_r$$

is real analytic to $\boldsymbol{r}$ and it is not 0 identically. Thus, by lemma 2, for $\boldsymbol{r} \sim \eta_r$, we have

$$\int q_{g_r}(\boldsymbol{g}_r)||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}} d\boldsymbol{g}_r > 0$$

Namely, $||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}} > 0$ for a set of $\boldsymbol{g}_r$ with non-zero measure. In the beginning of the proof, we show that this set of $\boldsymbol{g}_r$ is almost everywhere in $\mathbb{R}^D$ due to its real analyticity. Namely, $||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}} > 0$ for $\boldsymbol{r} \sim \eta_r$ and $\boldsymbol{g}_r \sim \eta_g$ if $p \neq q$. Thus, we can conclude that for $SK_{rg_r} = 0$ if and only if $p = q$ almost surely for $\boldsymbol{r} \sim \eta_r$ and $\boldsymbol{g}_r \sim \eta_g$. □

Corollary 6.1 (Normalizing $\boldsymbol{g}_r$)**.** *Assume the conditions in theorem 1 are satisfied, then the following operations do not violate the validity of SKSD-rg SK_{rg_r}. (1) For $\boldsymbol{g}'_r, \boldsymbol{r}' \in \mathbb{S}^{D-1}$, we define $\boldsymbol{g}_r = \boldsymbol{g}'_r + \boldsymbol{\gamma}_g$ and $\boldsymbol{r} = \boldsymbol{r}' + \boldsymbol{\gamma}_r$, where γ_r, γ_g are the noise from Gaussian distribution. (2) Define $\tilde{\boldsymbol{g}}_r = c_g \times \boldsymbol{g}_r$ and $\tilde{\boldsymbol{r}} = c_r \times \boldsymbol{r}$, where $\tilde{\boldsymbol{g}}_r, \tilde{\boldsymbol{r}}$ are unit vectors and $c_r, c_g > 0$. The resulting active slices $\tilde{\boldsymbol{r}}$ and $\tilde{\boldsymbol{g}}_r$ do not violate the validity of SK_{rg_r}.*

Proof. From the theorem 1 with $\boldsymbol{g}_r, \boldsymbol{r}$, when $p \neq q$, we have

$$\begin{aligned} SK_{rg_r} &= ||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}} \\ &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[f^*_r(\boldsymbol{x})k_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r, \boldsymbol{x}'^T\boldsymbol{g}_r)f^*_r(\boldsymbol{x}')] \\ &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[c_r^2 f^*_{\tilde{r}}(\boldsymbol{x})k_{rg_r}(c\boldsymbol{x}^T\tilde{\boldsymbol{g}}_r, c\boldsymbol{x}'^T\tilde{\boldsymbol{g}}_r)f^*_{\tilde{r}}(\boldsymbol{x}')] > 0 \end{aligned}$$

From the assumption 6 that $k_{rg_r}(c\boldsymbol{x}^T\tilde{\boldsymbol{g}}_r, c\boldsymbol{x}'^T\tilde{\boldsymbol{g}}_r) = k'_{rg_r}(\boldsymbol{x}^T\tilde{\boldsymbol{g}}_r, \boldsymbol{x}'^T\tilde{\boldsymbol{g}}_r)$. So this is equivalent to the *SKSD-rg* defined with a new c_0-universal kernel k'_{rg_r} and $\tilde{\boldsymbol{g}}_r, \tilde{\boldsymbol{r}} \in \mathbb{S}^{D-1}$. Thus, the corresponding *maxSKSD-rg* with $\tilde{\boldsymbol{g}}_r, \tilde{\boldsymbol{r}} \in \mathbb{S}^{D-1}$ is a valid discrepancy almost surely. □

E.2. Relationship beetween SSD and SKSD

Proof of proposition 1

Proof. We consider the *SSD-rg* (S_{rg_r}) without the optimal test function:

$$\mathbb{E}_q[s^r_p(\boldsymbol{x})h_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r) + \boldsymbol{r}^T\boldsymbol{g}_r\nabla_{\boldsymbol{x}^T\boldsymbol{g}_r}h_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r)] \tag{50}$$

From the Stein identity (Eq.23), we can let $\boldsymbol{f}(\boldsymbol{x}) = \left[r_1 h_{rg_r}\left(\boldsymbol{x}^T\boldsymbol{g}_r\right), r_2 h_{rg_r}\left(\boldsymbol{x}^T\boldsymbol{g}_r\right), \ldots, r_D h_{rg_r}\left(\boldsymbol{x}^T\boldsymbol{g}_r\right)\right]^T$ and then take the trace. Thus, we have

$$\mathbb{E}_q[s^r_q(\boldsymbol{x})h_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r)] = \mathbb{E}_q[\boldsymbol{r}^T\boldsymbol{g}_r\nabla_{\boldsymbol{x}^T\boldsymbol{g}_r}h_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r)]$$

Substitute it into Eq.50 and change the variable to $\boldsymbol{y} = \boldsymbol{G}_r\boldsymbol{x}$, we have

$$\begin{aligned} &\mathbb{E}_q[(s^r_p(\boldsymbol{x}) - s^r_q(\boldsymbol{x}))h_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r)] \\ =&\int q_{G_r}(y_d, \boldsymbol{y}_{-d}) \underbrace{\left(s^r_p(\boldsymbol{G}_r^{-1}\boldsymbol{y}) - s^r_q(\boldsymbol{G}_r^{-1}\boldsymbol{y})\right)}_{f^*_r(\boldsymbol{G}_r^{-1}\boldsymbol{y})} h_{rg_r}(y_d)d\boldsymbol{y} \\ =&\int q_{G_r}(y_d) \underbrace{\int q_{G_r}(\boldsymbol{y}_{-d}|y_d)f^*_r(\boldsymbol{G}_r^{-1}\boldsymbol{y})d\boldsymbol{y}_{-d}}_{h^*_{rg_r}(y_d)} h_{rg_r}(y_d)dy_d \\ \leq&\sqrt{\mathbb{E}_{q_{G_r}(y_d)}[h^*_{rg_r}(y_d)^2]}\sqrt{\mathbb{E}_{q_{G_r}(y_d)}[h_{rg_r}(y_d)^2]} \end{aligned}$$

where the last inequality is from Cauchy-Schwarz inequality, where the equality holds when

$$\begin{aligned} &h_{rg_r}(y_d) \propto h^*_{rg_r}(y_d) \\ &= \mathbb{E}_{q_{G_r}(\boldsymbol{y}_{-d}|y_d)}\left[\left(s^r_p\left(\boldsymbol{G}_r^{-1}\boldsymbol{y}\right) - s^r_q\left(\boldsymbol{G}_r^{-1}\boldsymbol{y}\right)\right)\right] \end{aligned}$$

where $y_d = \boldsymbol{x}^T\boldsymbol{g}_r$. □

Proof of theorem 2

Proof. Let's first re-write of S_{rg_r} and SK_{rg_r}.

$$\begin{aligned}
&S_{rg_r} = \sup_{h_{rg_r}\in\mathcal{F}_q} \mathbb{E}_q[(s_p^r(\boldsymbol{y}) - s_q^r(\boldsymbol{x}))h_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r)] \\
=&\mathbb{E}_{q_{G_r}(y_d)}[\underbrace{\int q_{G_r}(\boldsymbol{y}_{-d}|y_d)(s_p^r(\boldsymbol{G}_r^{-1}\boldsymbol{y}) - s_q^r(\boldsymbol{G}_r^{-1}\boldsymbol{y}))d\boldsymbol{y}_{-d}}_{h^*_{rg_r}(y_d)} \\
&\times h^*_{rg_r}(y_d)] \\
=&\mathbb{E}_{q_{G_r}(y_d)}[h^*_{rg_r}(y_d)^2]
\end{aligned}$$

where the second equality is from proposition 1.

$$SK_{rg_r} = \langle \mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})], \mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]\rangle_{\mathcal{H}_k}$$

where $\xi_{p,r,g_r}(\boldsymbol{x},\cdot)$ is defined in Eq.5, and $\langle\cdot,\cdot\rangle_{\mathcal{H}_{rg_r}}$ is the RKHS inner product induced by kernel k_{rg_r}. By simple algebraic manipulation and Stein identity (Eq.23), we have

$$\begin{aligned}
&\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x},\cdot)] \\
=&\mathbb{E}_q[(s_p^r(\boldsymbol{x}) - s_q^r(\boldsymbol{x}))k_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r,\cdot)] \\
=&\mathbb{E}_{q_{G_r}(y_d)}[\underbrace{\int q_{G_r}(\boldsymbol{y}_{-d}|y_d)(s_p^r(\boldsymbol{G}_r^{-1}\boldsymbol{y}) - s_q^r(\boldsymbol{G}_r^{-1}\boldsymbol{y}))d\boldsymbol{y}_{-d}}_{h^*_{rg_r}(y_d)} \\
&\times k_{rg_r}(y_d,\cdot)] \\
=&\mathbb{E}_{q_{G_r}(y_d)}[h^*_{rg_r}(y_d)k_{rg_r}(y_d,\cdot)]
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&SK_{rg_r} \\
=&\mathbb{E}_{y_d,y'_d\sim q_{G_r}(y_d)}[h^*_{rg_r}(y_d)k_{rg_r}(y_d,y'_d)h^*_{rg_r}(y'_d)] \\
\leq&\sqrt{\mathbb{E}_{y_d,y'_d}[k_{rg_r}(y_d,y'_d)^2]}\sqrt{\mathbb{E}_{y_d}[h^*_{rg_r}(y_d)^2]}\sqrt{\mathbb{E}_{y'_d}[h^*_{rg_r}(y'_d)^2]} \\
=&MS^*_{rg_r}
\end{aligned}$$

where constant M is from the bounded kernel assumption, and the inequality is from Cauchy-Schwarz inequality. Without the loss of generality, we can set $M = 1$. For other value of $M > 0$, one can always set the optimal test function $(h^*_{rg_r})$ for *SSD-rg* with coefficient M. The the new *SSD-g* will be M multiplied by the original *SSD-rg* with $M = 1$.

Thus, *SSD-rg* is an upper bound for *SKSD-rg*. From the assumption 1, we know that the induced set $\mathcal{K} = \{y \in \mathbb{R} | y = \boldsymbol{x}^T\boldsymbol{g}, ||g|| = 1, \boldsymbol{x} \in \mathcal{X}\}$ is LCH, and the kernel $k_{rg_r} : \mathcal{K}\times\mathcal{K} \to \mathbb{R}$ is c_0-universal. Then, from (Sriperumbudur et al., 2011), c_0-universal implies L_p-universal. Namely, the induced RKHS $\mathcal{H}_{rg_r}$ is dense in $L^p(\mathcal{K};\mu)$ with all Borel probability measure μ w.r.t. *p-norm*, defined as

$$||f||_p = \left(\int |f(\boldsymbol{x})|^p d\mu(\boldsymbol{x})\right)^{\frac{1}{p}}$$

Now, from the assumption 4, we know $h^*_{rg_r}(y_d)$ is bounded for all possible $\boldsymbol{g}_r$, we have

$$\int q_{G_r}(y_d)|h^*_{rg_r}(y_d)|^2 dy_d < \infty$$

This means $h^*_{rg_r} \in L^2(\mathcal{K},\mu_{G_r})$, where μ_{G_r} is the probability measure with density $q_{G_r}(y_d)$

From the L_p-universality, there exists a function $\widetilde{h^*_{rg_r}} \in \mathcal{H}_{rg_r}$, such that for any given $\epsilon > 0$,

$$||h^*_{rg_r} - \widetilde{h^*_{rg_r}}||_2 < \epsilon$$

Let's define $\widetilde{SK_{rg_r}}$ is the *SKSD-rg* with the specific kernelized test function $\widetilde{h^*_{rg_r}}$, and from the optimality of *SKSD-rg*, we have

$$SK_{rg_r} \geq \widetilde{SK_{rg_r}}$$

Therefore, we have

$$\begin{aligned}
0 \leq& S_{rg_r} - SK_{rg_r} \\
\leq& S_{rg_r} - \widetilde{SK_{rg_r}} \\
=& \mathbb{E}_q[(s_p^r(\boldsymbol{x}) - s_q^r(\boldsymbol{x}))(h^*_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r) - \widetilde{h^*_{rg_r}}(\boldsymbol{x}^T\boldsymbol{g}_r))] \\
\leq& \underbrace{\sqrt{\mathbb{E}_q[(s_p^r(\boldsymbol{x}) - s_q^r(\boldsymbol{x}))^2]}}_{C_r} \\
&\times \sqrt{\mathbb{E}_q[(h^*_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r) - \widetilde{h^*_{rg_r}}(\boldsymbol{x}^T\boldsymbol{g}_r))^2]} \\
=& C_r\sqrt{\int q_{G_r}(y_d,\boldsymbol{y}_{-d})(h^*_{rg_r}(y_d) - \widetilde{h^*_{rg_r}}(y_d))^2 d\boldsymbol{y}} \\
=& C_r||h^*_{rg_r} - \widetilde{h^*_{rg_r}}||_2 < C_r\epsilon
\end{aligned}$$

From assumption 2, we know $s_p^r(\boldsymbol{x}) - s_q^r(\boldsymbol{x})$ is square integrable for all possible $\boldsymbol{r}$. Therefore, let's define $C = \max_{\boldsymbol{r}\in\mathbb{S}^{D-1}} C_r$, then,

$$0 \leq S_{rg_r} - SK_{rg_r} < C\epsilon$$

□

F. Theory related to active slice g

F.1. Optimal test function for PSD

Proposition 4 (Optimality of PSD)**.** *Assume the assumption* $1-3$ *(density regularity) are satisfied, then the optimal test function for PSD given* O_r *is proportional to the projected score difference, i.e.*

$$f_r^*(\boldsymbol{x}) \propto \left(s_p^r(\boldsymbol{x}) - s_q^r(\boldsymbol{x})\right) \tag{51}$$

Thus,

$$PSD(q,p;O_r) = \sum_{\boldsymbol{r}\in O_r} \mathbb{E}_q[(s_p^r(\boldsymbol{x}) - s_q^r(\boldsymbol{x}))^2] \tag{52}$$

if the coefficient of f_r^* *to be 1.*

Proof. From the Stein identity (Eq.23), we can re-write the inner part of the supremum of Eq.9 as

$$\begin{aligned}
&\mathbb{E}_q[s_p^r(\boldsymbol{x})f_r(\boldsymbol{x}) + \boldsymbol{r}^T\nabla_{\boldsymbol{x}}f_r(\boldsymbol{x})] \\
=&\mathbb{E}_q[(s_p^r(\boldsymbol{x}) - s_q^r(\boldsymbol{x}))f_r(\boldsymbol{x})]
\end{aligned}$$

Then, we can upper bound the PSD (Eq.9) as the following

$$\sum_{\boldsymbol{r}\in O_r} \mathbb{E}_q[(s_p^r(\boldsymbol{x}) - s_q^r(\boldsymbol{x}))f_r(\boldsymbol{x})]$$
$$\leq \sum_{\boldsymbol{r}\in O_r} \sqrt{\mathbb{E}_q[(s_p^r(\boldsymbol{x}) - s_q^r(\boldsymbol{x}))^2]}\sqrt{\mathbb{E}_q[(f_r(\boldsymbol{x}))^2]}$$

by Cauchy-Schwarz inequality. It is well-known that the equality holds when $f_r(\boldsymbol{x}) \propto (s_p^r(\boldsymbol{x}) - s_q^r(\boldsymbol{x}))$ □

F.2. Proof of Theorem 3

Proof. The key to this proof is to notice that $h^*_{rg_r}$ is the conditional mean of f_r^* w.r.t. the transformed distribution q_{G_r}. By using the similar terminology of proposition 1, and let $s_p^r = s_p^r(\boldsymbol{x})$ for abbreviation. Then,

$$\begin{aligned}
&\mathbb{E}_q[(s_p^r - s_q^r)f_r^*(\boldsymbol{x})] - \mathbb{E}_q[(s_p^r - s_q^r)h^*_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r)] \\
=&\int q(\boldsymbol{x})[(s_p^r - s_q^r)^2 - (s_p^r - s_q^r)h^*_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r)]d\boldsymbol{x} \\
=&\int q_{G_r}(y_d)[\int q_{G_r}(\boldsymbol{y}_{-d}|y_d)(s_p^r(\boldsymbol{G}_r^{-1}\boldsymbol{y}) - s_q^r(\boldsymbol{G}_r^{-1}\boldsymbol{y}))^2 d\boldsymbol{y}_{-d} \\
&-\underbrace{\int q_{G_r}(s_p^r(\boldsymbol{G}_r^{-1}\boldsymbol{y}) - s_q^r(\boldsymbol{G}_r^{-1}\boldsymbol{y}))d\boldsymbol{y}_{-d}}_{h^*_{rg_r}(y_d)} h^*_{rg_r}(y_d)]dy_d \\
=&\int q_{G_r}(y_d)[\int q_{G_r}(\boldsymbol{y}_{-d}|y_d)(f_r^*(\boldsymbol{G}_r^{-1}\boldsymbol{y}) - h^*_{rg_r}(y_d))^2]d\boldsymbol{y} \\
=&\mathbb{E}_q[(f_r^*(\boldsymbol{x}) - h^*_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r))^2] \geq 0
\end{aligned}$$

where the 3rd equality is due to the fact that $h^*_{rg_r}$ is the conditional mean of f_r^*. Thus,

$$\begin{aligned}
&PSD - S_{g_r} \\
=&\sum_{\boldsymbol{r}\in O_r} \mathbb{E}_q[(s_p^r - s_q^r)f_r^*(\boldsymbol{x})] - \mathbb{E}_q[(s_p^r - s_q^r)h^*_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r)] \\
=&\sum_{\boldsymbol{r}\in O_r} \mathbb{E}_q[(f_r^*(\boldsymbol{x}) - h^*_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r))] \geq 0
\end{aligned}$$

□

F.3. Proof of Theorem 4

Before we give the details, we introduce the main inequality and its variant for the proof.

Lemma 3 (Poincaré Inequality)**.** *For a probabilistic distribution p that satisfies assumption 7, for all locally Lipschitz function $f(\boldsymbol{x}) : \mathcal{X} \subseteq \mathbb{R}^D \to \mathbb{R}$, we have the following inequality*

$$\mathrm{Var}_p(f(\boldsymbol{x})) \leq C_p \int p(\boldsymbol{x})||\nabla_{\boldsymbol{x}} f(\boldsymbol{x})||^2 d\boldsymbol{x}$$

where C_p is called Poincaré constant that is only related to p.

One should note that the assumption of log concavity of p is a sufficient condition for Poincaré inequality, which means it may be applied to a broader class of distributions. But it is beyond the scope of this work.

Due to the form of optimal test functions of *SSD-g*, we need to deal with the transformed distribution q_{G_r} and its conditional expectations (see Eq.8). Unfortunately, the original form of Poincaré inequality cannot be applied. In the following, we introduce its variant called *subspace* Poincaré inequality (Constantine et al., 2014; Zahm et al., 2020; Parente et al., 2020) to deal with the conditional expectation. But before that, we need to make sure the transformed distribution and its conditional density still satisfy the conditions of Poincaré inequality, i.e. log concavity.

Lemma 4 (Preservation of log concavity)**.** *Assume distribution $q(\boldsymbol{x}) = \exp(-V(\boldsymbol{x}))$ is log-concave. With arbitrary orthogonal matrix $\boldsymbol{G}$ and corresponding transformed distribution q_G, the conditional distribution $q_G(\boldsymbol{y}_{-d} \mid y_d)$ is also log-concave for all $d = 1, \ldots, D$.*

Proof. Assume we have $\boldsymbol{y} = \boldsymbol{G}\boldsymbol{x}$. Thus, by change of variable formula, $q_G(\boldsymbol{y}) = q(\boldsymbol{x}) = q\left(\boldsymbol{G}^{-1}\boldsymbol{y}\right) = \exp\left(-V\left(\boldsymbol{G}^{-1}\boldsymbol{y}\right)\right)$. Thus, the log conditional distribution

$$\log q_G(\boldsymbol{y}_{-d} \mid y_d) = -V\left(\boldsymbol{G}^{-1}\boldsymbol{y}\right) - \log q_G(y_d)$$

We inspect its Hessian w.r.t $\boldsymbol{y}_{-d}$

$$\begin{aligned}
&\nabla^2_{\boldsymbol{y}_{-d}}\left(V\left(\boldsymbol{G}^{-1}\boldsymbol{y}\right) + \log q_G(y_d)\right) \\
=&\nabla^2_{\boldsymbol{y}_{-d}}\left(V\left(\boldsymbol{G}^{-1}\boldsymbol{y}\right)\right) \\
=&\nabla_{\boldsymbol{y}_{-d}}\left(\boldsymbol{G}_{\setminus d}V'\left(\boldsymbol{G}^{-1}\boldsymbol{y}\right)\right) \\
=&\boldsymbol{G}_{\setminus d}V''\left(\boldsymbol{G}^{-1}\boldsymbol{y}\right)\boldsymbol{G}_{\setminus d}^T
\end{aligned}$$

where $\boldsymbol{G}_{\setminus d} = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{d-1}, \boldsymbol{g}_{d+1}, \ldots, \boldsymbol{g}_D]^T$ and $V'\left(\boldsymbol{G}^{-1}\boldsymbol{y}\right) = \nabla_{\boldsymbol{G}^{-1}\boldsymbol{y}}V\left(\boldsymbol{G}^{-1}\boldsymbol{y}\right)$. We already know that $V(\cdot)$ is a convex function. Thus, for all $\boldsymbol{u} \in \mathbb{R}^D$, $\boldsymbol{u}^T V''(\boldsymbol{x})\boldsymbol{u} \geq 0$, therefore,

$$\boldsymbol{u}^T\boldsymbol{G}_{\setminus d}V''\left(\boldsymbol{G}^{-1}\boldsymbol{y}\right)\boldsymbol{G}_{\setminus d}^T\boldsymbol{u} = \boldsymbol{l}^T V''\left(\boldsymbol{G}^{-1}\boldsymbol{y}\right)\boldsymbol{l} \geq 0$$

where $\boldsymbol{l} = \boldsymbol{G}_{\setminus d}^T\boldsymbol{u}$. □

Now, we can introduce the subspace Poincaré inequality

Lemma 5 (Poincaré inequality for conditional expectation)**.** *Assume the assumption 2,4 (density regularity), 7 (Poincaré inequality condition) are satisfied, with arbitrary orthogonal matrix $\boldsymbol{G}$, $\boldsymbol{y} = \boldsymbol{G}\boldsymbol{x}$ and $y_d = \boldsymbol{x}^T\boldsymbol{g}$, we have the following inequality*

$$\begin{aligned}
&\int q_G(\boldsymbol{y}_{-d} \mid y_d)\left[f_r^*\left(\boldsymbol{G}^{-1}\boldsymbol{y}\right) - h^*_{rg_r}(y_d)\right]^2 d\boldsymbol{y}_{-d} \\
\leq& C_{yd}\mathbb{E}_{q_G(\boldsymbol{y}_{-d}|y_d)}\left[\left\|\boldsymbol{G}_{\setminus d}\nabla f_r^*\right\|^2\right]
\end{aligned}$$

where C_{y_d} is the Poincaré constant, $\boldsymbol{G}_{\setminus d} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{d-1}, \boldsymbol{a}_{d+1}, \ldots, \boldsymbol{a}_D]^T$ is the orthogonal matrix $\boldsymbol{G}$ excluding $\boldsymbol{a}_d = \boldsymbol{g}$ and f_r^*, $h_{rg_r}^*$ are the optimal test functions defined in proposition 4, 1 respectively with coefficient 1.

Proof. From lemma 4, we know $q_G(\boldsymbol{y}_{-d} \mid y_d)$ is a log-concave distribution. Therefore, it satisfies the Poincaré inequality (lemma.3). We have

$$\begin{aligned}
&\int q_G(\boldsymbol{y}_{-d} \mid y_d) \left[f_r^*\left(\boldsymbol{G}^{-1}\boldsymbol{y}\right) - h_{rg_r}^*(y_d)\right]^2 d\boldsymbol{y}_{-d} \\
=&Var_{q_G(\boldsymbol{y}_{-d}|y_d)}(f_r^*(\boldsymbol{G}^{-1}\boldsymbol{y})) \\
\leq& C_{y_d} \int q_G(\boldsymbol{y}_{-d} \mid y_d) \left\|\nabla_{\boldsymbol{y}_{-d}} f_r^*\left(\boldsymbol{G}^{-1}\boldsymbol{y}\right)\right\|^2 d\boldsymbol{y}_{-d} \\
=& C_{y_d} \int q_G(\boldsymbol{y}_{-d} \mid y_d) \left\|\boldsymbol{G}_{\setminus d} \nabla_{\boldsymbol{G}^{-1}\boldsymbol{y}} f_r^*\left(\boldsymbol{G}^{-1}\boldsymbol{y}\right)\right\|^2 d\boldsymbol{y}_{-d}
\end{aligned}$$

The first equality comes from the fact that $h_{rg_r}^*(y_d)$ is actually a conditional mean of $f_r^*(\boldsymbol{G}^{-1}\boldsymbol{y})$, and the inequality comes from the direct application of Poincaré inequality on $q_G(\boldsymbol{y}_{-d}|y_d)$ and $f_r^*(\boldsymbol{G}^{-1}\boldsymbol{y})$. □

With the above tools, it is now easy to prove theorem 4.

Theorem 4

Proof. We can re-write the inner part of controlled approximation (Eq.11) in the following:

$$\begin{aligned}
&\int q_{G_r}(y_d, \boldsymbol{y}_{-d}) \left[f_r^*\left(\boldsymbol{G}_r^{-1}\boldsymbol{y}\right) - h_{rg_r}^*(y_d)\right]^2 d\boldsymbol{y} \\
=& \int q_{G_r}(y_d) \mathbb{E}_{q_{G_r}(\boldsymbol{y}_{-d}|y_d)}[(f_r^*(\boldsymbol{G}_r^{-1}\boldsymbol{y}) - h_{rg_r}(y_d))^2] d\boldsymbol{y} \\
\leq& \int q_{G_r}(y_d, \boldsymbol{y}_{-d})\, C_{y_d} \left\|\boldsymbol{G}_{r\setminus d} \nabla f_r^*\right\|^2 d\boldsymbol{y} \\
\leq& C_{sup} \int q_{G_r}(y_d, \boldsymbol{y}_{-d}) \left\|\boldsymbol{G}_{r\setminus d} \nabla f_r^*\right\|^2 d\boldsymbol{y} \\
=& C_{sup} \int q(\boldsymbol{x}) \operatorname{tr}\left[\left(\boldsymbol{G}_{r\setminus d} \nabla f_r^*\right)\left(\boldsymbol{G}_{r\setminus d} \nabla f_r^*\right)^T\right] d\boldsymbol{x} \\
=& C_{sup} \operatorname{tr}\left[\boldsymbol{G}_{r\setminus d} \boldsymbol{H}_r \boldsymbol{G}_{r\setminus d}^T\right]
\end{aligned}$$

where the first inequality is directly from lemma 5 and the second inequality is from the definition of C_{sup}.

To minimize this upper bound, we can directly use the theorem 2.1 (Sameh & Tong, 2000) by setting $B = I$ and $\boldsymbol{X} = \boldsymbol{G}_{r\setminus d}^T$. Therefore, we only need to check if $\boldsymbol{G}_{r\setminus d}\boldsymbol{G}_{r\setminus d}^T = \boldsymbol{I}$. This is trivial as $\boldsymbol{G}_r$ is an orthogonal matrix. Thus, the proof is complete. □

G. Theory related to active slice r

G.1. Proof of proposition 2

First, from the theorem 3, we have

$$\text{PSD}_r \geq S_{r,g}$$

Thus, we can establish the following lower bound

$$\begin{aligned}
&S_{r_1,g_{r_1}} - S_{r_2,g_{r_2}} \\
\geq& S_{r_1,g_{r_1}} - \text{PSD}_{r_2} \\
=& \underbrace{S_{r_1,g_{r_1}} - \text{PSD}_{r_1}}_{\text{controlled approximation}} + \text{PSD}_{r_1} - \text{PSD}_{r_2}
\end{aligned}$$

Thus, from theorem 4, we can obtain

$$\begin{aligned}
&S_{r_1,g_{r_1}} - \text{PSD}_{r_1} \\
=& -\mathbb{E}_q\left[(f_{r_1}^*(\boldsymbol{x}) - h_{r_1 g_{r_1}}^*(\boldsymbol{x}^T\boldsymbol{g}_{r_1})^2)\right] \\
\geq& -C_{\text{sup}}\text{tr}(\boldsymbol{G}_{r_1\setminus d}\boldsymbol{H}_{r_1}\boldsymbol{G}_{r_1\setminus d}^T) \\
=& -C_{\text{sup}}\text{tr}(\boldsymbol{H}) + \underbrace{\boldsymbol{g}_{r_1}^T\boldsymbol{H}_{r_1}\boldsymbol{g}_{r_1}}_{\geq 0} \\
\geq& -C_{\text{sup}}\text{tr}(\boldsymbol{H}_{r_1})
\end{aligned}$$

where the first inequality is from the upper bound of controlled approximation (theorem 4) and $\boldsymbol{g}_{r_1}^T\boldsymbol{H}_{r_1}\boldsymbol{g}_{r_1} \geq 0$ is due to the positive semi-definiteness of $\boldsymbol{H}_{r_1}$. Assume we have an orthogonal basis O_{r_1} that contains $\boldsymbol{r}_1$, thus, for each $\boldsymbol{r} \in O_{r_1}$, we have $\text{tr}(\boldsymbol{H}_r) \geq 0$. Then, we can show

$$\begin{aligned}
\text{tr}(\boldsymbol{H}_{r_1}) \leq& \sum_{\boldsymbol{r}\in O_{r_1}} \text{tr}(\boldsymbol{H}_r) \\
=& \sum_{\boldsymbol{r}\in O_{r_1}} \text{tr}(\mathbb{E}_q[\nabla_{\boldsymbol{x}}\boldsymbol{f}^*(\boldsymbol{x})\boldsymbol{r}\boldsymbol{r}^T\nabla_{\boldsymbol{x}}\boldsymbol{f}^*(\boldsymbol{x})^T]) \\
=& \text{tr}(\mathbb{E}_q[\nabla_{\boldsymbol{x}}\boldsymbol{f}^*(\boldsymbol{x}) \sum_{\boldsymbol{r}\in O_{r_1}} \boldsymbol{r}\boldsymbol{r}^T\nabla_{\boldsymbol{x}}\boldsymbol{f}^*(\boldsymbol{x})^T]) \\
=& \text{tr}(\mathbb{E}_q[\nabla_{\boldsymbol{x}}\boldsymbol{f}^*(\boldsymbol{x})\nabla_{\boldsymbol{x}}\boldsymbol{f}^*(\boldsymbol{x})^T]) \\
=& \sum_{i=1}^D \omega_i = \Omega
\end{aligned}$$

where $\{\omega_i\}_i^D$ are the eigenvalues of $\mathbb{E}_q[\nabla_{\boldsymbol{x}}\boldsymbol{f}^*(\boldsymbol{x})\nabla_{\boldsymbol{x}}\boldsymbol{f}^*(\boldsymbol{x})^T]$, $\boldsymbol{f}^*(\boldsymbol{x}) = \boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})$ and $\sum_{\boldsymbol{r}\in O_{r_1}} \boldsymbol{r}\boldsymbol{r}^T = \boldsymbol{I}$ since $\boldsymbol{r} \in O_{r_1}$ are orthogonal to each other.

Thus, we can substitute it back, we have

$$S_{r_1,g_{r_1}} - S_{r_2,g_{r_2}} \geq \text{PSD}_{r_1} - \text{PSD}_{r_2} - C_{\text{sup}}\Omega$$

G.2. Proof of theorem 5

Proof. From proposition 4 we know $f_r^*(\boldsymbol{x}) = \left(s_p^r(x) - s_q^r(x)\right)$, thus, we can substitute into *PSD*

(Eq.9), we get

$$\text{PSD}_r = \max_{\boldsymbol{r}\in\mathbb{S}^{D-1}} \mathbb{E}_q \left[\left(\left(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})\right)^T \boldsymbol{r} \right)^2 \right]$$

To maximize it, we consider the following constraint optimization problem.

$$\max_{\boldsymbol{r}} \mathbb{E}_q \left[\left(\left(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x)\right)^T \boldsymbol{r} \right)^2 \right] \quad \text{s.t.} \quad \|\boldsymbol{r}\|^2 = 1$$

We take the derivative of the corresponding Lagrange multiplier w.r.t. $\boldsymbol{r}$,

$$\begin{aligned}
&\mathbb{E}_q \left[\nabla_{\boldsymbol{r}} \left(\left(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})\right)^T \boldsymbol{r} \right)^2 \right] - 2\lambda \boldsymbol{r} = 0 \\
\Rightarrow &\mathbb{E}_q \left[\left(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})\right)^T \boldsymbol{r} \left(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})\right) \right] = \lambda \boldsymbol{r} \\
\Rightarrow &\underbrace{\mathbb{E}_q \left[\left(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})\right) \left(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})\right)^T \right]}_{\boldsymbol{S}=\mathbb{E}_q[\boldsymbol{f}^*(\boldsymbol{x})\boldsymbol{f}^*(\boldsymbol{x})^T]} \boldsymbol{r} = \lambda \boldsymbol{r} \\
\Rightarrow &\boldsymbol{S}\boldsymbol{r} = \lambda \boldsymbol{r}
\end{aligned}$$

This exactly the problem of finding eigenpair for matrix $\boldsymbol{S}$. Let's assume $\boldsymbol{r} = \boldsymbol{v}$ which is the eigenvector of $\boldsymbol{S}$ with corresponding eigenvalue λ. Substituting it back to *PSD*, we have

$$\begin{aligned}
&\mathbb{E}_q \left[\left(\left(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})\right)^T \boldsymbol{r} \right)^2 \right] \\
=&\mathbb{E}_q \left[\left(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})\right)^T \boldsymbol{r} \left(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})\right) \right]^T \boldsymbol{r} \\
=&\boldsymbol{r}^T \boldsymbol{S} \boldsymbol{r} \\
=&\lambda \boldsymbol{v}^T \boldsymbol{v} = \lambda
\end{aligned}$$

Thus, to obtain the active slice $\boldsymbol{r}$, we only need to find the eigenvector of $\boldsymbol{S}$ with the largest eigenvalue. □

G.3. Greedy algorithm is eigen-decomposition

Corollary 6.2 (Greedy algorithm is eigen-decomposition)**.** *Assume the conditions in theorem 5 are satisfied, then finding the orthogonal basis O_r from the greedy algorithm is equivalent to the eigen-decomposition of $\boldsymbol{S}$.*

Proof. Assume we have obtained the active slice $\boldsymbol{r}$ from theorem 5, thus, we have $\boldsymbol{S}\boldsymbol{r} = \lambda \boldsymbol{r}$. The greedy algorithm for $\boldsymbol{r}'$ can be translated into the following constrained optimization

$$\begin{aligned}
&\max_{\boldsymbol{r}'} \quad \mathbb{E}_q \left[\left(\left(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x)\right)^T \boldsymbol{r}' \right)^2 \right] \\
&s.t. \quad ||\boldsymbol{r}'||^2 = 1 \\
&\quad\; \boldsymbol{r}^T \boldsymbol{r}' = 0
\end{aligned}$$

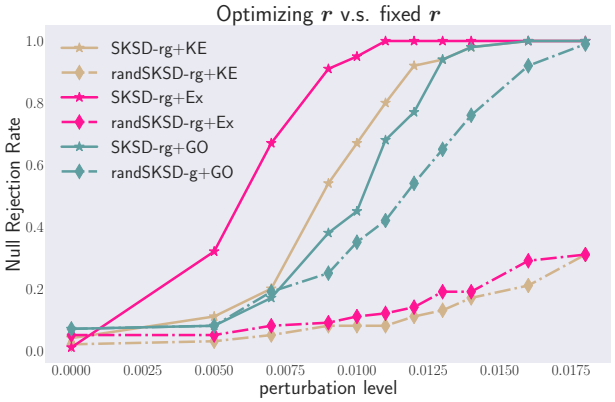

Figure 5. The test power difference with good $\boldsymbol{r}$ and fixed $\boldsymbol{r}$.

By using Lagrange multipliers (μ, γ), and then take derivative w.r.t. $\boldsymbol{r}'$,

$$\boldsymbol{S}\boldsymbol{r}' = \mu \boldsymbol{r}' + \gamma \boldsymbol{r}$$

Then taking the inner product with $\boldsymbol{r}$ in both side, and notice $\boldsymbol{S}$ is a symmetric matrix, we obtain

$$\begin{aligned}
\gamma &= \langle \boldsymbol{S}\boldsymbol{r}', \boldsymbol{r} \rangle \\
&= \langle \boldsymbol{r}', \boldsymbol{S}^T \boldsymbol{r} \rangle \\
&= \langle \boldsymbol{r}', \lambda \boldsymbol{r} \rangle = 0
\end{aligned}$$

Therefore, the constrained optimization is the same as the one in theorem 5, which is to find a eigenvector of $\boldsymbol{S}$ that is different from $\boldsymbol{r}$. Repeat the above procedure, the final resulting O_r is a group of eigenvectors of $\boldsymbol{S}$. □

H. Experiment Details

For all experiments in this paper, we use RBF kernel with median heuristics.

H.1. Benchmark GOF test

For gradient based optimization, we use Adam (Kingma & Ba, 2014) with learning rate 0.001 and $\beta = (0.9, 0.99)$. We use random initialization for SKSD-g+GO by drawing $\boldsymbol{g}_r$ from a Gaussian distribution before normalizing them to unit vectors. For kernel smooth and gradient estimator, we use RBF kernel with median heuristics. Although the algorithm 1 states that small Gaussian noise are needed for active slices, in practice, we found that active slices still have the satisfactory performance without the noise.

The significance level for GOF test $\alpha = 0.05$, and the dimensions of the benchmark problems grow from 2 to 100. We use 1000 bootstrap samples to estimate the threshold and run 100 trials for each benchmark problems.

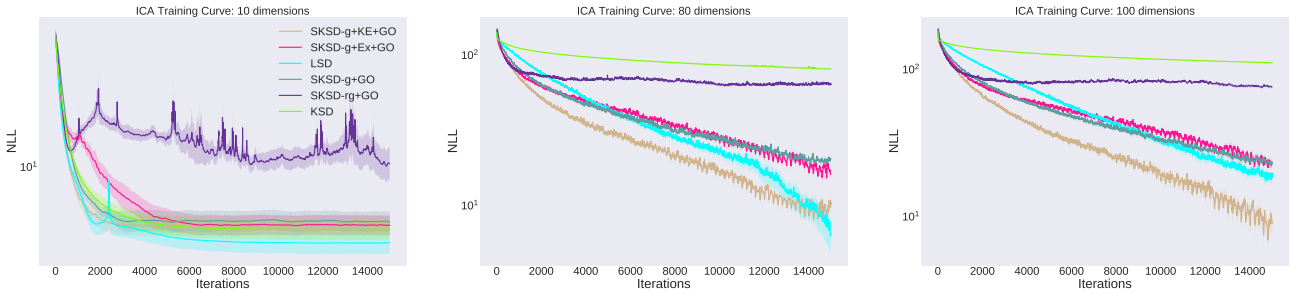

Figure 6. Training curve of ICA model with test NLL for different dimensions.

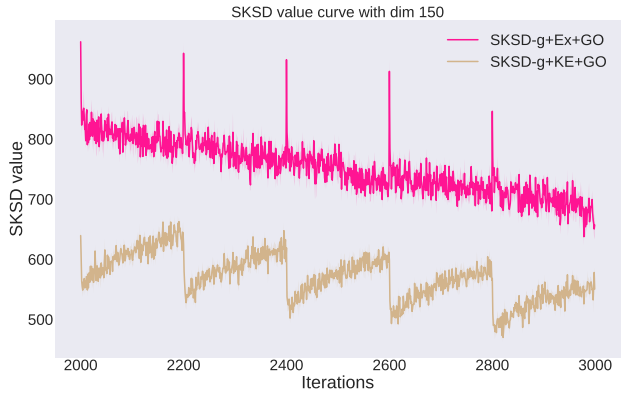

Figure 7. SKSD value curve with 150 dimensional ICA during iteration 5000 to 6000

H.2. RBM GOF test

We set significance level $\alpha = 0.05$ and use 1000 bootstrap samples to compute the threshold. For methods that require training (*SKSD* based method), we need to collect some training samples. Following the same settings as (Gong et al., 2021), to avoid over-fitting to small training set, we collect the pseudo-samples during the early burn-in stage. Note that these pseudo-samples should not be used for testing, as they are not drawn from the q. We collect 2000 samples. For gradient based optimization, we use the same optimizer as benchmark GOF test with the same hyper-parameters. The batch size is 100. For initialization of *SKSD+GO*, we found that if the slices are initialized randomly, the gradient optimization fails to find meaningful slices within a reasonable amount of time, therefore, we have initialize the $\boldsymbol{r}$ and $\boldsymbol{g}_r$ as one-hot vectors and set $\boldsymbol{r} = \boldsymbol{g}_r$. For pruning ablation study, if the pruning level is set to 50, we initialize $\boldsymbol{r}$ and $\boldsymbol{g}_r$ to be the identity matrix. The default number of gradient optimization for *SKSD+GO* is 50. For active slice method, we directly use the active slices without any further optimizations. We run 100 trials for GOF test with 1000 test samples per trial.

(Gong et al., 2021) reports *SKSD-rg+GO* has near optimal test power at perturbation level 0.01. The performance difference is because they train the *SKSD-rg* with 200 batch sizes per burn-in step. Namely, the training set size are $200 \times 2000 = 400000$, which is 200 times larger than ours. They also run 2000 iterations, which is equivalent to 100 epochs in our settings.

Figure 5 shows the test power difference with optimized $\boldsymbol{r}$ and fixed $\boldsymbol{r}$. The legend with *rand* annotation implies we randomly initialized $\boldsymbol{r}$ as one-hot vectors and fix them while updating $\boldsymbol{g}_r$ using *GO* or active slice. Without *rand*, it means both $\boldsymbol{r}$ and $\boldsymbol{g}_r$ are optimized. We only use 3 $\boldsymbol{r}$ for active slice method and 50 for gradient-based counterpart. For active slice method with pruning (*randSKSD-g+Ex* or *randSKSD-g+KE*), despite we show that any finite random slices define a valid discrepancy, it is clear that the performance is quite poor with random initialized $\boldsymbol{r}$'s. It indicates that using active slices of $\boldsymbol{g}_r$ alone cannot compensate the poor discriminating power of the random $\boldsymbol{r}$'s. Although *SKSD-rg+GO* demonstrates an advantage compared to *randSKSD-g+GO*, the performance boost is less clear compared to active slices method. This is because we do not use any pruning for *randSKSD-g+GO*, and adopt orthogonal basis $O_r = \boldsymbol{I}$. Despite the orthogonal basis may not capture the important directions, they can provide reasonable discriminating power due to their orthogonality from each other. In summary, using good directions for $\boldsymbol{r}$ is advantageous compared to fixed $\boldsymbol{r}$.

H.3. Model learning: Training ICA

We use Adam optimizer for the model and slice directions with learning rate 0.001 and $\beta = (0.9, 0.99)$. We totally run 15000 iterations. The batch size is 100. We evaluate our method in dimension 10, 80, 100 and 150. For more stable comparisons, we initialized the weight matrix $\boldsymbol{W}$ until its conditional number is smaller than its dimensions. For active slice method, we use randomly sampled 3000 data from training set to estimate the score difference and the matrices used for eigen-decomposition.

For *SKSD-rg+GO*, we initialize the $\boldsymbol{r}$ to be a group of one-

hot vectors to form identity matrix and $\boldsymbol{g}_r = \boldsymbol{r}$. We use an adversarial training procedure that updates both $\boldsymbol{r}$ and $\boldsymbol{g}_r$ using Adam once per iteration before we update the model. For *SKSD-g+GO*, we fix the orthogonal basis O_r to be the identity matrix and only update $\boldsymbol{g}_r$. Each results are the average of 5 runs of training.

As for the reason why *SKSD-g+Ex+GO* performs worse than *+KE+GO*, we suspect that *+Ex* only focus on directions with high discriminating power. However, high discriminating power is not necessarily good for model learning. It may focus on very small area that is different from the target but ignore the larger area with small difference. Because our algorithm for finding basis is greedy, this means it can ignore the generally good directions if they are not orthogonal to the directions with high discriminating power.

From figure 7, we can observe there is a spike of *SKSD-g+Ex+GO* value at every 200 iterations due to the new active slices found at the beginning of each training epoch. However, the value drops significantly fast to the one before new active slices. This indicates the *Ex* indeed finds directions with large discriminating power but they do not represents good directions for learning due to the fast drop of SKSD values. On the other hand, the directions provided by KE does not give the highest discriminating power, but it can find generally good directions of $\boldsymbol{g}_r$ using *GO* refinement steps within a few iterations. This means the directions found by *KE* indeed represents good directions for learning as the model cannot decrease this value quickly. We guess this is due to the smooth estimation of KE, where very small areas with high discriminating power are smoothed out.

Figure 6 shows the ICA training curve of other dimensions. We can observe the convergence speed of LSD deteriorates as the dimension increases due to the poor test function in early training stage, whereas *SKSD-g+KE+GO* maintains the fastest convergence in high dimensions.

I. Perturbation of eigenvectors

The active slice method (algorithm 1) is mainly based on the eigenvalue-decomposition of matrix $\boldsymbol{H}$, where

$$\boldsymbol{H} = \int q(\boldsymbol{x}) \nabla_{\boldsymbol{x}} f_r^*(\boldsymbol{x}) \nabla_{\boldsymbol{x}} f_r^*(\boldsymbol{x})^T d\boldsymbol{x}$$

Obtaining the analytic form of $\boldsymbol{H}$ involves complicated integration, so Monte Carlo estimation is often used for approximation. We denote it as $\hat{\boldsymbol{H}}$, with M being the number of samples:

$$\hat{\boldsymbol{H}} = \frac{1}{M} \sum_{i=1}^{M} [\nabla_{\boldsymbol{x}_i} f_r^*(\boldsymbol{x}_i) \nabla_{\boldsymbol{x}_i} f_r^*(\boldsymbol{x}_i)^T] \tag{53}$$

Let $\boldsymbol{g}$ be the top eigenvector of $\boldsymbol{H}$ and $\hat{\boldsymbol{g}}$ be the top eigenvector of $\hat{\boldsymbol{H}}$. Let λ_1, λ_2 be the top two eigenvalues of $\boldsymbol{H}$. Assuming the error matrix $\boldsymbol{E} = \hat{\boldsymbol{H}} - \boldsymbol{H}$ is deterministic, (Yu et al., 2015) proved that

$$||\boldsymbol{g}\boldsymbol{g}^T(\boldsymbol{I} - \hat{\boldsymbol{g}}\hat{\boldsymbol{g}}^T)||_F \leq \frac{2||\boldsymbol{E}||_{op}}{\lambda_1 - \lambda_2} \tag{54}$$

where we define the operator norm for a given $n \times n$ matrix $\boldsymbol{A}$ as

$$||\boldsymbol{A}||_{op} = \sup\{||\boldsymbol{A}\boldsymbol{x}|| : x \in \mathbb{R}^n \text{ with } ||\boldsymbol{x}|| = 1\}$$

We also have (with proof below)

$$\min_{\epsilon \in \{-1,1\}} ||\boldsymbol{g} - \epsilon\hat{\boldsymbol{g}}||_2 \leq \sqrt{2}||\boldsymbol{g}\boldsymbol{g}^T(\boldsymbol{I} - \hat{\boldsymbol{g}}\hat{\boldsymbol{g}}^T)||_F \tag{55}$$

Inequality 54 and 55 imply that,

$$\min_{\epsilon \in \{-1,1\}} ||\boldsymbol{g} - \epsilon\hat{\boldsymbol{g}}||_2 \leq 2^{3/2} \frac{||\hat{\boldsymbol{H}} - \boldsymbol{H}||_{op}}{\lambda_1 - \lambda_2} \tag{56}$$

I.1. Proof of inequality 55

Proposition 5. *Let $\boldsymbol{S}$ and $\boldsymbol{U}$ be two matrices with orthonormal columns and equal rank r. Let $\boldsymbol{\Pi}_S$ (resp. $\boldsymbol{\Pi}_U$) indicates the projection matrix to the column space of $\boldsymbol{S}$ (resp. $\boldsymbol{U}$). Then*

$$\min_{O \in \mathbb{R}^{r \times r} \text{ orthogonal}} ||\boldsymbol{S} - \boldsymbol{U}\boldsymbol{O}||_F \leq \sqrt{2}||\boldsymbol{\Pi}_S(\boldsymbol{I} - \boldsymbol{\Pi}_U)||_F \tag{57}$$

When $r = 1$, we denote $\boldsymbol{O}$ as ϵ. Following the definition of orthogonal matrix, we have $\epsilon^T\epsilon = \epsilon^2 = 1$, hence $\epsilon \in \{-1, 1\}$. Substituting $\boldsymbol{S} = \boldsymbol{g}$ and $\boldsymbol{U} = \hat{\boldsymbol{g}}$, we get inequality 55.

Proof. Let $\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be a singular value decomposition of $\boldsymbol{S}^T\boldsymbol{U}$, and use $\boldsymbol{O} = \boldsymbol{V}\boldsymbol{W}^T$.Now,

$$\begin{aligned} ||\boldsymbol{S} - \boldsymbol{U}\boldsymbol{O}||_F^2 &= \operatorname{Tr}((\boldsymbol{S} - \boldsymbol{U}\boldsymbol{O})^T(\boldsymbol{S} - \boldsymbol{U}\boldsymbol{O})) \\ &= ||\boldsymbol{S}||_F^2 + ||\boldsymbol{U}||_F^2 - 2\operatorname{Tr}(\boldsymbol{O}\boldsymbol{S}^T\boldsymbol{U}) \\ &= 2r - 2\operatorname{Tr}(\boldsymbol{\Sigma}) \end{aligned}$$

where r is the rank of $\boldsymbol{S}$ and $\boldsymbol{U}$. On the other hand, by Pythagora's theorem

$$\begin{aligned} ||\boldsymbol{\Pi}_S(\boldsymbol{I} - \boldsymbol{\Pi}_U)||_F^2 &= ||\boldsymbol{\Pi}_S||_F^2 - ||\boldsymbol{\Pi}_S\boldsymbol{\Pi}_U||_F^2 \\ &= r - ||\boldsymbol{\Pi}_S\boldsymbol{\Pi}_U||_F^2 \\ &= r - ||\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{U}\boldsymbol{U}^T||_F^2 \\ &= r - \operatorname{Tr}(\boldsymbol{\Sigma}^2) \end{aligned}$$

We claim that the entries of $\boldsymbol{\Sigma}$ are bounded above by 1, such that $\operatorname{Tr}(\boldsymbol{\Sigma}) \leq \operatorname{Tr}(\boldsymbol{\Sigma}^2)$, then

$$\begin{aligned} \min_{O \in \mathbb{R}^{r \times r} \text{ orthogonal}} ||\boldsymbol{S} - \boldsymbol{U}\boldsymbol{O}||_F^2 &\leq 2r - 2\operatorname{Tr}(\boldsymbol{\Sigma}) \\ &\leq 2r - 2\operatorname{Tr}(\boldsymbol{\Sigma}^2) \\ &= 2||\boldsymbol{\Pi}_S(\boldsymbol{I} - \boldsymbol{\Pi}_U)||_F^2 \end{aligned}$$

Taking the square root of both sides yields the desired inequality. To prove the claim, let $\boldsymbol{\omega} = [\boldsymbol{S}, \boldsymbol{S}']$ and $\tilde{\boldsymbol{U}} = [\boldsymbol{U}, \boldsymbol{U}']$ be orthogonal matrices. Then $\boldsymbol{S}^T\boldsymbol{U}$ is a diagonal block in $\boldsymbol{\omega}^T\tilde{\boldsymbol{U}}$. It follows that $\max_i \Sigma_{i,i} = ||\boldsymbol{S}^T\boldsymbol{U}||_{op} \leq ||\boldsymbol{\omega}^T\tilde{\boldsymbol{U}}||_{op} = 1$ □

From Eq.56, we can see if the top two eigenvalues are similar, then their corresponding eigenvectors can be arbitrary different. In terms of our active slice algorithm, it means if the most discriminating directions for two distributions q, p have similar "magnitude of difference", our algorithm may fail under Monte-carlo approximation. On the other hand, if the eigenvalues are different, Eq.56 guarantees that eigenvectors from $\hat{\boldsymbol{H}}$ are not far-away from the truth.