Active Slices for Sliced Stein Discrepancy

Wenbo Gong [1] **Kaibo Zhang** [1] **Yingzhen Li** [2] **José Miguel Hernández-Lobato** [1]

Abstract

Sliced Stein discrepancy (SSD) and its kernelized variants have demonstrated promising successes in goodness-of-fit tests and model learning in high dimensions. Despite their theoretical elegance, their empirical performance depends crucially on the search of optimal slicing directions to discriminate between two distributions. Unfortunately, previous gradient-based optimisation approaches for this task return sub-optimal results: they are computationally expensive, sensitive to initialization, and they lack theoretical guarantees for convergence. We address these issues in two steps. First, we provide theoretical results stating that the requirement of using optimal slicing directions in the kernelized version of SSD can be relaxed, validating the resulting discrepancy with *finite* random slicing directions. Second, given that good slicing directions are crucial for practical performance, we propose a fast algorithm for finding such slicing directions based on ideas of active sub-space construction and spectral decomposition. Experiments on goodness-of-fit tests and model learning show that our approach achieves both improved performance and faster convergence. Especially, we demonstrate a 14-80x speed-up in goodness-of-fit tests when comparing with gradient-based alternatives.

1. Introduction

Discrepancy measures between two distributions are critical tools in modern statistical machine learning. Among them, *Stein discrepancy* (SD) and its kernelized version, kernelized Stein discrepancy (KSD), have been extensively used for *goodness-of-fit* (GOF) testing (Liu et al., 2016; Chwialkowski et al., 2016; Huggins & Mackey, 2018; Jitkrittum et al., 2017; Gorham & Mackey, 2017) and model learning (Liu & Wang, 2016; Pu et al., 2017; Hu et al., 2018; Grathwohl et al., 2020). Despite their recent success, applications of Stein discrepancies to high-dimensional distribution testing and learning remains an unsolved challenge.

These "curse of dimensionality" issues have been recently addressed by the newly proposed Sliced Stein discrepancy (SSD) and its kernelized variants SKSD (Gong et al., 2021), which have demonstrated promising results in both high dimensional GOF tests and model learning. They work by first projecting the score function and the test inputs across two slice directions $\boldsymbol{r}$ and $\boldsymbol{g}_r$ and then comparing the two distributions using the resulting one dimensional slices. The performance of SSD and SKSD crucially depends on choosing slicing directions that are highly discriminative. Indeed, Gong et al. (2021) showed that such discrepancy can still be valid despite the information loss caused by the projections, if *optimal slices* – directions along which the two distributions differ the most – are used. Unfortunately, gradient-based optimization for searching such optimal slices often suffers from slow convergence and sub-optimal solutions. In practice, many gradient updates may be required to obtain a reasonable set of slice directions (Gong et al., 2021).

We aim to tackle the above practical challenges by proposing an efficient algorithm to find *good* slice directions with statistical guarantees. Our contributions are as follows:

- We propose a computationally efficient variant of SKSD using a *finite number of random slices*. This relaxes the restrictive constraint of having to use *optimal* slices, with the consequence that convergence during optimisation to a global optimum is no longer required.
- Given that *good* slices are still preferred in practice, we propose surrogate optimization tasks to find such directions. These are called *active slices* and have analytic solutions that can be computed very efficiently.
- Experiments on GOF test benchmarks (including testing on restricted Boltzmann machines) show that our algorithm outperforms alternative gradient-based approaches while achieving at least a 14x speed-up.
- In the task of learning high dimensional *independent component analysis* (ICA) models (Comon, 1994), our algorithm converges much faster and to significantly better solutions than other baselines.

[1]Department of Engineering, University of Cambridge, Cambridge, United Kingdom [2]Department of Computing, Imperial College London, London, United Kingdom. Correspondence to: Wenbo Gong <wg242@cam.ac.uk>, José Miguel Hernández-Lobato <jmh233@cam.ac.uk>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Road map: First, we give a brief background for *SD*, *SKSD* and its relevant variants (Section 2). Next, we show that the optimality of slices are not necessary. Instead, finite random slices are enough to ensure the validity of SKSD (3.1). Despite that relaxing the optimality constraint gives us huge freedom to select slice directions, choosing an appropriate objective for finding slices is still crucial. Unfortunately, analysing SKSD in RKHS is challenging. We thus propose to analyse SSD as a surrogate objective by showing SKSD can be well approximated by SSD (Section 3.2). Lastly, by analyzing SSD, we propose algorithms to find active slices for SKSD (Sections 4, 5, 6), and demonstrate the efficacy of our proposal in the experiments (Section 7). Assumptions and proofs of theoretical results as well as the experimental settings can be found in the appendix.

2. Background

For a distributions p on $\mathcal{X} \subset \mathbb{R}^D$ with differentiable density, we define its score function as $\boldsymbol{s}_p(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$. We also define the Stein operator $\mathcal{A}_p$ for distribution p as

$$\mathcal{A}_p \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{s}_p(\boldsymbol{x})^T \boldsymbol{f}(\boldsymbol{x}) + \nabla_{\boldsymbol{x}}^T \boldsymbol{f}(\boldsymbol{x}) \,, \tag{1}$$

where $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^D$ is a test function. Then the *Stein discrepancy* (SD) (Gorham & Mackey, 2015) between two distributions p, q with differentiable densities on $\mathcal{X}$ is

$$D_{SD}(q,p) = \sup_{\boldsymbol{f} \in \mathcal{F}_q} \mathbb{E}_q[\mathcal{A}_p \boldsymbol{f}(\boldsymbol{x})] \,, \tag{2}$$

where $\mathcal{F}_q$ is the Stein's class of q that contains test functions satisfying $\mathbb{E}_q[\mathcal{A}_q \boldsymbol{f}(\boldsymbol{x})] = 0$ (also see Definition B.2 in appendix B). The supremum can be obtained by choosing $\boldsymbol{f}^* \propto \boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})$ if $\mathcal{F}_q$ is rich (Hu et al., 2018).

Chwialkowski et al. (2016); Liu et al. (2016) further restricts the test function space $\mathcal{F}_q$ to be a unit ball in an RKHS induced by a c_0-universal kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. This results in the *kernelized Stein discrepancy* (KSD), which can be computed analytically:

$$\begin{aligned} D^2(q,p) &= \left(\sup_{\boldsymbol{f} \in \mathcal{H}_k, ||f||_{\mathcal{H}_k} \leq 1} \mathbb{E}_q[\mathcal{A}_p \boldsymbol{f}(\boldsymbol{x})] \right)^2 \\ &= ||\mathbb{E}_q[\boldsymbol{s}_p(\boldsymbol{x}) k(\boldsymbol{x}, \cdot) + \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \cdot)]||^2_{\mathcal{H}_k} \,, \end{aligned} \tag{3}$$

where $\mathcal{H}_k$ is the k induced RKHS with norm $|| \cdot ||_{\mathcal{H}_k}$.

2.1. Sliced kernelized Stein discrepancy

Despite the theoretical elegance of *KSD*, it often suffers from the curse-of-dimensionality in practice. To address this issue, Gong et al. (2021) proposed a divergence family called *sliced Stein discrepancy* (SSD) and its kernelized variants, under mild assumptions on the regularity of probability densities (**Assumptions 1-4** in appendix B) and the richness of the kernel (**Assumption 5** in appendix B). The key idea is to compare the distributions on their one dimensional slices by projecting the score $\boldsymbol{s}_p$ and test input $\boldsymbol{x}$ with two directions $\boldsymbol{r}$ and its corresponding $\boldsymbol{g}_r$, respectively. Readers are referred to appendix C for details. Despite that one cannot access all the information possessed by $\boldsymbol{s}_p$ and $\boldsymbol{x}$ due to the projections, the validity of the discrepancy can be ensured by using an orthogonal basis for $\boldsymbol{r}$ along with the corresponding most discriminative $\boldsymbol{g}_r$ directions. The resulting valid discrepancy is called *maxSSD-g*, which uses a set of orthogonal basis $\boldsymbol{r} \in O_r$ and their corresponding optimal $\boldsymbol{g}_r$ directions:

$$\begin{aligned} S_{\max_{gr}}(q,p) = \sum_{\boldsymbol{r} \in O_r} \sup_{\substack{h_{rg_r} \in \mathcal{F}_q \\ \boldsymbol{g}_r \in \mathbb{S}^{D-1}}} \mathbb{E}_q[s_p^r(\boldsymbol{x}) h_{rg_r}(\boldsymbol{x}^T \boldsymbol{g}_r) + \\ \boldsymbol{r}^T \boldsymbol{g}_r \nabla_{\boldsymbol{x}^T \boldsymbol{g}_r} h_{rg_r}(\boldsymbol{x}^T \boldsymbol{g}_r)] \,, \end{aligned} \tag{4}$$

where $h_{rg_r} : \mathcal{K} \subseteq \mathbb{R} \to \mathbb{R}$ is the test function, $\mathbb{S}^{D-1}$ is the D-dimensional unit sphere and $s_p^r(\boldsymbol{x}) = \boldsymbol{s}_p(\boldsymbol{x})^T \boldsymbol{r}$ is the projected score function. Under certain scenarios (Gong et al., 2021), i.e. GOF test, one can further improve the performance of *maxSSD-g* by replacing $\sum_{\boldsymbol{r} \in O_r}$ with the optimal $\sup_{\boldsymbol{r} \in \mathbb{S}^{D-1}}$ in Eq.4, resulting in another variant called *maxSSD-rg* ($S_{\max_{rg_r}}$). This increment in performance is due to the higher discriminative power provided by the optimal $\boldsymbol{r}$.

However, the optimal test functions $h^*_{rg_r}$ in *maxSSD-g* (or *-rg*) are intractable in practice. Gong et al. (2021) further proposed kernelized variants to address this issue by letting $\mathcal{F}_q$ to be in a unit ball of an RKHS induced by a c_0-universal kernel k_{rg_r}. With

$$\begin{aligned} \xi_{p,r,g_r}(\boldsymbol{x}, \cdot) = s_p^r(\boldsymbol{x}) k_{rg_r}(\boldsymbol{x}^T \boldsymbol{g}_r, \cdot) + \\ \boldsymbol{r}^T \boldsymbol{g}_r \nabla_{\boldsymbol{x}^T \boldsymbol{g}_r} k_{rg_r}(\boldsymbol{x}^T \boldsymbol{g}_r, \cdot) \,, \end{aligned} \tag{5}$$

the *maxSKSD-g* (the kernelized version of maxSSD-g) is

$$SK_{\max_{gr}}(q,p) = \sum_{\boldsymbol{r} \in O_r} \sup_{\boldsymbol{g}_r \in \mathbb{S}^{D-1}} ||\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg_r}} \,, \tag{6}$$

where $\mathcal{H}_{rg_r}$ is the RKHS induced by k_{rg_r} with the associated norm $|| \cdot ||_{\mathcal{H}_{rg_r}}$. Similarly, a kernelized version of maxSSD-rg, denoted by *maxSKSD-rg* ($SK_{\max_{rg_r}}$), is obtained by replacing $\sum_{\boldsymbol{r} \in O_r}$ with $\sup_{\boldsymbol{r} \in \mathbb{S}^{D-1}}$ in Eq.6.

Despite that *maxSKSD-g* (or *-rg*) addresses the tractability of test functions, the practical challenge of computing them is the computation of the optimal slice directions $\boldsymbol{r}$ and $\boldsymbol{g}_r$. Gradient-based optimization (Gong et al., 2021) for such computation suffers from slow convergence; even worse, it is sensitive to initialization and returns sub-optimal solutions only. In such case, it is unclear whether the resulting discrepancy is still valid, making the correctness of GOF test unverified. Therefore, the first important question to

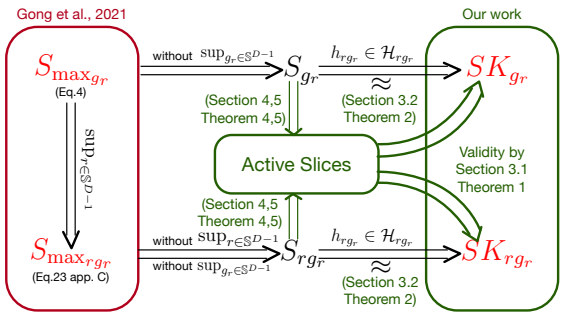

Figure 1. The relationship between different SSD discrepancies, where green texts indicate our contributions, red symbols indicate valid discrepancies and $\mathcal{H}_{rg_r}$ is the RKHS induced by kernel k_{rg_r}. The leftmost part are the discrepancies proposed by Gong et al. (2021), whereas the rightmost part + central "Active Slices" are our contributions. The arrows $\Rightarrow$ indicate the connections between Gong et al. (2021) and our work.

ask is: are the optimality of slices a necessary condition for the validity of *maxSKSD-g* (or *-rg*)? Remarkably, we show that the answer is **No** with mild assumptions on the kernel (**Assumption 5-6** in appendix B).

As the sliced Stein discrepancy defined previously involves a sup operator, making them difficult to analyze, we need to define notations for their "sub-optimal" versions. For example, *maxSSD-g* (Eq.4) involves a sup operator over slices $\boldsymbol{g}_r$. We thus define *SSD-g* (S_{g_r}) as Eq.4 with a given $\boldsymbol{g}_r$ instead of the sup:

$$S_{g_r} = \sum_{\boldsymbol{r}\in O_r} \sup_{h_{rg_r}\in\mathcal{F}_q} \mathbb{E}_q[s_p^r(\boldsymbol{x})h_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r) + \boldsymbol{r}^T\boldsymbol{g}_r\nabla_{\boldsymbol{x}^T\boldsymbol{g}_r}h_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r)] \tag{7}$$

Following similar logic, we define the "sub-optimal" version for each of the discrepancy mentioned in this section as table 1 and appendix A.

3. Relaxing constraints for the SKSD family

3.1. Is optimality necessary for validity?

As mentioned before, the discrepancy validity of max SKSD requires the optimality of slice directions, which restricts its application in practice. In the following, we show that these restrictions can be much relaxed with mild assumptions on the kernel. All proofs can be found in Appendix E.

The key idea is to use kernels such that the corresponding term SK_{rg_r} is *real analytic* w.r.t. both $\boldsymbol{r}$ and $\boldsymbol{g}_r$, which is detailed by **Assumption 6** (Appendix B). A nice property of any real analytic function is that, unless it is a constant function, otherwise the set of its roots has zero Lebesgue measure. This means the possible valid slices are almost everywhere in $\mathbb{R}^D$, giving us huge freedom to choose slices without worrying about violating validity.

Theorem 1 (Conditions for valid slices)**.** *Assuming assumptions 1-4 (density regularity), 5 (richness of kernel) and 6 (real analytic kernel) in Appendix B, let $\boldsymbol{g}_r \sim \eta_g$ for each $\boldsymbol{r} \sim \eta_r$, where η_g, η_r are distributions on $\mathbb{R}^D$ with a density, then $SK_{rg_r}(q,p) = 0$ iff. $p = q$ almost surely.*

The above theorem tells us that a *finite* number of *random* slices is enough to make SK_{rg_r} valid without the need of using optimal slices (c.f. $SK_{\max_{rg_r}}$). In practice, we often consider $\boldsymbol{r}, \boldsymbol{g}_r \in \mathbb{S}^{D-1}$ instead of $\mathbb{R}^D$. Fortunately, one can easily transform arbitrary slices to $\mathbb{S}^{D-1}$ without violating the validity. For any $\boldsymbol{r}, \boldsymbol{g}_r$, we (i) add Gaussian noises to them, and (2) re-normalize the noisy $\boldsymbol{r}, \boldsymbol{g}_r$ to unit vectors. We refer to corollary 6.1 in appendix E.1 for details.

3.2. Relationship between SSD and SKSD

Theorem 1 allows us to use random slices. However, it is still beneficial to find good ones in practice. Unfortunately, SK_{rg_r} is not a suitable objective for finding good slice directions. This is because, unlike the test function in a general function space ($h_{rg_r} \in \mathcal{F}_q$), the optimal kernel test function ($\mathbb{E}_q[\xi_{p,r,g_r}(\boldsymbol{x},\cdot)]$) can not be easily analyzed for finding good slices due to its restriction in RKHS.

Instead, we propose to use S_g (or S_{rg_r}) as the optimization objective. To justify S_{rg_r} as a good replacement for SK_{g_r}, we show that SK_{rg_r} approximates S_{rg_r} arbitrarily well if the corresponding RKHS of the chosen kernel is dense in continuous function space. Similar results for $SK_g \approx S_g$ can be derived accordingly as the only difference between S_{g_r} and S_{rg_r} is the summation over orthogonal basis O_r. However, S_{rg_r} still involves a sup operator over test functions h_{rg_r}, which hinders further analysis. To deal with this, we give an important proposition that are needed in almost every theoretical claims we made. This proposition characterises the optimal test functions for S_{rg_r} (or S_{g_r}).

Proposition 1 (Optimal test function given $\boldsymbol{r}, \boldsymbol{g}_r$)**.** *Assume assumptions 1-4 (density regularity) and given directions $\boldsymbol{r}, \boldsymbol{g}_r$. Assume an arbitrary orthogonal matrix $\boldsymbol{G}_r = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_D]^T$ where $\boldsymbol{a}_i \in \mathbb{S}^{D\times 1}$ and $\boldsymbol{a}_d = \boldsymbol{g}_r$. Denote $\boldsymbol{x} \sim q$ and $\boldsymbol{y} = \boldsymbol{G}_r\boldsymbol{x}$ which is also a random variable with the induced distribution q_{G_r}. Then, the optimal test function for S_{rg_r} is*

$$h^*_{rg_r}\left(\boldsymbol{x}^T\boldsymbol{g}_r\right) \propto \mathbb{E}_{q_{G_r}(\boldsymbol{y}_{-d}|y_d)}\left[\left(s_p^r\left(\boldsymbol{G}_r^{-1}\boldsymbol{y}\right) - s_q^r\left(\boldsymbol{G}_r^{-1}\boldsymbol{y}\right)\right)\right] \tag{8}$$

where $y_d = \boldsymbol{x}^T\boldsymbol{g}_r$ and $\boldsymbol{y}_{-d}$ contains other $\boldsymbol{y}$ elements.

Intuitively, assume $\boldsymbol{G}_r$ is a rotation matrix. Then $h^*_{rg_r}$ is the conditional expected score difference between two rotated p and q. This form is very similar to the optimal test function for SD, which is just the score difference between the origi-

Table 1. Notations for "sub-optimal" versions of SSD & SKSD.

Optimal form	*maxSSD-g* ($S_{\max_{g_r}}$)	*maxSSD-rg* ($S_{\max_{rg_r}}$)	*maxSKSD-g* ($SK_{\max_{g_r}}$)	*maxSKSD-rg* ($SK_{\max_{rg_r}}$)
Modifications	Change $\sup_{\boldsymbol{g}_r}$ to given $\boldsymbol{g}_r$ in Eq.4	Change $\sup_{\boldsymbol{r},\boldsymbol{g}_r}$ to given $\boldsymbol{r}$, $\boldsymbol{g}_r$ in Eq.37 (App. C)	Same as *maxSSD-g*. in Eq.6	Same as *maxSSD-rg* in Eq.41 (App. C)
"sub-optimum"	*SSD-g* (S_{g_r})	*SSD-rg* (S_{rg_r})	*SKSD-g* (SK_{g_r})	*SKSD-rg* (SK_{rg_r})

nal p, q. Knowing the optimal form of $h^*_{rg_r}$, we can show SK_{rg_r} can be well approximated by S_{rg_r}.

Theorem 2 ($SK_{rg_r} \approx S_{rg_r}$)**.** *Assume assumptions 1-4 (density regularity) and 5 (richness of kernel). Given $\boldsymbol{r}$ and $\boldsymbol{g}_r$, $\forall \epsilon > 0$ there exists a constant C such that*

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

As S_{rg_r} approximates SK_{rg_r} arbitrary well, the hope is that good slices for S_{rg_r} also correspond to good slices for SK_{rg_r} in practice. Therefore in the next section we focus on analyzing S_{rg_r} instead to propose an efficient algorithm for finding good slices.

4. Active slice direction g

Finding good slices involves alternating maximization of $\boldsymbol{r}$ and $\boldsymbol{g}_r$. To simplify the analysis, we focus on good directions $\boldsymbol{g}_r$ given fixed $\boldsymbol{r}$, e.g. the orthogonal basis $\boldsymbol{r} \in O_r$ for now. Finding good $\boldsymbol{g}_r$ is achieved in two steps: (i) Rewriting the problem of the maximizing S_{g_r} w.r.t $\boldsymbol{g}_r$ into an equivalent minimization problem, called *controlled approximation*; (ii) Establish an upper-bound of the controlled approximation objective such that its minimizer is analytic. This derivation is based on an important inequality: Poincaré inequality, which upper bounds the variances of a function by its gradient magnitude. Therefore, we need **Assumptions 7-8** (Appendix B) to make sure this inequality is valid. We name the resulting $\boldsymbol{g}_r$ that minimizes the upper bound as *active slices*. All proofs can be found in appendix F.

4.1. Controlled Approximation

To start with, we need an upper bound for S_{g_r} so that we can transform the maximization of S_{g_r} into the minimization of their gap. Hence, we propose a generalization of SD (Eq.2) called *projected Stein discrepancy* (PSD):

$$\text{PSD}(q, p; O_r) = \sum_{\boldsymbol{r}\in O_r} \sup_{f_r \in \mathcal{F}_q} \mathbb{E}_q[s_p^r(\boldsymbol{x}) f_r(\boldsymbol{x}) + \boldsymbol{r}^T \nabla_{\boldsymbol{x}} f_r(\boldsymbol{x})] \tag{9}$$

where $f_r : \mathcal{X} \subseteq \mathbb{R}^D \to \mathbb{R}$. SD is a special case of PSD by setting O_r as identity matrix $\boldsymbol{I}$. In proposition 4 of appendix F.1, we show that if $\mathcal{F}_q$ contains all bounded continuous functions, then the optimal test function in PSD is

$$f_r^*(\boldsymbol{x}) \propto \left(s_p^r(\boldsymbol{x}) - s_q^r(\boldsymbol{x})\right)\,. \tag{10}$$

It can also be shown that PSD is equivalent to the *Fisher divergence*, which has been extensively used in training energy based models (Song et al., 2020; Song & Ermon, 2019) and fitting kernel exponential families (Sriperumbudur et al., 2017; Sutherland et al., 2018; Wenliang et al., 2019).

We now prove that PSD upper-bounds S_{g_r}, with the gap as the expected square error between their optimal test functions f_r^* and $h^*_{rg_r}$ (Proposition 1). Since PSD is constant w.r.t. $\boldsymbol{g}_r$, maximization of S_{g_r} is equivalent to a minimization task, called *controlled approximation*.

Theorem 3 (Controlled Approximation)**.** *Assume assumptions 1-4 (density regularity), and the coefficient for the optimal test functions to be* 1 *w.l.o.g., then PSD $\geq S_{g_r}$ and*

$$\textit{PSD} - S_{g_r} = \sum_{\boldsymbol{r}\in O_r} \mathbb{E}_q[(f_r^*(\boldsymbol{x}) - h^*_{rg_r}(\boldsymbol{x}^T\boldsymbol{g}_r))^2], \tag{11}$$

with f_r^* and $h^*_{rg_r}$ are optimal test functions for PSD and S_{g_r} defined in Eq.10 and Eq.8 respectively.

Intuitively, minimizing the above gap can be regarded as a function approximation problem, where we want to approximate a multivariate function $f_r^* : \mathbb{R}^D \to \mathbb{R}$ by a univariate function $h^*_{rg_r} : \mathbb{R} \to \mathbb{R}$ with optimal parameters $\boldsymbol{g}_r$.

4.2. Upper-bounding the error

Solving the controlled approximation task directly may be difficult in practice. Instead, we propose an upper-bound of the approximation error, such that this upper-bound's minimizer $\boldsymbol{g}_r$ is analytic. The inspiration comes from the *active subspace method* for dimensionality reduction (Constantine et al., 2014; Zahm et al., 2020), therefore we name the corresponding minimizers as *active slices*.

Theorem 4 (Error upper-bound and active slices $\boldsymbol{g}_r$)**.** *Assume assumptions 2, 4 (density regularity) and 7-8 (Poincaré inequality conditions), we can upper bound the inner part of the controlled approximation error (Eq.11) by*

$$\mathbb{E}_q\left[\left(f_r^*(\boldsymbol{x}) - h^*_{rg_r}\left(\boldsymbol{x}^T\boldsymbol{g}_r\right)\right)^2\right] \leq C_{sup}\,\text{tr}\left(\boldsymbol{G}_{r\backslash d}\boldsymbol{H}_r\boldsymbol{G}_{r\backslash d}^T\right), \tag{12}$$

$$\boldsymbol{H}_r = \int q(\boldsymbol{x})\nabla_x f_r^*(\boldsymbol{x})\nabla_{\boldsymbol{x}} f_r^*(\boldsymbol{x})^T d\boldsymbol{x}. \tag{13}$$

Here C_{sup} is the Poincaré constant defined in assumption 8 and $\boldsymbol{G}_{r\backslash d} \in \mathbb{R}^{(D-1)\times D}$ is an arbitrary orthogonal matrix

$\boldsymbol{G}_r$ excluding the d^{th} row $\boldsymbol{g}_r$. The orthogonal matrix has the form $\boldsymbol{G}_r = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_D]^T$ where $\boldsymbol{a}_i \in \mathbb{S}^{D-1}$ and $\boldsymbol{a}_d = \boldsymbol{g}_r$.

The above upper-bound is minimized when the row space of $\boldsymbol{G}_{r\setminus d}$ is the span of the first $D-1$ eigenvectors of $\boldsymbol{H}_r$ (arranging eigenvalues in ascending order). One possible choice for active slice $\boldsymbol{g}_r$ is $\boldsymbol{v}_D$, where $(\lambda_i, \boldsymbol{v}_i)$ is the eigenpair of $\boldsymbol{H}_r$ and $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_D$.

Intuitively, the active slices $\boldsymbol{g}_r = \boldsymbol{v}_D$ are the directions where the test function f_r^* varies the most. Indeed, we have $\boldsymbol{v}_D^T \boldsymbol{H}_r \boldsymbol{v}_D = \mathbb{E}_q[||\nabla_{\boldsymbol{x}} f_r^*(\boldsymbol{x})^T \boldsymbol{v}_D||^2] = \lambda_D$, where the eigenvalue λ_D measures the averaged gradient variation in the direction defined by $\boldsymbol{v}_D$.

5. Active slice direction $\boldsymbol{r}$

The dependence of active slice $\boldsymbol{g}_r$ on $\boldsymbol{r}$ motivate us to consider the possible choices of $\boldsymbol{r}$. Although finite random slices $\boldsymbol{r}$ are sufficient for obtaining a valid discrepancy, in practice using sub-optimal $\boldsymbol{r}$ can result in weak discriminative power and poor active slices $\boldsymbol{g}_r$. We address this issue by proposing an efficient algorithm to search for good $\boldsymbol{r}$. Again all the proofs can be found in appendix G.

5.1. PSD Maximization for searching $\boldsymbol{r}$

Directly optimizing S_{rg_r} w.r.t. $\boldsymbol{r}$ is particularly difficult due to the alternated updates of $\boldsymbol{r}$ and $\boldsymbol{g}_r$. To simplify the analysis, we start from the task of finding a single direction $\boldsymbol{r}$. Our key idea to sidestep such alternation is based on the intuition that S_{rg_r} with active slices $\boldsymbol{g}_r$ should well approximate PSD_r (PSD with given $\boldsymbol{r}$) from theorem 4. The independence of PSD_r to $\boldsymbol{g}_r$ allows us to avoid the alternated update and the accurate approximation validates the direct usage of the resulting active slices in S_{rg_r}. Indeed, we will prove that maximizing PSD_r is equivalent to maximizing a lower-bound for S_{rg_r}.

Assume we have two slices $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$, with given $\boldsymbol{g}_{r_1}, \boldsymbol{g}_{r_2}$. Then finding good $\boldsymbol{r}_1$ is equivalent to maximizing the difference $S_{r_1,g_{r_1}} - S_{r_2,g_{r_2}}$. The following proposition establishes a lower-bound for this difference.

Proposition 2 (Lower-bound for the S_{rg_r} gap)**.** *Assume the conditions in theorem 4 are satisfied, then for any slices $\boldsymbol{r}_1$, $\boldsymbol{r}_2$ and $\boldsymbol{g}_{r_1}$, $\boldsymbol{g}_{r_2}$, we have*

$$S_{r_1,g_{r_1}} - S_{r_2,g_{r_2}} \geq \mathit{PSD}_{r_1} - \mathit{PSD}_{r_2} - C_{\mathit{sup}}\Omega, \qquad (14)$$

where C_{sup} is the Poincaré constant defined in assumption 8 and $\Omega = \sum_{i=1}^{D} \omega_i$ where $\{\omega_i\}_i^D$ is the eigenvalue of $\mathbb{E}_q[\nabla_{\boldsymbol{x}} \boldsymbol{f}^*(\boldsymbol{x}) \nabla_{\boldsymbol{x}} \boldsymbol{f}^*(\boldsymbol{x})^T]$, $\boldsymbol{f}^*(\boldsymbol{x}) = \boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})$.

Proposition 2 justifies the maximization of PSD_{r_1} w.r.t. $\boldsymbol{r}_1$ as a valid surrogate. But more importantly, this alternative objective admits an analytic maximizer of $\boldsymbol{r}$, which is then used as the active slice direction:

Theorem 5 (Active slice $\boldsymbol{r}$)**.** *Assuming assumptions 1-4 (density regularity), then the maximum of the PSD_r is achieved at $\boldsymbol{r}^* = \boldsymbol{v}_{max}$:*

$$\max_{\boldsymbol{r} \in \mathbb{S}^{D-1}} \mathbb{E}_q \left[s_p^r(\boldsymbol{x}) f_r^*(\boldsymbol{x}) + \boldsymbol{r}^T \nabla_{\boldsymbol{x}} f_r^*(\boldsymbol{x}) \right] = \lambda_{max}.$$

Here $(\lambda_{max}, \boldsymbol{v}_{max})$ is the largest eigenpair of the matrix $\boldsymbol{S} = \mathbb{E}_q \left[\boldsymbol{f}^*(\boldsymbol{x}) \boldsymbol{f}^*(\boldsymbol{x})^T \right]$

5.2. Constructing the orthogonal basis O_r

Under certain scenarios, e.g. model learning, we want to train the model to perform well in every directions instead of a particular one. Thus, using a good orthogonal basis is preferred over a single active slice $\boldsymbol{r}$. Here gradient-based optimization is less suited as it breaks the orthogonality constraint. Also proposition 2 is less useful here as well, as PSD is invariant to the choice of O_r, i.e. $\text{PSD}(q, p; O_{r_1}) = \text{PSD}(q, p; O_{r_2})$ and $O_{r_1} \neq O_{r_2}$.

Inspired by the analysis of single active $\boldsymbol{r}$, we propose to use the eigendecomposition of $\boldsymbol{S}$ to obtain a good orthogonal basis O_r. Theoretically, this operation also corresponds to a greedy algorithm, where in step i it searches for the optimal direction $\boldsymbol{r}_i$ that is orthogonal to $\{\boldsymbol{r}_{<i}\}$ and maximizes $\text{PSD}_{\boldsymbol{r}_i}$ (see Corollary 6.2 in appendix G.3). Although there is no guarantee for finding the *optimal* O_r due to its myopic behavior, in practice this greedy algorithm at least finds some good directions with high discriminative power (eigenvectors with large eigenvalues).

6. Practical algorithm

The proposed active slice method is summarized in Algorithm 1, which requires the intractable score difference $\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})$. Two types of approximations can be used. The first approach applies *gradient estimators* (GE) to estimate $\boldsymbol{s}_q(\boldsymbol{x})$ from $\boldsymbol{x}$ samples. We use the Stein gradient estimator (Li & Turner, 2017) for the GE approach, although other estimators (Sriperumbudur et al., 2017; Sutherland et al., 2018; Shi et al., 2018; Zhou et al., 2020) can also be employed. The second method directly estimates the score difference using a *kernel-smoothed estimator* (KE):

$$\begin{aligned} \boldsymbol{s}_p(\boldsymbol{y}) - \boldsymbol{s}_q(\boldsymbol{y}) \approx & \mathbb{E}_{\boldsymbol{x}\sim q}[(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x}))k(\boldsymbol{x}, \boldsymbol{y})] \\ = & \mathbb{E}_{\boldsymbol{x}\sim q}[\boldsymbol{s}_p(\boldsymbol{x})k(\boldsymbol{x}, \boldsymbol{y}) + \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{y})], \end{aligned} \qquad (15)$$

where the second expression comes from integration by part, and it can be computed in practice. Figure 1 summarizes the relationships between different SSD discrepancies and highlights our contributions. For GOF test specifically, we also derive the asymptotic distribution and propose an practical GOF algorithm in appendix D.

Algorithm 1 Active slice algorithm

Input: Samples $\boldsymbol{x} \sim q$, density p, kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, Gaussian noise γ, pruning factor m (optional)
Result: $\widetilde{O_r}$, $\boldsymbol{G}$
Estimate $\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})$ using KE or GE with kernel k and samples $\boldsymbol{x}$.
if Pruning **then**
  Top m eigenvectors of $\boldsymbol{S}$ to form $\widetilde{O_r}$ (Theorem 5)
else
  Getting all eigenvectors of $\boldsymbol{S}$ to form $\widetilde{O_r}$
end if
Add noise γ to $\widetilde{O_r}$, then normalize. (Section 3.1)
for $\boldsymbol{r} \in \widetilde{O_r}$ **do**
  $\boldsymbol{g}_r$ is the top 1 eigenvector of $\boldsymbol{H}_r$ (Theorem 4)
  Add noise γ to $\boldsymbol{g}_r$ then normalize (Section 3.1)
  Concatenate $\boldsymbol{g}_r$ to $\boldsymbol{G}$
end for
Further optimize $\widetilde{O_r}$, $\boldsymbol{G}$ with *SKSD-g* (SK_{g_r}) using gradient-based optimization (Optional)
Return: $\widetilde{O_r}$, $\boldsymbol{G}$

6.1. Computational cost

The overall complexity includes the cost for (1) finding active slices (algorithm 1) (2) applying the downstream test. For finding the active slices $\boldsymbol{r}$, one important fact is that we only need the m ($m \ll D$) most important $\boldsymbol{r}$ (importance characterised by eigenvalues). Luckily, fast eigenvalue-decomposition algorithm, e.g. randomized SVD from Saibaba et al. (2021), requires $O(m)$ matrix-vector product. For $\boldsymbol{g}$, from algorithm 1, we only need to solve m eigenvalue-decomposition, each only cares about the most important eigenvector. Therefore, $O(m \times 1)$ matrix-vector product are needed. So the overall complexity for finding slices is $O(mD^2)$, where D^2 comes from matrix-vector product. For gradient-based optimization (GO), the complexity is $O(l(D^2 + C_{\text{grad}}))$ (l is optimization step and C_{grad} is the back-prop cost, D^2 coms from evaluating SK_{g_r} or SK_{rg_r}). Our algorithm in general has lower training cost as $l \gg m$ and C_{grad} can be expensive. For (2), our method has $O(mD)$ cost compared to $O(D^2)$ for GO. As $m \ll D$, active slices have less complexity compared to pure GO based method proposed in Gong et al. (2021). For memory cost, our method costs $O(mD)$ to store $\boldsymbol{r}, \boldsymbol{g}$ whereas GO uses $O(D^2)$. Overall, our method requires nearly an order of magnitude less complexity in terms of computation and memory consumption.

7. Experiments

GOF test aims to test the fitness of the model to the target data. The test procedure roughly proceeds as: (1) Define null hypothesis (model matches the data distribution) and alternative hypothesis (model does not match the data distribution); (2) Compute test statistic (e.g. KSD) and threshold (e.g. bootstrap method); (3) Reject null hypothesis (statistic $>$ threshold) or not (statistic $\leq$ threshold). Refer to appendix D for more details.

7.1. Benchmark GOF tests

We demonstrate the improved test power results (in terms of null rejection rates) and significant speed-ups of the proposed active slice algorithm on 3 benchmark tasks, which have been extensively used for measuring GOF test performances (Jitkrittum et al., 2017; Huggins & Mackey, 2018; Chwialkowski et al., 2016; Gong et al., 2021). Here the test statistic is based on *SKSD-g* (SK_{g_r}) with fixed basis $O_r = \boldsymbol{I}$. Two practical approaches are considered for computing the active slice $\boldsymbol{g}_r$: (i) gradient estimation with the Stein gradient estimator (*SKSD-g+GE*), and (ii) gradient estimation with the kernel-smoothed estimator (*KE*), plus further gradient-based optimization (*SKSD-g+KE+GO*). For reference, we include a version of the algorithm with exact score difference (*SKSD-g+Ex*) as an ablation for the gradient estimation approaches.

In comparison, we include the following strong baselines: *KSD* with RBF kernel (Liu et al., 2016; Chwialkowski et al., 2016), maximum mean discrepancy (MMD, Gretton et al., 2012) with RBF kernel, random feature Stein discrepancy with L1 IMQ kernel (L1-IMQ, Huggins & Mackey, 2018), and the current state-of-the-art — *maxSKSD-g* with random initialized $\boldsymbol{g}_r$ followed by gradient optimization (SKSD-g+GO, Gong et al., 2021). For all methods requiring *GO* or active slices, we split the 1000 test samples from q into 800 test and 200 training data, where we run GO or active slice method on the training set.

The 3 GOF test benchmarks, with details in appendix H.1, are: (1) **Laplace**: $p(\boldsymbol{x}) = \mathcal{N}(0, \boldsymbol{I})$, $q(\boldsymbol{x}) = \prod_{d=1}^{D} \text{Lap}(x_d|0, 1/\sqrt{2})$; (2) **Multivariate-t**: $p(\boldsymbol{x}) = \mathcal{N}(0, \frac{5}{3}\boldsymbol{I})$, $q(\boldsymbol{x})$ is a fully factorized multivariate-t with 5 degrees of freedom, 0 mean and scale 1; (3) **Diffusion**: $p(\boldsymbol{x}) = \mathcal{N}(0, \boldsymbol{I})$, $q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_1)$ where in $q(\boldsymbol{x})$ the variance of 1^{st}-dim is 0.3 and the rest is $\boldsymbol{I}$.

The upper panels in Figure 2 show the test power results as the dimensions D increase. As expected, *KSD* and *MMD* with RBF kernel suffer from the curse-of-dimensionality. *L1-IMQ* performs relatively well in **Laplace** and **multivariate-t** but still fails in **diffusion**. For SKSD based approaches, *SKSD-g+GO* with 1000 training epochs still exhibits a decreasing test power in **Laplace** and **multivariate-t**. On the other hand, *SKSD-g+KE+GO* with 50 training epochs has nearly optimal performance. *SKSD-g+Ex* and *SKSD-g+GE* achieve the true optimal rejection rate without any *GO*. Specifically, Table 2 shows that the active slice method achieves significant computational savings

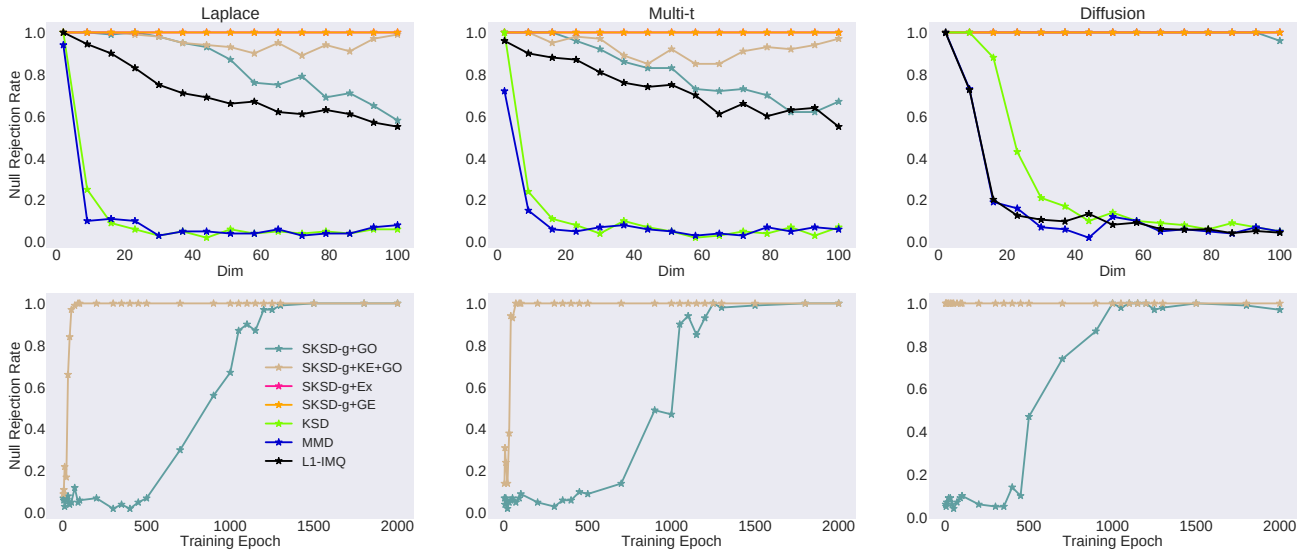

Figure 2. (**Upper panel**): The null rejection rate w.r.t. different dimensional benchmark problems. *SKSD-g+Ex* and *SKSD-g+GE* coincide at the optimal rejection rate (**Lower panel**): Null rejection rate with different number of gradient optimization epochs.

with **14x**-**80x** speed-up over *SKSD-g+GO*.

For approaches that require gradient optimization, the lower panels in Figure 2 show the test power as the number of training epochs increases. *SKSD-g+GO* with random slice initialization requires a huge number of gradient updates to obtain reasonable test power, and 1000 epochs achieves the best balance between run-time and performance. On the other hand, *SKSD-g+KE+GO* with active slice achieves significant speed-ups with near-optimal test power using around **50** epochs on **Laplace** and **Multivariate-t**. Remarkably, on **Diffusion** test, $\boldsymbol{g}_r$ initialized by the active slices achieves near-optimal results already, so that the later gradient refinements are not required.

7.2. RBM GOF test

Following Gong et al. (2021), we conduct a more complex GOF test using restrict Boltzman machines (RBMs, (Hinton & Salakhutdinov, 2006; Welling et al.)). Here the p distribution is an RBM: $p(\boldsymbol{x}) = \frac{1}{Z}\exp\left(\mathbf{x}^\top \mathbf{B}\mathbf{h} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{x} - \frac{1}{2}\|\mathbf{x}\|^2\right)$, where $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{h} \in \{\pm 1\}^{d_h}$ denotes the hidden variables. The q distribution is also an RBM with the same $\mathbf{b}, \mathbf{c}$ parameters as p but a different $\boldsymbol{B}$ matrix perturbed by different levels of Gaussian noise. We use $D = 50$ and $d_h = 40$, and block Gibbs sampler with 2000 burn-in steps. The test statistics for all the approaches are computed on a test set containing 1000 samples from q.

The test statistic is constructed using *SKSD-rg* (SK_{rg_r}) with $\boldsymbol{r}$, $\boldsymbol{g}_r$ obtained either by gradient-based optimization (*SKSD-rg+GO*) or the active slice algorithms (*+KE*, *+GE* and *+Ex*) without the gradient refinements. Specifically, *SKSD-rg+GO* runs 50 training epochs with $\boldsymbol{r}$ and $\boldsymbol{g}_r$ initialized to $\boldsymbol{I}$. For the active slice methods, we also prune away most slices and only keep the top-3 most important $\boldsymbol{r}$ slices.

The left panel of Figure 3 shows that *SKSD-rg+KE* achieves the best null rejection rates among all baselines, except for *SKSD-rg+Ex* whose performance is expected to upper-bound all other active slice methods. This shows the potential of our approach with an accurate score difference estimator. Although *SKSD-rg+GO* performs reasonably well, its run-time is **53x** longer than *SKSD-rg+KE* as shown in Table 4. Interestingly, *SKSD-rg+GE* performs worse than KSD due to the significant under-estimation of the magnitude of $\boldsymbol{s}_q(\boldsymbol{x})$. Therefore, we omit this approach in the following ablation studies.

Ablation studies The first ablation study, with results shown in the middle panel in Figure 3, considers pruning the active slices at different pruning levels, where the horizontal axis indicates the number of $\boldsymbol{r}$ slices used to construct the test statistic. We observe that the null rejection rates of active slice methods peak with pruning level 3, indicating their ability to select the most important directions. Their performances decrease when more $\boldsymbol{r}$ are considered since, in practice, those less important directions introduce extra noise to the test statistic. On the other hand, *SKSD-rg+GO* shows no pruning abilities due to its sensitivity to slice initialization. Remarkably, the final performance of *SKSD-rg+GO* without pruning is still worse than *SKSD-rg+KE* with pruning, showing the importance of finding 'good' instead of many 'average-quality' directions. Another advantage of

Table 2. Test power for 100 dimensional benchmarks and time consumption. The run-time for SKSD-g+KE+GO include both the active slice computation and the later gradient-based refinement steps. NRR stands for null rejection rate.

	Laplace			Multi-t			Diffusion		
Method	NRR	sec/trial	Speed-up	NRR	sec/trial	Speed-up	NRR	sec/trial	Speed-up
SKSD-g+Ex	**1**	**0.38**	**103x**	**1**	**0.49**	**90x**	**1**	**0.34**	**102x**
SKSD-g+GO	0.58	39.39	1x	0.67	44.24	1x	0.96	34.73	1x
SKSD-g+KE+GO	**0.99**	2.72	14x	**0.97**	2.38	19x	**1**	**0.43**	**81x**
SKSD-g+GE	**1**	**0.66**	**60x**	**1**	**0.67**	**66x**	**1**	**0.78**	**44x**

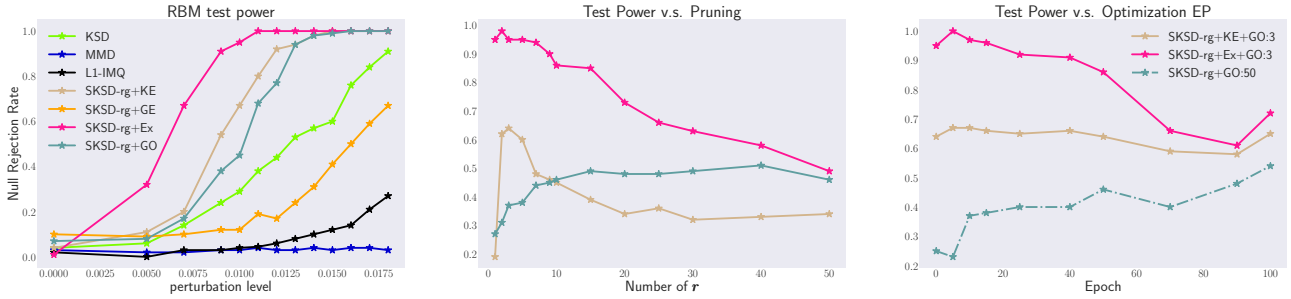

Figure 3. (**Left**): The GOF test power of each method with different level of noise perturbations (**Mid**): The effect of different pruning level towards the test power (**Right**): The effect of gradient based optimization epoch to the test power. 3 and 50 indicates the pruning level. *KE+GO* or *Ex+GO* means active slices with further gradient refinement steps.

pruning is to reduce the computational and memory costs from $O(MD)$ to $O(mD)$, where m and M are the number of pruned $\boldsymbol{r}$ and slice initializations, respectively ($m \ll M$).

The second ablation study investigates the quality of the obtained slices either by gradient-based optimization or by the active slice approaches. Results are shown in the right panel of Figure 3, where the horizontal axis indicates the number of training epochs, and the numbers annotated in the legend (3 and 50) indicate the pruning. We observe that the null rejection rate of *SKSD-rg+KE+GO* starts to improve only after 100 epochs, meaning that short run of *GO* refinements are redundant due to the good quality of active slices. The performance decrease of *SKSD-rg+Ex+GO* is due to the over-fitting of GO to the training set. The null rejection rate of *SKSD-rg+GO* gradually increases with larger training epochs as expected. However, even after 100 epochs, the test power is still lower than active slices without any GO.

In appendix H.2, another ablation study also shows the advantages of good $\boldsymbol{r}$ compared to using random slices.

7.3. Model learning: ICA

We evaluate the performance of the active slice methods in model learning by training an independent component analysis (ICA) model, which has been extensively used to evaluate algorithms for training energy-based models (Gutmann & Hyvärinen, 2010; Hyvärinen & Dayan, 2005; Ceylan & Gutmann, 2018; Grathwohl et al., 2020). ICA follows a simple generative process: it first samples a D-dimensional random variable $\boldsymbol{z}$ from a non-Gaussian p_z (we use multivariate-t), then transforms $\boldsymbol{z}$ to $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z}$ with a non-singular matrix $\boldsymbol{W} \in \mathbb{R}^{D\times D}$. The log-likelihood is $\log p(\boldsymbol{x}) = \log p_z(\boldsymbol{W}^{-1}\boldsymbol{x}) + C$ where C can be ignored if trained by minimizing Stein discrepancies. We follow Grathwohl et al. (2020); Gong et al. (2021) to sample 20000 training and 5000 test datapoints from a randomly initialized ICA model. The baselines considered include *KSD*, *SKSD-g+GO*, *SKSD-rg+GO* and the state-of-the-art *learned Stein discrepancy* (*LSD*) (Grathwohl et al., 2020), where the test function is parametrized by a neural network. For active slice approaches, one optimization epoch include the following two steps: (i) finding active slices for both orthogonal basis O_r and $\boldsymbol{g}_r$ at the beginning of the epoch, and (ii) refining the $\boldsymbol{g}_r$ directions and the $\boldsymbol{W}$ parameters in an adversarial manner with O_r fixed. For *SKSD-g+GO*, we fix basis $O_r = \boldsymbol{I}$ and only update $\boldsymbol{g}_r$ with *GO*. We refer to appendix H.3 for details on the setup and training procedure. We see from Figure 4 that *SKSD-g+KE+GO* converges significantly faster at 150 dimensions than all baselines; moreover, it has much better NLL (Table 3). We argue this performance gain is due to the use of the better orthogonal basis O_r found by the greedy algorithm, showing the advantages of better O_r in model learning. On the other hand, the importance of orthogonality in O_r is indicated by the poor performance of *SKSD-rg+GO*, as gradient updates for $\boldsymbol{r}$ violate the orthogonality constraint. The goal of learning is to train the model to match the data distribution along every slicing direction, and the orthogonality constraint can help prevent the model from ignoring important slicing directions.

Table 3. The test NLL of different dimensional ICA model

Dimensions	SKSD-g+KE+GO	SKSD-g+Ex+GO	SKSD-g+GO	SKSD-rg+GO	LSD	KSD
10	7.93±0.31	7.95±0.31	8.06±0.33	10.03±0.61	**7.42±0.31**	**7.82±0.31**
80	**7.88±0.77**	15.17±0.97	19.03±1.06	62.53±0.92	**6.26±1.49**	80.75±1.22
100	**6.93±1.36**	21.50±1.41	22.22±1.08	75.28±1.63	17.55±1.60	110.78±1.19
150	**11.67±2.46**	27.37±3.04	21.63±3.27	107.25±1.93	32.15±3.75	180.47±1.91

Table 4. Test power and time consumption at 0.01 perturbation

	Test Power	Opt. Time	Speed-up
SKSD-rg+Ex	0.95	0.04s	254x
SKSD-rg+KE	**0.67**	**0.19s**	**53x**
SKSD-rg+GO	0.45	10.15s	1x

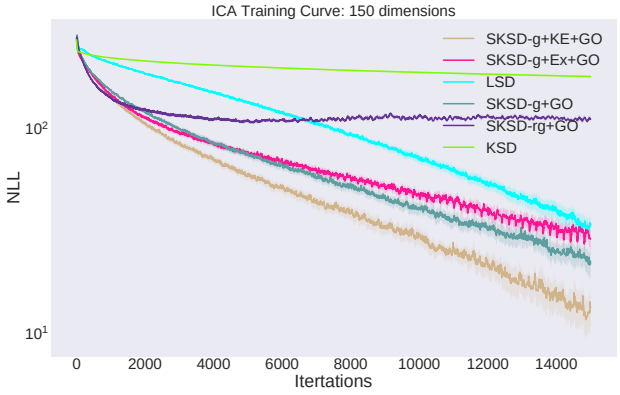

Figure 4. Training Curve of ICA model, where y-axes indicates the test NLL.

Interestingly, *SKSD-g+Ex+GO* performs worse than *+KE+GO*. We hypothesize that this is because the *+Ex+GO* approach often focuses on directions with large discriminative power but with less useful learning signal (see appendix H.3). *LSD* performs well in low dimensional problems. However, in high dimensional learning tasks it spends too much time on finding good test functions, which slows down the convergence significantly.

8. Related Work

Active subspace method (ASM): ASM is initially proposed as a dimensionality reduction method, which constructs a subspace with low-rank projectors (Constantine et al., 2014) according to the subspace Poincaré inequality. Zahm et al. (2020) showed promising results on the application of ASM to approximating multivariate functions with lower dimensional ones. However, they only considered the subspace Poincaré inequality under Gaussian measures, and a generalization to a broader family of famlity is proposed by Parente et al. (2020). Another closely related approach uses logarithmic Sobolev inequality instead to construct the active subspace (Zahm et al., 2018), which can be interpreted as finding the optimal subspace to minimize a KL-divergence. It has shown successes in Bayesian inverse problems and particle inference (Chen et al., 2019). However, as the ASM method is based on the eigen-decomposition of the sensitivity matrix, there is a potential limitation when the sensitivity matrix is estimated by Monte-Carlo method. We prove this limitation in appendix I.

Sliced discrepancies: Existing examples of sliced discrepancies can be roughly divided into two groups. Most of them belong to the first group, and they use the slicing idea to improve computational efficiency. For example, sliced Wasserstein distance projects distributions onto one dimensional slices so that the corresponding distance has an analytic form (Kolouri et al., 2019; Deshpande et al., 2019). Sliced score matching uses Hutchinson's trick to avoid the expensive computation of the Hessian matrix (Song et al., 2020). The second group focuses on the curse-of-dimensionality issue which remains to be addressed. To the best of our knowledge, existing integral probability metrics in this category include *SSD* (Gong et al., 2021) and *kernelized complete conditional Stein discrepancy* (KCC-SD, Singhal et al., 2019). The former is more general and requires less restrictive assumptions, while the latter requires samples from complete conditional distributions. Recent work has also investigated the statistical properties of sliced discrepancies (Nadjahi et al., 2020).

9. Conclusion

We have proposed the active slice method as a practical solution for searching good slices for *SKSD*. We first prove that the validity of the kernelized discrepancy only requires finite number of random slices instead of optimal ones, giving us huge freedom to select slice directions. Then by analyzing the approximation quality of *SSD* to *SKSD*, we proposed to find active slices by optimizing surrogate optimization tasks. Experiments on high-dimensional GOF tests and ICA training showed the active slice method performed the best across a number of competitive baselines in terms of both test performance and run-time. Future research directions include better score difference estimation methods, non-linear generalizations of slice projections, and the application of the active slice method to other discrepancies.

References

Arcones, M. A. and Gine, E. On the bootstrap of u and v statistics. *The Annals of Statistics*, pp. 655–674, 1992.

Ceylan, C. and Gutmann, M. U. Conditional noise-contrastive estimation of unnormalised models. In *International Conference on Machine Learning*, pp. 726–734. PMLR, 2018.

Chen, P., Wu, K., Chen, J., O'Leary-Roseberry, T., and Ghattas, O. Projected stein variational newton: A fast and scalable bayesian inference method in high dimensions. *arXiv preprint arXiv:1901.08659*, 2019.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. JMLR: Workshop and Conference Proceedings, 2016.

Comon, P. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Constantine, P. G., Dow, E., and Wang, Q. Active subspace methods in theory and practice: applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4): A1500–A1524, 2014.

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. G. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019.

Gong, W., Li, Y., and Hernández-Lobato, J. M. Sliced kernelized stein discrepancy. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=t0TaKv0Gx6Z.

Gorham, J. and Mackey, L. Measuring sample quality with stein's method. In *Advances in Neural Information Processing Systems*, pp. 226–234, 2015.

Gorham, J. and Mackey, L. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pp. 1292–1301. PMLR, 2017.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., and Zemel, R. Cutting out the middle-man: Training and evaluating energy-based models without sampling. *arXiv preprint arXiv:2002.05616*, 2020.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science*, 313 (5786):504–507, 2006.

Hoeffding, W. A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics*, pp. 308–334. Springer, 1992.

Hu, T., Chen, Z., Sun, H., Bai, J., Ye, M., and Cheng, G. Stein neural sampler. *arXiv preprint arXiv:1810.03545*, 2018.

Huggins, J. and Mackey, L. Random feature stein discrepancies. In *Advances in Neural Information Processing Systems*, pp. 1899–1909, 2018.

Huskova, M. and Janssen, P. Consistency of the generalized bootstrap for degenerate u-statistics. *The Annals of Statistics*, pp. 1811–1823, 1993.

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pp. 262–271, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. K. Generalized sliced wasserstein distances. *arXiv preprint arXiv:1902.00434*, 2019.

Li, Y. and Turner, R. E. Gradient estimators for implicit models. *arXiv preprint arXiv:1705.07107*, 2017.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. *arXiv preprint arXiv:1608.04471*, 2016.

Liu, Q., Lee, J., and Jordan, M. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pp. 276–284, 2016.

Mityagin, B. The zero set of a real analytic function. *arXiv preprint arXiv:1512.07276*, 2015.

Nadjahi, K., Durmus, A., Chizat, L., Kolouri, S., Shahrampour, S., and Şimşekli, U. Statistical and topological properties of sliced probability divergences. *arXiv preprint arXiv:2003.05783*, 2020.

Parente, M. T., Wallin, J., Wohlmuth, B., et al. Generalized bounds for active subspaces. *Electronic Journal of Statistics*, 14(1):917–943, 2020.

Pu, Y., Gan, Z., Henao, R., Li, C., Han, S., and Carin, L. Vae learning via stein variational gradient descent. *arXiv preprint arXiv:1704.05155*, 2017.

Saibaba, A. K., Hart, J., and van Bloemen Waanders, B. Randomized algorithms for generalized singular value decomposition with application to sensitivity analysis. *Numerical Linear Algebra with Applications*, pp. e2364, 2021.

Sameh, A. and Tong, Z. The trace minimization method for the symmetric generalized eigenvalue problem. *Journal of computational and applied mathematics*, 123(1-2):155–175, 2000.

Serfling, R. J. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

Shi, J., Sun, S., and Zhu, J. A spectral approach to gradient estimation for implicit distributions. *arXiv preprint arXiv:1806.02925*, 2018.

Singhal, R., Han, X., Lahlou, S., and Ranganath, R. Kernelized complete conditional stein discrepancy. *arXiv preprint arXiv:1904.04478*, 2019.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pp. 11918–11930, 2019.

Song, Y., Garg, S., Shi, J., and Ermon, S. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pp. 574–584. PMLR, 2020.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Hyvärinen, A., and Kumar, R. Density estimation in infinite dimensional exponential families. *The Journal of Machine Learning Research*, 18(1):1830–1888, 2017.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12 (7), 2011.

Sutherland, D., Strathmann, H., Arbel, M., and Gretton, A. Efficient and principled score estimation with nyström kernel exponential families. In *International Conference on Artificial Intelligence and Statistics*, pp. 652–660. PMLR, 2018.

Welling, M., Rosen-Zvi, M., and Hinton, G. E. Exponential family harmoniums with an application to information retrieval. Citeseer.

Wenliang, L., Sutherland, D., Strathmann, H., and Gretton, A. Learning deep kernels for exponential family densities. In *International Conference on Machine Learning*, pp. 6737–6746. PMLR, 2019.

Yu, Y., Wang, T., and Samworth, R. J. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Zahm, O., Cui, T., Law, K., Spantini, A., and Marzouk, Y. Certified dimension reduction in nonlinear bayesian inverse problems. *arXiv preprint arXiv:1807.03712*, 2018.

Zahm, O., Constantine, P. G., Prieur, C., and Marzouk, Y. M. Gradient-based dimension reduction of multivariate vector-valued functions. *SIAM Journal on Scientific Computing*, 42(1):A534–A558, 2020.

Zhou, Y., Shi, J., and Zhu, J. Nonparametric score estimators. *arXiv preprint arXiv:2005.10099*, 2020.