1 PROOFS

We first present another hardness result about the computation of expected kernels besides Theorem 2.2.

Theorem 1.1. There exist representations of distributions $p$ and $q$ that are smooth and compatible, yet computing the expected kernel of a simple kernel $k$ is the Kronecker delta is already $\#P$-hard.

Proof. (an alternative proof to the one in Section 4) Consider the case when the positive definite kernel $k$ is a Kronecker delta function defined as $k(x, x') = 1$ if and only if $x = x'$. Moreover, assume that the probabilistic circuit $p$ is smooth and decomposable, and that $q = p$. Then computing the expected kernel is equivalent to computing the power of a probabilistic circuit $p$, that is, $M_k(p, q) = \sum_{x \in \mathcal{X}} p^2(x)$ with $\mathcal{X}$ being the domain of variables $X$. [2021] proves that the task of computing $\sum_{x \in \mathcal{X}} p^2(x)$ is $\#P$-hard even when the PC $p$ is smooth and decomposable, which concludes our proof. 

Proposition 4.4 Let $p_n$ and $q_m$ be two compatible probabilistic circuits over variables $X$ whose output units $n$ and $m$ are sum units, denoted by $p_n(X) = \sum_{i \in \mathcal{X}} \theta_i p_i(X)$ and $q_m(X) = \sum_{j \in \mathcal{X}} \delta_j q_j(X)$ respectively. Let $k_l$ be a kernel circuit with its output unit being a sum unit $l$, denoted by $k_l(X) = \sum_{c \in \mathcal{X}} \gamma_c k_c(X)$. Then it holds that

$$M_{k_l}(p_n, q_m) = \sum_{i \in \mathcal{X}} \theta_i \sum_{j \in \mathcal{X}} \delta_j \sum_{c \in \mathcal{X}} \gamma_c M_{k_c}(p_i, q_j).$$

(1)

Proof. $M_{k_l}(p_n, q_m)$ can be expanded as

$$M_{k_l}(p_n, q_m) = \sum_{x \in \mathcal{X}} p_n(x) q_m(x') k_l(x, x')$$

$$= \sum_{x \in \mathcal{X}} \sum_{i \in \mathcal{X}} \theta_i p_i(x) \sum_{j \in \mathcal{X}} \delta_j q_j(x') \sum_{c \in \mathcal{X}} \gamma_c k_c(x, x')$$

$$= \sum_{i \in \mathcal{X}} \theta_i \sum_{j \in \mathcal{X}} \delta_j \sum_{c \in \mathcal{X}} \gamma_c M_{k_c}(p_i, q_j).$$

Proposition 4.5 Let $p_n$ and $q_m$ be two compatible probabilistic circuits over variables $X$ whose output units $n$ and $m$ are product units, denoted by $p_n(X) = p_{n_l}(X_L)p_{n_R}(X_R)$ and $q_m(X) = q_{m_l}(X_L)q_{m_R}(X_R)$. Let $k$ be a kernel circuit that is kernel-compatible with the circuit pair $p_n$ and $q_m$ with its output unit being a product unit denoted by $k(x, x') = k_L(X_L, X'_L)k_R(X_R, X'_R)$. Then it holds that

$$M_k(p_n, q_m) = M_{k_L}(p_{n_l}, q_{m_l}) \cdot M_{k_R}(p_{n_R}, q_{m_R}).$$

Proof. $M_k(p_n, q_m)$ can be expanded as

$$M_k(p_n, q_m) = \sum_{x \in \mathcal{X}} p_n(x) q_m(x') k(x, x')$$

$$= \sum_{x \in \mathcal{X}} \sum_{x' \in \mathcal{X}} p_{n_l}(x_L)p_{n_R}(x_R)q_{m_l}(x_L)q_{m_R}(x_R)k_L(x_L, x'_L)k_R(x_R, x'_R)$$

$$= M_{k_L}(p_{n_l}, q_{m_l}) \cdot M_{k_R}(p_{n_R}, q_{m_R}).$$

Corollary 4.6 Following the assumptions in Theorem 4.3, the squared maximum mean discrepancy $MMD[\mathcal{H}, p, q]$ in RKHS $\mathcal{H}$ associated with kernel $k$ as defined in [2012] can be tractably computed.
Proof. This is an immediate result following Theorem 4.3 by rewriting MMD as defined in [Gretton et al., 2012] in the form of a linear combination of expected kernels, that is, 
\[ MMD^2(\mathcal{H}, p, q) = M_k(p, p) + M_k(q, q) - 2M_k(p, q). \]

\[ \square \]

Corollary 4.7. Following the assumptions in Theorem 4.3 if the probabilistic circuit \( p \) further satisfies determinism, the kernelized discrete Stein discrepancy (KSD) 
\[ D^2(q \parallel p) = \mathbb{E}_{x \sim q}[k_p(x, x')] \] in the RKHS associated with kernel \( k \) as defined in [Yang et al., 2018] can be tractably computed.

Before showing the proof for Corollary 4.7, we first give definitions that are necessary for defining KSD as follows to be self-contained.

Definition 1.2 (Cyclic permutation). For a finite set \( \mathcal{X} \) and \( D = |\mathcal{X}| \), a cyclic permutation \( \gamma : \mathcal{X} \rightarrow \mathcal{X} \) is a bijective function such that for some ordering \( a_1, a_2, \cdots, a_D \) of the elements in \( \mathcal{X} \), \( \gamma = a_{(i+1) \mod D} \forall i = 1, 2, \cdots, D \).

Definition 1.3 (Partial difference operator). For any function \( f : \mathcal{X} \rightarrow \mathbb{R} \) with \( D = |\mathcal{X}| \), the partial difference operator is defined as

\[ \Delta_i^\gamma f(X) := f(X) - f(\gamma_i X), \forall i = 1, \cdots, D, \] (2)

with \( \gamma_i X := (X_1, \cdots, \gamma_i X, \cdots, X_D) \). Moreover, the difference operator is defined as \( \Delta^\gamma f(X) := (\Delta_1^\gamma f(X), \cdots, \Delta_D^\gamma f(X)) \). Similarly, let \( \gamma^{-1} \) be the inverse permutation of \( \gamma \), and \( \Delta \) denote the difference operator defined with respect to \( \gamma^{-1} \), i.e.,

\[ \Delta_i f(X) := f(X) - f(\gamma_i^{-1} X), i = 1, \cdots, D. \]

Definition 1.4 (Difference score function). The (difference) score function is defined as \( s_p(X) := \frac{\Delta^\gamma p(X)}{p(X)} \) on domain \( \mathcal{X} \) with \( D = |\mathcal{X}| \), a vector-valued function with its \( i \)-th dimension being

\[ s_{p,i}(X) := \frac{\Delta_i^\gamma p(X)}{p(X)} = 1 - \frac{p(\gamma_i X)}{p(X)}, i = 1, 2, \cdots, D. \] (3)

Given the above definitions, the discrete Stein discrepancy between two distributions \( p \) and \( q \) is defined as

\[ D(q \parallel p) := \sup_{f \in \mathcal{H}} \mathbb{E}_{x \sim q(x)}[T_p f(x)], \] (4)

where \( f : \mathcal{X} \rightarrow \mathbb{R}^D \) is a test function, belonging to some function space \( \mathcal{H} \) and \( T_p \) is the so-called Stein difference operator, which is defined as

\[ T_p f = s_p(x) f^\top - \Delta f(x). \] (5)

If the function space \( \mathcal{H} \) is an reproducing kernel Hilbert space (RKHS) on \( \mathcal{X} \) equipped with a kernel function \( k(\cdot, \cdot) \), then a kernelized discrete Stein discrepancy (KSD) is defined and admits a closed-form representation as

\[ S(q \parallel p) := D^2(q \parallel p) = \mathbb{E}_{x \sim \gamma \sim q}[k_p(x, x')]. \] (6)

Here, the kernel function \( k_p \) is defined as

\begin{align*}
&k_p(x, x') = s_p(x)^\top k(x, x') s_p(x') - s_p(x)^\top \Delta x^k(x, x') \\
&\quad - \Delta x^k(x, x')^\top s_p(x') + tr(\Delta x^k x'(x, x')), \nonumber
\end{align*}

where the difference operator \( \Delta x \) is as in Definition 1.3. The superscript \( x \) specifies the variables that it operates on.

Proof. [Corollary 4.7] By the definition of difference score functions, the close form of KSD can be further rewritten as follows.

\[ \mathbb{E}_{x \sim \gamma \sim q}[k_p(x, x')] = \sum_{i=1}^D \mathbb{E}_{x \sim \gamma \sim q}[\frac{p(\gamma_i x)p(\gamma_i x')}{p(x)p(x')} k(x, x') - \frac{p(\gamma_i x)}{p(x)} k(x, \gamma_i x')] \\
&= \sum_{i=1}^D [M_k(q \frac{\tilde{p}_i}{\tilde{p}} q \frac{\tilde{p}_i}{\tilde{p}}) - M_k(q \frac{\tilde{q}_i}{\tilde{p}}) \tilde{q}_i] \\
&= M_k(\tilde{q}_i, q \frac{\tilde{p}_i}{\tilde{p}}) + M_k(\tilde{q}_i, \tilde{q}_i) \] (7)

where \( D \) denotes the cardinality of the domain of variables \( X \), the probability \( \tilde{p}_i(X) := p(\gamma_i X) \) and the probability \( \tilde{q}_i(X) := q(\gamma_i X) \). Notice that the cyclic permutation \( \gamma_i \) operates on individual variable and the resulting PC \( \tilde{p}_i \) and \( \tilde{q}_i \) retains the same structure properties as PCs \( p \) and \( q \) respectively. To prove that KSD can be tractably computed, it suffices to prove that the expected kernel terms in Equation 7 can be tractably computed.

For a deterministic and structured-decomposable PC \( p \), since PC \( \tilde{p}_i \) retains the same structure, then resulting ratio \( \tilde{p}_i/p \) is again a smooth circuit compatible with \( p \) by [Vergari et al., 2021]. Moreover, since PC \( p \) and \( q \) are compatible, the circuit \( \tilde{p}_i/p \) is compatible with PC \( q \). Thus, the resulting product \( q \frac{\tilde{p}_i}{\tilde{p}} \) is a circuit that is smooth and compatible with both \( p \) and \( q \) by Theorem B.2 and thus compatible with \( \tilde{q}_i \). By similar arguments, we can verify that all the circuit pair in the expected kernel terms in Equation 7 satisfy the assumptions in Theorem 4.3 and thus they are amenable to the tractable computation we propose in Algorithm 1 which finishes our proof.

\[ \square \]

Proposition (convergence of Categorical BBIS). Let \( f(x) \) be a test function. Assume that \( f - \mathbb{E}_p[f] \in H_p \), with
\( \mathcal{H}_p \) being the RKHS associated with the kernel function \( k_p \), and \( \sum_i w_i = 1 \), then it holds that

\[
\left| \sum_{n=1}^{N} w_n f(x_n) - \mathbb{E}_p f \right| \leq C_f \sqrt{\mathbb{S}(\{x^{(n)}, w_n\} \parallel p)},
\]

where \( C_f := \| f - \mathbb{E}_p f \|_{\mathcal{H}_p} \). Moreover, the convergence rate is \( O(N^{-1/2}) \).

**Proof.** Let \( \hat{f}(x) := f(x) - \mathbb{E}_p f \), then it holds that

\[
\left| \sum_{n=1}^{N} w_n f(x^{(n)}) - \mathbb{E}_p f \right| = \left| \sum_{n=1}^{N} w_n \hat{f}(x^{(n)}) \right| \\
\leq \left\| \hat{f} \right\|_{\mathcal{H}_p} \cdot \left| \sum_{n=1}^{N} w_n k_p(\cdot, x^{(n)}) \right\|_{\mathcal{H}_p} \\
= \left\| \hat{f} \right\|_{\mathcal{H}_p} \cdot \sqrt{\mathbb{S}(\{x^{(n)}, w_n\} \parallel p)}.
\]

We further prove the convergence rate of the estimation error by using the importance weights as reference weights. Let \( v_{n}^* = \frac{1}{n} p(x^{(n)})/q(x^{(n)}) \). Then \( \mathbb{S}(\{x^{(n)}, v_{n}^*\} \parallel p) \) is a degenerate \( V \)-statistics [Liu and Lee, 2017] and it holds that \( \mathbb{S}(\{x^{(n)}, v_{n}^*\} \parallel p) = O(N^{-1}). \) Moreover, we have that \( \sum_{n=1}^{N} v_{n}^* = \frac{1}{n} + O(N^{-1/2}) \), which we denote by \( Z \), i.e., \( Z = \sum_{n=1}^{N} v_{n}^* \). Let \( w_n = v_{n}^*/Z \), then it holds that

\[
\mathbb{S}(\{x^{(n)}, w_n\} \parallel p) = \frac{\mathbb{S}(\{x^{(n)}, v_{n}^*\} \parallel p)}{Z^2} = O(N^{-1}).
\]

Therefore,

\[
\left| \sum_{n=1}^{N} w_n f(x^{(n)}) - \mathbb{E}_p f \right| \leq \left| \hat{f} \right|_{\mathcal{H}_p} \cdot \sqrt{\mathbb{S}(\{x^{(n)}, w_n\} \parallel p)} \\
\leq \left| \hat{f} \right|_{\mathcal{H}_p} \cdot \sqrt{\mathbb{S}(\{x^{(n)}, w_n\} \parallel p)} \\
= O(N^{-1/2}).
\]

**Proposition 5.5** Let \( p(X_e \mid x_a) \) be a PC that encodes a conditional distribution over variables \( X_e \) conditioned on \( X_s = x_a \), and \( k \) be a KC. If the PC \( p(X_e \mid x_a) \) and \( p(X_e \mid x_a') \) are compatible and \( k \) is kernel-compatible with the PC pair for any \( x_a, x_a' \), then the conditional kernel function \( k_{p,s} \) as defined in Proposition 5.4 can be tractably computed.

**Proof.** From Proposition 5.4, \( k_{p,s} \) can be written as

\[
k_{p,s} = \sum_{i=1}^{D} \mathbb{E}_{x_e \sim p(X_e \mid x_a), x_e' \sim p(X_e \mid x_a')} [k_{p,i}(x, x')],
\]

where \( k_{p,i} \) can be expanded as follows.

\[
k_{p,i}(x, x') = \frac{p(\gamma \mid x) p(\gamma \mid x')}{p(x) p(x')} k_{p,i}(x, x') \\
- \frac{p(\gamma \mid x)}{p(x)} k(x, \gamma \mid x') \\
- \frac{p(\gamma \mid x')}{p(x')} k(\gamma \mid x, x').
\]

for any \( i \in s \), given that none of the variables in \( X_s \) is flipped in the above formulation, kernel \( k_{p,i} \) can be further written as

\[
k_{p,i}(x, x') = \frac{p(\gamma \mid x_e \mid x_a) p(\gamma \mid x_e' \mid x_a')}{p(x_e \mid x_a) p(x_e' \mid x_a')} k(x, x') \\
- \frac{p(\gamma \mid x_e \mid x_a)}{p(x_e \mid x_a)} k(x, \gamma \mid x_a') \\
- \frac{p(\gamma \mid x_e' \mid x_a')}{p(x_e' \mid x_a')} k(\gamma \mid x, x_a') \\
+ k(\gamma \mid x, \gamma \mid x_a').
\]

By substituting \( k_{p,i} \) into the expected kernel in the expectation of \( k_{p,i} \) with respect to the conditional distributions can be simplified to be a constant zero, that is,

\[
\mathbb{E}_{x_e \sim p(X_e \mid x_a), x_e' \sim p(X_e \mid x_a')} [k_{p,i}(x, x')] = 0.
\]

Thus, \( k_{p,s} \) can be expanded as

\[
k_{p,s}(x, x') = \mathbb{E}_{x_e \sim p(X_e \mid x_a), x_e' \sim p(X_e \mid x_a')} \left( \sum_{i \in s} k_{p,i}(x, x') \right) \\
= \sum_{i \in s} \left[ \frac{p(\gamma \mid x_e \mid x_a) p(\gamma \mid x_e' \mid x_a')}{p(x_e \mid x_a) p(x_e' \mid x_a')} k(\cdot, \cdot) - \frac{p(\gamma \mid x_e \mid x_a)}{p(x_e \mid x_a)} k(\cdot, \gamma \mid x_a') - \frac{p(\gamma \mid x_e' \mid x_a')}{p(x_e' \mid x_a')} k(\gamma \mid \cdot, \cdot) + k(\gamma \mid \cdot, \gamma \mid x_a') \right].
\]

As Theorem 4.3 has shown that \( M_{q}(p, q) \) can be computed exactly in time linear in the size of each PC, \( k_{p,s}(x, x') \) can also be computed exactly in time \( O(|p_1| |p_2| |k|) \), where \( p_1 \) and \( p_2 \) denote circuits that represent the conditional probability distribution given the index set, i.e., \( p(\cdot \mid x_a) \) or \( p(\cdot \mid \gamma \mid x_a) \).

**2 ALGORITHMS**

Algorithm summarizes how to perform the BBIS scheme we propose for Categorical distributions, and generate a set of weighted samples.
Algorithm 1 CATEGORICALBBIS\((p, q, k, n)\)

**Input:** target distributions \(p\) over variables \(X\), a black-box mechanism \(q\), a kernel function \(k\) and number of samples \(n\)

**Output:** weighted samples \(\{(x^{(i)}, w^*_i)\}_{i=1}^n\)

1. Sample \(\{x^{(i)}\}_{i=1}^n\) from \(q\)
2. for \(i = 1, \ldots, n\) do
3.   for \(j = 1, \ldots, n\) do
4.     \([K_p]_{ij} = k_p(x^{(i)}, x^{(j)})\) \(\triangleright\) cf. Section 5.2
5.   \(w^* = \arg\min_w \{w^\top K_p w \mid \sum_{i=1}^n w_i = 1, w_i \geq 0\}\)
6. return \(\{(x^{(i)}, w^*_i)\}_{i=1}^n\)