Tractable Computation of Expected Kernels (Supplementary material)

Wenzhe Li[*1] **Zhe Zeng**[*2] **Antonio Vergari**[2] **Guy Van den Broeck**[2]

[1]Tsinghua University
[2]University of California, Los Angeles

scott.wenzhe.li@gmail.com, {zhezeng, aver, guyvdb}@cs.ucla.edu

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 that 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 $k(\mathbf{x}, \mathbf{x}') = 1$ if and only if $\mathbf{x} = \mathbf{x}'$. 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, $M_k(p, q) = \sum_{\mathbf{x}\in\mathcal{X}} p^2(\mathbf{x})$ with $\mathcal{X}$ being the domain of variables $\mathbf{X}$. Vergari et al. [2021] proves that the task of computing $\sum_{\mathbf{x}\in\mathcal{X}} p^2(\mathbf{x})$ 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 $\mathbf{X}$ whose output units n and m are sum units, denoted by $p_n(\mathbf{X}) = \sum_{i\in\mathsf{in}(n)} \theta_i p_i(\mathbf{X})$ and $q_m(\mathbf{X}) = \sum_{j\in\mathsf{in}(m)} \delta_j q_j(\mathbf{X})$ respectively. Let k_l be a kernel circuit with its output unit being a sum unit l, denoted by $k_l(\mathbf{X}) = \sum_{c\in\mathsf{in}(l)} \gamma_c k_c(\mathbf{X})$. Then it holds that

$$M_{k_l}(p_n, q_m) = \sum_{i\in\mathsf{in}(n)} \theta_i \sum_{j\in\mathsf{in}(m)} \delta_j \sum_{c\in\mathsf{in}(l)} \gamma_c \, M_{k_c}(p_i, q_j). \tag{1}$$

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

$$\begin{aligned} & M_{k_l}(p_n, q_m) \\ &= \sum_{\mathbf{x}} \sum_{\mathbf{x}'} p_n(\mathbf{x}) q_m(\mathbf{x}') k_l(\mathbf{x}, \mathbf{x}') \\ &= \sum_{\mathbf{x}} \sum_{\mathbf{x}'} \sum_{i\in\mathsf{in}(n)} \theta_i p_i(\mathbf{x}) \sum_{j\in\mathsf{in}(m)} \delta_j q_j(\mathbf{x}') \sum_{c\in\mathsf{in}(l)} \gamma_c k_c(\mathbf{x}, \mathbf{x}') \\ &= \sum_{i\in\mathsf{in}(n)} \theta_i \sum_{j\in\mathsf{in}(m)} \delta_j \sum_{c\in\mathsf{in}(l)} \gamma_c \, M_{k_c}(p_i, q_j). \end{aligned}$$

□

Proposition 4.5 Let p_n and q_m be two compatible probabilistic circuits over variables $\mathbf{X}$ whose output units n and m are product units, denoted by $p_n(\mathbf{X}) = p_{n_\mathsf{L}}(\mathbf{X}_\mathsf{L}) p_{n_\mathsf{R}}(\mathbf{X}_\mathsf{R})$ and $q_m(\mathbf{X}) = q_{m_\mathsf{L}}(\mathbf{X}_\mathsf{L}) q_{m_\mathsf{R}}(\mathbf{X}_\mathsf{R})$. 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 $k(\mathbf{X}, \mathbf{X}') = k_\mathsf{L}(\mathbf{X}_\mathsf{L}, \mathbf{X}'_\mathsf{L}) k_\mathsf{R}(\mathbf{X}_\mathsf{R}, \mathbf{X}'_\mathsf{R})$. Then it holds that

$$M_k(p_n, q_m) = M_{k_\mathsf{L}}(p_{n_\mathsf{L}}, q_{m_\mathsf{L}}) \cdot M_{k_\mathsf{R}}(p_{n_\mathsf{R}}, q_{m_\mathsf{R}}).$$

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

$$\begin{aligned} & M_k(p_n, q_m) \\ &= \sum_{\mathbf{x}} \sum_{\mathbf{x}'} p_n(\mathbf{x}) q_m(\mathbf{x}') k(\mathbf{x}, \mathbf{x}') \\ &= \sum_{\mathbf{x}} \sum_{\mathbf{x}'} p_{m_\mathsf{L}}(\mathbf{x}_\mathsf{L}) p_{m_\mathsf{R}}(\mathbf{x}_\mathsf{R}) q_{n_\mathsf{L}}(\mathbf{x}_\mathsf{L}) q_{n_\mathsf{R}}(\mathbf{x}_\mathsf{R}) k_\mathsf{L}(\mathbf{x}_\mathsf{L}, \mathbf{x}'_\mathsf{L}) k_\mathsf{R}(\mathbf{x}_\mathsf{R}, \mathbf{x}'_\mathsf{R}) \\ &= M_{k_\mathsf{L}}(p_{n_\mathsf{L}}, q_{m_\mathsf{L}}) \cdot M_{k_\mathsf{R}}(p_{n_\mathsf{R}}, q_{m_\mathsf{R}}). \end{aligned}$$

□

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 Gretton et al. [2012] can be tractably computed.

[*]Authors contributed equally. This research was performed while W.L. was visiting UCLA remotely.

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, $MMD^2[\mathcal{H}, p, q] = M_k(p,p) + M_k(q,q) - 2M_k(p,q)$. □

Corollary 4.7. Following the assumptions in Theorem 4.3, if the probabilistic circuit p further satisfies determinism, the kernelized discrete Stein discrepancy (KDSD) $\mathbb{D}^2(q \parallel p) = \mathbb{E}_{\mathbf{x},\mathbf{x}' \sim q}[k_p(\mathbf{x},\mathbf{x}')]$ 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 KDSD as follows to be self-contained.

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

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

$$\Delta_i^* f(\mathbf{X}) := f(\mathbf{X}) - f(\neg_i \mathbf{X}), \forall i = 1, \cdots, D, \tag{2}$$

with $\neg_i \mathbf{X} := (X_1, \cdots, \neg X_i, \cdots, X_D)$. Moreover, the difference operator is defined as $\Delta^* f(\mathbf{X}) := (\Delta_1^* f(\mathbf{X}), \cdots, \Delta_D^* f(\mathbf{X}))$. Similarly, let $\llcorner$ be the inverse permutation of $\neg$, and Δ denote the difference operator defined with respect to $\llcorner$, i.e.,

$$\Delta_i f(\mathbf{X}) := f(\mathbf{X}) - f(\llcorner_i \mathbf{X}), i = 1, \cdots, D.$$

Definition 1.4 (Difference score function). *The (difference) score function is defined as $\boldsymbol{s}_p(\mathbf{X}) := \frac{\Delta^* p(\mathbf{X})}{p(\mathbf{X})}$ on domain $\mathcal{X}$ with $D = \mid \mathcal{X} \mid$, a vector-valued function with its i-th dimension being*

$$\boldsymbol{s}_{p,i}(\mathbf{X}) := \frac{\Delta_i^* p(\mathbf{X})}{p(\mathbf{X})} = 1 - \frac{p(\neg_i \mathbf{X})}{p(\mathbf{X})}, i = 1, 2, \cdots, D. \tag{3}$$

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

$$\mathbb{D}(q \parallel p) := \sup_{\boldsymbol{f} \in \mathcal{H}} \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})}[\mathcal{T}_p \boldsymbol{f}(\mathbf{x})], \tag{4}$$

where $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^D$ is a *test* function, belonging to some function space $\mathcal{H}$ and $\mathcal{T}_p$ is the so-called *Stein difference operator*, which is defined as

$$\mathcal{T}_p \boldsymbol{f} = \boldsymbol{s}_p(\mathbf{x}) \boldsymbol{f}^\top - \Delta \boldsymbol{f}(\mathbf{x}). \tag{5}$$

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* (KDSD) is defined and admits a closed-form representation as

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

Here, the kernel function k_p is defined as

$$\begin{aligned} k_p(\mathbf{x}, \mathbf{x}') = \boldsymbol{s}_p(\mathbf{x})^\top k(\mathbf{x}, \mathbf{x}') \boldsymbol{s}_p(\mathbf{x}') - \boldsymbol{s}_p(\mathbf{x})^\top \Delta^{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') \\ - \Delta^{\mathbf{x}} k(\mathbf{x}, \mathbf{x}')^\top \boldsymbol{s}_p(\mathbf{x}') + tr(\Delta^{\mathbf{x},\mathbf{x}'} k(\mathbf{x}, \mathbf{x}')), \end{aligned}$$

where the difference operator $\Delta^{\mathbf{x}}$ is as in Definition 1.3. The superscript $\mathbf{x}$ specifies the variables that it operates on.

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

$$\begin{aligned} &\mathbb{E}_{\mathbf{x},\mathbf{x}' \sim q}[k_p(\mathbf{x}, \mathbf{x}')] \\ =& \sum_{i=1}^{D} \mathbb{E}_{\mathbf{x},\mathbf{x}' \sim q}[\frac{p(\neg_i \mathbf{x}) p(\neg_i \mathbf{x}')}{p(\mathbf{x}) p(\mathbf{x}')} k(\mathbf{x}, \mathbf{x}') - \frac{p(\neg_i \mathbf{x})}{p(\mathbf{x})} k(\mathbf{x}, \neg_i \mathbf{x}') \\ &- \frac{p(\neg_i \mathbf{x}')}{p(\mathbf{x}')} k(\neg_i \mathbf{x}, \mathbf{x}') + k(\neg_i \mathbf{x}, \neg_i \mathbf{x}')] \\ =& \sum_{i=1}^{D} [M_k(q \frac{\tilde{p}_i}{p}, q \frac{\tilde{p}_i}{p}) - M_k(q \frac{\tilde{p}_i}{p}, \tilde{q}_i) \\ &- M_k(\tilde{q}_i, q \frac{\tilde{p}_i}{p}) + M_k(\tilde{q}_i, \tilde{q}_i)] \end{aligned} \tag{7}$$

where D denotes the cardinality of the domain of variables $\mathbf{X}$, the probablity $\tilde{p}_i(\mathbf{X}) := p(\neg_i \mathbf{X})$ and the probablity $\tilde{q}_i(\mathbf{X}) := q(\neg_i \mathbf{X})$. Notice that the cyclic permutation $\neg_i$ 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 KDSD can be tratably 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 $q \frac{\tilde{p}_i}{p}$ 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.

□

Proposition (convergence of Categorical BBIS). Let $f(\mathbf{x})$ be a test function. Assume that $f - \mathbb{E}_p[f] \in \mathcal{H}_p$, with

$\mathcal{H}_p$ being the RKHS associated with the kernel function k_p, and $\sum_i w_i = 1$, then it holds that

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

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

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

$$\begin{aligned}
\left| \sum_{n=1}^{N} w_n f(\mathbf{x}^{(n)}) - \mathbb{E}_p f \right| &= \left| \sum_{n=1}^{N} w_n \hat{f}(\mathbf{x}^{(n)}) \right| \\
&= \left| \sum_{n=1}^{N} w_n \langle \hat{f}, k_p(\cdot, \mathbf{x}^{(n)}) \rangle \right| \\
&= \left| \langle \hat{f}, \sum_{n=1}^{N} w_n k_p(\cdot, \mathbf{x}^{(n)}) \rangle_{\mathcal{H}_p} \right| \\
&\leq \parallel \hat{f} \parallel_{\mathcal{H}_p} \cdot \parallel \sum_{n=1}^{N} w_n k_p(\cdot, \mathbf{x}^{(n)}) \parallel_{\mathcal{H}_p} \\
&= \parallel \hat{f} \parallel_{\mathcal{H}_p} \cdot \sqrt{\mathbb{S}(\{\mathbf{x}^{(n)}, w_n\} \parallel p)}.
\end{aligned}$$

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

$$\mathbb{S}(\{\mathbf{x}^{(n)}, w_n^*\} \parallel p) = \frac{\mathbb{S}(\{\mathbf{x}^{(n)}, v_n^*\} \parallel p)}{Z^2} = \mathcal{O}(N^{-1}).$$

Therefore,

$$\begin{aligned}
\left| \sum_{n=1}^{N} w_n f(\mathbf{x}^{(n)}) - \mathbb{E}_p f \right| &\leq \parallel \hat{f} \parallel_{\mathcal{H}_p} \cdot \sqrt{\mathbb{S}(\{\mathbf{x}^{(n)}, w_n\} \parallel p)} \\
&\leq \parallel \hat{f} \parallel_{\mathcal{H}_p} \cdot \sqrt{\mathbb{S}(\{\mathbf{x}^{(n)}, w_n^*\} \parallel p)} \\
&= \mathcal{O}(N^{-1/2}).
\end{aligned}$$

□

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

Proof. From Proposition 5.4, $k_{p,\mathbf{s}}$ can be written as

$$k_{p,\mathbf{s}} = \sum_{i=1}^{D} \mathbb{E}_{\mathbf{x_c} \sim p(\mathbf{X_c}|\mathbf{x_s}), \mathbf{x}'_\mathbf{c} \sim p(\mathbf{X_c}|\mathbf{x}'_\mathbf{s})} [k_{p,i}(\mathbf{x}, \mathbf{x}')],$$

where $k_{p,i}$ can be expanded as follows.

$$\begin{aligned}
k_{p,i}(\mathbf{x}, \mathbf{x}') =& \frac{p(\neg_i \mathbf{x}) p(\neg_i \mathbf{x}')}{p(\mathbf{x}) p(\mathbf{x}')} k(\mathbf{x}, \mathbf{x}') - \frac{p(\neg_i \mathbf{x})}{p(\mathbf{x})} k(\mathbf{x}, \neg_i \mathbf{x}') \\
&- \frac{p(\neg_i \mathbf{x}')}{p(\mathbf{x}')} k(\neg_i \mathbf{x}, \mathbf{x}') + k(\neg_i \mathbf{x}, \neg_i \mathbf{x}').
\end{aligned}$$

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

$$\begin{aligned}
k_{p,i}(\mathbf{x}, \mathbf{x}') =& \frac{p(\neg_i \mathbf{x_c} \mid \mathbf{x_s}) p(\neg_i \mathbf{x}'_\mathbf{c} \mid \mathbf{x}'_\mathbf{s})}{p(\mathbf{x_c} \mid \mathbf{x_s}) p(\mathbf{x}'_\mathbf{c} \mid \mathbf{x}'_\mathbf{s})} k(\mathbf{x}, \mathbf{x}') \\
&- \frac{p(\neg_i \mathbf{x_c} \mid \mathbf{x_s})}{p(\mathbf{x_c} \mid \mathbf{x_s})} k(\mathbf{x}, \neg_i \mathbf{x}') \\
&- \frac{p(\neg_i \mathbf{x}'_\mathbf{c} \mid \mathbf{x}'_\mathbf{s})}{p(\mathbf{x}'_\mathbf{c} \mid \mathbf{x}'_\mathbf{s})} k(\neg_i \mathbf{x}, \mathbf{x}') \\
&+ k(\neg_i \mathbf{x}, \neg_i \mathbf{x}').
\end{aligned}$$

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,

$$\mathbb{E}_{\mathbf{x_c} \sim p(\mathbf{X_c}|\mathbf{x_s}), \mathbf{x}'_\mathbf{c} \sim p(\mathbf{X}'_\mathbf{c}|\mathbf{x}'_\mathbf{s})} [k_{p,i}(\mathbf{x}, \mathbf{x}')] = 0.$$

Thus, $k_{p,\mathbf{s}}$ can be expanded as

$$\begin{aligned}
k_{p,\mathbf{s}}(\mathbf{x}, \mathbf{x}') &= \mathbb{E}_{\mathbf{x_c} \sim p(\mathbf{X_c}|\mathbf{x_s}), \mathbf{x}'_\mathbf{c} \sim p(\mathbf{X_c}|\mathbf{x}'_\mathbf{s})} [\sum_{i \in \mathbf{s}} k_{p,i}(\mathbf{x}, \mathbf{x}')] \\
&= \sum_{i \in \mathbf{s}} [\frac{p(\neg_i \mathbf{x_s}) p(\neg_i \mathbf{x}'_\mathbf{s})}{p(\mathbf{x_s}) p(\mathbf{x}'_\mathbf{s})} \cdot M_{k(\cdot,\cdot)}(p(\cdot \mid \neg_i \mathbf{x_s}), p(\cdot \mid \neg_i \mathbf{x}'_\mathbf{s})) \\
&\quad - \frac{p(\neg_i \mathbf{x_s})}{p(\mathbf{x_s})} \cdot M_{k(\cdot,\neg_i \cdot)}(p(\cdot \mid \neg_i \mathbf{x_s}), p(\cdot \mid \mathbf{x}'_\mathbf{s})) \\
&\quad - \frac{p(\neg_i \mathbf{x}'_\mathbf{s})}{p(\mathbf{x}'_\mathbf{s})} \cdot M_{k(\neg_i \cdot,\cdot)}(p(\cdot \mid \mathbf{x_s}), p(\cdot \mid \neg_i \mathbf{x}'_\mathbf{s})) \\
&\quad + M_{k(\neg_i \cdot,\neg_i \cdot)}(p(\cdot \mid \mathbf{x_s}), p(\cdot \mid \mathbf{x}'_\mathbf{s}))].
\end{aligned}$$

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

2 ALGORITHMS

Algorithm 1 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 $\mathbf{X}$, a black-box mechanism q, a kernel function k and number of samples n
Output: weighted samples $\{(\mathbf{x}^{(i)}, w_i^*)\}_{i=1}^n$

Sample $\{\mathbf{x}^{(i)}\}_{i=1}^n$ from q
for $i = 1, \ldots, n$ **do**
    for $j = 1, \ldots, n$ **do**
        $[\boldsymbol{K}_p]_{ij} = k_p(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ ▷ cf. Section 5.2
$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \left\{ \boldsymbol{w}^\top \boldsymbol{K_p} \boldsymbol{w} \,\middle|\, \sum_{i=1}^n w_i = 1,\ w_i \geq 0 \right\}$
return $\{(\mathbf{x}^{(i)}, w_i^*)\}_{i=1}^n$