Appendix to
Goodness-of-fit Testing for Discrete Distributions via Stein Discrepancy

Proof of Theorem 1. Clearly, $p = q$ implies that $\mathbf{s}_p(\mathbf{x}) = \mathbf{s}_q(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}^d$. It remains to be shown that the converse is true. By Eq. (1), $\mathbf{s}_p(\mathbf{x}) = \mathbf{s}_q(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}^d$ implies that ${}^{p(\neg_i\mathbf{x})}/_{p(\mathbf{x})} = {}^{q(\neg_i\mathbf{x})}/_{q(\mathbf{x})}$ for all $\mathbf{x} \in \mathcal{X}^d$ and all $i = 1, \ldots, d$. We show that the latter implies that all the singleton conditional distributions of p and q must match, *i.e.*, $p(x_i|\mathbf{x}_{-i}) = q(x_i|\mathbf{x}_{-i})$ for all $x_i \in \mathcal{X}$ and for all $i = 1, \ldots, d$, where $\mathbf{x}_{-i} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$.

Specifically, using the fact that $\neg$ is a cyclic permutation on $\mathcal{X}$, we can write

$$\begin{aligned}
\frac{1}{p(x_i|\mathbf{x}_{-i})} &= \frac{\sum_{\xi_i \in \mathcal{X}} p(x_1, \ldots, x_{i-1}, \xi_i, x_{i+1}, \ldots, x_d)}{p(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)} = \sum_{\xi_i \in \mathcal{X}} \frac{p(x_1, \ldots, x_{i-1}, \xi_i, x_{i+1}, \ldots, x_d)}{p(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)} \\
&= \sum_{\ell=1}^{|\mathcal{X}|} \frac{p(x_1, \ldots, x_{i-1}, \neg^{(\ell)} x_i, x_{i+1}, \ldots, x_d)}{p(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)} \\
&= \sum_{\ell=1}^{|\mathcal{X}|} \frac{p(\neg_i^{(\ell)}\mathbf{x})}{p(\mathbf{x})} = \sum_{\ell=1}^{|\mathcal{X}|} \prod_{j=0}^{\ell-1} \frac{p(\neg_i^{(j+1)}\mathbf{x})}{p(\neg_i^{(j)}\mathbf{x})} = \sum_{\ell=1}^{|\mathcal{X}|} \prod_{j=0}^{\ell-1} \frac{p(\neg_i \mathbf{y}_{ij})}{p(\mathbf{y}_{ij})}\,,
\end{aligned} \tag{17}$$

where we adopted the convention that $\neg^{(0)}\mathbf{x} = \mathbf{x}$ and written $\mathbf{y}_{ij} := \neg_i^{(j)}\mathbf{x}$ in the last term. By Eq. (1), all the terms on the right-hand-side of Eq. (17) will be determined by the score function $\mathbf{s}_p(\mathbf{x})$, and thus $\mathbf{s}_p(\mathbf{x}) = \mathbf{s}_q(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}^d$ implies that all the singleton conditional distributions must match: $p(x_i|\mathbf{x}_{-i}) = q(x_i|\mathbf{x}_{-i})$, $\forall \mathbf{x} \in \mathcal{X}^d$. By Brook's lemma (Brook, 1964; see Lemma 9 for a self-contained proof), the joint probability distribution is fully specified by the collection of singleton conditional distributions, and thus we must have $p(\mathbf{x}) = q(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}^d$. □

Lemma 9 (Brook, 1964)**.** *Assume that $p(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}^d$. The joint distribution $p(\mathbf{x})$ is completely determined by the collection of singleton conditional distributions $p(x_i|\mathbf{x}_{-i})$, where $\mathbf{x}_{-i} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$, $i = 1, \ldots, d$.*

Proof. Let $p(x_1, \ldots, x_d)$ and $p(y_1, \ldots, y_d)$ denote the joint densities (pmfs or pdfs) for $(x_1, \ldots, x_d)$ and $(y_1, \ldots, y_d)$, respectively. We can write

$$\begin{aligned}
\frac{p(x_1, x_2, \ldots, x_d)}{p(y_1, y_2, \ldots, y_d)} &= \frac{p(x_1, x_2, \ldots, x_d)}{p(y_1, x_2, \ldots, x_d)} \cdot \frac{p(y_1, x_2, \ldots, x_d)}{p(y_1, y_2, \ldots, x_d)} \cdots \frac{p(y_1, y_2, \ldots, y_{d-1}, x_d)}{p(y_1, y_2, \ldots, y_{d-1}, y_d)} \\
&= \frac{p(x_1|x_2, \ldots, x_d)}{p(y_1|x_2, \ldots, x_d)} \cdot \frac{p(x_2|y_1, x_3, \ldots, x_d)}{p(y_2|y_1, x_3, \ldots, x_d)} \cdots \frac{p(x_d|y_1, \ldots, y_{d-1})}{p(y_d|y_1, \ldots, y_{d-1})}\,.
\end{aligned}$$

Thus, the collection of all singleton conditional distributions completely determine the ratios of joint probability densities, which in turn completely determine the joint densities themselves, since they have to sum to one. □

The following result provides more convenient expressions for evaluating $\mathbb{E}_{\mathbf{x}\sim p}\left[\mathcal{A}_p \mathbf{f}(\mathbf{x})\right]$ and $\mathbb{E}_{\mathbf{x}\sim p}\left[\mathrm{tr}\left(\mathcal{A}_p \mathbf{f}(\mathbf{x})\right)\right]$.

Lemma 10. (Ley & Swan, 2013) *For positive pmfs p and q,*

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim q}\left[\mathcal{A}_p \mathbf{f}(\mathbf{x})\right] &= \mathbb{E}_{\mathbf{x}\sim q}\left[\left(\mathbf{s}_p(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})\right)\mathbf{f}(\mathbf{x})^\mathsf{T}\right]; \\
\mathbb{E}_{\mathbf{x}\sim q}\left[\mathrm{tr}\left(\mathcal{A}_p \mathbf{f}(\mathbf{x})\right)\right] &= \mathbb{E}_{\mathbf{x}\sim q}\left[\left(\mathbf{s}_p(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})\right)^\mathsf{T}\mathbf{f}(\mathbf{x})\right].
\end{aligned}$$

Proof. Lemma 2 states that $\mathbb{E}_{\mathbf{x}\sim q}\left[\mathcal{A}_q \mathbf{f}(\mathbf{x})\right] = 0$. Thus, writing $\mathbb{E}_{\mathbf{x}\sim q}\left[\mathcal{A}_p \mathbf{f}(\mathbf{x})\right] = \mathbb{E}_{\mathbf{x}\sim q}\left[\mathcal{A}_p \mathbf{f}(\mathbf{x}) - \mathcal{A}_q \mathbf{f}(\mathbf{x})\right] = \mathbb{E}_{\mathbf{x}\sim q}[\left(\mathbf{s}_p(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})\right)\mathbf{f}(\mathbf{x})^\mathsf{T}]$ and taking the trace on both sides completes the proof. □

Proof of Theorem 3 (Continued). *Necessity:* Assume that a linear operator $\mathcal{T}$ satisfies Eq. (7); we show that it can be written in the form of Eq. (8) for some linear operators $\mathcal{L}$ and $\mathcal{L}^*$ of the forms (5) and (6). Recall that for a finite set $\mathcal{X}$, any function $f : \mathcal{X}^d \to \mathbb{R}$ can be represented by a vector $\mathbf{f} \in \mathbb{R}^{|\mathcal{X}|^d}$, and any linear operator $\mathcal{T}$ on the set of functions f can be represented via a matrix $\mathbf{T} \in \mathbb{R}^{|\mathcal{X}|^d \times |\mathcal{X}|^d}$ under the standard basis of $\mathbb{R}^{|\mathcal{X}|^d}$. Under these notations, $\mathcal{T}f$ can be represented by $\mathbf{T}\mathbf{f}$, and Eq. (7) can be rewritten in matrix form as

$$\mathbb{E}_{\mathbf{x}\sim p}\left[\mathcal{T}_p f(\mathbf{x})\right] = \sum_{\mathbf{x}\in\mathcal{X}^d} p(\mathbf{x})\mathcal{T}_p f(\mathbf{x}) = \mathbf{p}^\mathsf{T}(\mathbf{T}_p\mathbf{f}) = 0\,,$$

which holds for any function f (*i.e.*, for any vector $\mathbf{f}$) if and only if $\mathbf{p}^\mathsf{T}\mathbf{T}_p = \mathbf{0}$. We can always find a diagonal matrix $\mathbf{D}$ and a matrix $\mathbf{L}$ such that $\mathbf{T}_p = \mathbf{D} - \mathbf{L}$. Observe that $\mathbf{p}^\mathsf{T}\mathbf{T}_p = \mathbf{0}$, *i.e.*, $\mathbf{p}^\mathsf{T}\mathbf{D} = \mathbf{p}^\mathsf{T}\mathbf{L}$ if and only if $d_{ii} = \mathbf{p}^\mathsf{T}\mathbf{L}_{*i}/p_i$ for all i, where d_{ii} is the i-th diagonal element of $\mathbf{D}$ and $\mathbf{L}_{*i}$ is the i-th column of $\mathbf{L}$. Thus, Eq. (7) holds if and only if

$$\mathbf{T}_p = \operatorname{diag}\left\{\mathbf{p}\right\}^{-1}\operatorname{diag}\left\{\mathbf{L}^\mathsf{T}\mathbf{p}\right\} - \mathbf{L}$$

for some matrix $\mathbf{L}$, where $\operatorname{diag}\left\{\mathbf{p}\right\}$ denotes the diagonal matrix whose i-th diagonal entry equals p_i. Rewriting, we have

$$\operatorname{diag}\left\{\mathbf{p}\right\}\mathbf{T}_p = \operatorname{diag}\left\{\mathbf{L}^\mathsf{T}\mathbf{p}\right\} - \operatorname{diag}\left\{\mathbf{p}\right\}\mathbf{L}\,.$$

Right-multiplying both sides by an arbitrary vector $\mathbf{f} \in \mathbb{R}^{|\mathcal{X}|^d}$, we obtain

$$\mathbf{p} \odot (\mathbf{T}_p\mathbf{f}) = (\mathbf{L}^\mathsf{T}\mathbf{p}) \odot \mathbf{f} - \mathbf{p} \odot (\mathbf{L}^\mathsf{T}\mathbf{f})\,, \tag{18}$$

where $\odot$ denotes the Hadamard product. Let $\mathcal{L}$ and $\mathcal{L}^*$ be the linear operators with matrices $\mathbf{L}^\mathsf{T}$ and $\mathbf{L}$ under the standard basis, Eq. (18) can be re-written as

$$p(\mathbf{x})\mathcal{T}_p f(\mathbf{x}) = \mathcal{L}p(\mathbf{x})f(\mathbf{x}) - p(\mathbf{x})\mathcal{L}^* f(\mathbf{x})$$

for all $\mathbf{x} \in \mathcal{X}^d$. Finally, dividing by $p(\mathbf{x})$ on both sides yields Eq. (8). □

Proof of Theorem 6. Observe that

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim q}\left[\operatorname{tr}\left(\mathcal{A}_p\mathbf{f}(\mathbf{x})\right)\right] &= \sum_{\ell=1}^{d}\mathbb{E}_{\mathbf{x}\sim q}\left[s_p^\ell(\mathbf{x})\, f_\ell(\mathbf{x}) - \Delta^*_{x_\ell} f_\ell(\mathbf{x})\right] \\
&= \sum_{\ell=1}^{d}\mathbb{E}_{\mathbf{x}\sim q}\left[s_p^\ell(\mathbf{x})\left\langle f_\ell, k(\cdot,\mathbf{x})\right\rangle_{\mathcal{H}} - \left\langle f_\ell, \Delta^*_{x_\ell} k(\cdot,\mathbf{x})\right\rangle_{\mathcal{H}}\right] \\
&= \sum_{\ell=1}^{d}\left\langle f_\ell, \mathbb{E}_{\mathbf{x}\sim q}\left[s_p^\ell(\mathbf{x})\, k(\cdot,\mathbf{x}) - \Delta^*_{x_\ell} k(\cdot,\mathbf{x})\right]\right\rangle_{\mathcal{H}}\,,
\end{aligned}$$

where we used the reproducing property $\left\langle f_\ell, k(\cdot,\mathbf{x})\right\rangle_{\mathcal{H}} = f_\ell(\mathbf{x})$ and the fact that

$$\Delta^*_{x_j} f_i(\mathbf{x}) = f_i(\mathbf{x}) - f_i(\neg_j\mathbf{x}) = \left\langle f_i, k(\cdot,\mathbf{x})\right\rangle - \left\langle f_i, k(\cdot,\neg_j\mathbf{x})\right\rangle = \left\langle f_i, k(\cdot,\mathbf{x}) - k(\cdot,\neg_j\mathbf{x})\right\rangle = \left\langle f_j, \Delta^*_{x_j} k(\cdot,\mathbf{x})\right\rangle.$$

Denoting $\boldsymbol{\beta}(\cdot) := \mathbb{E}_{\mathbf{x}\sim q}\left[\mathbf{s}_p(\mathbf{x})k(\cdot,\mathbf{x}) - \Delta^* k(\cdot,\mathbf{x})\right] \in \mathcal{H}^m$, we have

$$\mathbb{E}_{\mathbf{x}\sim q}\left[\operatorname{tr}\left(\mathcal{A}_p\mathbf{f}(\mathbf{x})\right)\right] = \sum_{\ell=1}^{d}\left\langle f_\ell, \beta_\ell\right\rangle_{\mathcal{H}} = \left\langle \mathbf{f}, \boldsymbol{\beta}\right\rangle_{\mathcal{H}^m}\,.$$

Thus, we can rewrite the kernelized discrete Stein discrepancy as

$$\mathbb{D}(q\,\|\,p) = \sup_{\mathbf{f}\in\mathcal{H}^m,\,\|\mathbf{f}\|_{\mathcal{H}^m}\le 1}\left\langle \mathbf{f}, \boldsymbol{\beta}\right\rangle_{\mathcal{H}^m}\,,$$

which immediately implies that $\mathbb{D}(q\,\|\,p) = \|\boldsymbol{\beta}\|_{\mathcal{H}^m}$ since the supremum will be attained by $\mathbf{f} = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|_{\mathcal{H}^m}$.

By Lemma 10, we have

$$\boldsymbol{\beta}(\cdot) = \mathbb{E}_{\mathbf{x}\sim q}\left[\mathbf{s}_p(\mathbf{x})k(\cdot,\mathbf{x}) - \Delta^* k(\cdot,\mathbf{x})\right] = \mathbb{E}_{\mathbf{x}\sim q}\left[(\mathbf{s}_p(\mathbf{x}) - \mathbf{s}_q(\mathbf{x}))k(\cdot,\mathbf{x})\right].$$

Writing $\boldsymbol{\delta}_{p,q}(\mathbf{x}) := \mathbf{s}_p(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})$, we have

$$\begin{aligned}
\mathbb{D}(q \,\|\, p)^2 = \|\boldsymbol{\beta}\|^2_{\mathcal{H}^m} = \sum_{\ell=1}^{d} \langle \beta_\ell, \beta_\ell \rangle_{\mathcal{H}} &= \sum_{\ell=1}^{d} \left\langle \mathbb{E}_{\mathbf{x}\sim q}\left[\delta^\ell_{p,q}(\mathbf{x})\, k(\cdot,\mathbf{x})\right], \mathbb{E}_{\mathbf{x}'\sim q}\left[\delta^\ell_{p,q}(\mathbf{x}')\, k(\cdot,\mathbf{x}')\right] \right\rangle_{\mathcal{H}} \\
&= \sum_{\ell=1}^{d} \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\delta^\ell_{p,q}(\mathbf{x})\, \langle k(\cdot,\mathbf{x}), k(\cdot,\mathbf{x}')\rangle_{\mathcal{H}}\, \delta^\ell_{p,q}(\mathbf{x}')\right] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\boldsymbol{\delta}_{p,q}(\mathbf{x})^\top \langle k(\cdot,\mathbf{x}), k(\cdot,\mathbf{x}')\rangle_{\mathcal{H}}\, \boldsymbol{\delta}_{p,q}(\mathbf{x}')\right] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\boldsymbol{\delta}_{p,q}(\mathbf{x})^\top k(\mathbf{x},\mathbf{x}')\, \boldsymbol{\delta}_{p,q}(\mathbf{x}')\right],
\end{aligned}$$

where we used the reproducing property, $k(\mathbf{x},\mathbf{x}') = \langle k(\cdot,\mathbf{x}), k(\cdot,\mathbf{x}')\rangle_{\mathcal{H}}$. This concludes the proof. □

Proof of Theorem 7. Expanding the expression for $\boldsymbol{\delta}_{p,q}(x)$ and applying Lemma 10 twice, we obtain

$$\begin{aligned}
\mathbb{D}(q \,\|\, p)^2 &= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\boldsymbol{\delta}_{p,q}(\mathbf{x})^\top k(\mathbf{x},\mathbf{x}')\boldsymbol{\delta}_{p,q}(\mathbf{x}')\right] \\
&= \mathbb{E}_{\mathbf{x}\sim q}\left[\boldsymbol{\delta}_{p,q}(\mathbf{x})^\top \mathbb{E}_{\mathbf{x}'\sim q}\left[k(\mathbf{x},\mathbf{x}')\boldsymbol{\delta}_{p,q}(\mathbf{x}')\right]\right] \\
&= \mathbb{E}_{\mathbf{x}\sim q}\left[\boldsymbol{\delta}_{p,q}(\mathbf{x})^\top \mathbb{E}_{\mathbf{x}'\sim q}\left[k(\mathbf{x},\mathbf{x}')\mathbf{s}_p(\mathbf{x}') - \Delta^*_{\mathbf{x}'} k(\mathbf{x},\mathbf{x}')\right]\right] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\mathbf{s}_p(\mathbf{x})^\top k(\mathbf{x},\mathbf{x}')\,\mathbf{s}_p(\mathbf{x}') - \mathbf{s}_p(\mathbf{x})^\top \Delta^*_{\mathbf{x}'} k(\mathbf{x},\mathbf{x}') - \Delta^*_{\mathbf{x}} k(\mathbf{x},\mathbf{x}')^\top \mathbf{s}_p(\mathbf{x}') + \operatorname{tr}\left(\Delta^*_{\mathbf{x},\mathbf{x}'} k(\mathbf{x},\mathbf{x}')\right)\right] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\kappa_p(\mathbf{x},\mathbf{x}')\right],
\end{aligned}$$

which completes the proof. □

Theorem 11 (Adapted from Liu et al., 2016)**.** *Let $k(x,x')$ be a strictly positive definite kernel on $\mathcal{X}^d$, and assume that $\mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\kappa_p(\mathbf{x},\mathbf{x}')^2\right] < \infty$. We have the following two cases:*

(i) If $q \neq p$, then $\widehat{\mathbb{S}}(q \,\|\, p)$ is asymptotically Normal:

$$\sqrt{n}\left(\widehat{\mathbb{S}}(q \,\|\, p) - \mathbb{S}(q \,\|\, p)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma^2),$$

where $\sigma^2 = \operatorname{Var}_{\mathbf{x}\sim q}(\mathbb{E}_{\mathbf{x}'\sim q}\left[\kappa_p(\mathbf{x},\mathbf{x}')\right]) > 0$.

(ii) If $q = p$, then $\sigma^2 = 0$, and the U-statistic is degenerate:

$$n\,\widehat{\mathbb{S}}(q \,\|\, p) \xrightarrow{\mathcal{D}} \sum_{j=1}^{|\mathcal{X}|^d} c_j (Z_j^2 - 1),$$

where $\{Z_j\} \overset{iid}{\sim} \mathcal{N}(0,1)$ and $\{c_j\}$ are the eigenvalues of the kernel $\kappa_p(\cdot,\cdot)$ under q.

Lemma 12. *The exponentiated Hamming kernel*

$$k(\mathbf{x},\mathbf{x}') = \exp\{-H(\mathbf{x},\mathbf{x}')\},$$

where $H(\mathbf{x},\mathbf{x}') := \frac{1}{d}\sum_{i=1}^{d} \mathbb{I}\{x_i \neq x'_i\}$ is the normalized Hamming distance, is positive definite.

Proof. Without loss of generality, assume that $\mathcal{X} = \{0,1\}$ is a binary set; the general case can be easily accommodated by modifying the feature map to be described next. Define the feature map $\boldsymbol{\phi} : \mathcal{X}^d \to \mathcal{X}^{2d}$, $\mathbf{x} \mapsto \widetilde{\mathbf{x}}$, where $\widetilde{x}_{2i-1} = \mathbb{I}\{x_i = 0\}$ and $\widetilde{x}_{2i} = \mathbb{I}\{x_i = 1\}$ for $i = 1,\ldots,d$. Then, the normalized Hamming distance can be expressed as

$$H(\mathbf{x},\mathbf{x}') = 1 - \frac{1}{d}\sum_{i=1}^{d} \mathbb{I}\{x_i = x'_i\} = 1 - \frac{1}{2d}\sum_{j=1}^{2d} \widetilde{x}_j \widetilde{x}'_j = 1 - \frac{1}{2d}\widetilde{\mathbf{x}}^\top \widetilde{\mathbf{x}}' = 1 - \frac{1}{2d}\boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\phi}(\mathbf{x}').$$

Thus, $1 - H(\mathbf{x},\mathbf{x}')$ is a positive definite kernel. By Taylor expansion, $\exp\{1 - H(\mathbf{x},\mathbf{x}')\}$ (and hence $\exp\{-H(\mathbf{x},\mathbf{x}')\}$) also constitutes a positive definite kernel on $\mathcal{X}^d$. □