SUPPLEMENTARY MATERIAL

Gangzhen Qian[1] *Hong Zhou*[2] *Qi Zhang*[3,*]

[1]UC Berkeley [2]Peking University [3]Xi'an Jiaotong University

The original expression of Theorem 1 in the main text contains typographical errors (highlighted in red). We apologize for the oversight; the corrected version of Theorem 1, along with its complete proof, is provided in the supplementary material.

Theorem 1 (Empirical Estimator of $TC(U^1, \ldots, U^M)$)**.** *Given a mini-batch of N observations $\left\{\left(u_1^i, u_2^i, \ldots, u_M^i\right)\right\}_{i=1}^N$. Let $Q_k \in \mathbb{R}^{N \times N}$ denote the Gram matrix for the k-th $(1 \le k \le M)$ modality, i.e., $Q_k(i,j) = G_\sigma\left(u_k^i - u_k^j\right)$, in which $G_\sigma(\cdot) = \exp\left(-\frac{\|\cdot\|^2}{2\sigma^2}\right)$ refers to a Gaussian kernel of width σ. The empirical estimator of CS-TC is given by:*

$$
\begin{aligned}
\widehat{TC}(u_1, u_2, \ldots, u_M) &= \log\left(\frac{1}{N^2} \sum_{(i,j) \in \mathbf{i}_2^N} \prod_{k=1}^{M} Q_k(i,j)\right) \\
&+ \log\left(\frac{1}{N^{2M}} \sum_{(i_1, j_1, i_2, j_2, \cdots, i_M, j_M) \in \mathbf{i}_{2M}^N} \prod_{k=1}^{M} Q_k(i_k, j_k)\right) \\
&- 2\log\left(\frac{1}{N^{M+1}} \sum_{(i, j_1, j_2, \cdots, j_M) \in \mathbf{i}_{M+1}^N} \prod_{k=1}^{M} Q_k(i, j_k)\right),
\end{aligned}
\tag{1}
$$

where the index set $\mathbf{i}_r^N$ denotes the set of all r-tuples drawn with replacement from $\{1, 2, \cdots, N\}$.

Proof. The Cauchy-Schwarz divergence based total correlation (CS-TC) is defined as [1]:

$$
\begin{aligned}
TC_{\text{CS}}(u_1, u_2, \ldots, u_M) &:= D_{CS}\left(p(u_1, u_2, \ldots, u_M); \prod_{i=1}^{M} p(u_i)\right) \\
&= \log\left(\int p(u_1, u_2, \ldots, u_M)^2 \, du_1 du_2 \ldots du_M\right) + \log\left(\int \left(\prod_{i=1}^{M} p(u_i)\right)^2 du_1 du_2 \ldots du_M\right) \\
&- 2\log\left(\int p(u_1, u_2, \ldots, u_M) \prod_{i=1}^{M} p(u_i) du_1 du_2 \ldots du_M\right),
\end{aligned}
\tag{2}
$$

where D_{CS} denotes the CS divergence [1, 2], defined as:

$$
D_{CS}(p; q) = -\log\left(\frac{\left(\int p(x) q(x) \, dx\right)^2}{\int p(x)^2 \, dx \int q(x)^2 \, dx}\right).
\tag{3}
$$

The CS divergence formulation directly follows from the classic CS inequality for two probability density functions $p(x)$ and $q(x)$:

$$
\left(\int p(x) q(x) \, dx\right)^2 \le \int p(x)^2 \, dx \int q(x)^2 \, dx,
\tag{4}
$$

with equality if and only if $p(x)$ and $q(x)$ are linearly dependent.

Corresponding author: zhangqi@xjtu.edu.cn

Let us discuss the three terms inside the "log" of Eq. (2):

$$\begin{aligned}
\int p(u_1, u_2, \ldots, u_M)^2 \, du_1 du_2 \ldots du_M &= \mathbb{E}_{p(u_1, u_2, \ldots, u_M)} \left[p(u_1, u_2, \ldots, u_M) \right] \\
&= \frac{1}{N} \sum_{i=1}^{N} p(u_1^i, u_2^i, \ldots, u_M^i) \\
&= \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{N} \sum_{j=1}^{N} \kappa([u_1^i, u_2^i, \ldots, u_M^i]^T - [u_1^j, u_2^j, \ldots, u_M^j]^T) \right) \\
&= \frac{1}{N^2} \sum_{(i,j) \in \mathbf{i}_2^N} \kappa([u_1^i, u_2^i, \ldots, u_M^i]^T - [u_1^j, u_2^j, \ldots, u_M^j]^T) \\
&= \frac{1}{N^2} \sum_{(i,j) \in \mathbf{i}_2^N} \kappa(u_1^i - u_1^j) \kappa(u_2^i - u_2^j) \cdots \kappa(u_M^i - u_M^j) \\
&= \frac{1}{N^2} \sum_{(i,j) \in \mathbf{i}_2^N} \prod_{k=1}^{M} Q_k(i,j),
\end{aligned} \tag{5}$$

where the index set $\mathbf{i}_r^N$ denotes the set of all r-tuples drawn with replacement from $\{1, 2, \cdots, N\}$.

The third line in Eq. (5) follows from the kernel density estimation (KDE) formulation [3], where κ denotes a Gaussian kernel with bandwidth σ, expressed as $\kappa(x-y) = \exp\left(-\frac{|x-y|^2}{2\sigma^2}\right)$. The fifth line is derived under the assumption that the covariance matrix of $[u_1, u_2, \ldots, u_M]^\top$ is diagonal, a common simplification in KDE. Under this assumption, the multivariate kernel factorizes into a product of univariate kernels.

Similarly,

$$\begin{aligned}
&\int p(u_1, u_2, \ldots, u_M) p(u_1) p(u_2) \ldots p(u_M) \, du_1 du_2 \ldots du_M \\
&= \mathbb{E}_{p(u_1, u_2, \ldots, u_M)} \left[p(u_1) p(u_2) \ldots p(u_M) \right] \\
&= \frac{1}{N} \sum_{i=1}^{N} p(u_1^i) p(u_2^i) \ldots p(u_M^i) \\
&= \frac{1}{N} \sum_{i=1}^{N} \left[\left(\frac{1}{N} \sum_{j_1=1}^{N} \kappa(u_1^i - u_1^{j_1}) \right) \left(\frac{1}{N} \sum_{j_2=1}^{N} \kappa(u_2^i - u_2^{j_2}) \right) \cdots \left(\frac{1}{N} \sum_{j_M=1}^{N} \kappa(u_M^i - u_M^{j_M}) \right) \right] \\
&= \frac{1}{N^{M+1}} \sum_{(i, j_1, j_2, \cdots, j_M) \in \mathbf{i}_{M+1}^N} \prod_{k=1}^{M} Q_k(i, j_k),
\end{aligned} \tag{6}$$

and

$$\begin{aligned}
&\int \left(p(u_1) p(u_2) \ldots p(u_M) \right)^2 \, du_1 du_2 \ldots du_M \\
&= \int p^2(u_1) p^2(u_2) \ldots p^2(u_M) \, du_1 du_2 \ldots du_M \\
&= \left[\frac{1}{N^2} \sum_{i_1=1}^{N} \sum_{j_1=1}^{N} \kappa(u_1^{i_1} - u_1^{j_1}) \right] \left[\frac{1}{N^2} \sum_{i_2=1}^{N} \sum_{j_2=1}^{N} \kappa(u_2^{i_2} - u_2^{j_2}) \right] \cdots \left[\frac{1}{N^2} \sum_{i_M=1}^{N} \sum_{j_M=1}^{N} \kappa(s_M^{i_M} - s_M^{j_M}) \right] \\
&= \frac{1}{N^{2M}} \sum_{i_1=1}^{N} \sum_{j_1=1}^{N} \sum_{i_2=1}^{N} \sum_{j_2=1}^{N} \cdots \sum_{i_M=1}^{N} \sum_{j_M=1}^{N} \kappa(u_1^{i_1} - u_1^{j_1}) \kappa(u_2^{i_2} - u_2^{j_2}) \cdots \kappa(u_M^{i_M} - u_M^{j_M}) \\
&= \frac{1}{N^{2M}} \sum_{(i_1, j_1, i_2, j_2, \cdots, i_M, j_M) \in \mathbf{i}_{2M}^N} \prod_{k=1}^{M} Q_k(i_k, j_k).
\end{aligned} \tag{7}$$

Combining Eqs. (5)-(7), we obtain Eq. (1).

□

1. REFERENCES

[1] Jose C Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*, Springer Science & Business Media, 2010.

[2] Shujian Yu, Xi Yu, Sigurd Løkse, Robert Jenssen, and Jose C Principe, "Cauchy-schwarz divergence information bottleneck for regression," in *The Twelfth International Conference on Learning Representations*, 2024.

[3] Emanuel Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.