In *Estimating Mutual Information*, Kraskov, Stögbauer, and Grassberger (KSG) extended the classical nearest-neighbors Kozachenko-Leonenko (KL) entropy estimator by proposing make the number of nearest neighbors k dependent on each observation in a sequence $x_i, i = 1, \dots, N$. To do this, KSG utilized the max-norm distance from x_i to its k^{th} point instead of the euclidean distance. This results in the formation of a hyper-square with radius $\frac{\epsilon}{2}$. Then, the resulting volume V of the hyper-square can be simply found by taking the product of each side-length; that is, $V = \epsilon^d$, where d is the dimension of the hyper-square.

KSG propose a second algorithm based on hyper-rectangles based on the product of d hyper-squares. The d-dimensional hyper-rectangle is constructed by *constraining* the sides lengths of the hyper-cube until one of the k nearest-neighbors lies on the border or a corner between two borders of each side-length. This is equivalent to taking the maximum norm for each dimension of the k nearest-neighbors. Here, the volume becomes $V = \prod_{i=1}^{d} \epsilon_i$, where ϵ_i is the i^{th} side-length of the hyper-rectangle.

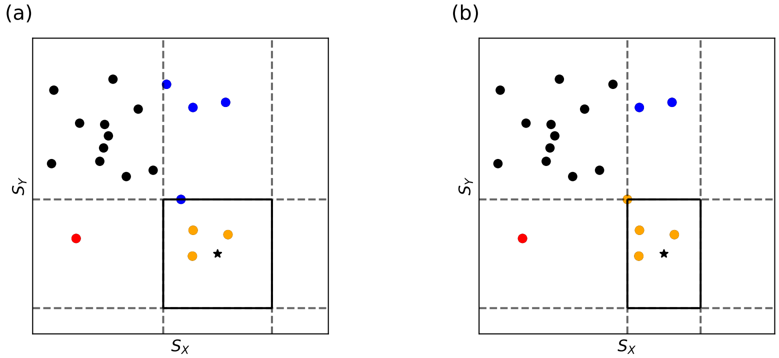

Figure 1: Nearest-neighboring distance strategies in 2-dimensional joint space $S_{(X,Y)}$. The star indicates the current reference point;Blue points indicate points present in the S_X sub-space, while red points indicate points present in the S_Y sub-space. Yellow points are points present in both the S_X and S_Y sub-spaces; that is, the joint space $S_{(X,Y)}$. **(a)** demonstrates the hyper-cube nearest-neighboring strategy. Note that border points are excluded. **(b)** demonstrates the hyper-rectangle nearest-neighboring strategy achieved by *constraining* the $\frac{\epsilon_X}{2}$ radius until 1 point lies on a corner or 2 points lie on a border. Here, border points are included. While it is possible to have more than 2 points present on the border, the probability of this happening in continuous space is 0.

We obtain 4 equations: 2 KL entropies $H(X)^{(1)}_{\text{KL}}$ and $H(X)^{(2)}_{\text{KL}}$ used to estimate the entropy in the full joint space $S_{(X_1,\dots,X_N)}$ for the hyper-square and hyper-rectangle neighboring strategies, respectively; and 2 KSG entropies $H(X)^{(1)}_{\text{KSG}}$ and $H(X)^{(2)}_{\text{KSG}}$ used to estimate marginal entropies from sub-spaces within the joint space for the respective neighboring strategy. They are as follows:

$$H(X)^{(1)}_{\text{KL}} = \psi(N) - \psi(k) + \langle \log(V_{X,i}) \rangle \tag{1}$$

$$H(X)^{(2)}_{\text{KL}} = \psi(N) - \psi(k) + \langle \log(V_{X,i}) \rangle + \frac{d_X - 1}{k} \tag{2}$$

$$H(X)^{(1)}_{\text{KSG}} = \psi(N) + \langle \log(V_{X,i}) - \psi(n_{X,i} + 1) \rangle \tag{3}$$

$$H(X)^{(2)}_{\text{KSG}} = \psi(N) + \left\langle \log(V_{X,i}) - \psi(n_{X,i}) + \frac{d_X - 1}{n_{X,i}} \right\rangle, \tag{4}$$

where $\psi(\cdot)$ is the digamma function, $\psi(x) = \frac{1}{\Gamma(x)} \frac{d\Gamma(x)}{dx}$, k is the number of nearest neighbors in the joint space, $n_{X,i}$ is the i^{th} number of nearest neighbors contained within the marginal space S_X, and $V_{X,i}$ is the i^{th} volume of the d_X-dimensional hyper-cube $^{(1)}$ or hyper-rectangle $^{(2)}$.

Using the chain rule, we can clearly see that the transfer entropy is a sum of 4 joint entropies:

$$\begin{aligned} TE_{Y \to X} &= I(X^p; X^- \mid Y^-) \\ &= H(X^p \mid X^-) - H(X^p \mid X^-, Y^-) \\ &= H(X^p, X^-) + H(X^-, Y^-) - H(X^p, X^-, Y^-) - H(X^-), \end{aligned} \tag{5}$$

where X^p is the present sequence of the destination variable X, X^- is the delayed sequence of the destination, and Y^- is the delayed sequence of the source variable Y. Thus, working from the joint space $S_{(X^p,X^-,Y^-)}$, we can estimate the transfer entropy in terms of one KN entropy (for the joint space) and three KSG entropies (for the marginal spaces $S_{(X^p,X^-)}$, $S_{(X^-,Y^-)}$, and S_{X^-}). The construction for the transfer entropy of the first KSG algorithm utilizing hyper-squares $TE^{(1)}_{Y\to X}$ proceeds as follows:

$$\begin{aligned}
TE^{(1)}_{Y\to X} &= H(X^-,Y^-)^{(1)}_{\text{KSG}} + H(X^p,X^-)^{(1)}_{\text{KSG}} - H(X^p,X^-,Y^-)^{(1)}_{\text{KL}} - H(X^-)^{(1)}_{\text{KSG}} \\
&= \psi(N) + \big\langle \log(V_{(X^-,Y^-),i}) - \psi(n_{(X^-,Y^-),i}+1)\big\rangle \\
&\quad + \psi(N) + \big\langle \log(V_{(X^p,X^-),i}) - \psi(n_{(X^p,X^-),i}+1)\big\rangle \\
&\quad - \psi(N) - \big\langle \log(V_{(X^p,X^-,Y^-),i}) - \psi(k)\big\rangle \\
&\quad - \psi(N) - \big\langle \log(V_{X^-,i}) - \psi(n_{X^-,i}+1)\big\rangle \\
&= \left\langle \log\left(\frac{V_{(X^-,Y^-),i}V_{(X^p,X^-),i}}{V_{(X^p,X^-,Y^-),i}V_{X^-,i}}\right)\right\rangle \\
&\quad + \big\langle \psi(k) + \psi(n_{X^-,i}+1) - \psi(n_{(X^-,Y^-),i}+1) - \psi(n_{(X^p,X^-),i}+1)\big\rangle \\
&\boxed{= \big\langle \psi(k) + \psi(n_{X^-,i}+1) - \psi(n_{(X^-,Y^-),i}+1) - \psi(n_{(X^p,X^-),i}+1)\big\rangle}.
\end{aligned} \tag{6}$$

In the same manner, we create a definition of the transfer entropy for the second KSG algorithm $TE^{(2)}_{Y\to X}$:

$$\begin{aligned}
TE^{(2)}_{Y\to X} &= H(X^-,Y^-)^{(2)}_{\text{KSG}} + H(X^p,X^-)^{(2)}_{\text{KSG}} - H(X^p,X^-,Y^-)^{(2)}_{\text{KL}} - H(X^-)^{(2)}_{\text{KSG}} \\
&= \psi(N) + \left\langle \log(V_{(X^-,Y^-),i}) - \psi(n_{(X^-,Y^-),i}) + \frac{d_{(X^-,Y^-)}-1}{n_{(X^-,Y^-),i}}\right\rangle \\
&\quad + \psi(N) + \left\langle \log(V_{(X^p,X^-),i}) - \psi(n_{(X^p,X^-),i}) + \frac{d_{(X^P,X^-)}-1}{n_{(X^p,X^-)},i}\right\rangle \\
&\quad - \psi(N) - \left\langle \log(V_{(X^p,X^-,Y^-),i}) - \psi(k) + \frac{d_{(X^p,X^-,Y^-)}-1}{k}\right\rangle \\
&\quad - \psi(N) - \left\langle \log(V_{X^-,i}) - \psi(n_{X^-,i}) + \frac{d_{X^-}-1}{n_{X^-,i}}\right\rangle \\
&= \left\langle \psi(k) - \frac{d_{(X^p,X^-,Y^-)}-1}{k} + \psi(n_{X^-,i}) - \frac{d_{X^-}-1}{n_{X^-,i}}\right. \\
&\qquad \left. - \psi(n_{(X^-,Y^-),i}) + \frac{d_{(X^-,Y^-)}-1}{n_{(X^-,Y^-),i}} - \psi(n_{(X^p,X^-),i}) + \frac{d_{(X^p,X^-)}-1}{n_{(X^p,X^-),i}}\right\rangle \\
&\boxed{= \left\langle \psi(k) - \frac{2}{k} + \psi(n_{X^-,i}) - \psi(n_{(X^-,Y^-),i}) + \frac{1}{n_{(X^-,Y^-),i}} - \psi(n_{(X^p,X^-),i}) + \frac{1}{n_{(X^p,X^-),i}}\right\rangle}.
\end{aligned} \tag{7}$$

To meaningfully compare information flows for different proteins or between residues of the same protein, it is important to account for the differing dynamics of each time series. We account for that here, by normalizing the KSG transfer entropy between 0 and 1 in a manner similar to min-max feature scaling:

$$x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \in [0,1], \tag{8}$$

where $x_{\min}$ and $x_{\max}$ represent the minimum and maximum values of the sequence $x_i, i = 1, \ldots, N$, and x'_i is the normalized mapping of the value x_i.

Theoretical minimum and maximum values for the transfer entropy can be solved for by erroneously optimizing the number of nearest neighbors in the joint space and the respective subspaces. For KSG transfer entropy algorithm 1, the number of nearest neighbors are bounded between $k-1$ and $N-2$ since border points are excluded. Thus, solving for the theoretical minimum transfer entropy $TE^{(1),\min}_{Y\to X}$, we obtain

$$\begin{aligned} TE^{(1),\min}_{Y\to X} &= \psi(k) + \psi(N-1) - \psi(N-1) - \psi(N-1) \\ &= \psi(k) - \psi(N-1). \end{aligned} \tag{9}$$

It is important to note, we plug in $N-1$ for the subspace S_{X^-}, since it must contain at-least as many points as found in the subspaces $S_{(X^p,X^-)}$ and $S_{(X^-,Y^-)}$. The maximum transfer entropy $TE^{(1),\max}_{Y\to X}$ is found by maximizing the subspace S_{X^-} and minimizing the subspaces $S_{(X^p,X^-)}$ and $S_{(X^-,Y^-)}$.

$$\begin{aligned} TE^{(1),\max}_{Y\to X} &= \psi(k) + \psi(N-1) - \psi(k) - \psi(k) \\ &= \psi(N-1) - \psi(k) \end{aligned} \tag{10}$$

Finally, we solve for the normalized transfer entropy for KSG algorithm 1 $NTE^{(1)}_{Y\to X}$:

$$NTE^{(1)}_{Y\to X} = \frac{TE^{(1)}_{Y\to X} + \psi(N-1) - \psi(k)}{2(\psi(N-1) - \psi(k))} \in [0,1]. \tag{11}$$

The normalized transfer entropy for KSG algorithm 2 is obtained in a similar manner by first solving for the maximum and minimum theoretical transfer entropies

$$\begin{aligned} TE^{(2),\min}_{Y\to X} &= \psi(k) - \frac{2}{k} + \psi(N-1) - \psi(N-1) + \frac{1}{N-1} - \psi(N-1) + \frac{1}{N-1} \\ &= \psi(k) - \frac{2}{k} - \psi(N-1) + \frac{2}{N-1} \end{aligned} \tag{12}$$

and

$$\begin{aligned} TE^{(2),\max}_{Y\to X} &= \psi(k) - \frac{2}{k} + \psi(N-1) - \psi(k) + \frac{1}{k} - \psi(k) + \frac{1}{k} \\ &= \psi(N-1) - \psi(k). \end{aligned} \tag{13}$$

By plugging the respective values into the min-max feature scaling formula, we obtain the normalized transfer entropy for KSG algorithm 2 $NTE^{(2)}_{Y\to X}$:

$$\begin{aligned} NTE^{(2)}_{Y\to X} &= \frac{TE_{Y\to X} + \psi(N-1) - \frac{2}{N-1} - \psi(k) + \frac{2}{k}}{\psi(N-1) - \psi(k) + \psi(N-1) - \frac{2}{N-1} - \psi(k) + \frac{2}{k}} \\ &= \frac{TE_{Y\to X} + \psi(N-1) - \frac{2}{N-1} - \psi(k) + \frac{2}{k}}{2(\psi(N-1) - \psi(k)) + \frac{2}{k} - \frac{2}{N-1}} \in [0,1]. \end{aligned} \tag{14}$$

Since entropy scales logarithmically, we propose to alter the normalization to a non-linear variant for each algorithm such that the values obtained are proportional to the percentage of uncertainty reduced.

$$\boxed{NTE^{(1)}_{Y\to X} = \frac{e^{TE^{(1)}_{Y\to X}} - e^{\psi(k)-\psi(N-1)}}{e^{\psi(N-1)-\psi(k)} - e^{\psi(k)-\psi(N-1)}} \in [0,1]} \tag{15}$$

$$\boxed{NTE^{(2)}_{Y\to X} = \frac{e^{TE^{(2)}_{Y\to X}} - e^{\psi(k)-\frac{2}{k}-\psi(N-1)+\frac{2}{N-1}}}{e^{\psi(N-1)-\psi(k)} - e^{\psi(k)-\frac{2}{k}-\psi(N-1)+\frac{2}{N-1}}} \in [0,1]} \tag{16}$$

The NTE is dependent upon the choice for k and the sample size N. Specifically, as k approaches N or N approaches k, the difference between $TE^{\min}_{Y\to X}$ and $TE^{\max}_{Y\to X}$ approaches 0.

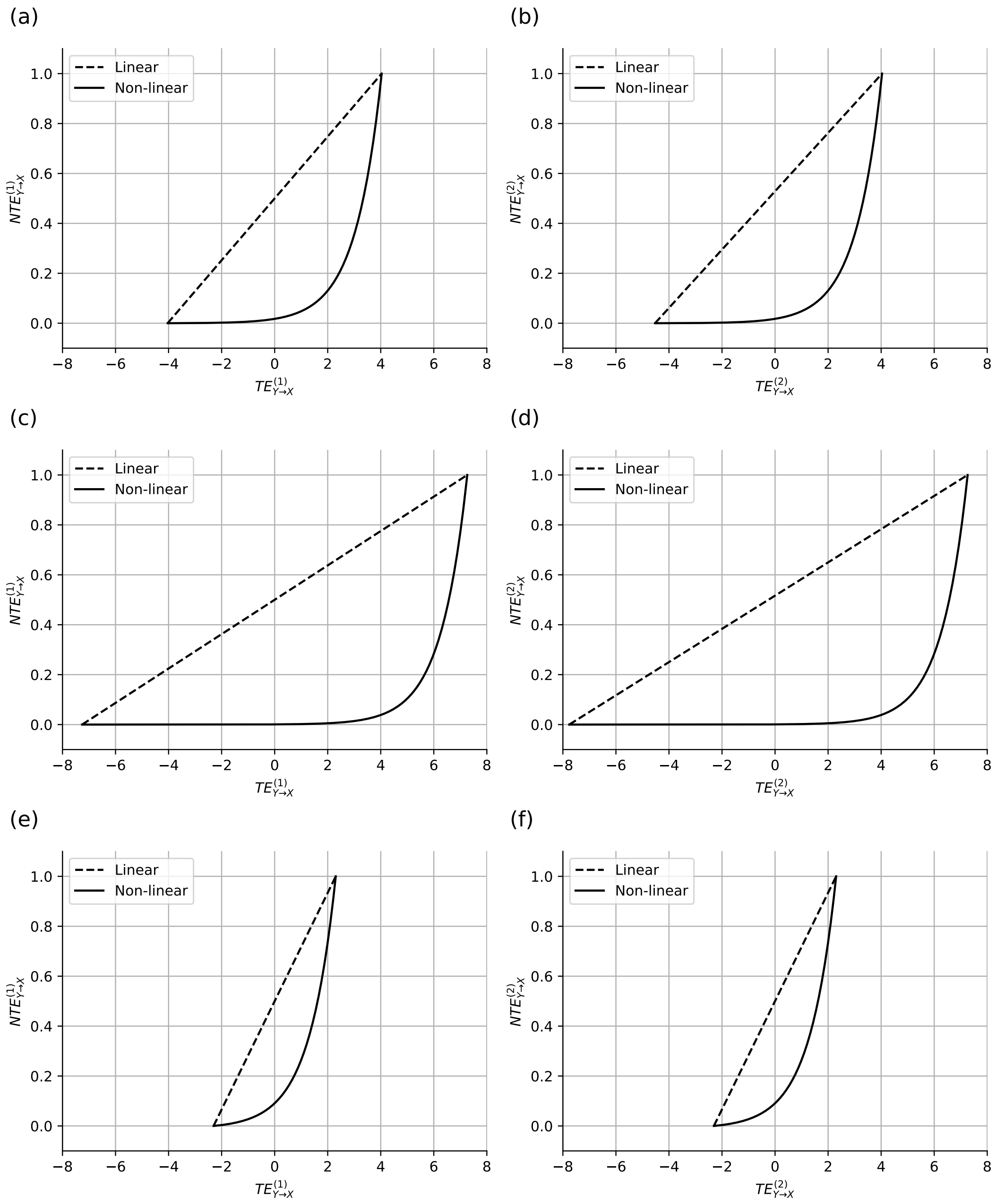

Figure 2: Normalized transfer entropy (NTE) displaying the linear and non-linear formulations. **(a, c, e)** utilize the first KSG algorithm; **(b, d, f)** utilize the second KSG algorithm. The figure shows the effect of the number of samples and number of nearest neighbors on the NTE: **(a, b)** $N = 200$, $k = 4$; **(c, d)** $N = 5000$, $k = 4$; **(e, f)** $N = 5000$, $k = 500$.

We also note the increased variance found in KSG algorithm 2. While $TE^{(1),\min}_{Y\to X}$ and $TE^{(1),\max}_{Y\to X}$ are equidistant from a $TE^{(1)}_{Y\to X}$ of 0, $TE^{(2),\min}_{Y\to X}$ is further from a $TE^{(2)}_{Y\to X}$ of 0 than $TE^{(2),\max}_{Y\to X}$. Thus, KSG algorithm 1 is best used when testing against a null hypothesis $H_0 : TE_{Y\to X} = 0$, while KSG algorithm 2, which displays decreased bias, is preferable when interpreting raw transfer entropy or normalized transfer entropy values. For this reason, we chose to use KSG algorithm 2 in the remaining calculations.