Exercise — SNE vs t-SNE on a 4-point toy example

Given

High-dimensional points (corners of unit square):

$$x_1 = (0,0),\ x_2 = (1,0),\ x_3 = (0,1),\ x_4 = (1,1).$$

Low-dimensional placement to evaluate:

$$y_1 = (0,0),\ y_2 = (1.5,0),\ y_3 = (0,1.5),\ y_4 = (1.5,1.5).$$

Use Gaussian kernel with $\sigma = 1$ for high-dim similarities. Compare two low-dim similarity models:

- SNE-style Gaussian (same $\sigma = 1$), and
- t-SNE Student-t (d.o.f. = 1) kernel: $\tilde{q}_{ij} = (1 + \|y_i - y_j\|^2)^{-1}$, normalized.

We will compute:

1. high-dim conditional probabilities $p_{j|i}$,
2. symmetrized ordered-pair p_{ij} (so $\sum_{i \neq j} p_{ij} = 1$),
3. low-dim $Q^{(G)}$ (Gaussian) and $Q^{(t)}$ (Student-t),
4. KL divergences $KL(P\|Q^{(G)})$ and $KL(P\|Q^{(t)})$,
5. interpretation.

Throughout let $N = 4$ (number of points).

Step 1 — High-dim squared distances $\|x_i - x_j\|^2$

Compute squared Euclidean distances (trivial geometry):

- adjacent (edge) pairs: $\|x_1 - x_2\|^2 = \|x_1 - x_3\|^2 = \cdots = 1.$
- diagonal pairs: $\|x_1 - x_4\|^2 = \|x_2 - x_3\|^2 = 2.$

Distance matrix (rows i, columns j):

$$D_{high} = \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix}.$$

Step 2 — Conditional high-dim probabilities $p_{j|i}$ (Gaussian, $\sigma = 1$)

Definition (for fixed i):

$$p_{j|i} = \frac{\exp\big(-\|x_i - x_j\|^2/(2\sigma^2)\big)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2/(2\sigma^2))}.$$

With $\sigma = 1$:

- for squared distance 1: exponent $e^{-1/2} = e^{-0.5}$.
- for squared distance 2: exponent e^{-1}.

Numeric values (rounded to 18 significant digits for clarity):

- $e^{-0.5} \approx 0.606530659712633424$.
- $e^{-1} \approx 0.367879441171442322$.

For any i the denominator (sum over three other points):

$$S = e^{-0.5} + e^{-0.5} + e^{-1} = 0.606530659712633424 + 0.606530659712633424 + 0.367879441171442322 = 1.580940760596709170.$$

Thus (same for every i by symmetry):

- adjacent conditional:

$$p_{adj|i} = \frac{e^{-0.5}}{S} = \frac{0.606530659712633424}{1.580940760596709170} \approx 0.383651731190550693.$$

- diagonal conditional:

$$p_{diag|i} = \frac{e^{-1}}{S} = \frac{0.367879441171442322}{1.580940760596709170} \approx 0.232696537618898613.$$

We can arrange the conditional matrix $P_{\text{cond}} = [p_{j|i}]$ (rows sum to 1), but the two numbers above are enough.

Step 3 — Symmetrize to get ordered-pair p_{ij} (so total sums to 1)

Common symmetrization used in SNE/t-SNE:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}.$$

Because the high-dim distances are symmetric, $p_{j|i} = p_{i|j}$. Therefore:

$$p_{ij} = \frac{p_{j|i}}{N}.$$

Thus (with $N = 4$):

- ordered adjacent pair:

$$p_{adj} = \frac{p_{adj|i}}{4} = \frac{0.383651731190550693}{4} \approx 0.095912932797637673.$$

- ordered diagonal pair:

$$p_{diag} = \frac{p_{diag|i}}{4} = \frac{0.232696537618898613}{4} \approx 0.058174134404724653.$$

Sanity check (sums to 1):

- There are 8 ordered adjacent pairs (each undirected edge counted twice) and 4 ordered diagonal pairs:

$$8 \cdot p_{adj} + 4 \cdot p_{diag} = 8(0.095912932797637673) + 4(0.058174134404724653) \approx 0.999999999999999999 \text{ (numerical rounding)} \approx 1.$$

So the ordered-pair probability matrix $P = [p_{ij}]$ (with zeros on diagonal) is fully defined by these two values.

Step 4 — Low-dim squared distances $\|y_i - y_j\|^2$

Given y coordinates:

- adjacent-like pairs: $\|y_1 - y_2\|^2 = 1.5^2 + 0^2 = 2.25.$
- diagonal pairs: $\|y_1 - y_4\|^2 = 1.5^2 + 1.5^2 = 4.5.$

Low-dim distance matrix:

$$D_{low} = \begin{pmatrix} 0 & 2.25 & 2.25 & 4.5 \\ 2.25 & 0 & 4.5 & 2.25 \\ 2.25 & 4.5 & 0 & 2.25 \\ 4.5 & 2.25 & 2.25 & 0 \end{pmatrix}.$$

Step 5 — SNE-style Gaussian low-dim conditional probabilities $q^{(G)}_{j|i}$ and symmetrized $q^{(G)}_{ij}$

Using Gaussian conditional with $\sigma = 1$:

Exponents:

- for squared distance 2.25: $\exp(-2.25/2) = \exp(-1.125) \approx 0.324652467358349730$.
- for squared distance 4.5: $\exp(-4.5/2) = \exp(-2.25) \approx 0.105399224561864337$.

Denominator per row:

$S_{low} = 0.324652467358349730 + 0.324652467358349730 + 0.105399224561864337 = 0.754704159278563797.$

Thus conditional values (per row):

- adjacent:

$$q^{(G)}_{adj|i} = \frac{0.324652467358349730}{0.754704159278563797} \approx 0.430171827420020131.$$

- diagonal:

$$q^{(G)}_{diag|i} = \frac{0.105399224561864337}{0.754704159278563797} \approx 0.139656345159959739.$$

Symmetrize to ordered-pair $q^{(G)}_{ij}$ exactly as for p:

$$q^{(G)}_{ij} = \frac{q^{(G)}_{j|i}}{N} \quad (\text{because } q_{j|i} = q_{i|j}).$$

So:

- ordered adjacent:

$$q^{(G)}_{adj} = \frac{0.430171827420020131}{4} \approx 0.107542956855005033.$$

- ordered diagonal:

$$q^{(G)}_{diag} = \frac{0.139656345159959739}{4} \approx 0.034914086289989935.$$

Check sum: $8 \cdot q^{(G)}_{adj} + 4 \cdot q^{(G)}_{diag} \approx 1$.

Step 6 — t-SNE Student-t low-dim probabilities $q^{(t)}_{ij}$

t-SNE numerator for ordered pair (unnormalized):

$$\tilde{q}_{ij} = \frac{1}{1 + \|y_i - y_j\|^2}.$$

Compute numerators:

- for $d^2 = 2.25$: $\tilde{q}_{adj} = 1/(1+2.25) = 1/3.25 = 0.307692307692307692.$
- for $d^2 = 4.5$: $\tilde{q}_{diag} = 1/(1+4.5) = 1/5.5 = 0.181818181818181818.$

Normalization constant Z sums these over all ordered pairs:

- 8 ordered adjacent pairs contribute $8 \times 0.307692307692307692$.
- 4 ordered diagonal pairs contribute $4 \times 0.181818181818181818$.

So

$$Z = 8 \cdot 0.307692307692307692 + 4 \cdot 0.181818181818181818 \approx 3.188811188811188811.$$

Finally normalized ordered-pair probabilities:

- adjacent:

$$q_{adj}^{(t)} = \frac{0.307692307692307692}{3.188811188811188811} \approx 0.096491228070175439.$$

- diagonal:

$$q_{diag}^{(t)} = \frac{0.181818181818181818}{3.188811188811188811} \approx 0.057017543859649123.$$

Check: $8 \cdot q_{adj}^{(t)} + 4 \cdot q_{diag}^{(t)} \approx 1.$

Step 7 — Compute KL divergences $KL(P\|Q)$

KL divergence (ordered-pairs):

$$KL(P\|Q) = \sum_{i \neq j} p_{ij} \ln \frac{p_{ij}}{q_{ij}}.$$

Because there are only two distinct types of pairs (adjacent and diagonal), group terms:

- 8 ordered adjacent pairs: each contributes $p_{adj} \ln(p_{adj}/q_{adj})$.
- 4 ordered diagonal pairs: each contributes $p_{diag} \ln(p_{diag}/q_{diag})$.

So

$$KL = 8 \cdot p_{adj} \ln \frac{p_{adj}}{q_{adj}} + 4 \cdot p_{diag} \ln \frac{p_{diag}}{q_{diag}}.$$

We compute two KLs: with $Q^{(G)}$ and with $Q^{(t)}$. Use values from above.

7A — KL with Gaussian low-dim $Q^{(G)}$

Plug numbers (keep many digits):

- $p_{adj} = 0.095912932797637673$
- $q_{adj}^{(G)} = 0.107542956855005033$
- $p_{diag} = 0.058174134404724653$
- $q_{diag}^{(G)} = 0.034914086289989935$

Compute the two grouped terms separately.

1. Adjacent term:

$$T_{adj} = 8 \cdot p_{adj} \cdot \ln \frac{p_{adj}}{q_{adj}^{(G)}}.$$

First compute ratio and log:

$$\frac{p_{adj}}{q^{(G)}_{adj}} = \frac{0.095912932797637673}{0.107542956855005033} \approx 0.892001.$$

$$\ln\left(\frac{p_{adj}}{q^{(G)}_{adj}}\right) \approx \ln(0.892001) \approx -0.114385.$$

Now multiply:

$$T_{adj} \approx 8 \times 0.095912932797637673 \times (-0.114385) \approx -0.087428.$$

(negative because $p_{adj} < q^{(G)}_{adj}$.)

2. Diagonal term:

$$T_{diag} = 4 \cdot p_{diag} \cdot \ln\frac{p_{diag}}{q^{(G)}_{diag}}.$$

Ratio:

$$\frac{p_{diag}}{q^{(G)}_{diag}} = \frac{0.058174134404724653}{0.034914086289989935} \approx 1.666666.$$

$$\ln(1.666666) \approx 0.510826.$$

Multiply:

$$T_{diag} \approx 4 \times 0.058174134404724653 \times 0.510826 \approx 0.118414.$$

Add terms:

$$KL(P\|Q^{(G)}) \approx T_{adj} + T_{diag} \approx -0.087428 + 0.118414 = 0.030986.$$

Accurate numeric (more precision) result:

$$\boxed{KL(P\|Q^{(G)}) \approx 0.03098579952637548.}$$

7B — KL with Student-t low-dim $Q^{(t)}$

Plug numbers:

- $q^{(t)}_{adj} \approx 0.096491228070175439$
- $q^{(t)}_{diag} \approx 0.057017543859649123$

Compute grouped terms.

1. Adjacent:

$$\frac{p_{adj}}{q^{(t)}_{adj}} = \frac{0.095912932797637673}{0.096491228070175439} \approx 0.994119.$$

$$\ln(0.994119) \approx -0.005896.$$

$$T_{adj} = 8 \times 0.095912932797637673 \times (-0.005896) \approx -0.004523.$$

2. Diagonal:

$$\frac{p_{diag}}{q^{(t)}_{diag}} = \frac{0.058174134404724653}{0.057017543859649123} \approx 1.020207.$$

$$\ln(1.020207) \approx 0.020006.$$

$$T_{diag} = 4 \times 0.058174134404724653 \times 0.020006 \approx 0.004584.$$

Add:

$$KL(P\|Q^{(t)}) \approx -0.004523 + 0.004584 = 0.000061.$$

Accurate numeric:

$$\boxed{KL(P\|Q^{(t)}) \approx 0.00006049960013897914 \ (\approx 6.05 \times 10^{-5}).}$$

Step 8 — Final numeric summary (rounded reasonably)

- $p_{adj} = 0.095912932797637673$
- $p_{diag} = 0.058174134404724653$
- Gaussian low-dim:
  - $q_{adj}^{(G)} = 0.107542956855005033$
  - $q_{diag}^{(G)} = 0.034914086289989935$
  - $KL(P\|Q^{(G)}) \approx 0.03098579952637548$
- Student-t low-dim:
  - $q_{adj}^{(t)} = 0.096491228070175439$
  - $q_{diag}^{(t)} = 0.057017543859649123$
  - $KL(P\|Q^{(t)}) \approx 0.00006049960013897914$

Step 9 — Clear interpretation (why these numbers matter)

- The KL divergence of the t-distribution case is **much smaller** than the Gaussian low-dim case:
  $KL(P\|Q^{(t)}) \ll KL(P\|Q^{(G)})$.
- This shows that, for the chosen low-dim layout Y, the **Student-t kernel** produces similarity values q_{ij} that match the high-dim neighborhood distribution p_{ij} **far better** than a Gaussian low-dim kernel does.
- Intuition: the Student-t kernel has **heavy tails**, which allows points moderately far apart in the 2-D map to still have non-negligible similarity. This **reduces crowding** in low dimensions and improves the match to high-dim local neighborhoods — exactly the motivation for t-SNE.

how u calculate high dim Distance matrix (rows i, columns j):

$$D_{high} = \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix}.$$

D
high

=

0
1
1

Given $x_1 = (0,0),\ x_2 = (1,0),\ x_3 = (0,1),\ x_4 = (1,1)$. Compute squared Euclidean distance $d^2(x_i, x_j) = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2$.

Calculations (each entry $D_{ij} = d^2(x_i, x_j)$):

- $D_{11} = (0-0)^2 + (0-0)^2 = 0$
- $D_{12} = (0-1)^2 + (0-0)^2 = (-1)^2 + 0 = 1$
- $D_{13} = (0-0)^2 + (0-1)^2 = 0 + (-1)^2 = 1$
- $D_{14} = (0-1)^2 + (0-1)^2 = 1 + 1 = 2$
- $D_{21} = (1-0)^2 + (0-0)^2 = 1 + 0 = 1$
- $D_{22} = (1-1)^2 + (0-0)^2 = 0$
- $D_{23} = (1-0)^2 + (0-1)^2 = 1 + 1 = 2$
- $D_{24} = (1-1)^2 + (0-1)^2 = 0 + 1 = 1$
- $D_{31} = (0-0)^2 + (1-0)^2 = 0 + 1 = 1$
- $D_{32} = (0-1)^2 + (1-0)^2 = 1 + 1 = 2$
- $D_{33} = (0-0)^2 + (1-1)^2 = 0$
- $D_{34} = (0-1)^2 + (1-1)^2 = 1 + 0 = 1$
- $D_{41} = (1-0)^2 + (1-0)^2 = 1 + 1 = 2$
- $D_{42} = (1-1)^2 + (1-0)^2 = 0 + 1 = 1$
- $D_{43} = (1-0)^2 + (1-1)^2 = 1 + 0 = 1$
- $D_{44} = (1-1)^2 + (1-1)^2 = 0$

Thus the matrix is

$$D_{\text{high}} = \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix}.$$