UNIVERSITY of WASHINGTON

Extending t-SNE

MEGAN MORRISON AND BENJAMIN LIU
DEPARTMENT OF APPLIED MATHEMATICS, UNIVERSITY OF WASHINGTON

OVERVIEW OF T-SNE

t-distributed Stochastic Neighbor Embedding (t-SNE) is a dimension reduction technique for data visualization.

The similarity of N data points in the origin space is first encoded in a probability distribution P. N new points in the low-dimensional embedding space are then randomly distributed and their similarity Q is calculated. Through gradient descent, the positions of these points are adjusted to bring P and Q into as close agreement as possible.

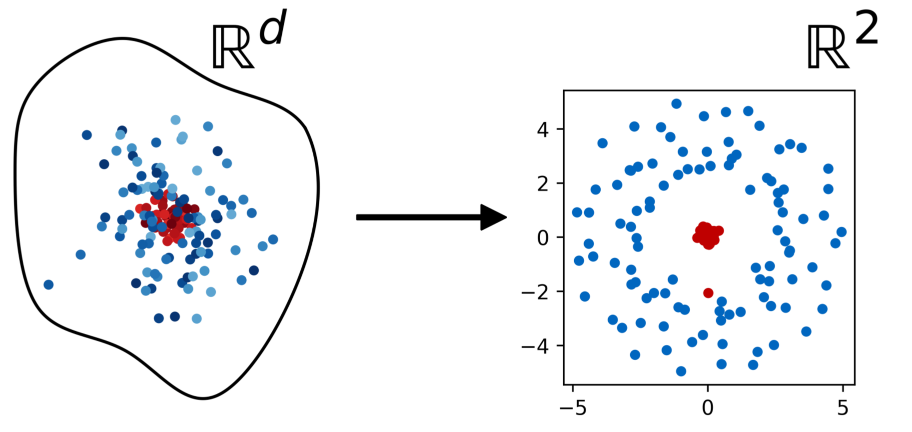

Figure 1. Illustration of t-SNE. A representation of data in a high-dimensional space $\mathbb{R}^d$ is created in a low-dimensional space such as $\mathbb{R}^2$.

We show several of variants of the t-SNE algorithm that aim to produce more accurate representations of the data and improve the quality of clusters.

JOINT AND CONDITIONAL DISTRIBUTIONS

t-SNE begins by defining a family of conditional distributions P_i. The probability $p_{j|i}$ measures how similar or related point x_j is to point x_i,

$$p_{j|i} = \frac{\exp(-\|x_j - x_i\|_2^2/\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_k - x_i\|_2^2/\sigma_i^2)}.$$

Each σ_i measures the scale of the neighborhood of point x_i. Tightly packed points are assigned smaller values and more loosely packed points are given larger values.

The probabilities $p_{j|i}$ and $p_{i|j}$ are not, in general, equal. A single symmetric joint distribution P is defined by

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

To measure similarity in the embedding space, a joint distribution Q is defined directly,

$$q_{i,j} = \frac{(1 + \|x_k - x_l\|_2^2)^{-1}}{\sum_{k\neq l}(1 + \|x_k - x_l\|_2^2)^{-1}}.$$

We propose a variant of t-SNE called conditional—σ_i that modifies the definition of Q to be defined by a conditional distribution.

Figure 2. shows snapshots of t-SNE embeddings for a data set consisting of three nested Gaussian-distributed clouds of data. The three clouds have differing variances and are successfully separated by t-SNE and the conditional—σ_i variant, but the variant does so by rapidly producing a transient clustering that distinguishing the three groups of points in a colinear arrangement.

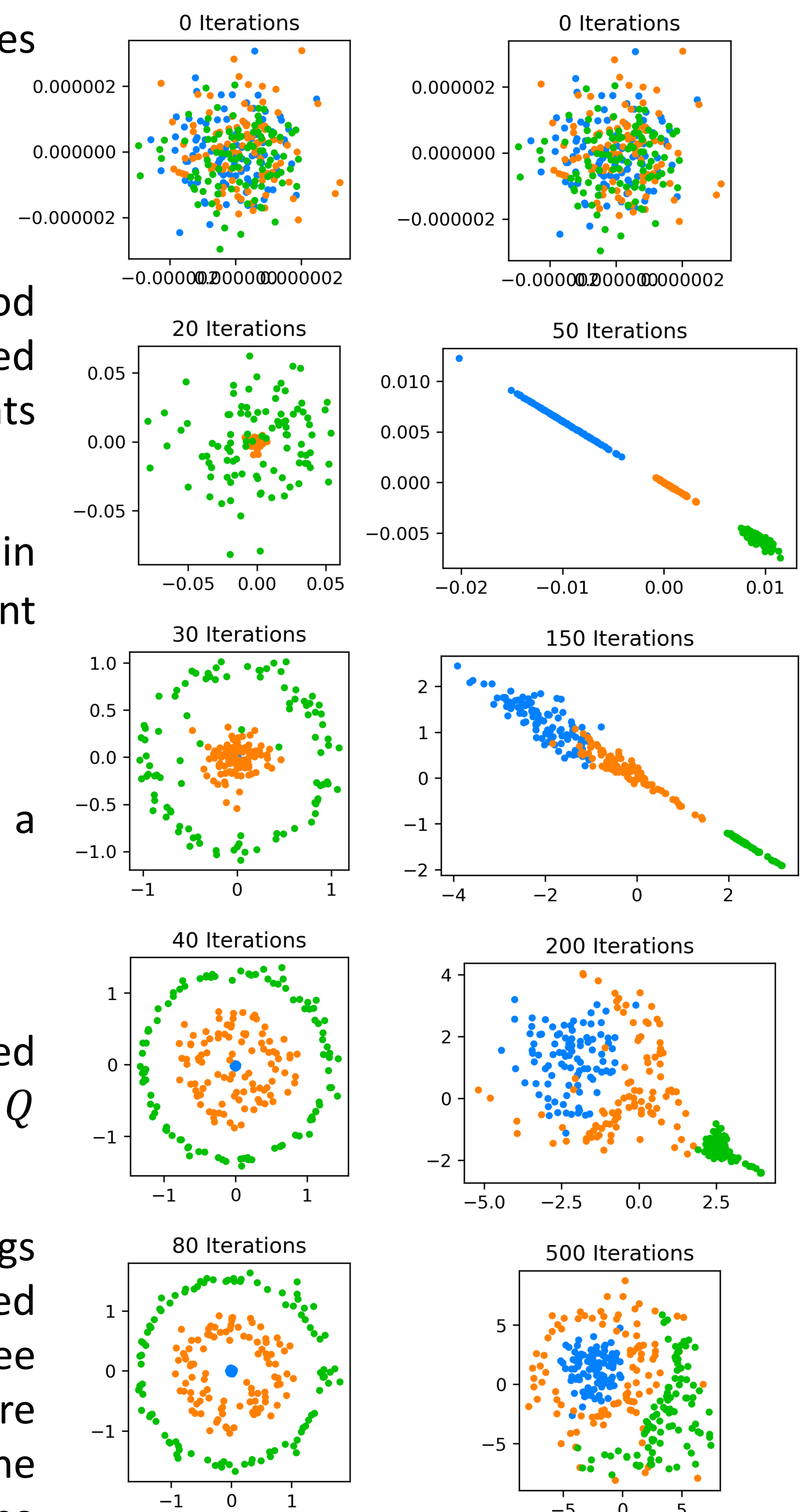

Figure 2. Evolution of embeddings for standard t-SNE (left) and the conditional--σ_i variant (right). The variant rapidly produces transient clusters that distinguish features in the data better than standard t-SNE.

TAIL FATNESS

We propose a variant of t-SNE called fat-tailed t-SNE that introduces a parameter ν_{target} that controls the degree of tail fatness of the distribution used to measure similarity in the embedding space. Figure 4 shows sample embeddings for fat-tailed t-SNE, and shows that values of $\nu_{target} < 1$ can produce clusters that are superior to standard t-SNE.

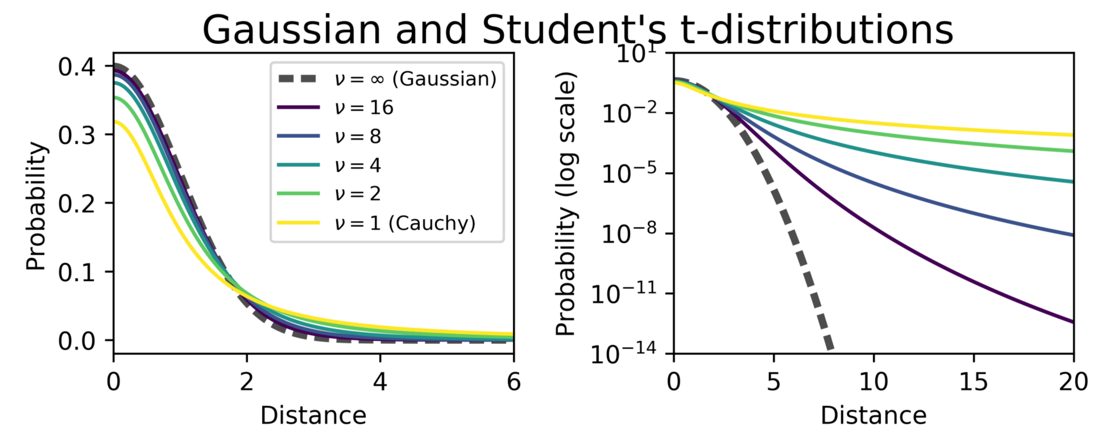

Figure 3. Comparison of Gaussian and Student's t-distributions.

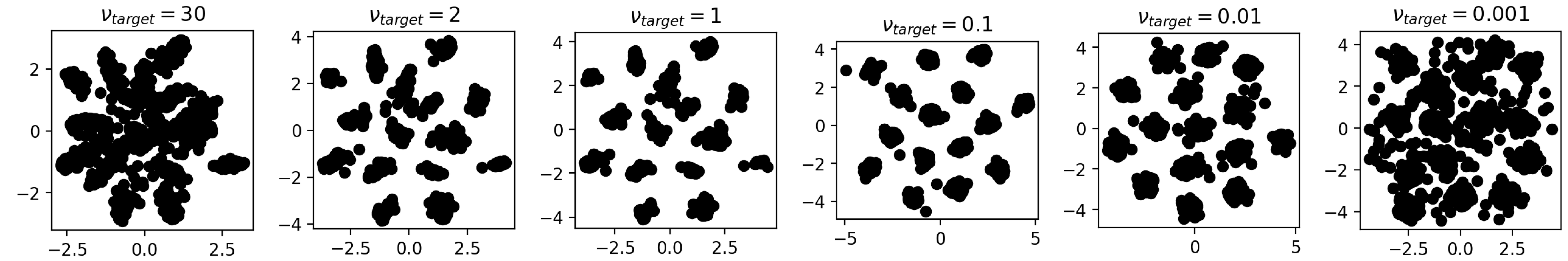

Figure 4. t-SNE embeddings for different values of ν_{target}. Standard t-SNE ($\nu_{target} = 1$) performs well, but not best.

MNIST

Figure 5 shows how fat-tailed t-SNE generates tighter clusters than standard t-SNE for a subset of MNIST, leading to higher classification accuracy when using k-means clustering to classify the low-dimensional data.

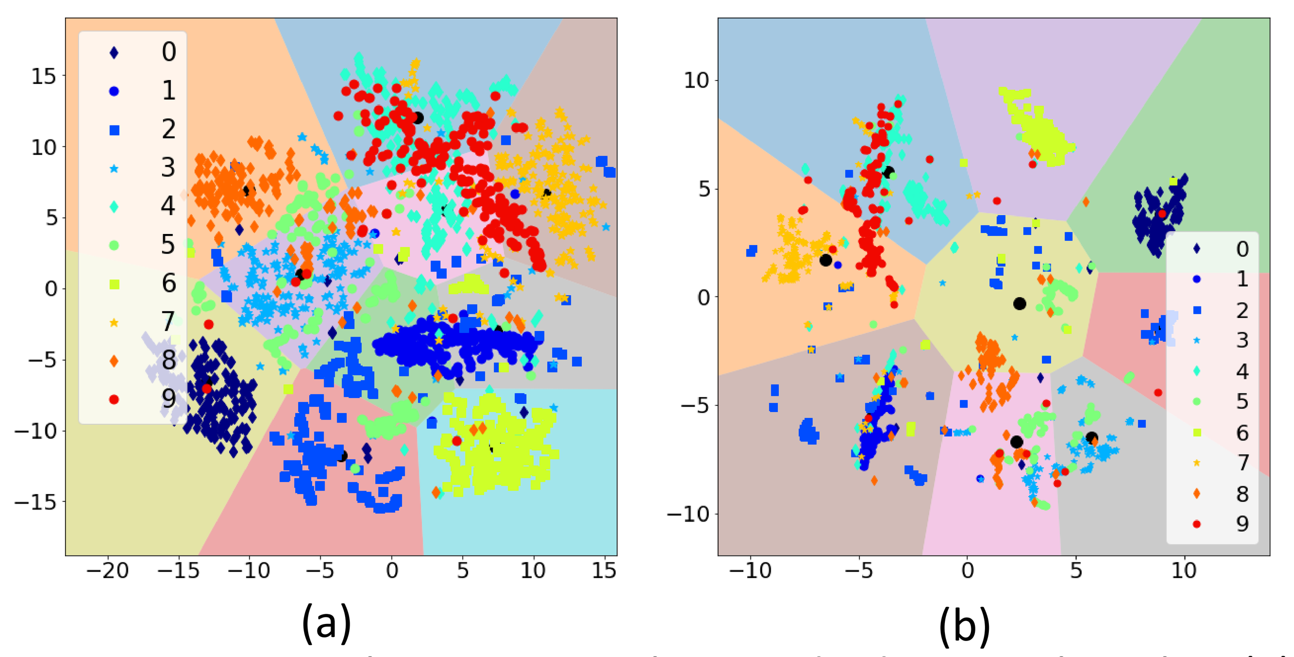

Figure 5. MNIST clusters using the standard t-SNE algorithm (a), and the fat-tailed t-SNE algorithm with $\nu_{target} = 0.1$ (b).

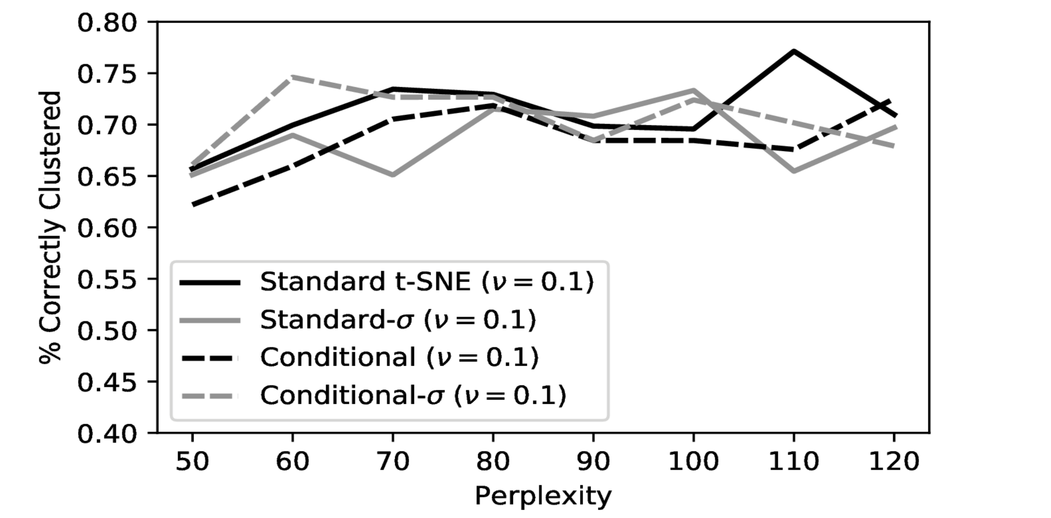

Figure 6. K-means classification accuracy as a function of perplexity, using four variations of the fat-tailed t-SNE algorithm. Although each of these algorithms generates different cluster shapes, the resulting classification accuracy is similar for all modifications.

BLACK FRIDAY

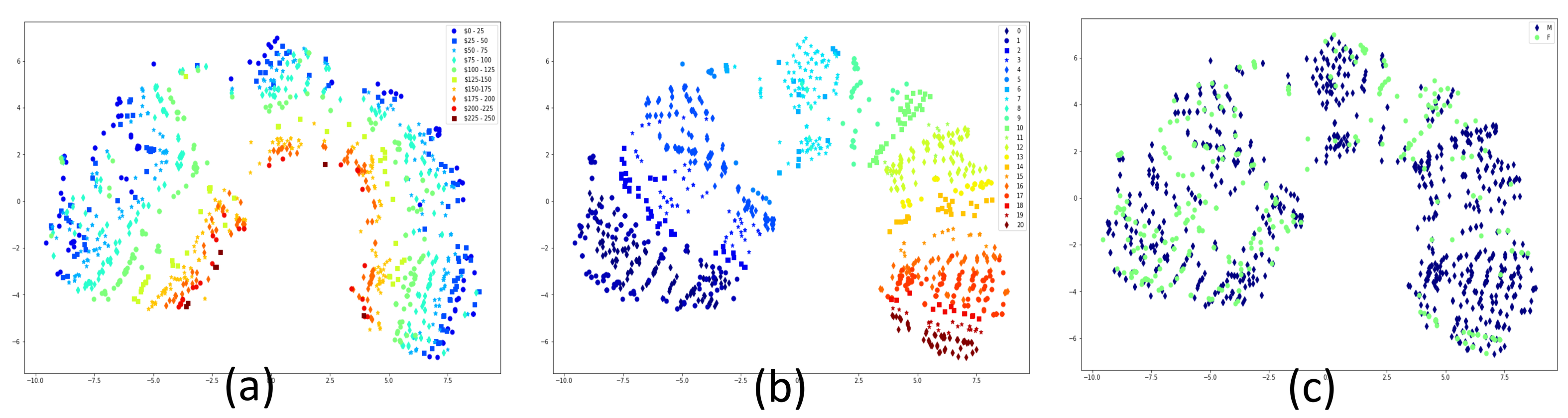

Figure 7. Black Friday shopper data differentiates according to amount spent (a), and occupation level (b) in the low-dimensional space but not gender (c). Shopper demographics data failed to produce meaningful clusters due to categorical features in the dataset.

CONCLUSION

- Variants of the t-SNE algorithm can generate cleaner clusters than standard t-SNE
- Clusters are maximally visually separable when using the t-SNE conditional—σ_i variant
- t-SNE and variations of this algorithm generated meaningful clusters when applied to the high-dimensional MNSIT dataset which we used to generate low-dimensional digit regions.
- t-SNE did not generate meaningful clusters when applied to a low-dimensional dataset with non-numerical features, perhaps because distances between labels of non-numerical features do not carry any intrinsic meaning.