Supplementary Materials: Alpha-Beta Divergences Discover Micro and Macro Structures in Data

We first discuss details regarding continuity in Alpha-Beta SNE which allow us to make settings, e.g., $(\alpha, \alpha+\beta) = (1, 1)$ used in t-SNE, while bypassing issues arising from singularity. We then display a smattering of ABSNE plots across a large selection of datasets.

1. Continuity and Gradients for the Alpha-Beta Objective

The form of the original Alpha-Beta divergence used in the paper, defined $\mathcal{J}(\mathcal{E}; \alpha, \beta) = D_{AB}^{\alpha\beta}(\mathbf{P}\|\mathbf{Q})$, is computed as:

$$\frac{1}{\alpha\beta} \sum_{i \neq j} \left(-\mathbf{P}_{ij}^{\alpha} \mathbf{Q}_{ij}^{\beta} + \frac{\alpha}{\alpha+\beta} \mathbf{P}_{ij}^{\alpha+\beta} + \frac{\beta}{\alpha+\beta} \mathbf{Q}_{ij}^{\alpha+\beta} \right), \tag{1}$$

where $\alpha \in \mathbb{R} \setminus \{0\}, \beta \in \mathbb{R}$ are hyperparameters. According to (Cichocki et al., 2011), to account for cases such as $\beta = 0$ and $\alpha + \beta = 0$ in this objective, we can re-define:

$$D_{AB}^{\alpha\beta}(\mathbf{P}\|\mathbf{Q}) = \sum_{ij} d_{AB}^{(\alpha,\beta)}(\mathbf{P}_{ij}, \mathbf{Q}_{ij}), \tag{2}$$

where:

$$d_{AB}^{(\alpha,\beta)}(\mathbf{P}_{ij}, \mathbf{Q}_{ij}) = \begin{cases} -\dfrac{1}{\alpha\beta} \left(\mathbf{P}_{ij}^{\alpha} \mathbf{Q}_{ij}^{\beta} - \dfrac{\alpha}{\alpha+\beta} \mathbf{P}_{ij}^{\alpha+\beta} - \dfrac{\alpha}{\alpha+\beta} \mathbf{Q}_{ij}^{\alpha+\beta} \right), & \alpha, \beta, \alpha+\beta \neq 0 \\ \dfrac{1}{\alpha^2} \left(\mathbf{P}_{ij}^{\alpha} \ln \dfrac{\mathbf{P}_{ij}^{\alpha}}{\mathbf{Q}_{ij}^{\alpha}} - \mathbf{P}_{ij}^{\alpha} + \mathbf{Q}_{ij}^{\alpha} \right), & \alpha \neq 0, \beta = 0 \\ \dfrac{1}{\alpha^2} \left(\ln \dfrac{\mathbf{Q}_{ij}^{\alpha}}{\mathbf{P}_{ij}^{\alpha}} + \dfrac{\mathbf{P}_{ij}^{\alpha}}{\mathbf{Q}_{ij}^{\alpha}} - 1 \right), & \alpha = -\beta \neq 0 \\ \dfrac{1}{\beta^2} \left(\mathbf{Q}_{ij}^{\beta} \ln \dfrac{\mathbf{Q}_{ij}^{\beta}}{\mathbf{P}_{ij}^{\beta}} + \mathbf{P}_{ij}^{\beta} \right), & \alpha = 0, \beta \neq 0 \\ \dfrac{1}{2} (\ln \mathbf{P}_{ij} - \ln \mathbf{Q}_{ij})^2, & \alpha = \beta = 0 \end{cases} \tag{3}$$

As described in the paper, note that we obtain the KL-divergence by setting $\alpha = 1, \beta = 0$ and the Itakura-Saito divergence by setting $\alpha = 1, \beta = -1$, since $\mathbf{P}$ and $\mathbf{Q}$ are probability distributions. Given these various definitions, we would like to ensure that the gradient described in the paper is the same particularly for the first three cases (namely where $\alpha \neq 0$), as these are the ones that arise in our reductions. Specifically, we need to ensure that the gradients $\partial D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) / \partial \mathbf{y}_i$ all match.

Case 1: $\alpha, \beta \neq 0$. We have that:

$$\begin{aligned}
\frac{\partial D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})}{\partial \mathbf{y}_i} &= \frac{\partial}{\partial \mathbf{y}_i}\left[-\frac{1}{\alpha\beta}\left(\mathbf{P}_{ij}^{\alpha}\mathbf{Q}_{ij}^{\beta} - \frac{\alpha}{\alpha+\beta}\mathbf{P}_{ij}^{\alpha+\beta} - \frac{\beta}{\alpha+\beta}\mathbf{Q}_{ij}^{\alpha+\beta}\right)\right] && (4)\\
&= \frac{\partial}{\partial \mathbf{y}_i}\left[-\frac{1}{\alpha\beta}\mathbf{P}_{ij}^{\alpha}\mathbf{Q}_{ij}^{\beta} + \frac{1}{\alpha(\alpha+\beta)}\mathbf{Q}_{ij}^{\alpha+\beta}\right] && (5)\\
&= \frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i}\cdot\left[-\frac{1}{\alpha\beta}\mathbf{P}_{ij}^{\alpha}\cdot\beta\cdot\mathbf{Q}_{ij}^{\beta-1} + \frac{1}{\alpha(\alpha+\beta)}\cdot(\alpha+\beta)\cdot\mathbf{Q}_{ij}^{\alpha+\beta-1}\right] && (6)\\
&= \frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i}\cdot\left[-\frac{1}{\alpha}\mathbf{P}_{ij}^{\alpha}\cdot\mathbf{Q}_{ij}^{\beta-1} + \frac{1}{\alpha}\cdot\mathbf{Q}_{ij}^{\alpha+\beta-1}\right] && (7)\\
&= -\frac{1}{\alpha}\cdot\frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i}\cdot\left[\mathbf{P}_{ij}^{\alpha}\mathbf{Q}_{ij}^{\beta-1} - \mathbf{Q}_{ij}^{\alpha+\beta-1}\right] && (8)
\end{aligned}$$

In the remainder of the cases, we set β to the appropriate value to demonstrate that the gradients match.

Case 2: $\alpha \neq 0, \beta = 0$. We have that:

$$\begin{aligned}
\frac{\partial D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})}{\partial \mathbf{y}_i} &= \frac{\partial}{\partial \mathbf{y}_i}\left[\frac{1}{\alpha^2}\left(\mathbf{P}_{ij}^{\alpha}\ln\frac{\mathbf{P}_{ij}^{\alpha}}{\mathbf{Q}_{ij}^{\alpha}} - \mathbf{P}_{ij}^{\alpha} + \mathbf{Q}_{ij}^{\alpha}\right)\right] && (9)\\
&= \frac{\partial}{\partial \mathbf{y}_i}\left[-\frac{1}{\alpha^2}\mathbf{P}_{ij}^{\alpha}\ln\mathbf{Q}_{ij}^{\alpha} + \frac{1}{\alpha^2}\mathbf{Q}_{ij}^{\alpha}\right] && (10)\\
&= \frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i}\left[-\frac{1}{\alpha^2}\mathbf{P}_{ij}^{\alpha}\cdot\alpha\mathbf{Q}_{ij}^{\alpha-1}\cdot\frac{1}{\mathbf{Q}_{ij}^{\alpha}} + \frac{1}{\alpha}\mathbf{Q}_{ij}^{\alpha-1}\right] && (11)\\
&= \frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i}\left[-\frac{1}{\alpha}\frac{\mathbf{P}_{ij}^{\alpha}}{\mathbf{Q}_{ij}^{\alpha}} + \frac{1}{\alpha}\mathbf{Q}_{ij}^{\alpha-1}\right] && (12)\\
&= -\frac{1}{\alpha}\cdot\frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i}\cdot\left[\frac{\mathbf{P}_{ij}^{\alpha}}{\mathbf{Q}_{ij}^{\alpha}} - \mathbf{Q}_{ij}^{\alpha-1}\right] && (13)
\end{aligned}$$

Setting $\beta = 0$ in Equation (8), we observe that the gradients match, as desired.

Case 3: $\alpha = -\beta \neq 0$. We have that:

$$\begin{aligned}
\frac{\partial D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})}{\partial \mathbf{y}_i} &= \frac{\partial}{\partial \mathbf{y}_i}\left[\frac{1}{\alpha^2}\left(\ln\frac{\mathbf{Q}_{ij}^{\alpha}}{\mathbf{P}_{ij}^{\alpha}} + \frac{\mathbf{P}_{ij}^{\alpha}}{\mathbf{Q}_{ij}^{\alpha}} - 1\right)\right] && (14)\\
&= \frac{\partial}{\partial \mathbf{y}_i}\left[\frac{1}{\alpha^2}\left(\ln\mathbf{Q}_{ij}^{\alpha} + \frac{\mathbf{P}_{ij}^{\alpha}}{\mathbf{Q}_{ij}^{\alpha}}\right)\right] && (15)\\
&= \frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i}\cdot\left[\frac{1}{\alpha^2}\cdot\alpha\mathbf{Q}_{ij}^{\alpha-1}\cdot\frac{1}{\mathbf{Q}_{ij}^{\alpha}} + \frac{1}{\alpha^2}\cdot\mathbf{P}_{ij}^{\alpha}\cdot -\alpha\cdot\frac{1}{\mathbf{Q}_{ij}^{\alpha+1}}\right] && (16)\\
&= -\frac{1}{\alpha}\cdot\frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i}\cdot\left[\frac{\mathbf{P}_{ij}^{\alpha}}{\mathbf{Q}_{ij}^{\alpha+1}} - \frac{1}{\mathbf{Q}_{ij}}\right] && (17)
\end{aligned}$$

Setting $\beta = -\alpha$ in Equation (8), we again observe that the gradients match, as desired. It follows that for all cases in the paper that we explore, we can simply use the gradient in Equation (8).

2. Embedding Plots of Various Datasets

In Figures 1, 2 and 3 we showcase visualizations of datasets on which we presented some results in the main paper, but did not include plots due to lack of space.

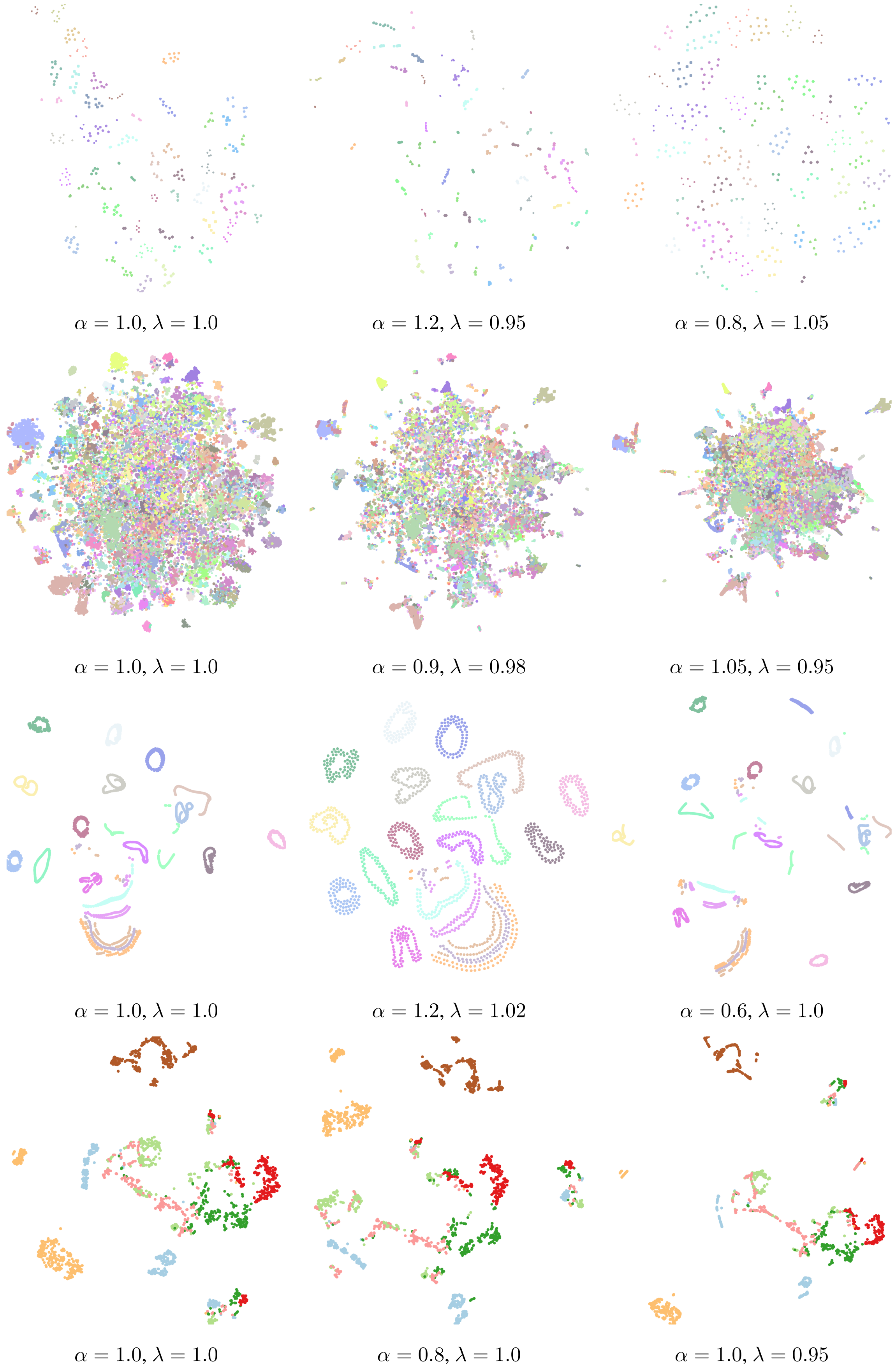
These plots aim to show the impact of α and β on qualitative properties of the embedding. As discussed in the paper: *changing λ should primarily affect global over local structure, where (i) $\lambda < 1$ should lead to greater cluster separation* *while (ii) $\lambda > 1$ should lead to low separation. Further, ABSNE should tend to produce lots of small, fine-grained clusters for $\alpha < 1$ with few global changes in visualization while $\alpha > 1$ should lead to fewer, larger clusters with more global visualization changes.*

We directly use pixels for the ORL and COIL-20 vision datasets, while we compute fc7 features yielded by Caffe's ImageNet model(Jia et al., 2014) for Caltech256. We use raw features for the other datasets. For all datasets with over 100 dimensions, we first apply 100 dimensional PCA before computing neighborhood scores.

References

Cichocki, A., Cruces, S., and Amari, S. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13:134–170, 2011.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. Caffe: Convolutional architecture for fast feature embedding, 2014.



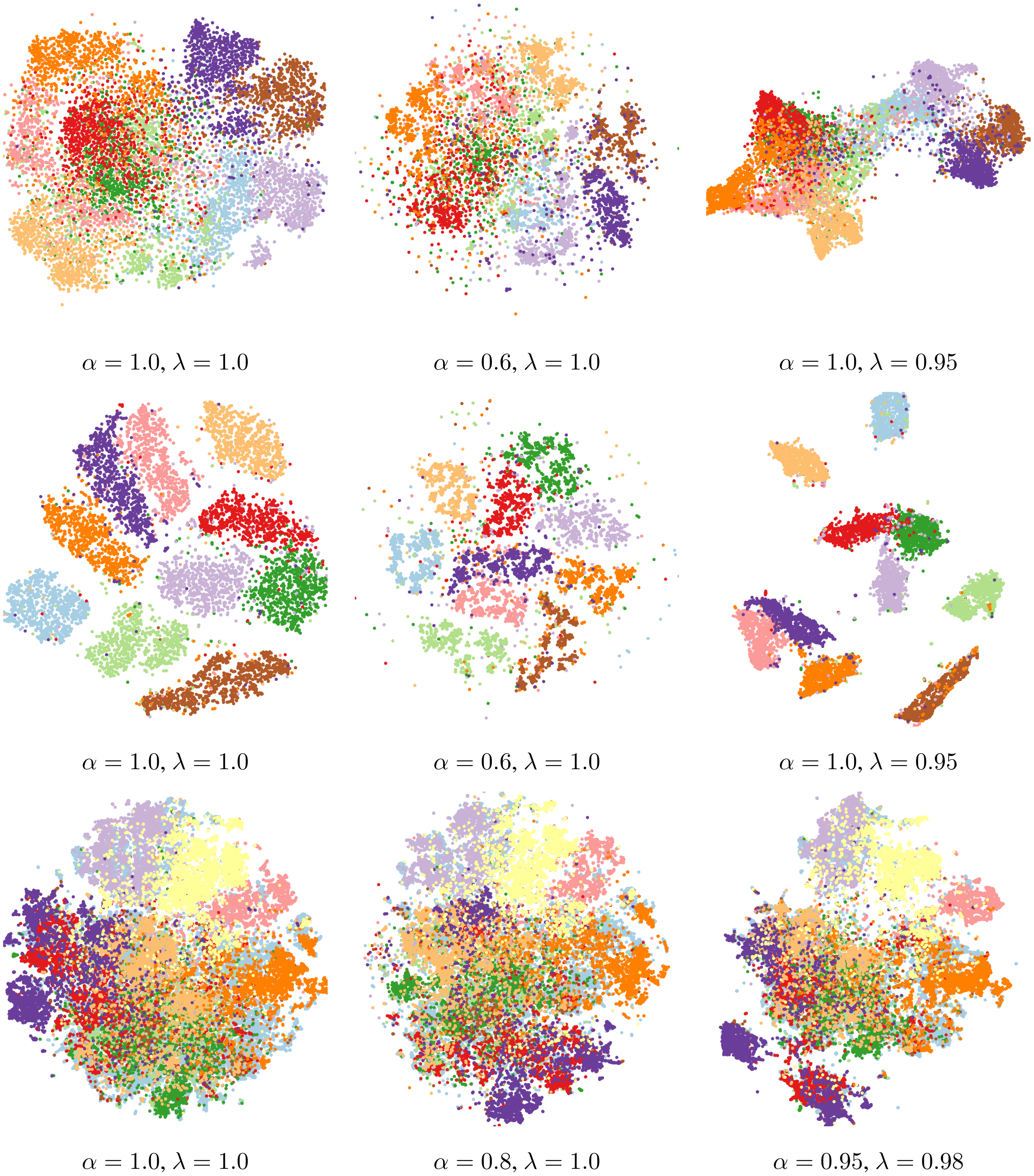
$\alpha = 1.0, \lambda = 1.0$ $\alpha = 1.2, \lambda = 0.95$ $\alpha = 0.8, \lambda = 1.05$

$\alpha = 1.0, \lambda = 1.0$ $\alpha = 0.9, \lambda = 0.98$ $\alpha = 1.05, \lambda = 0.95$

$\alpha = 1.0, \lambda = 1.0$ $\alpha = 1.2, \lambda = 1.02$ $\alpha = 0.6, \lambda = 1.0$

$\alpha = 1.0, \lambda = 1.0$ $\alpha = 0.8, \lambda = 1.0$ $\alpha = 1.0, \lambda = 0.95$

Figure 1. ABSNE visualizations for (rows from the top) ATT-Faces, Caltech256, COIL20, Segmentation Datasets. The left column corresponds to t-SNE. The center and right column contain visualizations with α and λ set manually to emphasize local or global clustering. These two parameters can be used by a data scientist for goal-driven exploratory data visualization.

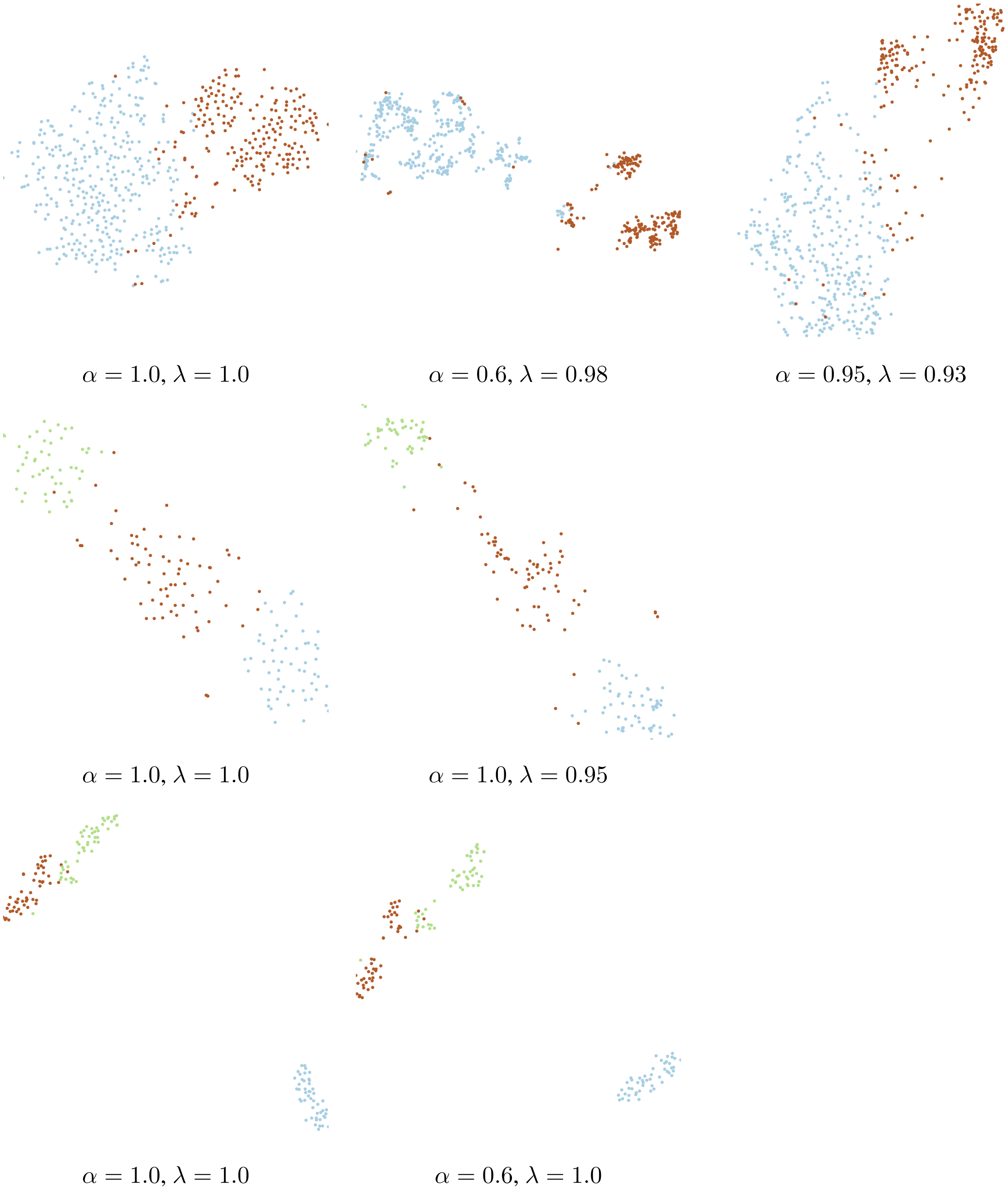


$\alpha = 1.0$, $\lambda = 1.0$ $\alpha = 0.6$, $\lambda = 1.0$ $\alpha = 1.0$, $\lambda = 0.95$

$\alpha = 1.0$, $\lambda = 1.0$ $\alpha = 0.6$, $\lambda = 1.0$ $\alpha = 1.0$, $\lambda = 0.95$

$\alpha = 1.0$, $\lambda = 1.0$ $\alpha = 0.8$, $\lambda = 1.0$ $\alpha = 0.95$, $\lambda = 0.98$

Figure 2. ABSNE visualizations for (rows from the top) CIFAR10, MNIST, ILSVRC2012 (validation) Datasets. The left column corresponds to t-SNE. The center and right column contain visualizations with α and λ set manually to emphasize local or global clustering. These two parameters can be used by a data scientist for goal-driven exploratory data visualization.



$\alpha = 1.0, \lambda = 1.0$

$\alpha = 0.6, \lambda = 0.98$

$\alpha = 0.95, \lambda = 0.93$

$\alpha = 1.0, \lambda = 1.0$

$\alpha = 1.0, \lambda = 0.95$

$\alpha = 1.0, \lambda = 1.0$

$\alpha = 0.6, \lambda = 1.0$

Figure 3. BSNE visualizations for (rows from the top) WDBC, WINE, IRIS Datasets. The left column corresponds to t-SNE. The center and right column contain visualizations with α and λ set manually to emphasize local or global clustering. WINE and IRIS are very small datasets with clear clusters, and ABSNE plots do not differ widely from t-SNE