Majorization-Minimization for Manifold Embedding (Supplemental Document)

Zhirong Yang[1,2,3], **Jaakko Peltonen**[1,3,4] **and Samuel Kaski**[1,2,3]
[1]Helsinki Institute for Information Technology HIIT, [2]University of Helsinki,
[3]Aalto University, [4]University of Tampere

Abstract

This supplemental document provides additional information that does not fit in the paper due to the space limit. Section 1 gives related mathematical basics, including majorizations of convex and concave functions, as well as the definitions of information divergences. Section 2 presents more examples on developing MM algorithms for manifold embedding, in addition to the t-SNE included in the paper. Section 3 illustrates that many existing manifold embedding methods can be unified into the Neighbor Embedding framework given in Section 5 of the paper. Section 4 provides proofs of the theorems in the paper. Section 5 gives an example of QL beyond manifold embedding. Section 6 gives the statistics and source of the experimented datasets. Section 7 provides supplemental experiment results.

1 Preliminaries

Here we provide two lemmas related to majorization of concave/convex functions and definitions of information divergences.

1.1 Majorization of concave/convex functions

1) A concave function can be upper-bounded by its tangent.

Lemma 1. *If $f(z)$ is concave in z, then*

$$\begin{aligned} f(\tilde{x}) \leq & f(x) + \left\langle \frac{\partial f(\tilde{x})}{\partial \tilde{x}}\Big|_{\tilde{x}=x}, \tilde{x} - x \right\rangle \\ = & \left\langle \frac{\partial f(\tilde{x})}{\partial \tilde{x}}\Big|_{\tilde{x}=x}, \tilde{x} \right\rangle + \textit{constant} \stackrel{def}{=} G(\tilde{x}, x). \end{aligned}$$

For a scalar concave function $f()$, this simplifies to $f(\tilde{x}) \leq G(\tilde{x}, x) = \tilde{x} f'(x) + \text{constant}$. Obviously, $G(x, x) = f(x)$ and $\frac{\partial f(\tilde{x})}{\partial \tilde{x}}\Big|_{\tilde{x}=x} = \frac{\partial G(\tilde{x}, x)}{\partial \tilde{x}}\Big|_{\tilde{x}=x}$.

2) A convex function can be upper-bounded by using the Jensen's inequality.

Lemma 2. *If $f(z)$ is convex in z, and $\tilde{x} = [\tilde{x}_1, \ldots, \tilde{x}_n]$ as well as $x = [x_1, \ldots, x_n]$ are nonnegative,*

$$\begin{aligned} f\left(\sum_{i=1}^n \tilde{x}_i\right) \leq & \sum_{i=1}^n \frac{x_i}{\sum_j x_j} f\left(\frac{\tilde{x}_i}{\frac{x_i}{\sum_j x_j}}\right) \\ = & \sum_{i=1}^n \frac{x_i}{\sum_j x_j} f\left(\frac{\tilde{x}_i}{x_i} \sum_j x_j\right) \stackrel{def}{=} G(\tilde{x}, x). \end{aligned}$$

Obviously $G(x, x) = f\left(\sum_{i=1}^n x_i\right)$. Their first derivatives also match:

$$\begin{aligned} \frac{\partial G(\tilde{x}, x)}{\partial \tilde{x}_i}\Big|_{\tilde{x}=x} = & \frac{x_i}{\sum_j x_j} f'\left(\frac{\tilde{x}_i}{x_i} \sum_j x_j\right) \frac{\sum_j x_j}{x_i}\Big|_{\tilde{x}=x} \\ = & f'\left(\sum_j x_j\right) \\ = & \frac{\partial f\left(\sum_{i=1}^n \tilde{x}_i\right)}{\partial \tilde{x}_i}\Big|_{\tilde{x}=x}. \end{aligned}$$

1.2 Information divergences

Information divergences, denoted by $D(p||q)$, were originally defined for probabilities and later extended to measure the difference between two (usually nonnegative) tensors p and q, where $D(p||q) \geq 0$, and

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

$D(p||q) = 0$ iff $p = q$. To avoid notational clutter we only give vectorial definitions; it is straightforward to extend the formulae to matrices and higher-order tensors.

Let D_α, D_β, D_γ, and D_r respectively denote α-, β-, γ-, and Rényi-divergences. Their definitions are (see e.g. [3])

$$
\begin{aligned}
D_\alpha(p||q) =& \frac{1}{\alpha(\alpha-1)} \sum_i \left[p_i^\alpha q_i^{1-\alpha} - \alpha p_i + (\alpha-1) q_i\right], \\
D_\beta(p||q) =& \frac{1}{\beta(\beta-1)} \sum_i \left[p_i^\beta + (\beta-1) q_i^\beta - \beta p_i q_i^{\beta-1}\right], \\
D_\gamma(p||q) =& \frac{\ln\left(\sum_i p_i^\gamma\right)}{\gamma(\gamma-1)} + \frac{\ln\left(\sum_i q_i^\gamma\right)}{\gamma} - \frac{\ln\left(\sum_i p_i q_i^{\gamma-1}\right)}{\gamma-1}, \\
D_r(p||q) =& \frac{1}{r-1} \ln\left(\sum_i \tilde{p}_i^r \tilde{q}_i^{1-r}\right),
\end{aligned}
$$

where p_i and q_i are the entries in p and q respectively, $\tilde{p}_i = p_i / \sum_j p_j$, and $\tilde{q}_i = q_i / \sum_j q_j$. To handle p's containing zero entries, we only consider nonnegative α, β, γ and r. These families are rich as they cover most commonly used divergences in machine learning such as normalized Kullback-Leibler divergence (obtained from D_r with $r \to 1$ or D_γ with $\gamma \to 1$), non-normalized KL-divergence ($\alpha \to 1$ or $\beta \to 1$), Itakura-Saito divergence ($\beta \to 0$), squared Euclidean distance ($\beta = 2$), Hellinger distance ($\alpha = 0.5$), and Chi-square divergence ($\alpha = 2$). Different divergences have become widespread in different domains. For example, D_{KL} is widely used for text documents (e.g. [7]) and D_{IS} is popular for audio signals (e.g. [4]). In general, estimation using α-divergence is more exclusive with larger α's, and more inclusive with smaller α's (e.g. [10]). For β-divergence, the estimation becomes more robust but less efficient with larger β's.

2 More Development Examples

The paper provides an example of developing a MM algorithm for t-Distributed Stochastic Neighbor Embedding (t-SNE); here we provide additional examples of developing MM algorithms for other manifold embedding methods.

2.1 Elastic Embedding (EE)

Given a symmetric and nonnegative matrix P and $\lambda > 0$, Elastic Embedding [2] minimizes

$$
\begin{aligned}
\mathcal{J}_{\mathrm{EE}}(\widetilde{Y}) =& \sum_{ij} P_{ij} \|\tilde{y}_i - \tilde{y}_j\|^2 + \lambda \sum_{ij} \exp\left(-\|\tilde{y}_i - \tilde{y}_j\|^2\right) \\
=& -\sum_{ij} P_{ij} \ln \widetilde{Q}_{ij} + \lambda \sum_{ij} \widetilde{Q}_{ij}
\end{aligned}
$$

where $\widetilde{Q}_{ij} = \exp\left(-\|\tilde{y}_i - \tilde{y}_j\|^2\right)$.

The EE objective is naturally decomposed into

$$
\begin{aligned}
A(P, \widetilde{Q}) =& -\sum_{ij} P_{ij} \ln \widetilde{Q}_{ij} \\
B(\widetilde{Q}) =& \lambda \sum_{ij} \widetilde{Q}_{ij}.
\end{aligned}
$$

The quadratification of $A(P, \widetilde{Q})$ is simply identical, with $W = P$. The final majorization function is

$$
\begin{aligned}
G(\widetilde{Y}, Y) =& \sum_{ij} P_{ij} \|\tilde{y}_i - \tilde{y}_j\|^2 + \left\langle \Psi, \widetilde{Y} \right\rangle \\
&+ \frac{\rho}{2} \|\widetilde{Y} - Y\|^2 + \text{constant},
\end{aligned}
$$

where

$$
\Psi = \frac{\partial B}{\partial \widetilde{Y}}\Big|_{\widetilde{Y}=Y} = -4\lambda \mathcal{L}_Q Y.
$$

Thus the MM update rule of EE is

$$
Y^{\mathrm{new}} = \left(\mathcal{L}_P + \frac{\rho}{4} I\right)^{-1} \left(\lambda \mathcal{L}_Q Y + \frac{\rho}{4} Y\right).
$$

2.2 Stochastic Neighbor Embedding (SNE)

Suppose the input matrix is row-stochastic, i.e. $P_{ij} \geq 0$ and $\sum_j P_{ij} = 1$. Denote $\tilde{q}_{ij} = \exp\left(-\|\tilde{y}_i - \tilde{y}_j\|^2\right)$ and $\widetilde{Q}_{ij} = \tilde{q}_{ij} / \sum_b \tilde{q}_{ib}$. Stochastic Neighbor Embedding [6] minimizes total Kullback-Leibler (KL) divergence between P and $\widetilde{Q}$ rows:

$$
\begin{aligned}
&\mathcal{J}_{\mathrm{SNE}}(\widetilde{Y}) \\
=& \sum_i \sum_j P_{ij} \ln \frac{P_{ij}}{\widetilde{Q}_{ij}} \\
=& -\sum_{ij} P_{ij} \ln \tilde{q}_{ij} + \sum_i \ln \sum_j \tilde{q}_{ij} + \text{constant} \\
=& \sum_{ij} P_{ij} \|\tilde{y}_i - \tilde{y}_j\|^2 + \sum_i \ln \sum_j \tilde{q}_{ij} + \text{constant}
\end{aligned}
$$

Thus we can decompose the SNE objective into $A(P, \tilde{q}) + B(\tilde{q}) + \text{constant}$, where

$$
\begin{aligned}
A(P, \tilde{q}) =& \sum_{ij} P_{ij} \|\tilde{y}_i - \tilde{y}_j\|^2 \\
B(\tilde{q}) =& \sum_i \ln \sum_j \tilde{q}_{ij}.
\end{aligned}
$$

Again the quadratification of $A(P,\tilde{q})$ is simply identical, with $W = P$. The final majorization function is

$$G(\widetilde{Y},Y) = \sum_{ij} P_{ij}\|\tilde{y}_i - \tilde{y}_j\|^2 + \left\langle \Psi, \widetilde{Y}\right\rangle + \frac{\rho}{2}\|\widetilde{Y} - Y\|^2 + \text{constant},$$

where

$$\Psi = \frac{\partial B}{\partial \widetilde{Y}}\Big|_{\widetilde{Y}=Y} = -2\mathcal{L}_{Q+Q^T}Y.$$

Thus the MM update rule of SNE is

$$Y^{\text{new}} = \left(\mathcal{L}_{P+P^T} + \frac{\rho}{2}I\right)^{-1}\left(\mathcal{L}_{Q+Q^T}Y + \frac{\rho}{2}Y\right).$$

Similarly, we can develop the rule for symmetric SNE (s-SNE) [12]:

$$\mathcal{J}_{\text{s-SNE}}(Y) = \sum_{ij} P_{ij}\ln\frac{P_{ij}}{Q_{ij}},$$

$$Y^{\text{new}} = \left(\mathcal{L}_P + \frac{\rho}{4}I\right)^{-1}\left(\mathcal{L}_Q Y + \frac{\rho}{4}Y\right),$$

where $\sum_{ij} P_{ij} = 1$, $Q_{ij} = q_{ij}/\sum_{ab} q_{ab}$ and $q_{ij} = \exp\left(-\|y_i - y_j\|^2\right)$; and for Neighbor Retrieval Visualizer (NeRV) [13]:

$$\mathcal{J}_{\text{NeRV}}(Y) = \lambda\sum_i\sum_j P_{ij}\ln\frac{P_{ij}}{Q_{ij}} + (1-\lambda)\sum_i\sum_j Q_{ij}\ln\frac{Q_{ij}}{P_{ij}},$$

$$Y^{\text{new}} = \left(\mathcal{L}_{P+P^T} + \frac{\rho}{2}I\right)^{-1}\left(\Psi + \frac{\rho}{2}Y\right)$$

where P and Q are defined same manner as in SNE, and

$$\Psi = \frac{\partial \mathcal{J}_{\text{NeRV}} + \lambda\sum_i\sum_j P_{ij}\ln q_{ij}}{\partial Y}.$$

2.3 LinLog

Suppose a weighted undirected graph is encoded in a symmetric and nonnegative matrix P (i.e. the weighted adjacency matrix). The node-repulsive LinLog graph layout method [11] minimizes the following energy function:

$$\mathcal{J}_{\text{LinLog}}(\widetilde{Y}) = \sum_{ij} P_{ij}\|\tilde{y}_i - \tilde{y}_j\| - \lambda\sum_{ij}\ln\|\tilde{y}_i - \tilde{y}_j\|.$$

We write $\widetilde{Q} = \|\tilde{y}_i - \tilde{y}_j\|^{-1}$ and then decompose the LinLog objective function into $A(P,\widetilde{Q}) + B(P,\widetilde{Q})$, where

$$A(P,\widetilde{Q}) = \sum_{ij} P_{ij}\|\tilde{y}_i - \tilde{y}_j\|$$

$$B(P,\widetilde{Q}) = -\lambda\sum_{ij}\ln\|\tilde{y}_i - \tilde{y}_j\|.$$

Since the square root function is concave, by Lemma 1 we can upper bound $A(P,\widetilde{Q})$ by

$$A(P,\widetilde{Q}) \le \sum_{ij}\frac{P_{ij}}{2\|y_i - y_j\|}\|\tilde{y}_i - \tilde{y}_j\|^2 + \text{constant}$$

That is, $W_{ij} = \frac{P_{ij}}{2\|y_i - y_j\|} = \frac{1}{2}P_{ij}Q_{ij}$ in the quadratification phase. The final majorization function of the LinLog objective is

$$G(\widetilde{Y},Y) = \sum_{ij}\frac{P_{ij}}{2\|y_i - y_j\|}\|\tilde{y}_i - \tilde{y}_j\|^2 + \left\langle \Psi, \widetilde{Y}\right\rangle + \frac{\rho}{2}\|\widetilde{Y} - Y\|^2 + \text{constant},$$

where

$$\Psi = \frac{\partial B}{\partial \widetilde{Y}}\Big|_{\widetilde{Y}=Y} = -2\lambda\mathcal{L}_{Q\circ Q}Y,$$

where $\circ$ is the elementwise product. Then the MM update rule of LinLog is

$$Y^{\text{new}} = \left(\mathcal{L}_{P\circ Q} + \frac{\rho}{2}I\right)^{-1}\left(\mathcal{L}_{\lambda\circ Q\circ Q}Y + \frac{\rho}{2}Y\right).$$

2.4 Multidimensional Scaling with Kernel Strain (MDS-KS)

Strain-based multidimensional scaling maximizes the cosine between the inner products in the input and output spaces (see e.g. [1]):

$$\mathcal{J}_{\text{MDS-S}}(\widetilde{Y}) = \frac{\sum_{ij} P_{ij}\langle\tilde{y}_i,\tilde{y}_j\rangle}{\sqrt{\sum_{ij}P_{ij}^2}\cdot\sqrt{\sum_{ij}\langle\tilde{y}_i,\tilde{y}_j\rangle^2}}$$

where $P_{ij} = \langle x_i, x_j\rangle$. The inner products $\langle\tilde{y}_i,\tilde{y}_j\rangle$ above can be seen as a linear kernel $\widetilde{Q}_{ij} = \langle\tilde{y}_i,\tilde{y}_j\rangle$ and simply lead to kernel PCA of P. In this example we instead consider a nonlinear embedding kernel $\widetilde{Q}_{ij} = \exp(-\|\tilde{y}_i - \tilde{y}_j\|^2)$, where the corresponding objective function is

$$\frac{\sum_{ij} P_{ij}\widetilde{Q}_{ij}}{\sqrt{\sum_{ij}P_{ij}^2}\cdot\sqrt{\sum_{ij}\widetilde{Q}_{ij}^2}},$$

and maximizing it is equivalent to minimizing its negative logarithm

$$\mathcal{J}_{\text{MDS-KS}}(\widetilde{Y}) = -\ln\sum_{ij}P_{ij}\widetilde{Q}_{ij} + \frac{1}{2}\ln\sum_{ij}\widetilde{Q}_{ij}^2 + \frac{1}{2}\ln\sum_{ij}P_{ij}^2.$$

We decompose the MDS-KS objective function into $A(P,\widetilde{Q}) + B(P,\widetilde{Q}) + \text{constant}$, where

$$A(P,\widetilde{Q}) = -\ln\sum_{ij}P_{ij}\widetilde{Q}_{ij}$$

$$B(P,\widetilde{Q}) = \frac{1}{2}\ln\sum_{ij}\widetilde{Q}_{ij}^2.$$

We can upper bound $A(P,\widetilde{Q})$ by the Jensen's inequality (Lemma 2):

$$\begin{aligned}A(P,\widetilde{Q}) \leq & -\sum_{ij}\frac{P_{ij}Q_{ij}}{\sum_{ab}P_{ab}Q_{ab}}\ln\left(\frac{P_{ij}\widetilde{Q}_{ij}}{\frac{P_{ij}Q_{ij}}{\sum_{ab}P_{ab}Q_{ab}}}\right)\\ =&\sum_{ij}\frac{P_{ij}Q_{ij}}{\sum_{ab}P_{ab}Q_{ab}}\|\tilde{y}_i-\tilde{y}_j\|^2\\ &+\sum_{ij}\frac{P_{ij}Q_{ij}}{\sum_{ab}P_{ab}Q_{ab}}\ln\frac{Q_{ij}}{\sum_{ab}P_{ab}Q_{ab}}.\end{aligned}$$

That is, $W_{ij}=\frac{P_{ij}Q_{ij}}{\sum_{ab}P_{ab}Q_{ab}}$ in the quadratification phase.

The final majorization function of the MDS-KS objective is

$$\begin{aligned}G(\widetilde{Y},Y)=&\sum_{ij}\frac{P_{ij}Q_{ij}}{\sum_{ab}P_{ab}Q_{ab}}\|\tilde{y}_i-\tilde{y}_j\|^2+\left\langle\Psi,\widetilde{Y}\right\rangle\\ &+\frac{\rho}{2}\|\widetilde{Y}-Y\|^2+\text{constant},\end{aligned}$$

where

$$\Psi=\frac{\partial B}{\partial\widetilde{Y}}\Big|_{\widetilde{Y}=Y}=-\mathcal{L}_U Y,$$

with $U_{ij}=\frac{Q_{ij}^2}{\sum_{ab}Q_{ab}^2}$. Then the MM update rule of MDS-KS is

$$Y^{\text{new}}=\left(\mathcal{L}_W+\frac{\rho}{4}I\right)^{-1}\left(\mathcal{L}_U Y+\frac{\rho}{4}Y\right).$$

3 Neighbor Embedding

In Section 5 of the paper, we review a framework for manifold embedding which is called Neighbor Embedding (NE) [16, 17]. Here we demonstrate that many manifold embedding objectives, including the above examples, can be equivalently formulated as an NE problem.

NE minimizes $D(P||\widetilde{Q})$, where D is a divergence from α-, β-, γ-, or Rényi-divergence families, and the embedding kernel in the paper is parameterized as $\widetilde{Q}_{ij}=\left(c+a\|\tilde{y}_i-\tilde{y}_j\|^2\right)^{-b/a}$ for $a\geq 0$, $b>0$, and $c\geq 0$ (adapted from [15]).

First we show the parameterized form of $\widetilde{Q}$ includes the most popularly used embedding kernels. When $a=1$, $b=1$, and $c=1$, $\widetilde{Q}_{ij}=\left(1+\|\tilde{y}_i-\tilde{y}_j\|^2\right)^{-1}$ is the Cauchy kernel (i.e. the Student-t kernel with a single degree of freedom); when $a\to 0$ and $b=1$, $\widetilde{Q}_{ij}=\exp\left(-\|\tilde{y}_i-\tilde{y}_j\|^2\right)$ is the Gaussian kernel; when $a=1$, $b=1/2$, $c=0$, $\widetilde{Q}_{ij}=\|\tilde{y}_i-\tilde{y}_j\|^{-1}$ is the inverse to the Euclidean distance.

Next we show some other existing manifold embedding objectives can equivalently expressed as NE. Obviously SNE and its variants belong to NE. Moreover, we have

- $\arg\min_{\widetilde{Y}}\mathcal{J}_{\text{EE}}(\widetilde{Y})=\arg\min_{\widetilde{Y}}D_{\beta\to 1}(P||\lambda\widetilde{Q})$, with $\widetilde{Q}_{ij}=\exp\left(-\|\tilde{y}_i-\tilde{y}_j\|^2\right)$ [17].

  Proof.

  $$\begin{aligned}&\mathcal{J}_{\text{EE}}(\widetilde{Y})\\ =&\sum_{ij}P_{ij}\|\tilde{y}_i-\tilde{y}_j\|^2+\lambda\sum_{ij}\exp\left(-\|\tilde{y}_i-\tilde{y}_j\|^2\right)\\ =&-\sum_{ij}P_{ij}\ln\widetilde{Q}_{ij}+\lambda\sum_{ij}\widetilde{Q}_{ij}\\ =&\sum_{ij}P_{ij}\ln\frac{P_{ij}}{\lambda\widetilde{Q}_{ij}}+\lambda\sum_{ij}\widetilde{Q}_{ij}-\sum_{ij}P_{ij}\\ &-\sum_{ij}P_{ij}\ln P_{ij}+(\ln\lambda+1)\sum_{ij}P_{ij}\\ =&D_{\beta\to 1}(P||\lambda\widetilde{Q})+\text{constant}.\end{aligned}$$

  □

- $\arg\min_{\widetilde{Y}}\mathcal{J}_{\text{LinLog}}(\widetilde{Y})=\arg\min_{\widetilde{Y}}D_{\beta\to 0}(P||\lambda\widetilde{Q})$, with $\widetilde{Q}_{ij}=\|\tilde{y}_i-\tilde{y}_j\|^{-1}$ [17].

  Proof.

  $$\begin{aligned}&\frac{1}{\lambda}\mathcal{J}_{\text{LinLog}}(\widetilde{Y})\\ =&\frac{1}{\lambda}\sum_{ij}P_{ij}\|\tilde{y}_i-\tilde{y}_j\|-\sum_{ij}\ln\|\tilde{y}_i-\tilde{y}_j\|,\\ =&\sum_{ij}\left[\frac{P_{ij}}{\lambda\widetilde{Q}_{ij}}-\ln\frac{1}{\widetilde{Q}_{ij}}\right]\\ =&\sum_{ij}\left[\frac{P_{ij}}{\lambda\widetilde{Q}_{ij}}-\ln\frac{P_{ij}}{\lambda\widetilde{Q}_{ij}}-1\right]+\sum_{ij}\left[\ln\frac{P_{ij}}{\lambda}+1\right]\\ =&D_{\beta\to 0}(P||\lambda\widetilde{Q})+\text{constant}.\end{aligned}$$

  □

- $\arg\min_{\widetilde{Y}}\mathcal{J}_{\text{MDS-KS}}(\widetilde{Y})=\arg\min_{\widetilde{Y}}D_{\gamma=2}(P||\widetilde{Q})$ by their definitions.

4 Proofs

Here we provide proofs of the five theorems in the paper.

4.1 Proof of Theorem 1

Proof. Since B is upper-bounded by its Lipschitz surrogate

$$B(P,\widetilde{Q}) = B(\widetilde{Y}) \leq B(Y) + \langle \Psi, \widetilde{Y} - Y\rangle + \frac{\rho}{2}\|\widetilde{Y} - Y\|_F^2,$$

we have $H(Y,Y) = G(Y,Y)$. Therefore $\mathcal{J}(Y) = H(Y,Y) = G(Y,Y) \geq G(Y^{\text{new}},Y) \geq \mathcal{J}(Y^{\text{new}})$, where the first inequality comes from minimization and the second is ensured by the backtracking. □

4.2 Proof of Theorem 2

The proposed MM updates share many convergence properties with the Expectation-Maximization (EM) algorithm. In this section we follow the steps in [14] to show that the MM updates will converge a stationary point of $\mathcal{J}$. A stationary point can be a local optimum or a saddle point.

The convergence result is a special case of the Global Convergence Theorem (GCT; [19]) which is quoted below. A map $\mathcal{A}$ from points of X to subsets of X is called a point-to-set map on X. It is said to be closed at x if $x_k \to x$, $x_k \in X$ and $y_k \to y$, $y_k \in \mathcal{A}(x_k)$, imply $y \in \mathcal{A}(x)$. For point-to-point map, continuity implies closedness.

Global Convergence Theorem. *(GCT; from [14]) Let the sequence $\{x_k\}_{k=0}^{\infty}$ be generated by $x_{k+1} \in \mathcal{M}(x_k)$, where $\mathcal{M}$ is a point-to-set map on X. Let a solution set $\Gamma \in X$ be given, and suppose that:*

i all points x_k are contained in a compact set $S \subset X$;

ii $\mathcal{M}$ is closed over the complement of Γ;

iii there is a continuous function α on X such that

(a) if $x \notin \Gamma$, $\alpha(x) > \alpha(y)$ for all $y \in \mathcal{M}(x)$, and
(b) if $x \in \Gamma$, $\alpha(x) \geq \alpha(y)$ for all $y \in \mathcal{M}(x)$

Then all the limit points of $\{x_k\}$ are in the solution set Γ and $\alpha(x_k)$ converges monotonically to $\alpha(x)$ for some $x \in \Gamma$.

The proof can be found in [19].

Before showing the convergence of MM, we need the following Lemmas. For brevity, denote $\mathcal{M} : Y \to Y^{\text{new}}$ the map by using the MM update Eq. 4 in the paper. Let S and F be the sets of stationary points of $\mathcal{J}(Y)$ and fixed points of $\mathcal{M}$, respectively.

Lemma 3. *$S = F$.*

Proof. The fixed points of the MM update rule appear when

$$\begin{aligned} Y &= \left(2\mathcal{L}_{W+W^T} + \rho I\right)^{-1}\left(-\Psi + \rho Y\right) \\ \left(2\mathcal{L}_{W+W^T} + \rho I\right) Y &= \left(-\Psi + \rho Y\right) \\ 2\mathcal{L}_{W+W^T} Y + \Psi &= 0, \end{aligned} \tag{1}$$

which is recognized as $\frac{\partial \mathcal{H}}{\partial \widetilde{Y}}\Big|_{\widetilde{Y}=Y} = 0$. Since we require the majorization function $\mathcal{H}$ shares the tangent with $\mathcal{J}$, i.e. $\frac{\partial \mathcal{J}}{\partial \widetilde{Y}}\Big|_{\widetilde{Y}=Y} = \frac{\partial \mathcal{H}}{\partial \widetilde{Y}}\Big|_{\widetilde{Y}=Y}$, Eq. 1 is equivalent to $\frac{\partial \mathcal{J}}{\partial \widetilde{Y}}\Big|_{\widetilde{Y}=Y} = 0$, the condition of stationary points of $\mathcal{J}$. Therefore $F \subseteq S$.

On the other hand, because QL requires that G and $\mathcal{J}$ share the same tangent at Y, and thus $\frac{\partial \mathcal{J}}{\partial \widetilde{Y}}\Big|_{\widetilde{Y}=Y} = 0$ implies $\frac{\partial \mathcal{G}}{\partial \widetilde{Y}}\Big|_{\widetilde{Y}=Y} = 0$, i.e. $Y = \mathcal{M}(Y)$. Therfore $S \subseteq F$ □

Lemma 4. *$\mathcal{J}(Y^{new}) < \mathcal{J}(Y)$ if $Y \notin F$.*

Proof. Because G is convexly quadratic, it has a unique minimum $Y^{\text{new}} = \mathcal{M}(Y)$. If $Y \notin F$, i.e. $Y \neq \mathcal{M}(Y)$, we have $Y \neq Y^{\text{new}}$, which implies $\mathcal{J}(Y^{\text{new}}) \leq G(Y^{\text{new}},Y) < G(Y,Y) = \mathcal{J}(Y)$. □

Now we are ready to prove the convergence to stationary points (Theorem 2).

Proof. Consider S the solution set Γ, and $\mathcal{J}$ the continuous function α in the GCT theorem. Lemma 3 shows that this is equivalent to considering F the solution set. Next we show that the QL-majorization and its resulting map $\mathcal{M}$ fulfill the conditions of the GCT theorem: $\mathcal{J}(Y^{\text{new}}) \geq \mathcal{J}(Y)$ and the boundedness assumption of $\mathcal{J}$ imply Condition i; $\mathcal{M}$ is a point-to-point map and thus the continuity of G over both $\widetilde{Y}$ and Y implies the closedness condition ii; Lemma 4 implies iii(a); Theorem 1 in the paper implies iii(b). Therefore, the proposed MM updates are a special case of GCT and thus converge to a stationary point of $\mathcal{J}$. □

4.3 Proof of Theorem 3

Proof. Since A_{ij} is concave to $\|\tilde{y}_i - \tilde{y}_j\|^2$, it can be upper-bounded by its tangent:

$$\begin{aligned} & A_{ij}(P_{ij},\widetilde{Q}_{ij}) \\ \leq & A_{ij}(P_{ij},Q_{ij}) \\ & + \left\langle \frac{\partial A_{ij}}{\partial\|\tilde{y}_i - \tilde{y}_j\|^2}\Big|_{\widetilde{Y}=Y}, \|\tilde{y}_i - \tilde{y}_j\|^2 - \|y_i - y_j\|^2 \right\rangle \\ = & \left\langle \frac{\partial A_{ij}}{\partial\|\tilde{y}_i - \tilde{y}_j\|^2}\Big|_{\widetilde{Y}=Y}, \|\tilde{y}_i - \tilde{y}_j\|^2 \right\rangle + \text{constant}. \end{aligned}$$

That is, $W_{ij} = \frac{\partial A_{ij}}{\partial \|\tilde{y}_i - \tilde{y}_j\|^2}\Big|_{\widetilde{Y}=Y}$.

For the tangent sharing, first we have $\mathcal{H}(Y,Y) = \mathcal{J}(Y)$ because obviously $A_{ij}(P_{ij}, \widetilde{Q}_{ij}) = A_{ij}(P_{ij}, Q_{ij})$ when $\widetilde{Y} = Y$. Next, because

$$\begin{aligned}\frac{\partial A}{\partial \widetilde{Y}_{st}} =& \sum_{ij} \frac{A_{ij}}{\|\tilde{y}_i - \tilde{y}_j\|^2} \frac{\partial \|\tilde{y}_i - \tilde{y}_j\|^2}{\partial \widetilde{Y}_{st}} \\ =& 2\sum_j W_{si}\left(\tilde{y}_{st} - \tilde{y}_{it}\right) \\ =& 2\left(\mathcal{L}_{W+W^T}\widetilde{Y}\right)_{st}\end{aligned}$$

is the same as

$$\frac{\partial \sum_{ij} W_{ij}\|\tilde{y}_i - \tilde{y}_j\|^2 + \text{constant}}{\partial \widetilde{Y}_{st}} = 2\left(\mathcal{L}_{W+W^T}\widetilde{Y}\right)_{st},$$

we have $\frac{\partial H}{\partial \widetilde{Y}}\Big|_{\widetilde{Y}=Y} = \frac{\partial \mathcal{J}}{\partial \widetilde{Y}}\Big|_{\widetilde{Y}=Y}$. □

4.4 Proof of Theorem 4

Proof. Obviously all α- and β-divergences are additively separable. Next we show the concavity in the given range. Denote $\xi_{ij} = \|\tilde{y}_i - \tilde{y}_j\|^2$ for brevity.

α-divergence. We decompose an α-divergence into $D_\alpha(P||\widetilde{Q}) = A(P,\widetilde{Q}) + B(P,\widetilde{Q}) + \text{constant}$, where

$$\begin{aligned}A(P,\widetilde{Q}) =& \sum_{ij} A_{ij}(P_{ij},\widetilde{Q}_{ij}) = \sum_{ij} \frac{1}{\alpha(\alpha-1)} P_{ij}^\alpha \widetilde{Q}_{ij}^{1-\alpha} \\ =& \sum_{ij} \frac{1}{\alpha(\alpha-1)} P_{ij}^\alpha \left(c + a\xi_{ij}\right)^{-b(1-\alpha)/a} \\ B(P,\widetilde{Q}) =& \frac{1}{\alpha}\sum_{ij}\widetilde{Q}_{ij}\end{aligned}$$

The second derivative of

$$A_{ij} = \frac{1}{\alpha(\alpha-1)} P_{ij}^\alpha \left(c + a\xi_{ij}\right)^{-b(1-\alpha)/a}$$

to ξ_{ij} is

$$\frac{\partial^2 A_{ij}}{\partial \xi_{ij}^2} = \frac{bP_{ij}^\alpha((\alpha-1)b - a)\left((a\xi_{ij}+c)^{-\frac{b}{a}}\right)^{1-\alpha}}{\alpha(ax+c)^2}.$$

Since a, b, c, and P_{ij} are nonnegative, we have $\frac{\partial^2 A_{ij}}{\partial \xi_{ij}^2} \le 0$ iff $\alpha((\alpha-1)b - a) \le 0$ and $\alpha \ne 0$. That is, A_{ij} is concave in ξ_{ij} iff $\alpha \in (0, 1 + a/b]$.

β-divergence. We decompose a β-divergence into $D_\beta(P||\widetilde{Q}) = A(P,\widetilde{Q}) + B(P,\widetilde{Q}) + \text{constant}$, where

$$\begin{aligned}A(P,\widetilde{Q}) =& \sum_{ij} A_{ij}(P_{ij},\widetilde{Q}_{ij}) = \sum_{ij} \frac{1}{1-\beta} P_{ij}\widetilde{Q}_{ij}^{\beta-1} \\ =& \sum_{ij} \frac{1}{1-\beta} P_{ij}\left(c + a\xi_{ij}\right)^{-b(\beta-1)/a} \\ B(P,\widetilde{Q}) =& \frac{1}{\beta}\sum_{ij}\widetilde{Q}_{ij}^{\beta}\end{aligned}$$

The second derivative of $A_{ij} = \frac{1}{1-\beta} P_{ij}\left(c + a\xi_{ij}\right)^{-b(\beta-1)/a}$ with respect to ξ_{ij} is

$$-\frac{bP_{ij}(a + b(\beta-1))\left((a\xi_{ij}+c)^{-\frac{b}{a}}\right)^{\beta-1}}{(a\xi_{ij}+c)^2} \tag{2}$$

Since a, b, c, and P_{ij} are nonnegative, we have $\frac{\partial^2 A_{ij}}{\partial \xi_{ij}^2} \le 0$ iff $-(a + b(\beta-1)) \le 0$. That is A_{ij} is concave in ξ_{ij} iff $\beta \in [1 - a/b, \infty)$.

Special case 1 ($\alpha \to 1$ or $\beta \to 1$): $D_{\mathrm{I}}(P||\widetilde{Q}) = -\sum_{ij} P_{ij}\ln\widetilde{Q}_{ij} + \sum_{ij}\widetilde{Q}_{ij} + \text{constant}$. We write

$$\begin{aligned}A(P,\widetilde{Q}) =& \sum_{ij} A_{ij}(P_{ij},\widetilde{Q}_{ij}) = -\sum_{ij} P_{ij}\ln\widetilde{Q}_{ij} \\ =& \sum_{ij}\frac{b}{a}P_{ij}\ln\left(c + a\xi_{ij}\right) \\ B(P,\widetilde{Q}) =& \sum_{ij}\widetilde{Q}_{ij}\end{aligned}$$

The second derivative of $A_{ij} = \frac{b}{a}P_{ij}\ln\left(c + a\xi_{ij}\right)$ with respect to ξ_{ij} is $-\frac{abP_{ij}}{(a\xi_{ij}+c)^2}$, which is always non-positive.

Special case 2 ($\alpha \to 0$ for $P_{ij} > 0$): $D_{\text{dual-I}}(P||\widetilde{Q}) = \sum_{ij}\widetilde{Q}_{ij}\ln P_{ij} + \sum_{ij}\widetilde{Q}_{ij}\ln\widetilde{Q}_{ij} - \sum_{ij}\widetilde{Q}_{ij}$. We write

$$\begin{aligned}A(P,\widetilde{Q}) =& \sum_{ij} A_{ij}(P_{ij},\widetilde{Q}_{ij}) = \sum_{ij}\widetilde{Q}_{ij}\ln P_{ij} \\ =& \sum_{ij}\left(c + a\xi_{ij}\right)^{-b/a}\ln P_{ij} \\ B(P,\widetilde{Q}) =& \sum_{ij}\widetilde{Q}_{ij}\ln\widetilde{Q}_{ij} - \sum_{ij}\widetilde{Q}_{ij}\end{aligned}$$

The second derivative of $A_{ij} = \left(c + a\xi_{ij}\right)^{-b/a}\ln P_{ij}$ with respect to ξ_{ij} is $b(a+b)\ln P_{ij}(a\xi_{ij}+c)^{-\frac{b}{a}-2}$, which is non-positive iff $P_{ij} \in (0, 1]$.

Special case 3 ($\beta \to 0$): $D_{\mathrm{IS}}(P||\widetilde{Q}) = \sum_{ij} P_{ij}\widetilde{Q}_{ij}^{-1} +$

$\sum_{ij} \ln \widetilde{Q}_{ij} + \text{constant}$. We write

$$\begin{aligned} A(P, \widetilde{Q}) &= \sum_{ij} A_{ij}(P_{ij}, \widetilde{Q}_{ij}) = \sum_{ij} P_{ij} \widetilde{Q}_{ij}^{-1} \\ &= \sum_{ij} P_{ij} \left(c + a\xi_{ij}\right)^{b/a} \\ B(P, \widetilde{Q}) &= \sum_{ij} \ln \widetilde{Q}_{ij}. \end{aligned}$$

The second derivative of $A_{ij} = P_{ij} \left(c + a\xi_{ij}\right)^{b/a}$ with respect to ξ_{ij} is $(b-a)bP_{ij}(a\xi_{ij}+c)^{\frac{b}{a}-2}$, which is non-positive iff $a \geq b$, or equivalently $0 \in [1 - a/b, \infty]$.

□

4.5 Proof of Theorem 5

The proofs are done by zeroing the derivative of the right hand side with respect to λ (see [17]). The closed-form solutions of λ at the current estimate for $\tau \geq 0$ are

$$\lambda^* = \arg\min_{\lambda} D_{\alpha \to \tau}(P||\lambda Q) = \left(\frac{\sum_{ij} P_{ij}^{\tau} Q_{ij}^{1-\tau}}{\sum_{ij} Q_{ij}} \right)^{\frac{1}{\tau}},$$

$$\lambda^* = \arg\min_{\lambda} D_{\beta \to \tau}(P||\lambda Q) = \frac{\sum_{ij} P_{ij} Q_{ij}^{\tau-1}}{\sum_{ij} Q_{ij}^{\tau}},$$

with the special case

$$\lambda^* = \exp\left(- \frac{\sum_{ij} Q_{ij} \ln(Q_{ij}/P_{ij})}{\sum_{ij} Q_{ij}} \right)$$

for $\alpha \to 0$.

5 Examples of QL beyond Manifold Embedding

To avoid vagueness we defined the scope to be manifold embedding, which already is a broad field and one of listed AISTATS research areas. Within this scope the work is general.

It is also naturally applicable to any other optimization problems amenable to QL majorization. QL is applicable to any cost function $\mathcal{J} = A + B$, where A can be tightly and quadratically upper bounded by Eq. 2 in the paper, and B is smooth. Next we give an example beyond visualization and discuss its potential extensions.

Consider a semi-supervised problem: given a training set comprising a supervised subset $\{(x_i, y_i)\}_{i=1}^n$ and an unsupervised subset $\{x_i\}_{i=n+1}^N$, where $x_i \in \mathbb{R}^D$ are vectorial primary data and $y_i \in \{-1, 1\}$ are supervised labels, the task is to learn a linear function $z = \langle w, x \rangle + b$ for predicting y.

In this example a composite cost function is used:

$$\mathcal{J}(w, b) = A(w, b) + B(w, b)$$

where $A(w)$ is a locality preserving regularizer (see e.g. [5])

$$A(w, b) = \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} S_{ij} (z_i - z_j)^2$$

and $B(w, b)$ is an empirical loss function [9]

$$B(w, b) = (1 - \lambda) \sum_{i=1}^{n} [1 - \tanh(y_i z_i)]$$

with $\lambda \in [0, 1]$ the tradeoff parameter and S_{ij} the local similarity between x_i and x_j (e.g. a Gaussian kernel).

Because B is non-convex and non-concave, conventional majorization techniques that require convexity/concavity such as CCCP [18] are not applicable. However, we can apply QL here (Lipschitzation to B and quadratification to A), which gives the update rule for w (denote $X = [x_1 \dots x_N]$):

$$w^{\text{new}} = \left(\lambda X \mathcal{L}_{\frac{S+S^T}{2}} X^T + \rho I \right)^{-1} \left(\frac{\partial B}{\partial w} + \rho w \right).$$

This example can be further extended with the same spirit, where B is replaced by another smooth loss function which is non-convex and non-concave, e.g. [8]

$$B(w, b) = (1 - \lambda) \sum_{i=1}^{n} \left[1 - \frac{1}{1 + \exp(-y_i z_i)} \right]^2,$$

and $X \mathcal{L}_{\frac{S+S^T}{2}} X^T$ can be replaced by other positive semi-definite matrices.

6 Datasets

Twelve datasets have been used in our experiments. Their statistics and sources are given in Table 1. Below are brief descriptions of the datasets.

- `SCOTLAND`: It lists the (136) multiple directors of the 108 largest joint stock companies in Scotland in 1904-5: 64 non-financial firms, 8 banks, 14 insurance companies, and 22 investment and property companies (Scotland.net).
- `COIL20`: the *COIL-20* dataset from Columbia University Image Library, images of toys from different angles, each image of size 128×128.
- `7SECTORS`: the *4 Universities* dataset from CMU Text Learning group, text documents classified to 7 sectors; 10,000 words with maximum information gain are preserved.

Table 1: Dataset statistics and sources

Dataset	#samples	#classes	Domain	Source
SCOTLAND	108	8	network	PAJEK
COIL20	1440	20	image	COIL
7SECTORS	4556	7	text	CMUTE
RCV1	9625	4	text	RCV1
PENDIGITS	10992	10	image	UCI
MAGIC	19020	2	telescope	UCI
20NEWS	19938	20	text	20NEWS
LETTERS	20000	26	image	UCI
SHUTTLE	58000	7	aerospace	UCI
MNIST	70000	10	image	MNIST
SEISMIC	98528	3	sensor	LIBSVM
MINIBOONE	130064	2	physics	UCI

PAJEK	`http://vlado.fmf.uni-lj.si/pub/networks/data/`
UCI	`http://archive.ics.uci.edu/ml/`
COIL	`http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php`
CMUTE	`http://www.cs.cmu.edu/~TextLearning/datasets.html`
RCV1	`http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/`
20NEWS	`http://people.csail.mit.edu/jrennie/20Newsgroups/`
MNIST	`http://yann.lecun.com/exdb/mnist/`
LIBSVM	`http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`

- `RCV1`: text documents from four classes, with 29992 words.
- `PENDIGITS`: the UCI *pen-based recognition of handwritten digits* dataset, originally with 16 dimensions.
- `MAGIC`: the UCI *MAGIC Gamma Telescope Data Set*, 11 numerical features.
- `20NEWS`: text documents from 20 newsgroups; 10,000 words with maximum information gain are preserved.
- `LETTERS`: the UCI *Letter Recognition Data Set*, 16 numerical features.
- `SHUTTLE`: the UCI *Statlog (Shuttle) Data Set*, 9 numerical features. The classes are imbalanced. Approximately 80% of the data belongs to class 1.
- `MNIST`: handwritten digit images, each of size 28×28.
- `SEISMIC`: the LIBSVM *SensIT Vehicle (seismic)* data, with 50 numerical features (the seismic signals) from the sensors on vehicles.
- `MINIBOONE`: the UCI *MiniBooNE particle identification Data Set*, 50 numerical features. This dataset is taken from the MiniBooNE experiment and is used to distinguish electron neutrinos (signal) from muon neutrinos (background).

7 Supplemental experiment results

7.1 Evolution curves: objective vs. iteration

Figure 1 shows the evolution curves of the t-SNE objective (cost function value) as a function of iteration for the compared algorithms. This supplements the results in the paper of objective vs. running time.

7.2 Average number of MM trials

The proposed backtracking algorithm for MM involves an inner loop for searching for ρ. We have recorded the numbers of trials in each outer loop iteration. The average number of trials is then calculated. The means and standard deviations across multiple runs are reported in Table 2.

We can see that for all datasets, the average number of trials is around two. The number is generally and slightly decreased with larger datasets. Two trials in an iteration means that ρ remains unchanged after the inner loop. This indicates potential speedups may be achieved in future work by keeping ρ constant in most iterations.

We have also recorded the average of number of func-

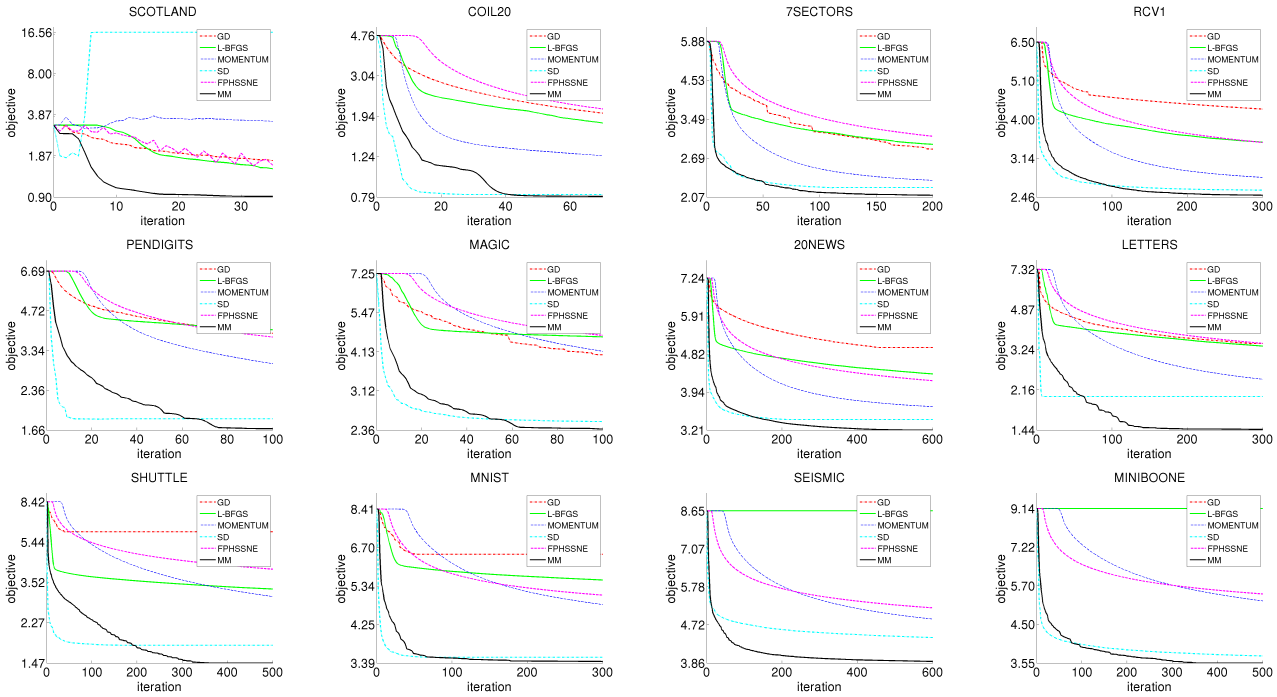

Figure 1: Evolution of the t-SNE objective (cost function value) as a function of iteration for the compared algorithms. The first and second rows were exactly calculated, while the third row uses Barnes-Hut approximation.

Table 2: Average number of MM trials and SD function calls over all iterations (mean ± standard deviation over 10 runs)

dataname	N	MM trials	SD fun.calls
SCOTLAND	108	2.00±0.00	21.91±1.88
COIL20	1.4K	1.94±0.11	1.92±0.29
7SECTORS	4.6K	2.00±0.00	3.06±0.34
RCV1	9.6K	2.00±0.00	2.56±0.09
PENDIGITS	11K	2.00±0.00	6.30±5.51
MAGIC	19K	2.00±0.00	2.12±0.06
20NEWS	20K	2.00±0.00	3.85±0.33
LETTERS	20K	2.00±0.01	13.30±4.00
SHUTTLE	58K	2.07±0.02	4.16±0.66
MNIST	70K	2.10±0.04	2.90±0.23
SEISMIC	99K	2.02±0.00	3.64±0.42
MINIBOONE	130K	2.07±0.01	3.29±0.20

tion calls to t-SNE objectives by the SD algorithm (last column of Table 2). It can be seen that the SD often requires more cost function calls than MM, which is one of the reasons that SD is slower.

7.3 ρ value and number of MM trials vs. iteration

However, the above average number of trials does not mean that ρ will remain nearly constant around its initial value. Actually ρ can vary greatly in different iterations. See Figure 2 for example, where the ranges of ρ in the first 300 iterations are $[3.6 \times 10^{-21}, 2.6 \times 10^{-4}]$, $[6.2 \times 10^{-8}, 1.6 \times 10^{-5}]$, and $[3.9 \times 10^{-9}, 4.0 \times 10^{-6}]$ for `COIL20`, `20NEWS`, and `MNIST`, respectively.

References

[1] A. Buja, D. Swayne, M. Littman, N. Dean, H. Hofmann, and L. Chen. Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, pages 444–472, 2008.

[2] M. Carreira-Perpiñán. The elastic embedding algorithm for dimensionality reduction. In *ICML*, pages 167–174, 2010.

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

[4] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-

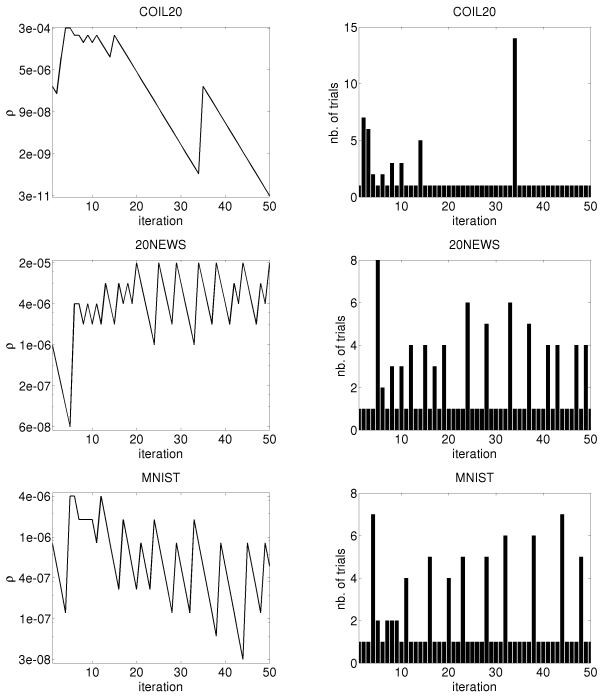

Figure 2: Backtracking behavior statistics of MM for t-SNE: (left) ρ values vs. iteration, and (right) number of trials in the backtracking algorithm. We only show the first 50 iterations for better visibility.

Saito divergence with application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[5] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

[6] G. Hinton and S. Roweis. Stochastic neighbor embedding. In *NIPS*, pages 833–840, 2002.

[7] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[8] F. Li and Y. Yang. A loss function analysis for classification methods in text categorization. In *ICML*, pages 472–479, 2003.

[9] L. Mason, J. Baxter, P. Bartlett, and M. Frean. Boosting algorithms as gradient descent in function space. In *NIPS*, pages 512–518, 1999.

[10] T. Minka. Divergence measures and message passing. Technical report, Microsoft Research, 2005.

[11] A. Noack. Energy models for graph clustering. *Journal of Graph Algorithms and Applications*, 11(2):453–480, 2007.

[12] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[13] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490, 2010.

[14] J. Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[15] Z. Yang, I. King, Z. Xu, and E. Oja. Heavy-tailed symmetric stochastic neighbor embedding. In *NIPS*, pages 2169–2177, 2009.

[16] Z. Yang, J. Peltonen, and S. Kaski. Scalable optimization of neighbor embedding for visualization. In *ICML*, pages 127–135, 2013.

[17] Z. Yang, J. Peltonen, and S. Kaski. Optimization equivalence of divergences improves neighbor embedding. In *ICML*, 2014.

[18] A. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15:915–936, 2003.

[19] W. Zangwill. *Nonlinear Programming: A Unified Approach*. Prentice Hall, Englewood Cliffs, New Jersey, 1969.