Supplemental Proofs for the ICML 2010 paper: The Elastic Embedding algorithm for dimensionality reduction

Miguel Á. Carreira-Perpiñán
EECS, School of Engineering, UC Merced

February 1, 2010

1 Bounds on the eigenvalues of $\mathbf{L} = \mathbf{L}^+ - \lambda\mathbf{L}^-$

The bounds in this section are useful to estimate the value $\lambda = \lambda_1^*$ above which $\mathbf{L} = \mathbf{L}^+ - \lambda\mathbf{L}^-$ stops being positive semidefinite. All the bounds hold even if $\mathbf{W}^+$ and $\mathbf{W}^-$ are sparse and not positive definite (as long as they are nonnegative), because $\mathbf{L}^+$ and $\mathbf{L}^-$ are always symmetric positive semidefinite. Some of the bounds are trivial to compute, as they are simple functions of the entries of $\mathbf{L}^+$ and $\mathbf{L}^-$; others require the eigenvalues of $\mathbf{L}^+$ and $\mathbf{L}^-$.

Theorem 1.1. *Let $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$, $0 = \lambda_1^+ \leq \lambda_2^+ \leq \cdots \leq \lambda_N^+$, $0 = \lambda_1^- \leq \lambda_2^- \leq \cdots \leq \lambda_N^-$ be the eigenvalues of $\mathbf{L} = \mathbf{L}^+ - \lambda\mathbf{L}^-$, $\mathbf{L}^+$ and $\mathbf{L}^-$, respectively (where $\mathbf{L}^+ = \mathbf{D}^+ - \mathbf{W}^+$ and $\mathbf{L}^- = \mathbf{D}^- - \mathbf{W}^-$, $\mathbf{D}^+ = \operatorname{diag}(\mathbf{W}^+\mathbf{1})$, $\mathbf{D}^- = \operatorname{diag}(\mathbf{W}^-\mathbf{1})$, and $\mathbf{W}^+$ and $\mathbf{W}^-$ are symmetric with nonnegative entries). Then we have the following bounds:*

- $\lambda_n(\mathbf{L}) \in [\lambda_n^+ - \lambda\lambda_N^-, \lambda_n^+]$, *so corresponding eigenvalues decrease by up to $-\lambda\lambda_N$, and if $\lambda < \lambda_2^+/\lambda_N^-$ then $\lambda_1(\mathbf{L}) = 0$ and $\mathbf{L}$ is positive semidefinite.*
- $\lambda_n(\mathbf{L}) \leq \min(\lambda_n^+, \lambda_{n+1}^+ - \lambda\lambda_2^-)$, $1 \leq n \leq N-1$ *and* $\lambda_1(\mathbf{L}) < 0$ *if* $\lambda > \min_{n=2,\ldots,N}(\lambda_n^+/\lambda_n^-)$.
- $\lambda_1(\mathbf{L}) < 0$ *if* $\lambda > \min_{n=1,\ldots,N}(L_{nn}^+/L_{nn}^-)$.

Hence:

- *If* $\lambda > \min\left(\frac{\lambda_2^+}{\lambda_2^-}, \ldots, \frac{\lambda_N^+}{\lambda_N^-}, \frac{L_{11}^+}{L_{11}^-}, \ldots, \frac{L_{NN}^+}{L_{NN}^-}\right)$ *then* $\lambda_1(\mathbf{L}) < 0$ *so* $\mathbf{L}$ *is not positive semidefinite.*
- *If* $\lambda < \lambda_2^+/\lambda_N^-$ *then* $\lambda_2(\mathbf{L}) > 0$ *so* $\mathbf{L}$ *is positive semidefinite.*
- $1 \leq n \leq N-1$: $\lambda_n(\mathbf{L}) \in [\lambda_n^+ - \lambda\lambda_N^-, \min(\lambda_n^+, \lambda_{n+1}^+ - \lambda\lambda_2^-)]$.

Proof. We will use the following properties (Horn and Johnson, 1986, pp. 181, 193 and 199), where $\mathbf{A}$ and $\mathbf{B}$ are $N \times N$ real symmetric matrices and $\lambda_n(\mathbf{A})$ is the nth algebraically largest eigenvalue of $\mathbf{A}$:

$$\lambda_n(\mathbf{A}) + \lambda_1(\mathbf{B}) \leq \lambda_n(\mathbf{A}+\mathbf{B}) \leq \lambda_n(\mathbf{A}) + \lambda_N(\mathbf{B}) \quad \text{(Weyl's theorem)} \tag{1}$$

$$\text{For } 1 \leq n \leq N: \ \lambda_n(\mathbf{A}+\mathbf{B}) \leq \min\{\lambda_i(\mathbf{A}) + \lambda_j(\mathbf{B}) : \ i+j = n+N\} \tag{2}$$

$\operatorname{diag}(\mathbf{A})$ majorises $\lambda(\mathbf{A})$, i.e., letting $a_{11}, \ldots, a_{NN}$ be sorted in ascending order:

$$\sum_{n=1}^{m} a_{nn} \geq \sum_{n=1}^{m} \lambda_n(\mathbf{A}) \text{ if } 1 \leq m \leq N \text{ and } \sum_{n=1}^{N} a_{nn} = \sum_{n=1}^{N} \lambda_n(\mathbf{A}). \tag{3}$$

Call $\mathbf{A} = \mathbf{L}^+$ and $\mathbf{B} = -\lambda\mathbf{L}^-$ in eqs. (1)–(3), so the eigenvalues of $\mathbf{B}$ are $-\lambda\lambda_N^- \leq \cdots \leq 0$ in increasing order.

- From eq. (1): $\lambda_n^+ - \lambda\lambda_N^- \leq \lambda_n \leq \lambda_n^+$. In particular for $n = 2$: if $\lambda < \lambda_2^+/\lambda_N^-$ then $0 < \lambda_2^+ - \lambda\lambda_N^- \leq \lambda_2$ so $\lambda_1 = 0$ and $\mathbf{L}$ is positive semidefinite.
- From eq. (2): $\lambda_n \leq \min\{\lambda_i^+ - \lambda\lambda_{i-n+1}^-, \ i = n, \ldots, N\}$. In particular, $\lambda_1 \leq \min(0, \min_{n=2,\ldots,N}(\lambda_n^+ - \lambda\lambda_n^-))$ and $\lambda_n \leq \min(\lambda_n^+, \lambda_{n+1}^+ - \lambda\lambda_2^-)$. Hence $\mathbf{L}$ is not positive semidefinite if $\lambda > \min_{n=2,\ldots,N}(\lambda_n^+/\lambda_n^-)$; and $\lambda_n < 0$ if $\lambda > \min_{i=n+1,\ldots,N}(\lambda_i^+/\lambda_{i-n+1}^-)$ where $\lambda_{i-n+1}^- > 0$ (and this bound seems very tight).

- From eq. (3): $\lambda_1 \leq \min_{n=1,\dots,N}(L^+_{nn} - \lambda L^-_{nn})$, hence $\mathbf{L}$ is not positive semidefinite if $\lambda > \min_{n=1,\dots,N} L^+_{nn}/L^-_{nn}$. This can also be obtained noting that a positive semidefinite matrix cannot have negative elements in the diagonal.

□

For the particular case where $\mathbf{W}^- = \mathbf{1}\mathbf{1}^T$ (a matrix of ones), $\mathbf{L}^- = N\mathbf{I} - \mathbf{1}\mathbf{1}^T$ has a single eigenvalue 0 (associated with eigenvector $\mathbf{1}$) and an eigenvalue N (associated with eigenvectors orthogonal to $\mathbf{1}$) with multiplicity $N-1$. We have the following.

Corollary 1.2. *If $\mathbf{W}^- = \mathbf{1}\mathbf{1}^T$ then:*

- $\lambda_n(\mathbf{L}) \in [\lambda^+_n - \lambda N, \lambda^+_n]$, *so corresponding eigenvalues decrease by up to $-\lambda N$.*
- $\lambda_n(\mathbf{L}) \leq \min(\lambda^+_n, \lambda^+_{n+1} - \lambda N),\ 1 \leq n \leq N-1$. *Hence $\lambda_1(\mathbf{L}) < 0$ if $\lambda > \lambda^+_2/N$.*
- $\lambda_1(\mathbf{L}) < 0$ *if* $\lambda > \min_{n=1,\dots,N} L^+_{nn}/(N-1)$.

Hence:

- *If* $\lambda > \min\left(\frac{\lambda^+_2}{N}, \frac{L^+_{11}}{N-1}, \dots \frac{L^+_{NN}}{N-1}\right)$ *then $\lambda_1(\mathbf{L}) < 0$ so $\mathbf{L}$ is not positive semidefinite.*
- $1 \leq n \leq N-1:\ \lambda_n(\mathbf{L}) \in [\lambda^+_n - \lambda N, \min(\lambda^+_n, \lambda^+_{n+1} - \lambda N)]$.

2 Another lower bound on the critical $\lambda = \lambda^*_1$

The stationary point equation (6) implies

$$\mathbf{X}(\mathbf{L}^+ - \lambda\tilde{\mathbf{L}}^-)\mathbf{X}^T = \frac{1}{2}\sum_{n,m=1}^{N}(w^+_{nm} - \lambda\tilde{w}^-_{nm})(\mathbf{x}_n - \mathbf{x}_m)(\mathbf{x}_n - \mathbf{x}_m)^T = \mathbf{0}$$

and in particular

$$\operatorname{tr}(\mathbf{X}(\mathbf{L}^+ - \lambda\tilde{\mathbf{L}}^-)\mathbf{X}^T) = \frac{1}{2}\sum_{n,m=1}^{N}(w^+_{nm} - \lambda\tilde{w}^-_{nm})\|\mathbf{x}_n - \mathbf{x}_m\|^2 = \mathbf{0}. \tag{4}$$

Consider $\lambda > 0$ and assume that all points $\mathbf{Y}$ are distinct and that there is at least one distinct pair (n,m) whose entry in $\mathbf{W}^-$ is positive. Note the following four cases for each pair (n,m):

1. $w^+_{nm} = 0,\ w^-_{nm} = 0$: $w^+_{nm} - \lambda\tilde{w}^-_{nm} = 0$
2. $w^+_{nm} > 0,\ w^-_{nm} = 0$: $w^+_{nm} - \lambda\tilde{w}^-_{nm} > 0$
3. $w^+_{nm} = 0,\ w^-_{nm} > 0$: $w^+_{nm} - \lambda\tilde{w}^-_{nm} < 0$
4. $w^+_{nm} > 0,\ w^-_{nm} > 0$: $w^+_{nm} - \lambda\tilde{w}^-_{nm}$ can be $=, >, < 0$.

If $w^+_{nm} > 0$ for all pairs (i.e., no pairs in case 1 or 3) and any one pair separates, then for the sum in eq. (4) to be zero there must exist at least one pair in case 4 with $w^+_{nm} - \lambda\tilde{w}^-_{nm} \leq 0 \Leftrightarrow 0 \leq \|\mathbf{x}_n - \mathbf{x}_m\|^2 \leq \log(\lambda w^-_{nm}/w^+_{nm})$. This is also true if $w^+_{nm} = 0$ only when $w^-_{nm} = 0$ and at least one pair with $w^-_{nm} > 0$ separates. If there are pairs where $w^+_{nm} = 0$ and $w^-_{nm} > 0$ then it does not follow that there must exist at least one pair in case 4 with $w^+_{nm} - \lambda\tilde{w}^-_{nm} \leq 0$. In summary, defining

$$\mathcal{S} = \left\{(n,m) \text{ with } 1 \leq n,m \leq N:\ w^-_{nm} > 0\right\} \qquad l'_1 = \min_{n,m\in\mathcal{S}}(w^+_{nm}/w^-_{nm})$$

then $\lambda \geq l'_1$ if at least one pair with $w^-_{nm} > 0$ separates, so l'_1 is a lower bound for λ^*_1 in that case. However, it is a very coarse bound unless $\mathbf{W}^-$ has some zero entries: if we sparsify $\mathbf{W}^-$ (by making zero some of its elements), the set $\mathcal{S}$ loses pairs and l'_1 increases or stays the same.

If $\lambda < l'_1$ and $w^+_{nm} > 0$ for all pairs then $\|\mathbf{x}_n - \mathbf{x}_m\|^2 = 0$ and $w^+_{nm} - \lambda\tilde{w}^-_{nm} > 0\ \forall n,m$, so $\mathbf{D}^+ - \lambda\tilde{\mathbf{D}}^-$ is positive definite and $\mathbf{D}^+ - \lambda\tilde{\mathbf{W}}^-$ is strictly diagonally dominant. Besides, if $\mathbf{A}$ is an $N \times N$ symmetric, strictly diagonally dominant matrix with positive diagonal elements and negative or zero off-diagonal elements, then for any $\mathbf{x} \neq \mathbf{0}$:

$$\sum_{n,m=1}^{N} a_{nm}x_nx_m = \sum_{n=1}^{N} a_{nn}x_n^2 + \sum_{n\neq m}^{N} a_{nm}x_nx_m > \sum_{n\neq m}^{N} -a_{nm}x_n^2 + a_{nm}x_nx_m = -\frac{1}{2}\sum_{n\neq m}^{N} a_{nm}(x_n - x_m)^2 \geq 0$$

so $\mathbf{D}^+ - \lambda\tilde{\mathbf{W}}^-$ is positive definite.

3 Upper and lower bounds on the critical $\lambda = \lambda_1^*$

From theorem 1.1 and the lower bound l_1' from the previous section we have that $\lambda_1^* \in [l_1, u_1]$ with

$$l_1 = \max\left(\frac{\lambda_2^+}{\lambda_N^-}, l_1'\right) \qquad u_1 = \min\left(\frac{\lambda_2^+}{\lambda_2^-}, \ldots, \frac{\lambda_N^+}{\lambda_N^-}, \frac{L_{11}^+}{L_{11}^-}, \ldots, \frac{L_{NN}^+}{L_{NN}^-}\right). \tag{5}$$

Note that $L_{nn}^{\pm} = d_n^{\pm}$ (the degree matrix).

There is one particular case where we can determine λ_1^* exactly, namely when $\mathbf{W}^- = \mathbf{1}\mathbf{1}^T$ (i.e., it does not contain $\|\mathbf{y}_n - \mathbf{y}_m\|^2$). Then $\mathbf{L} = \mathbf{L}^+ - \lambda(N\mathbf{I} - \mathbf{1}\mathbf{1}^T)$, which has a null eigenvalue associated with the eigenvector $\mathbf{1}$, and its remaining eigenvectors are the same ones as those of $\mathbf{L}^+$ and associated with the corresponding eigenvalues of $\mathbf{L}^+$ but shifted by λN. Thus $\lambda_1^* = \lambda_2^+/N$, and for λ just larger than λ_1^*, each dimension of $\mathbf{X}$ would expand along the second trailing eigenvector of $\mathbf{L}^+$ (which is the 1D LE solution using the unnormalised graph Laplacian). In this case the lower bound λ_2^+/λ_N^- and the upper bound λ_2^+/λ_2^- are tight.

4 Design of search directions and proof of convergence

(This follows from section 4.3 in the paper.) The stationary points of $E(\mathbf{X}; \lambda)$ for fixed λ satisfy

$$\mathbf{G}(\mathbf{X}; \lambda) = \frac{\partial E}{\partial \mathbf{X}} = 4\mathbf{X}(\mathbf{L}^+ - \lambda\tilde{\mathbf{L}}^-) = 4\mathbf{X}(\mathbf{D}^+ - \mathbf{W}^+ - \lambda\tilde{\mathbf{D}}^- + \lambda\tilde{\mathbf{W}}^-) = \mathbf{0}. \tag{6}$$

Rearranging the gradient (a matrix of $N \times L$) as a splitting $\mathbf{G} = \mathbf{X}(\mathbf{A} + \mathbf{B}) = \mathbf{0}$, where $\mathbf{A}$ is symmetric positive definite and $\mathbf{B}$ is symmetric, we obtain a fixed point iteration $\mathbf{X} \leftarrow -\mathbf{X}\mathbf{B}\mathbf{A}^{-1}$. Although this iteration does not always converge, it does suggest a search direction $\mathbf{\Delta} = -\mathbf{X}\mathbf{B}\mathbf{A}^{-1} - \mathbf{X} = -\mathbf{X}(\mathbf{B}\mathbf{A}^{-1} + \mathbf{I}) = -\mathbf{G}\mathbf{A}^{-1}$ along which we can decrease E with a line search $\mathbf{X} \leftarrow \mathbf{X} + \eta\mathbf{\Delta}$ for $\eta \geq 0$. To prove that this algorithm converges, we first show that the direction $\mathbf{\Delta}$ is descent, and that it never becomes too close to being orthogonal to the gradient.

Proposition 4.1. *The direction $\mathbf{\Delta} = -\mathbf{G}\mathbf{A}^{-1}$ is descent if $\mathbf{G} \neq \mathbf{0}$.*

Proof. Using the identity $\operatorname{vec}(\mathbf{M})^T \operatorname{vec}(\mathbf{N}) = \operatorname{tr}(\mathbf{M}^T\mathbf{N})$ (where $\operatorname{vec}(\mathbf{M})$ concatenates the columns of matrix $\mathbf{M}$ into a column vector), the scalar product of the search direction with the gradient is $\operatorname{vec}(\mathbf{\Delta})^T \operatorname{vec}(\mathbf{G}) = -\operatorname{tr}(\mathbf{G}\mathbf{A}^{-1}\mathbf{G}^T) < 0$, so $\cos(\widehat{\mathbf{\Delta}, \mathbf{G}}) < 0$ and $\mathbf{\Delta}$ is a descent direction. □

Proposition 4.2. *Let $M \geq 1$ and $\delta = 1/M^2$. If $\operatorname{cond}(\mathbf{A}) \leq M$, then $-\cos(\widehat{\mathbf{\Delta}, \mathbf{G}}) \geq \delta$.*

Proof. Write $\mathbf{A}^{-1} = \mathbf{U}\mathbf{S}\mathbf{U}^T$ (SVD decomposition) where $\mathbf{U}$ is orthogonal and $\mathbf{S}$ is diagonal positive definite (with $s_N \geq \cdots \geq s_1 > 0$), all of $N \times N$; and let $\mathbf{Z} = (\mathbf{z}_1 \ldots \mathbf{z}_N) = \mathbf{G}\mathbf{U}$ of $L \times N$. Then he have:

$$-\cos(\widehat{\mathbf{\Delta}, \mathbf{G}}) = \frac{\operatorname{tr}(\mathbf{G}\mathbf{A}^{-1}\mathbf{G}^T)}{\sqrt{\operatorname{tr}(\mathbf{G}\mathbf{G}^T)\operatorname{tr}(\mathbf{G}\mathbf{A}^{-2}\mathbf{G}^T)}} = \frac{\operatorname{tr}(\mathbf{Z}\mathbf{S}\mathbf{Z}^T)}{\sqrt{\operatorname{tr}(\mathbf{Z}\mathbf{Z}^T)\operatorname{tr}(\mathbf{Z}\mathbf{S}^2\mathbf{Z}^T)}} = \frac{\sum_{n=1}^N s_n t_n^2}{\sqrt{\left(\sum_{n=1}^N t_n^2\right)\left(\sum_{n=1}^N s_n^2 t_n^2\right)}}$$

where we have written $\operatorname{tr}(\mathbf{Z}\mathbf{S}\mathbf{Z}^T) = \sum_{n=1}^N s_n \operatorname{tr}(\mathbf{z}_n\mathbf{z}_n^T) = \sum_{n=1}^N s_n t_n^2$, with $t_n^2 = \operatorname{tr}(\mathbf{z}_n\mathbf{z}_n^T) = \operatorname{tr}(\mathbf{z}_n^T\mathbf{z}_n) \geq 0$. Let $v_n^2 = t_n^2 / \sum_{m=1}^N t_m^2 \in [0, 1]$ for each $n = 1, \ldots, N$. Then we have

$$-\cos(\widehat{\mathbf{\Delta}, \mathbf{G}}) = \frac{\sum_{n=1}^N s_n v_n^2}{\sqrt{\sum_{n=1}^N s_n^2 v_n^2}} \geq \frac{s_1}{s_N} = \frac{1}{\operatorname{cond}(\mathbf{A})^2} \geq \delta > 0 \qquad \forall \mathbf{v} \in \mathbb{R}^N \text{ with } \|\mathbf{v}\| = 1$$

with $\delta = 1/M^2$, since $\operatorname{cond}(\mathbf{A}) \leq M$ with $M \geq 1$. □

Now, consider the sequence of iterates $\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \ldots$ where $\mathbf{X}_i$ is obtained from a line search using the direction $\mathbf{\Delta}_i = -\mathbf{G}_i\mathbf{A}_i^{-1}$ and satisfying the Wolfe conditions, and $\mathbf{A}_i$ is positive definite with $\operatorname{cond}(\mathbf{A}_i) \leq M$ for a fixed $M \geq 1$. E is bounded below by 0 and $\partial E/\partial \mathbf{X}$ is Lipschitz continuous in $\mathbb{R}^{N \times L}$. Then, from Zoutendijk's theorem (th. 3.2 in Nocedal and Wright, 2006), the sequence converges to a stationary point of E from any initial $\mathbf{X}_0 \in \mathbb{R}^{L \times N}$.

References

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, U.K., 1986.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer-Verlag, New York, second edition, 2006.