

Supplementary material for: Fast Training of Nonlinear Embedding Algorithms

1 Equations for Hessians

In the paper we have used general form (2) and (3) to write the equations for the Hessian. Here we want to define explicitly the parts that participate in those equations:

The first graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is represented by $N \times N$ matrix and defined on the positive data-independent weights. It is, thus, always positive semidefinite. This graph Laplacian defines the spectral direction (SD) - the leading method in optimizing all of the methods described in the paper. The $Nd \times Nd$ matrix $4\mathbf{L} \otimes \mathbf{I}_d$ represents a zero matrix that has d equal matrices \mathbf{L} placed along the diagonal.

The matrix \mathbf{L}^{xx} is $Nd \times Nd$ dimensional data-dependent matrix. For a given dimensions (i, j) this matrix decouples into d^2 $N \times N$ dimensional matrices $\mathbf{L}_{i*,j*}^{xx}$ each of which is a graph Laplacian defined for a weights $\mathbf{W}_{i*,j*}^{xx}$ and degree matrix $\mathbf{D}_{i*,j*}^{xx}$. Because of data-dependency, the weights $\mathbf{W}_{i*,j*}^{xx}$ can be negative. However, as shown below for each case separately, the graph Laplacians $\mathbf{L}_{i*,i*}^{xx}$ taken along the diagonal of \mathbf{L}^{xx} (i.e. when $i = j$) all have positive weights.

Finally, the graph Laplacian \mathbf{L}_q that appears in the Hessian for normalized symmetric models is $N \times N$ matrix defined for positive data-dependent weights \mathbf{W}_q and degree matrix \mathbf{D}_q .

We give the expressions for the weights of the graph Laplacians above.

For EE:

- $w_{nm} = w_{nm}^+ - \lambda w_{nm}^- e^{-\|\mathbf{x}_n - \mathbf{x}_m\|^2}.$
- $w_{in,jm}^{xx} = \lambda w_{nm}^- e^{-\|\mathbf{x}_n - \mathbf{x}_m\|^2} (x_{in} - x_{im})(x_{jn} - x_{jm})^1.$

For s-SNE:

- $w_{nm} = p_{nm} - \lambda q_{nm}.$
- $w_{in,jm}^{xx} = \lambda q_{nm} (x_{mj} - x_{nj})(x_{mi} - x_{ni})^1.$
- $w_{nm}^q = q_{nm}.$

For t-SNE:

- $w_{nm} = (p_{nm} - q_{nm})(1 + \|\mathbf{x}_m - \mathbf{x}_n\|^2)^{-1}.$
- $w_{in,jm}^{xx} = -(p_{nm} - 2\lambda q_{nm})(1 + \|\mathbf{x}_n - \mathbf{x}_m\|^2)^{-2} (x_{mi} - x_{ni})(x_{mj} - x_{nj})^2.$
- $w_{nm}^q = q_{nm}(1 + \|\mathbf{x}_n - \mathbf{x}_m\|^2)^{-2}.$

2 Algorithms

Here we give more details description of how the search direction is computed for the algorithms in the experimental section:

- **The diagonal of the Hessian (DiagH).** The Hessian approximation is given by $\mathbf{B}_k = 4\mathbf{D} \otimes \mathbf{I}_d - 8\mathbf{D}_{i*,i*}^{xx}$. This matrix is diagonal, so the computation of the search direction involves dividing every element of the gradient by corresponding element of the \mathbf{B}_k matrix.

¹Note, that when $i = j$ the weights are positive.

²The weights are positive if we ignore p_{nm} part.

- **The spectral direction (SD).** The Hessian approximation is given by $\mathbf{B}_k = 4\mathbf{L} \otimes \mathbf{I}_d$. Before the optimization, the Cholesky factorization is precomputed and cached. The search direction is given by solving two triangular systems.
- **The partial Hessian method (SD-).** The Hessian approximation is given by $\mathbf{B}_k = 4\mathbf{L} \otimes \mathbf{I}_d - 8\mathbf{L}_{i^*, i^*}^{xx}$. The search direction involves solving d linear systems each of the size $N \times N$. We can further approximate this computation by solving this system iteratively with CG algorithm.

3 Proofs

We give a derivation of the result for the rate of linear convergence in p. 4 of the submission. Consider an objective function f to be minimized using a search direction obtained from $\mathbf{B}_k \mathbf{p}_k = -\nabla f(\mathbf{x}_k)$ where $\mathbf{B}_k = \mathbf{B}(\mathbf{x}_k)$ is a positive definite partial Hessian for $k = 0, 1, 2, \dots$. Under the assumptions of theorem 3.1 in the paper, \mathbf{x}_k converges to a stationary point \mathbf{x}^* . Assume that $\mathbf{B}(\mathbf{x})$ is Lipschitz continuous in the region of interest (that is, $\exists L > 0$: $\|\mathbf{B}(\mathbf{x}) - \mathbf{B}(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|$ for any two points \mathbf{x}, \mathbf{y} in the region) with bounded condition number, and that we take unit steps in the line search. Then:

$$\mathbf{x}_k + \mathbf{p}_k - \mathbf{x}^* = \mathbf{x}_k - \mathbf{x}^* - \mathbf{B}_k^{-1} \nabla f(\mathbf{x}_k) = \mathbf{B}_k^{-1} (\mathbf{B}_k(\mathbf{x}_k - \mathbf{x}^*) - (\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*)))$$

since $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Applying Taylor's theorem (Nocedal and Wright, 2006, th. 2.1) to $\nabla f(\mathbf{x}_k)$ we have

$$\begin{aligned} \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*) &= \int_0^1 \nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) (\mathbf{x}_k - \mathbf{x}^*) dt \\ &= \int_0^1 \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) (\mathbf{x}_k - \mathbf{x}^*) dt + \int_0^1 (\nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))) (\mathbf{x}_k - \mathbf{x}^*) dt. \end{aligned}$$

Hence:

$$\begin{aligned} &\|\mathbf{B}_k(\mathbf{x}_k - \mathbf{x}^*) - (\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*))\| = \\ &\left\| \int_0^1 (\mathbf{B}(\mathbf{x}_k) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))) (\mathbf{x}_k - \mathbf{x}^*) dt - \int_0^1 (\nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))) (\mathbf{x}_k - \mathbf{x}^*) dt \right\| \\ &\leq \left\| \int_0^1 (\mathbf{B}(\mathbf{x}_k) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))) (\mathbf{x}_k - \mathbf{x}^*) dt \right\| + \left\| \int_0^1 (\nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))) (\mathbf{x}_k - \mathbf{x}^*) dt \right\| \\ &\leq \int_0^1 \|\mathbf{B}(\mathbf{x}_k) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))\| \|\mathbf{x}_k - \mathbf{x}^*\| dt + \int_0^1 \|\nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))\| \|\mathbf{x}_k - \mathbf{x}^*\| dt \\ &\leq \int_0^1 L \|\mathbf{x}_k - \mathbf{x}^*\|^2 t dt + \int_0^1 \|\nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))\| \|\mathbf{x}_k - \mathbf{x}^*\| dt \\ &= \frac{1}{2} L \|\mathbf{x}_k - \mathbf{x}^*\|^2 + \left(\int_0^1 \|\nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))\| dt \right) \|\mathbf{x}_k - \mathbf{x}^*\| \end{aligned}$$

so the upper bound contains a second-order term (on $\|\mathbf{x}_k - \mathbf{x}^*\|$) and a first-order term. The latter vanishes if using $\mathbf{B}_k = \nabla^2 f(\mathbf{x}_k)$ (i.e., Newton's method) and results in quadratic convergence. Otherwise, convergence is linear and we can estimate its rate as follows:

$$\begin{aligned} \|\mathbf{x}_k + \mathbf{p}_k - \mathbf{x}^*\| &\leq \|\mathbf{B}_k^{-1}\| \|\mathbf{B}_k(\mathbf{x}_k - \mathbf{x}^*) - (\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*))\| \\ &\leq \mathcal{O}(\|\mathbf{x}_k - \mathbf{x}^*\|^2) + \|\mathbf{B}_k^{-1}\| \left(\int_0^1 \|\nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))\| dt \right) \|\mathbf{x}_k - \mathbf{x}^*\|. \end{aligned}$$

When $\mathbf{x}_k - \mathbf{x}^*$ is small, the second-order term is negligible and the first-order term becomes approximately $r \|\mathbf{x}_k - \mathbf{x}^*\|$ with $r = \|\mathbf{B}^{-1}(\mathbf{x}^*) \nabla^2 f(\mathbf{x}^*) - \mathbf{I}\|$.

References

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