ENTROPIC AFFINITIES: PROPERTIES AND EFFICIENT NUMERICAL COMPUTATION

Max Vladymyrov and **Miguel Á. Carreira-Perpiñán.** EECS, UC Merced, USA

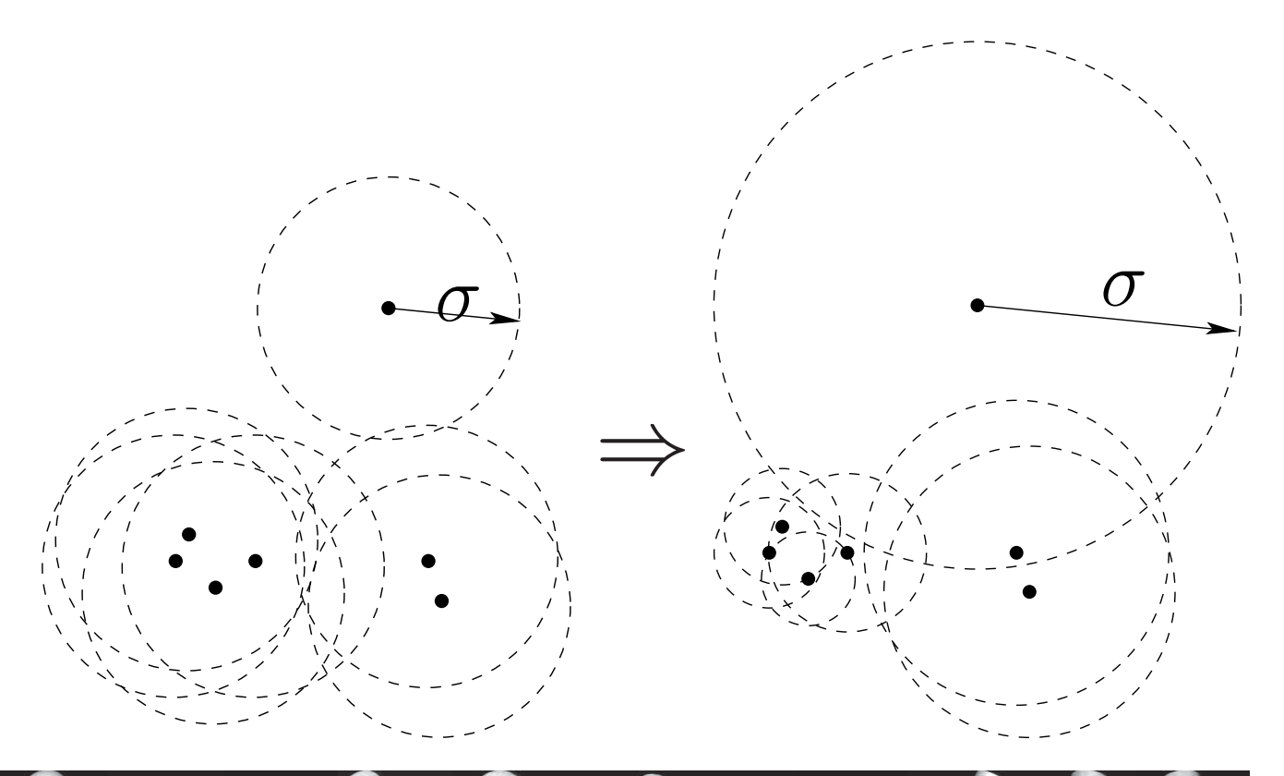
1 Abstract

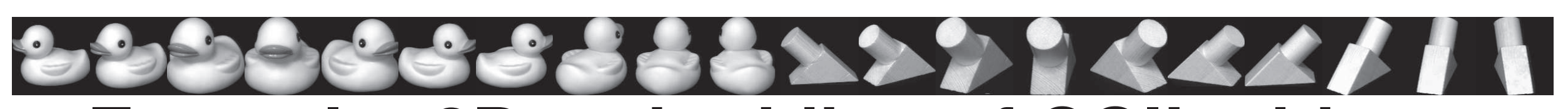
Gaussian affinities are commonly used in graph-based methods such as spectral clustering or nonlinear embedding. Hinton and Roweis (2003) introduced a way to set the scale individually for each point so that it has a distribution over neighbors with a desired perplexity, or effective number of neighbors. This gives very good affinities that adapt locally to the data but are harder to compute. We study the mathematical properties of these entropic affinities and show that they implicitly define a continuously differentiable function in the input space and give bounds for it. We then devise a fast algorithm to compute the widths and affinities, based on robustified, quickly convergent root-finding methods combined with a tree- or density-based initialization scheme that exploits the slowly-varying behavior of this function. This algorithm is nearly optimal and much more accurate and fast than the existing bisection-based approach, particularly with large datasets, as we show with image and text data.

2 Motivation

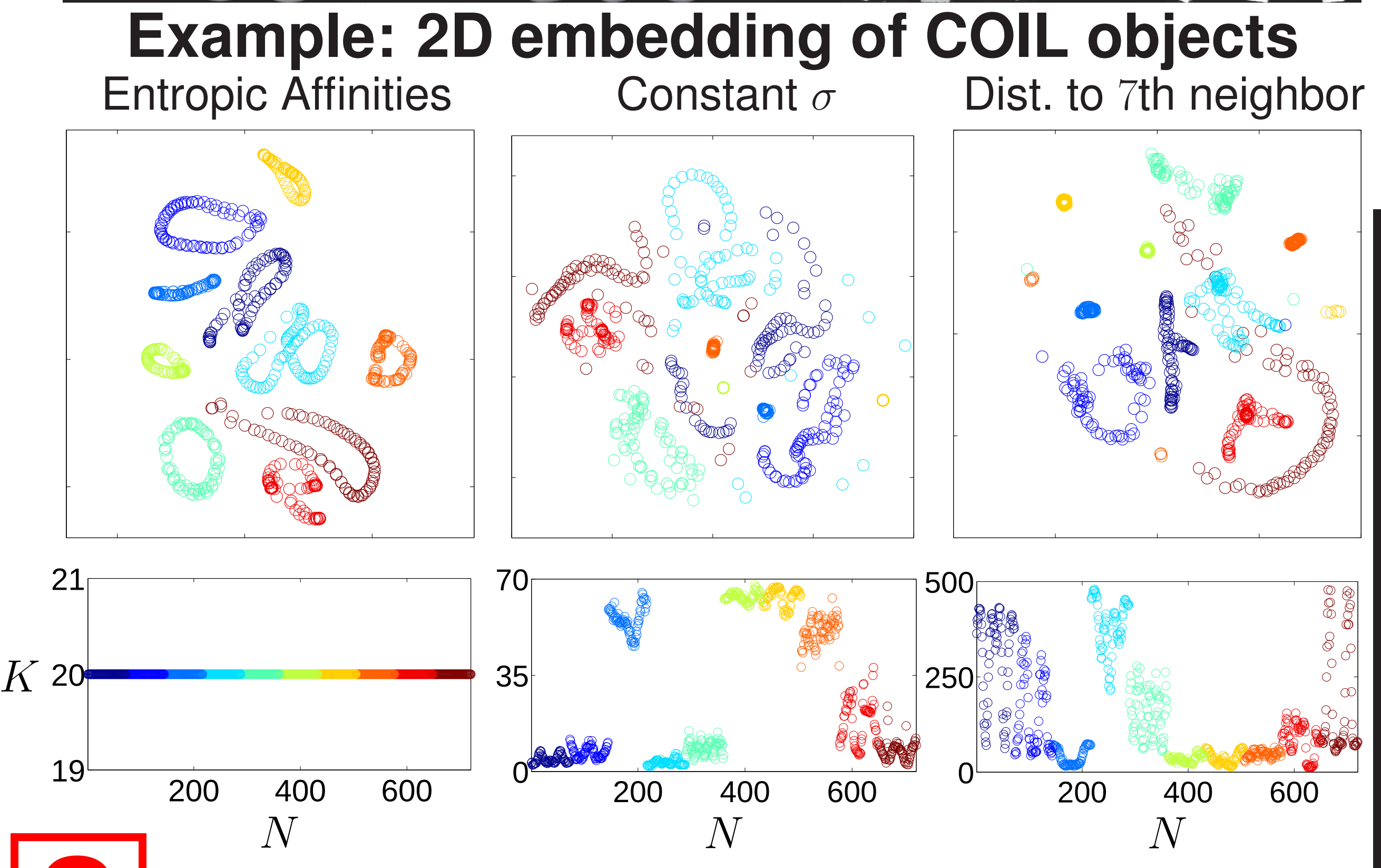
The bandwidth σ parameter in Gaussian affinities is data-dependent and usually set using some rule-of-thumb. Too high values result in almost identical interaction between neighboring points, while too low values result in almost zero interaction.

Better results are achieved when σ is set separately for each data point and takes into account the whole distribution of distances.





Example: 2D embedding of COIL objects



3 Entropic affinities

In the entropic affinities, the bandwidth is set individually for each point such that it has a distribution over neighbors with fixed perplexity K (Hinton & Roweis, 2003).

For a point $\mathbf{x} \in \mathbb{R}^D$, consider the posterior distribution of an isotropic kernel density estimator of width σ defined on a set $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^D$. We have a discrete distribution $p(\mathbf{x};\sigma)$ with probabilities for $n = 1, \dots, N$

$$p_n(\mathbf{x};\sigma) = \frac{K\left(\|(\mathbf{x}-\mathbf{x}_n)/\sigma\|^2\right)}{\sum_{k=1}^N K\left(\|(\mathbf{x}-\mathbf{x}_k)/\sigma\|^2\right)} = \frac{K\left((d_n/\sigma)^2\right)}{\sum_{k=1}^N K\left((d_k/\sigma)^2\right)}$$

The entropy of the distribution is defined as

$$\begin{aligned} H(\mathbf{x},\sigma) &= -\textstyle\sum_{n=1}^N p_n(\mathbf{x},\sigma)\log(p_n(\mathbf{x},\sigma)) \\ &= -\textstyle\sum_{n=1}^N p_n(\mathbf{x},\sigma)\log K_n + \log\sum_{n=1}^N K_n \end{aligned}$$

For the Gaussian kernel it becomes (define $\beta = 1/2\sigma^2$):

$$H(\mathbf{x},\beta) = \beta\sum_{n=1}^N p_n(\mathbf{x},\beta)d_n^2 + \log\sum_{n=1}^N \exp(-d_n^2\beta).$$

Define the partition function $Z(\beta) = \sum_{n=1}^N \exp(-d_n^2\beta)$ and moments $m_k(\beta) = \sum_{n=1}^N p_n d_n^{2k}$. Then:

$$H(\mathbf{x},\beta) = \beta m_1 + \log Z, \quad H'_\beta(\mathbf{x},\beta) = -\beta(m_2 - m_1^2)$$
$$H''_\beta(\mathbf{x},\beta) = \beta(m_3 - 3m_2m_1 + 2m_1^3) + m_1^2 - m_2.$$

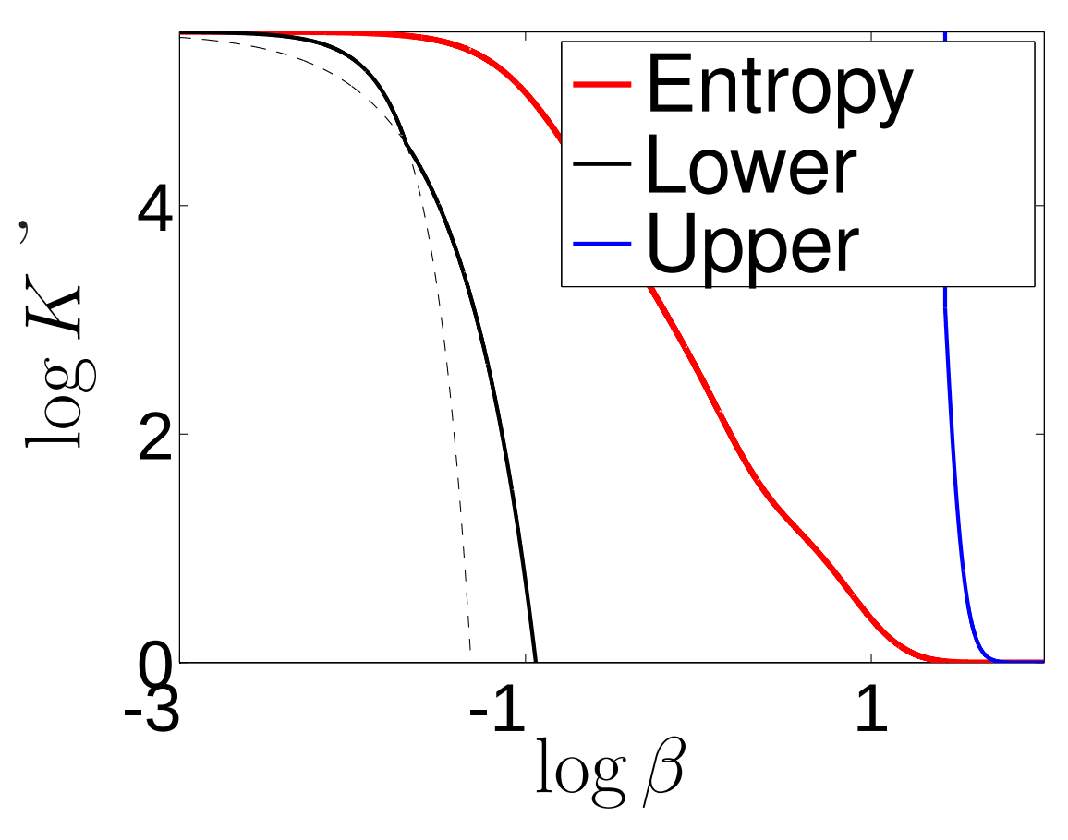
Setting σ (or β) given the perplexity K implies finding the root of the nonlinear equation: $H(\mathbf{x},\beta) - \log K = 0$.

4 Entropic affinities: Properties

1. The root-finding problem is well defined for any value of $\beta > 0$ and has a unique root $\beta(\mathbf{x})$ for any $K \in (0, N)$, so $H(\mathbf{x},\beta)$ is invertible over β.
2. This unique inverse is a continuously differentiable function $\beta(\mathbf{x},K)$ for all $K \in (0, \log N)$ and $\mathbf{x} \in \mathbb{R}^D$.
3. The root is in the interval $[\beta_L, \beta_U]$ where the bounds are, for every $\mathbf{x}$ and K:

$$\beta_L = \max\left(\frac{N\log\frac{N}{K}}{(N-1)\Delta_N^2}, \sqrt{\frac{\log\frac{N}{K}}{d_N^4 - d_1^4}}\right),$$
$$\beta_U = \frac{1}{\Delta_2^2}\log\left(\frac{p_1}{1-p_1}(N-1)\right),$$

where $d_1^2 < d_2^2 < \dots < d_N^2$, $\Delta_N^2 = d_N^2 - d_1^2$, $\Delta_2^2 = d_2^2 - d_1^2$ and p_1 is the unique solution in the interval $[3/4, 1]$ of the equation: $2(1-p_1)\log\frac{N}{2(1-p_1)} = \log\left(\min(\sqrt{2N}, K)\right)$. The bounds can be computed in $\mathcal{O}(1)$ for each point.



5 Entropic affinities: Computation

For each point, we run a quickly-convergent, robustified root-finding algorithm with a clever initialization.

1. Choice of root-finding algorithm

We focus of the following methods:

1. Derivative-free methods. They define an interval around the root and iteratively shrink it. Slow but guaranteed convergence.
   - Bisection - first-order convergence.
   - Brent - superlinear convergence.
   - Ridder - quadratic convergence.
2. Derivative-based methods. Generally do not have convergence guarantees. High-order convergence.
   - Newton - second order, approximation with a line.
   - Euler - third order, approximation with a parabola.
   - Halley - third order, approximation with a hyperbola.

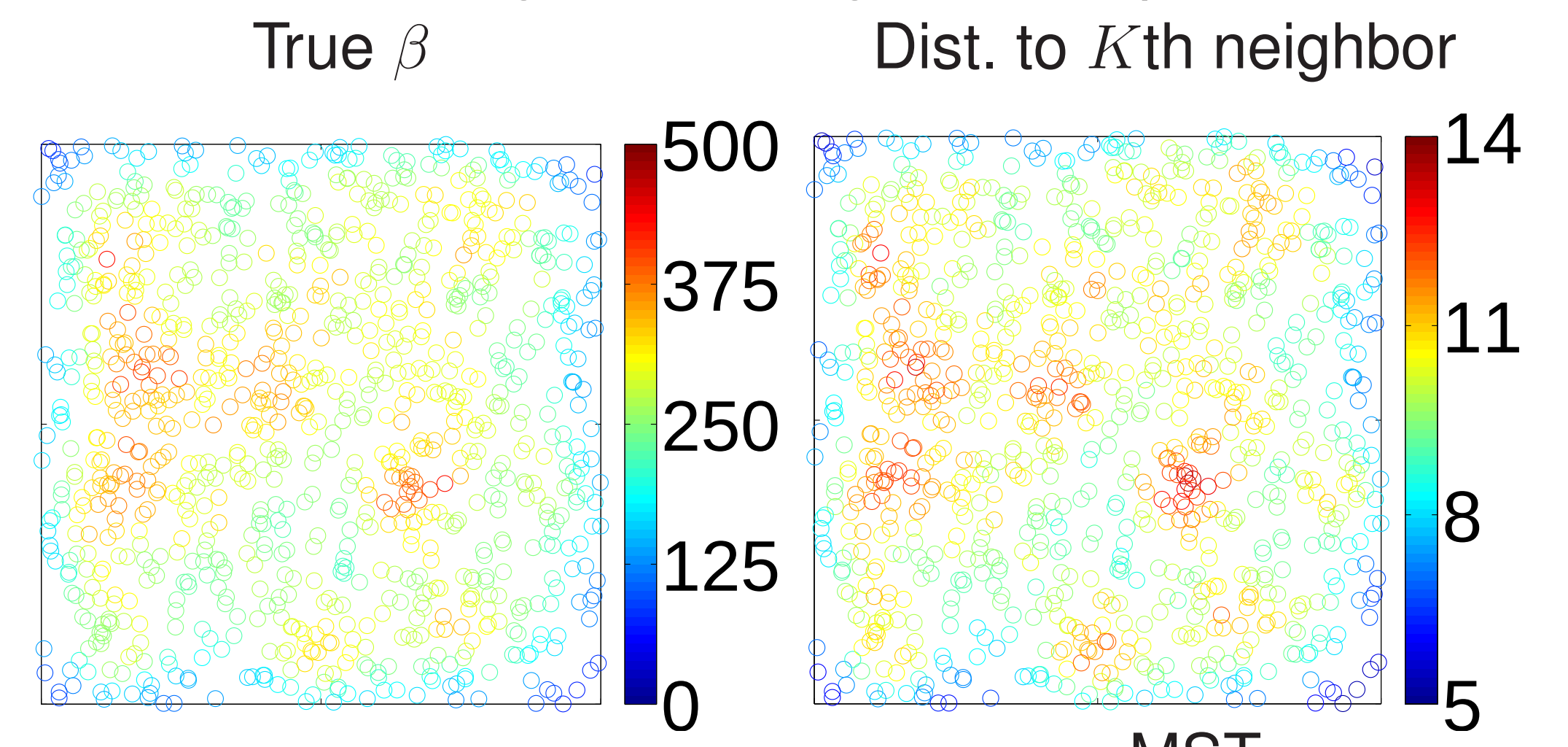
We robustify the derivative-based methods to achieve convergence from any starting point by embedding them in a bisection loop:

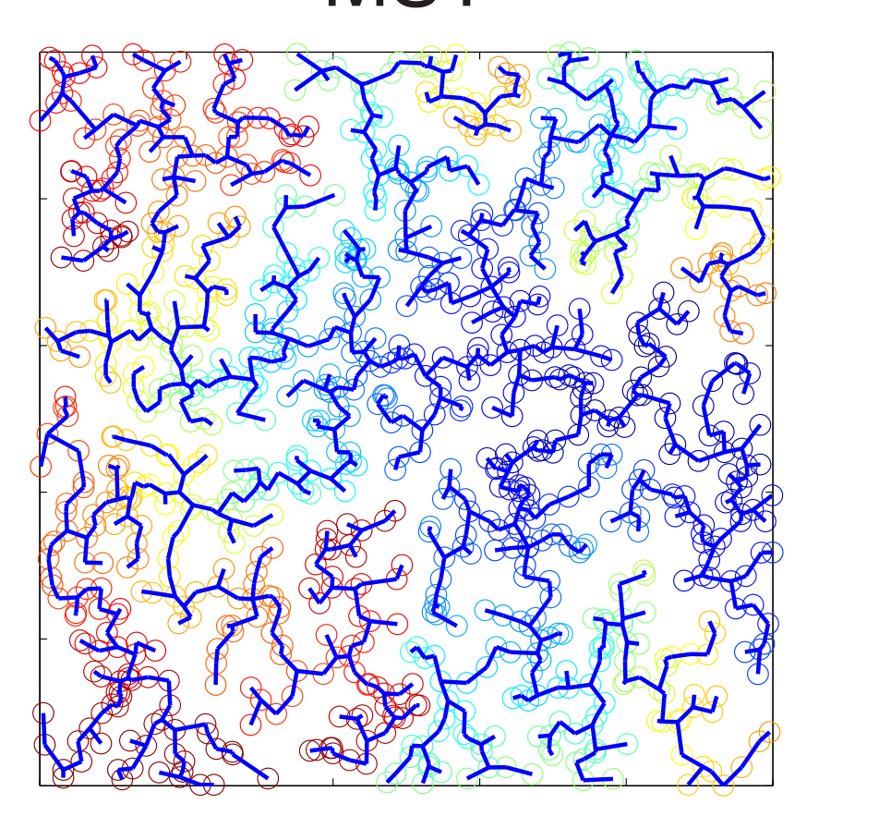
```
Input: initial $\beta$, perplexity $K$, distances $d_1^2, \dots, d_N^2$
compute bounds $\mathcal{B}$
while $true$ do
  for $k = 1$ to maxit do
    compute $\beta$ using a derivative-based method
    if tolerance achieved return $\beta$
    if $\beta \notin \mathcal{B}$ exit for loop
    update $\mathcal{B}$
  end for
  compute $\beta$ using bisection
  update $\mathcal{B}$
end while
```

2. Initialization

The initialization should be close to the root:

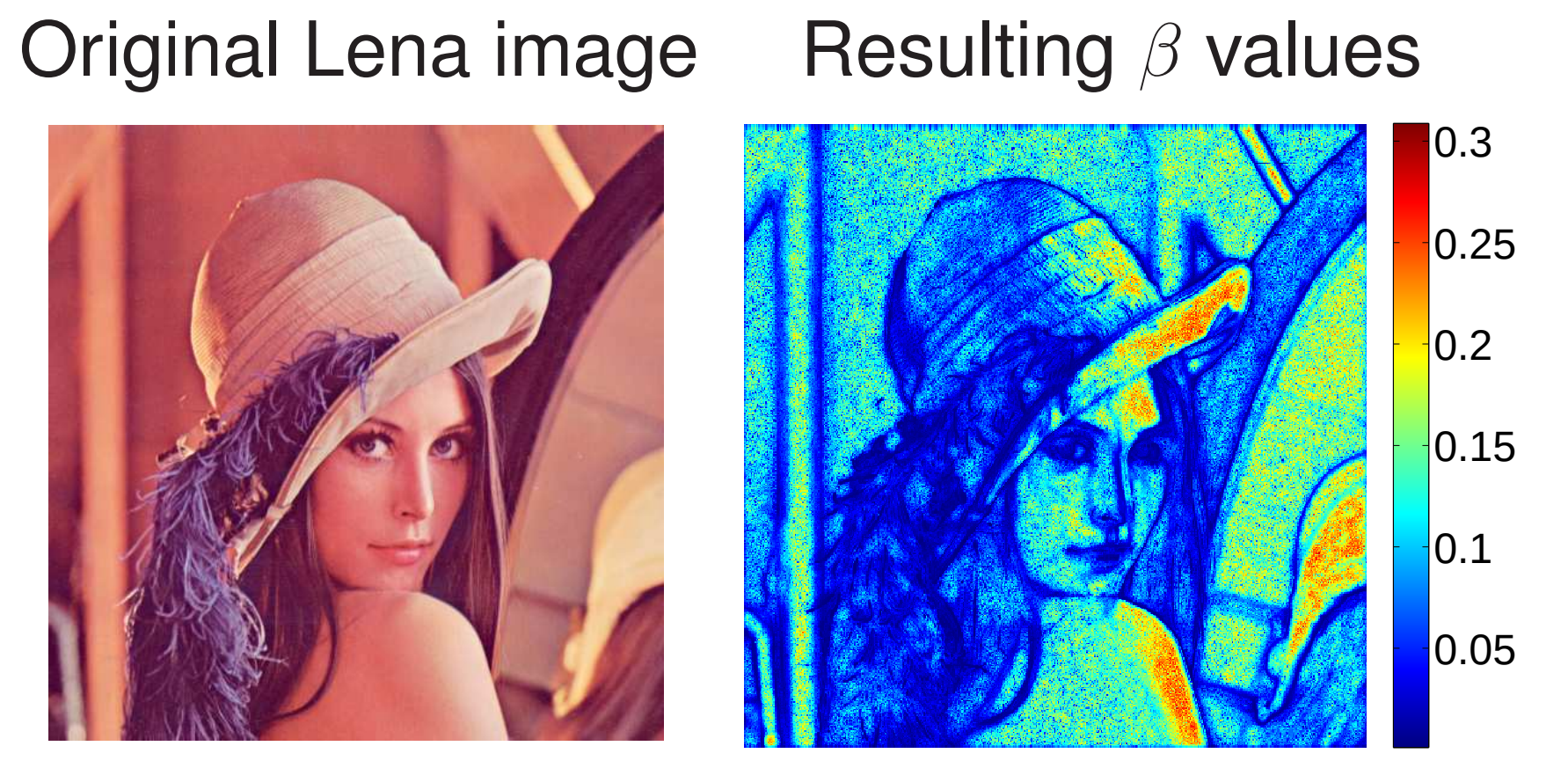
1. **Simple initializations,** e.g. the middle of the bounds or the distance to kth nearest neighbor, are typically far from the root and require more iterations.
2. **Sequential or tree order,** based on the correlation between β and the structure of the dataset. Each new point is initialized from the solution to its predecessor. Good orders require just over one iteration on average:
   - MST, *local* strategy: continuous changes in $\mathbf{x}$ lead to continuous changes in β, so nearby points have similar β values. Build a minimum spanning tree around the data and process the points in level-order, so parents are initialized before children.
   - $\mathcal{D}_K$, *density* strategy: for the entropy to remain constant, the resulting β must be larger in dense regions and smaller in sparser ones. We sort the points $\mathbf{x}_1, \dots, \mathbf{x}_N$ according to the distance to their kth nearest neighbor. This gives a sequential order.



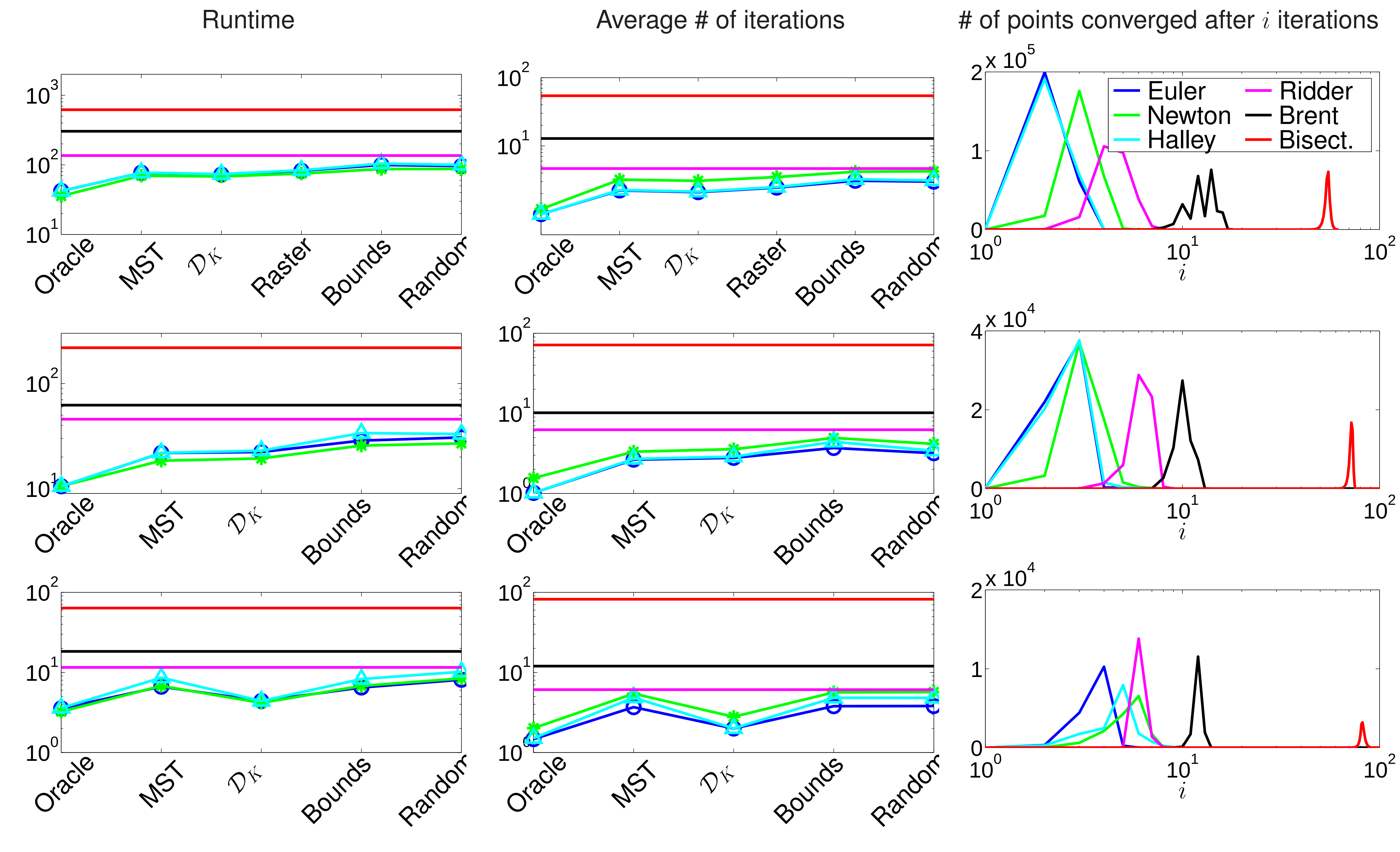


6 Experimental evaluation

1. 512×512 Lena image. Each data point is a pixel represented by spatial and range features $(i,j,L,u,v) \in \mathbb{R}^5$ where (i,j) is the pixel location and (L,u,v) the pixel value (overall $N = 262\,144$ points in $D = 5$ dimensions).
2. $60\,000$ handwritten digits from the MNIST dataset. Each datapoint is a 28×28 grayscale image.
3. $30\,991$ articles from Grolier's encyclopedia. Each point is a word count of the most popular words from $30\,991$ articles.



Bisection: > 10 min. Our method: 1 min.
Computing just the affinities given βs: 20 s.



7 Conclusions

We studied the properties and computation of entropic affinities. Using (1) root-finding methods with high-order convergence, (2) warm-start initialization based on local or density orders, and (3) bounds for the root, we find the root almost to machine precision in just over one iteration per point on average. In applications such as spectral clustering and embeddings, semi-supervised learning, etc. using entropic affinities should give better results than fixing the bandwidth to a single value or using a rule-of-thumb. You just need to give the global perplexity value K.

MATLAB code: `http://eecs.ucmerced.edu`. Run it simply like `[W,s] = ea(X,K)`

ICML | Atlanta
International Conference on Machine Learning

Partially supported by NSF CAREER award IIS–0754089.