Fast nearest neighbor retrieval for bregman divergences

Lawrence Cayton
UC San Diego
Nearest neighbor search

query: \[ q = \] find best match in \[ X = \]

- NN methods ubiquitous, but expensive
- Many NN data structures designed to reduce the complexity, mostly for metrics
- In learning, vision, text, use many non-metric measures; a prominent example is the KL-divergence.

This work: a data structure designed for bregman divergences.
For strictly convex \( f : \mathbb{R}^d \to \mathbb{R} \),

\[
d_f(x, y) \equiv f(x) - f(y) - \langle \nabla f(y), x - y \rangle
\]
Bregman divergence examples

\[ d_f(x, y) = \frac{1}{2}\|x - y\|_2^2 \]

Mahalanobis (\(Q \succ 0\))

\[ d_f(x, y) = \frac{1}{2}(x - y)^\top Q(x - y) \]

KL-divergence

\[ d_f(x, y) = \sum x_i \log \frac{x_i}{y_i} \]

Itakura-Saito

\[ d_f(x, y) = \sum \left( \frac{x_i}{y_i} - \log \frac{x_i}{y_i} - 1 \right) \]
Bregman divergences VS metrics

**Metrics:**

- **non-negativity**
  
  \[ d(x, y) \geq 0 \]

- **symmetry**
  
  \[ d(x, y) = d(y, x) \]

- **triangle inequality**
  
  \[ d(x, y) + d(y, z) \geq d(x, z) \]
Bregman divergences VS metrics

**Metrics:**

- non-negativity: $d(x, y) \geq 0$
- symmetry: $d(x, y) = d(y, x)$
- triangle inequality: $d(x, y) + d(y, z) \geq d(x, z)$

**Bregman divergences:**

- non-negativity: $d_f(x, y) \geq 0$
- symmetry: $d_f(x, y) = d_f(y, x)$
- triangle inequality: $d_f(x, y) + d_f(y, z) \geq d_f(x, z)$
Review: tree-based NN retrieval

* e.g. kd-trees, metric trees, many many variants

Hierarchical space decomposition

Search via branch and bound exploration
Bregman ball trees

- Fundamental geometric unit: bregman ball.
  
  \[ B(\mu, R) \equiv \{ x : d_f(x, \mu) \leq R \} \]

- Need a reasonable build heuristic.

- Can’t use the triangle inequality for bounds.

- Need to handle asymmetry of divergence.
  
  (Not covered here -- see paper)
Intuition: at each level, want balls that are well separated & compact.

Can prune left node

VS

Have to search both
bbtree -- build

Intuition: at each level, want balls that are well separated & compact.

Build method: Deploy \( k \)-means hierarchically (top-down).

\( (k\text{-means was generalized to bregman divergences in } \text{Banerjee et al. 2005}) \)
bbtree -- search

Want to find the left NN:

$$\arg\min_{x \in X} d_f(x, q)$$

Branch & bound search:

1. Descend tree, choosing child whose mean is closest to $q$. Ignore the sibling.

2. At leaf, compute distances to all points; call the nearest the candidate NN $x_c$.

3. Traverse back up tree; check the ignored nodes. If

$$d_f(x_c, q) > \min_{x \in B(\mu, R)} d_f(x, q)$$

need to explore it.
Computing the bound

Need to check if

\[ d_f(x_c, q) > \min_{x \in B(\mu, R)} d_f(x, q) \]
Computing the bound

Need to check if

$$d_f(x_c, q) > \min_{x \in B(\mu, R)} d_f(x, q)$$

The bregman projection onto a bregman ball

Convex, but need to compute it in time comparable to evaluating an analytic expression
The $\ell^2_2$ case

\[
\min_x \frac{1}{2} \|x - q\|^2 \\
\text{subject to: } \frac{1}{2} \|x - \mu\|^2 \leq R
\]

Can compute projection analytically:

\[
x_p = \theta \mu + (1 - \theta)q
\]

where \( \theta = \frac{\sqrt{2R}}{\|q - \mu\|} \)

Easy because

\(x_p\) is on line between \(\mu\) and \(q\)
The general case

\[ \min_x d_f(x, q) \]

subject to: \[ d_f(x, \mu) \leq R \]

Something similar holds..
The general case

\[
\min_x d_f(x, q)
\]
subject to: \(d_f(x, \mu) \leq R\)

Something similar holds..

Claim 1: \(\nabla f(x_p) = \theta \nabla f(\mu) + (1 - \theta) \nabla f(q)\).

The \(l_2^2\) relationship is a special case since \(\nabla f(x) = x\).

Nearly as useful....
Since $f$ strictly convex,

$\nabla f$ is one-to-one.

Moreover, its inverse is given by the gradient of

$$f^*(y) \equiv \sup_x \{ \langle x, y \rangle - f(x) \}.$$ 

Thus can recover $x_p$ from $\nabla f(x_p)$.
Notation:

\[ \mu' \equiv \nabla f(\mu) \]

\[ q' \equiv \nabla f(q) \]

\[ x'_\theta \equiv \theta \mu' + (1 - \theta)q' \]

\[ \nabla f^*(x'_\theta), \quad \theta \in [0, 1] \]

Solution lies on this curve.
Algorithm

Bisection search on \( \theta \) for \( x_\theta \) satisfying \( d_f(x_\theta, \mu) = R \).

1. \( \theta_1 = \frac{1}{2} \)
2. \( \theta_2 = \frac{1}{4} \)
3. \( \theta_3 = \frac{3}{8} \)
4. \( \theta_4 = \frac{5}{16} \)
Algorithm

Bisection search on $\theta$ for $x_\theta$ satisfying $d_f(x_\theta, \mu) = R$.

1. $\theta_1 = \frac{1}{2}$
2. $\theta_2 = \frac{1}{4}$
3. $\theta_3 = \frac{3}{8}$
4. $\theta_4 = \frac{5}{16}$

- Can compute a solution to accuracy $\epsilon$ in $\log \frac{1}{\epsilon}$ steps.
- Each step requires 1 gradient evaluation and 1 divergence evaluation.

Very fast.
But: Don’t actually need an exact solution.

Only need to determine if:

\[ d_f(x_c, q) > \min_{x \in B(\mu, R)} d_f(x, q) \]

\((x_c \text{ is the current candidate NN})\)
**But:** Don’t actually need an **exact** solution.

Only need to determine if:  
\[ d_f(x_c, q) > \min_{x \in B(\mu, R)} d_f(x, q) \]

\((x_c \text{ is the current candidate NN})\)

**i.e.** upper and lower bounds suffice

**Lower bound:** **weak duality**  
\[ \mathcal{L}(\theta) \equiv d_f(x_\theta, q) + \frac{\theta}{1 - \theta} \left( d_f(x_\theta, \mu) - R \right) \]

\[ \leq \min_{x \in B(\mu, R)} d_f(x, q) \]

**Upper bound:** **primal**  
\[ d_f(x_\theta, q) \geq \min_{x \in B(\mu, R)} d_f(x, q) \]

for feasible \(x_\theta\)

Evaluate bounds at each step of bisection to stop early.
Experiments: KL-divergence

Why KL divergence?

- Used extensively to compare histograms (e.g. text, vision).
- No (correct) NN schemes out there for it.
- Mahalanobis, $\ell_2^2$ can be handled by metric methods.
Data sets

- **rcv\text{-}$D$:** 500k documents from the RCV corpus represented as a $D$-dimensional distribution over topic (generated using LDA).

- **Corel histograms:** 60k color histograms, 64-dimensional.

- **Semantic space:** 371-dimensional representation of 5000 images (from CBIR literature)

- **SIFT signatures:** 1111-dimensional quantized histogram representation of 10k images from PASCAL 2007 dataset
Approx search experiments

Stop search early (after examining only a few leaves)
-- standard practice with metric, kd-trees, etc.

Evaluation

Speedup over brute-force search in execution time.

VS

NC for *number closer*: how many closer points are there? *e.g.* if NC=3, the bbtree returned the fourth NN.
Approximate search

cv-128

Speedup (exponent)

num closer (exponent)
corel, semantic space, SIFT
## Exact search

<table>
<thead>
<tr>
<th>dataset</th>
<th>dimensionality</th>
<th>speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>rcv-8</td>
<td>8</td>
<td>64.5</td>
</tr>
<tr>
<td>rcv-16</td>
<td>16</td>
<td>36.7</td>
</tr>
<tr>
<td>rcv-32</td>
<td>32</td>
<td>21.9</td>
</tr>
<tr>
<td>rcv-64</td>
<td>64</td>
<td>12.0</td>
</tr>
<tr>
<td>corel histograms</td>
<td>64</td>
<td>2.4</td>
</tr>
<tr>
<td>rcv-128</td>
<td>128</td>
<td>5.3</td>
</tr>
<tr>
<td>rcv-256</td>
<td>256</td>
<td>3.3</td>
</tr>
<tr>
<td>semantic space</td>
<td>371</td>
<td>1.0</td>
</tr>
<tr>
<td>SIFT signatures</td>
<td>1111</td>
<td>0.9</td>
</tr>
</tbody>
</table>
Thanks..

- Serge Belongie
- Sanjoy Dasgupta
- Charles Elkan
- Carolina Galleguillos
- Daniel Hsu
- Nikhil Rasiwasia
- Lawrence Saul