A Bregman near neighbor lower bound via directed isoperimetry

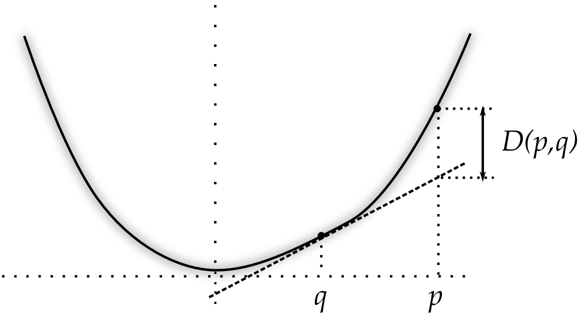
Amirali Abdullah
Suresh Venkatasubramanian

University of Utah

Bregman Divergences

For convex $\phi : \mathbb{R}^d \to \mathbb{R}$

$$D_\phi(\mathbf{p}, \mathbf{q}) = \phi(\mathbf{p}) - \phi(\mathbf{q}) - \langle \nabla\phi(\mathbf{q}), \mathbf{p} - \mathbf{q} \rangle$$



Examples

$\phi(x) = \|x\|^2$ (Squared Euclidean):

$$D_\phi(p,q) = \|p\|^2 - \|q\|^2 - 2\langle q, p-q\rangle = \|p-q\|^2$$

$\phi(x) = \sum_i x_i \ln x_i$ (Kullback-Leibler):

$$D_\phi(p,q) = \sum_i p_i \ln \frac{p_i}{q_i} - p_i + q_i$$

$\phi(x) = -\ln x$ (Itakura-Saito):

$$D_\phi(p,q) = \sum_i \frac{p_i}{q_i} - \ln \frac{p_i}{q_i} - 1$$

Where do they come from ?

Exponential family:

$$p_{(\psi,\theta)}(x) = \exp(\langle x, \theta \rangle - \psi(\theta))p_0(x)$$

can be written [BMDG06] as

$$p_{(\psi,\theta)}(x) = \exp(-D_\phi(x, \mu))b_\phi(x)$$

Distribution	Distance
Gaussian	Squared Euclidean
Multinomial	Kullback-Leibler
Exponential	Itakura-Saito

Bregman divergences generalize methods like AdaBoost, MAP estimation, clustering, and mixture model estimation.

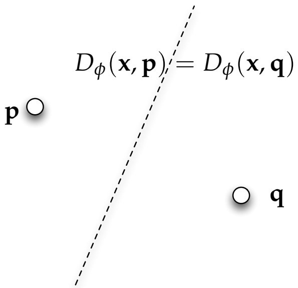
Exact Geometry of Bregman Divergences

We can generalize projective duality to Bregman divergences:

$$\phi^*(\mathbf{u}) = \max_{\mathbf{p}} \langle \mathbf{p}, \mathbf{u} \rangle - \phi(\mathbf{p})$$

$$\mathbf{p}^* = \arg\max_{\mathbf{p}} \langle \mathbf{p}, \mathbf{u} \rangle - \phi(\mathbf{p}) = \nabla\phi(p)$$

Bregman hyperplanes are linear (or dually linear) [BNN07]:



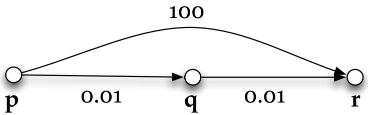
Exact Geometry of Bregman Divergences

Exact algorithms based on duality and arrangements carry over:

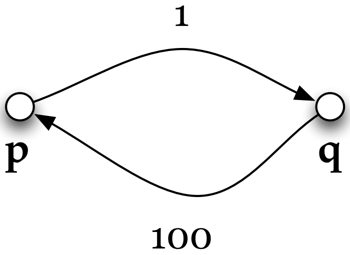
$$\begin{tikzcd}[column sep=huge, row sep=huge]
\text{Arrangement of hyperplanes} \arrow[r, leftrightarrow, "\mathbf{p} \mapsto \mathbf{p}^*"] \arrow[d] & \text{Convex hull} \\
\text{Voronoi diagram} \arrow[r, leftrightarrow] & \text{Delaunay triangulation} \arrow[u, "\mathbf{p} \mapsto (\mathbf{p}, \phi(\mathbf{p}))"']
\end{tikzcd}$$

We can solve exact nearest neighbor problem (modulo algebraic operations)

But this doesn't work for *approximate* algorithms:
No triangle inequality:



No symmetry



Where does the asymmetry come from?

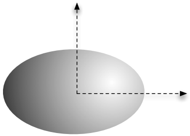
Reformulating the Bregman divergence:

$$\begin{aligned}
D_\phi(\mathbf{p},\mathbf{q}) &= \phi(\mathbf{p}) - \phi(\mathbf{q}) - \langle \nabla\phi(\mathbf{q}), \mathbf{p}-\mathbf{q}\rangle \\
&= \phi(\mathbf{p}) - \Big(\phi(\mathbf{q}) + \langle \nabla\phi(\mathbf{q}), \mathbf{p}-\mathbf{q}\rangle\Big) \\
&= \phi(\mathbf{p}) - \tilde{\phi}_{\mathbf{q}}(\mathbf{p}) \\
&= (\mathbf{p}-\mathbf{q})^\top \nabla^2\phi(\mathbf{r})(\mathbf{p}-\mathbf{q}), \mathbf{r} \in [\mathbf{p},\mathbf{q}]
\end{aligned}$$

As $\mathbf{p} \to \mathbf{q}$,

$$D_\phi(\mathbf{p},\mathbf{q}) \simeq (\mathbf{p}-\mathbf{q})^\top A(\mathbf{p}-\mathbf{q})$$

is called a Mahalanobis distance.

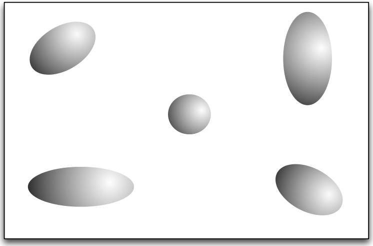


Where does the asymmetry come from?

If A is fixed and positive definite, then $A = U^\top U$:

$$\begin{aligned}(\mathbf{p}-\mathbf{q})^\top A(\mathbf{p}-\mathbf{q}) &= (\mathbf{p}-\mathbf{q})^\top U^\top U(\mathbf{p}-\mathbf{q}) \\ &= \|\mathbf{p}'-\mathbf{q}'\|^2\end{aligned}$$

where $\mathbf{p}' = U\mathbf{p}$.
So the problem arises when the Hessian varies across the domain of interest:



Quantifying the asymmetry

Let Δ be a domain of interest.

***μ*-asymmetry:**

$$\mu = \max_{\mathbf{p},\mathbf{q} \in \Delta} \frac{D_\phi(\mathbf{p},\mathbf{q})}{D_\phi(\mathbf{q},\mathbf{p})}$$

***μ*-similarity:**

$$\mu = \max_{\mathbf{p},\mathbf{q},\mathbf{r} \in \Delta} \frac{D_\phi(\mathbf{p},\mathbf{r})}{D_\phi(\mathbf{p},\mathbf{q}) + D_\phi(\mathbf{q},\mathbf{r})}$$

***μ*-defectiveness:**

$$\mu = \max_{\mathbf{p},\mathbf{q},\mathbf{r} \in \Delta} \frac{D_\phi(\mathbf{p},\mathbf{q}) - D_\phi(\mathbf{r},\mathbf{q})}{D_\phi(\mathbf{p},\mathbf{r})}$$

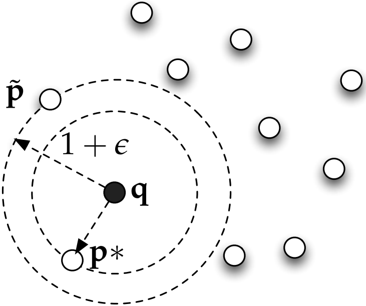
- If $\max_x \lambda_{\max}/\lambda_{\min}$ is bounded, then all of above are bounded.
- If μ-asymmetry is unbounded, then all are.

There are different flavors of results for approximate algorithms for Bregman divergences

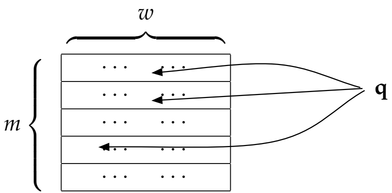
- Assume that μ is bounded and get $f(\mu, \epsilon)$-approximations for clustering: [Manthey-Röglin, Ackermann-Blömer, Feldman-Schmidt-Sohler]
- Assume that μ is bounded and get $(1+\epsilon)$-approximation in time dependent on μ for approximate near neigbor: [Abdullah-V]
- Assume nothing about μ and get unconditional (but weaker) bounds for clustering: [McGregor-Chaudhuri]
- Use heuristics inspired by Euclidean algorithms without guarantees [Nielsen-Nock for MEB, [Cayton,Zhang et al for approximate NN]

Is μ intrinsic to the (approximate) study of Bregman divergences

Process a data set on n points in $\mathbb{R}^d$ to answer $(1+\epsilon)$-approximate near neighbor queries in $\log n$ time using space near-linear in n, with *polynomial dependence* on $d, 1/\epsilon$.



We work within the cell probe model:



- Data structure takes space mw and processes queries using r probes. Call it a (m, w, r)-structure.
- We will work in the *non-adaptive* setting: probes are a function of q

Our Result

Theorem

Any (m, w, r)-nonadaptive data structure for c-approximate near-neighbor search for n points in $\mathbb{R}^d$ under a uniform Bregman divergence with μ-asymmetry (where $\mu \leq d/\log n$) must have

$$mw = \Omega(dn^{1+\Omega(\mu/cr)})$$

Comparing this to a result for ℓ_1 [Panigrahy/Talwar/Wieder]:

Theorem

Any (m, w, r)-nonadaptive data structure for c-approximate near-neighbor search for n points in $\mathbb{R}^d$ under ℓ_1 must have

$$mw = \Omega(dn^{1+\Omega(1/cr)})$$

Our Result

Theorem

Any (m, w, r)-nonadaptive data structure for c-approximate near-neighbor search for n points in $\mathbb{R}^d$ under a uniform Bregman divergence with μ-asymmetry (where $\mu \le d/\log n$) must have

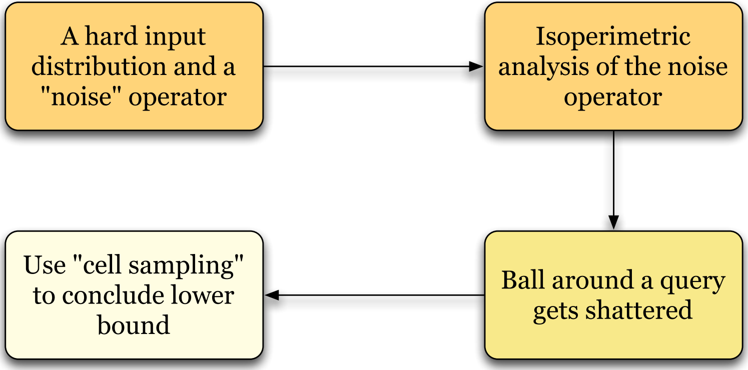
$$mw = \Omega(dn^{1+\Omega(\mu/cr)})$$

- It applies to *uniform* Bregman divergences:

$$D_\phi(\mathbf{p}, \mathbf{q}) = \sum_i D_\phi(p_i, q_i)$$

- Works generally for any divergence that has a lower bound on asymmetry: only need two points in $\mathbb{R}$ to generate the instance.
- $\mu = d/\log n$ is "best possible" in a sense: requiring linear space with $\mu = d/\log n$ implies that $t = \Omega(d/\log n)$ [Barkol-Rabani]

Overview of proof



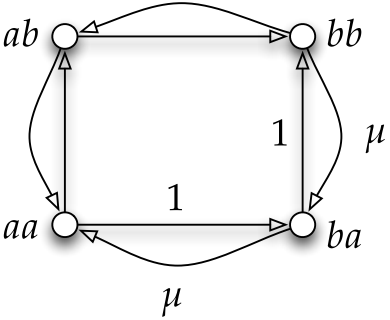
Follows the framework of [Panigrahy-Talwar-Wieder], except when we don't.

Related Work

- Deterministic lower bounds [CCGL,L, PT]
- Exact lower bounds [BOR, BR]
- Randomized lower bounds (poly space) [CR, AIP]
- Randomized lower bounds (near-linear space) [PTW]
- Lower bounds for LSH [MNP, OWZ, AIP]

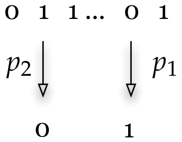
Fix points a, b such that

$$D_\phi(a,b) = 1, D_\phi(b,a) = \mu$$



A directed noise operator

We perturb a vector asymmetrically:



$$\mathbf{x} \overset{v_{p_1,p_2}}{\longmapsto} \mathbf{y}$$

The *directed noise operator*

$$R_{p_1,p_2}(f) = E_{y \sim v_{p_1,p_2}(x)}[f(y)]$$

If we set $p_1 = p_2 = \rho$, we get the *symmetric* noise operator T_ρ.

Lemma

If $p_1 > p_2$, then $R_{p_1,p_2} = T_{p_2} R_{\frac{p_1 - p_2}{1 - 2p_2},0}$

Constructing the instance

1. Take a random set S of n points.
2. Let $P = \{p_i = v_{\epsilon,\epsilon/\mu}(s_i)\}$
3. Let $Q = \{q_i = v_{\epsilon/\mu,\epsilon}(s_i)\}$
4. Pick $q \in_R Q$

Properties: Let $q = q_i$:

1. For all $j \neq i, D(q, p_j) = \Omega(\mu d)$
2. $D(q, p_i) = \Theta(\epsilon d)$
3. If $\mu \leq \epsilon d / \log n$, these hold w.h.p

Noise and the Bonami-Beckner inequality

Fix the uniform measure over the hypercube: $\|f\|_2 = \sqrt{E[f^2(x)]}$
The symmetric noise operator "expands":

$$\|\tau_\rho(f)\|_2 \leq \|f\|_{1+\rho^2}$$

even if the underlying space has a biased measure ($\Pr[x_i = 1] = p \neq 0.5$)

$$\|\tau_\rho(f)\|_{2,p} \leq \|f\|_{1+g(\rho,p),p}$$

We would like to show that the asymmetric noise operator "expands" in the same way:

$$\|R_{p_1,p_2}(f)\|_2 \leq \|f\|_{1+g(p_1,p_2)}$$

Noise and the Bonami-Beckner inequality

Fix the uniform measure over the hypercube: $\|f\|_2 = \sqrt{E[f^2(x)]}$
The symmetric noise operator "expands":

$$\|\tau_\rho(f)\|_2 \leq \|f\|_{1+\rho^2}$$

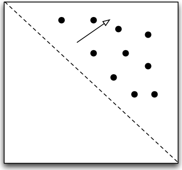
even if the underlying space has a biased measure ($\Pr[x_i = 1] = p \neq 0.5$)

$$\|\tau_\rho(f)\|_{2,p} \leq \|f\|_{1+g(\rho,p),p}$$

We would like to show that the asymmetric noise operator "expands" in the same way:

$$\|R_{p_1,p_2}(f)\|_2 \leq \|f\|_{1+g(p_1,p_2)}$$

It's not actually true !



We will assume that f has support over the lower half of the hypercube.

Proof Sketch

Analyze asymmetric operator over uniform measure by analyzing symmetric operator over biased measure.

$$\|R_{p,0}f\|_2$$

Proof Sketch

Analyze asymmetric operator over uniform measure by analyzing symmetric operator over biased measure.

$$\|R_{p,0}f\|_2 \xrightarrow{\text{[Ahlberg et al]}} \|\tau_{\sqrt{\frac{1-p}{1+p}}}f\|_{2,\frac{1+p}{2}}$$

Proof Sketch

Analyze asymmetric operator over uniform measure by analyzing symmetric operator over biased measure.

$$\begin{tikzcd} \|R_{p,0}f\|_2 \arrow[r, "\text{[Ahlberg et al]}"] & \|\tau_{\sqrt{\frac{1-p}{1+p}}}f\|_{2,\frac{1+p}{2}} \arrow[d, "\text{Biased Bonami-Beckner}"] \\ & \|f\|_{1+\frac{1}{1-\log(1-p)},\frac{1+p}{2}} \end{tikzcd}$$

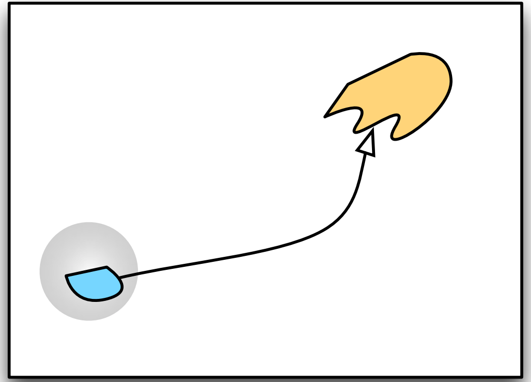
Proof Sketch

Analyze asymmetric operator over uniform measure by analyzing symmetric operator over biased measure.

$$
\begin{tikzcd}
\|R_{p,0}f\|_2 \arrow[r, "\text{[Ahlberg et al]}"] & \|\tau_{\sqrt{\frac{1-p}{1+p}}}f\|_{2,\frac{1+p}{2}} \arrow[d, "\text{Biased Bonami-Beckner}"] \\
\|f\|_{1+\frac{1}{1-\log(1-p)}} & \|f\|_{1+\frac{1}{1-\log(1-p)},\frac{1+p}{2}} \arrow[l, "\text{Restriction to lower half-cube}"']
\end{tikzcd}
$$

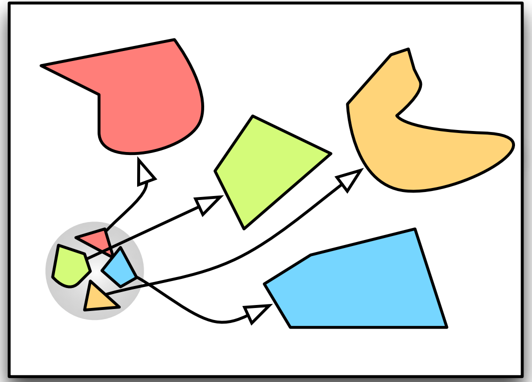
From hypercontractivity to shattering I

For any small fixed region of the hypercube, only a small portion of the ball around a point is sent there by the noise operator.



Proof is based on hypercontractivity and Cauchy-Schwarz.

If we partition the hypercube into small enough regions (each corresponding to a hash table entry) then a ball gets shattered among many pieces.



The cell sampling technique

Suppose you have a data structure with space S that can answer NN queries with t probes.

- Fix a (random) input point that you want to reconstruct.

The cell sampling technique

Suppose you have a data structure with space S that can answer NN queries with t probes.

- Fix a (random) input point that you want to reconstruct.
- Sample a fraction of the cells of the structure

The cell sampling technique

Suppose you have a data structure with space S that can answer NN queries with t probes.

- Fix a (random) input point that you want to reconstruct.
- Sample a fraction of the cells of the structure
- Determine which queries still "work" (only access cells from the sample)

The cell sampling technique

Suppose you have a data structure with space S that can answer NN queries with t probes.

- Fix a (random) input point that you want to reconstruct.
- Sample a fraction of the cells of the structure
- Determine which queries still "work" (only access cells from the sample)
- Suppose one of these works: then we've reconstructed the input point using a small sample (with some probability)

The cell sampling technique

Suppose you have a data structure with space S that can answer NN queries with t probes.

- Fix a (random) input point that you want to reconstruct.
- Sample a fraction of the cells of the structure
- Determine which queries still "work" (only access cells from the sample)
- Suppose one of these works: then we've reconstructed the input point using a small sample (with some probability)
- By Fano's inequality, the size of this sample must be reasonably large.

The cell sampling technique

Suppose you have a data structure with space S that can answer NN queries with t probes.

- Fix a (random) input point that you want to reconstruct.
- Sample a fraction of the cells of the structure
- Determine which queries still "work" (only access cells from the sample)
- Suppose one of these works: then we've reconstructed the input point using a small sample (with some probability)
- By Fano's inequality, the size of this sample must be reasonably large.
- Therefore, the data structure is large

The cell sampling technique

Suppose you have a data structure with space S that can answer NN queries with t probes.

- Fix a (random) input point that you want to reconstruct.
- Sample a fraction of the cells of the structure
- Determine which queries still "work" (only access cells from the sample)
- Suppose one of these works: then we've reconstructed the input point using a small sample (with some probability)
- By Fano's inequality, the size of this sample must be reasonably large.
- Therefore, the data structure is large

The hypercontractivity-based shattering property implies that many of the "working" queries are sent to different cells, so there's a high chance that one of them will succeed.

Conclusions

- The measure of asymmetry μ appears to play an important role in the design of algorithms for Bregman divergences.
- Can these measures quantify asymmetry ? In particular, what about Bregman k-center clustering ?
- Are there any other applications for an "on average" asymmetric hypercontractivity result ?