Mixed Bregman Clustering with Approximation Guarantees

Richard Nock[1] Panu Luosto[2] Jyrki Kivinen[2]

[1] Université Antilles-Guyane, Schoelcher, France

[2] University of Helsinki, Finland

September 15, 2008

Previous work and our contribution

- we introduce mixed Bregman clustering which generalizes Bregman hard clustering (Banerjee, Merugo, Dillon and Ghosh 2005 [BGW05])

Previous work and our contribution

- we introduce mixed Bregman clustering which generalizes Bregman hard clustering (Banerjee, Merugo, Dillon and Ghosh 2005 [BGW05])
- we use a seeding method similar to D^2 seeding (Arthur and Vassilvitskii [AV07])
- Ostrovsky, Rabani, Schulman and Swamy have introduced a seeding method almost like D^2 seeding in [ORSS06]

Previous work and our contribution

- we introduce mixed Bregman clustering which generalizes Bregman hard clustering (Banerjee, Merugo, Dillon and Ghosh 2005 [BGW05])
- we use a seeding method similar to D^2 seeding (Arthur and Vassilvitskii [AV07])
- Ostrovsky, Rabani, Schulman and Swamy have introduced a seeding method almost like D^2 seeding in [ORSS06]
- ...and get some approximation guarantees like in [AV07]

Contents

Bregman divergences

- let $\psi : X \to \mathbb{R}$ be a strictly convex and differentiable function defined on a convex set $X \subset \mathbb{R}^d$
- the corresponding Bregman divergence is then $D_\psi(x\|y) \doteq \psi(x) - \psi(y) - \langle x - y, \nabla\psi(y)\rangle$
- examples
  - Mahalanobis distance (M symmetric positive definite)

    $$\psi_M(x) = x^T M x, \qquad D_M(x\|y) = (x-y)^T M (x-y)$$

  - Kullback-Leibler divergence

    $$\psi_{KL}(x) = \sum_{i=1}^d x_i \log x_i, \qquad D_{KL}(x\|y) = \sum_{i=1}^d x_i \log(x_i/y_i)$$

  - Itakura-Saito distance

    $$\psi_{IS}(x) = -\log x, \quad D_{IS}(x\|y) = \sum_{i=1}^d (x_i/y_i - \log(x_i/y_i) - 1)$$

Properties of Bregman divergences

- $D_\psi(x\|y) \geq 0$ for all x, y
- $D_\psi(x\|y) = 0$ if and only if $x = y$
- usually $D_\psi(x\|y) \neq D_\psi(y\|x)$
- Mahalanobis distance is the only symmetrical Bregman divergence

Exponential families and Bregman divergences

- Bregman divergences are important because of the exponential families
- probability density of a regular exponential family:

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

- Kullback-Leibler divergence between distributions is equal to the Bregman divergence between their natural parameters:

$$D_{KL}(p_{(\psi,\theta)} \| p_{(\psi,\theta')}) = D_\psi(\theta \| \theta') = D_{\psi^*}(\mu' \| \mu)$$

Exponential families and Bregman divergences

- Bregman divergences are important because of the exponential families
- probability density of a regular exponential family:

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

- Kullback-Leibler divergence between distributions is equal to the Bregman divergence between their natural parameters:

$$D_{KL}(p_{(\psi,\theta)} \| p_{(\psi,\theta')}) = D_\psi(\theta \| \theta') = D_{\psi^*}(\mu' \| \mu)$$

- k-means type methods work with Bregman divergences

k-means problem

- Find a k-clustering which minimizes the potential

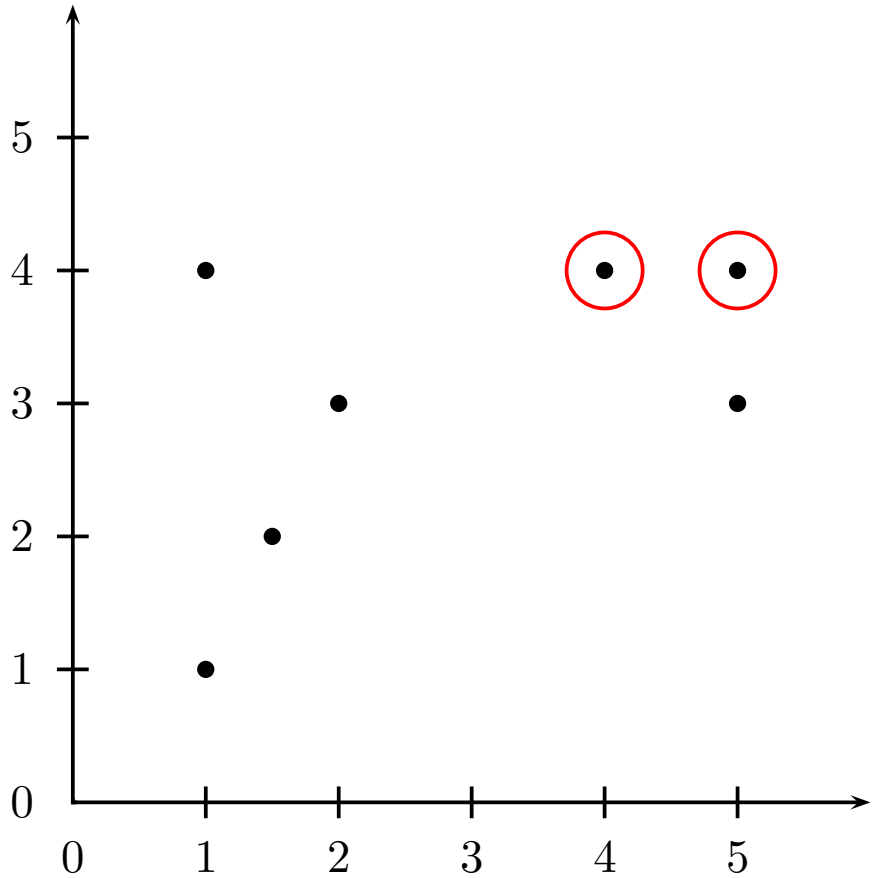
$$\sum_{x \in S} \min_{c \in C} \|x - c\|^2 = \sum_{A \in \mathcal{P}} \sum_{x \in A} \|x - c_A\|^2,$$

$$\text{where } c_A = \frac{1}{|A|} \sum_{x \in A} x$$

- the problem is NP-hard
- approximation algorithms exist but usually k-means heuristic is used

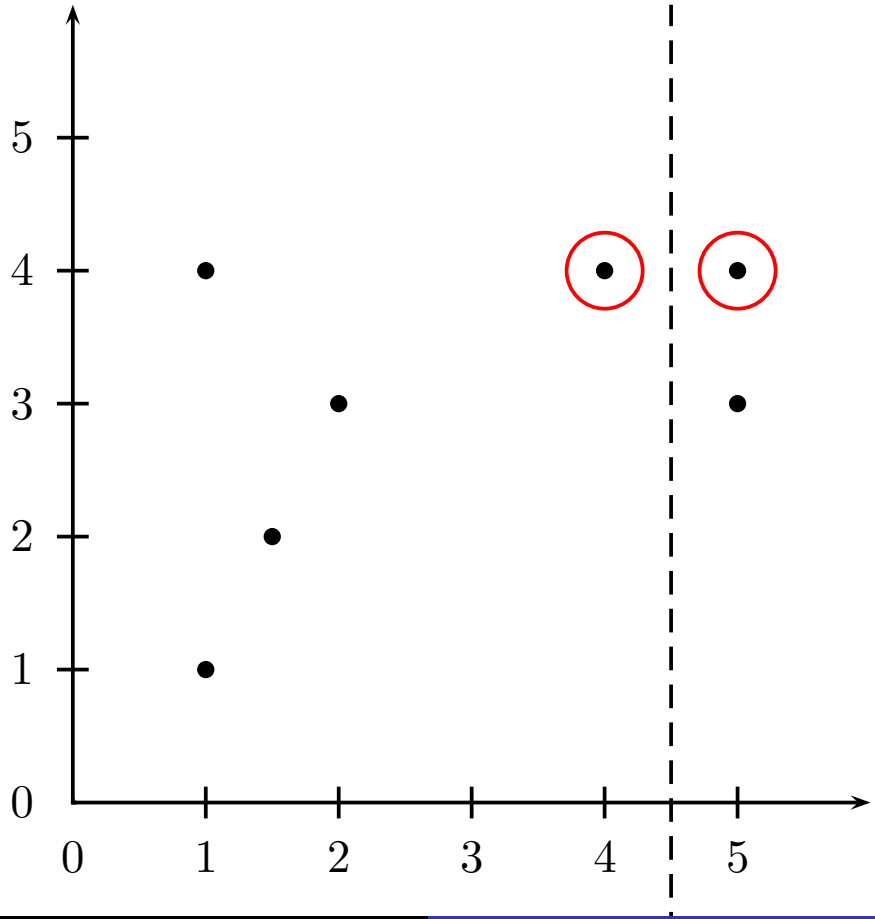
k-means method

Pick the initial seeds



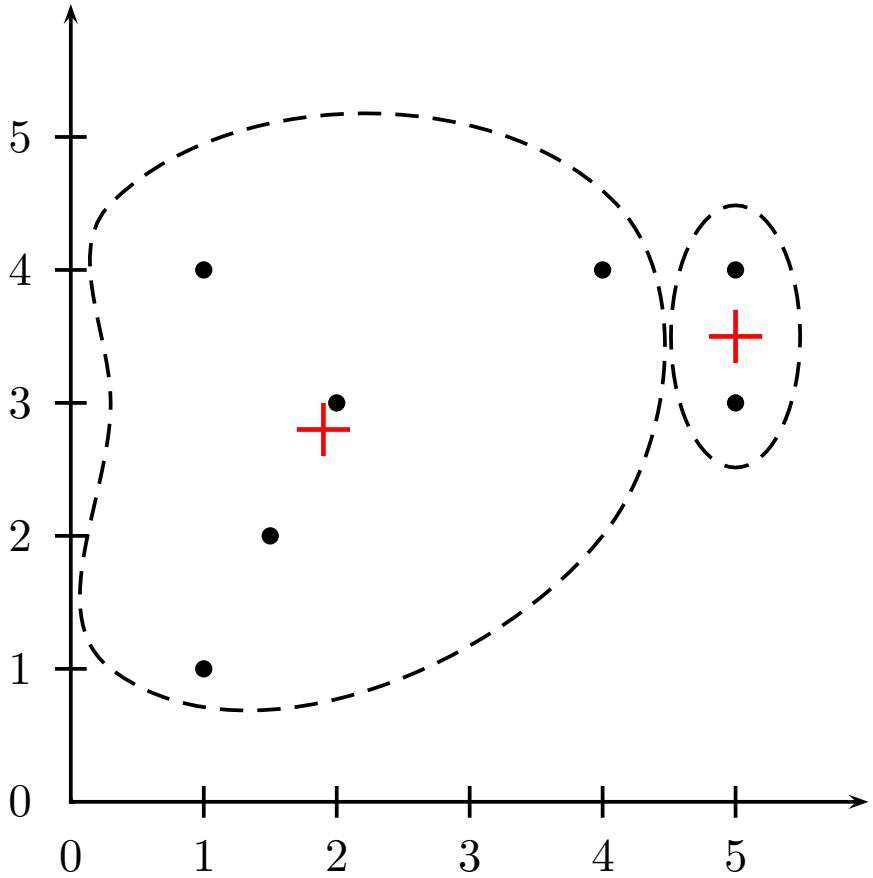
k-means method

Assign points to nearest centres



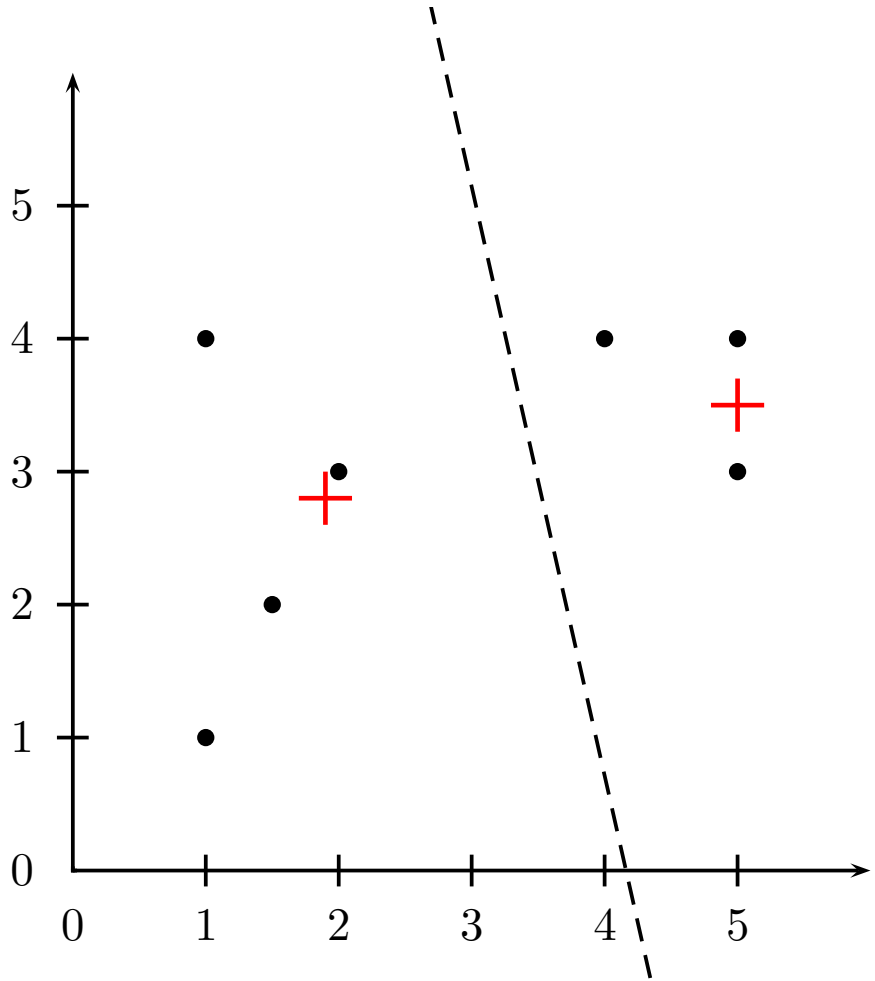
k-means method

Optimize the centres



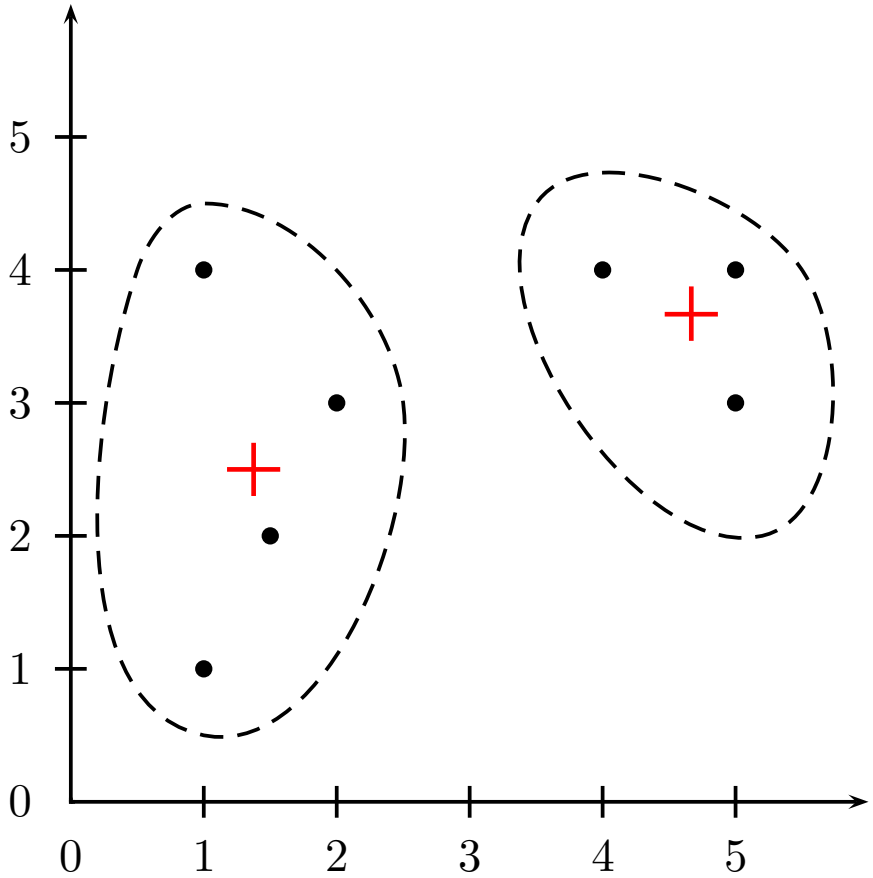
k-means method

Assign points to nearest centres (2. round)



k-means method

Optimize the centres (2. round) . . .



D^2 seeding [AV07]

- seeding is the crucial phase in k-means clustering

D^2 seeding [AV07]

- seeding is the crucial phase in k-means clustering
- How to get at least some quality guarantees?
  – Use D^2 seeding.

D^2 seeding [AV07]

- seeding is the crucial phase in k-means clustering
- How to get at least some quality guarantees?
  – Use D^2 seeding.
- pick the first centre uniformly at random

D^2 seeding [AV07]

- seeding is the crucial phase in k-means clustering
- How to get at least some quality guarantees?
  – Use D^2 seeding.
- pick the first centre uniformly at random
- then favour points that are far from the points picked so far

D^2 seeding [AV07]

- seeding is the crucial phase in k-means clustering
- How to get at least some quality guarantees?
  – Use D^2 seeding.
- pick the first centre uniformly at random
- then favour points that are far from the points picked so far
- pick point x as the next centre with a probability

$$\pi(x) = \frac{D^2(x)}{\sum_{x \in S} D^2(x)},$$

where $D^2(x)$ is the squared Euclidean distance from x to the set of centres picked already

D^2 seeding [AV07]

- in other words: pick point x with a probability which is equal to the relative contribution of x to the clustering cost around the centres which have been chosen before

$$\pi(x) = \frac{D^2(x)}{\sum_{y \in S} D^2(y)}$$

D^2 seeding [AV07]

- in other words: pick point x with a probability which is equal to the relative contribution of x to the clustering cost around the centres which have been chosen before

$$\pi(x) = \frac{D^2(x)}{\sum_{y \in S} D^2(y)}$$

- we get approximation guarantees

$$\mathbb{E}[Cost] \leq 8(2 + \ln k)\ Optcost$$

D^2 seeding [AV07]

- in other words: pick point x with a probability which is equal to the relative contribution of x to the clustering cost around the centres which have been chosen before

$$\pi(x) = \frac{D^2(x)}{\sum_{y \in S} D^2(y)}$$

- we get approximation guarantees

$$\mathbb{E}[Cost] \leq 8(2 + \ln k)\ Optcost$$

- Does the idea work with other Bregman divergences?

Contents

Mixed Bregman divergence

- let ψ be a Bregman function and α a number in range $[0, 1]$
- the corresponding Bregman α-divergence between the points of the triplet (c^*, x, c) is

$$\triangle_{\psi,\alpha}(c^* \| x \| c) \doteq (1 - \alpha) D_\psi(c^* \| x) + \alpha D_\psi(x \| c)$$

Mixed Bregman divergence

- let ψ be a Bregman function and α a number in range $[0, 1]$
- the corresponding Bregman α-divergence between the points of the triplet (c^*, x, c) is

$$\triangle_{\psi,\alpha}(c^* \| x \| c) \doteq (1 - \alpha) D_\psi(c^* \| x) + \alpha D_\psi(x \| c)$$

- this is one way to handle the asymmetry of Bregman divergences

Mixed Bregman divergence

- let ψ be a Bregman function and α a number in range $[0, 1]$
- the corresponding Bregman α-divergence between the points of the triplet (c^*, x, c) is

$$\triangle_{\psi,\alpha}(c^* \| x \| c) \doteq (1 - \alpha) D_\psi(c^* \| x) + \alpha D_\psi(x \| c)$$

- this is one way to handle the asymmetry of Bregman divergences
- now we need two centres per cluster

Optimal pair of centres

- What is the optimal centre pair (c_A^*, c_A) for a given cluster A?
- pair $(a, b) \in X \times X$ yields the potential

$$Cost_{\psi,\alpha}(\{(a,b)\}, A) = \sum_{x \in A} \triangle_{\psi,\alpha}(a \| x \| b)$$
$$= (1-\alpha) \sum_{x \in A} D_\psi(a \| x) \; + \; \alpha \sum_{x \in A} D_\psi(x \| b)$$

Optimal pair of centres

- finding the best left and right centres turns out to be easy:

$$c_A = \frac{1}{|A|} \sum_{x \in A} x$$

$$c_A^* = (\nabla \psi)^{-1} \left(\frac{1}{|A|} \sum_{x \in A} \nabla \psi(x) \right)$$

Optimal pair of centres

- finding the best left and right centres turns out to be easy:

$$c_A = \frac{1}{|A|} \sum_{x \in A} x$$

$$c_A^* = (\nabla\psi)^{-1} \left(\frac{1}{|A|} \sum_{x \in A} \nabla\psi(x) \right)$$

- ... because for every $c \in X$

$$\sum_{x \in A} D_\psi(x \| c) - \sum_{x \in A} D_\psi(x \| c_A) = |A| \; D_\psi(c_A \| c)$$

$$\sum_{x \in A} D_\psi(c \| x) - \sum_{x \in A} D_\psi(c_A^* \| x) = |A| \; D_\psi(c \| c_A^*)$$

Potential of Bregman clustering

- suppose we are given set $\mathcal{C}$ of centroid pairs

Potential of Bregman clustering

- suppose we are given set C of centroid pairs
- corresponding potential of the mixed Bregman clustering is

$$\begin{aligned} Cost_{\psi,\alpha}(C,S) &\doteq \sum_{x\in S} \min_{(a,b)\in C} \triangle_{\psi,\alpha}(a\|x\|b) \\ &= \sum_{x\in S} \min_{(a,b)\in C} \Big((1-\alpha)D_\psi(a\|x) + \alpha D_\psi(x\|b)\Big) \end{aligned}$$

Potential of Bregman clustering

- suppose we are given set C of centroid pairs
- corresponding potential of the mixed Bregman clustering is

$$\begin{aligned} Cost_{\psi,\alpha}(C,S) &\doteq \sum_{x\in S} \min_{(a,b)\in C} \triangle_{\psi,\alpha}(a\|x\|b) \\ &= \sum_{x\in S} \min_{(a,b)\in C} \Big((1-\alpha)D_\psi(a\|x) + \alpha D_\psi(x\|b)\Big) \end{aligned}$$

- potential of the optimal k-clustering

$$Optcost_{\psi,\alpha}(S) \doteq \min_{C\subset X^2, |C|=k} Cost_{\psi,\alpha}(C,S)$$

Dual potential

- potential of the optimal clustering was

$$Optcost_{\psi,\alpha}(S) = \sum_{x \in S} \min_{(a,b) \in \mathbf{C}_{\text{opt}}} \Big((1-\alpha) D_\psi(a \| x) + \alpha D_\psi(x \| b) \Big)$$

Dual potential

- potential of the optimal clustering was

$$Optcost_{\psi,\alpha}(S) = \sum_{x \in S} \min_{(a,b) \in C_{\text{opt}}} \Big((1-\alpha) D_\psi(a \| x) + \alpha D_\psi(x \| b) \Big)$$

- dual potential of the optimal clustering C_{opt} is

$$\begin{aligned} Optcost^*_{\psi,\alpha}(S) &= \sum_{x \in S} \min_{(a,b) \in C_{\text{opt}}} \triangle_{\psi,1-\alpha}(b \| x \| a) \\ &= \sum_{x \in S} \min_{(a,b) \in C_{\text{opt}}} \Big((1-\alpha) D_\psi(x \| a) + \alpha D_\psi(b \| x) \Big) \end{aligned}$$

Mixed Bregman seeding

Let $C \leftarrow \{(x, x)\}$,

where x is chosen uniformly at random from S;

Repeat $k - 1$ times:

Pick point $x \in S$ with probability

$$\frac{\triangle_{\psi,\alpha}(c_x \| x \| c_x)}{\sum_{y \in S} \triangle_{\psi,\alpha}(c_y \| y \| c_y)},$$

where $(c_x, c_x) = \arg\min_{(z,z) \in C} \triangle_{\psi,\alpha}(z \| x \| z)$;

$C \leftarrow C \cup \{(x, x)\}$;

Mixed Bregman clustering

Choose the set of initial centroids $C = \{(a_i, b_i) \mid i \in \{1, \ldots, k\}\}$ using mixed Bregman seeding;

Repeat until convergence:

for i **in** $1, 2, \ldots, k$ **do**

$$A_i \leftarrow \{x \in S \mid i = \arg\min_{j \in \{1,\ldots,k\}} \triangle_{\psi,\alpha}(a_j \| x \| b_j)\}$$

$$a_i \leftarrow (\nabla\psi)^{-1}\Big(\frac{1}{|A_i|}\sum_{x \in A_i} \nabla\psi(x)\Big)$$

$$b_i \leftarrow \frac{1}{|A_i|}\sum_{x \in A_i} x$$

Approximation guarantees

- [AV07] proves when squared Euclidean distance is used: after D^2 seeding

$$\mathbb{E}[Cost] \leq 8(2 + \ln k)\ Optcost$$

Approximation guarantees

- [AV07] proves when squared Euclidean distance is used: after D^2 seeding

$$\mathbb{E}[Cost] \leq 8(2 + \ln k)\ Optcost$$

- it can be shown that after mixed Bregman α-seeding

$$\mathbb{E}[Cost_{\psi,\alpha}] \leq 4\rho_\psi^2(2 + \ln k)(Optcost_{\psi,\alpha} + Optcost^*_{\psi,\alpha})$$

- the factor ρ_ψ will be defined soon...

Approximation guarantees

- [AV07] proves when squared Euclidean distance is used: after D^2 seeding

$$\mathbb{E}[Cost] \leq 8(2 + \ln k)\ Optcost$$

- it can be shown that after mixed Bregman α-seeding

$$\mathbb{E}[Cost_{\psi,\alpha}] \leq 4\rho_\psi^2(2 + \ln k)(Optcost_{\psi,\alpha} + Optcost^*_{\psi,\alpha})$$

- the factor ρ_ψ will be defined soon...
- in the case of Mahalanobis distances the bound becomes

$$\mathbb{E}[Cost_{\psi,\alpha}] \leq 8(2 + \ln k)\ Optcost_{\psi,\alpha}$$

Factor ρ_ψ

- factor ρ_ψ is needed for the estimate:
  for any points x, y, z of the convex closure of S

$$D_\psi(x,z) \leq 2\rho_\psi^2 \left(D_\psi(x,y) + D_\psi(y,z) \right) ,$$

  where

$$\rho_\psi^2 \doteq \sup_{s,t,u,v \in \mathrm{co}(S)} \frac{(u-v)^T \mathrm{H}_s (u-v)}{(u-v)^T \mathrm{H}_t (u-v)}$$

  and H_s is the Hessian matrix of ψ at s
- ρ_ψ depends both on ψ and the set to be clustered

Factor ρ_ψ

- factor ρ_ψ is needed for the estimate:
  for any points x, y, z of the convex closure of S

$$D_\psi(x, z) \leq 2\rho_\psi^2 \big(D_\psi(x, y) + D_\psi(y, z)\big) \ ,$$

  where

$$\rho_\psi^2 \doteq \sup_{s,t,u,v \in \mathrm{co}(S)} \frac{(u-v)^T \mathrm{H}_s (u-v)}{(u-v)^T \mathrm{H}_t (u-v)}$$

  and H_s is the Hessian matrix of ψ at s
- ρ_ψ depends both on ψ and the set to be clustered
- this bound is tighter than the one in [CKW07] (Crammer, Kearns, Wortman 2007)

Contents

A simple task?

- 100 dimensions
- 10 distributions, all coordinates independently exponentially distributed
- every distribution has its unique 10 coordinates which have the expectation 10, other coordinates have expectation 1/100
- 100 points are sampled from every distribution

Proportion of perfect runs (known clustering is found)

k-means		Itakura-Saito ($\alpha = 1$)	
seeding	%	**seeding**	%
uniform	0.42	uniform	12.63
D^2	1.13	D^2	16.44
Kullback-Leibler-0.5	18.88	Kullback-Leibler-0.5	36.39
Kullback-Leibler-1.0	22.57	Kullback-Leibler-1.0	39.15
Itakura-Saito-0.5	47.07	Itakura-Saito-0.5	51.56
Itakura-Saito-1.0	**51.57**	Itakura-Saito-1.0	**59.18**

Open problems

- Is the factor ρ_ψ really needed in the upper bound? Would an analysis without our "triangle inequality" be possible?
- What about a lower bound?
- other divergences (the bounds hold after the seeding phase)
- How can we utilize the factor α?

References

AV07 D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Proc. of the 18th ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.

BMDG05 A. Banerjee, S. Merugo, I. Dhillon and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

CKW07 K. Crammer, M. Kearns and J. Wortman. Learning form multiple sources. In *Advances in Neural Information Processing Systems 19*, pages 321–328. MIT Press, 2007.

ORSS06 R. Ostrovsky, Y. Rabani, L.J. Schulman and C. Swamy. The effectiveness of Lloyd-type methods for the k-means problem. In *Proc. of the 47th IEEE Symposium on the Foundations of Computer Science*, pages 165–176, IEEE Computer Society, 2006.