Paper Reading Seminar

Learning Probabilistic Submodular Diversity Models Via Noise Contrastive Estimation

Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS), 2016

Our Roadmap

3 main topics and 2 off-track topics

Introduction
"Diversity"
"Summary"

DPPs
Model
Problems

FLID
Model
Results Discussion

Set Theory
Quick Review
Notations

NCE
Z Estimation
My favorite!

Diversification and its application

“Bad” search result example

Google Images

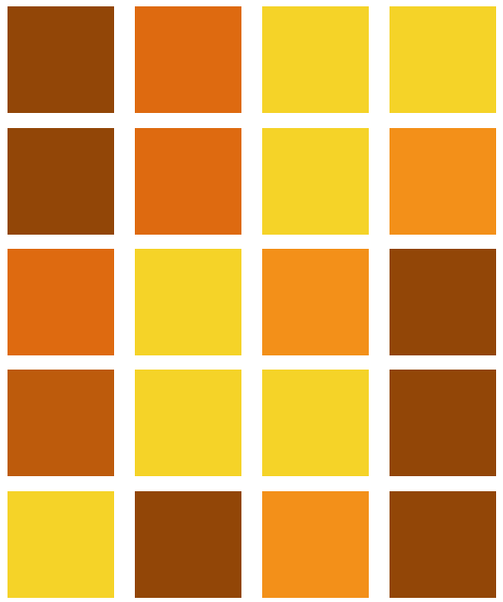
Car

Diversification and its application

Facility location problem and coverage problem

Customers

Location	1	2	3	4
1	1.0	0.5	0.0	0.0
2	1.0	0.6	0.0	0.2
3	0.5	0.0	0.1	1.0
4	0.7	0.0	0.0	1.0
5	0.0	1.0	0.1	1.0

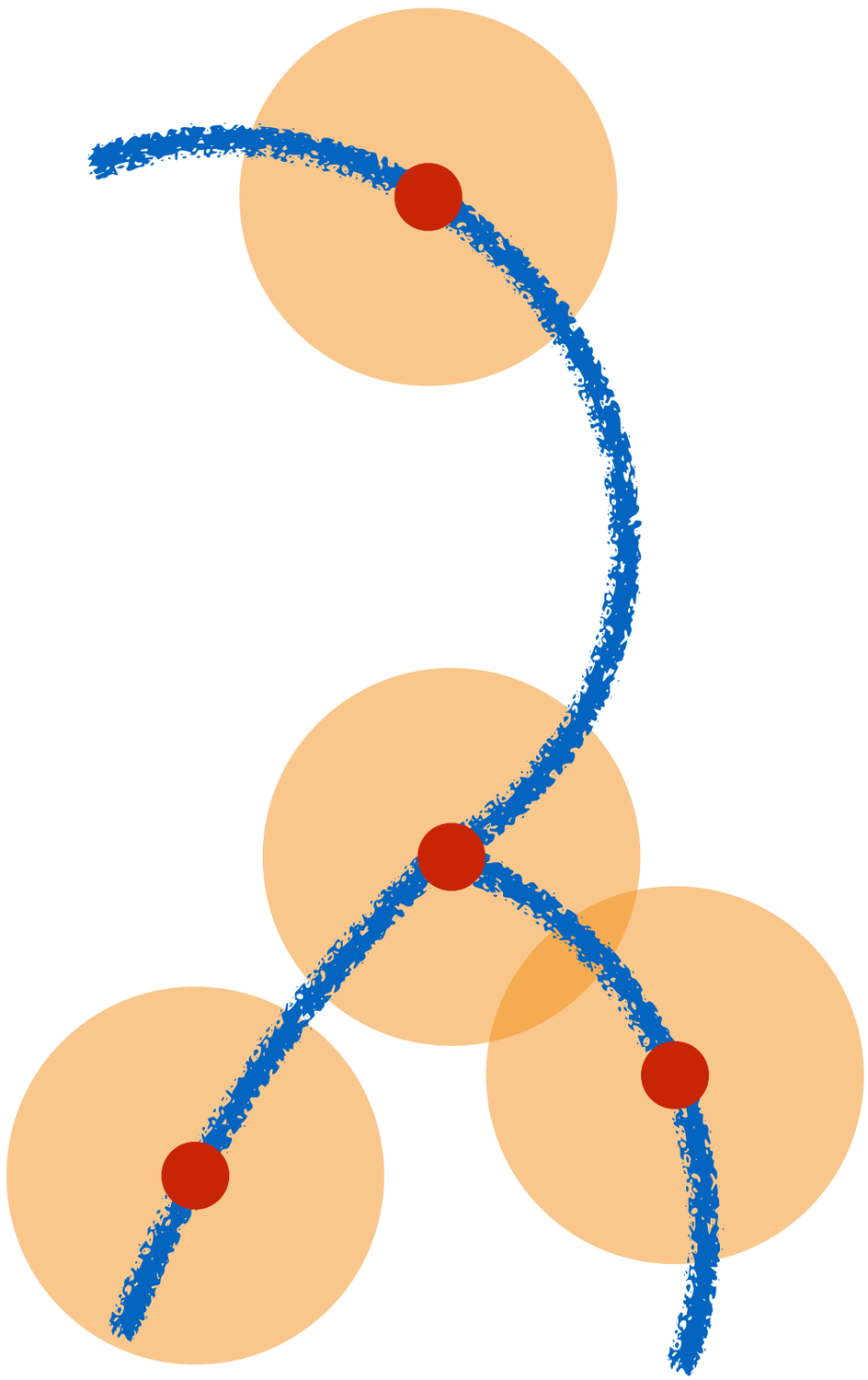


$k = 3$

- {1,4,5}
- {2, 5}

Diversification and its application

The nature of coverage and submodular functions



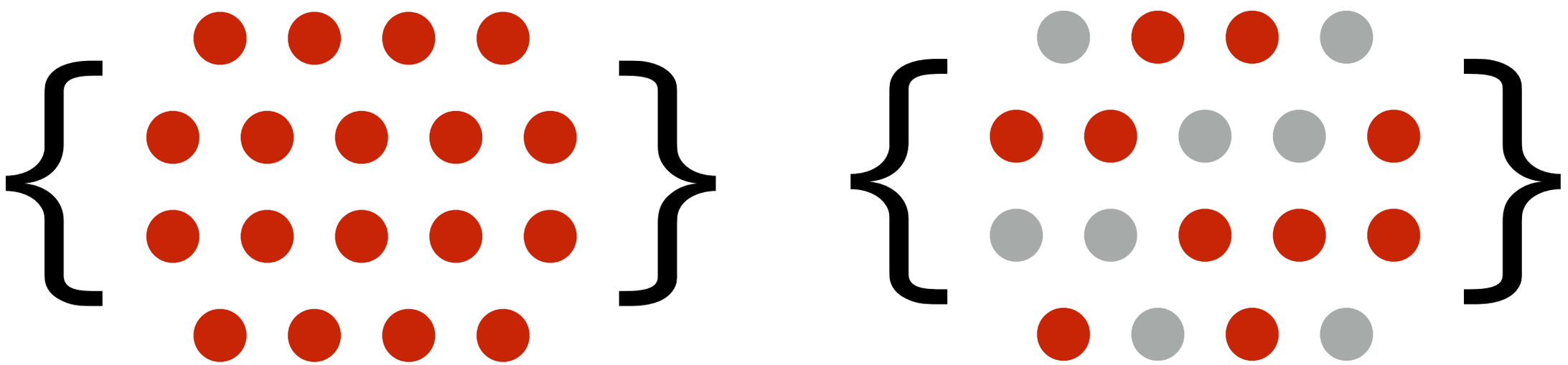
$S = \{ \bullet \bullet \bullet \bullet \}$

$f(S) = \langle coverage \rangle$

Tiny course for set theory

Basic notation

Set S is a collection of “things”.



Ground set (informally) is where we pick elements from.

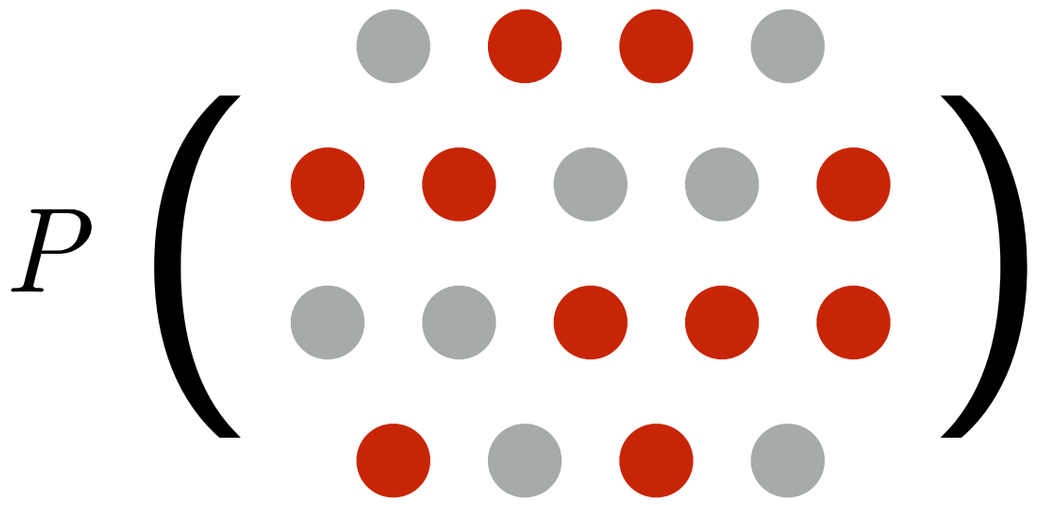
(conf. Matroids and Partially Ordered Sets for more.)

2^V is the power set, or set of subsets of ground set V.

Tiny course for set theory

Point process

Given a set S, a (simple) point process is a random subset A of S. *(Terry Tao, Determinantal processes, 2009)*



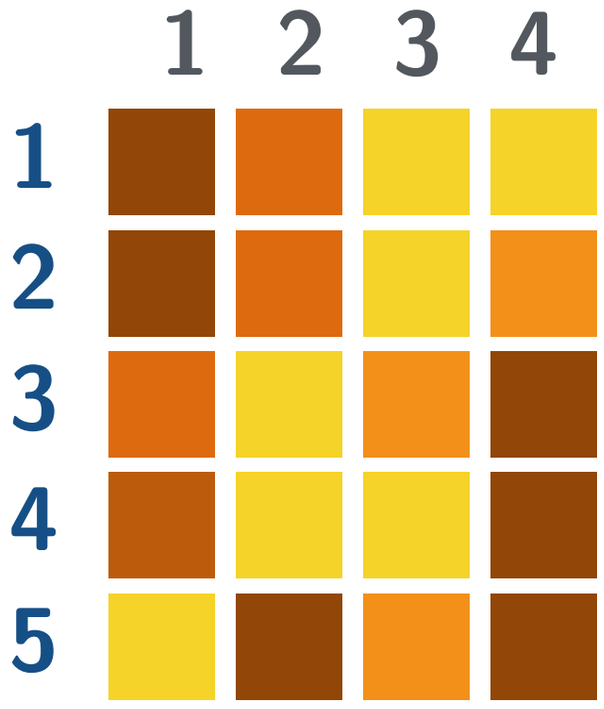
Tiny course for set theory

Submodular functions (also modular and supermodular)

$$F\colon 2^V \to \mathbb{R}$$

$$F(A \cup \{i\}) - F(A) \geq F(B \cup \{i\}) - F(B),$$

$$A \subseteq B \subseteq V \setminus \{i\}.$$



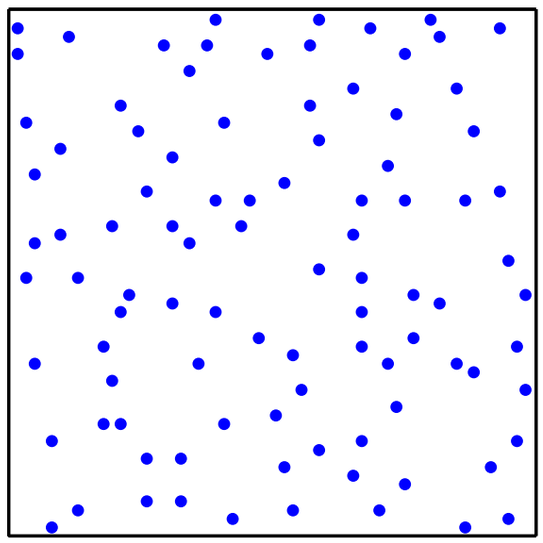
$$F(S) = [\ \max(S_{\mathrm{i,d}})\ \text{for}\ S_{\mathrm{i}}\ \text{in}\ S\ ]$$

$F(\{2 \mid \{1\}\})$ vs. $F(\{2 \mid \{1,3\}\})$

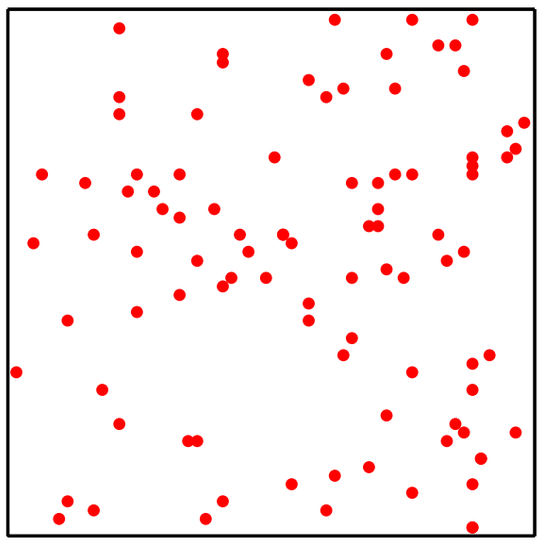
Determinantal Point Processes

Solution to diversity sampling problem

$$P\left(\begin{array}{c}\bullet\,\bullet\,\bullet\,\bullet\\ \bullet\,\bullet\,\bullet\,\bullet\,\bullet\\ \bullet\,\bullet\,\bullet\,\bullet\,\bullet\\ \bullet\,\bullet\,\bullet\,\bullet\end{array}\right) = \det(K_{\mathrm{A}})$$



DPP



Independent

(Kulesza, UAI'11)

Determinantal Point Processes

Real data modeling: L-ensemble

For data modeling, L-ensembles has advantages.

(Borodin, 2009)

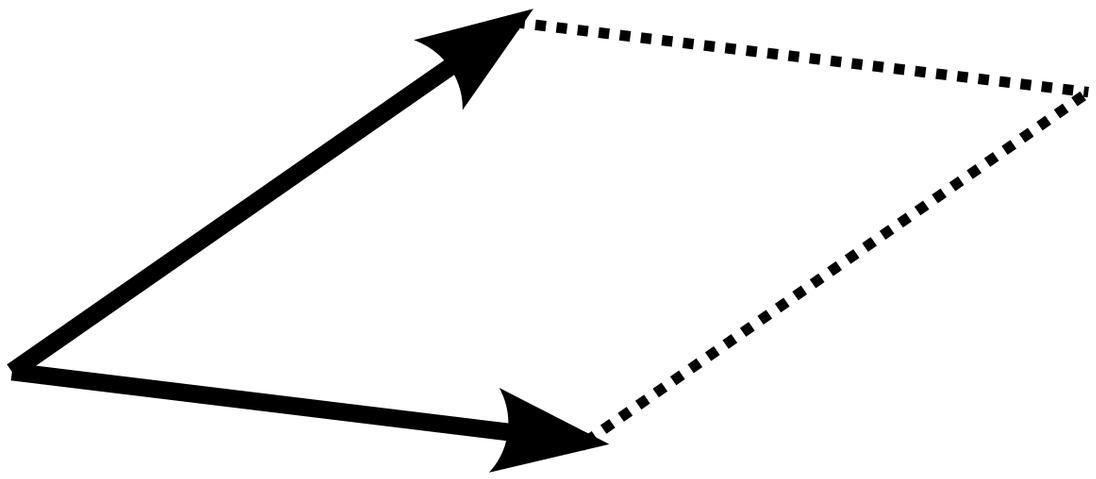
$$\mathcal{P}_L(\mathbf{Y} = Y) = \frac{\det(L_Y)}{\det(L + I)}$$

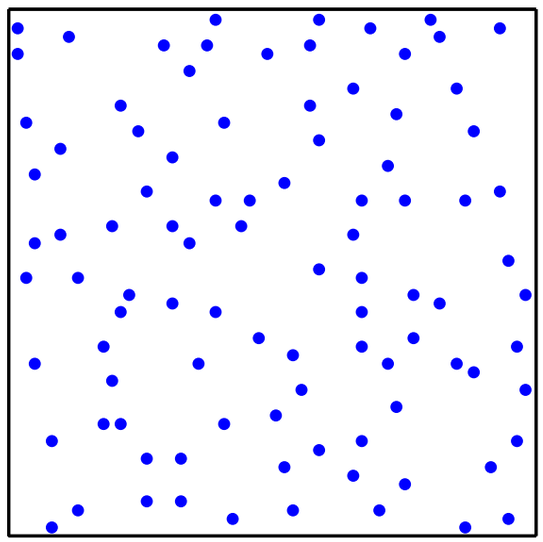
$$K = (L + I)^{-1} L$$

$$L_{ij} = q_i \phi_i^\top \phi_j q_j$$

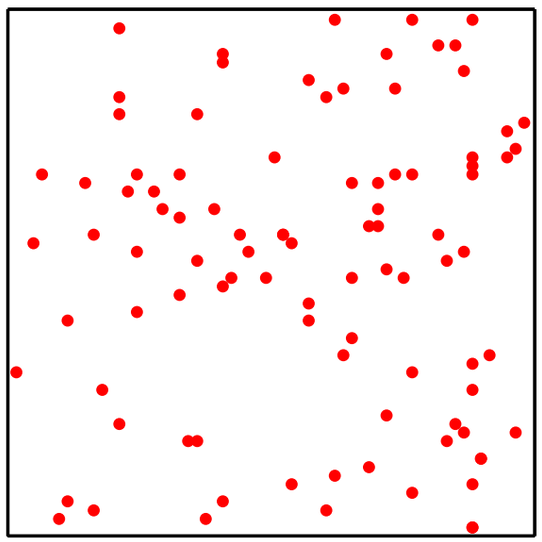
Determinantal Point Processes

Some intuition and conditional DPPs





DPP



Independent

(Kulesza, UAI'11)

Noise Contrastive Estimation

Fast approximation for partition function

(Guntmann, AISTATS'10)

$$\int p_m(\mathbf{u};\hat{\alpha})\mathrm{d}\mathbf{u} = 1.$$

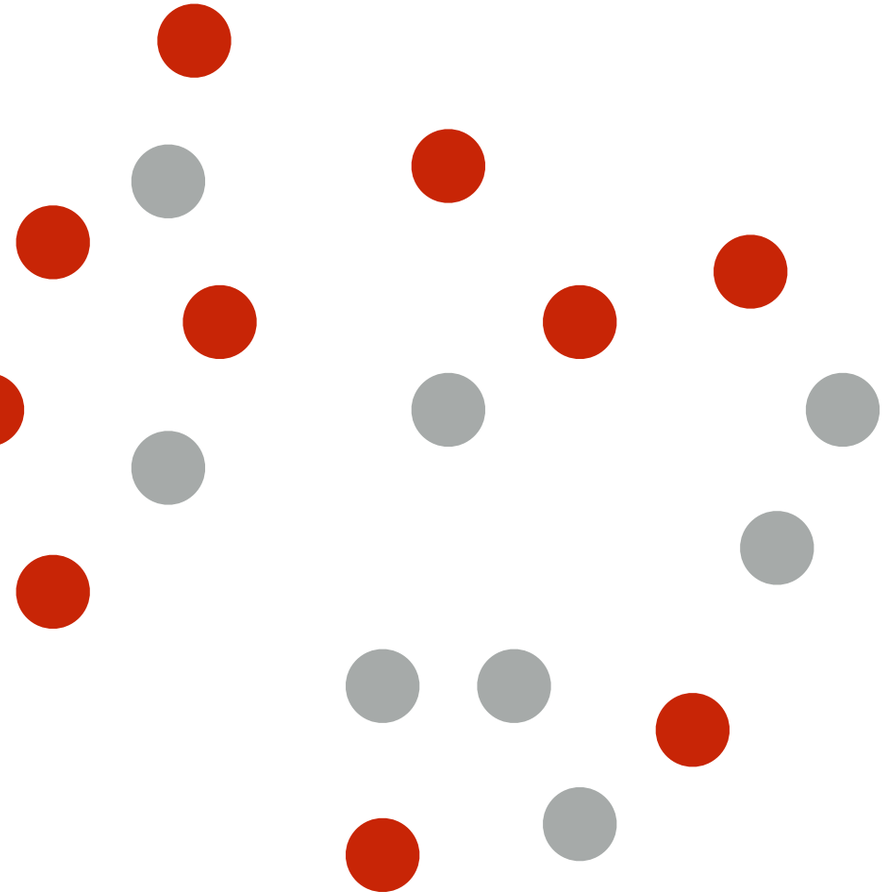
$$p_m(.;\alpha) = \frac{p_m^0(.;\alpha)}{Z(\alpha)}, \quad Z(\alpha) = \int p_m^0(\mathbf{u};\alpha)\mathrm{d}\mathbf{u},$$

$$\ln p_m(.;\theta) = \ln p_m^0(.;\alpha) + c$$

$$\begin{aligned} P(C=1|\mathbf{u};\theta) &= \frac{p_m(\mathbf{u};\theta)}{p_m(\mathbf{u};\theta)+p_n(\mathbf{u})} \\ &= h(\mathbf{u};\theta) \\ P(C=0|\mathbf{u};\theta) &= 1-h(\mathbf{u};\theta). \end{aligned}$$

Noise Contrastive Estimation

Classification between “positive” and “negative” samples



Facility LocatIon Diversity

NCE-inspired learning algorithm

(Tschiatschek, AISTATS'16)

Log-probabilistic model:

$$P\left(\begin{array}{ccccc} \bullet & \bullet & \circ & \circ & \bullet \\ \circ & \circ & \bullet & \bullet & \bullet \\ & \bullet & \circ & \bullet & \end{array}\right) \propto \exp(\textit{something})$$

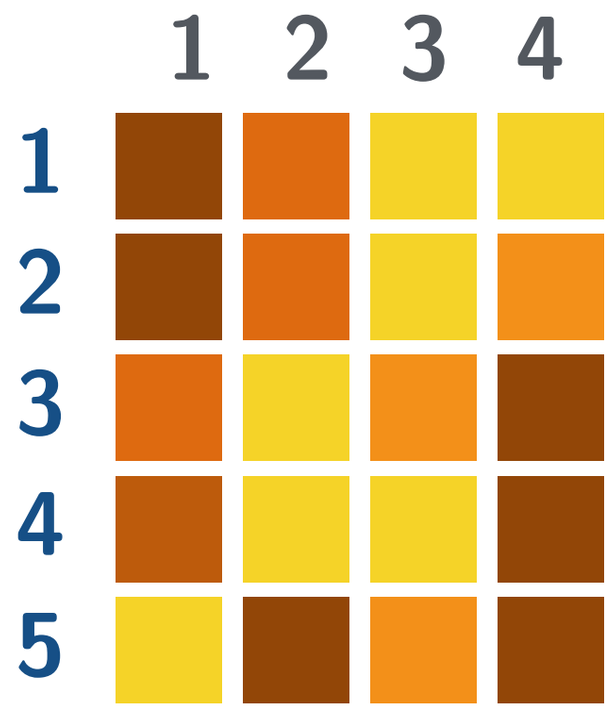
$$P(S \mid \mathbf{u}, \mathbf{W}) = \underbrace{\frac{1}{Z} \exp\left(\sum_{i \in S} u_i + \underbrace{\sum_{d=1}^{L} \left(\max_{i \in S} w_{i,d} - \sum_{i \in S} w_{i,d} \right)}_{\mathrm{Div}(S)} \right)}_{\widetilde{P}(S|\mathbf{u},\mathbf{W})},$$

Facility Location Diversity

Modular and submodular formulation

Modular

$$P(S) \propto \exp\left(\sum_{i \in S} u_i\right)$$



Div({1,2}) = -1.5
Div({2,4}) = -0.9

Submodular

$$P(S \mid \mathbf{u}, \mathbf{W})$$

$$= \frac{1}{Z} \exp\left(\underbrace{\sum_{i \in S} u_i + \underbrace{\sum_{d=1}^{L} \left(\max_{i \in S} w_{i,d} - \sum_{i \in S} w_{i,d}\right)}_{\mathrm{Div}(S)}}_{\widetilde{P}(S|\mathbf{u},\mathbf{W})}\right),$$

index-d of a L-dim vector

Facility LocatIon Diversity

Computational complexity for partition function

Straight-forward way to compute Z is $\mathcal{O}(|V|^{L+1})$

Therefore, NCE comes to the rescue!

$$P(Y_S = 1 \mid S) = \frac{P(S \mid Y_S = 1)}{P(S \mid Y_S = 1) + \eta P(S \mid Y_S = 0)} = \frac{P_d(S)}{P_d(S) + \eta P_n(S)},$$

$$g(\theta) = \sum_{S \in \mathcal{D}} \log P(Y_S = 1 \,|\, S, \theta) + \sum_{S \in \mathcal{N}} \log P(Y_S = 0 \,|\, S, \theta)$$

$$\theta = [\mathbf{u}_{\mathrm{NCE}}, \mathbf{W}_{\mathrm{NCE}}, \hat{Z}]$$

Facility Location Diversity

A small note on gradient

$$\nabla \log P(Y_S = Y \mid S)$$
$$= \left(Y - \frac{1}{1 + \eta \frac{P_n(S)}{\frac{1}{\hat{Z}} \widetilde{P}(S|\mathbf{u},\mathbf{W})}} \right) \nabla \log \frac{1}{\hat{Z}} \widetilde{P}(S \mid \mathbf{u}, \mathbf{W})$$

$$\nabla \log\left(\frac{Y P_d(S) + (1-Y)\eta P_n(S)}{P_d(S) + \eta P_n(S)}\right) = \frac{(2Y-1)\eta P_n(S)}{(P_d(S) + \eta P_n(S))(Y P_d(S) + (1-Y)\eta P_n(S))} \nabla P_d(S)$$

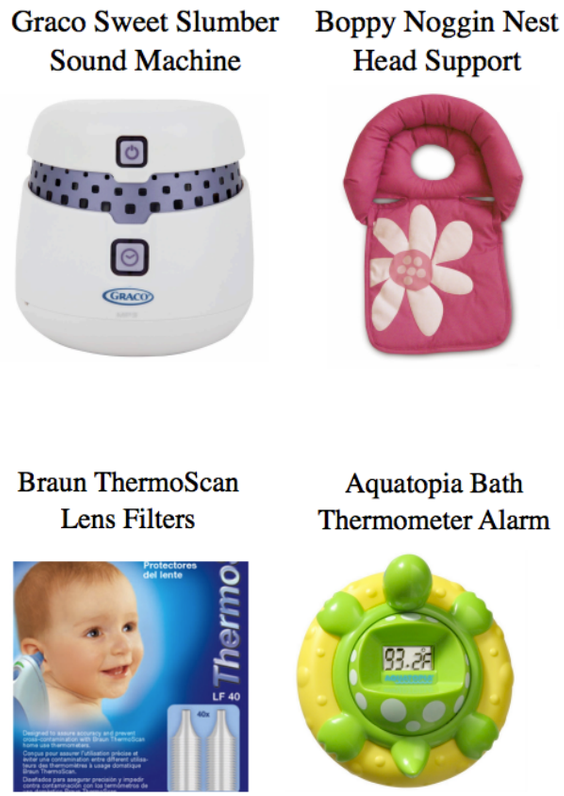
$$= \frac{(2Y-1)\eta P_n(S)}{(P_d(S) + \eta P_n(S))(Y + (1-Y)\eta \frac{P_n(S)}{P_d(S)})} \frac{1}{P_d(S)} \nabla P_d(S)$$

** Y only takes values of 0 or 1*

Facility LocatIon Diversity

Data for evaluation

- 29,632 baby registries from amazon.com
- 18 categories
- Filtered out listing with fewer than 5 or more than 100 products.
- Create sub-registries by categories. Further filtered out items that did not occur in at least 100 sub-registries.
- Result: 13 categories (avg. 71 products each), 8585 sub-registries. 70% for fitting model. 30% for testing.



Facility Locatlon Diversity

Evaluation metrics

Model fitting:

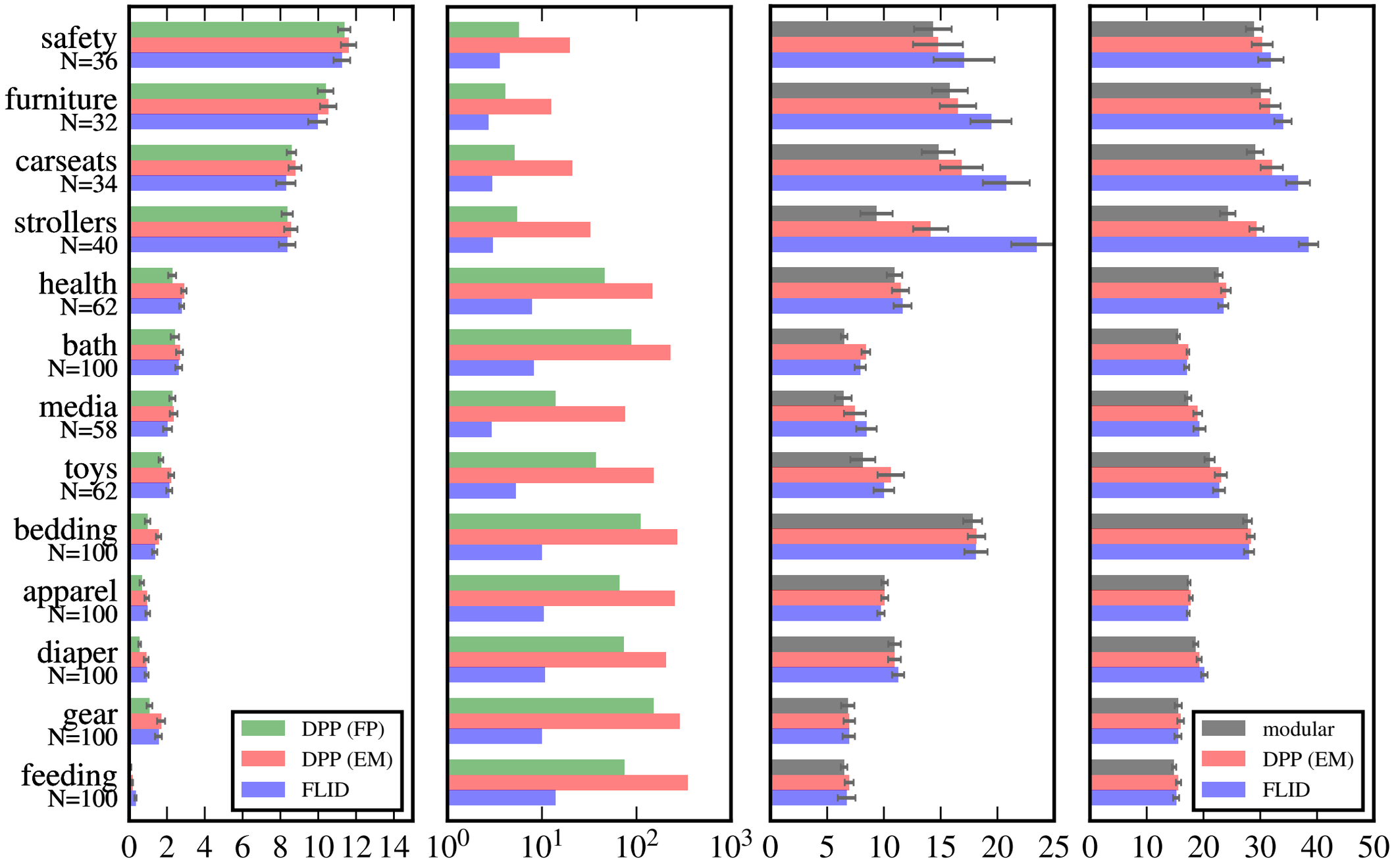
$$\mathrm{LLRI} = 100 \cdot \frac{\mathcal{L}_{\mathrm{method}} - \mathcal{L}_{\mathrm{modular}}}{|\mathcal{L}_{\mathrm{modular}}|},$$

Mean reciprocal rank (MRR)

$$\mathrm{MRR} = \frac{100}{|\mathcal{T}'|} \sum_{\check{S}_i \in \mathcal{T}'} \frac{1}{\mathrm{rank}_i^{\check{S}_i}}.$$

Facility LocatIon Diversity

Result for Amazon baby registry



(a) Log-likelihood relative improvement

(b) Cumulative runtime in seconds

(c) Accuracy

(d) MRR

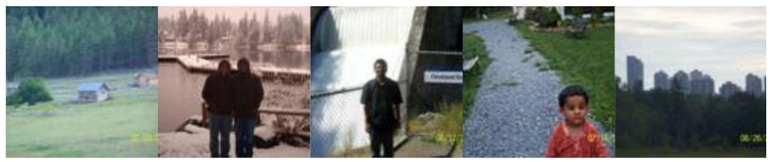
FLID

Facility LocatIon Diversity

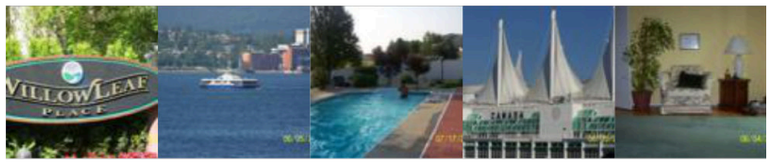
Result for Image summarization

(a) Image collection

Given:



H



1

2

3

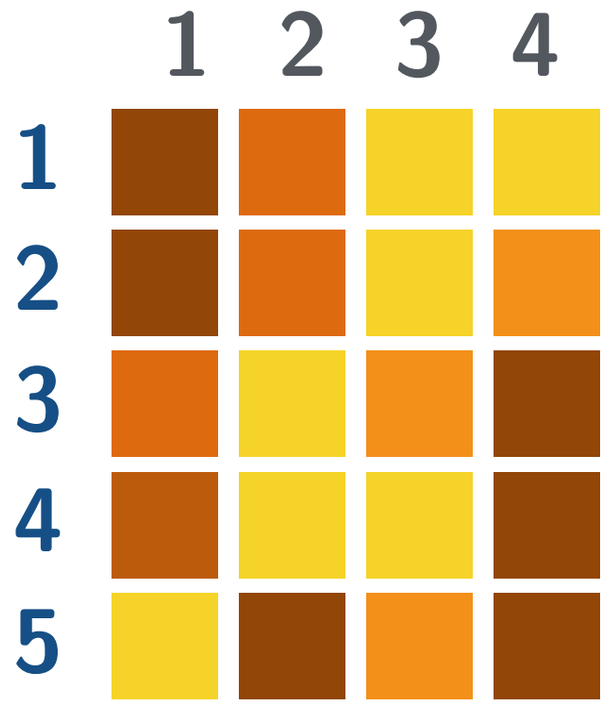
4

(b) Completed summaries

Facility Location Diversity

Discussion

$$P(S \mid \mathbf{u}, \mathbf{W}) = \underbrace{\frac{1}{Z} \exp\left(\sum_{i \in S} u_i + \underbrace{\sum_{d=1}^{L} \left(\max_{i \in S} w_{i,d} - \sum_{i \in S} w_{i,d} \right)}_{\mathrm{Div}(S)} \right)}_{\widetilde{P}(S|\mathbf{u},\mathbf{W})},$$



$$g(\theta) = \sum_{S \in \mathcal{D}} \log P(Y_S = 1 \,|\, S, \theta) + \sum_{S \in \mathcal{N}} \log P(Y_S = 0 \,|\, S, \theta)$$

The importance / sensitivity of the parameter eta.
Transitivity of FLID compared to DPPs.
Other application / usable dataset for FLID.

References

(Kulesza, UAI'11) Learning Determinantal Point Processes.

(Guntmann, AISTATS'10) Noise-contrastive estimation: A new estimation principle for unnormalized statistical models.

(Tschiatschek, AISTATS'16) Learning Probabilistic Submodular Diversity Models Via Noise-Contrastive Estimation.

(Terry Tao, 2009) Determinantal Point Processes.