Negative Dependence, Stable Polynomials etc in ML

Part 1

STEFANIE JEGELKA & SUVRIT SRA

Dept of EECS & CSAIL

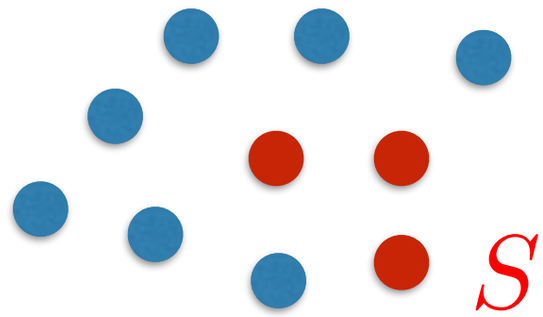
Massachusetts Institute of Technology

Neural information Processing Systems, 2018

ml.mit.edu

Negative Dependence

Discrete probability measure $\mu(S), S \subseteq V$
Equivalently: n binary random variables X_i



negative dependence

0 1 0 1 1 0

$$\mathbb{P}(\in S) \leq \mathbb{P}(\in S)\mathbb{P}(\in S)$$

$$\mathbb{P}(\in S \mid \in S) \leq \mathbb{P}(\in S)$$

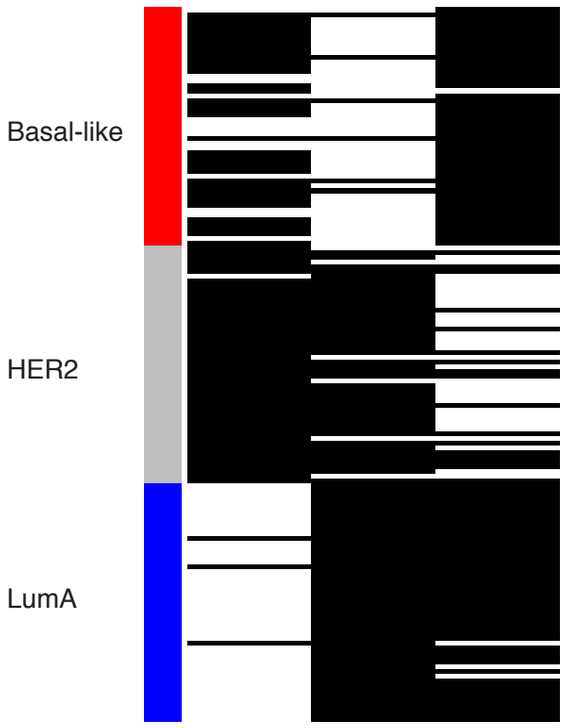
Negative dependence - where?



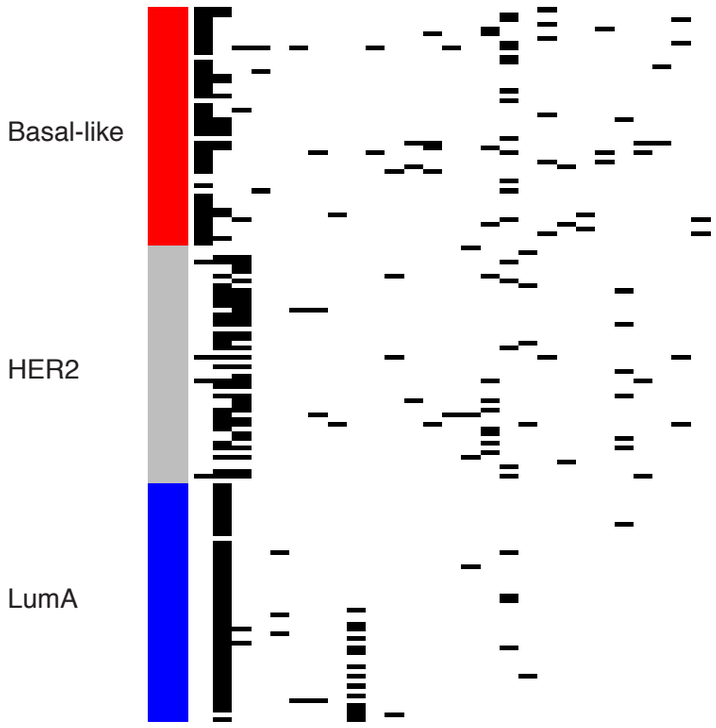



Priors on diversity
Interpretability

Negative Dependence

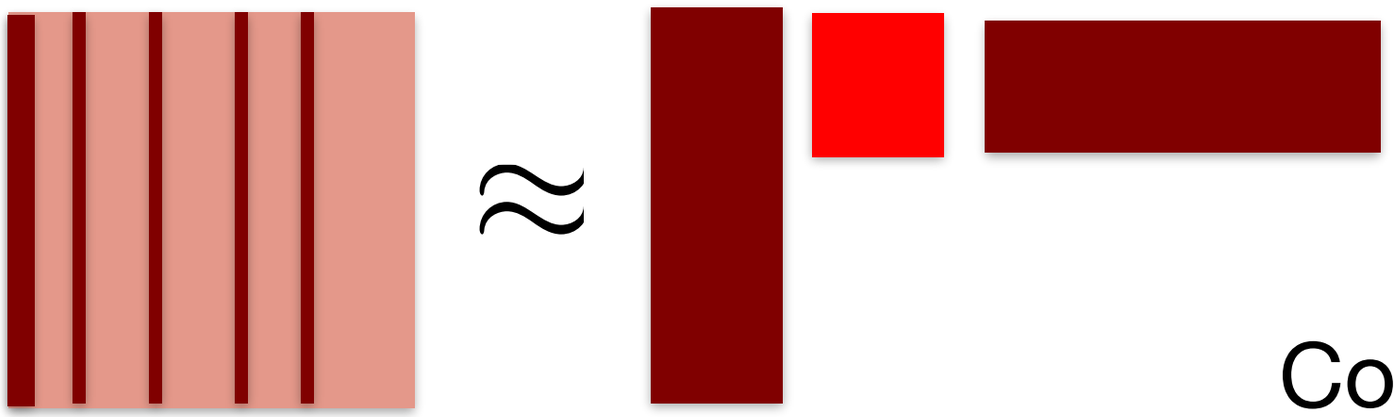


Indian Buffet Process

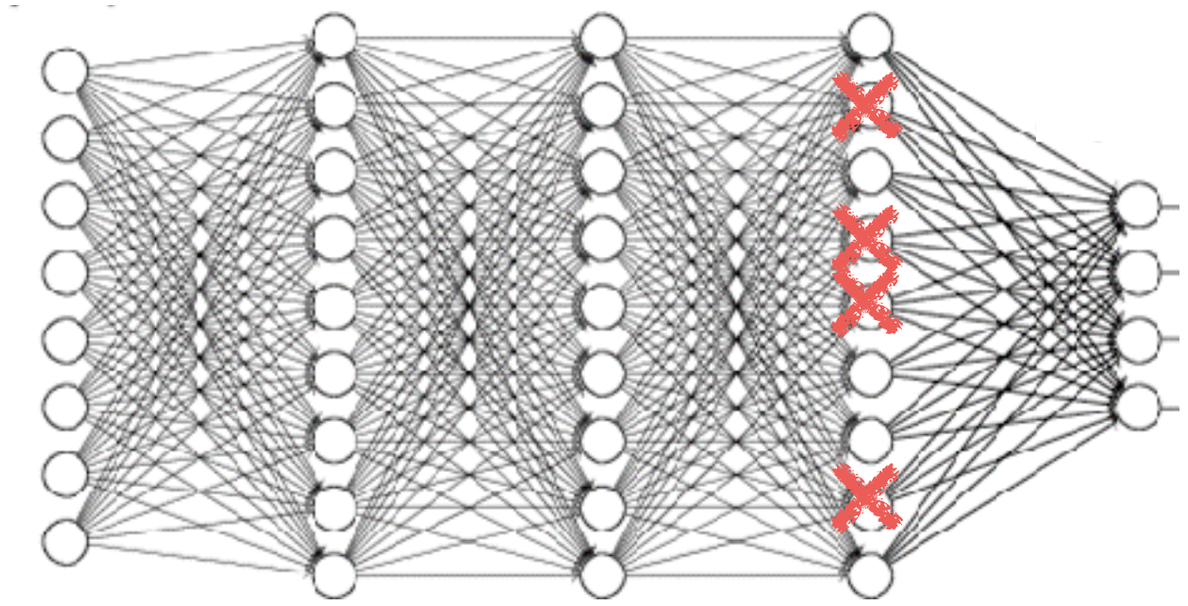


(Xu, Müller, Telesca 2015)

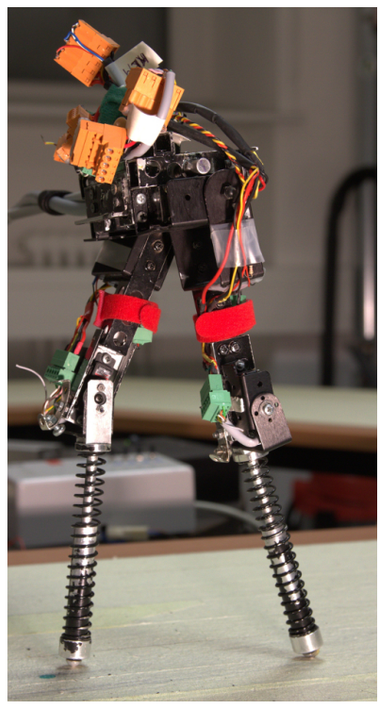
Negative dependence - where?



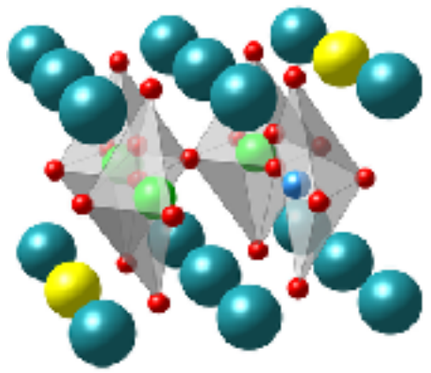
Compressing data
and models



Negative dependence - where?



Exploration,
active learning,
experimental design,
Bayesian Optimization



Outline

1 Intro & Theory

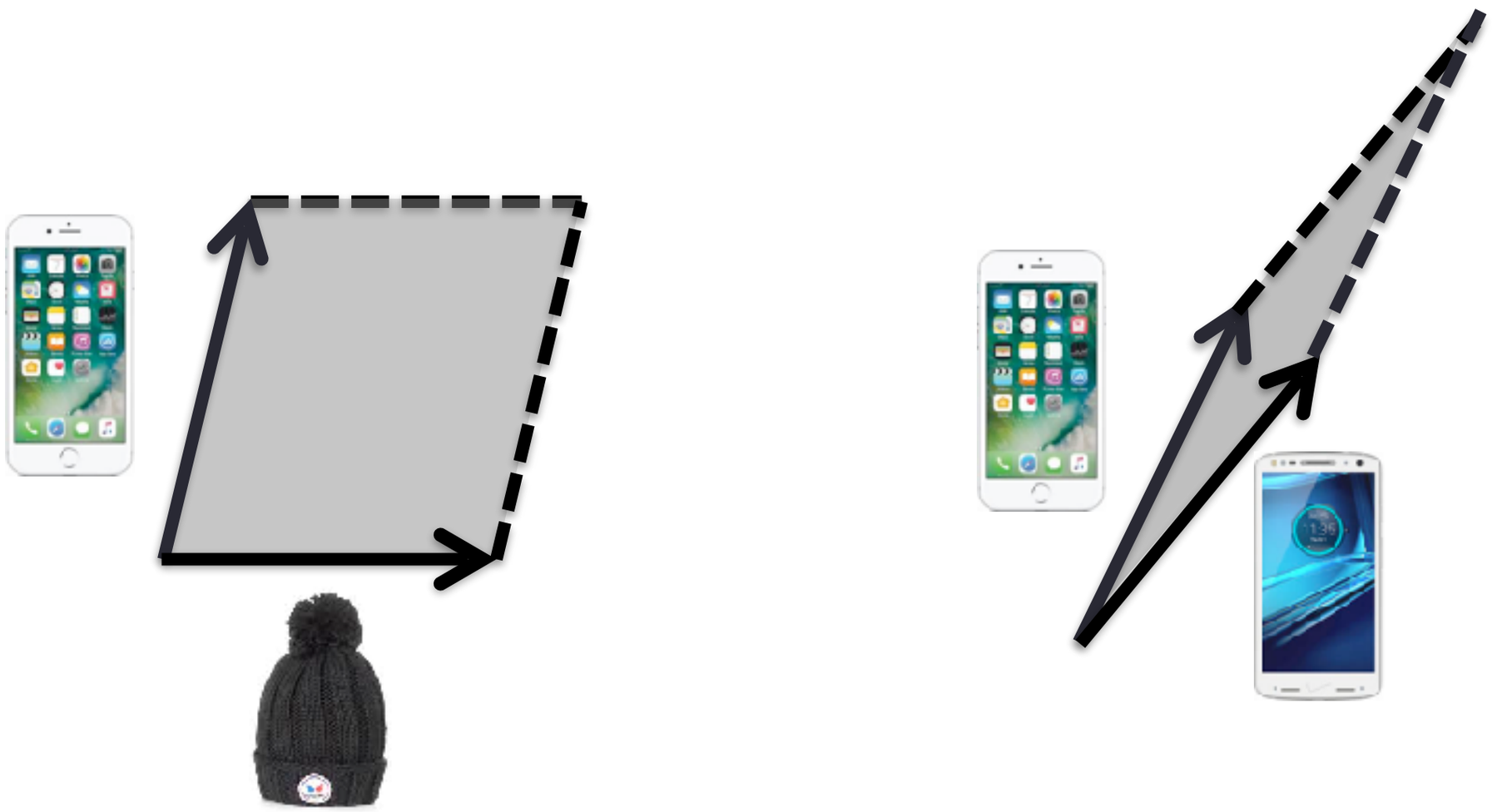
- **Introduction**
  Prominent example: Determinantal Point Processes
- **Stronger notions of negative dependence**
- **Implications: Sampling**

—BREAK—

2 Theory & Applications

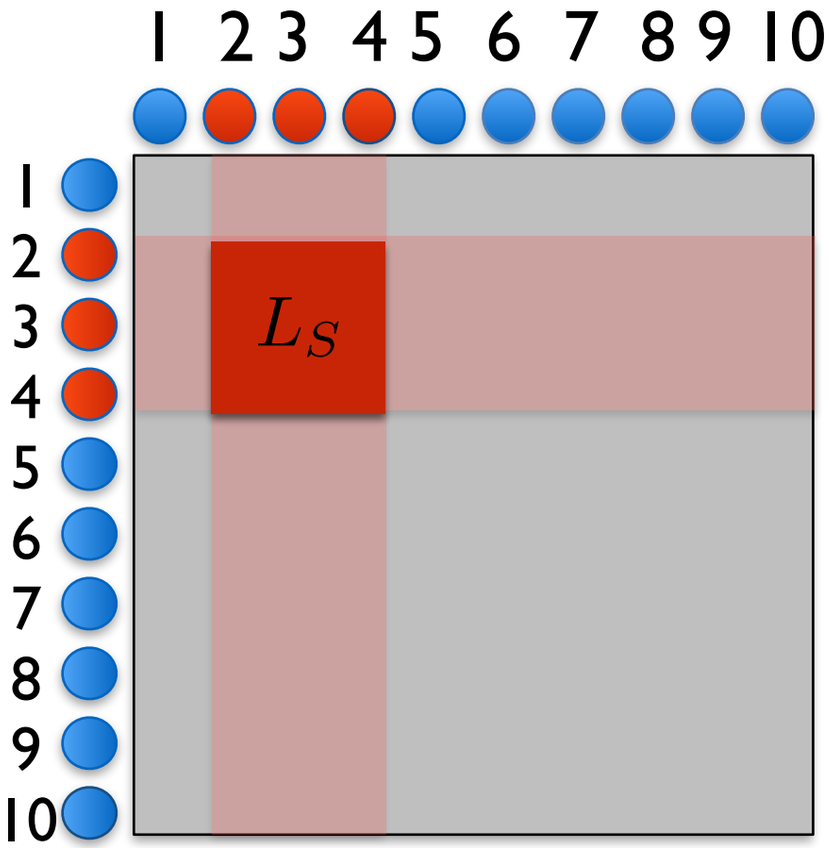
- **Approximating partition functions**
- **Learning a DPP (and some variants)**
- **Applications**
- **Perspectives and wrap-up**

Capturing Diversity: Determinantal Point Processes



$$\mu(S) \propto \mathrm{Vol}^2(\{v_i\}_{i \in S})$$

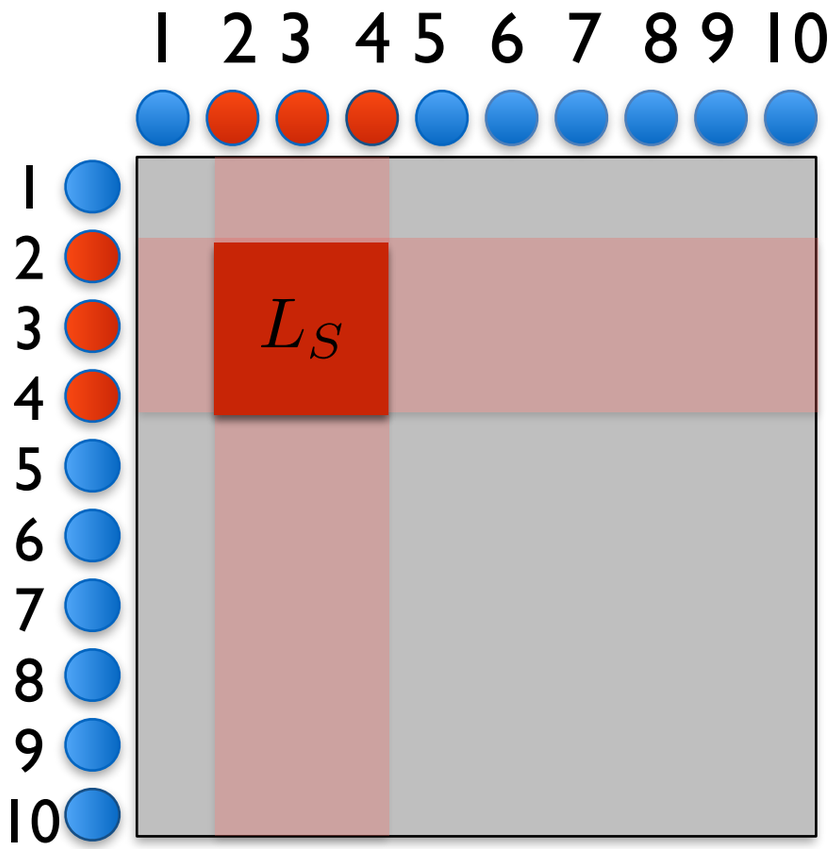
Determinantal Point Processes



PSD similarity matrix L

$$\mu(S) \propto \det(L_S)$$

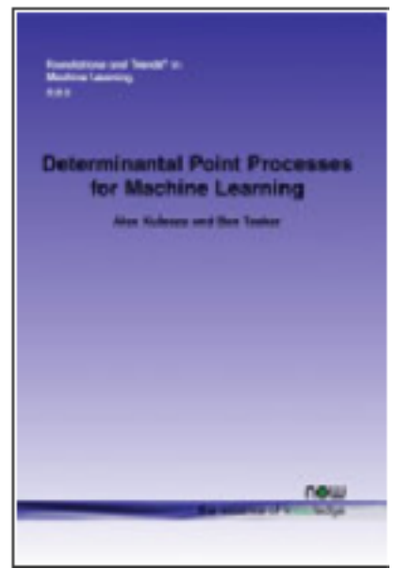
Determinantal Point Processes (DPPs)



PSD similarity matrix L

$$\mu(S) = \frac{\det(L_S)}{\det(L+I)}$$

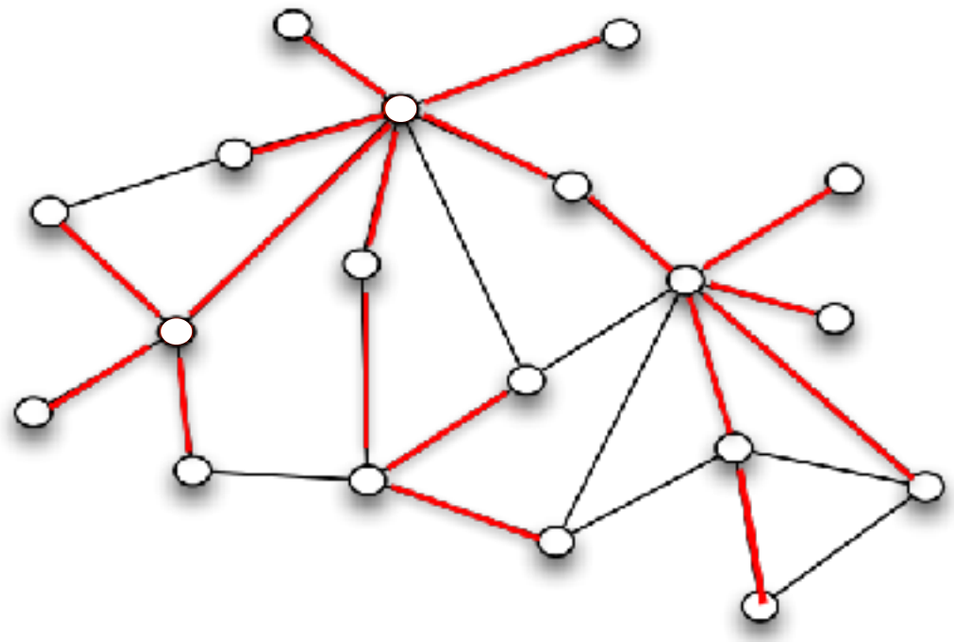
- *Macchi 1975*: “fermion processes”
- *Borodin & Olshanski 2000*: “Determinantal Point Process”
- Introduction to ML: *Kulesza & Taskar*



(Hough, Krishnapur, Peres, Virag 2006; Lyons 2014; Lyons & Peres 2016; Pemantle 2000)

Combinatorial Examples

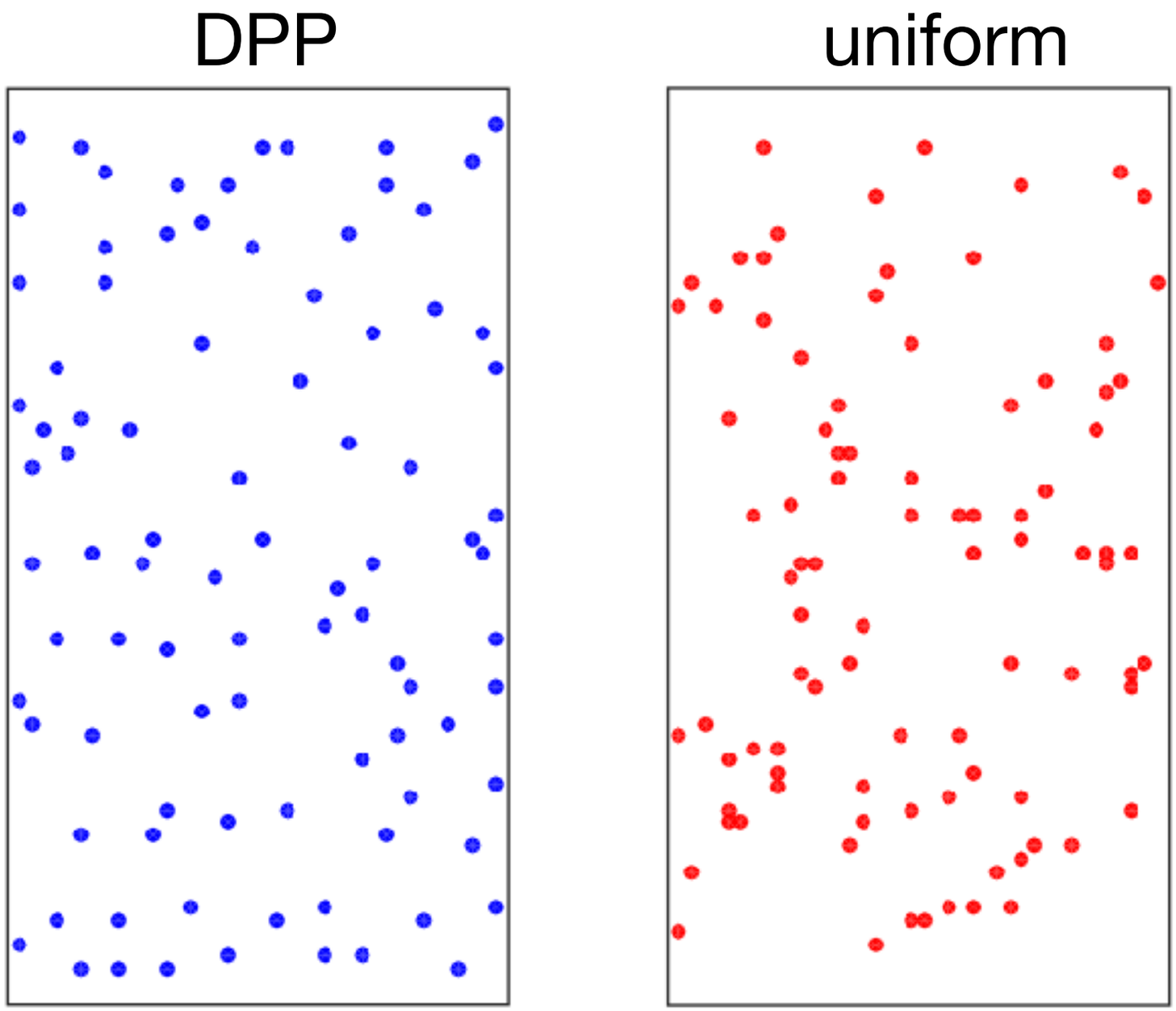
random spanning trees



ascents in random sequences

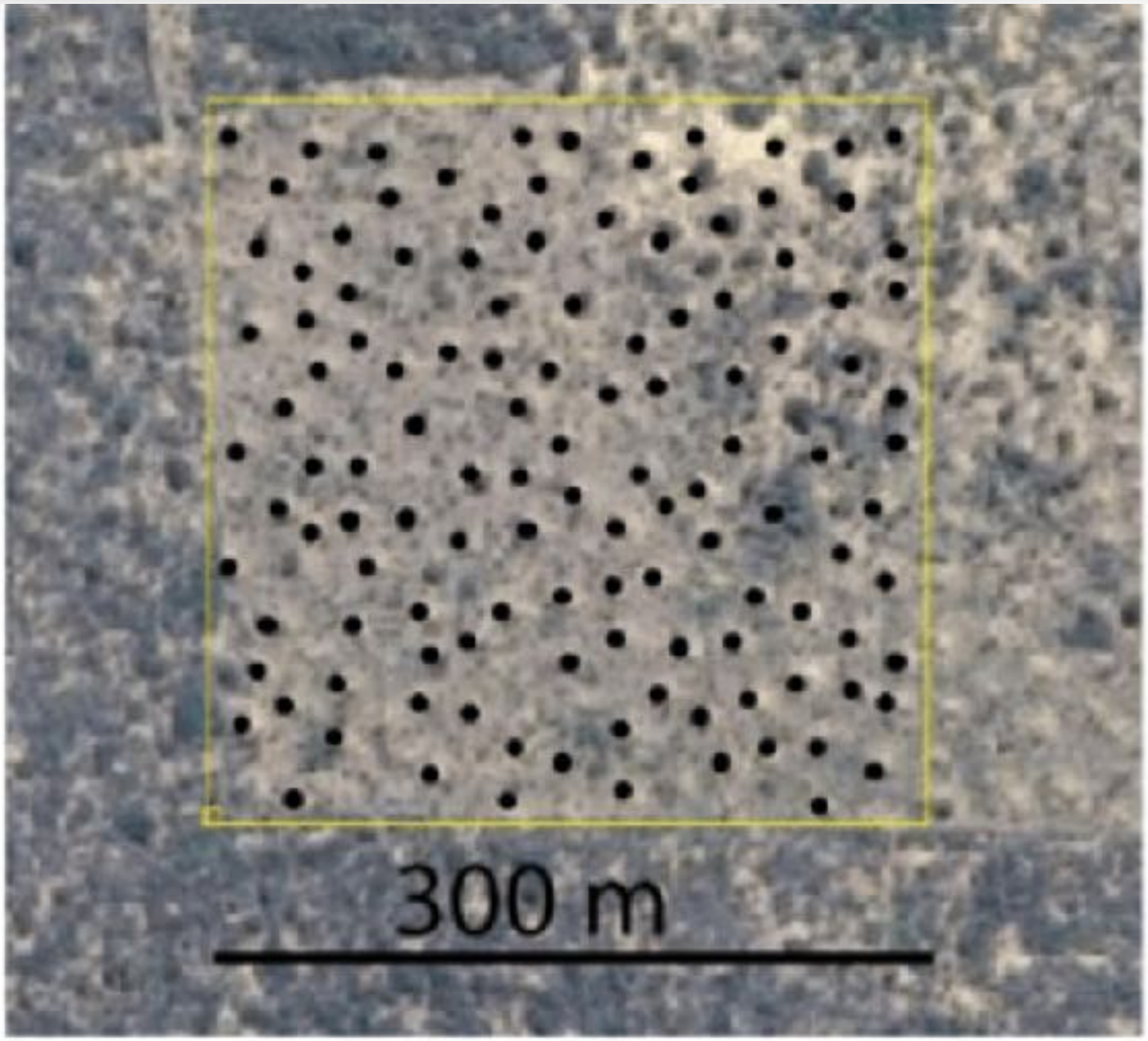
0 2 8 0 9 7 4 5 2 4 9 5 5 2 4

DPP samples



$$l_{ij} = \exp(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2)$$

Repulsion in nature



termite mounds in Brazil

(Martin, Funch, Hanson, Yoo, Current Biology 2018,
http://djalil.chafai.net/blog/2018/11/23/yet-another-determinantal-point-process-in-nature/,)

Outline

Introduction

Examples

Determinantal Point Processes

1

Intro & Theory

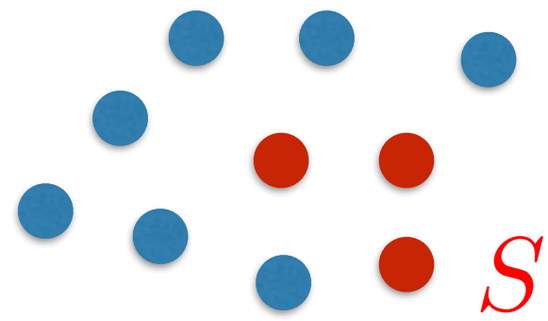
Stronger notions of negative dependence

Relations to polynomials

Implications: Sampling

Negative Dependence

Discrete probability measure $\mu(S), S \subseteq V$
Equivalently: n binary random variables X_i



negative dependence

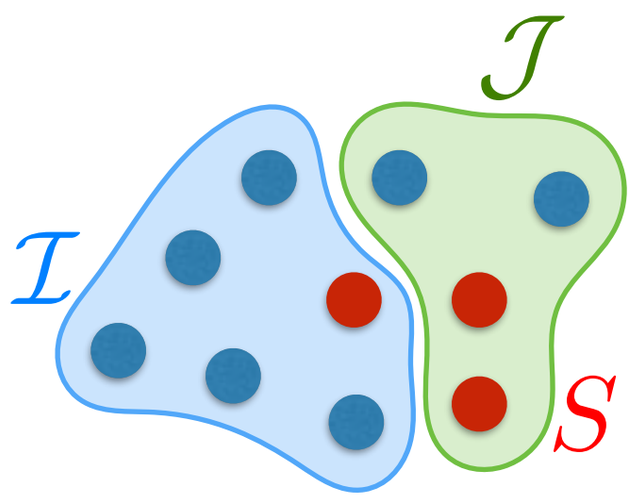
0 1 0 1 1 0

$$\mathbb{P}(\text{📱}\,\text{📱} \in S) \leq \mathbb{P}(\text{📱} \in S)\mathbb{P}(\text{📱} \in S)$$

Stronger Notions

$$\mathbb{P}(\text{[phone]} \in S) \leq \mathbb{P}(\text{[phone]} \in S)\mathbb{P}(\text{[phone]} \in S)$$

equivalently: $\mathbb{E}X_iX_j \leq \mathbb{E}X_i\,\mathbb{E}X_j$

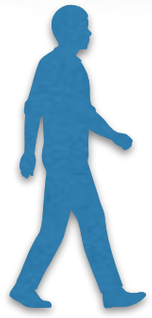


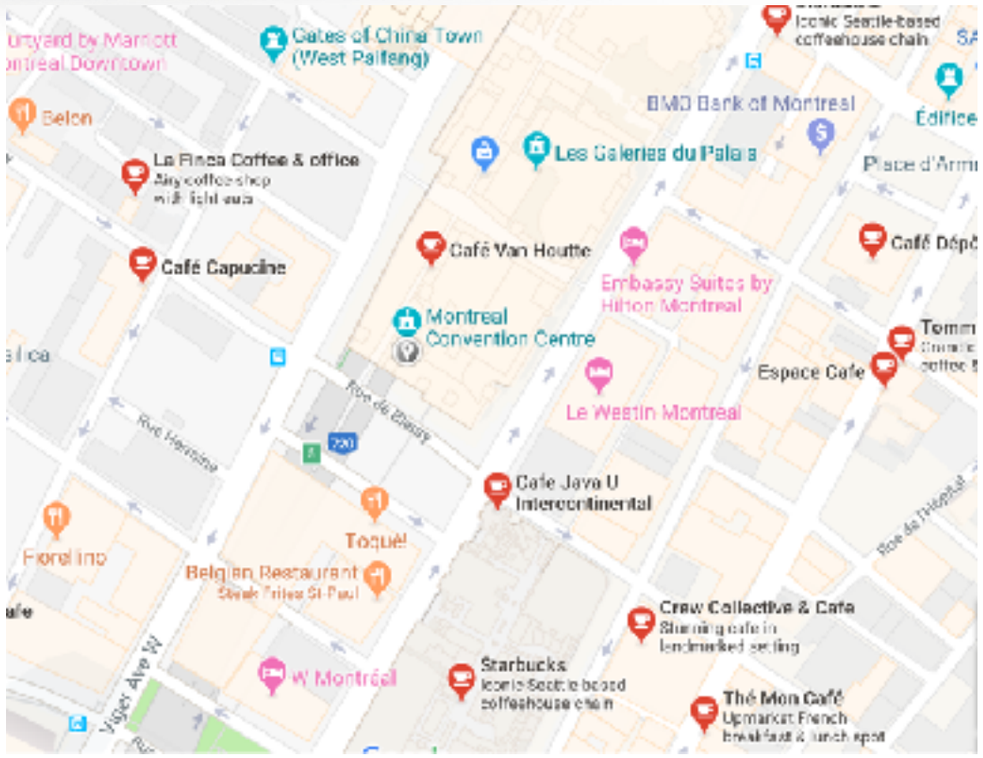
Negative Association

For all monotone increasing functions $G(S), H(S)$

$$\mathbb{E}[G(S \cap \mathcal{I})H(S \cap \mathcal{J})] \leq \mathbb{E}G(S \cap \mathcal{I})\,\mathbb{E}H(S \cap \mathcal{J})$$

Example





$$B_i = \begin{cases} 1 & \text{if thirsty researcher goes to coffee shop } i \\ 0 & \text{otherwise} \end{cases}$$

Detour: Positive Dependence

Positive Association:

For all monotone increasing functions $G(S), H(S)$

$$\mathbb{E}[G(S)H(S)] \quad \geq \quad \mathbb{E}G(S)\,\mathbb{E}H(S)$$

FKG Lattice Condition, Multivariate Totally Positive:

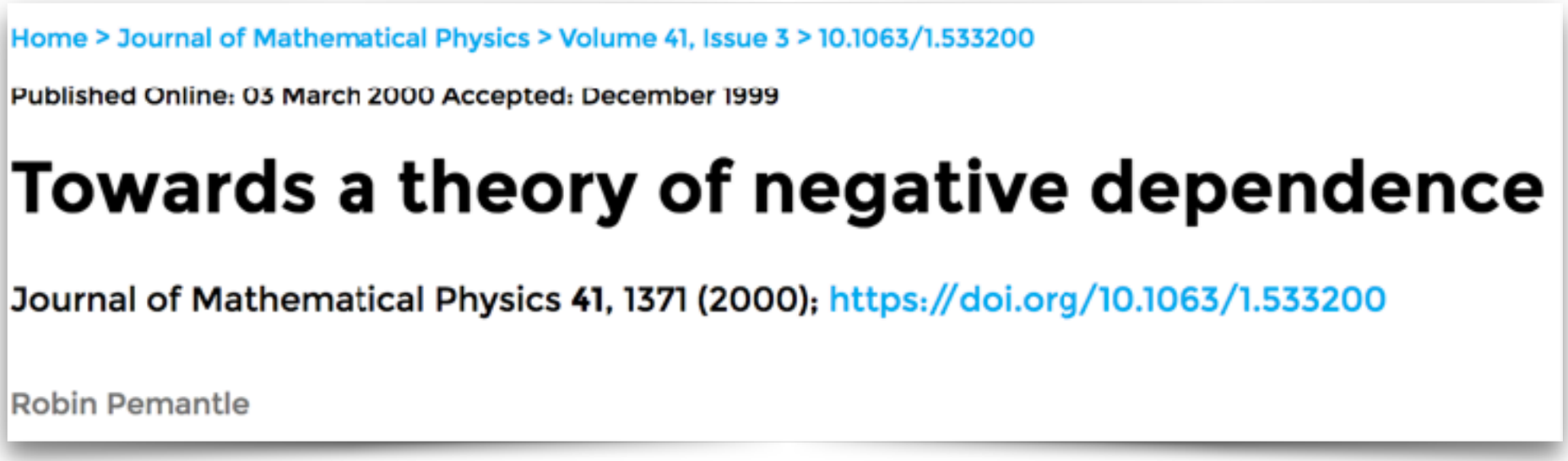
log-supermodularity

$$\mu(S)\mu(T) \;\leq\; \mu(S \cup T)\mu(S \cap T) \quad \forall S, T \subseteq V$$

implies positive association *(Fortuin,Kasteleyn,Ginibre 1971)*

Analog does not hold for negative dependence!

negative association implies log-submodularity, but not conversely


Home > Journal of Mathematical Physics > Volume 41, Issue 3 > 10.1063/1.533200

Published Online: 03 March 2000 Accepted: December 1999

Towards a theory of negative dependence

Journal of Mathematical Physics 41, 1371 (2000); https://doi.org/10.1063/1.533200

Robin Pemantle

The “right notion” should imply:

- Negative Association
- Stochastic Covering
- Log-concave rank sequences $a_k = \mathbb{P}(|S| = k)$
- Closed under “natural” operations
  - conditioning
  - marginalization
  - products …

Generating Polynomial

$$q_\mu(z) = \sum_{S \subseteq V} \mu(S) \prod_{i \in S} z_i, \quad z \in \mathbb{C}^n$$

Example:
2 items

$$q_\mu(z) = \mu_\emptyset + \mu_1 z_1 + \mu_2 z_2 + \mu_{1,2} z_1 z_2$$

Operations on polynomial = operations on distribution

Obtain coefficient $\mu_1 = \mu(\{1\})$

1. differentiate wrt z_1: $\dfrac{\partial}{\partial z_1} q_\mu(z) = \mu_1 + \mu_{1,2} z_2$

2. set $z = 0$

$$\left.\frac{\partial}{\partial z_1} q_\mu(z)\right|_{z=0}$$

Generating Polynomial

$$q_\mu(z) = \sum_{S \subseteq V} \mu(S) \prod_{i \in S} z_i, \quad z \in \mathbb{C}^n$$

Example:
2 items

$$q_\mu(z) = \mu_\emptyset + \mu_1 z_1 + \mu_2 z_2 + \mu_{1,2} z_1 z_2$$

Marginalization: $\pi(\{1\}) = \mathbb{P}(1 \in S)$

Set $z_2 = 1$

$$q_\mu(z_1, 1) = [\mu_\emptyset + \mu_2] + [\mu_1 + \mu_{1,2}] z_1 \qquad = q_\pi(z_1)$$

$\mathbb{P}(1 \notin S)$ $\qquad$ $\mathbb{P}(1 \in S)$

Properties of polynomial $\Leftrightarrow$ Properties of distribution

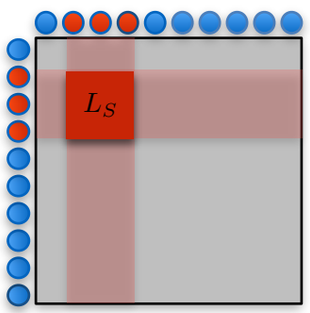
Strongly Rayleigh Measures

$$q_\mu(z) = \sum_{S \subseteq V} \mu(S) \prod_{i \in S} z_i, \quad z \in \mathbb{C}^n$$

$\mu(S)$ is **Strongly Rayleigh (SR)** if $q_\mu(z)$ is *real stable:*

$$\mathrm{Im}(z_i) > 0 \ \forall i \quad \Rightarrow \quad q_\mu(z) \neq 0$$

SR implies almost all conditions laid out by Pemantle!
(Borcea, Bränden, Liggett 2009)



Determinantal Point Process: $q_\mu(z)$ essentially multivariate variant of the characteristic polynomial $\det(z_i I - L)$

Operations on polynomial $\Leftrightarrow$ Operations on distribution

Real Stable Polynomials

$$\mathrm{Im}(z_i) > 0 \ \forall i \quad \Rightarrow \quad q_\mu(z) \neq 0$$

- ✦ Deep mathematical connections
- ✦ Univariate case goes back to Newton (at least)
- ✦ Powerful class of closure properties

Generating polynomial $q(z)$ is SR is equivalent to

$$\frac{\partial q(x)}{\partial z_i}\frac{\partial q(x)}{\partial z_j} \geq q(x)\frac{\partial^2 q(x)}{\partial z_i \partial z_j}, \quad x \in \mathbb{R}^n \ \ \forall i,j$$

$$\Rightarrow \quad \mu(S)\mu(T) \ \geq \ \mu(S \cup T)\mu(S \cap T) \quad \forall S, T \subseteq V$$

Nice properties of SR

- **Closed** under marginalization,
  conditioning on $|S| = k, X_i = 1, X_i = 0$
  ...

- Implies many other types of **negative dependence**,
  e.g. negative association and $\mathbb{E} \prod_i X_i \leq \prod_i \mathbb{E} X_i$

- **Concentration of measure**: X_i behave like independent random variables
  sum $\sum_{i=1}^{n} X_i$, Lipschitz functions $F(S)$, matrices
  (Panconesi,Srinivasan 1997, Dubhashi,Ranjan 1998, Farcomeni 2008, Pemantle,Peres 2011, Kyng,Song 2018, Garbe,Vondrak 2018,…)

Algorithmic Implications

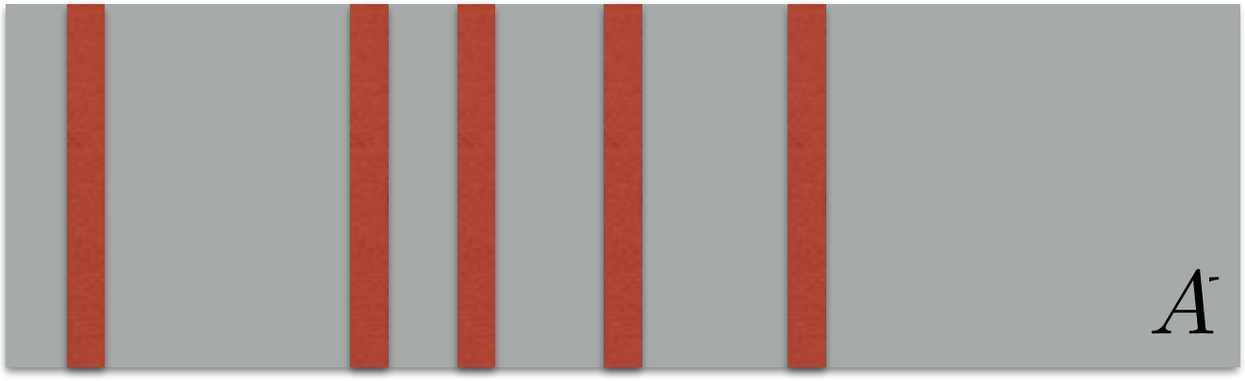
- Sampling
  (Féder,Mihail 1992, Jerrum,Son 2002, Jerrum,Son,Tetali,Vigoda 2004, Anari,Gharan,Rezaei 2016, Li,Jegelka,Sra 2016)

- Approximate partition functions, permanents, volumes, counting
  (Gurvits 2006, Nikolov-Singh 2016, Straszak-Vishnoi 2016, Anari,Gharan, Saberi,Singh 2016, Anari,Gharan 2017…)

- Approximation Algorithms
  (Panconesi,Srinivasan 1997, Hayes 2003, Considine,Byers,Mitzenmacher 2004, Asadpour,Goemans,Madry,Oveis-Gharan,Saberi 2010, Gharan, Saberi,Singh 2011, Spielman,Srivastava 2011, …)

Other SR measures

Dual Volume Sampling



$$P(S) \propto \det(A_S A_S^\top)$$

NOT a DPP … but **Strongly Rayleigh!**

proof: closeness properties of polynomials

(Avron,Boutsidis 2013, Li,Jegelka,Sra 2017, Derezinski,Warmuth 2017)

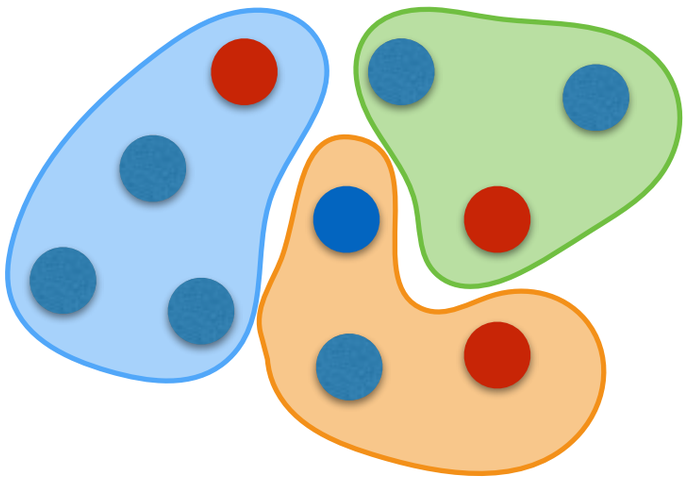
… and limits

$\mu(S)$ is SR. What about $\mu(S)^p$? In general not SR!

(Kulesza,Taskar 2012, Zou,Adams 2012, Gillenwater 2014, Anari, Gharan 2017, Mariet,Sra,Jegelka 2018)

Conditioning on combinatorial constraints: at most one item per group. In general not SR!

(Celis,Deshpande,Kathuria,Straszak,Vishnoi 2017, Celis,Keswani, Straszak,Deshpande,Kathuria,Vishnoi 2018)



Outline

Introduction

Examples

Determinantal Point Processes

1 **Intro & Theory**

Stronger notions of negative dependence

Relations to polynomials

Strongly Rayleigh measures

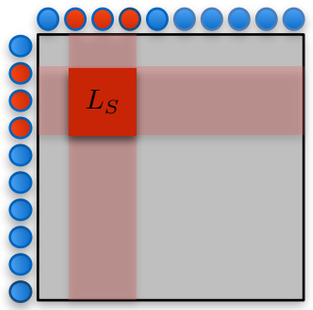
→ **Implications: Sampling**

Sampling for Determinantal Point Processes

EVD/SVD *(Hough, Krishnapur, Peres, Virág 2006)* $O(n^3)$

- ✦ sample subspace via eigenvalues
- ✦ sequentially sample items

Faster via *A*: "dual" sampling *(Kulesza, Taskar 2010)*



$$L = A^\top A$$

$n \times d$

Acceleration: approximate L or A

volume-preserving sketching *(Magen, Zouzias 2008; Deshpande, Rademacher 2010, Gillenwater, Kulesza, Taskar 2012)*
Nyström *(Affandi, Kulesza, Fox, Taskar 2013)*
MCMC *(Bardenet, Hardy 2016)*
Coresets *(Li, Jegelka, Sra 2016)*
R-DPP *(Derezinski 2018)*

Other approaches *(Derezinski, Warmuth 2018, Derezinski, Warmuth, Hsu 2018, Mariet, Sra 2016,…)*

Sampling for general SR measures

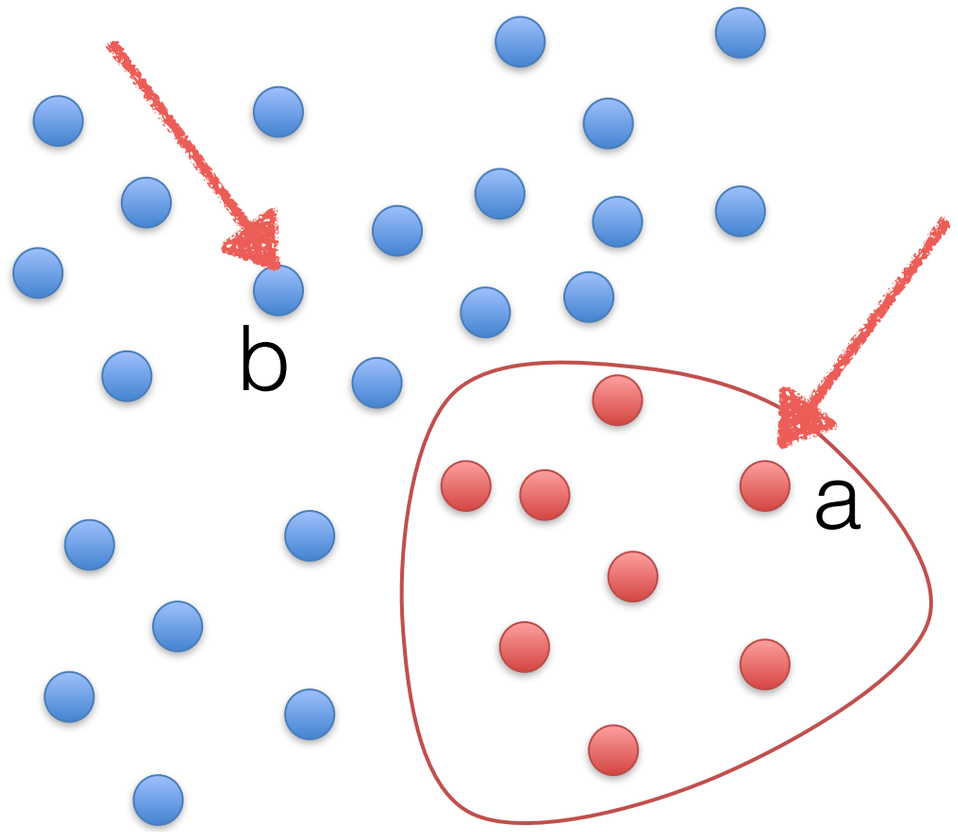
Markov Chain Monte Carlo (MCMC)

in iteration t:

sample $a \in S_{t-1}, b \notin S_{t-1}$ uniformly at random

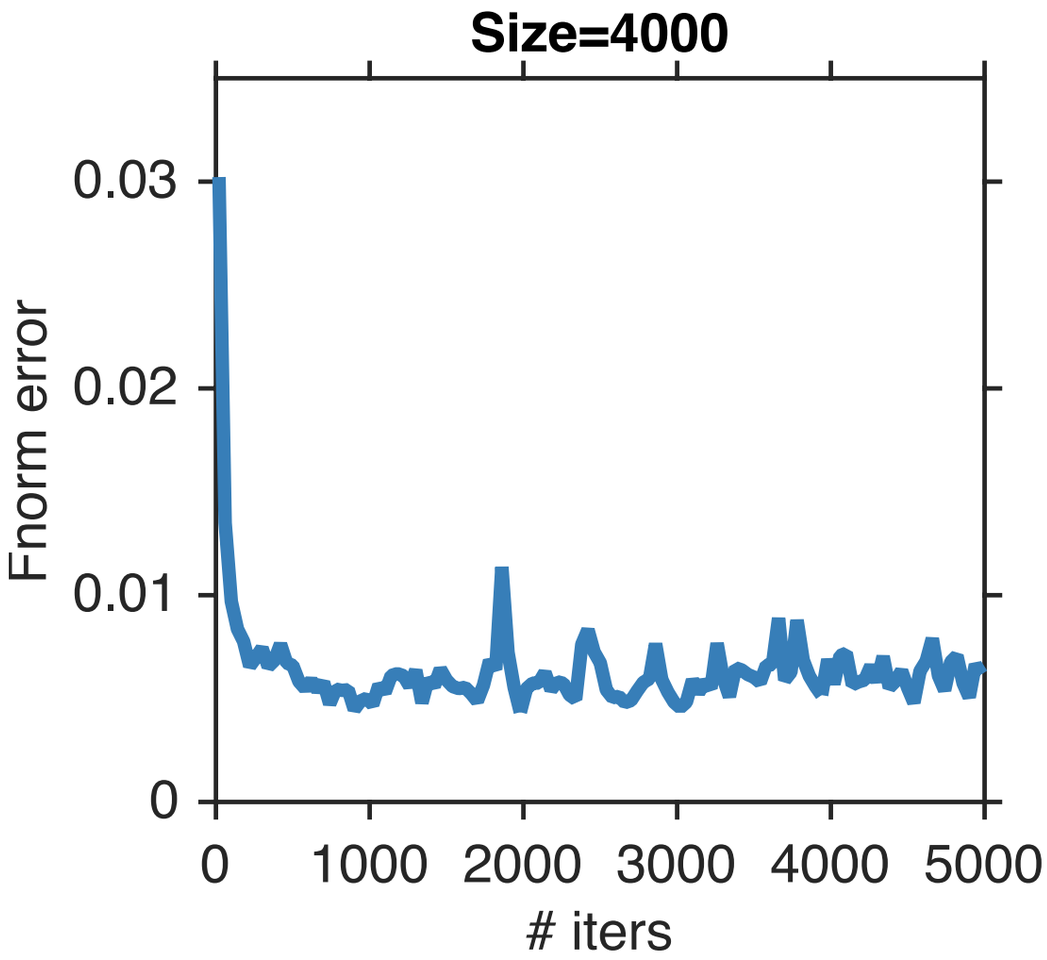
randomly do one of

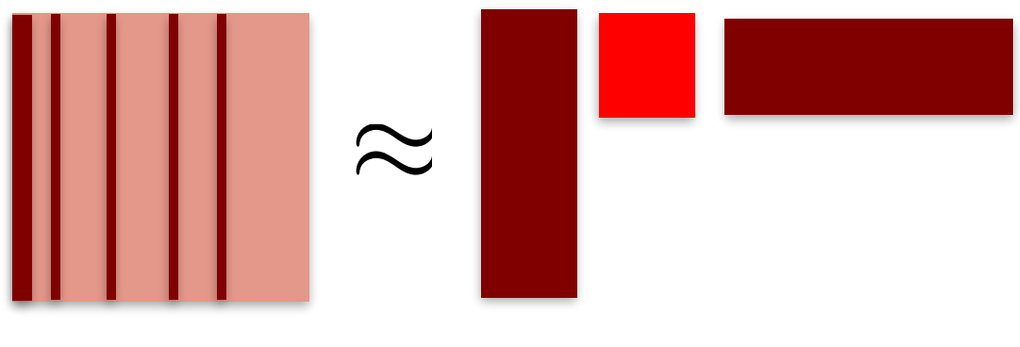
- swap: remove b, add a
- add b
- remove a
- retain $S_t = S_{t-1}$



if conditioning on $|S| = k$: only swaps

Sampling empirically





How many iterations?

Mixing time: #iterations until $\|\mu^t - \mu\|_{\mathrm{TV}} \leq \epsilon$

balanced matroids:

(Féder,Mihail 1992, Jerrum,Son 2002
Jerrum,Son,Tetali,Vigoda, 2004)

$$O\big(nk \log(\tfrac{1}{\epsilon \mu(S_0)})\big) \quad \Big\} |S| = k$$

fixed-cardinality SR:

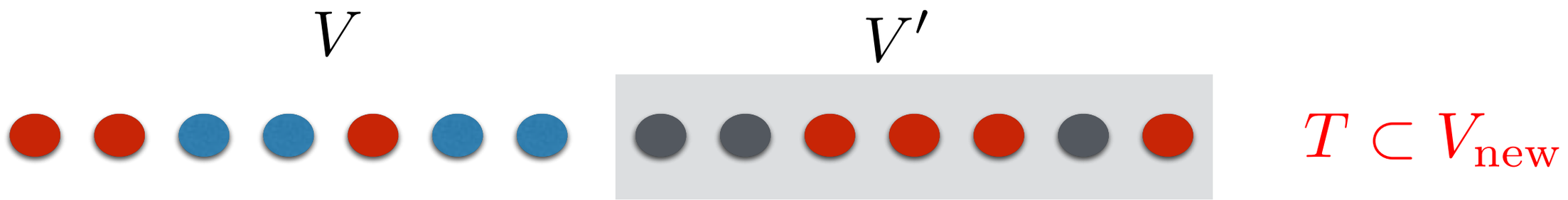
(Anari,Oveis-Gharan,Rezaei 2016)

arbitrary SR:

(Li,Jegelka,Sra 2016)

$$O\big(n^2 \log(\tfrac{1}{\epsilon \mu(S_0)})\big)$$

Sampling: from fixed-cardinality to general SR



$$V_{\text{new}} = V \cup V'$$

- sample n out of 2n, but use only $T \cap V$
- extend measure to “shadow set”
- ***Key:*** measure on 2n is Strongly Rayleigh by closedness properties of real stable polynomials (“symmetric homogenization”)

(Li, Jegelka, Sra 2016)

Sampling

- generic method: MCMC;
  different algorithms for special cases (DPP, dual volume)
- Accelerating each iteration of DPP-MCMC: lazy computations with Gauss quadrature *(Li,Sra,Jegelka 2016)*
- Continuous DPP sampling
  (Affandi,Fox,Taskar 2013, Gharan,Rezaei 2018, …)
- SR sufficient for fast mixing but not necessary: e.g., complete log concavity *(Anari,Liu,Gharan,Vinzant 2018)*
- Sampling log-submodular distributions (negative lattice condition) *(Rebeschini,Karbasi 2015, Gotovos,Hassani, Krause 2018, Gotovos,Hassani,Krause,Jegelka 2018)*

Outline

1 **Intro & Theory**

Introduction

Examples, Determinantal Point Processes

Stronger notions of negative dependence

Strongly Rayleigh measures and real stable polynomials

Implications: Sampling

—BREAK—

2 **Theory & Applications**

Approximating partition functions

Learning a DPP (and some variants)

Applications

Perspectives and wrap-up

Negative Dependence, Stable Polynomials etc in ML

Part 2

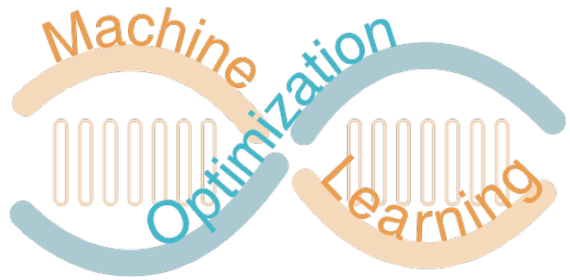
SUVRIT SRA & STEFANIE JEGELKA

Laboratory for Information and Decision Systems

Massachusetts Institute of Technology

Neural information Processing Systems, 2018

ml.mit.edu



Outline

1 Intro & Theory

Introduction

Prominent example: Determinantal Point Processes

Stronger notions of negative dependence

Implications: Sampling

2 Theory & Applications

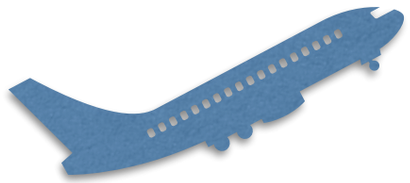
Approximating partition functions

Learning a DPP (and some variants)

Applications

Recommender systems, Nyström method, optimal design, regression, neural net pruning, negative mining, anomaly detection, etc.

Perspectives and wrap-up



Theory

Partition functions

Learning DPPs

Computing Partition functions

Aim: Estimate Z_μ, i.e., normalization const / partition function

$$\Pr(S) = \frac{1}{Z_\mu} \mu(S)$$

Typically intractable and often even hard to approximate

(exponential number of terms to sum over, or evaluation of high-dimensional integrals / volumes)

but...

Computing Partition functions

Nature makes an exception for DPPs!

$$Z_L = \sum_{S \subseteq [n]} \det(L_S) \quad = \det(I + L)$$

What about?

$$Z_\mu = \sum_{S \subseteq [n]} \mu(S) \qquad \text{(SR)}$$

$$Z_{\mu,p} = \sum_{S \subseteq [n]} \mu(S)^p \qquad \text{(ESR)}$$

Computing Partition functions

$$Z_\mu = \sum_{S \subseteq [n]} \mu(S), \qquad Z_{\mu,p} = \sum_{S \subseteq [n]} \mu(S)^p$$

Using properties of stable polynomials, these can be approximated within factor e^n (e^k for k-homogeneous, e.g., k-DPP): *[Straszak, Vishnoi, 2016; Nikolov, Singh, 2016; Anari, Gharan, Saberi, Singh, 2016; Anari, Gharan 2017]*

Key: Build on Leonid Gurvits' fundamental work (2006) on approximating permanents of nonnegative matrices using convex relaxation afforded by stable polynomials

$$\inf_{z>0} \frac{p(z_1, \ldots, z_n)}{z_1 z_2 \cdots z_n}$$

z=exp(y): yields convex optim.

(a geometric program - GP)

Example: matrix permanents

$$\operatorname{per}(A) = \sum_{\sigma \in \mathfrak{S}_n} \prod_{i=1}^{n} a_{i,\sigma(i)}$$

Eg: counts number of perfect matchings in a bipartite graph

A is doubly stochastic

Permanents via stable polynomials (Gurvits 2006)

$$\operatorname{per}(A) = \frac{\partial^n p(0)}{\partial z_1 \cdots \partial z_n}$$

$$p(z_1, \ldots, z_n) = \prod_{i=1}^{n} \Big(\sum\nolimits_{j=1}^{n} a_{ij} z_j \Big)$$

$$\frac{\partial^n p}{\partial z_1 \cdots \partial z_n} \geq \frac{n!}{n^n} \inf_{z>0} \frac{p(z_1, \ldots, z_n)}{z_1 z_2 \cdots z_n}$$

Learning

Learning a DPP from data

Aim: Learn a DPP kernel matrix from data

More generally: Learn an SR measure from data (how?)

Application: Learn from observed subsets to be able to "recommend" or perform "subset selection"

Originally studied in:

Kulesza, Taskar ICML 2011, UAI 2011
Affandi, Fox, Adams, Taskar, ICML 2014
Gillenwater, Kulesza, Fox, Taskar, NIPS 2014

MLE for learning a DPP

Given observations $Y_1, \ldots, Y_N$ (subsets of [n])

$$\max_{L \succ 0} \phi(L) := \sum_{i=1}^{N} \log \Pr(Y_i) = \sum_{i=1}^{N} \log \frac{\det(L_{Y_i})}{\det(I + L)}$$

Amazingly simple algorithm *[Mariet, Sra, 2015]*

$$L \leftarrow L + L \nabla \phi(L) L$$

Related recent work

- Asymptotic properties of MLE for DPPs: *[Brunel, Moitra, Rigollet, Urschel, 2017]*
- Learning a DPP via method of moments to achieve near optimal sample complexity: *[Urschel, Brunel, Moitra, Rigollet, ICML 2017]*

Speeding up DPP learning

Challenge: Basic $L+L\phi'(L)L$ iteration costs n^3, avoid?

k-DPP: Restrict DPP to subsets of size exactly 'k'

[Kulesza, Taskar, 2011]

LR-DPP: Write $L=VV^T$ for low-rank V (can sample size ≤ k)

[Gartrell, Paquet, Koenigstein, 2017]

Kron-DPP: Write $L = L_1 \otimes L_2$ (can sample any size)

[Mariet, Sra, 2017]

among others…

Open problems: learning

Problem 1: Learning parametrized classes of other SR measures

Problem 2: Efficiently learn a "Power-DPP", i.e., $\mu(S) = det(L_S)^p$

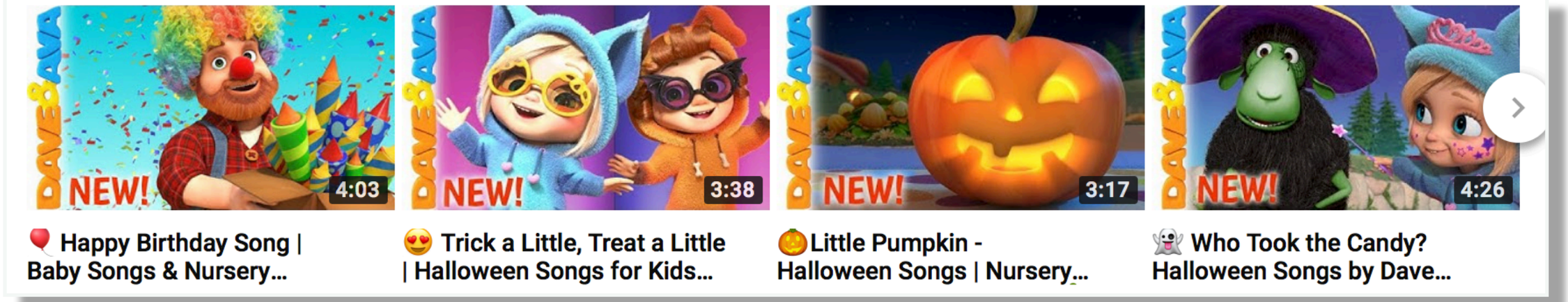
Problem 3: Learn the diversity tuning parameter 'p' in Power-DPPs and more generally in Exponentiated SR measures

Problem 4: Explore other learning models; e.g. Deep-DPP to learn nonlinear features for a DPP *[Gartrell, Dohmatob, 2018]*, or "negative mining" for reducing overfitting *[Mariet, Gartrell, Sra, 2018]*

Applications

- Recommender systems
- Model compression
- Nyström approximation
- Outlier detection
- Optimal design

Recommender systems



Practical Diversified Recommendations on YouTube with Determinantal Point Processes

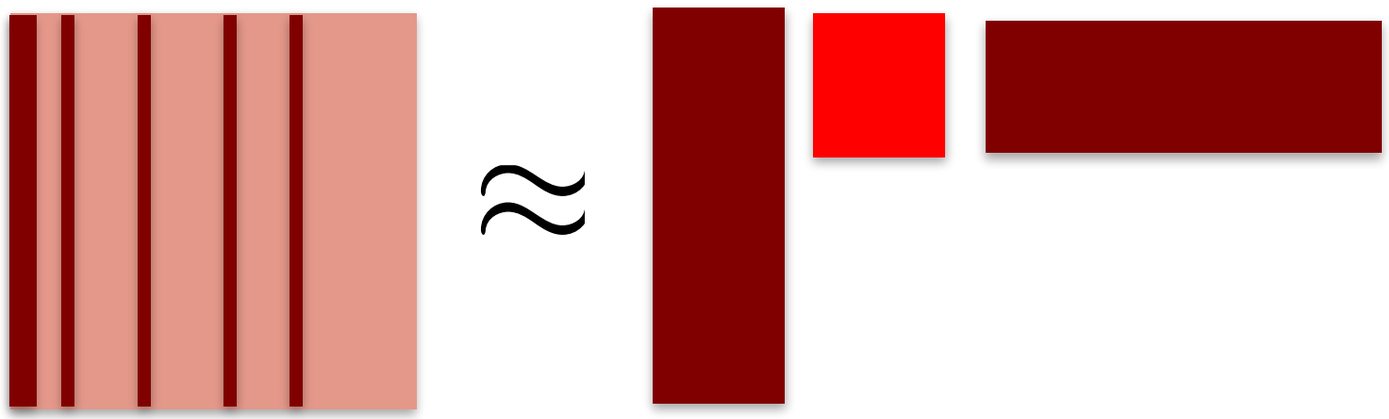
Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, Jennifer Gillenwater

Challenges: • Handling mismatch between model's notion of diversity versus user's perception of diversity (true for other applications too)
• Scalability to large-scale data
• Integrating within existing recommender ecosystems
(e.g. existing pointwise recommenders vs DPP's setwise!)

See also monograph and tutorial by A. Kulesza for more!

Nyström approximation

- Fundamental tool for scaling up kernel methods



- Which columns (data points)?

(Williams & Seeger 01, Zhang et al 08, Belabbas & Wolfe 09, Gittens & Mahoney 13, Alaoui & Mahoney 15, Deshpande et al 06, Smola & Schölkopf 00, Drineas & Mahoney 05, Drineas et al 06, …)

- Sample subset S from k-DPP

$$\widehat{K} = K_{:,S} K_{S,S}^{\dagger} K_{S,:}$$

Nyström approximation

- Sketching matrices/kernel methods

$$\widehat{K} = K_{:,S} K_{S,S}^{\dagger} K_{S,:}$$

Theorems. *(Li, Jegelka, Sra 2016)*

$$\frac{\mathbb{E}[\|K - \widehat{K}\|_F]}{\|K - K_k\|_F} \leq \frac{c+1}{c+1-k}\sqrt{N-k}$$

Approx quality
c ≥ k landmarks

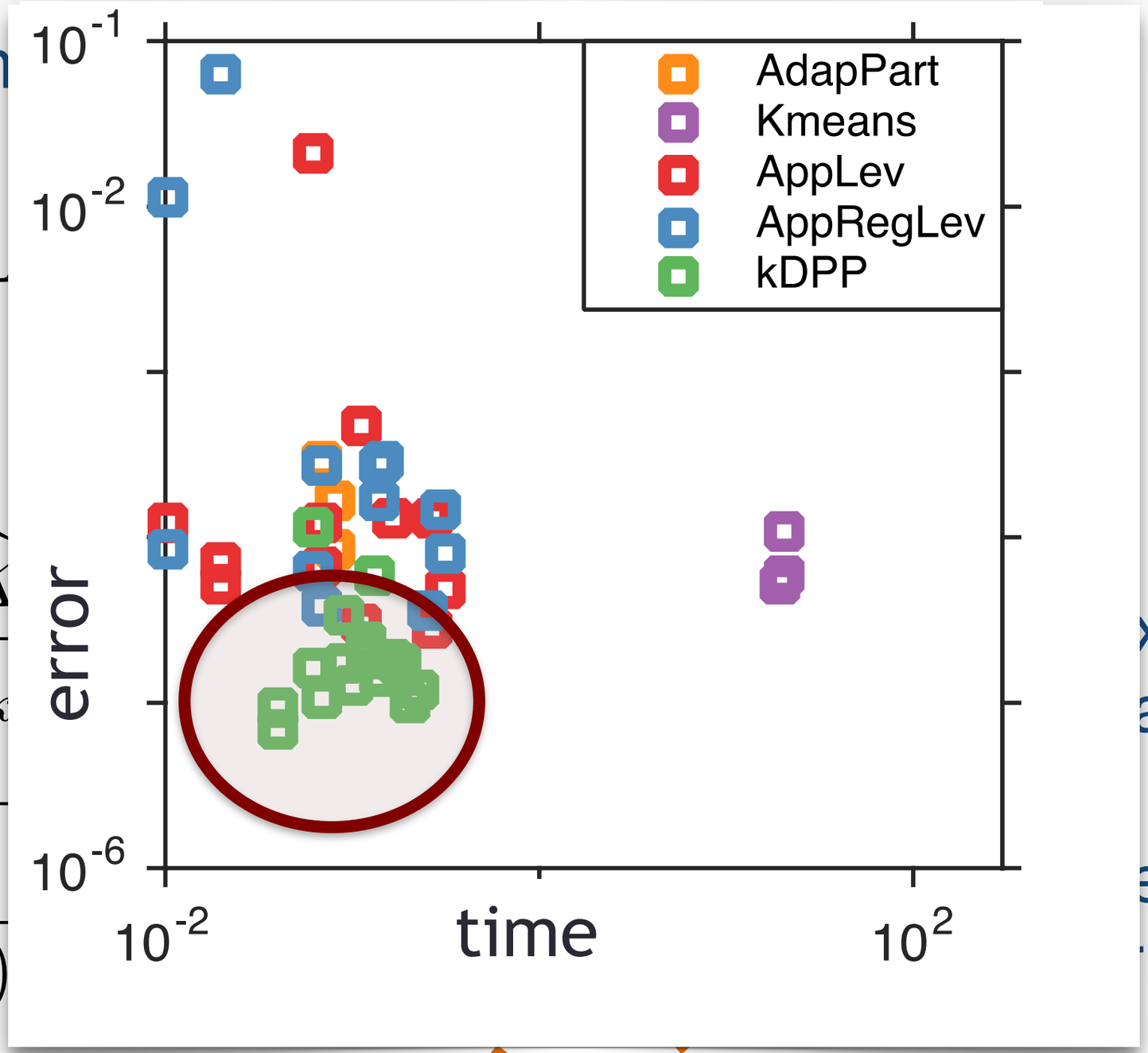
$$\mathbb{E}\sqrt{\frac{\mathcal{R}(\hat{z})}{\mathcal{R}(\hat{z}_S)}} \geq 1 - \frac{c+1}{N\gamma}\frac{e_{c+1}(K)}{e_c(K)}$$

Expected risk
kernel ridge regression

ratio of elementary symm. polynomials

Nyström approximation

- Sketch



Theorems.

$$\mathbb{E}[\|K - \hat{K}$$
$$\|K - K_k$$

x quality
andmarks

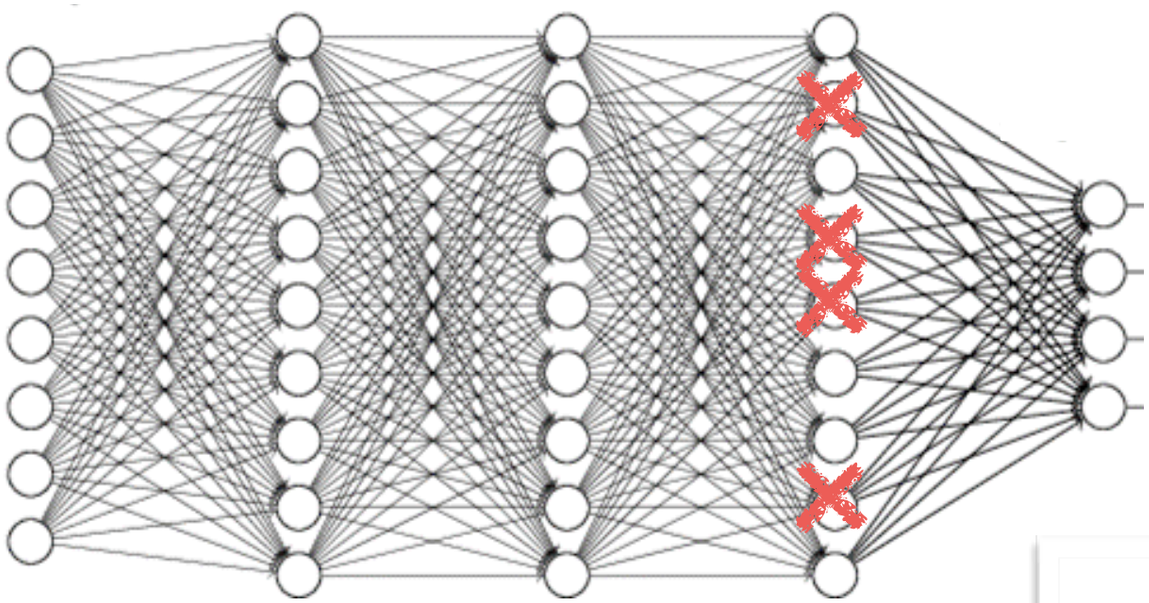
$$\mathbb{E}\sqrt{\frac{\mathcal{R}(\hat{z})}{\mathcal{R}(\hat{z}_S)}}$$

ed risk
idge regression

(Li, Jegelka, Sra 2016)

ratio of elementary symm. polynomials

Neural network pruning

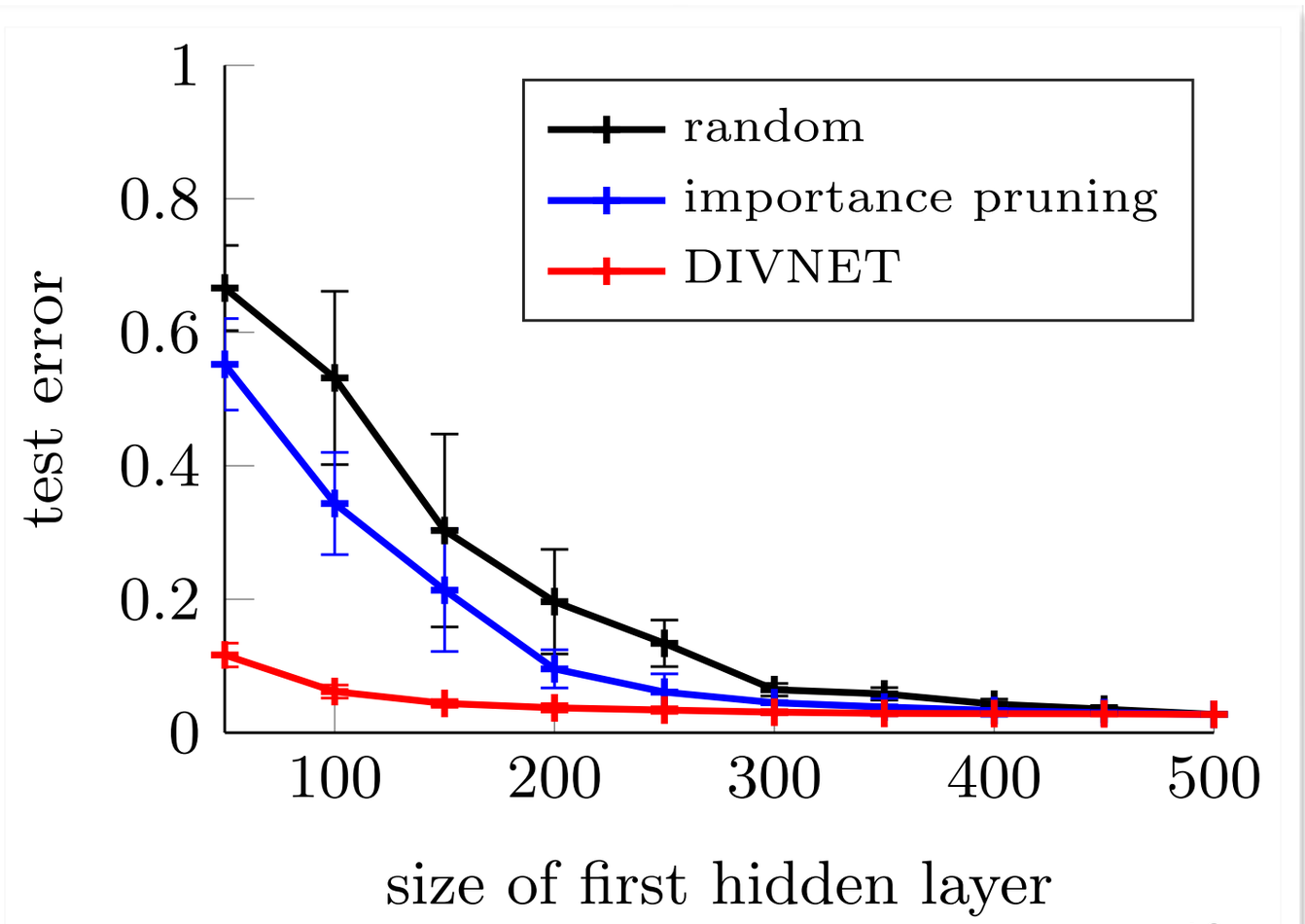


Challenge: Which measure to use for sampling?

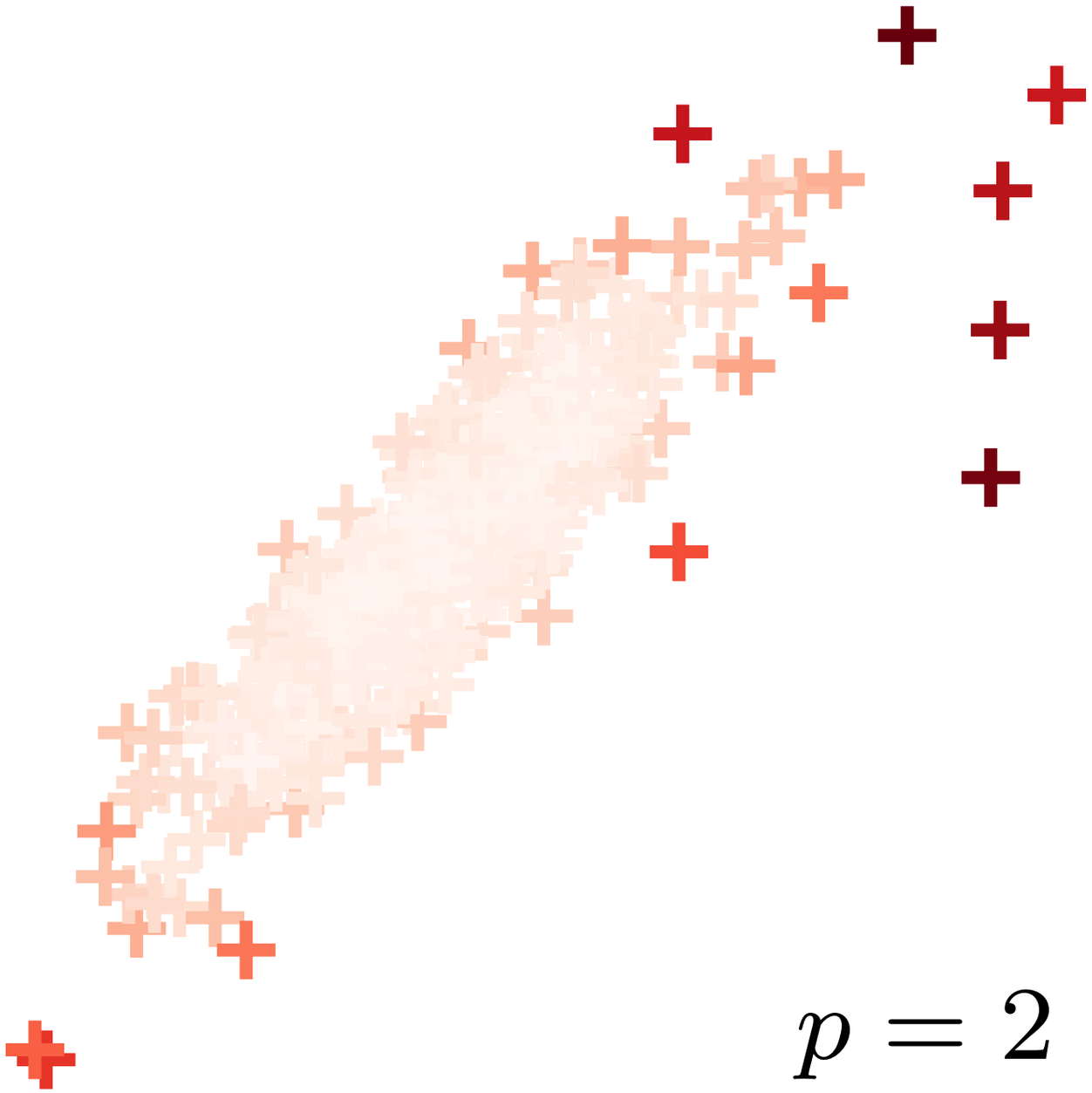
“Diversity networks”

1. Sample diverse neurons
2. Delete redundant ones
3. Rebalance layer output

(Mariet, Sra 2016)

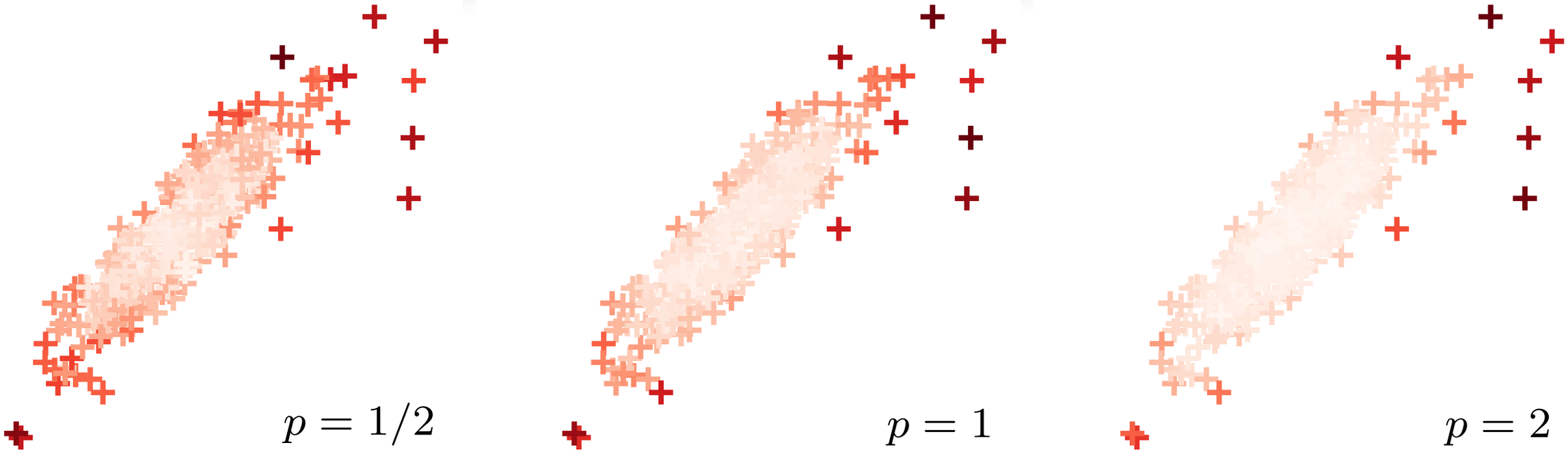


Outlier detection



$p = 2$

Outlier detection



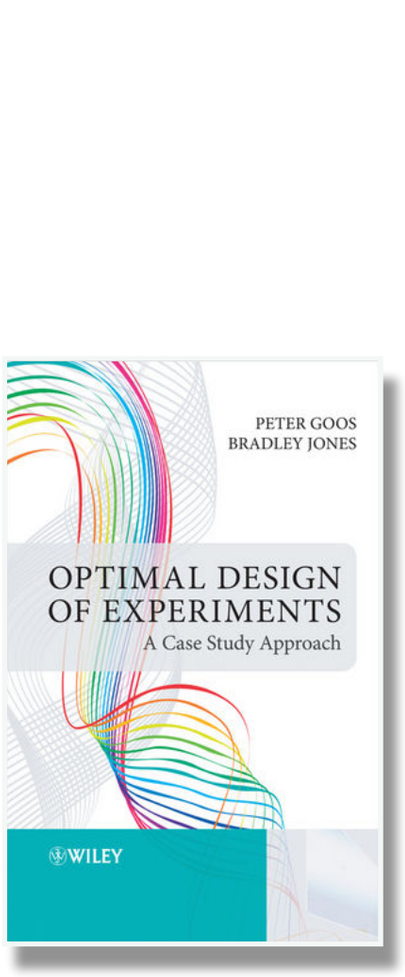
p=0 increasing sensitivity p

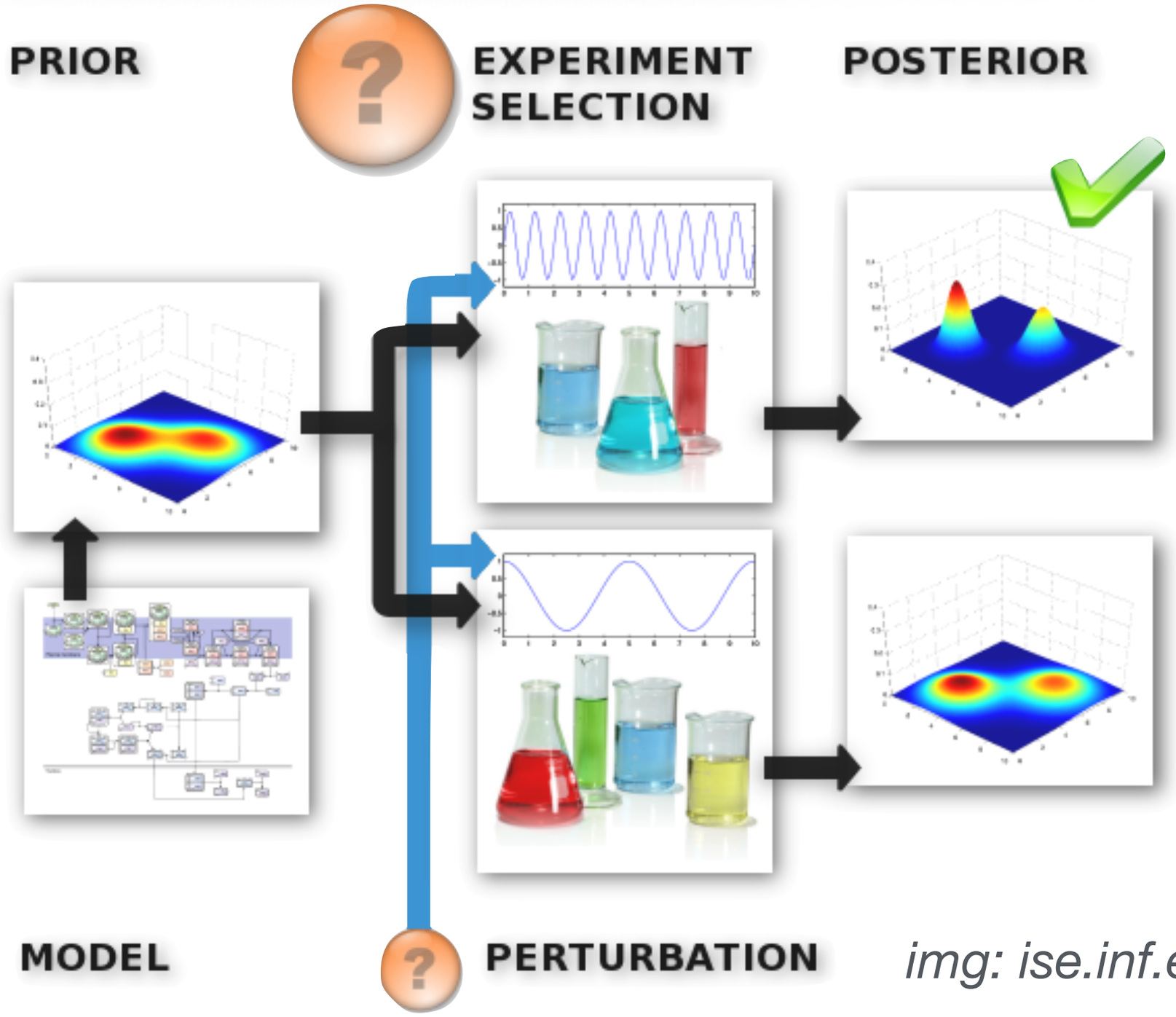
uniform distribution

$$\nu(S) = \mu(S)^p$$

[Mariet, Sra, Jegelka, 2018]

Optimal design & active learning





img: ise.inf.eth.ch

Optimal design & active learning

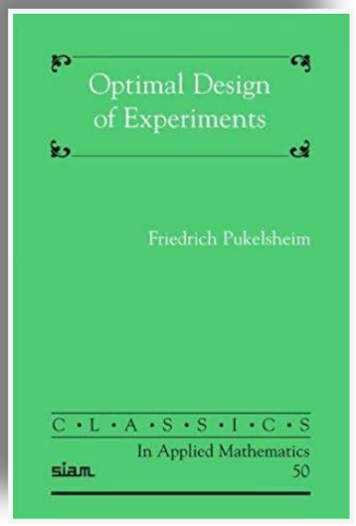
Setup: Say ‘m’ possible experiments with measurements $x_1, \ldots, x_m$, (with x_i in $\mathbb{R}^n$), and scalar outcomes $y_1, \ldots, y_m$

$$y_i = \theta^T x_i + \epsilon$$

Aim: Pick a subset S of [m] to “minimize” uncertainty

$$\min_{S \subseteq [m], |S| = k} \quad \Phi\left(\left(\sum_{i \in S} x_i x_i^T \right)^{-1} \right)$$

What is this?



Ref. Pukelsheim, *Optimal design of experiments.*

Optimal design & active learning

$$\min_{S \subseteq [m], |S|=k} \quad \Phi\left(\left(\sum_{i \in S} x_i x_i^T\right)^{-1}\right)$$

Φ=trace gives A-optimal, Φ=det gives D-optimal design

(Wang, Yu, Singh, 2016)

(Bayesian A-opt: Golovin,Krause,Ray, 2013)

(Chamon, Ribeiro, 2017)

(Chen, Hassani, Karbasi, 2018)

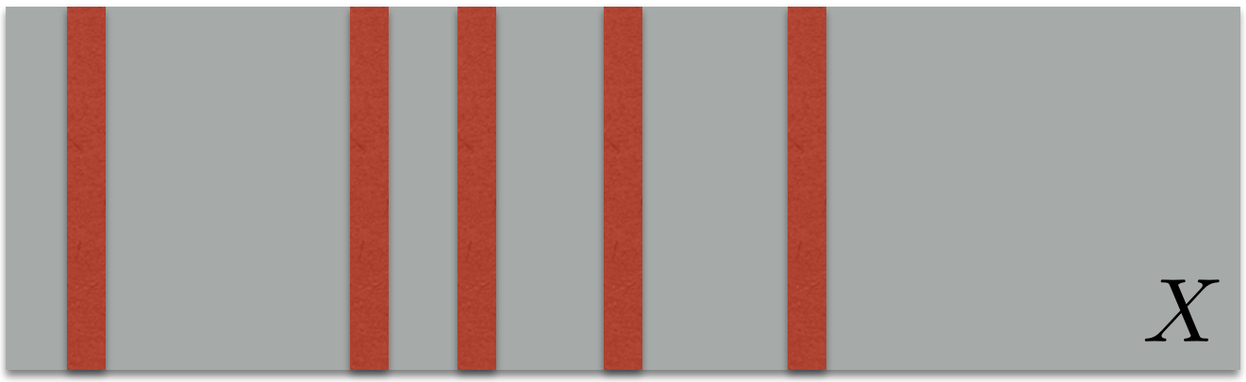
(Singh, Xie, 2018)

…and many more

(Mariet, Sra, 2017): Φ=Elemenetary Symmetric Polynomial
(recovers A- and D-optimal case extreme cases)

Thm. Greedy algo and convex relaxation both work.
Success of greedy uses "Dual" volume sampling!

“Dual” volume sampling



$$P(S) \propto \det(X_S X_S^\top)$$

NOT a DPP
…but SR

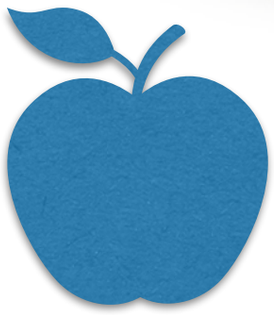
n rows, $m \gg n$ columns. Sample $k > n$ columns.

(Avron & Boutsidis 2013): approximation bounds on Frobenius norms for A-/E-optimal experimental design from sampling.

(Mariet, Sra, 2017) generalize to E-Symm. Polynomials

Note: *(Derezinski, Warmuth, 2017)* and *(Li, Jegelka, Sra, 2017)* provide efficient algorithms to sample from P(S)

Optimal design & active learning



An aside for convex optimization folks

Dual of convex relaxation to D-optimal design is the famous MVCE problem (Todd, *Minimum Volume Ellipsoids* SIAM 2016)

$$\max \ \log\det(M), \quad M \succ 0, \ \|Ma_i - z\| \leq 1, \ 1 \leq i \leq N$$

Uncovers a connection between geometry, optimization, and optimal-design (and hence stable polynomials!)

Hence, similar geometric problems via duals of convex relaxations of the Φ-optimal design problems (prev. slide)

Other ML applications

- ★ See past tutorials on submodular models in ML (various authors)
- ★ Reinforcement learning (diversity based exploration)
  https://arxiv.org/abs/1802.04564
- ★ Fairness and diversity
  https://arxiv.org/abs/1610.07183
- ★ Video Summarization
  https://arxiv.org/abs/1807.10957
- ★ Diversified minibatches for SGD
  https://arxiv.org/abs/1705.00607
- ★ Diverse sampling in Bayesian optimization
  (Kathuria, Deshpande, Kohli, 2016; Wang, Li, Jegelka, Kohli, 2017)
- ★ and of course, many more (see tutorial website for more…)

Related work at this conference

- **Derezinski, Warmuth, Hsu.** *Leveraged volume sampling for linear regression*
- **Zhang, Galley, Gao, Gan, Li, Brockett, Dolan.** *Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization* (based on MI)
- **Chen, Zhang, Zhou.** *Fast Greedy MAP Inference for Determinantal Point Process to Improve Recommendation Diversity*
- **Zhou, Wang, Bilmes.** *Diverse Ensemble Evolution: Curriculum Data-Model Marriage*
- **Hong, Shann, Su, Chang, Fu, Lee.** *Diversity-Driven Exploration Strategy for Deep Reinforcement Learning* (adds a distance based control)
- **Gillenwater, Kulesza, Vassilvitskii, Mariet.** *Maximizing Induced Cardinality Under a Determinantal Point Process*
- **Brunel.** *Learning Signed Determinantal Point Processes through the Principal Minor Assignment Problem*
- **Mariet, Sra, Jegelka.** *Exponentiated Strongly Rayleigh Distributions*
- **Djolonga, Jegelka, Krause.** *Provable Variational Inference for Constrained Log-Submodular Models*

Perspectives

Recent results!

- Strongly log-concave (SLC) polynomials — introduced by Gurvits in 2009, many properties laid out. Aim: approximate partition functions over combinatorially large sample spaces
- Properties further developed by Anari, Gharan, Vinzant *(Oct & Nov 2018)* and used to solve: Mason's conjecture and more!
- Matroid Base Exchange Walk: Fast Mixing – so in particular, the SR property is not necessary for fast mixing.
- Exponentiated SR measures *(Mariet, Sra, Jegelka, 2018)*, with an approximate mixing time analysis and few applications
- The ESR case $0 < \alpha < 1$ falls under the SLC framework, hence fast MCMC sampling *(Anari, Liu, Gharan, Vinzant, Nov 2018)*

Summary and outlook

We saw:

- Negative dependence as a paradigm in ML
- Foundations of strong ND = Strongly Rayleigh
- Connections to real stable polynomials
- Fast MCMC sampling
- Fast approx of partition functions
- Many applications

Outlook:

- Deeper connections to optimization
- Modeling diversity (semi-supervised)
- Richer theory of ND sampling
- Proving stability of numerous polys still wide-open
- Additional applications: from active to interactive
- Mixing positive and negative dependence

Thanks

Chengtao Li

Zelda Mariet

Thanks