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

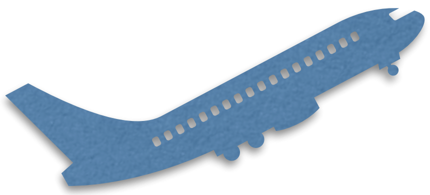
Outline

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

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

Eg: counts number of perfect matchings in a bipartite graph

A is doubly stochastic

Permanents via stable polynomials (Gurvits 2006)

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

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

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

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., $\mu(S)=det(L_S)^p$

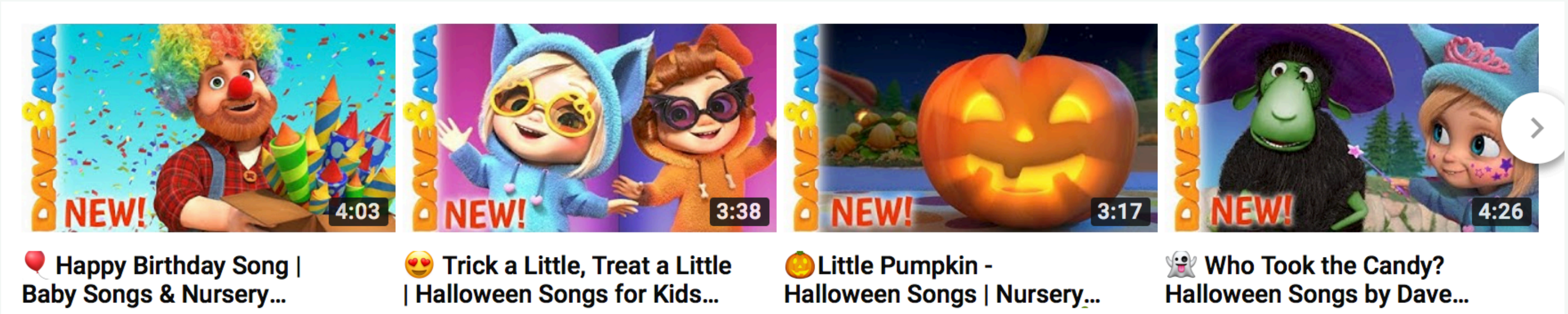
Problem 3: Learn the diversity tuning parameter 'p' in Power-DPPs and more generally in Exponentiated SR measures

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

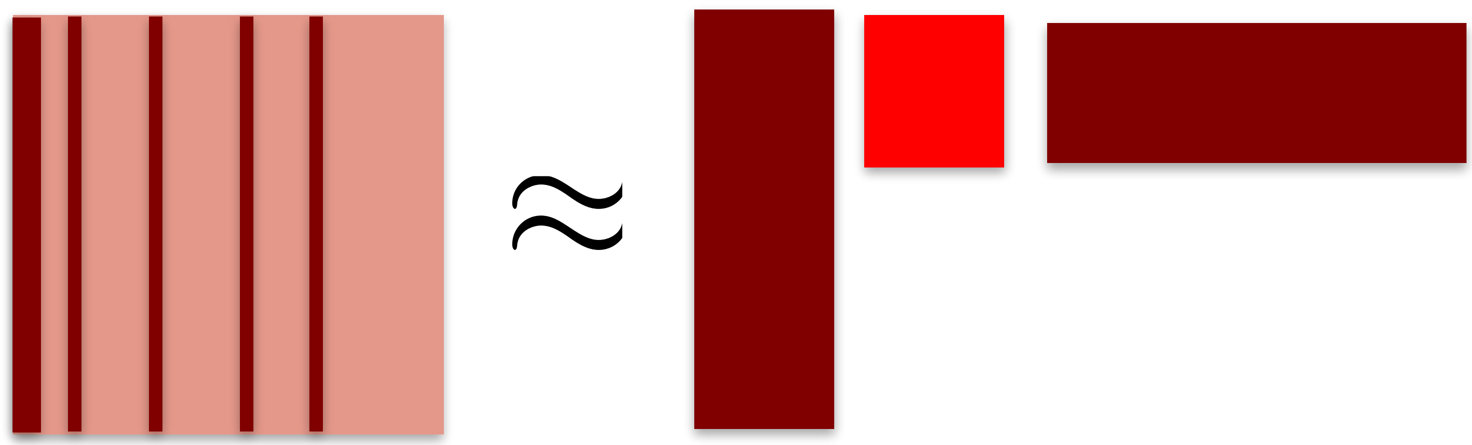
Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, Jennifer Gillenwater

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)*

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

Approx quality
c ≥ k landmarks

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

Expected risk
kernel ridge regression

ratio of elementary symm. polynomials

Nyström approximation

- Sketch

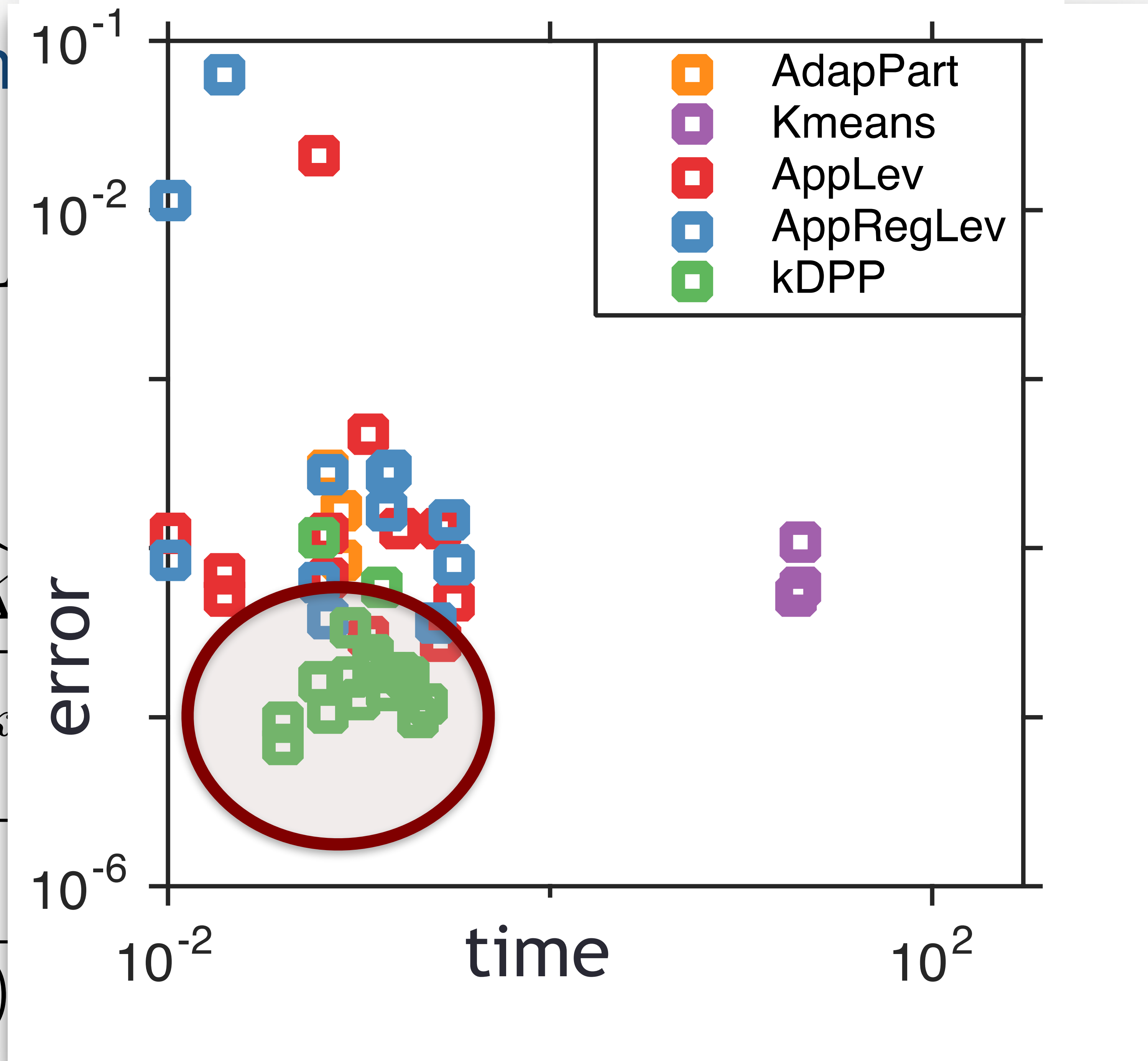
Theorems.

$$\frac{\mathbb{E}[\|K - \widehat{K}}{\|K - K_k}$$

x quality
andmarks

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

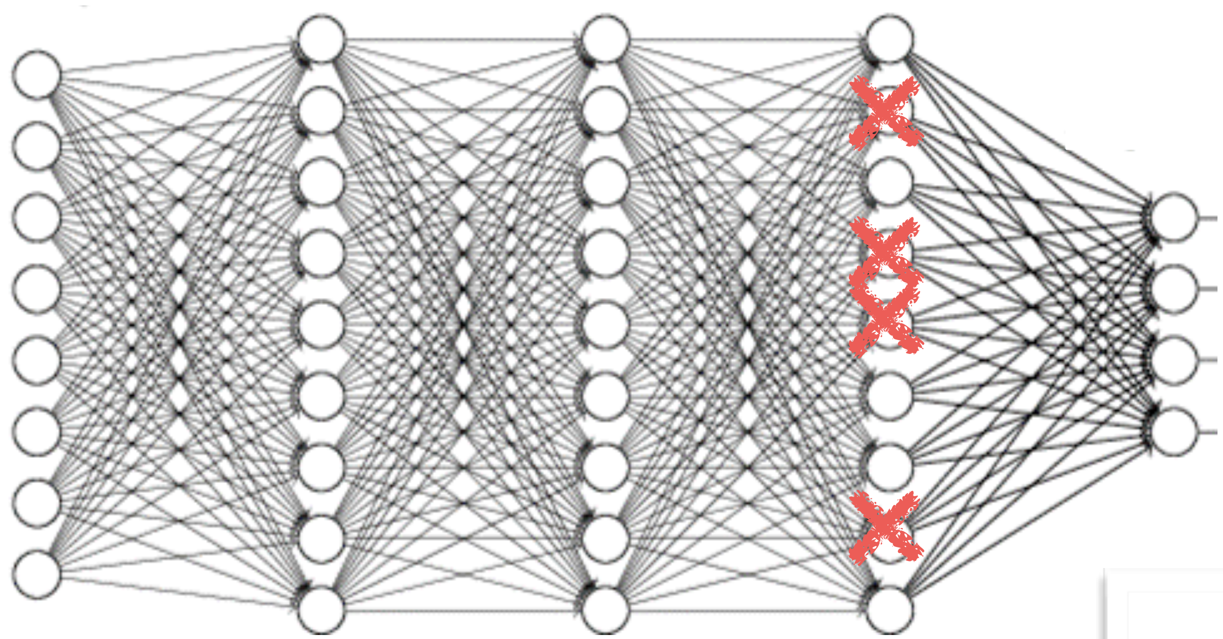
ed risk
idge regression



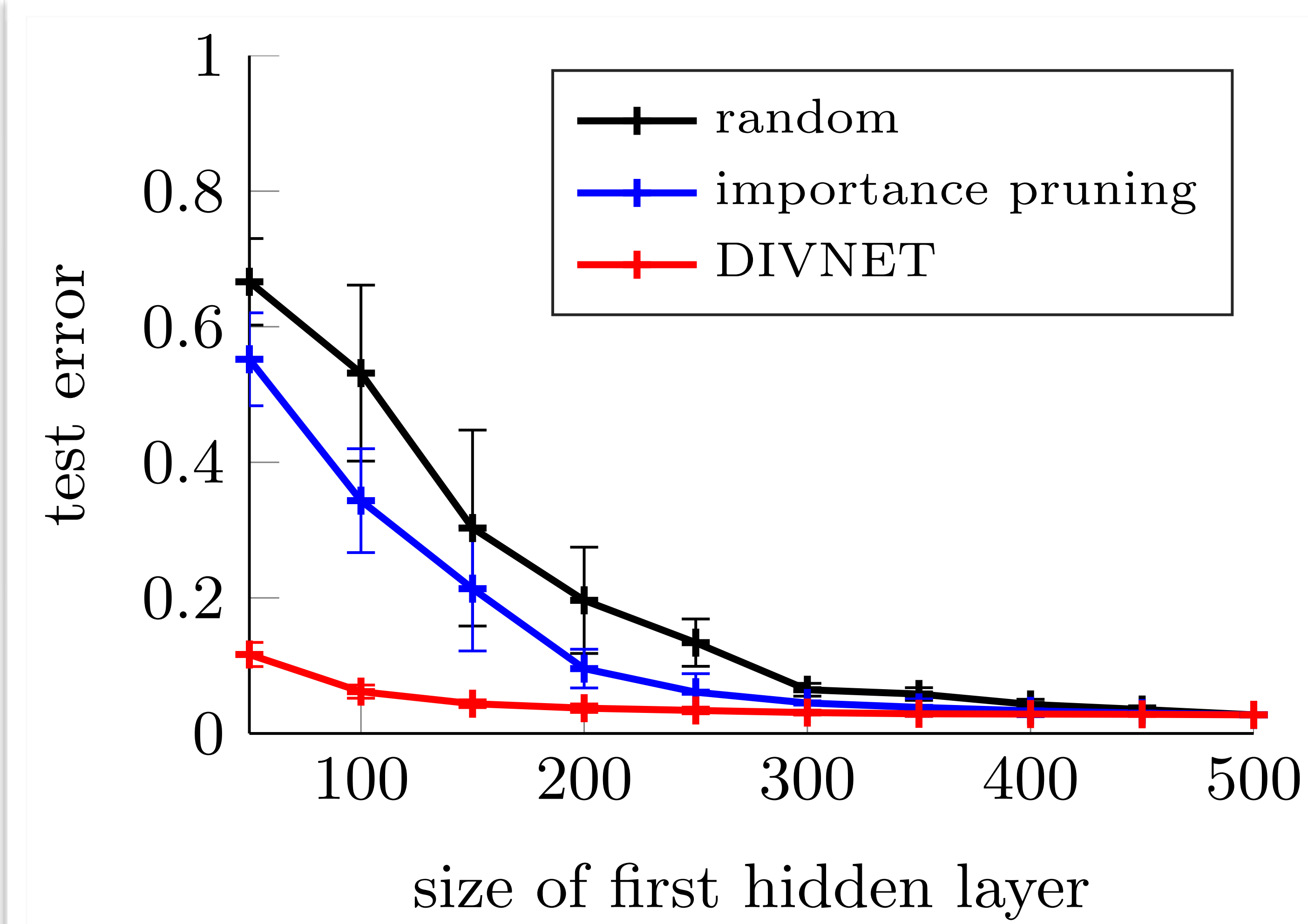
ratio of elementary symm. polynomials

(Li, Jegelka, Sra 2016)

Neural network pruning



Challenge: Which measure to use for sampling?

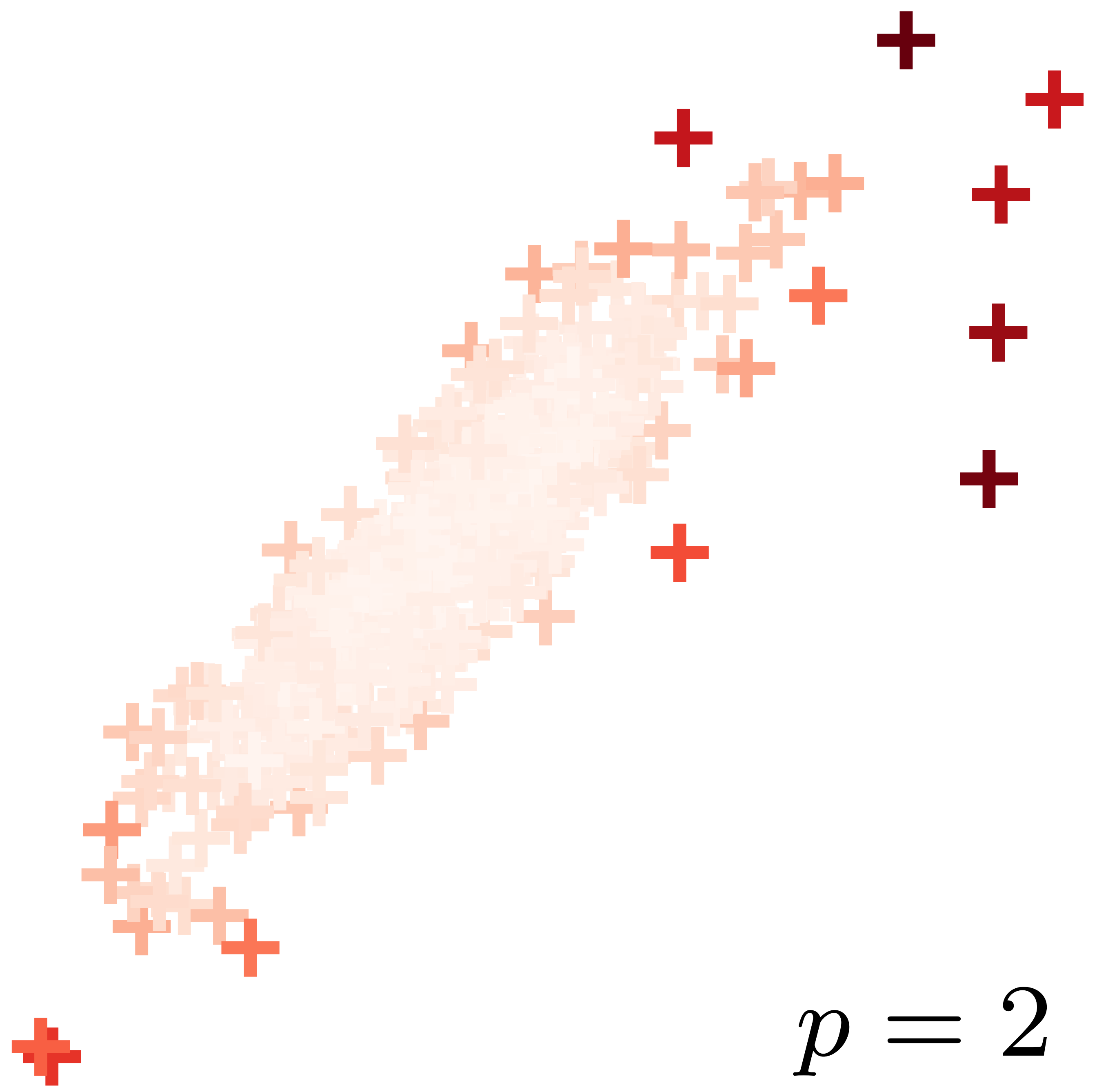


"Diversity networks"

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

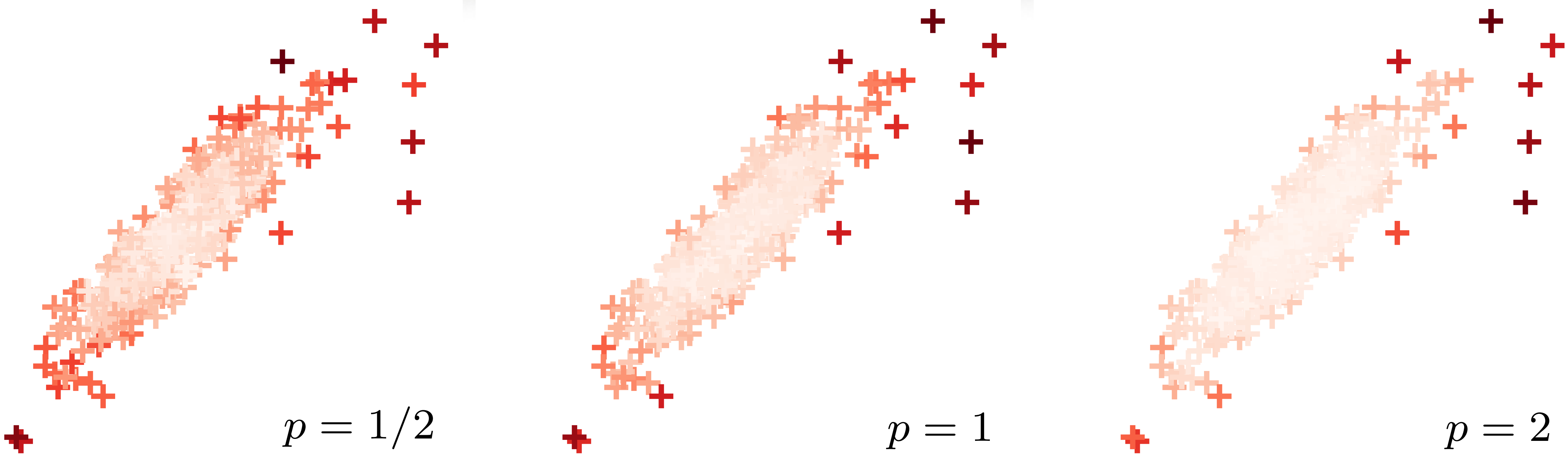
(Mariet, Sra 2016)

Outlier detection



$$p = 2$$

Outlier detection



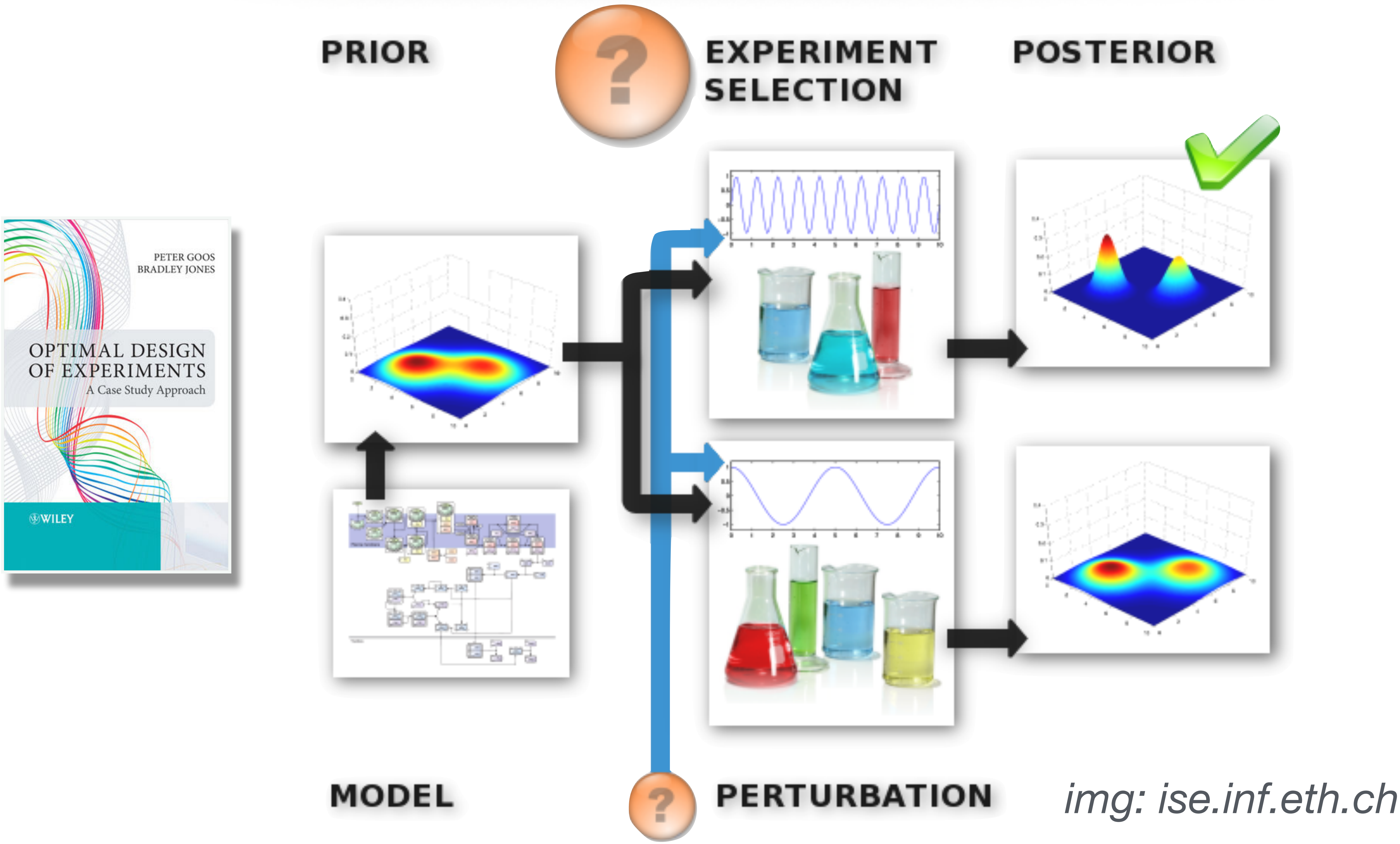
p=0 increasing sensitivity p

uniform distribution

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

[Mariet, Sra, Jegelka, 2018]

Optimal design & active learning



img: ise.inf.eth.ch

Optimal design & active learning

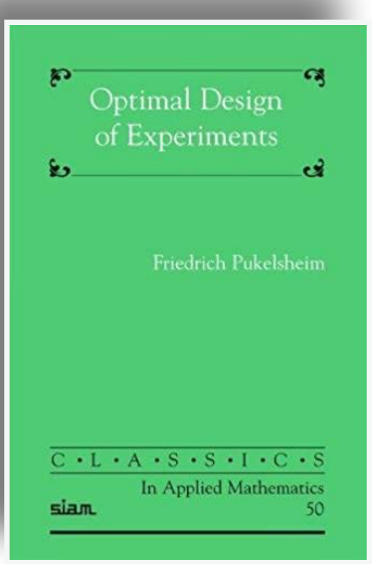
Setup: Say ‘m’ possible experiments with measurements $x_1, \ldots, x_m$, (with x_i in $\mathbb{R}^n$), and scalar outcomes $y_1, \ldots, y_m$

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

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

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

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

(Wang, Yu, Singh, 2016)

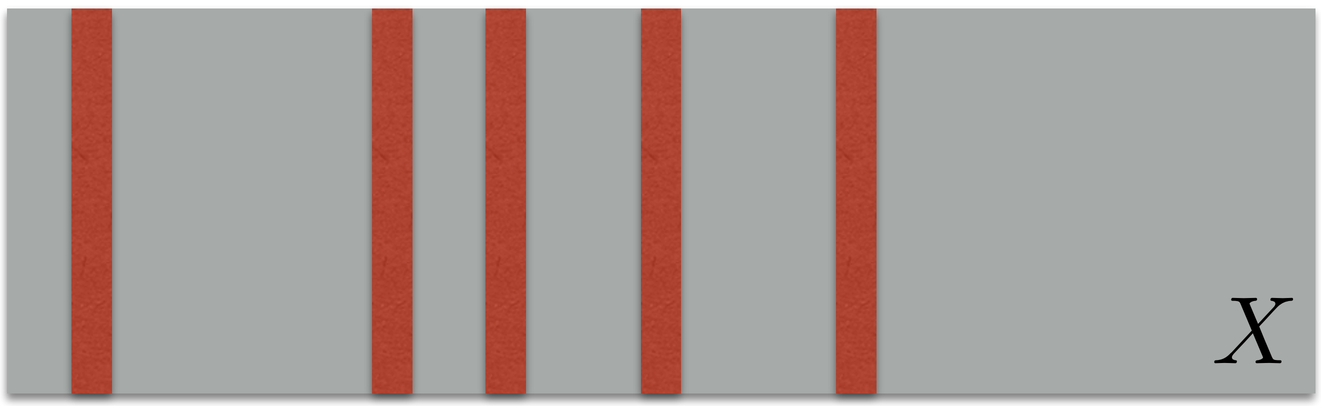
(Bayesian A-opt: Golovin,Krause,Ray, 2013)

(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

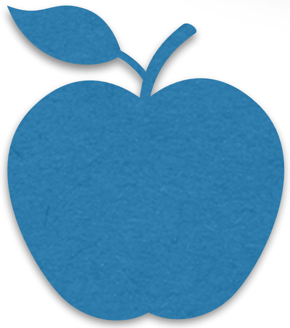
n rows, $m \gg n$ columns. Sample $k > n$ columns.

(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