Learning with Generalized Negative Dependence

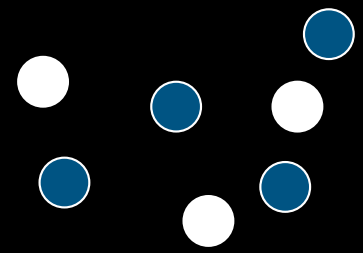
Probabilistic Models of Diversity for Machine Learning

Zelda Mariet

MIT, April 24th, 2019

What is negative dependence?

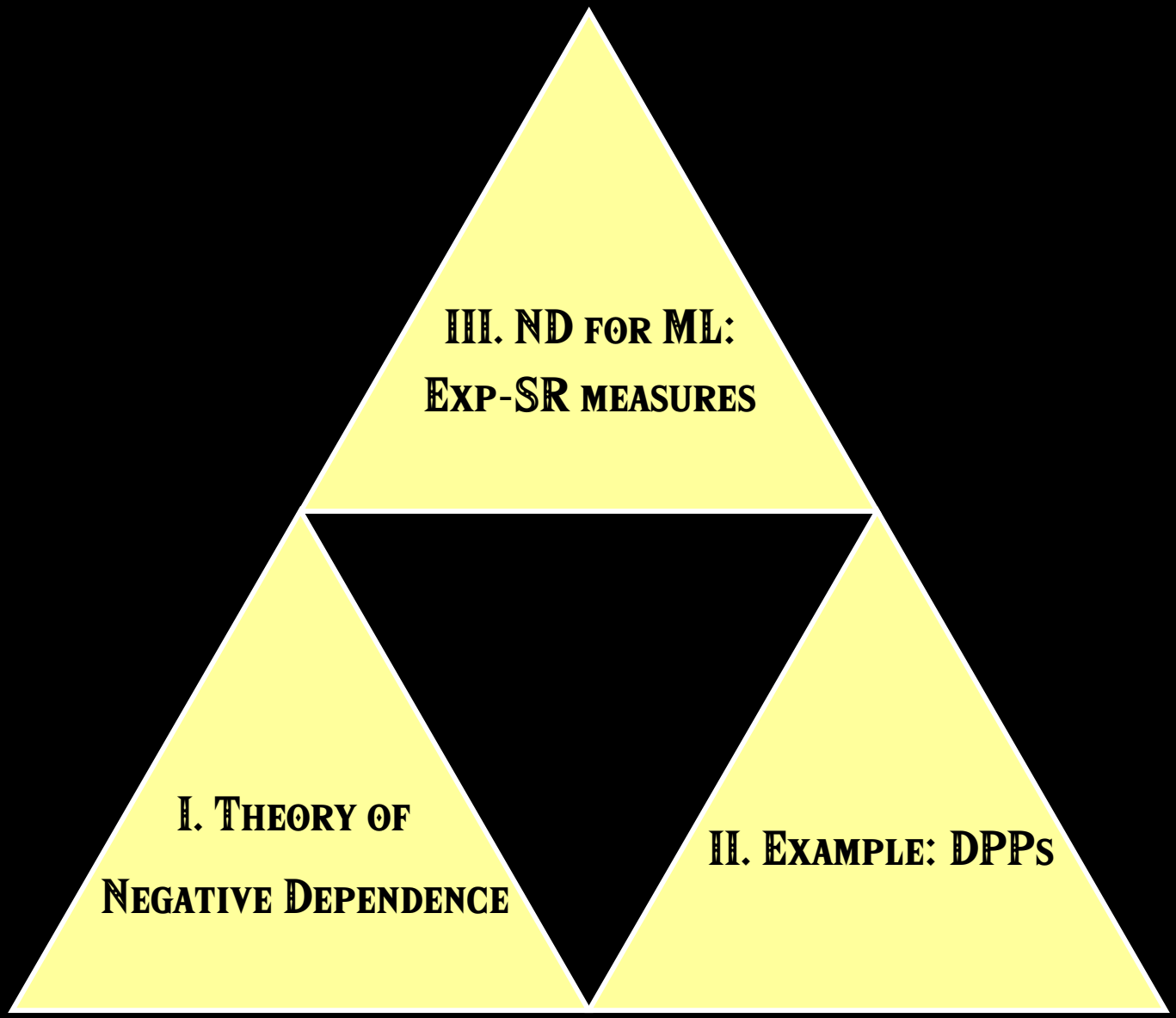
High level: similar items are unlikely to co-occur

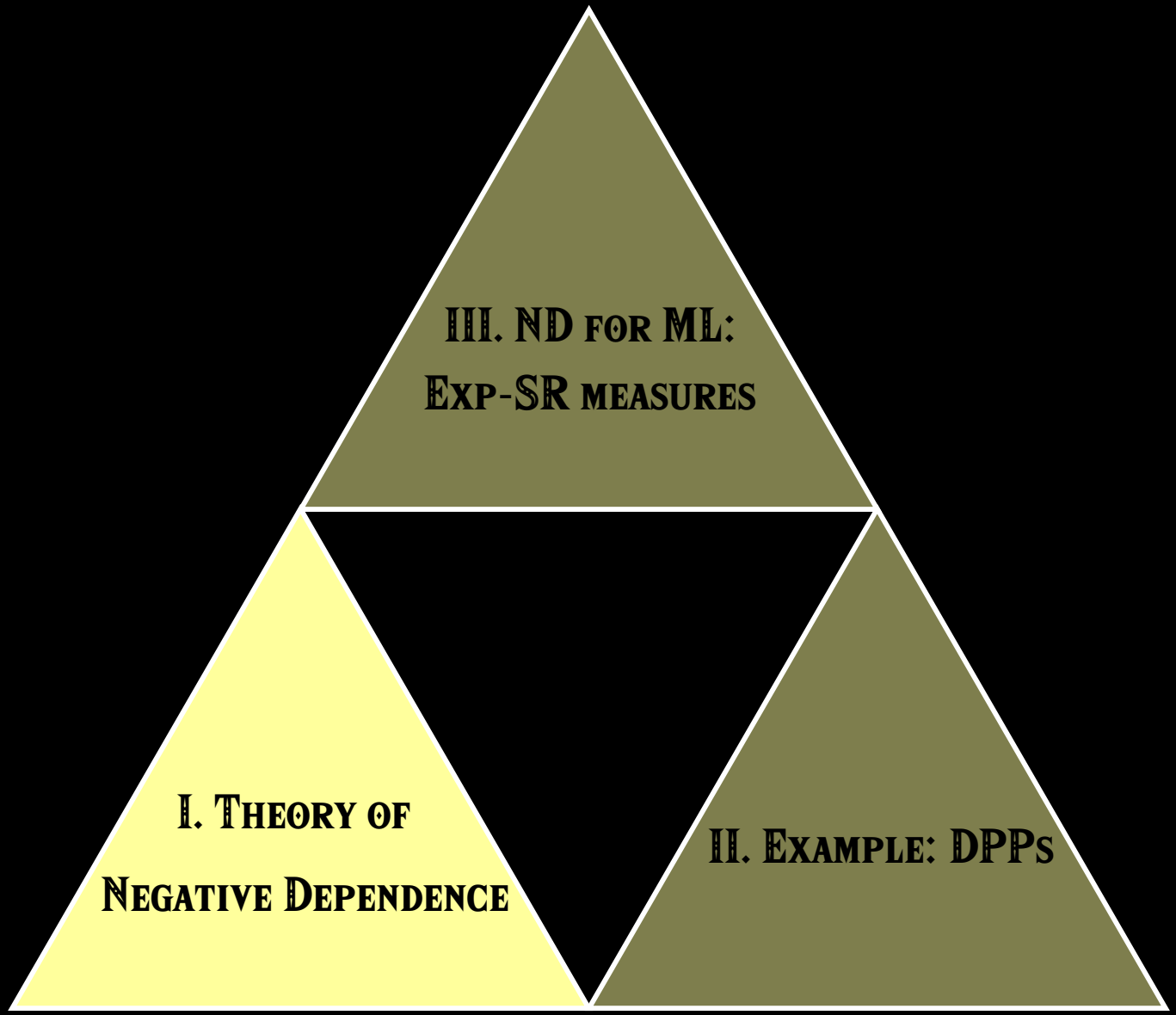


- Moviegoers: groups prefer to sit far from other groups
- Fermions: at most one fermion per quantum state
- Neurons: inhibitive interactions between firing neurons

In machine learning:

- Summarizing a document
- Recommender systems


III. ND for ML:
Exp-SR measures
I. Theory of
Negative Dependence
II. Example: DPPs


III. ND for ML:
Exp-SR measures
I. Theory of
Negative Dependence
II. Example: DPPs

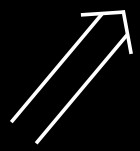
Defining negative dependence

Measure μ over subsets of $\{1, \ldots, n\}$.

Pairwise negative correlation

$$\mu(i \in S)\mu(j \in S) \geq \mu(i, j \in S)$$

$$\mathbb{E}_\mu[X_i]\mathbb{E}_\mu[X_j] \geq \mathbb{E}_\mu[X_i X_j]$$



Negative association

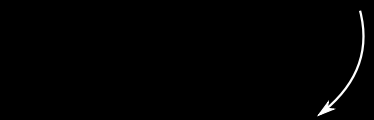
$$\mathbb{E}_\mu[f]\mathbb{E}_\mu[g] \geq \mathbb{E}_\mu[fg]$$

$f(S), g(S)$ increasing w/ disjoint supports

Log-submodularity

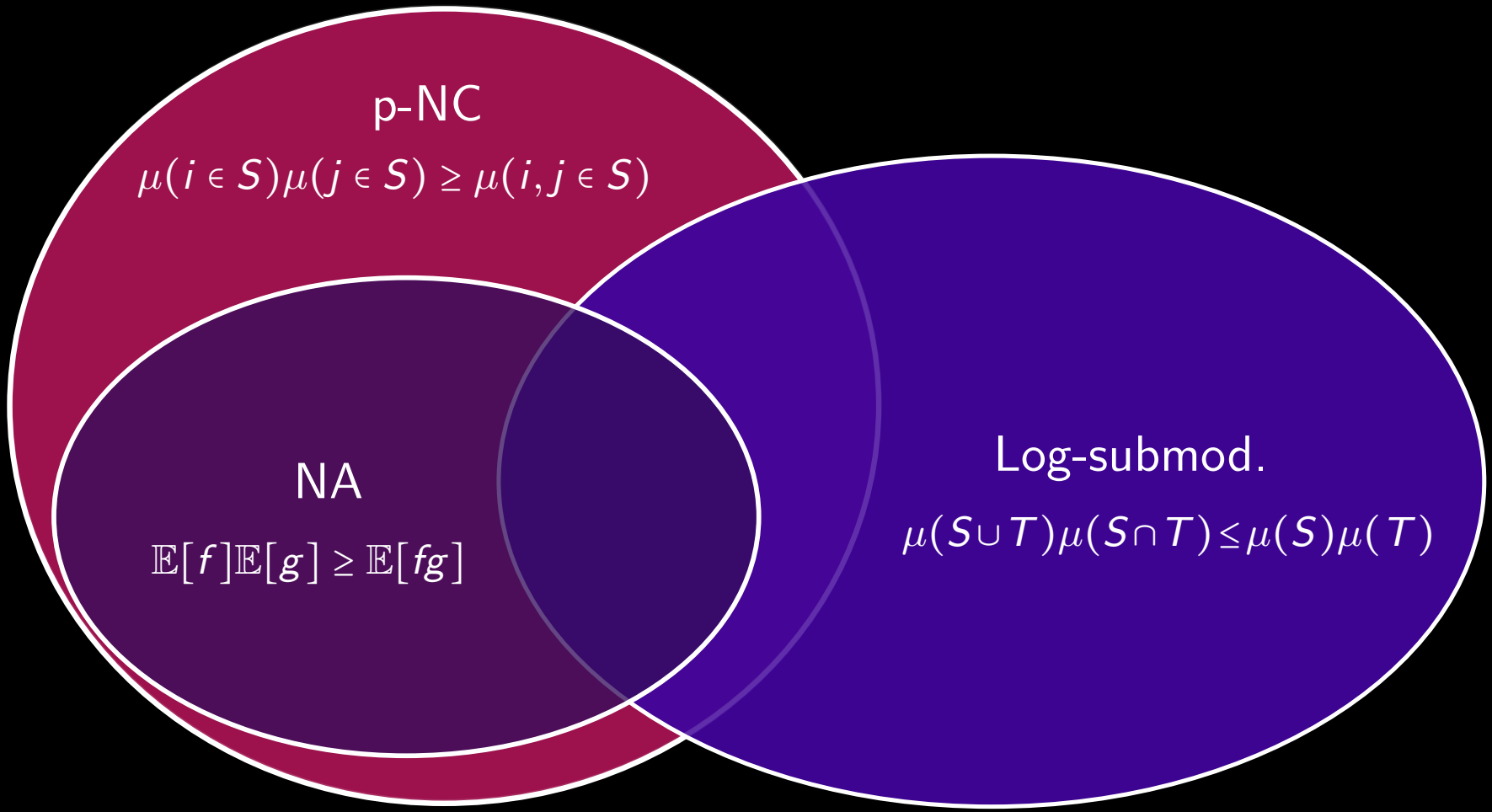
$$\mu(S \cup T)\mu(S \cap T) \leq \mu(S)\mu(T)$$

Nice optimization properties

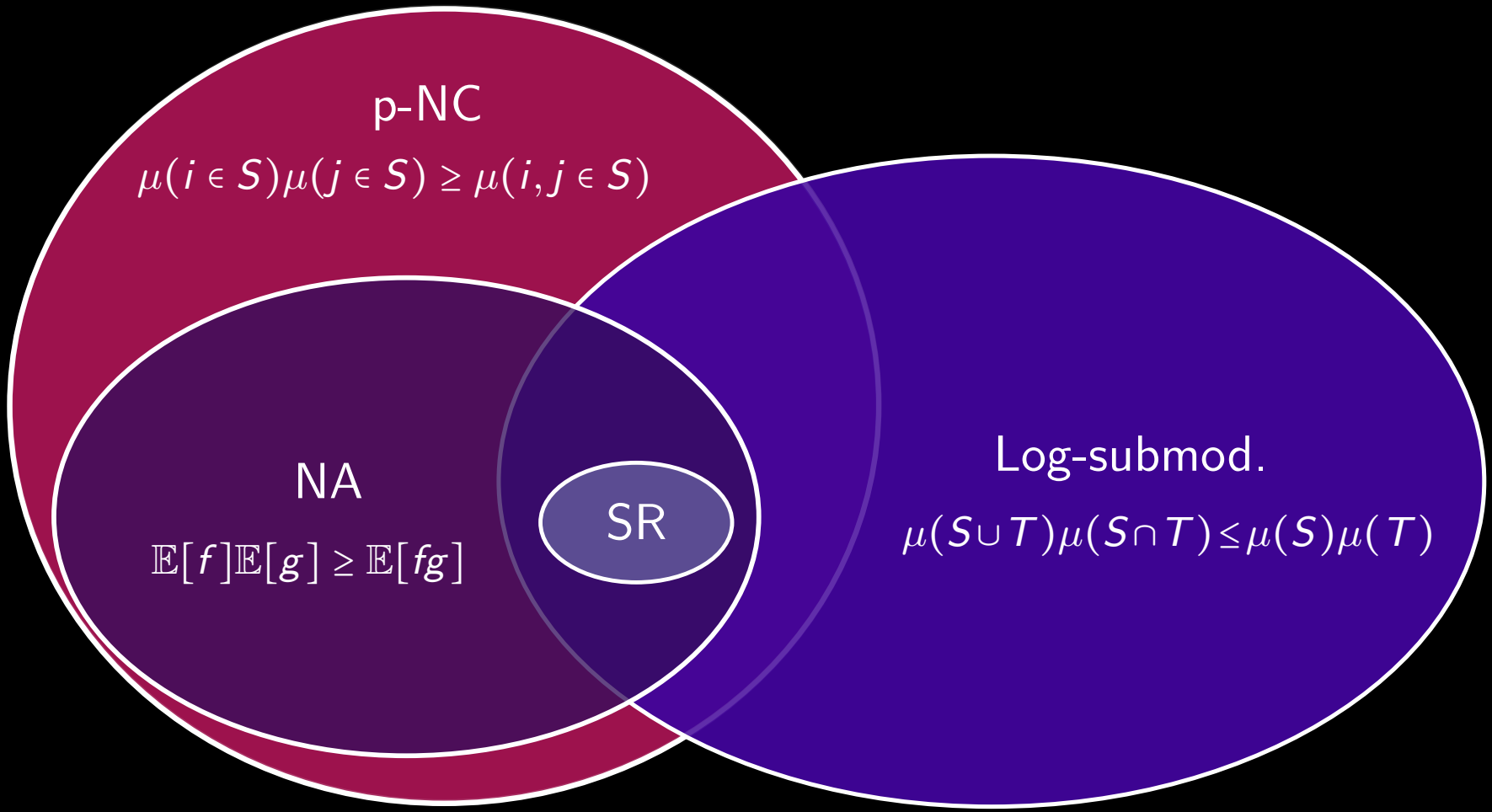


Fast sampling

Defining negative dependence



Defining negative dependence



Borcea et al. (2009): Measure μ is *Strongly Rayleigh* if

$$\forall i \in [n], \mathfrak{Im}(z_i) > 0 \implies g_\mu(z) = \sum_{S \subseteq [n]} \mu(S) \prod_{i \in S} z_i \neq 0$$

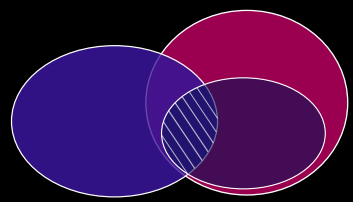
Theorem (Borcea et al. [2009])

μ is Strongly Rayleigh $\iff$ its generating polynomial g_μ verifies

$$\forall i \neq j \quad \partial_i g_\mu \partial_j g_\mu \geq g \partial_{ij} g_\mu$$

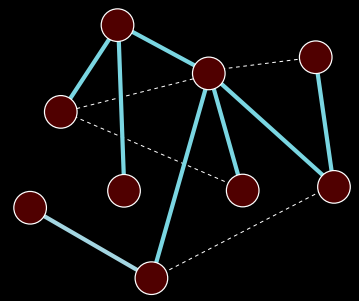
Theorem (Borcea et al. [2009])

SR measures are also negatively associated, log-submodular, pairwise negatively dependent.



Strongly Rayleigh measures in the wild

- Product measure $\Pr_\mu(i \in S) = q_i$

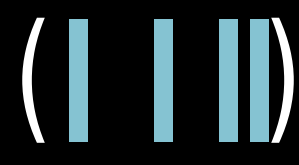


- Distribution of empty urns

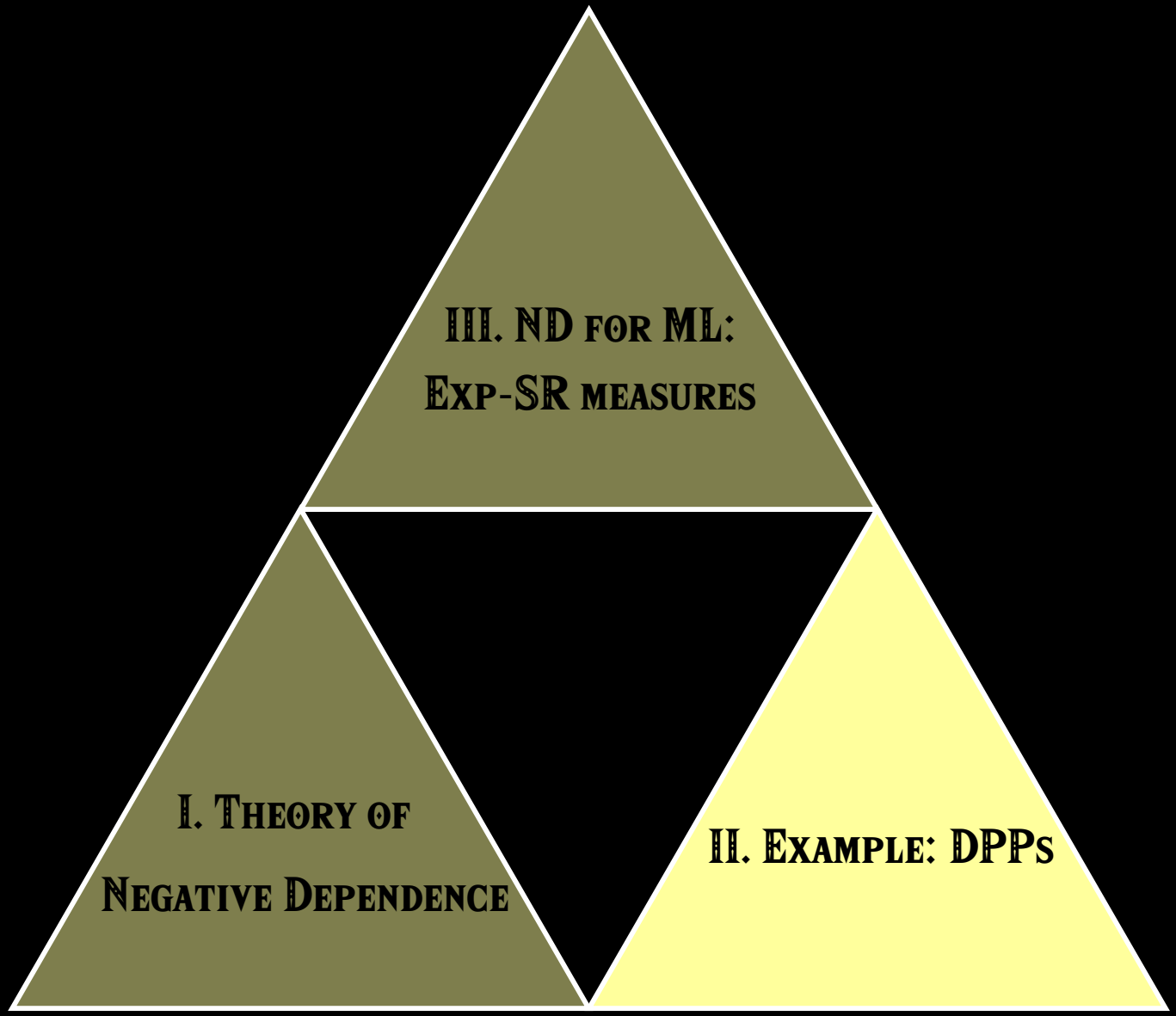
- Uniform random spanning tree measure

- Volume sampling $\mu(S) \propto e_k(X_{S,:}^\top X_{S,:})$

→ Experimental design [Mariet and Sra, NIPS'17]



- Determinantal point processes $\mu(S) \propto \det(L_S)$


III. ND for ML:
Exp-SR measures
I. Theory of
Negative Dependence
II. Example: DPPs

Example: Determinantal Point Processes

A measure μ over $2^{[n]}$ is a DPP if there exists $L \succeq 0 \in \mathbb{R}^{n \times n}$ s.t.

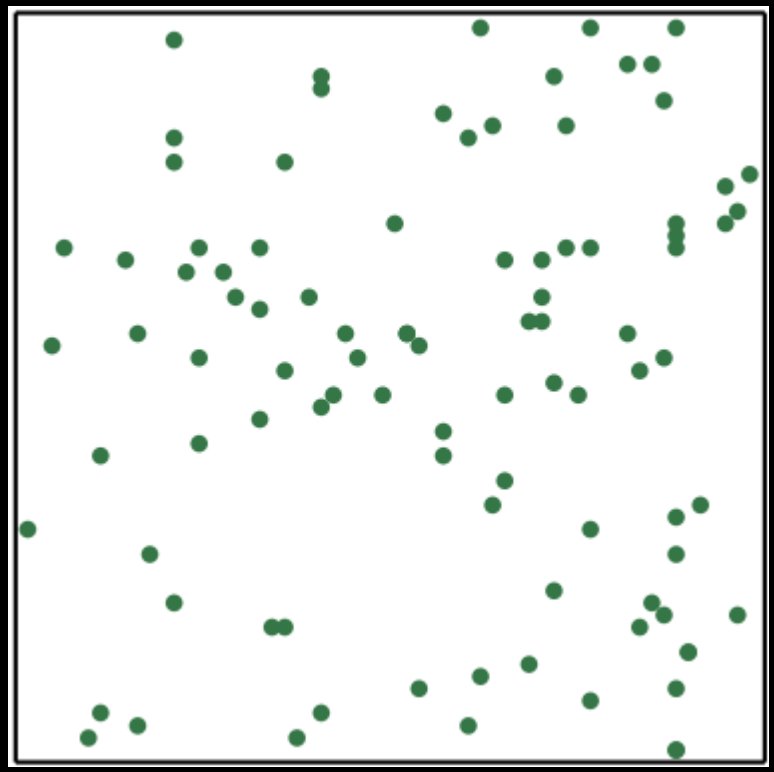
$$\forall S \subseteq [n], \quad \mu(S) \propto \det(L_S)$$

$$S = \{2, 3, 4\}, \quad L = \begin{pmatrix} 6 & 3 & 2 & 1 & \cdots & 4 \\ 3 & 3 & 0 & 2 & \cdots & 1 \\ 2 & 0 & 6 & 1 & \cdots & 7 \\ 1 & 2 & 1 & 5 & \cdots & 5 \\ \vdots & & & & \ddots & \vdots \\ 4 & 0 & 0 & 1 & \cdots & 9 \end{pmatrix}$$

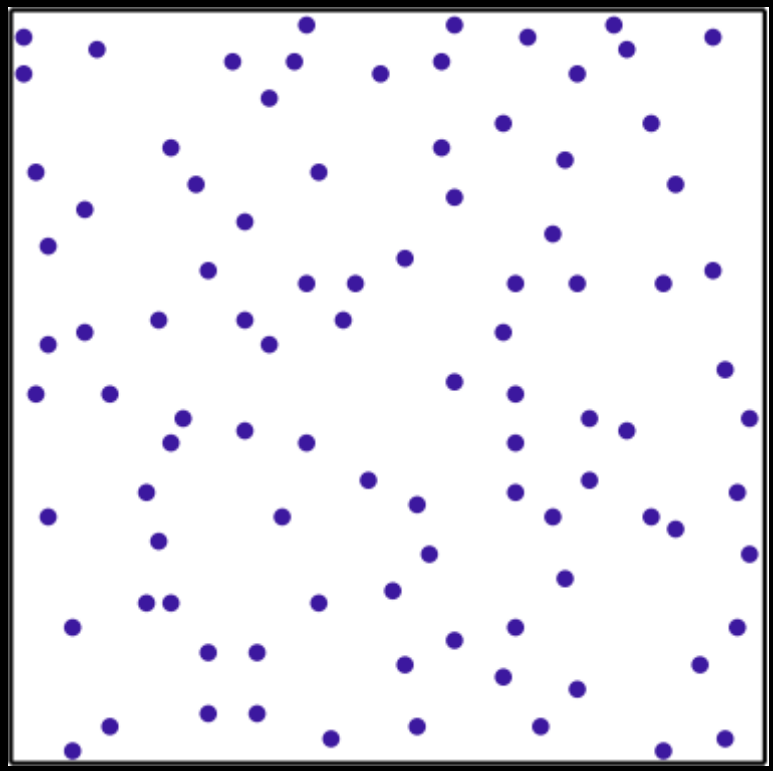
DPPs are strongly Rayleigh.

DPPs on the unit square

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



Sampling uniformly

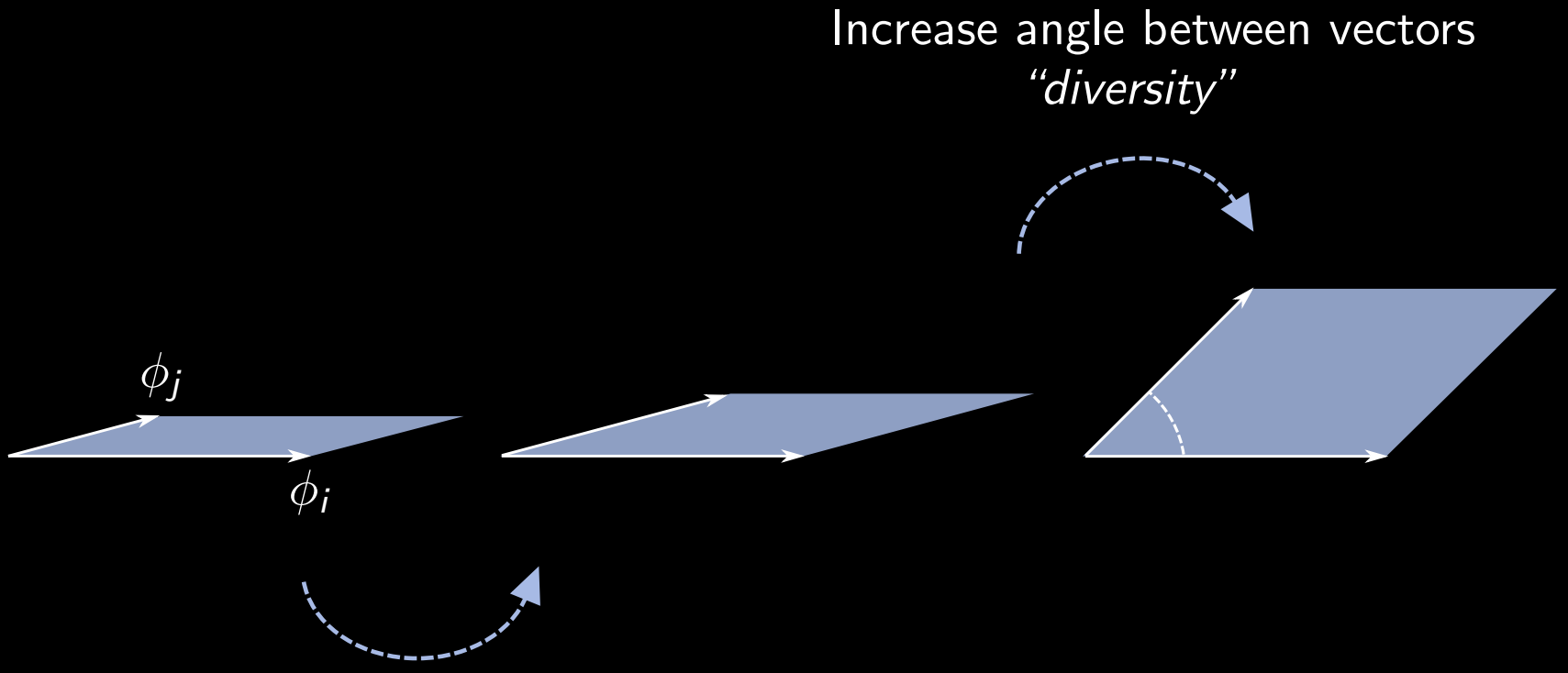


Sampling with a DPP

Alex Kulesza and Ben Taskar, 2012

Interpretable: volume-based interpretation, if $L = \Phi\Phi^\top$,

$$\mu(S) \propto \det L_S = \mathrm{Vol}(\mathrm{span}(\phi_{i_1}, \ldots, \phi_{i_k}))^2$$



Summary of contributions for DPPs

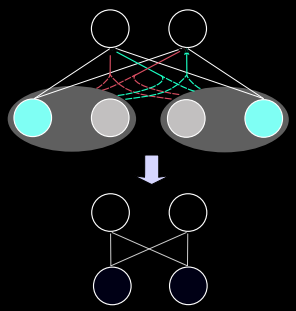
Learning — L?

- ⋆ Fixed point method [Mariet & Sra, ICML'15]
- Embedding negatives [Mariet, Gartrell, Sra, AISTATS'19]
- Kronecker structures [Mariet & Sra, NIPS'16]

Sampling — $\mathcal{O}(n^3)$

- Kronecker structures [Mariet & Sra, NIPS'16]
- Tree-based sublinear sampling [Gillenwater, Kulesza, Mariet, Vassilvitskii, ICML'19]
- Generative networks [Mariet, Ovadia, Snoek, draft]

Applications



- ⋆ Neural network pruning [Mariet & Sra, ICLR'16]
- Recommender systems [Mariet, Gartrell, Sra, AISTATS'19] [Gillenwater, Kulesza, Mariet, Vassilvitskii, NeurIPS'18]
- Kernel reconstruction [Mariet, Ovadia, Snoek, draft]

⋆ Master's thesis

Summary of contributions for DPPs

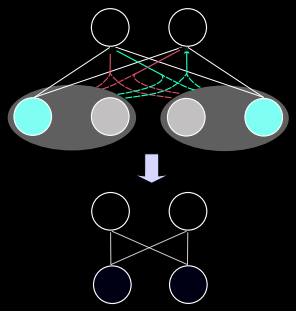
Learning — L?

- ⋆ Fixed point method [Mariet & Sra, ICML'15]
- Embedding negatives [Mariet, Gartrell, Sra, AISTATS'19]
- Kronecker structures [Mariet & Sra, NIPS'16]

Sampling — $\mathcal{O}(n^3)$

- Kronecker structures [Mariet & Sra, NIPS'16]
- Tree-based sublinear sampling [Gillenwater, Kulesza, Mariet, Vassilvitskii, ICML'19]
- Generative networks [Mariet, Ovadia, Snoek, draft]

Applications



- ⋆ Neural network pruning [Mariet & Sra, ICLR'16]
- Recommender systems [Mariet, Gartrell, Sra, AISTATS'19] [Gillenwater, Kulesza, Mariet, Vassilvitskii, NeurIPS'18]
- Kernel reconstruction [Mariet, Ovadia, Snoek, draft]

⋆ Master's thesis

Applications of DPPs in machine learning

- Neural network pruning [Mariet and Sra, 2016]
- Kernel reconstruction [Li et al., 2016a]
- Minibatch selection [Zhang et al., 2017]
- Fairness [Celis et al., 2018]

Bigger datasets + GPU/TPU operations
$\rightarrow$ using DPPs is a bottleneck ☹

DppNet: Motivation

Need: Fast, parallelizable sampling for varying ground sets
$\rightarrow$ generative network for DPP-like samples

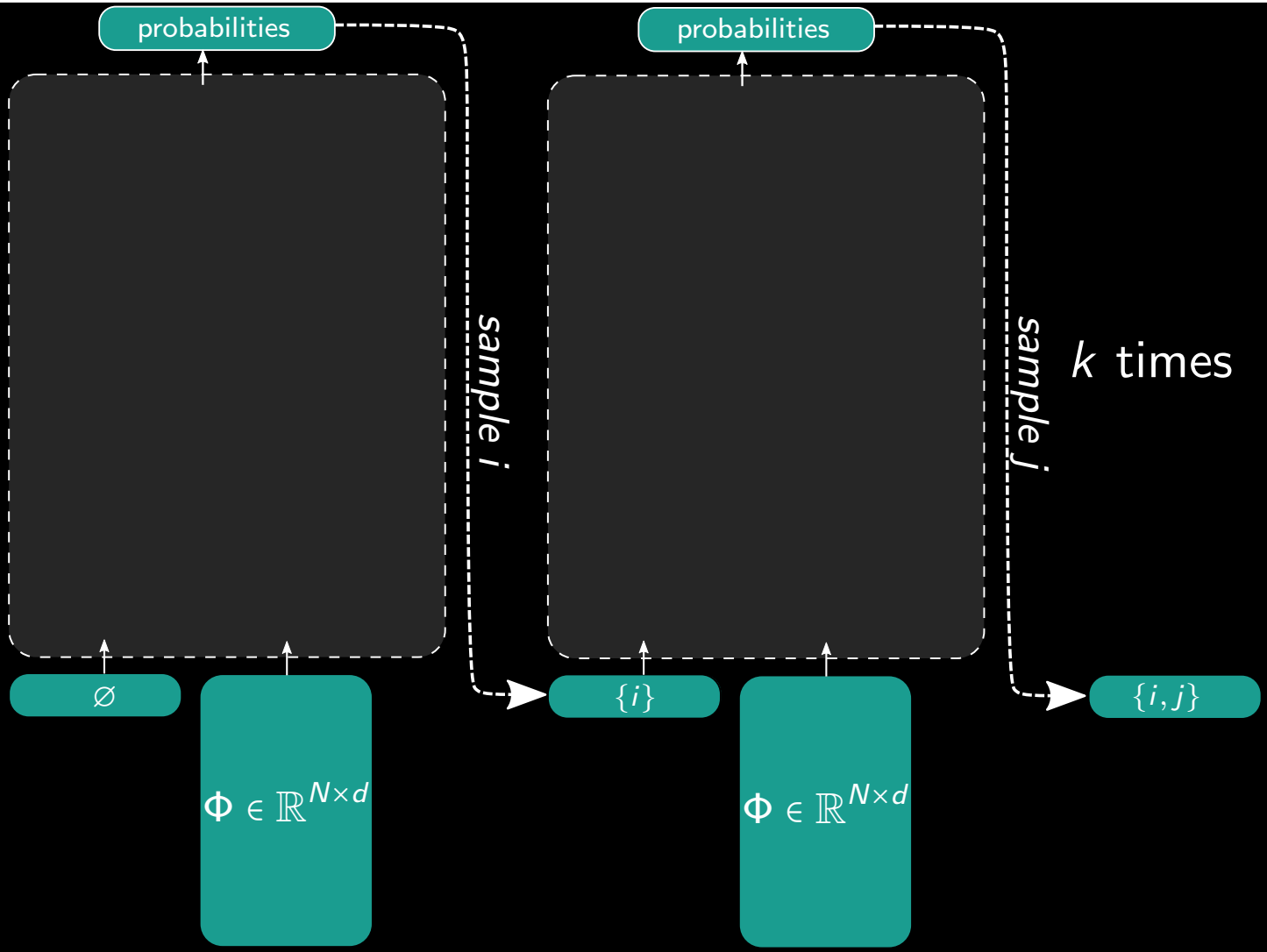
- Given a distribution over Φ, can generate ∞ training data
- Sampling from a DPP generates a sequential set build-up
- Theoretical guarantees under certain training assumptions

For small enough $\max_S \|\Pr(S) - \widehat{\Pr}(S)\|$,
$\widehat{\Pr}$ is also log-submodular.

[Zelda Mariet, Yaniv Ovadia, Jasper Snoek, 2019]

DppNet: Motivation

Need: Fast, parallelizable sampling for varying ground sets
→ generative network for DPP-like samples



[Zelda Mariet, Yaniv Ovadia, Jasper Snoek, 2019]

Experimental results: data summarization

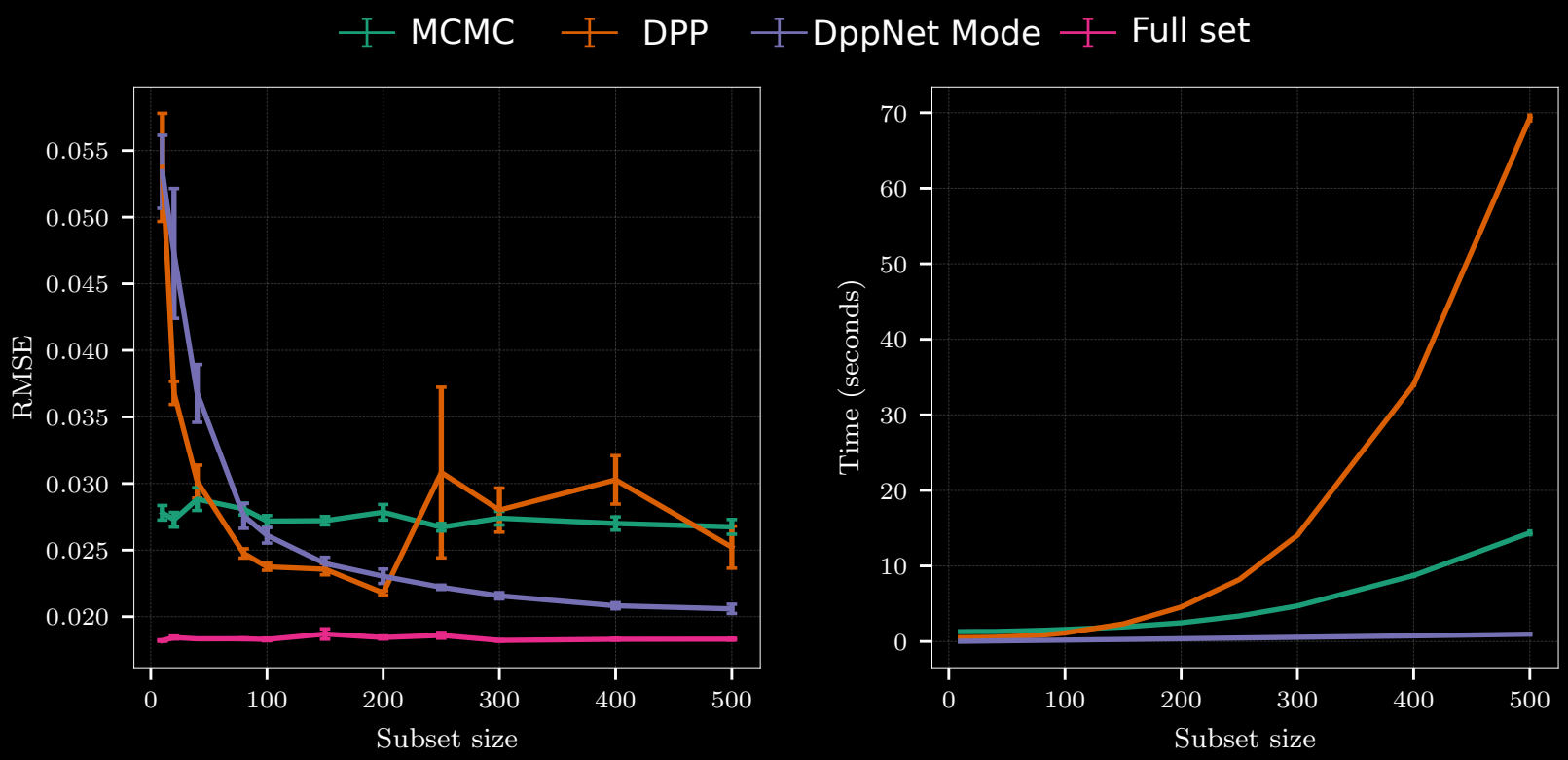
Train on features from MNIST training set; summarize test set.

[Zelda Mariet, Yaniv Ovadia, Jasper Snoek, 2019]

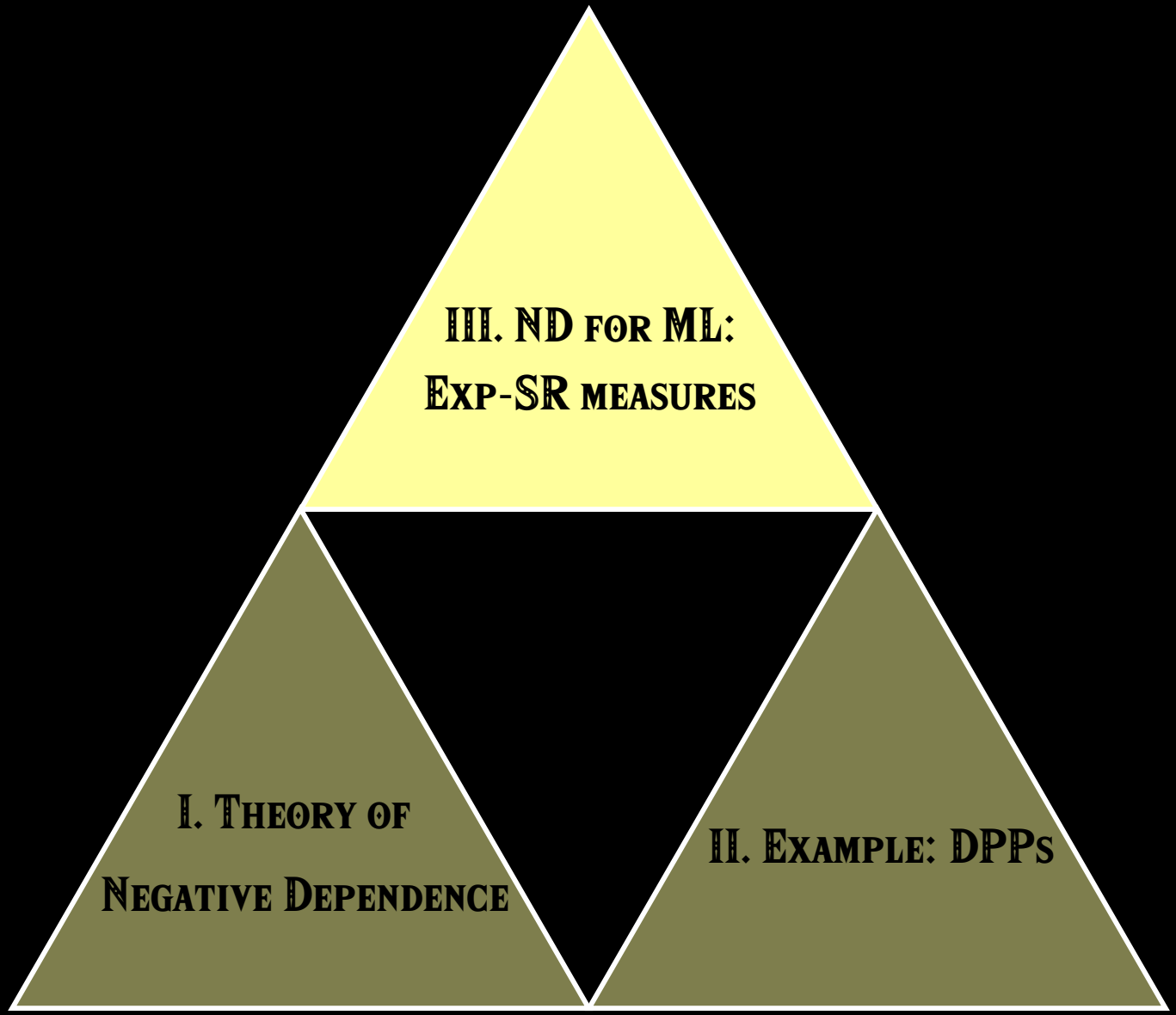
Experimental results: kernel reconstruction

Nyström method: given large kernel $K \in \mathbb{R}^{n \times n}$, approximate K via

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



[Zelda Mariet, Yaniv Ovadia, Jasper Snoek, 2019]


III. ND for ML:
Exp-SR measures
I. Theory of
Negative Dependence
II. Example: DPPs

Some observations

- ☺ SR measures allow fast sampling algorithms [Anari et al., 2016, Li et al., 2016b]
- ☹ No way to tune diversity strength

Can we have both?

- Fast learning/sampling
- Tune-able diversity strength

Exponentiated SR measures

Idea: augment SR measures to maintain fast sampling but control diversity

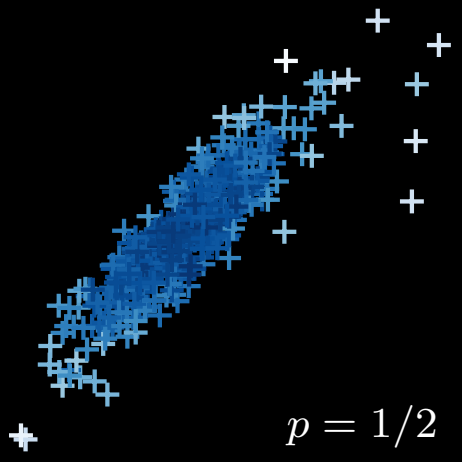
Mariet et al. [2018]: μ is *Exponentiated Strongly Rayleigh* if

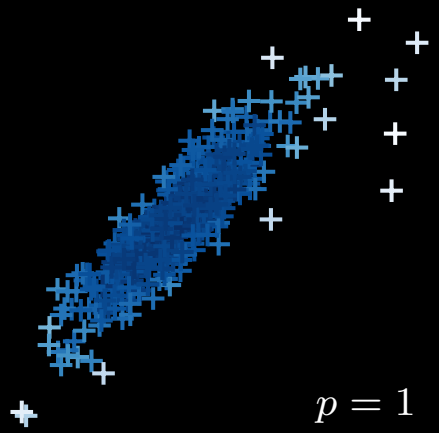
$$\mu \propto \nu^p \quad p \geq 0 \text{ and } \nu \text{ SR.}$$

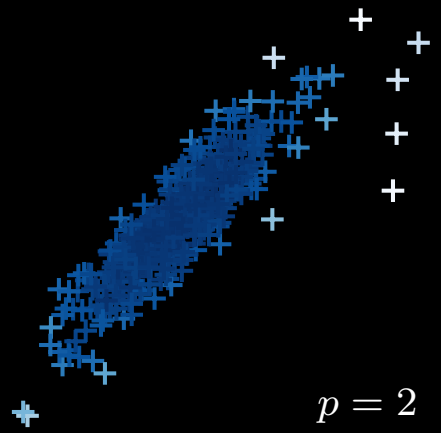
$p \to 0$: μ becomes uniform

$p \to \infty$: μ converges towards the mode

Controlling diversity







$$\mu(S) = \textsc{Dpp}(S)^p$$

[Zelda Mariet, Suvrit Sra, Stefanie Jegelka, NeurIPS 2018]

Fast sampling for ESR measures

Mariet et al. [2018]: μ is *Exponentiated Strongly Rayleigh* if

$$\mu \propto \nu^p \qquad\qquad p \geq 0 \text{ and } \nu \text{ SR.}$$

Fast sampling for ESR measures

Mariet et al. [2018]: μ is *Exponentiated Strongly Rayleigh* if

$$\mu(S) = \frac{1}{Z_p}\nu(S)^p \quad p \geq 0 \text{ and } \nu \text{ SR.}$$

Computing Z_p is NP-hard ☹

→ take advantage of fast MCMC sampling for SR measures

Algorithm 1: Approximate sampling for ESRs

Input: target ESR $\mu \propto \nu^p$

Draw $S \sim \nu$

while not mixed **do**

$S' \sim \nu$

$S \leftarrow S'$ with prob. $\min\left\{1, \frac{\mu(S')\nu(S)}{\mu(S)\nu(S')}\right\} = \min\left\{1, \frac{\nu(S)^{1-p}}{\nu(S')^{1-p}}\right\}$

return S

- No need for Z_p
- Fast sampling for ν

[Zelda Mariet, Suvrit Sra, Stefanie Jegelka, NeurIPS 2018]

Mixing time: required number of inner loops for a good approximation

$$\tau_S(\epsilon) := \min\{t : \|\tilde{\mu}_{t,S} - \mu\|_{TV}\} \leq \epsilon$$

Theorem: The previous algorithm mixes in time

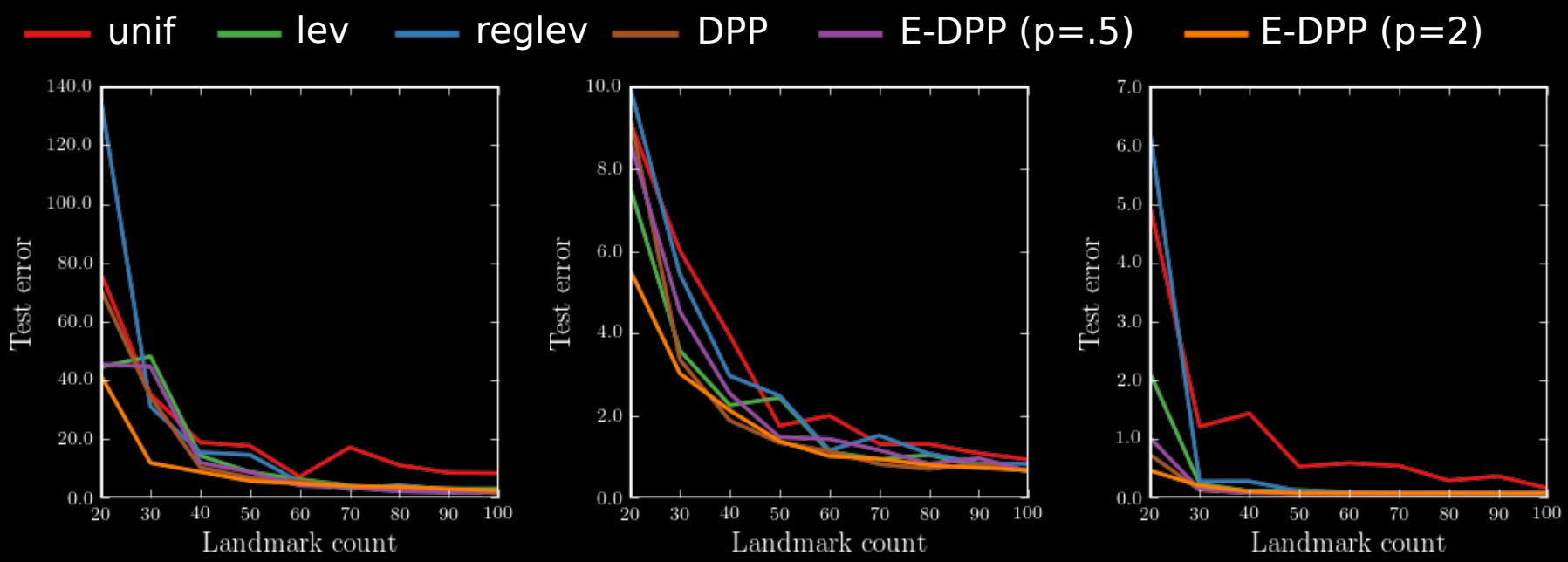
$$\tau(\epsilon) \leq 2 \max_{S \in \mathrm{Supp}(\nu)} \nu(S)^{-|p-1|} \log \frac{1}{\epsilon}$$

[Zelda Mariet, Suvrit Sra, Stefanie Jegelka, NeurIPS 2018]

Application: kernel reconstruction

Nyström method [Nyström, 1930]: given large kernel $K \in \mathbb{R}^{n \times n}$, approximate K via

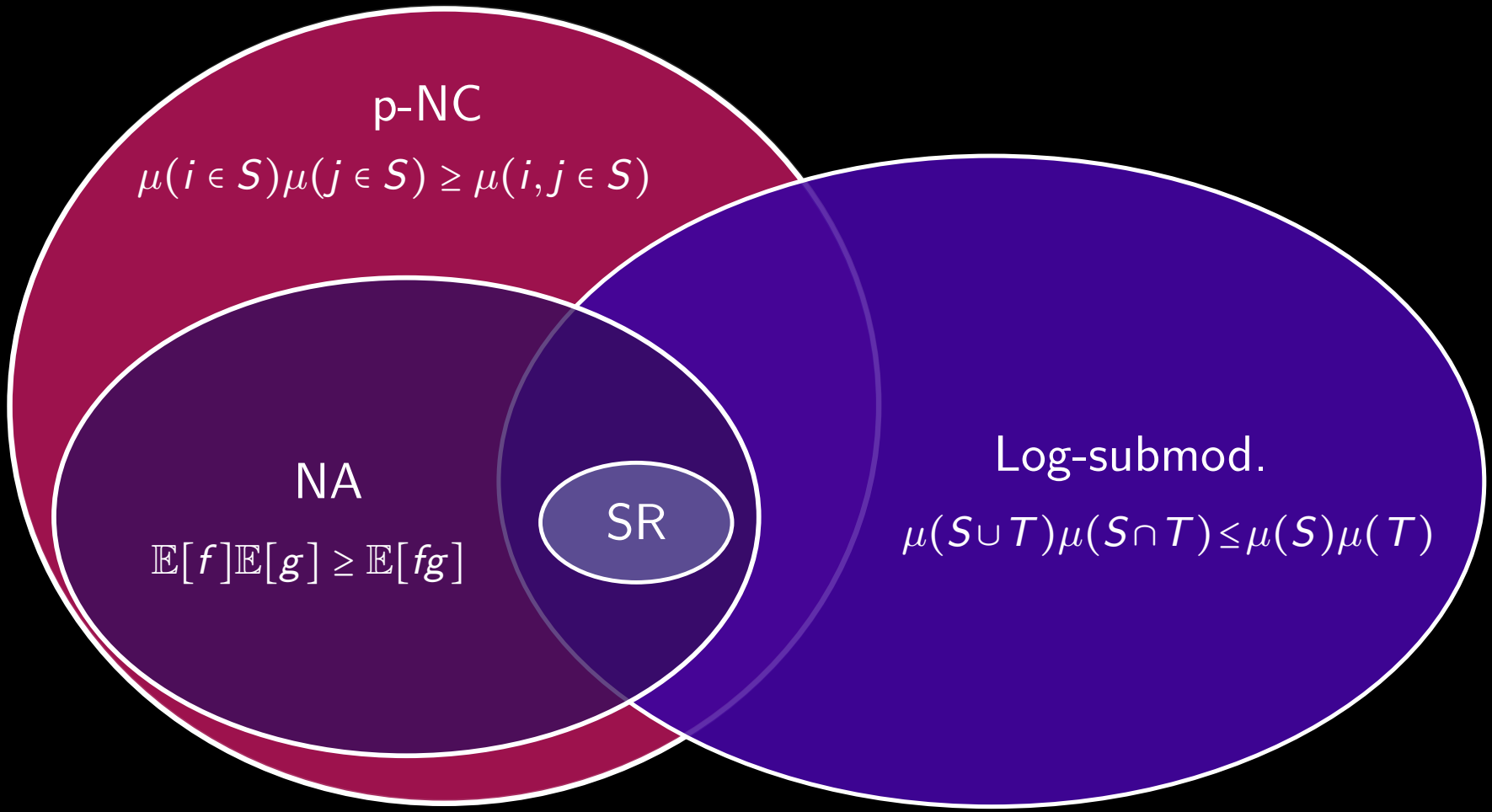
$$\hat{K} = K_{:,S} K_{S,S}^{\dagger} K_{S,:}$$



(a) Ailerons dataset (b) Bank32NH dataset (c) CPU dataset

[Zelda Mariet, Suvrit Sra, Stefanie Jegelka, NeurIPS 2018]

For the mathematically curious...



[Zelda Mariet, Suvrit Sra, Stefanie Jegelka, NeurIPS 2018]

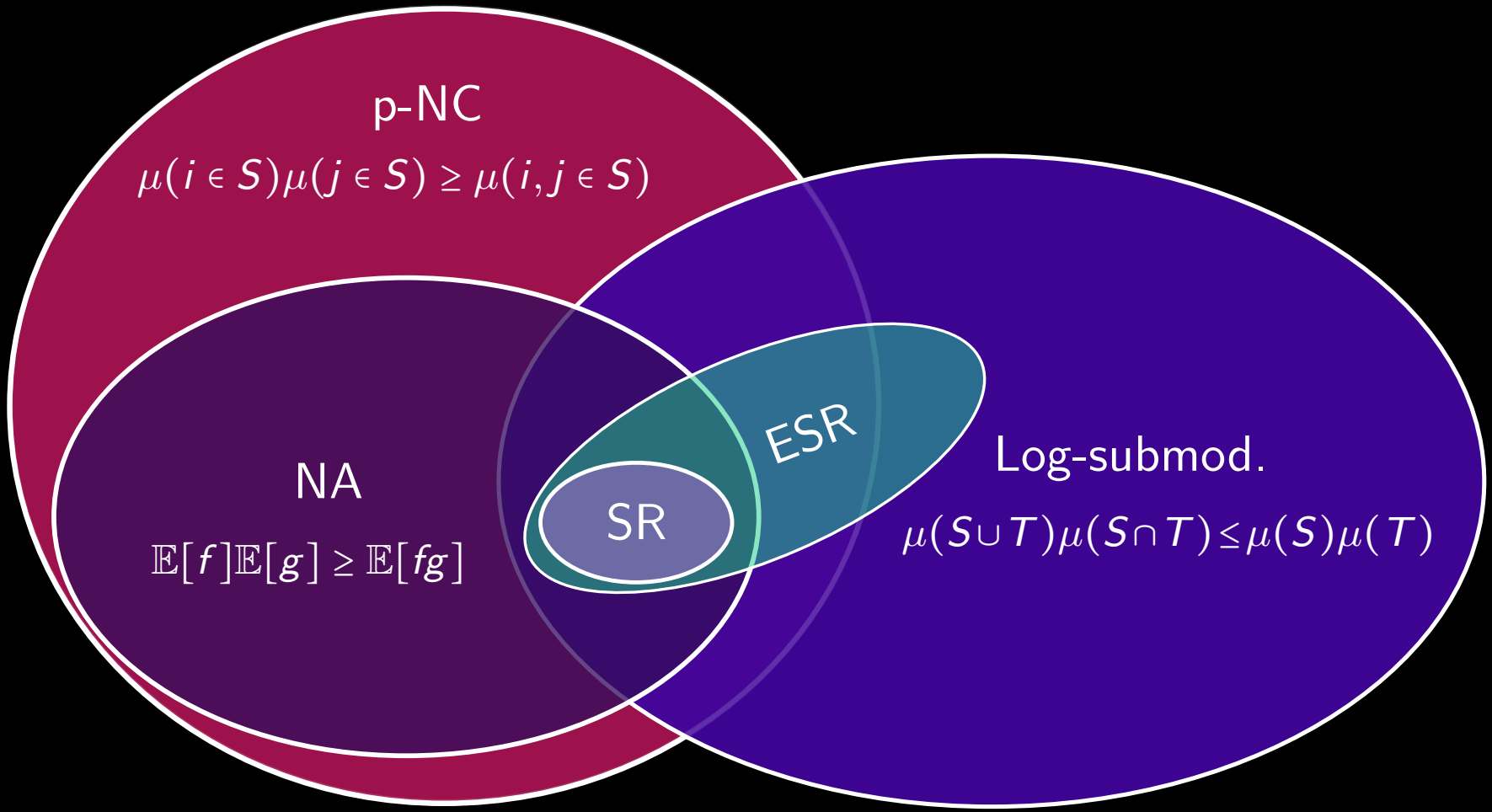
What about ESR measures?

- ESR measures are log-submodular
- ESR measures are not always SR

Theorem: There exist non-trivial ESR measures that are SR

[Zelda Mariet, Suvrit Sra, Stefanie Jegelka, NeurIPS 2018]

For the mathematically curious...



[Zelda Mariet, Suvrit Sra, Stefanie Jegelka, NeurIPS 2018]

Currently:

- Rich theory of negative dependence
- With many punctual applications to ML

New paradigm: ML through the lens of negative dependence

→ Efficient algorithms for negatively dependent models

→ Negative dependence theory specific to ML problems

Next steps?

- Reinforcement learning
- Hyperparameter space exploration
- Non-uniform sampling for SGD
- Fairness

N. Anari, S. Oveis Gharan, and A. Rezaei. Monte Carlo Markov Chain algorithms for sampling strongly Rayleigh distributions and Determinantal Point Processes. In *Conference on Learning Theory (COLT)*, 2016.

Julius Borcea, Petter Brändén, and Thomas Liggett. Negative dependence and the geometry of polynomials. *Journal of the American Mathematical Society*, 22(2), 2009.

Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth Vishnoi. Fair and diverse DPP-based data summarization. In *Proc. Int. Conference on Machine Learning (ICML)*, 2018.

Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Fast DPP sampling for Nyström with application to kernel methods. In *Proc. Int. Conference on Machine Learning (ICML)*, ICML'16. JMLR.org, 2016a.

Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Fast mixing Markov chains for strongly Rayleigh measures, DPPs, and constrained sampling. In *Advances in Neural Information Processing Systems*, 2016b.

Zelda Mariet and Suvrit Sra. Diversity networks: Neural network compression using Determinantal Point Processes. *Int. Conf. on Learning Representations (ICLR)*, 2016.

Zelda Mariet, Suvrit Sra, and Stefanie Jegelka. Exponentiated Strongly Rayleigh distributions. In *Advances in Neural Information Processing Systems*. 2018.

E.J. Nyström. Über die praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben. *Acta Mathematica*, 54(1), 1930.

Cheng Zhang, Hedvig Kjellström, and Stephan Mandt. Stochastic learning on imbalanced data: Determinantal Point Processes for mini-batch diversification. *CoRR*, abs/1705.00607, 2017.