Flexible modeling of diversity with strongly log-concave distributions

Joshua Robinson, Suvrit Sra, and Stefanie Jegelka

Massachusetts Institute of Technology

Josh Robinson
1st November 2019

A brief history of negative dependence

- Negative dependence is intimately connected to combinatorics

A brief history of negative dependence

- It appears frequently in combinatorial phenomena, e.g. random spanning trees, random cluster models, percolation, matroid theory etc.

A brief history of negative dependence

- "Disparate problems in combinatorics, ranging from problems in statistical mechanics to the problem of coloring a map, seem to bear no common features. However, they do have at least one common feature: **their solution can be reduced to the problem of finding the roots of some polynomial or analytic function.**" - Gian-Carlo Rota

A brief history of negative dependence

- In the spirit of Rota's world view, theories of negative dependence have been largely codified in terms of polynomials and their properties

A brief history of negative dependence

- Real stable polynomials are, perhaps the most famous example of a class of polynomials with negative dependence properties

- Famously used to prove the **Kadison-Singer conjecture**, which had been open for over 60 years and is known to have deep connections to many fields of mathematics, including quantum mechanics and C*-algebras.

A brief history of SLC polynomials

- This talk is about strong log-concavity
- SLC polynomials include all real stable polynomials
- Strong log-concavity was proposed by Gurvit's in 2009 as a property enabling approximation algorithms for discrete problems

A brief history of SLC polynomials

- SLC polynomials rose to prominence since 2018 when they were used by Anari et al., and Brändén and Huh to resolve several problems in matroid theory, including Mason's conjecture

LORENTZIAN POLYNOMIALS

PETTER BRÄNDÉN AND JUNE HUH

Log-Concave Polynomials II: High-Dimensional Walks and an FPRAS for Counting Bases of a Matroid

Nima Anari[1], Kuikui Liu[2], Shayan Oveis Gharan[2], and Cynthia Vinzant[3]

Log-Concave Polynomials I: Entropy and a Deterministic Approximation Algorithm for Counting Bases of Matroids

Nima Anari[1], Shayan Oveis Gharan[2], and Cynthia Vinzant[3]

Log-Concave Polynomials III: Mason's Ultra-Log-Concavity Conjecture for Independent Sets of Matroids

Nima Anari[1], Kuikui Liu[2], Shayan Oveis Gharan[2], and Cynthia Vinzant[3]

A brief history of SLC polynomials

- Convexity/concavity makes continuous optimization tractable
- Matroid property makes discrete optimization tractable
- Strong log-concavity was shown to connect these two idea in the work of Anari et al., and Brändén and Huh

Our focus

- Today we focus on SLC in the context of **diversity inducing probability distributions**

What do we mean by diversity?

- We have a collection of items $[n] = \{1,\ldots,n\}$
- We are interested in assigning a probability $\pi(S)$ to each $S \subset [n]$
- High level idea:

  If $i, j \in [n]$ are similar, then they are unlikely to co-occur

- E.g. Zelda's cinema / students studying in a library

Defining diversity

- Diversity inducing properties:
  - Pairwise negative correlation
    $\pi(i, j \in S) \leq \pi(i \in S)\pi(j \in S)$
  - Log-submodularity
    $\pi(S)\pi(S \cup \{i, j\}) \leq \pi(S \cup i)\pi(S \cup j)$

The usefulness of diversity in ML

- Video summarization
- Model compression
- Avoiding mode collapse in generative models
- Fairness
- Accelerated coordinate descent
- SGD minibatch selection

Previous models for diversity

- Strongly Rayleigh measures (aka have real stable generating polynomial)
- In particular, determinantal point processes
- **But they do not allow easy control over the nature and strength of the induced diversity**

Agenda

- What are SLC distributions?
- Some comparison to SR
- Basic computational tools:
  - Sampling (with guarantees)?
  - Mode finding (with guarantees)?

Generating polynomial

- There is a one to one correspondence between discrete distributions polynomials with *non-negative coefficients* (up to normalization)

$$\pi \longleftrightarrow f_\pi$$

Where $f_\pi(z_1, \ldots, z_n) = \sum_{S \subset [n]} \pi(S) \prod_{i \in S} z_i$

Definition of strongly log-concave (SLC)

Definition:

A polynomial $f(z_1, \ldots, z_n)$ with non-negative coefficients is said to be *strongly log-concave* if for any $\alpha \in \mathbb{N}^n$, the function $\log(\partial^\alpha f(z))$ is concave for all $z \in \mathbb{R}_+^n$.

Definition of strongly log-concave (SLC)

Definition:

A distribution $\pi : 2^{[n]} \to \mathbb{R}_+$ is strongly log-concave if its generating polynomial f_π is.

Why is SLC more flexible than SR?

- Because SR $\subset$ SLC (see e.g. Brändén and Huh 2019)
- There are interesting things in SLC$\setminus$SR

Why is SLC more flexible than SR?

- Budget constrained distribution

Theorem:

If π is SLC, then $\nu(S) \propto \pi(S)\mathbf{1}\{|S| \leq k\}/(n - |S|)!$ is too

Moral:

If π is SLC, then $\nu(S) \propto \pi(S)\mathbf{1}\{|S| \leq k\}$ is too

Interesting?

k can act as a "maximum budget"

Why is SLC more flexible than SR?

- Smoothed distribution

Theorem:

If π is SLC, then $\nu(S) \propto \pi(S)^\alpha / (n - |S|)!$ is too for $0 \leq \alpha \leq 1$

Moral:

If π is SLC, then $\nu(S) \propto \pi(S)^\alpha$ is too for $0 \leq \alpha \leq 1$

Interesting?

As α decreases from 1, the distribution becomes closer to uniform

What other SLC distributions are there?

- The uniform distribution on bases of a matroid. This was critical in the work by Anari et al.

- This is an open question

Sampling

- Anari et al. (log-concave polys II) gave a sampler for *homogenous* SLC distributions
- What about the general case?

Sampling

- From now, suppose $\nu := \pi$, or $\nu \propto \pi^\alpha$ with $0 \leq \alpha \leq 1$, or $\nu \propto \pi \mathbf{1}\{|S| \leq k\}$.
- More generally, let ν be any distribution such that $\nu(S)/(n - |S|)!$ is SLC
- Assume that the support of ν is on sets of cardinality less than or equal to d

Sampling

- Strategy: use the homogenous sampler somehow
- To sample from ν we devise a sampler for

$$\nu_{\mathrm{sh}}(S) \propto \begin{cases} \binom{k}{|S \cap [n]|}^{-1} \nu(S \cap [n]), & S \subset [n+d], |S| = d \\ 0, & \text{otherwise} \end{cases}$$

Symmetric homogenization

Sampling

- Properties of ν_{sh}:
  - d-homogenous
  - Marginal distribution on $[n]$ is exactly ν
  - Symmetric in variables $n+1, n+2, \ldots, n+d$
- Consequence: a simple recipe for sampling from ν:
  - Sample $S \sim \nu_{\text{sh}}$
  - Define $T := S \cap [n]$
- So our problem reduces to sampling from ν_{sh}

Sampling

- ν_{sh} is
  - Homogenous (good)
  - Not necessarily SLC (bad)
- But $\nu_{\text{sh}}(S)/(n - |S|)!$ is SLC (Theorem)
- We can sample from it using Anari et al.'s homogenous MCMC kernel, Q
- $Q =$ drop element uniformly at random, add new element proportionally to the probability of the resulting set

Sampling

Algorithm 1 Metropolis-Hastings sampler for ν_{sh} with proposal Q

Initialize $S \subseteq [n+d]$
while not mixed **do**
    Set $k \leftarrow \left|S \cap [n]\right|$
    Propose move $T \sim Q(S, \cdot)$
    if $\left|T \cap [n]\right| = k - 1$ **then**
        $R \leftarrow T$ with probability $\min\{1, \frac{e}{d}(d-k+1)\}$, otherwise stay at S
    if $\left|T \cap [n]\right| = k$ **then**
        $R \leftarrow T$
    if $\left|T \cap [n]\right| = k + 1$ **then**
        $R \leftarrow T$ with probability $\min\{1, \frac{d}{e}\frac{1}{(d-k)}\}$, otherwise stay at S

Sampling

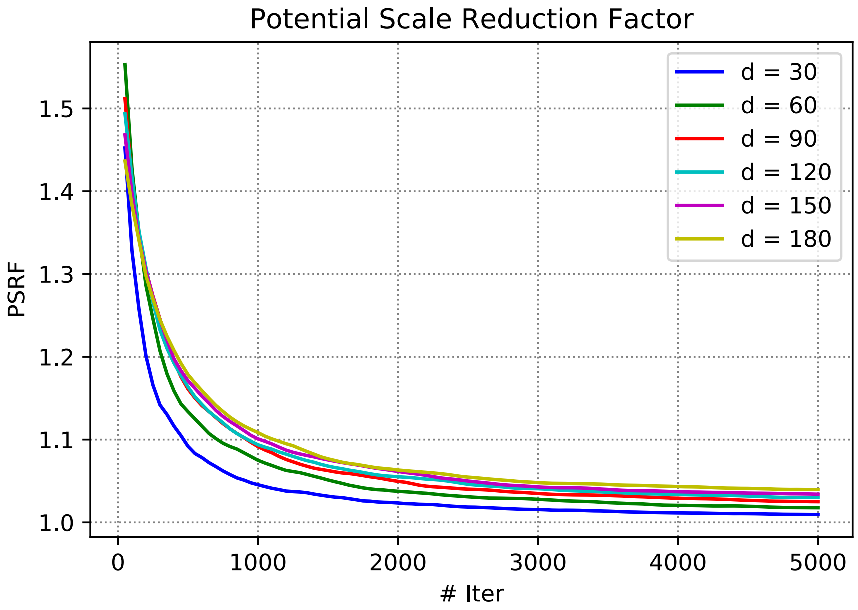
- The mixing time of (Q, ν_{sh}) started at $S_0 \subset [n]$ is

$$t_{S_0}(\varepsilon) = \min\{t \in \mathbb{N} \,|\, \|Q^t(S_0, \cdot\,) - \nu_{\text{sh}}\|_1 \leq \varepsilon\}$$

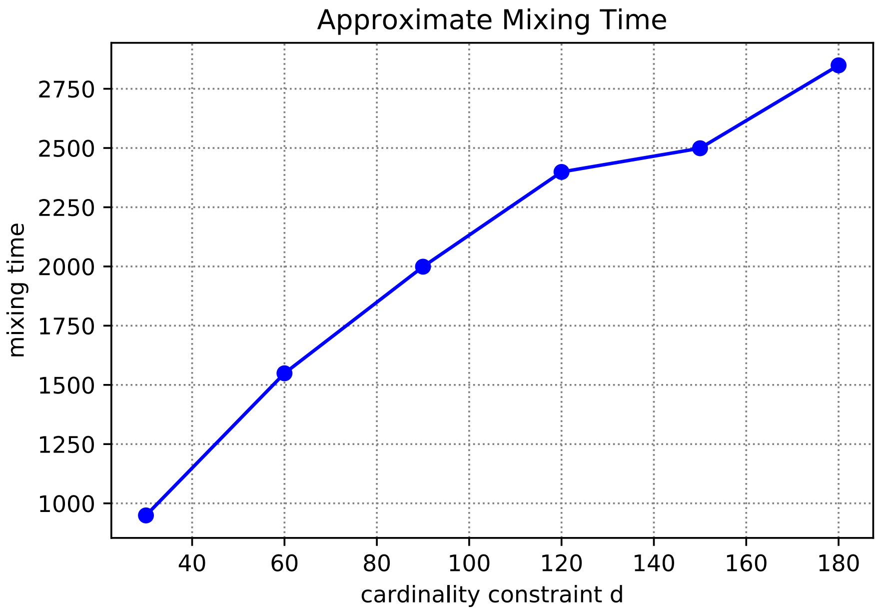
Theorem *For $d \geq 8$ the mixing time of the chain in Algorithm 1 started at S_0 satisfies the bound*

$$t_{S_0}(\varepsilon) \leq \frac{1}{e\sqrt{2\pi}} d^{5/2} 2^d \left(\log\log\left\{ \binom{d}{|S_0|} \frac{1}{\nu(S_0)} \right\} + \log \frac{1}{2\varepsilon^2} \right).$$

Sampling



Potential Scale Reduction Factor
PSRF
Iter
1.0
1.1
1.2
1.3
1.4
1.5
0
1000
2000
3000
4000
5000
d = 30
d = 60
d = 90
d = 120
d = 150
d = 180



Approximate Mixing Time
mixing time
cardinality constraint d
1000
1250
1500
1750
2000
2250
2500
2750
40
60
80
100
120
140
160
180

Mode finding

- We would like to find

$$\text{OPT} \in \arg\max_{|S| \leq k} \pi(S)$$

- But we don't want to check all $\sum_{j=1}^{k} \binom{n}{j}$ possibilities

- Aim: use submodularity - a nice property that yields fast algorithms with optimization guarantees

Mode finding

- Alkis recently proved that SLC distributions do not have to be log-submodular
- From an optimization perspective this is unfortunate :(
- However, SLC distribution are weakly log-submodular

Mode finding

Theorem (weak log-submodularity)**:**

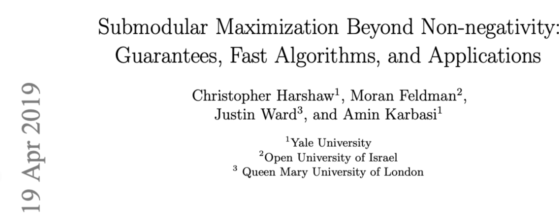
$$\nu(S) \cdot \nu(S \cup \{i, j\}) \leq \gamma \cdot \nu(S \cup i) \cdot \nu(S \cup j)$$

For any $S \subset [n]$, and $i, j \in [n]$ with $i \neq j$, where $\gamma = 4\left(1 - \frac{1}{d}\right)$

(Note, this is the same ν as before…)

Mode finding

- Algorithmically we apply submodular-type algorithms to $\rho := \log \nu$
- But ρ can be non-negative, and most submodular algorithms assume non-negativity.
- Fortunately, there is a recent algorithm, the distorted greedy algorithm, that works for any sign


Submodular Maximization Beyond Non-negativity: Guarantees, Fast Algorithms, and Applications

Christopher Harshaw[1], Moran Feldman[2], Justin Ward[3], and Amin Karbasi[1]

[1]Yale University
[2]Open University of Israel
[3] Queen Mary University of London

19 Apr 2019

Mode finding

- Decompose $\rho = \eta - c$, where η is non-negative, and c is modular - i.e. $c(S) = \sum_{i \in S} c_i$ for some c_i

Lemma (you can actually do this)**:**

First set $c_i = \max\{\rho([n]\backslash i) - \rho([n],0\}$, then define $\eta := \rho + c$. This gives the desired decomposition.

Mode finding

Algorithm 2 Distorted greedy weak submodular constrained maximization of $\nu = \eta - c$

1: Let $S_0 = \varnothing$
2: **for** $i = 0, \ldots, k-1$ **do**
3: Set $e_i = \arg\max_{e \in [n]} \Phi_{i+1}(S_i \cup e) - \Phi_{i+1}(S_i)$
4: **if** $\Phi_{i+1}(S_i \cup e_i) - \Phi_{i+1}(S_i) > 0$ **then**
5: $S_{i+1} \leftarrow S_i \cup e_i$
6: **else** $S_{i+1} \leftarrow S_i$
7: **return** $R = S_k$

- Build a sequence of sets $S_0, S_1, \ldots, S_k$
- Where we greedily maximize the distorted objective:

$$\Phi_i(S) = (1 - 1/k)^{k-i}\eta(S) - c(S)$$

Mode finding

Theorem 12. *Suppose* $\rho : 2^{[n]} \to \mathbb{R}$ *is* γ*-weakly submodular and* $\rho(\varnothing) = 0$*. Then the solution* $R = S_k$ *obtained by the distorted greedy algorithm satisfies*

$$\rho(R) = \eta(R) - c(R) \geq \left(1 - \frac{1}{e}\right)\left(\eta(OPT) - \frac{1}{2}\ell(\ell-1)\gamma\right) - c(OPT),$$

where $OPT \in \arg\max_{|S| \leq k} \rho(S)$ *and* $\ell := |OPT| \leq k$*.*

Open questions

- Learning an SLC distribution from data?
- What else is in SLC\SR?
- Negative dependence properties of SLC
- Close gap between mixing time bounds and practice

Summary

- Introduced the class of SLC distributions
- Exploration of what is in the class
- Sampling
- Mode finding