Discrete Sampling using Semigradient-based Product Mixtures

Alkis Gotovos — ETH Zurich

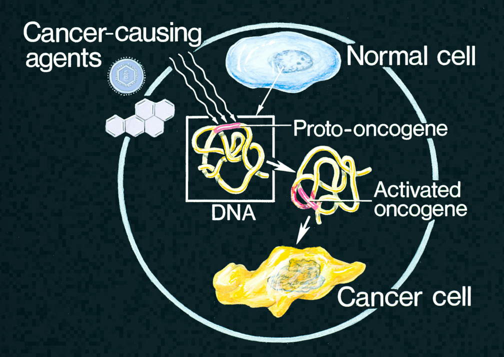
Hamed Hassani — UPenn

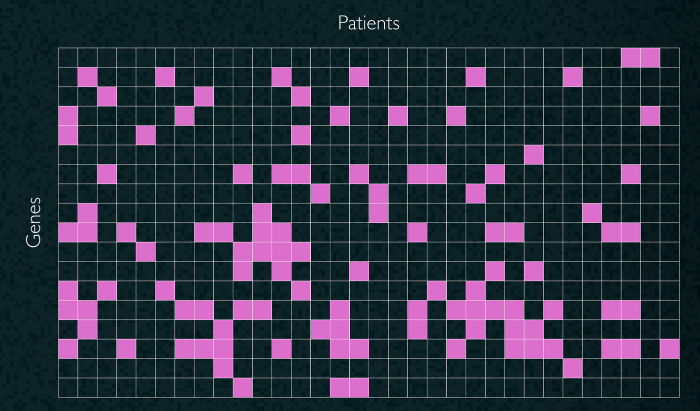
Andreas Krause — ETH Zurich

Stefanie Jegelka — MIT

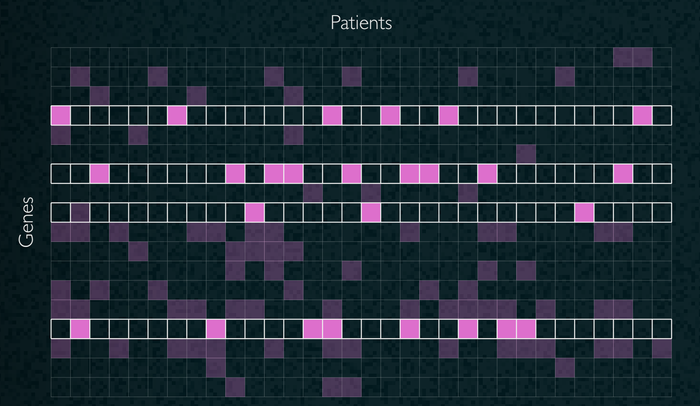
Motivation

Modeling gene alterations

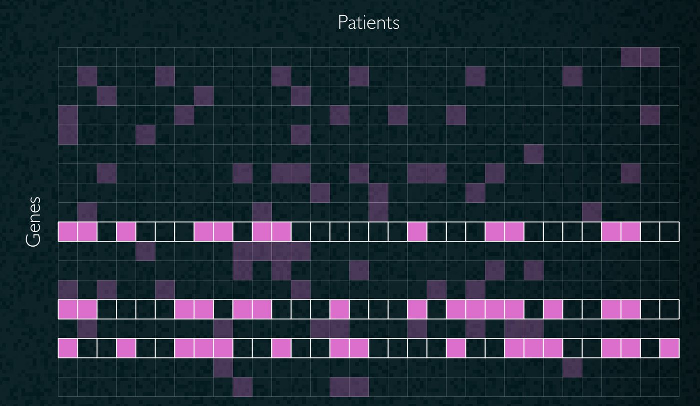




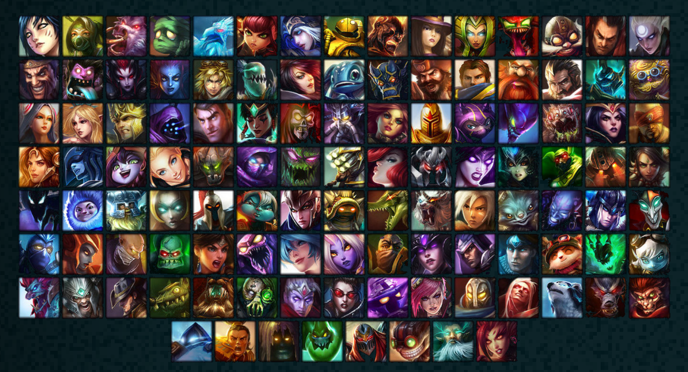
Mutual exclusivity

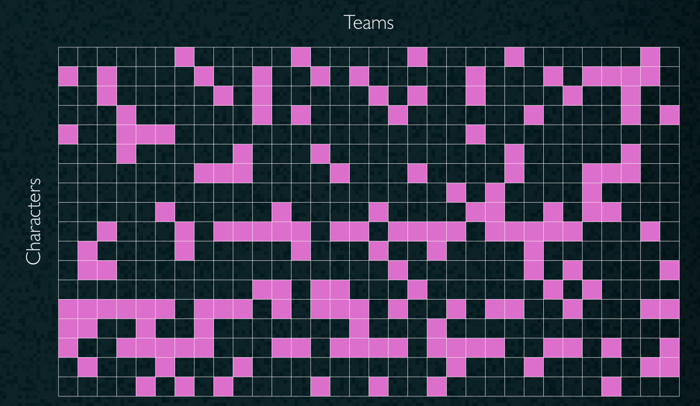


Co-occurence



Modeling teams





Discrete Probabilistic Models

Ground set $V = \{1, \ldots, n\}$

Data $\mathcal{D} = \{S_i\}_{i=0}^m,\ S_i \subseteq V$

Model higher-order interactions

$$\boxed{p(S;\theta) = \frac{1}{Z(\theta)} \exp\left(F(S;\theta)\right)}$$

- $F(S) =$ graph-cut(S) $\rightarrow$ Ising model
- $F(S) = \log|K_S|$ $\rightarrow$ DPP

Sampling and Inference

Learning

Learn θ

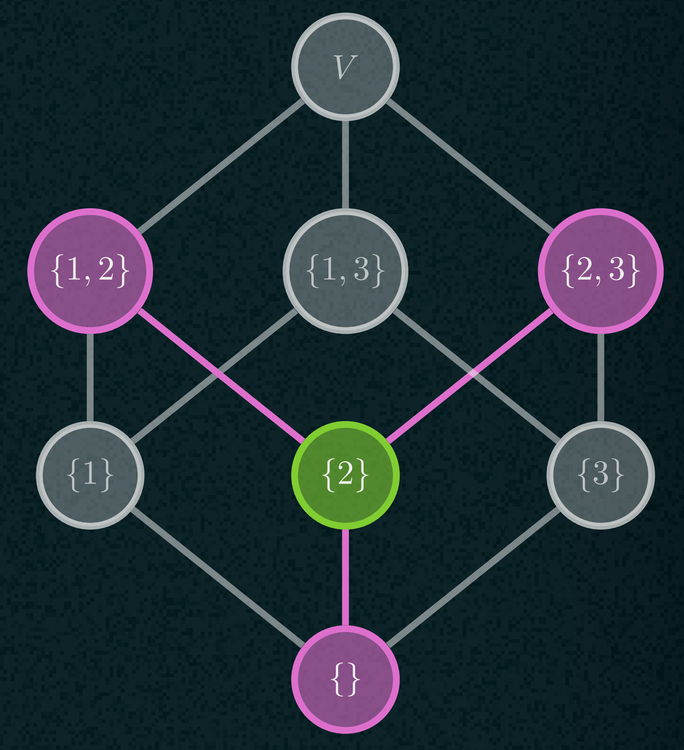
Max. likelihood

Approximate $\nabla_\theta Z(\theta)$

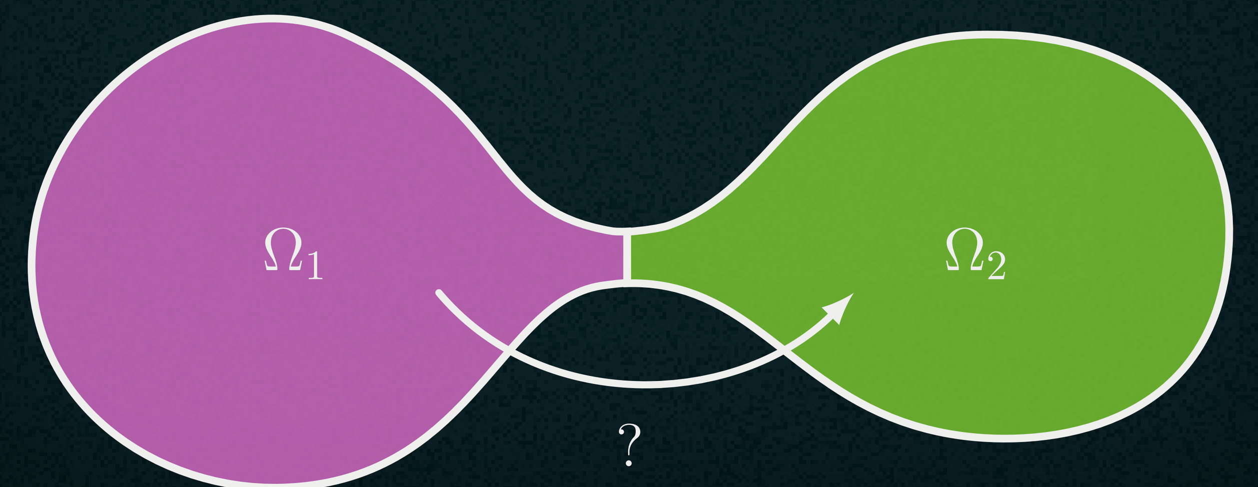
Exact computation
#P-hard in general

Sample from $p(\cdot\,;\theta)$

Gibbs sampling



When Gibbs fails



Contributions

- We propose the M^3 sampler, which makes global moves to avoid bottlenecks.
- For a specific class of Ising models, we show that combining Gibbs and M^3 results in an exponential mixing time improvement over Gibbs.
- We propose a semigradient-based mixture construction and demonstrate its effectiveness on three models learned from real-world data.

The M^3 Chain

M^3 = Mixture of Log-Modulars Metropolis

1. Mixture $\quad q(S,T) = q(T) = \dfrac{1}{Z_q}\displaystyle\sum_{i=1}^r w_i \exp\left(m_i(T)\right)$
2. Log-Modulars $\quad m_i(T) = \displaystyle\sum_{v\in T} m_{iv}$
3. Metropolis
   - Target $p(S) \propto \exp(F(S))$
   - Proposal $q(T)$
   - Accept with probability $\min\left\{1, \frac{p(T)q(S)}{p(S)q(T)}\right\}$

Proposition 1

Mixture q can approximate any distribution p arbitrarily well.

BUT may need an exponential (in n) number of components r

Constructing the Mixture

Ideally would want to minimize

$$E_1(q) := \left\| \frac{\exp(F(\cdot))}{Z} - \frac{1}{Z_q}\sum_{i=1}^r w_i \exp(m_i(\cdot)) \right\|$$

$$E_2^{(i)}(m_i) := \left\| \exp(F(\cdot)) - \sum_{j=1}^{i-1} \exp(m_i(\cdot)) \right\|,\quad i \in \{1,\ldots,r\}$$

Heuristic construction of q

Input: Set function F, mixture size r
for $i = 1$ **to** r **do**
  $\sigma \leftarrow$ Perm(V)
  $m_i \leftarrow$ SemiGradient(F, σ)
return $\{m_1, \ldots, m_r\}$

Perm

- Randomized construction

  $\sigma \leftarrow$ random permutation of V

- Iterative construction

  $\sigma \leftarrow$ Greedy$\left(F(\cdot) - \log\sum_{j=1}^{i-1}\exp(m_j(\cdot))\right)$

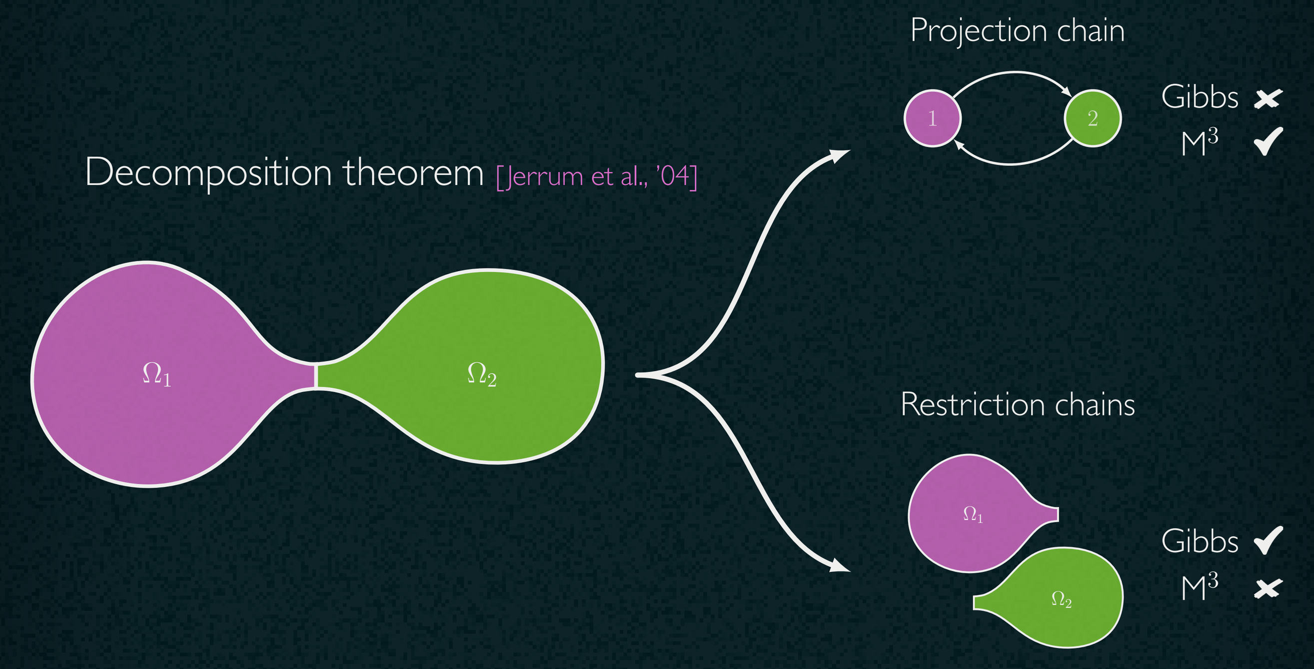
SemiGradient

- Submodularity $\rightarrow$ natural diminishing returns property
- Sub-/supergradients $\rightarrow$ modular lower/upper approx. [Iyer et al., '13]
- Construction works for general set functions F

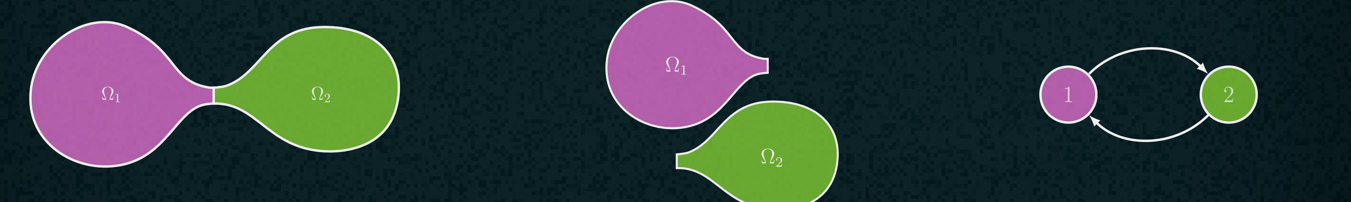
The Combined Chain

Gibbs step with prob. α | M^3 step with prob. $1-\alpha$

Decomposition theorem [Jerrum et al., '04]



Class of Ising models on the complete graph (Curie-Weiss)

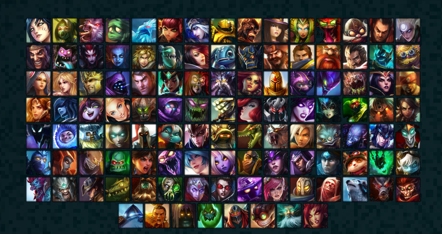


- Gibbs : $\gamma = \mathcal{O}\left(e^{-cn}\right)$ [Levin et al., '08]
- M^3 : ?
- Combo : $\gamma = \Omega\left(\frac{\log n}{n}\right)$ [Theorem 2]

- Gibbs : $\gamma = \Theta\left(\frac{\log n}{n}\right)$ [Ding et al., '09]

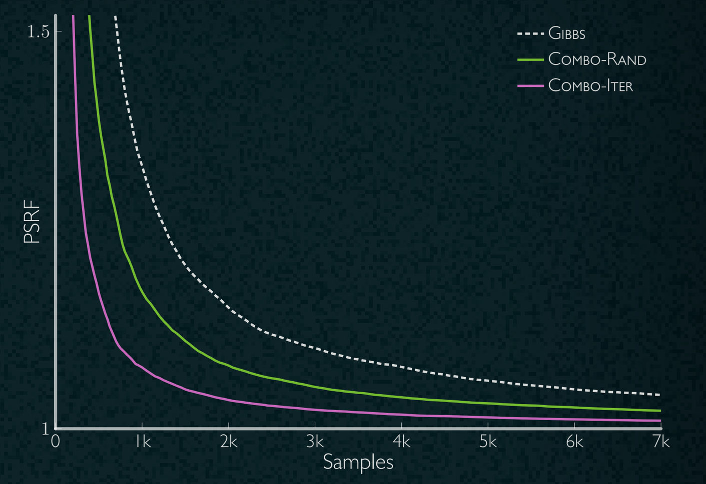
- M^3 : $\gamma = \Omega(1)$ [Lemma 1]

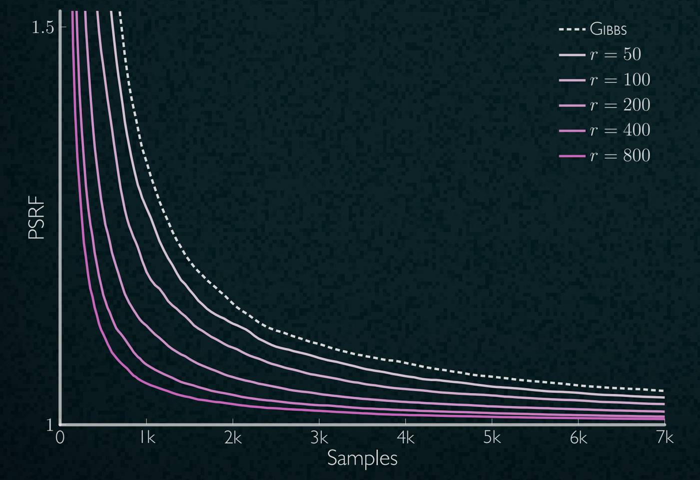
Experiments

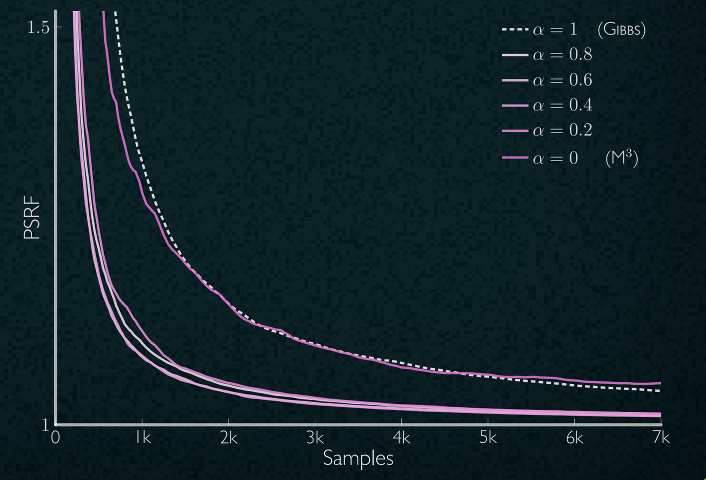


- $|V| = 48$
- 8.5k teams of 5 characters
- F is a (submodular) facility location diversity model [Tschiatschek et al., '16]

$$F(S) = \textstyle\sum_{i\in S} w_i + \sum_{j=1}^L \max_{i\in S} c_{ij}$$







[1] Jian Ding, Eyal Lubetzky, and Yuval Peres. Censored Glauber dynamics for the mean field Ising model. Journal of Statistical Physics, 2009.

[2] Rishabh Iyer, Stefanie Jegelka, and Jeff Bilmes. Fast semidifferential-based submodular function optimization. In International Conference on Machine Learning (ICML), 2013.

[3] Mark Jerrum, Jung-Bae Son, Prasad Tetali, and Eric Vigoda. Elementary bounds on Poincaré and log-Sobolev constants for decomposable Markov chains. Annals of Applied Probability, 2004.

[4] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. Markov Chains and Mixing Times. American Mathematical Society, 2008.

[5] Sebastian Tschiatschek, Josip Djolonga, and Andreas Krause. Learning probabilistic submodular diversity models via noise contrastive estimation. In International Conference on Artificial Intelligence and Statistics (AISTATS), 2016.