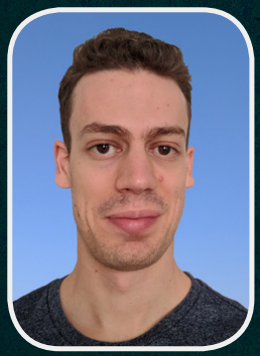
Discrete Sampling using Semigradient-based Product Mixtures

Alkis Gotovos
ETH Zurich



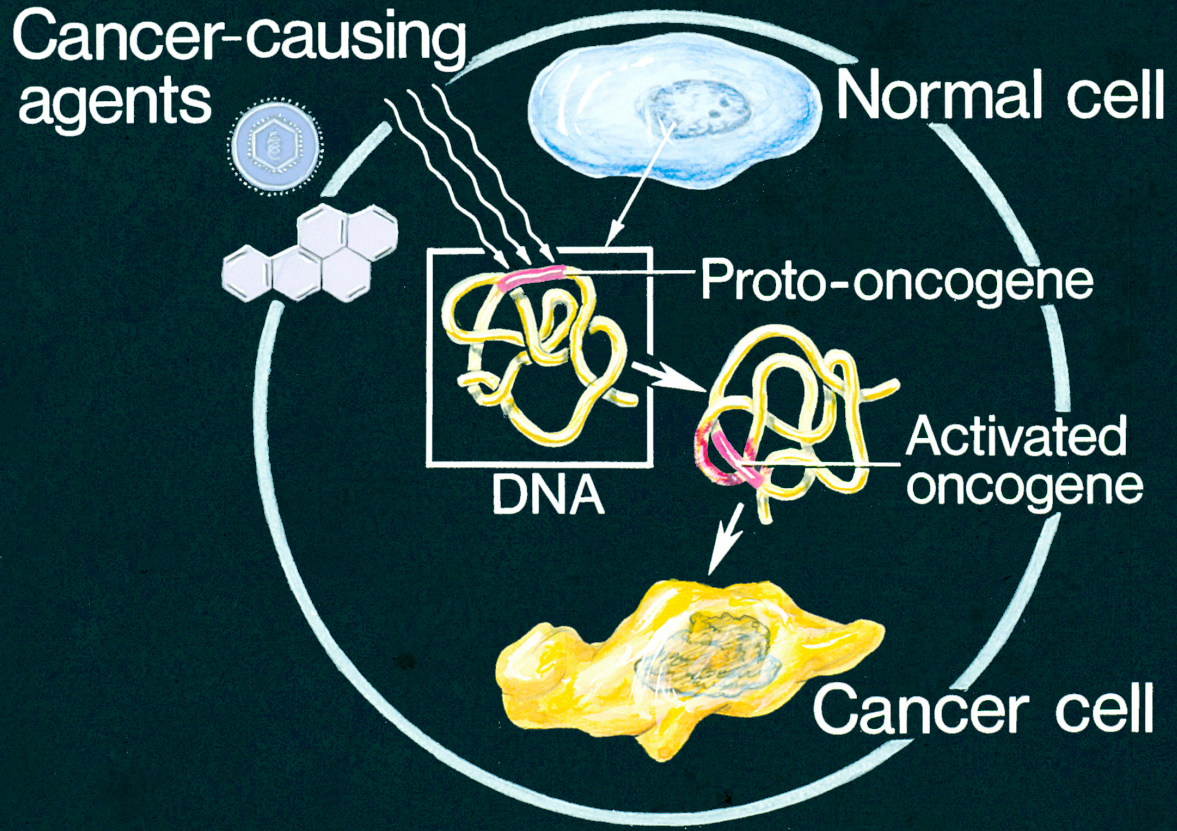
Hamed Hassani
UPenn

Andreas Krause
ETH Zurich

Stefanie Jegelka
MIT

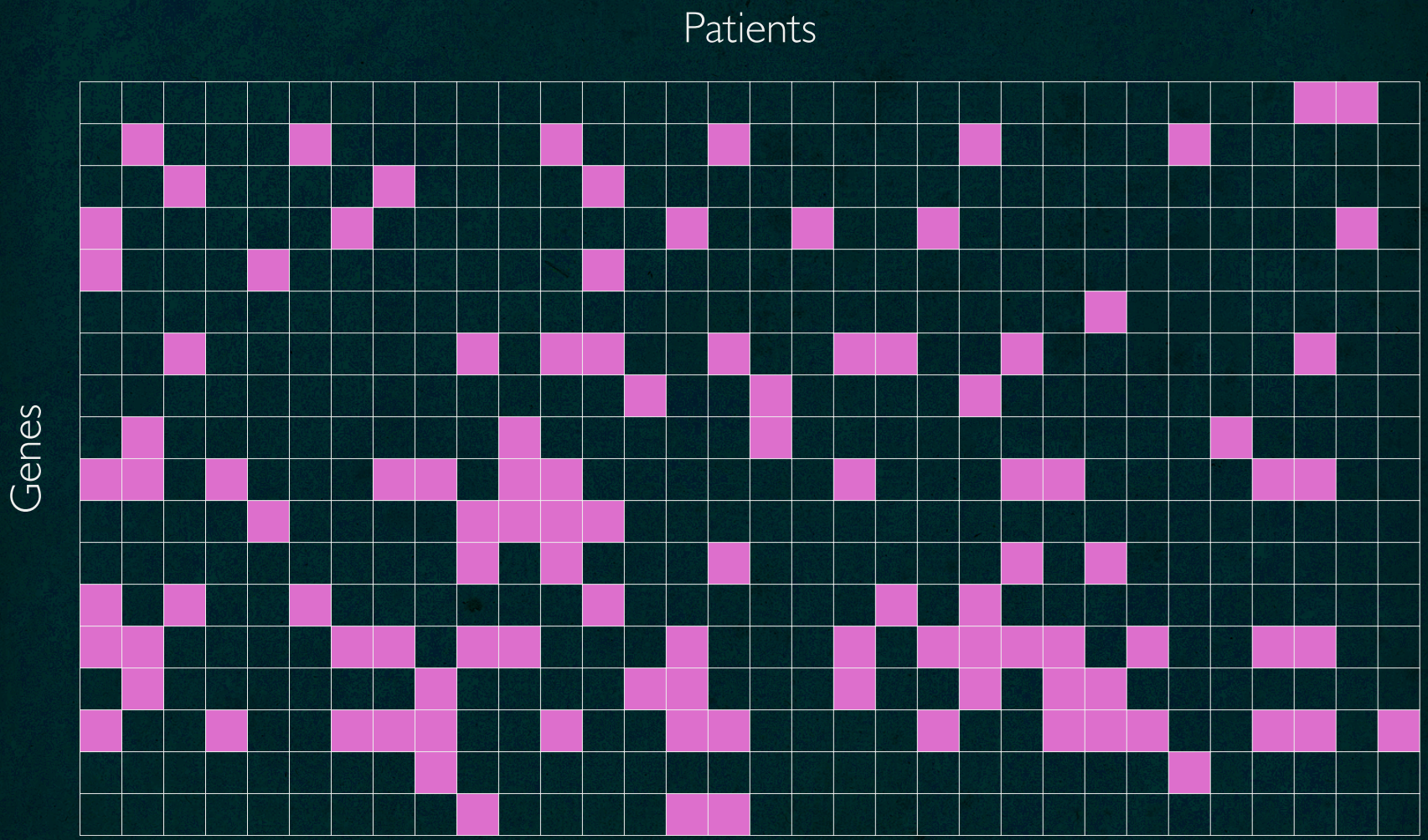
UAI 2018

Modeling gene alterations

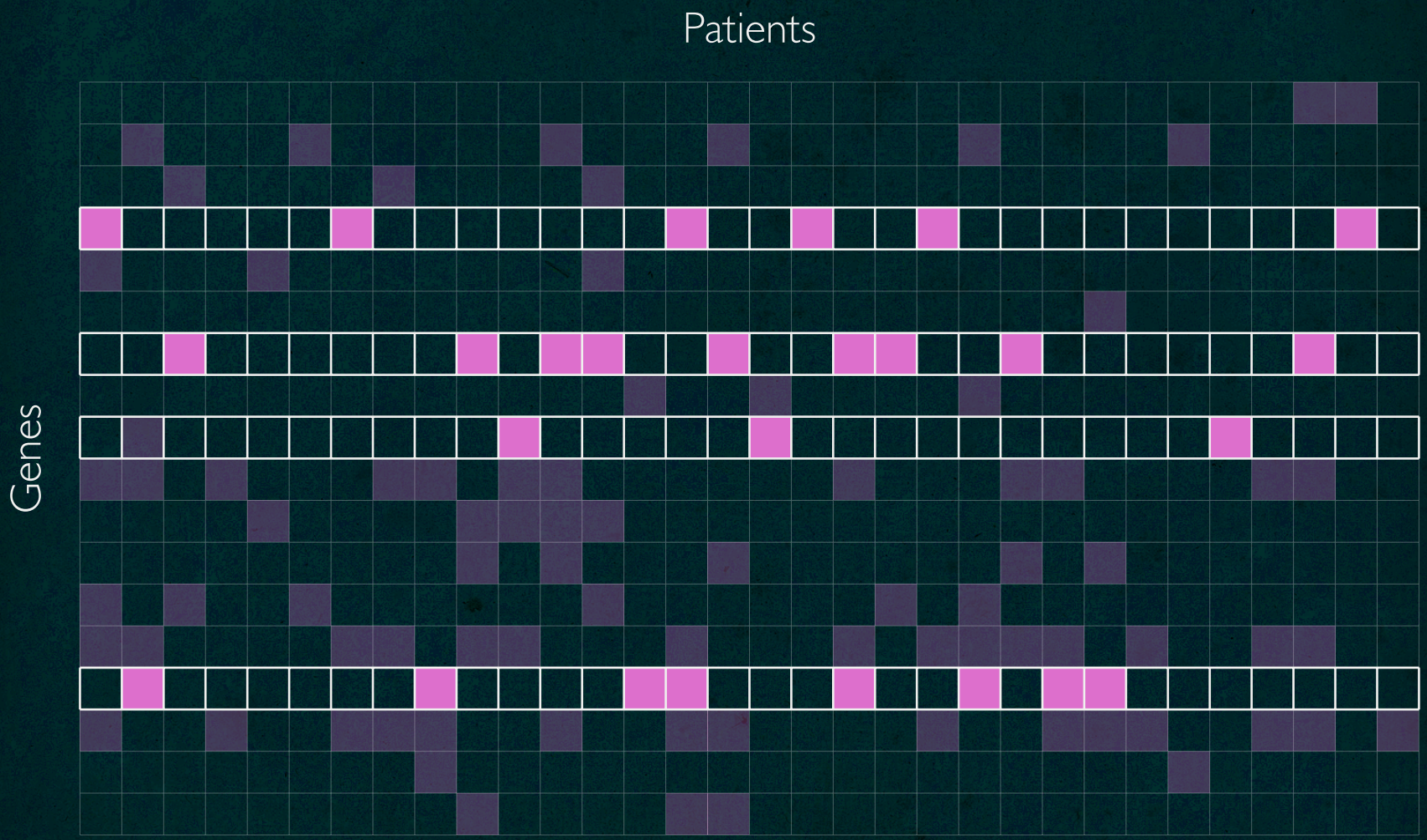


[cancergenome.nih.gov]

Modeling gene alterations



Modeling gene alterations



Modeling gene alterations

Modeling teams in online games

The OpenAI Dota 2 bots just defeated a team of former pros

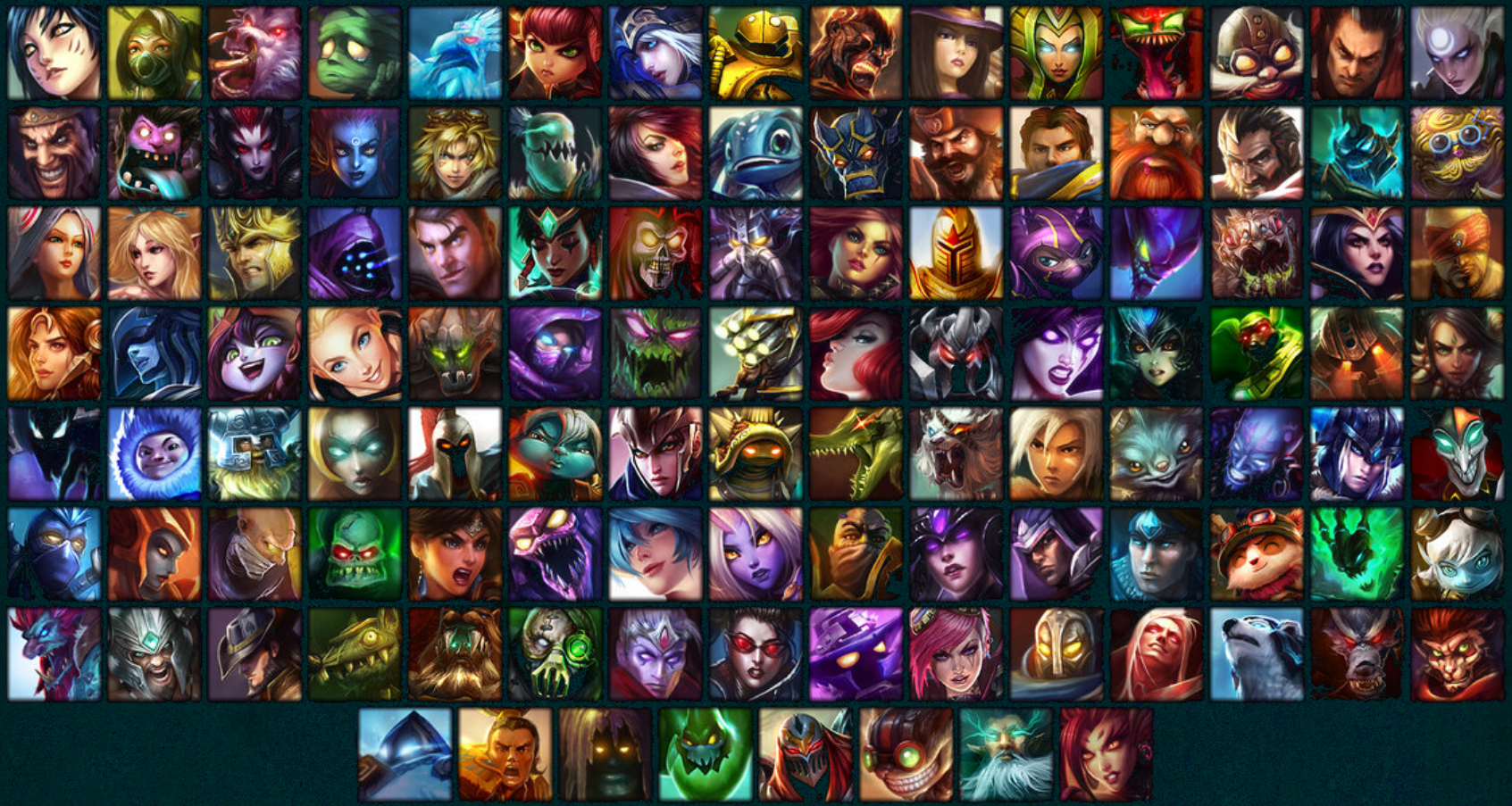
And it wasn't even close

By Vlad Savov | @vladsavov | Aug 6, 2018, 6:41am EDT

SHARE

[www.theverge.com]

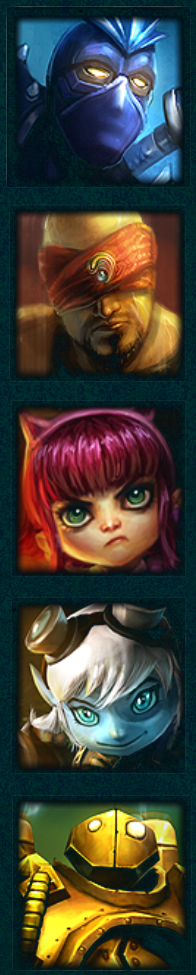
Modeling teams in online games



[euw.leagueoflegends.com]

Modeling teams in online games

Team 1



VS

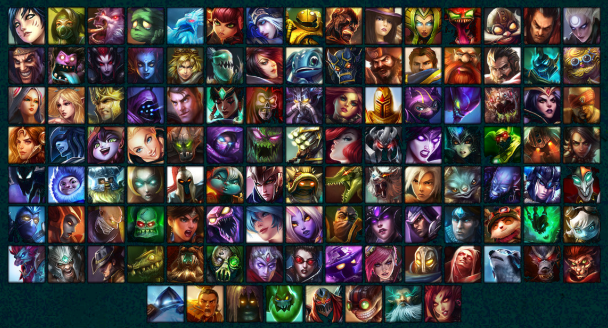
Team 2

Modeling teams in online games

Discrete probabilistic models

Ground set

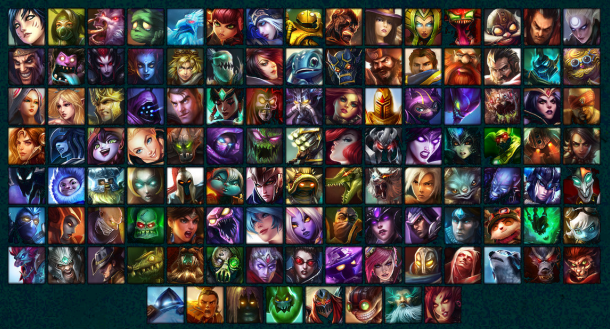
$$V = \{1, \ldots, n\}$$



Discrete probabilistic models

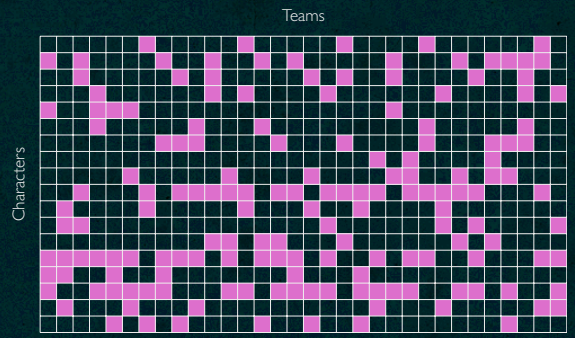
Ground set

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



Data

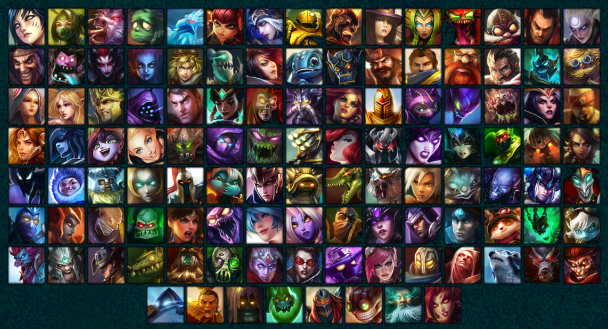
$$\mathcal{D} = \{S_i\}_{i=0}^{m}, S_i \subseteq V$$



Discrete probabilistic models

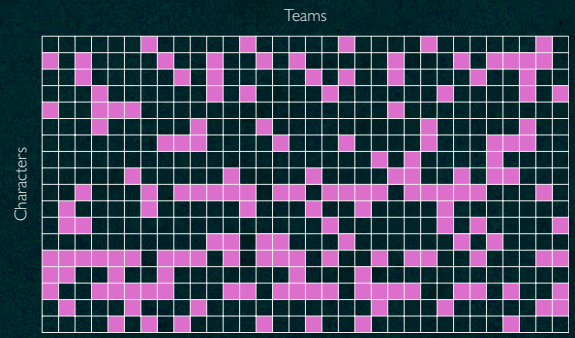
Ground set

$$V = \{1, \dots, n\}$$



Data

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



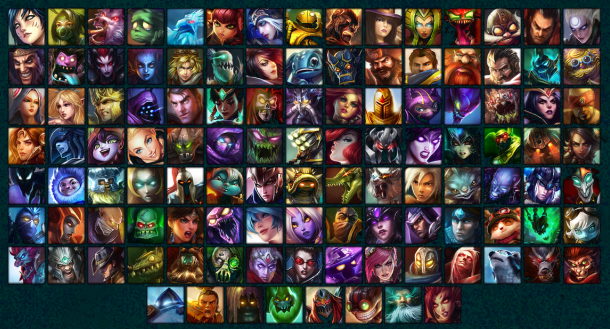
Model higher-order interactions

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

Discrete probabilistic models

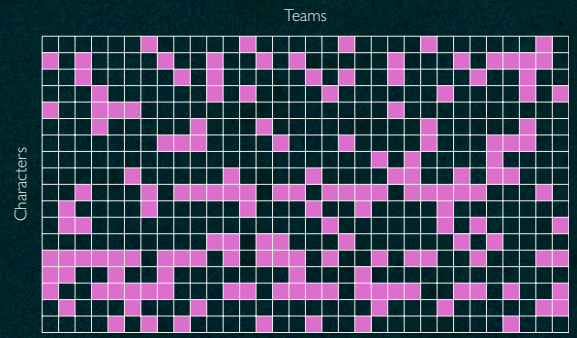
Ground set

$$V = \{1, \dots, n\}$$



Data

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



Model higher-order interactions

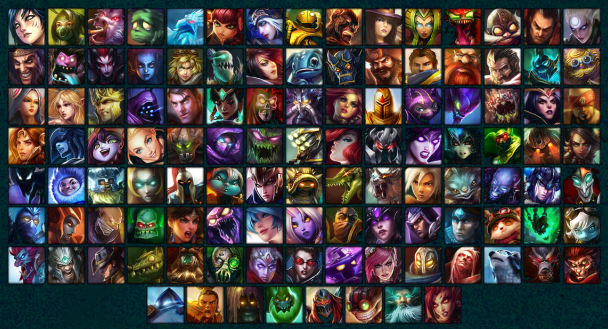
$$p(S;\theta) = \frac{1}{Z(\theta)} \exp\big(F(S;\theta)\big)$$

- $F(S) = \text{graph-cut}(S) \rightarrow$ Ising model

Discrete probabilistic models

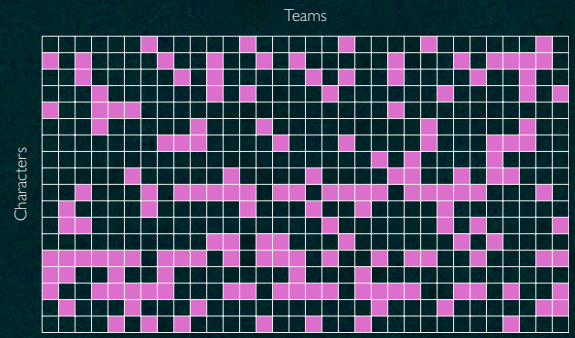
Ground set

$$V = \{1, \dots, n\}$$



Data

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



Model higher-order interactions

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

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

Discrete probabilistic models

Learn θ

Discrete probabilistic models

Learn θ

Max. likelihood

Discrete probabilistic models

Learn θ

Max. likelihood

Compute $\nabla_\theta Z(\theta)$

Discrete probabilistic models

Learn θ

Max. likelihood

Compute $\nabla_\theta Z(\theta)$

#P-hard in general

Discrete probabilistic models

Learn θ

Max. likelihood

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

#P-hard in general

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

Discrete probabilistic models

Learn θ

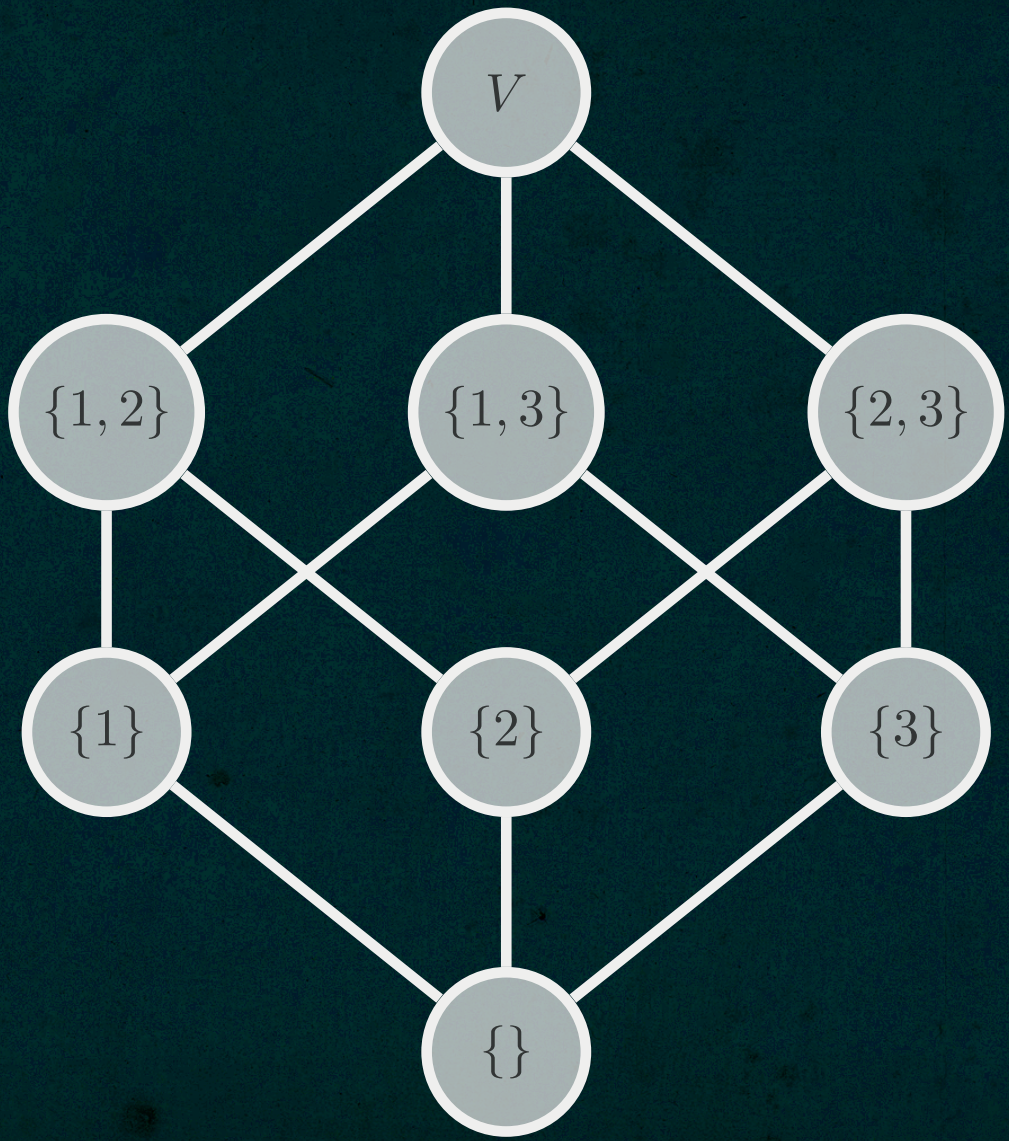
Max. likelihood

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

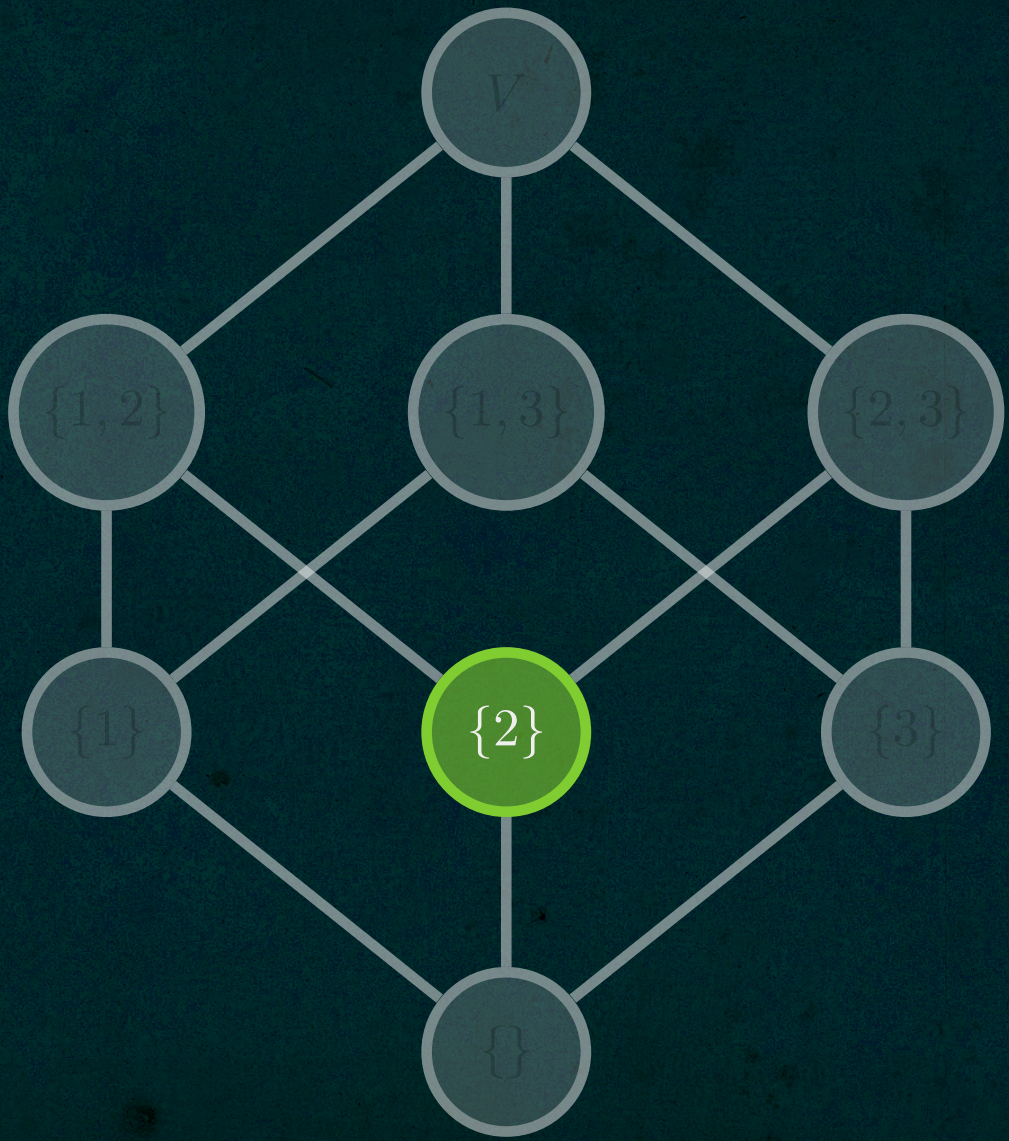
#P-hard in general

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

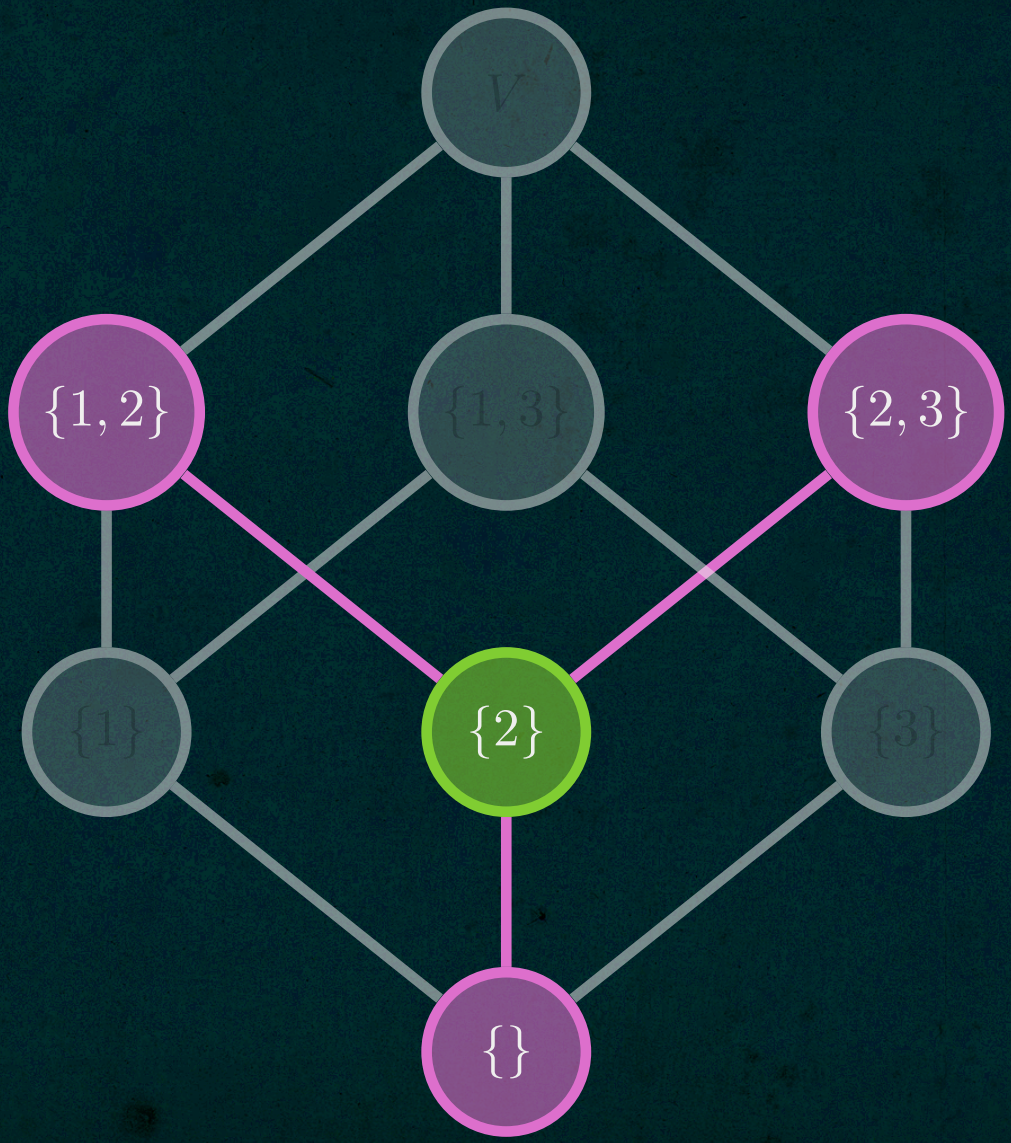
The Gibbs sampler



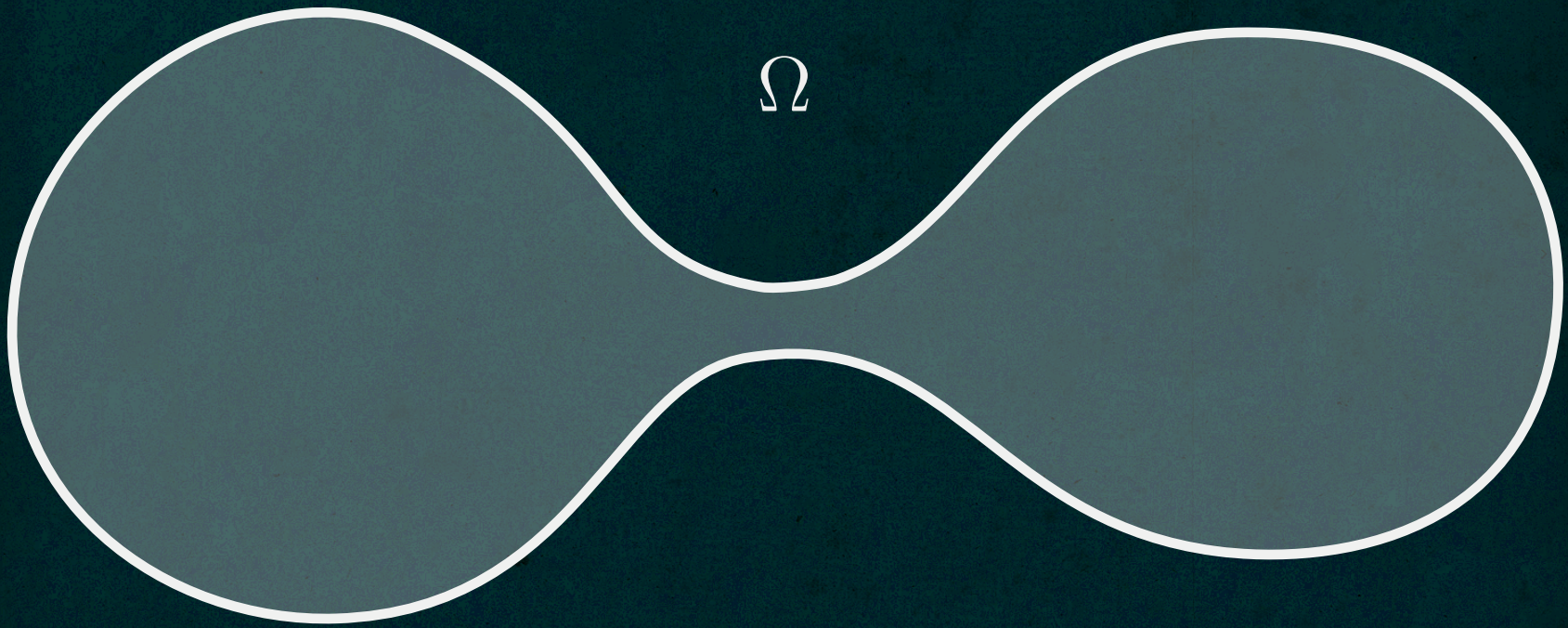
The Gibbs sampler



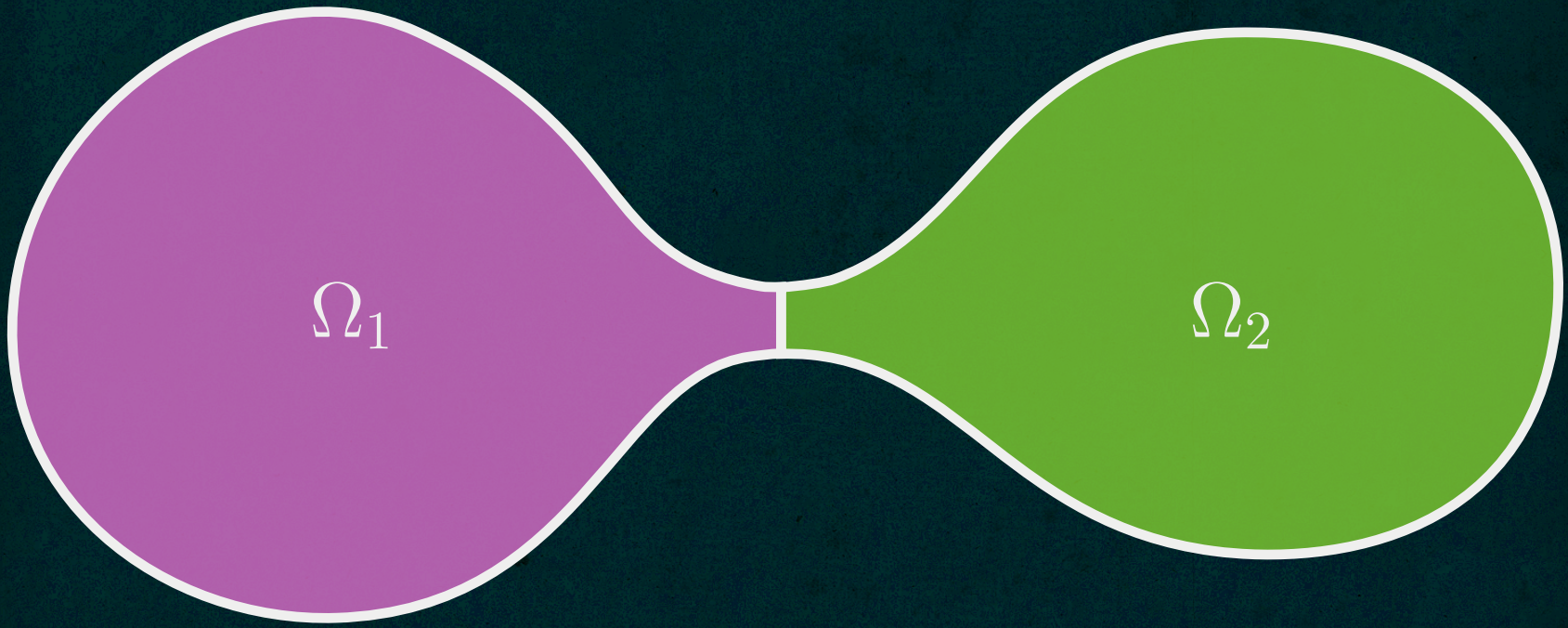
The Gibbs sampler



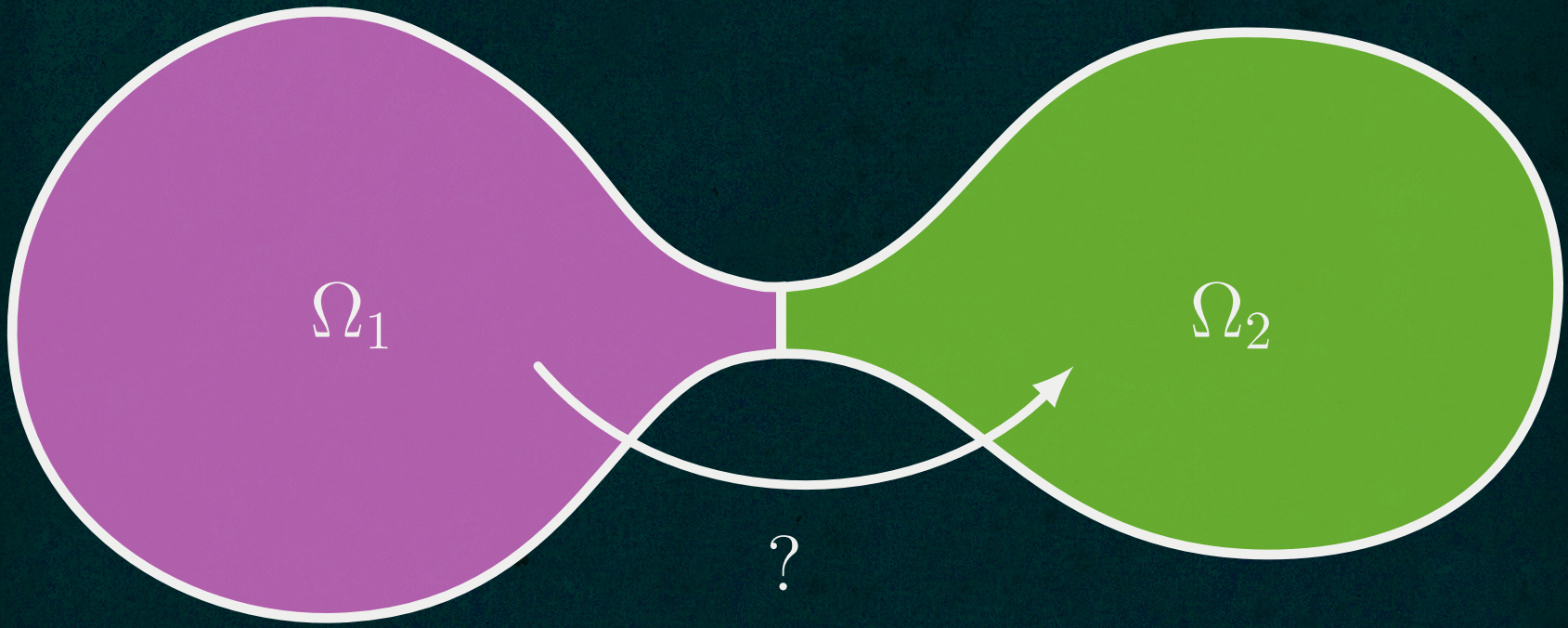
When Gibbs fails



When Gibbs fails



When Gibbs fails



The M^3 chain

$\rightarrow$ M^3 = Mixture of Log-Modulars Metropolis

The M^3 chain

$\rightarrow$ M^3 = Mixture of Log-Modulars Metropolis

1. Mixture

---

2. Log-Modulars

---

3. Metropolis

The M^3 chain

$\rightarrow$ M^3 = Mixture of Log-Modulars Metropolis

1. Mixture

---

2. Log-Modulars

---

3. Metropolis

The M^3 chain

$\rightarrow$ M^3 = Mixture of Log-Modulars Metropolis

1. Mixture

---

2. Log-Modulars

---

3. Metropolis
   - Target $p(S) \propto \exp(F(S))$

The M^3 chain

$\rightarrow$ M^3 = Mixture of Log-Modulars Metropolis

1. Mixture

---

2. Log-Modulars

---

3. Metropolis
   - Target $p(S) \propto \exp(F(S))$
   - Proposal $q(S, T)$

The M^3 chain

$\rightarrow$ M^3 = Mixture of Log-Modulars Metropolis

1. Mixture

---

2. Log-Modulars

---

3. Metropolis
   - Target $p(S) \propto \exp(F(S))$
   - Proposal $q(S, T)$
   - Accept with probability $\min\left\{1, \frac{p(T)q(T,S)}{p(S)q(S,T)}\right\}$

The M^3 chain

$\rightarrow$ M^3 = Mixture of Log-Modulars Metropolis

1. Mixture

---

2. Log-Modulars

---

3. Metropolis
   - Target $p(S) \propto \exp(F(S))$
   - Proposal $q(S, T)$
   - Accept with probability $\min\left\{1, \frac{p(T)q(T,S)}{p(S)q(S,T)}\right\}$

The M^3 chain

$\rightarrow$ M^3 = Mixture of Log-Modulars Metropolis

① Mixture $$q(S,T) = \frac{1}{Z_q} \sum_{i=1}^{r} w_i \exp\left(m_i(T)\right)$$

---

② Log-Modulars

---

③ Metropolis

- Target $p(S) \propto \exp(F(S))$
- Proposal $q(S,T)$
- Accept with probability $\min\left\{1, \frac{p(T)q(T,S)}{p(S)q(S,T)}\right\}$

The M^3 chain

$\rightarrow$ M^3 = Mixture of Log-Modulars Metropolis

① Mixture $$q(S,T) = \frac{1}{Z_q} \sum_{i=1}^{r} w_i \exp\left(m_i(T)\right)$$

---

② Log-Modulars

---

③ Metropolis

- Target $p(S) \propto \exp(F(S))$
- Proposal $q(S,T)$
- Accept with probability $\min\left\{1, \frac{p(T)q(T,S)}{p(S)q(S,T)}\right\}$

The M^3 chain

$\rightarrow$ M^3 = Mixture of Log-Modulars Metropolis

① Mixture $$q(S,T) = \frac{1}{Z_q} \sum_{i=1}^{r} w_i \exp\left(m_i(T)\right)$$

---

② Log-Modulars $$m_i(T) = \sum_{v \in T} m_{iv}$$

---

③ Metropolis

- Target $p(S) \propto \exp(F(S))$
- Proposal $q(S,T)$
- Accept with probability $\min\left\{1, \frac{p(T)q(T,S)}{p(S)q(S,T)}\right\}$

The M^3 chain

$\rightarrow$ M^3 = Mixture of Log-Modulars Metropolis

① Mixture $$q(T) = \frac{1}{Z_q} \sum_{i=1}^{r} w_i \exp\left(m_i(T)\right)$$

---

② Log-Modulars $$m_i(T) = \sum_{v \in T} m_{iv}$$

---

③ Metropolis

- Target $p(S) \propto \exp(F(S))$
- Proposal $q(T)$
- Accept with probability min $\left\{1, \frac{p(T)q(S)}{p(S)q(T)}\right\}$

The M^3 chain

Proposal $q(T) = \frac{1}{Z_q} \sum_{i=1}^{r} w_i \exp\left(m_i(T)\right)$

The M^3 chain

Proposal $q(T) = \frac{1}{Z_q} \sum_{i=1}^{r} w_i \exp\left(m_i(T)\right)$

$\rightarrow$ Can sample from q in $\mathcal{O}(n)$ time

The M^3 chain

Proposal $q(T) = \frac{1}{Z_q} \sum_{i=1}^{r} w_i \exp\left(m_i(T)\right)$

$\rightarrow$ Can sample from q in $\mathcal{O}(n)$ time

Proposition 1

Mixture q can approximate any distribution p arbitrarily well.

The M^3 chain

Proposal $q(T) = \frac{1}{Z_q} \sum_{i=1}^{r} w_i \exp\left(m_i(T)\right)$

$\rightarrow$ Can sample from q in $\mathcal{O}(n)$ time

Proposition 1

Mixture q can approximate any distribution p arbitrarily well.

BUT

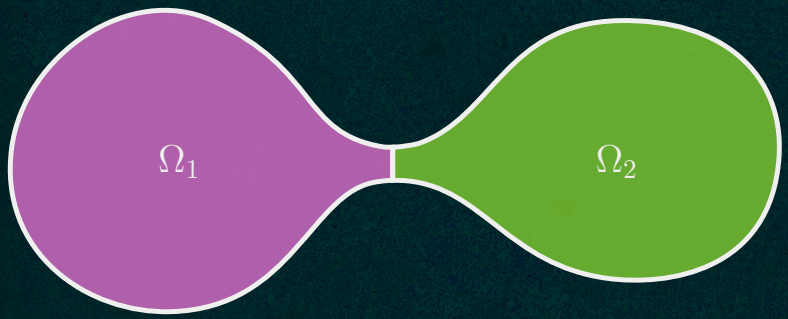
May need an exponential (in n) number of components r

The combined chain

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

The combined chain

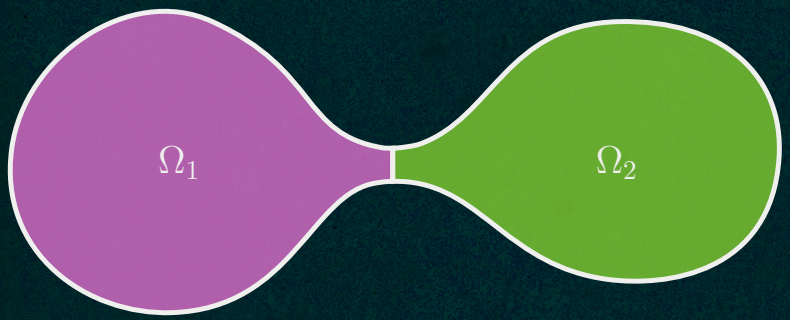
Gibbs step with probability α | M^3 step with prob. $1 - \alpha$



The combined chain

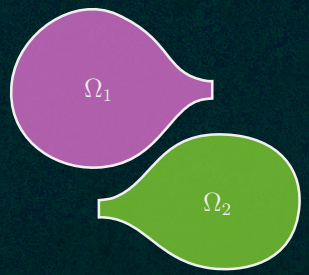
Gibbs step with probability α | M^3 step with prob. $1 - \alpha$

Decomposition theorem [Jerrum et al., '04]



Projection chain

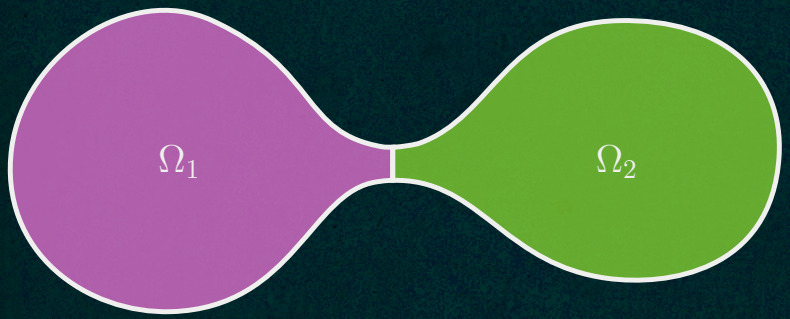
Restriction chains



The combined chain

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

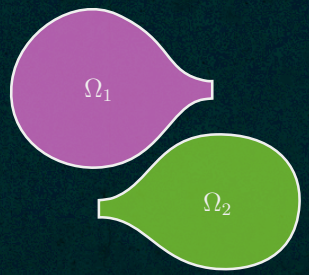
Decomposition theorem [Jerrum et al., '04]



Projection chain

Gibbs ✗
M^3 ✓

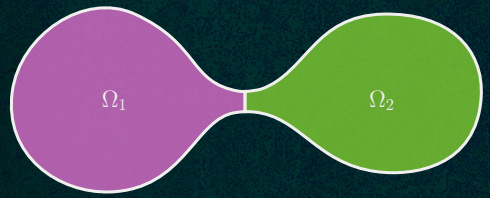
Restriction chains

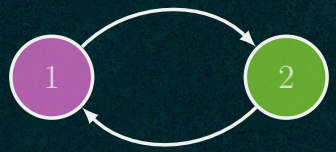


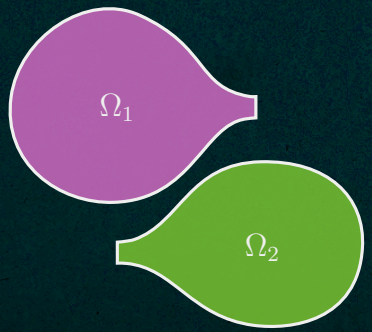
Gibbs ✓
M^3 ✗

The combined chain

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

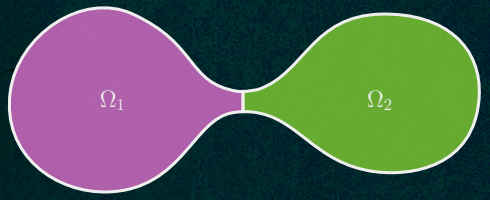




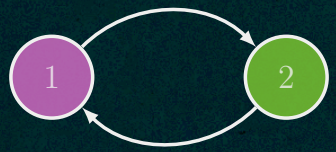


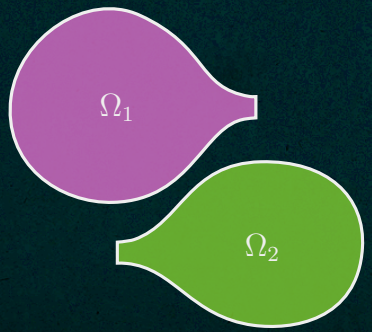
The combined chain

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



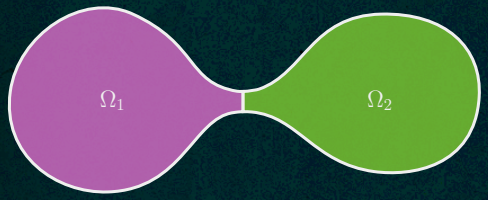
- Gibbs : $t_{\text{mix}} = \Omega\left(e^{cn}\right)$ [Levin et al., '08]
- M^3 :
- Combo :



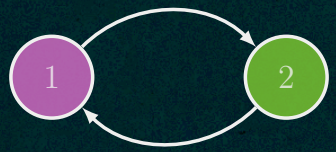


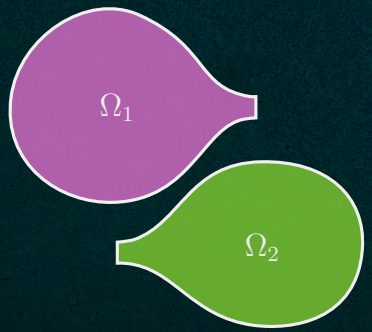
The combined chain

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



- Gibbs : $t_{\text{mix}} = \Omega\left(e^{cn}\right)$ [Levin et al., '08]
- M^3 :
- Combo :

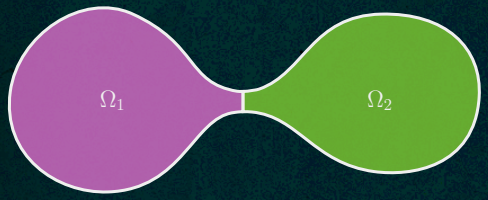




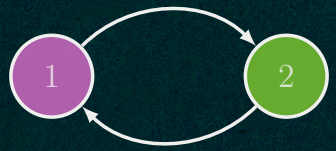
- Gibbs : $t_{\text{mix}} = \Theta\left(\frac{n^2}{\log n}\right)$ [Ding et al., '09]

The combined chain

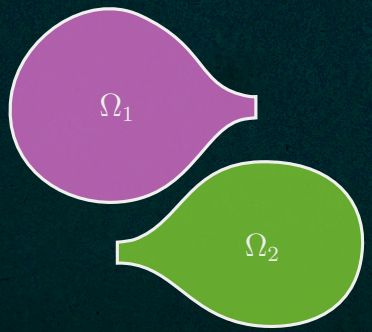
Class of Ising models on the complete graph (Curie-Weiss)



- Gibbs : $t_{\text{mix}} = \Omega\left(e^{cn}\right)$ [Levin et al., '08]
- M^3 :
- Combo :



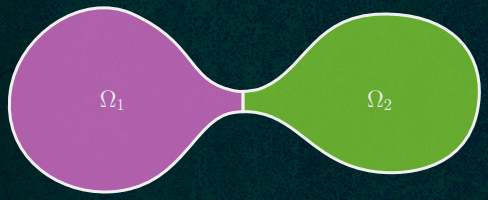
- M^3 : $t_{\text{mix}} = \mathcal{O}(1)$ [Lemma 1]



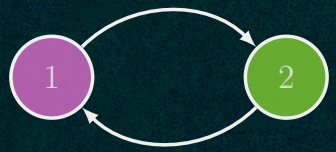
- Gibbs : $t_{\text{mix}} = \Theta\left(\frac{n^2}{\log n}\right)$ [Ding et al., '09]

The combined chain

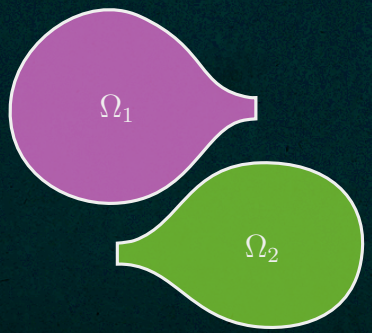
Class of Ising models on the complete graph (Curie-Weiss)



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



- M^3 : $t_{\text{mix}} = \mathcal{O}(1)$ [Lemma 1]



- Gibbs : $t_{\text{mix}} = \Theta\left(\frac{n^2}{\log n}\right)$ [Ding et al., '09]

Constructing the mixture

Construction of $q(\cdot) \propto \sum_{i=1}^{r} \exp\left(m_i(\cdot)\right)$

Constructing the mixture

Construction of $q(\cdot) \propto \sum_{i=1}^{r} \exp\left(m_i(\cdot)\right)$

Input: Set function F, mixture size r

for $i = 1$ **to** r **do**

$\sigma \leftarrow$ Permutation of V

$m_i \leftarrow$ Modular function that approximates F at σ

return $\{m_1, \ldots, m_r\}$

Constructing the mixture

Construction of $q(\cdot) \propto \sum_{i=1}^{r} \exp\left(m_i(\cdot)\right)$

Input: Set function F, mixture size r

for $i = 1$ **to** r **do**

$\sigma \leftarrow$ Permutation of V

$m_i \leftarrow \textsc{SemiGradient}(F, \sigma)$

return $\{m_1, \ldots, m_r\}$

Constructing the mixture

Construction of $q(\cdot) \propto \sum_{i=1}^{r} \exp\left(m_i(\cdot)\right)$

Input: Set function F, mixture size r

for $i = 1$ **to** r **do**

$\sigma \leftarrow$ Permutation of V

$m_i \leftarrow$ SemiGradient(F, σ)

return $\{m_1, \ldots, m_r\}$

- Submodularity $\rightarrow$ natural diminishing returns property

Constructing the mixture

Construction of $q(\cdot) \propto \sum_{i=1}^{r} \exp\left(m_i(\cdot)\right)$

Input: Set function F, mixture size r

for $i = 1$ **to** r **do**

$\sigma \leftarrow$ Permutation of V

$m_i \leftarrow \textsc{SemiGradient}(F, \sigma)$

return $\{m_1, \ldots, m_r\}$

- Submodularity $\rightarrow$ natural diminishing returns property
- Sub-/supergradients $\rightarrow$ modular lower/upper approx. [Iyer et al., '13]

Constructing the mixture

Construction of $q(\cdot) \propto \sum_{i=1}^{r} \exp\left(m_i(\cdot)\right)$

Input: Set function F, mixture size r

for $i = 1$ **to** r **do**

$\sigma \leftarrow$ Permutation of V

$m_i \leftarrow \textsc{SemiGradient}(F, \sigma)$

return $\{m_1, \ldots, m_r\}$

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

Constructing the mixture

Randomized construction of $q(\cdot) \propto \sum_{i=1}^{r} \exp\left(m_i(\cdot)\right)$

Input: Set function F, mixture size r

for $i = 1$ **to** r **do**

$\sigma \leftarrow$ Random permutation of V

$m_i \leftarrow \textsc{SemiGradient}(F, \sigma)$

return $\{m_1, \ldots, m_r\}$

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

Constructing the mixture

Iterative construction of $q(\cdot) \propto \sum_{i=1}^{r} \exp\left(m_i(\cdot)\right)$

Input: Set function F, mixture size r

for $i = 1$ **to** r **do**

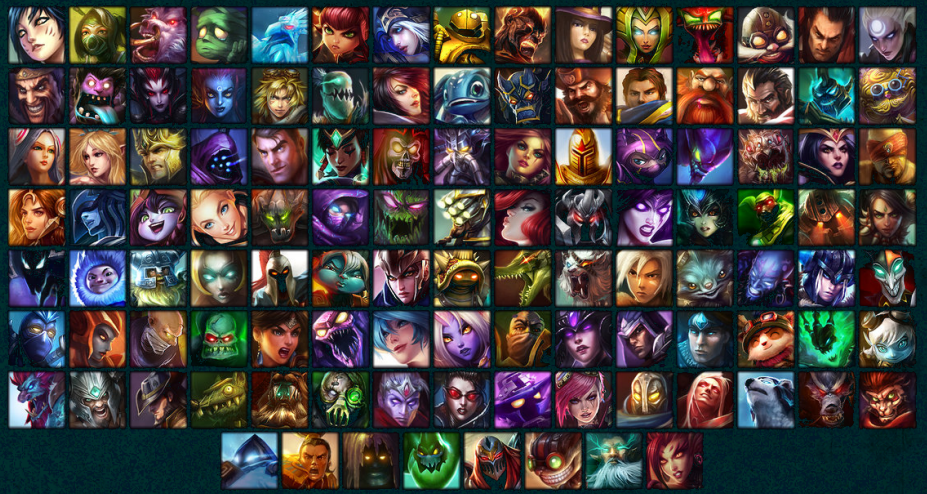
$\sigma \leftarrow \textsc{Greedy}\Big(F(\cdot) - \log \sum_{j=1}^{i-1} \exp(m_j(\cdot))\Big)$

$m_i \leftarrow \textsc{SemiGradient}(F, \sigma)$

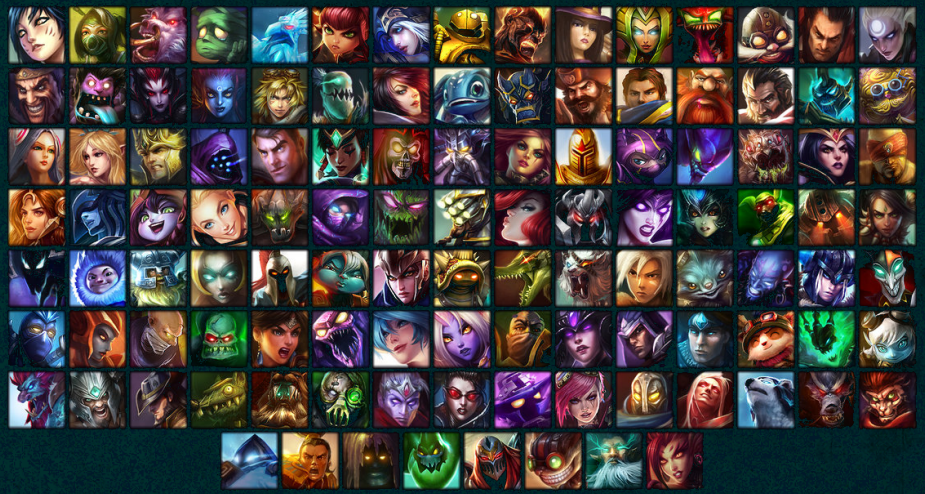
return $\{m_1, \ldots, m_r\}$

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

Experiments

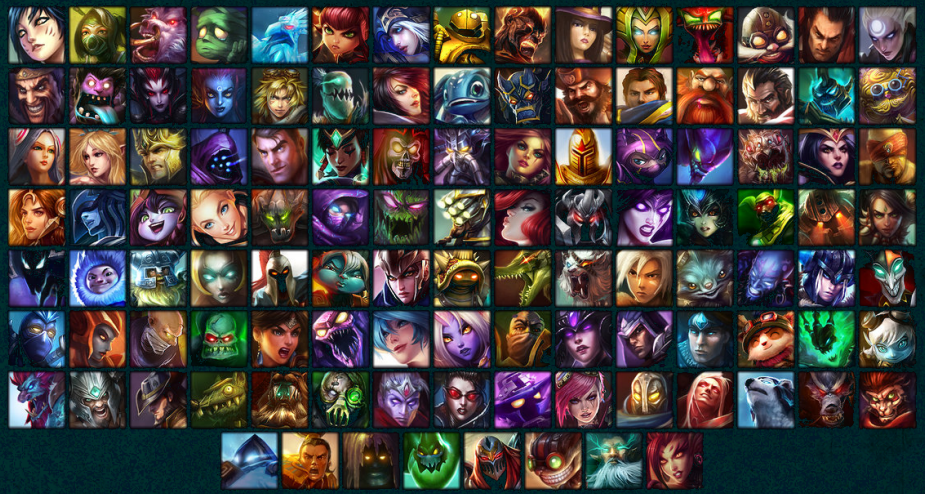


Experiments



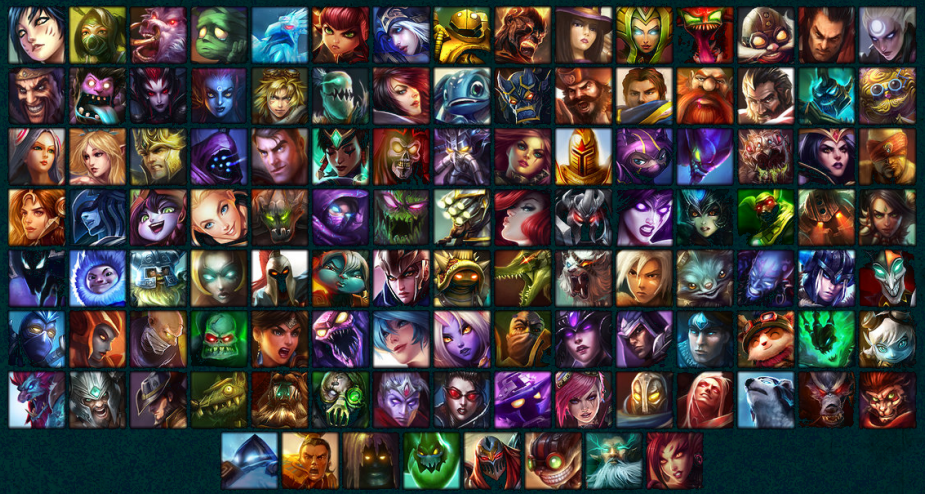
- $|V| = 48$

Experiments



- $|V| = 48$
- 8.5k teams of 5 characters

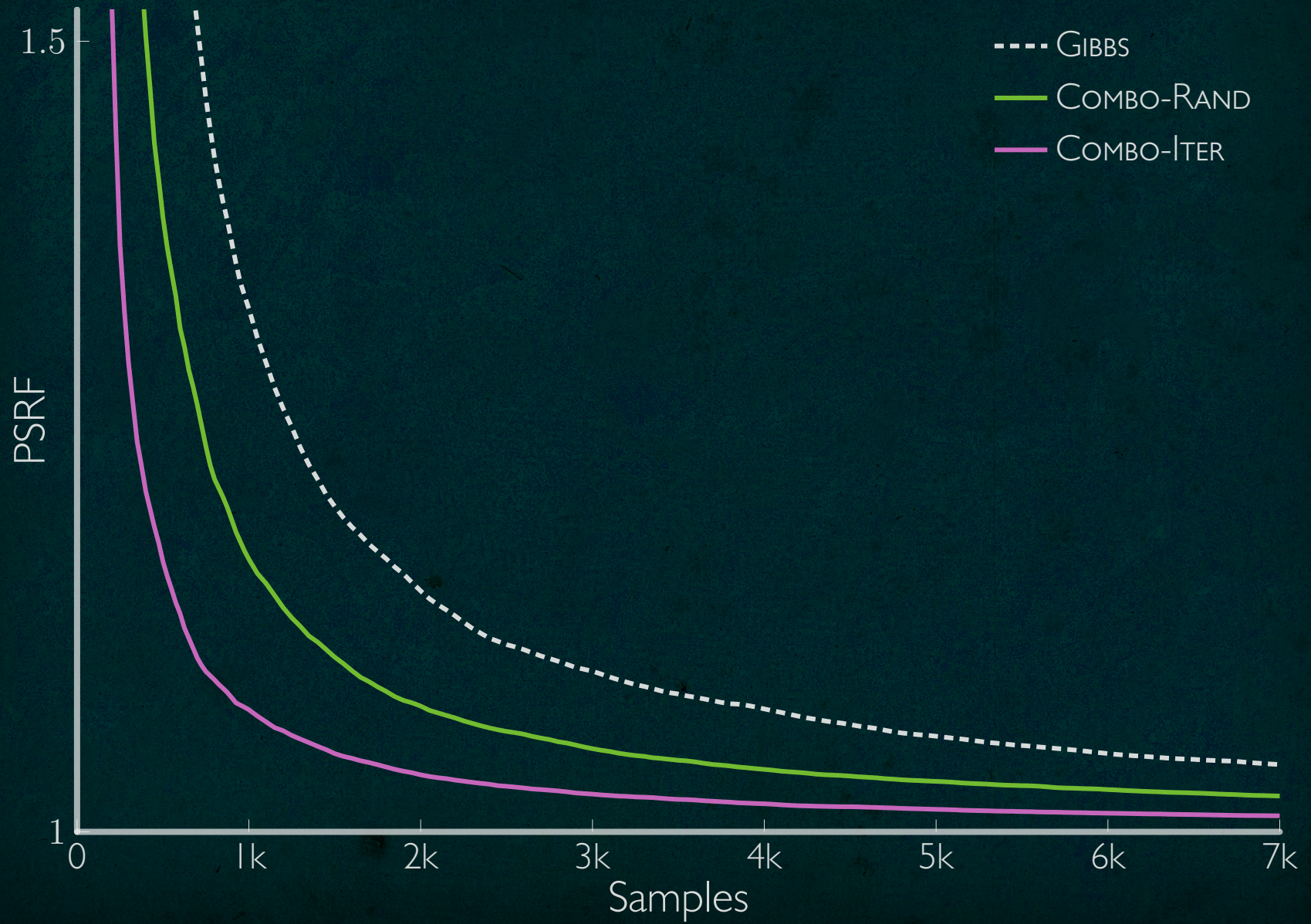
Experiments



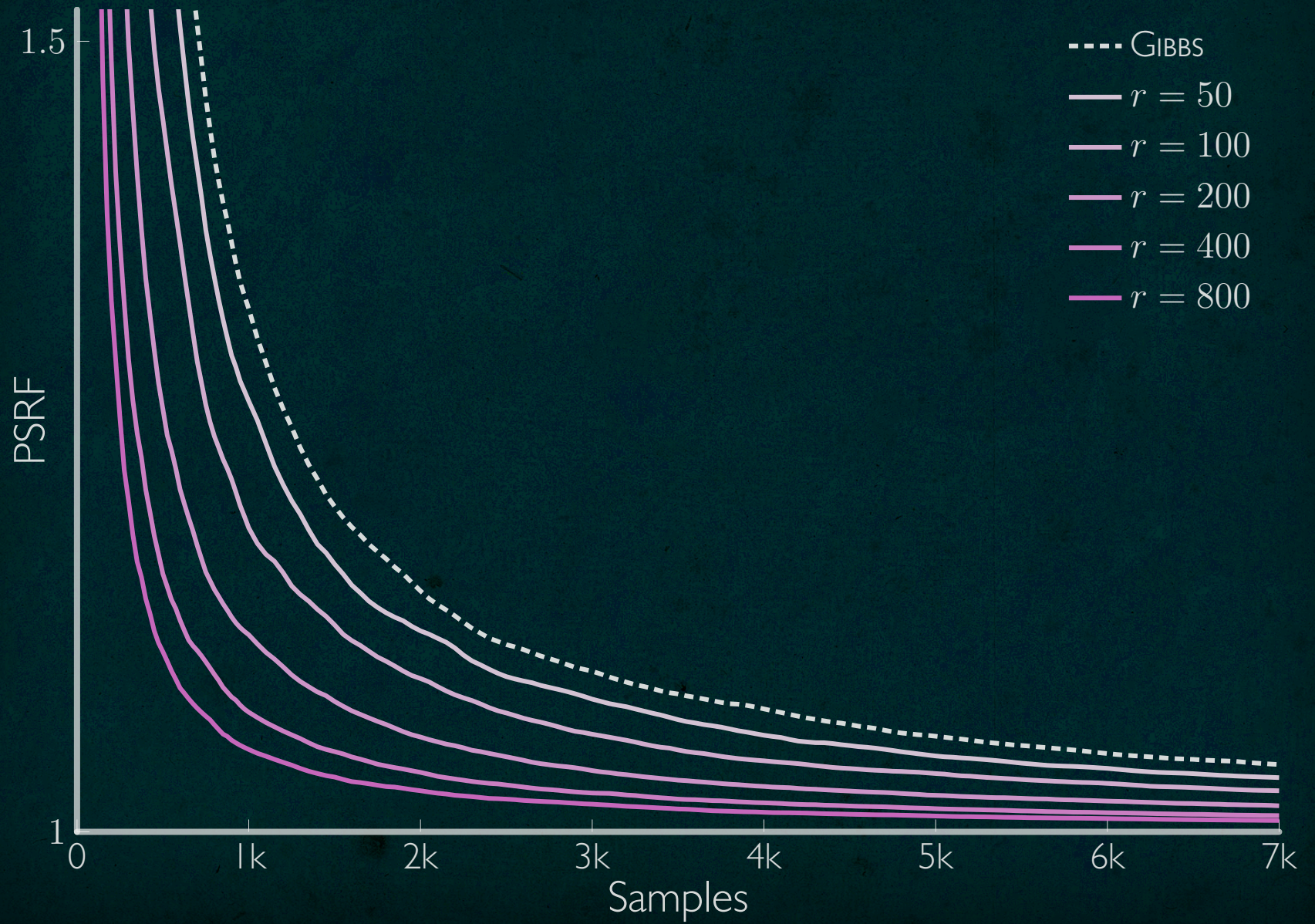
- $|V| = 48$
- 8.5k teams of 5 characters
- F is a (submodular) facility location diversity model [Tschiatschek et al., '16]

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

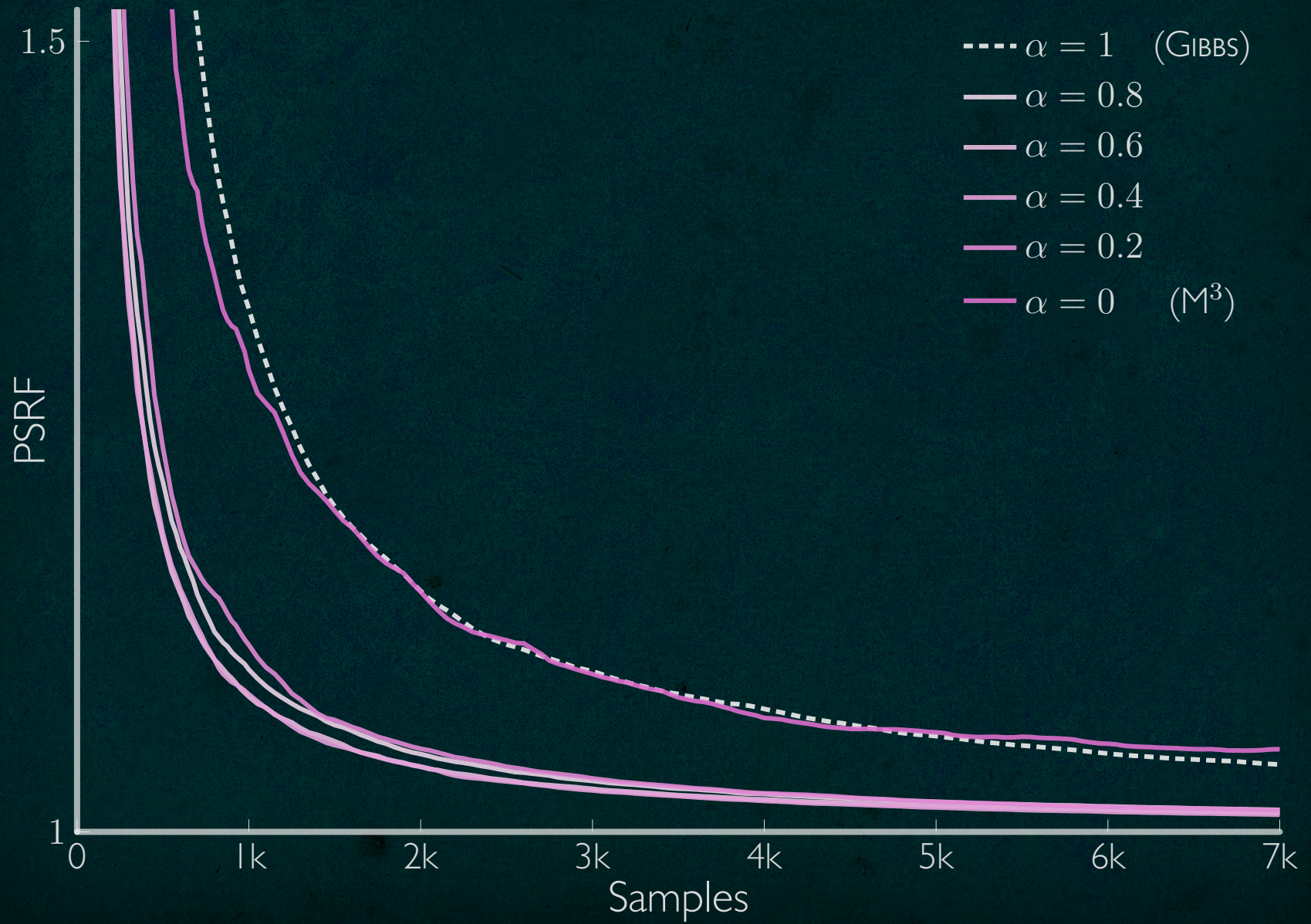
Experiments



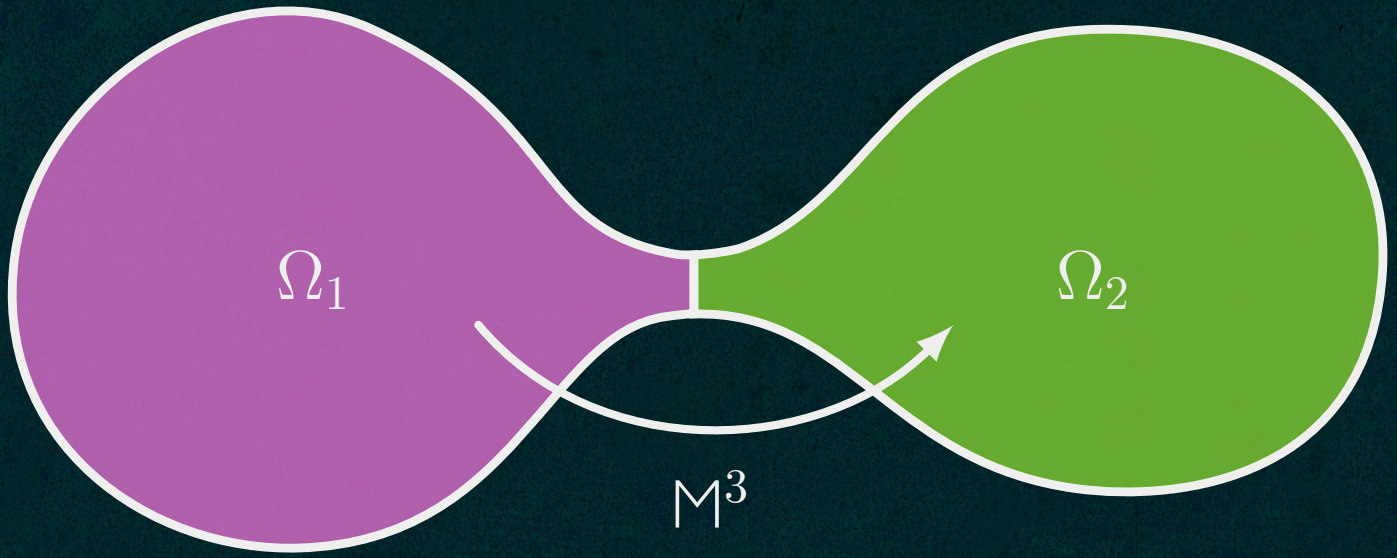
Experiments



Experiments

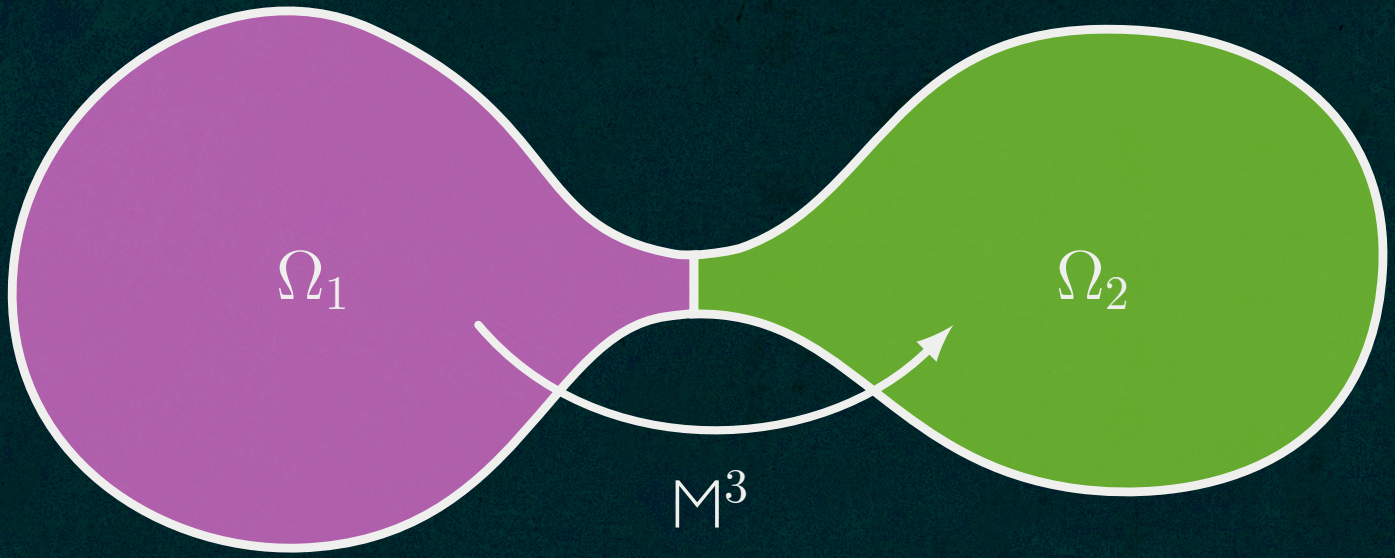


Conclusion



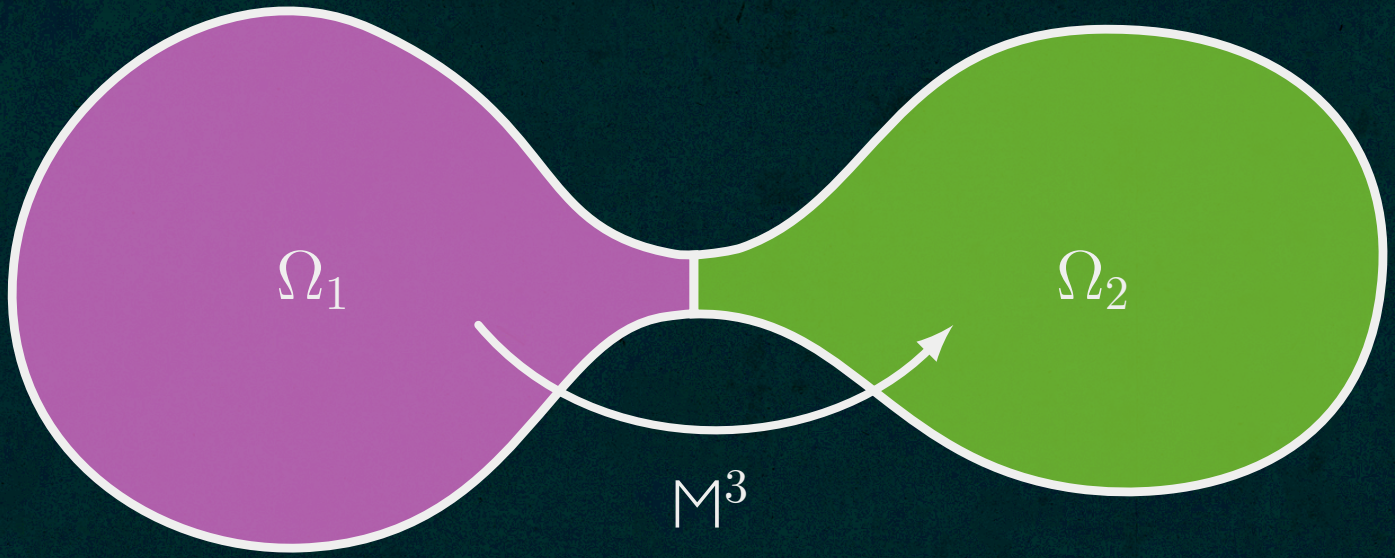
- M^3 sampler $\rightarrow$ propose global moves to overcome bottlenecks

Conclusion



- M^3 sampler $\rightarrow$ propose global moves to overcome bottlenecks
- Combined sampler $\rightarrow$ analysis based on decomposition theorem

Conclusion



- M^3 sampler $\rightarrow$ propose global moves to overcome bottlenecks
- Combined sampler $\rightarrow$ analysis based on decomposition theorem
- Semigradient construction $\rightarrow$ incorporate ideas from optimization