Sampling from Probabilistic Submodular Models

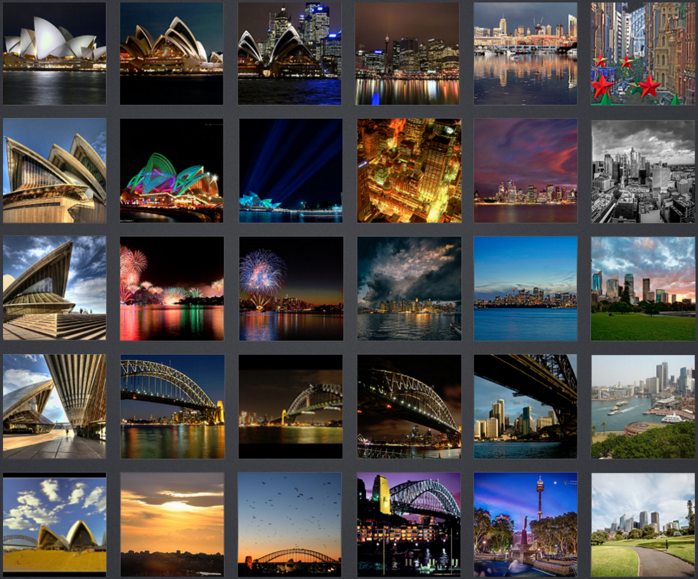
Alkis Gotovos

Hamed Hassani

Andreas Krause

ETH zürich

Image Collection Summarization



Submodularity

- Facility location objective [Lin and Bilmes, '12] [Tschiatschek et al., '14]

Submodularity

- Facility location objective [Lin and Bilmes, '12] [Tschiatschek et al., '14]
- Encourage coverage and diversity of the summary

Submodularity

- Facility location objective [Lin and Bilmes, '12] [Tschiatschek et al., '14]
- Encourage coverage and diversity of the summary
- Set of all images V

Submodularity

- Facility location objective [Lin and Bilmes, '12] [Tschiatschek et al., '14]
- Encourage coverage and diversity of the summary
- Set of all images V
- For any summary $S \subseteq V \quad \longrightarrow \quad F(S) \in \mathbb{R}$

Submodularity

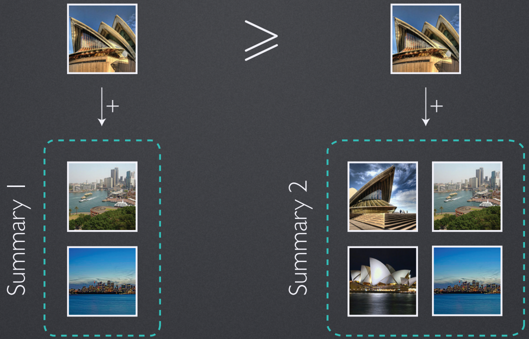
- Facility location objective [Lin and Bilmes, '12] [Tschiatschek et al., '14]
- Encourage coverage and diversity of the summary
- Set of all images V
- For any summary $S \subseteq V \quad \longrightarrow \quad F(S) \in \mathbb{R}$

Submodularity

- Facility location objective [Lin and Bilmes, '12] [Tschiatschek et al., '14]
- Encourage coverage and diversity of the summary
- Set of all images V
- For any summary $S \subseteq V \quad \longrightarrow \quad F(S) \in \mathbb{R}$

Submodularity

- Facility location objective [Lin and Bilmes, '12] [Tschiatschek et al., '14]
- Encourage coverage and diversity of the summary
- Set of all images V
- For any summary $S \subseteq V \quad \longrightarrow \quad F(S) \in \mathbb{R}$



Submodularity

F is submodular $\iff$ $-F$ is supermodular

$\downarrow$ $\downarrow$

$\boxed{\text{coverage / diversity}}$ $\boxed{\text{smoothness / cooperation}}$

Submodularity

F is submodular $\iff$ $-F$ is supermodular

$\downarrow$ $\downarrow$

$\boxed{\text{coverage / diversity}}$ $\boxed{\text{smoothness / cooperation}}$

- Submodular optimization is well-studied

Submodularity

F is submodular $\iff$ $-F$ is supermodular

$\downarrow$ $\downarrow$

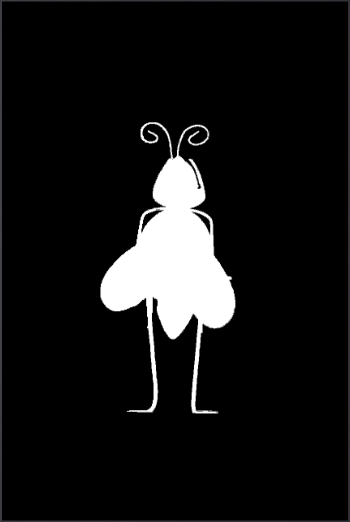
$\boxed{\text{coverage / diversity}}$ $\boxed{\text{smoothness / cooperation}}$

- Submodular optimization is well-studied
- Little existing work on probabilistic models

Sampling Summaries

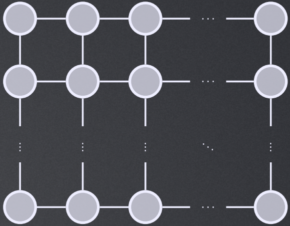
Foreground / Background Segmentation

Foreground / Background Segmentation

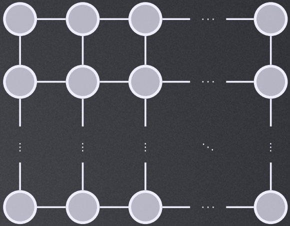


Foreground / Background Segmentation

Foreground / Background Segmentation

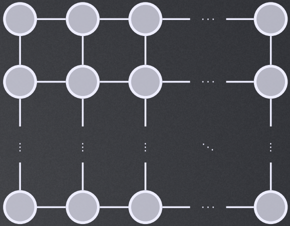


Foreground / Background Segmentation



- Set of all pixels V

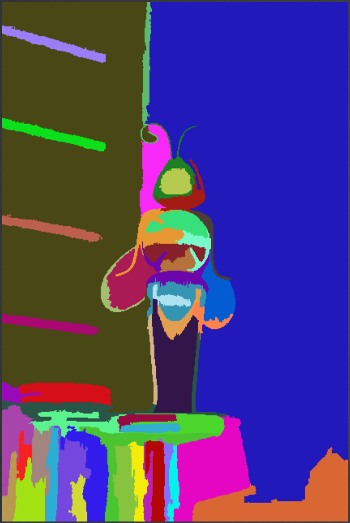
Foreground / Background Segmentation



- Set of all pixels V
- For $S \subseteq V$ of foreground pixels,

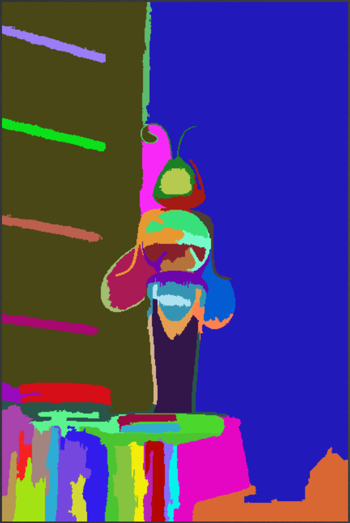
$$p(S) \propto \exp\left(\sum_{v \sim w} F_{v,w}(S)\right)$$

Higher-order Models



Superpixel potentials [Kohli et al., '08]

Higher-order Models



$$V = V_1 \cup V_2 \cup \cdots \cup V_L$$

Superpixel potentials [Kohli et al., '08]

Higher-order Models

$$V = V_1 \cup V_2 \cup \cdots \cup V_L$$

Superpixel potentials [Kohli et al., '08]

Higher-order Models

$$V = V_1 \cup V_2 \cup \cdots \cup V_L$$

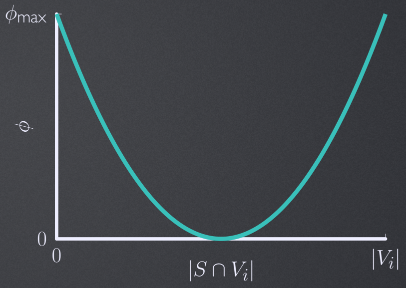
$$F_i(S) = \phi\left(|S \cap V_i|\right)$$

Superpixel potentials [Kohli et al., '08]

Higher-order Models

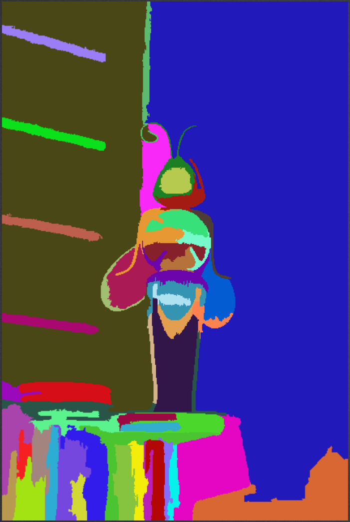
$$V = V_1 \cup V_2 \cup \cdots \cup V_L$$

$$F_i(S) = \phi\left(|S \cap V_i|\right)$$



Superpixel potentials [Kohli et al., '08]

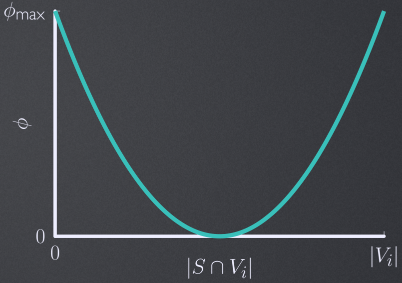
Higher-order Models



Superpixel potentials [Kohli et al., '08]

$$V = V_1 \cup V_2 \cup \cdots \cup V_L$$

$$F_i(S) = \phi\left(|S \cap V_i|\right)$$



$$p(S) \propto \exp\left(\sum_{i=1}^{L} F_i(S)\right)$$

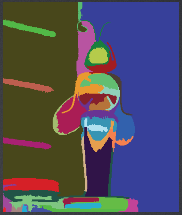
Higher-order Models [Djolonga and Krause, '15]

Pairwise

Higher-order

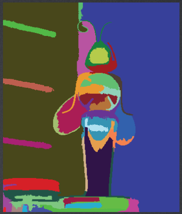
Probabilistic Submodular Models

Use submodular functions in probabilistic models



Equip existing models with higher-order interactions

Probabilistic Submodular Models

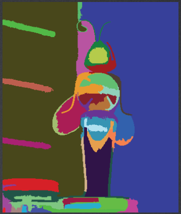


Use submodular functions in probabilistic models

Equip existing models with higher-order interactions

Probabilistic Submodular Models

Probabilistic Submodular Models



Use submodular functions in probabilistic models

Equip existing models with higher-order interactions

Probabilistic Submodular Models

$$\boxed{p(S) = \frac{1}{Z}\exp(F(S))}$$

$F : 2^V \to \mathbb{R}$ is a submodular or supermodular function

Probabilistic Submodular Models

PSMs	Markov Random Fields

Probabilistic Submodular Models

PSMs	Markov Random Fields
∘ Ground set V with $|V| = n$	

Probabilistic Submodular Models

PSMs	Markov Random Fields
∘ Ground set V with $\|V\| = n$	
∘ Sub- or supermodular function $F : 2^V \rightarrow \mathbb{R}$	

Probabilistic Submodular Models

PSMs	Markov Random Fields
◦ Ground set V with $\|V\| = n$	
◦ Sub- or supermodular function $F : 2^V \to \mathbb{R}$	
◦ Distribution over subsets $p(S) \propto \exp(F(S))$	

Probabilistic Submodular Models

PSMs	Markov Random Fields
◦ Ground set V with $\|V\| = n$	◦ Binary random vector $X = (X_1, \ldots, X_n)$
◦ Sub- or supermodular function $F : 2^V \to \mathbb{R}$	
◦ Distribution over subsets $p(S) \propto \exp(F(S))$	

Probabilistic Submodular Models

PSMs	Markov Random Fields
∘ Ground set V with $\lvert V \rvert = n$	∘ Binary random vector $X = (X_1, \ldots, X_n)$
∘ Sub- or supermodular function $F : 2^V \to \mathbb{R}$	∘ Set of factors $\phi_i : \{0,1\}^{\mathcal{C}_i} \to \mathbb{R}$
∘ Distribution over subsets $p(S) \propto \exp(F(S))$	

Probabilistic Submodular Models

PSMs	Markov Random Fields
◦ Ground set V with $\|V\| = n$	◦ Binary random vector $X = (X_1, \ldots, X_n)$
◦ Sub- or supermodular function $F : 2^V \to \mathbb{R}$	◦ Set of factors $\phi_i : \{0,1\}^{\mathcal{C}_i} \to \mathbb{R}$
◦ Distribution over subsets $p(S) \propto \exp(F(S))$	◦ Distribution over binary vectors $p(X) \propto \exp\left(\sum_i \phi_i(X_{\mathcal{C}_i})\right)$

Probabilistic Submodular Models

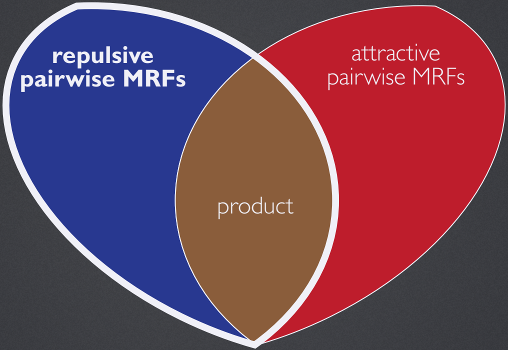
PSMs	Markov Random Fields
∘ Ground set V with $\lvert V\rvert = n$	∘ Binary random vector $X = (X_1, \ldots, X_n)$
∘ Sub- or supermodular function $F : 2^V \to \mathbb{R}$	∘ Set of factors $\phi_i : \{0,1\}^{\mathcal{C}_i} \to \mathbb{R}$
∘ Distribution over subsets $p(S) \propto \exp(F(S))$	∘ Distribution over binary vectors $p(X) \propto \exp\left(\sum_i \phi_i\left(X_{\mathcal{C}_i}\right)\right)$

$\boxed{\text{Model order: } \max_i \lvert\mathcal{C}_i\rvert}$

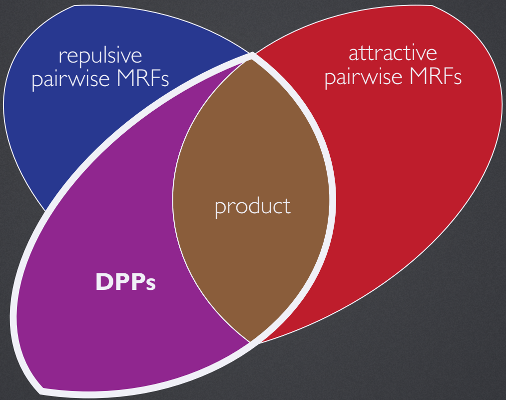
Landscape of Models

Landscape of Models

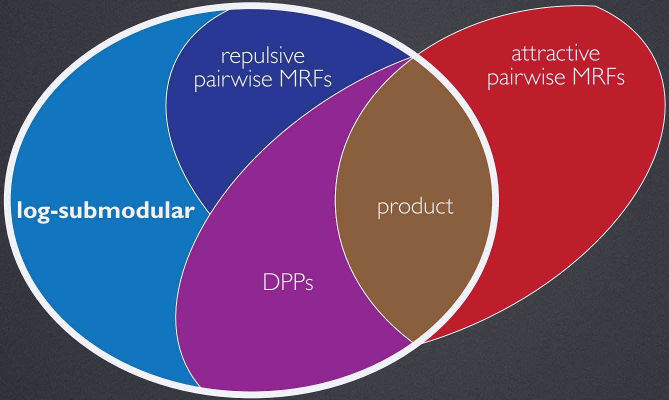
Landscape of Models



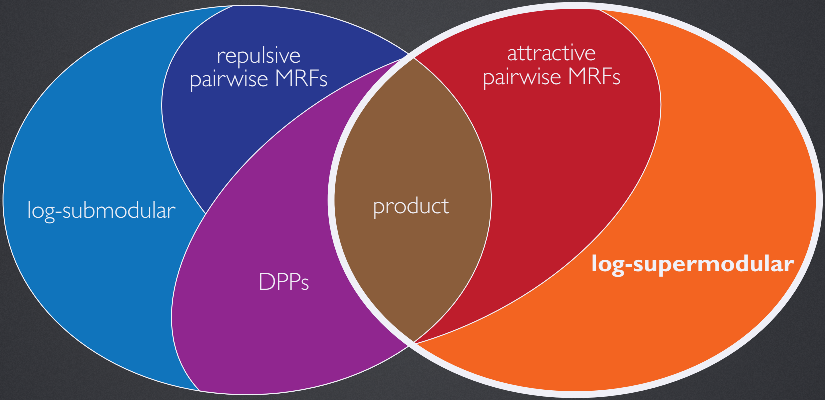
Landscape of Models



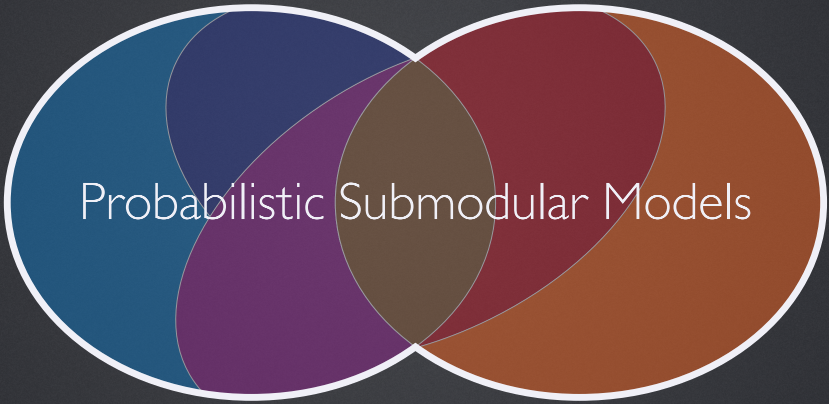
Landscape of Models



Landscape of Models



Landscape of Models



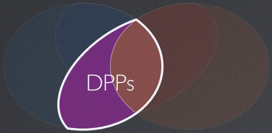
Inference

$\mathbb{P}(\text{pixel label})$

$\mathbb{P}(\text{image} \in \text{summary} \mid \text{selected})$

Inference

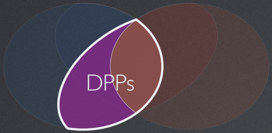
EXACT



- Tractable only for limited subclasses
- #P-hard even for Ising models

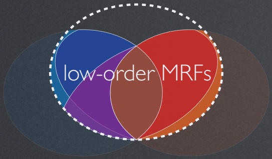
Inference

EXACT



- Tractable only for limited subclasses
- #P-hard even for Ising models

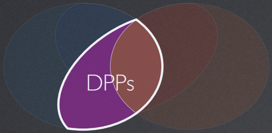
BP, MF, ...



- Extensively studied model class
- Complexity exponential in model order

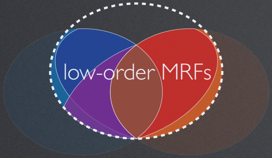
Inference

EXACT



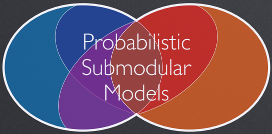
- Tractable only for limited subclasses
- #P-hard even for Ising models

BP, MF, …



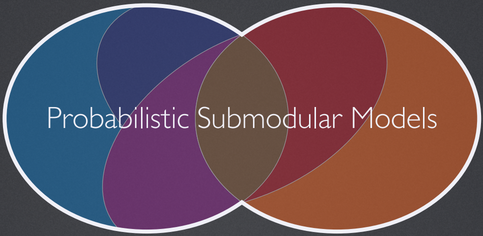
- Extensively studied model class
- Complexity exponential in model order

L-FIELD



- Variational approach for general PSMs
  [Djolonga and Krause, '14]

Inference



What about sampling?

Markov Chain Monte Carlo

- State space Ω
- Transition matrix P
- Stationary distribution π

Markov Chain Monte Carlo

- State space Ω
- Transition matrix P
- Stationary distribution π

Markov chain $(S_t)_{t \geq 0}$ that moves according to P

Markov Chain Monte Carlo

- State space Ω powerset of V
- Transition matrix P
- Stationary distribution π

Markov chain $(S_t)_{t \geq 0}$ that moves according to P

Markov Chain Monte Carlo

- State space Ω — powerset of V
- Transition matrix P — Gibbs sampler
- Stationary distribution π

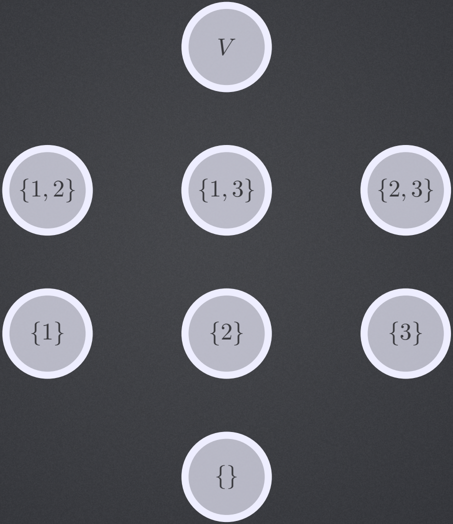
Markov chain $(S_t)_{t \geq 0}$ that moves according to P

Markov Chain Monte Carlo

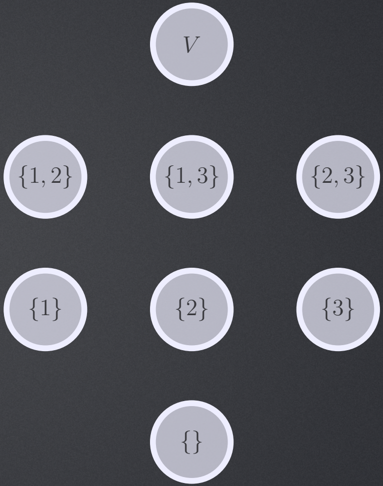
- State space Ω — powerset of V
- Transition matrix P — Gibbs sampler
- Stationary distribution π — PSM distribution

Markov chain $(S_t)_{t \geq 0}$ that moves according to P

State Space of $V = \{1, 2, 3\}$



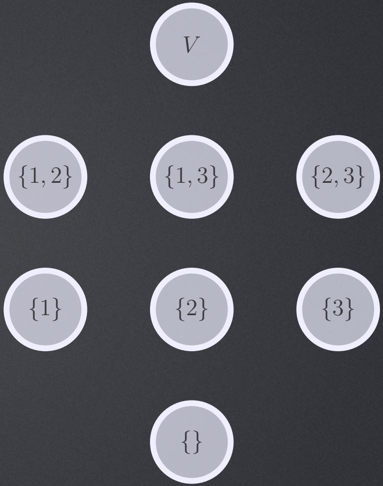
Gibbs Sampler



Gibbs Sampler

Start at S_0

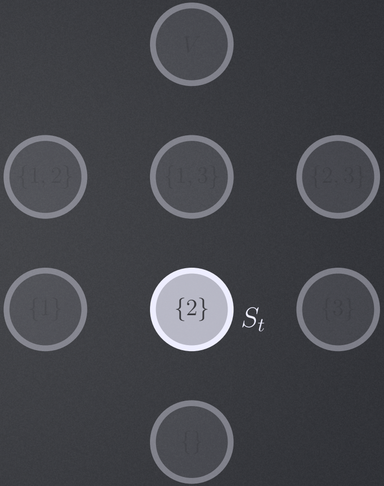
For $t = 1, 2, \ldots$



Gibbs Sampler

Start at S_0

For $t = 1, 2, \ldots$

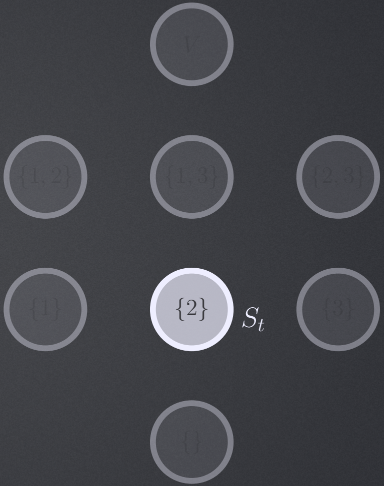


Gibbs Sampler

Start at S_0

For $t = 1, 2, \ldots$

- Select random $v \in V$

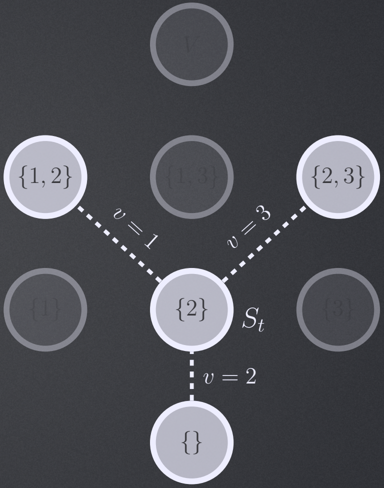


Gibbs Sampler

Start at S_0

For $t = 1, 2, \ldots$

- Select random $v \in V$

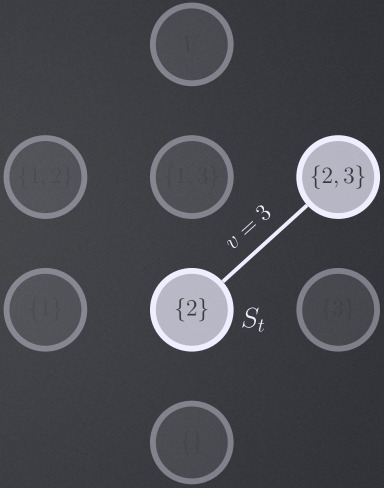


Gibbs Sampler

Start at S_0

For $t = 1, 2, \ldots$

- Select random $v \in V$

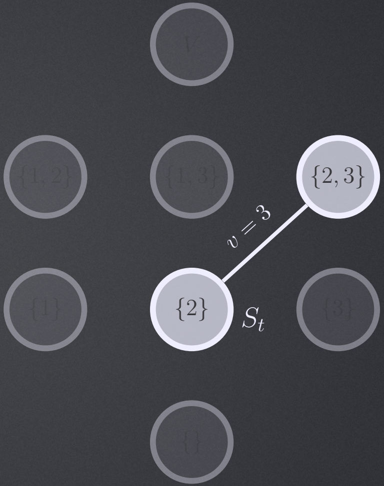


Gibbs Sampler

Start at S_0

For $t = 1, 2, \ldots$

- Select random $v \in V$
- Compute conditional p_{add}



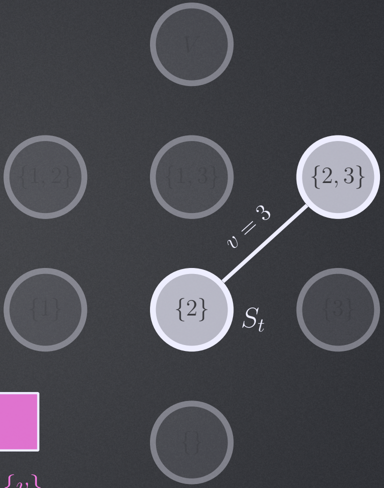
Gibbs Sampler

Start at S_0

For $t = 1, 2, \ldots$

- Select random $v \in V$
- Compute conditional p_{add}
- Flip biased coin

p_{add}	$1 - p_{\text{add}}$
$S_{t+1} \leftarrow S_t \cup \{v\}$	$S_{t+1} \leftarrow S_t \setminus \{v\}$



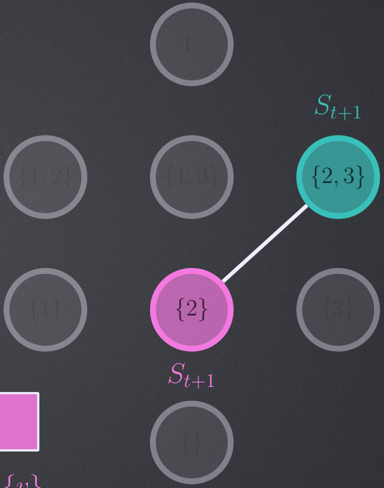
Gibbs Sampler

Start at S_0

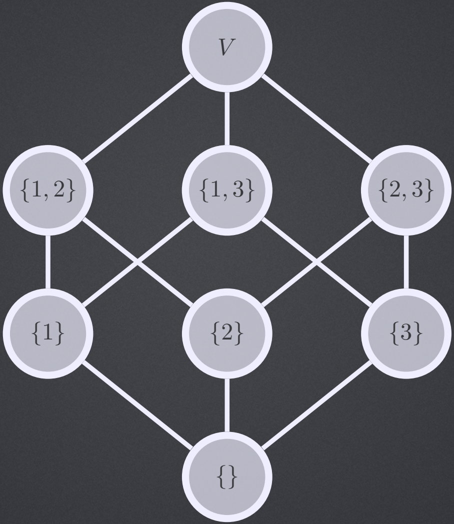
For $t = 1, 2, \ldots$

- Select random $v \in V$
- Compute conditional p_{add}
- Flip biased coin

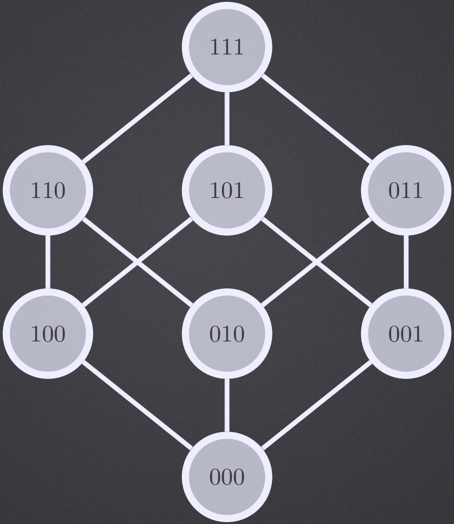
p_{add}	$1 - p_{\text{add}}$
$S_{t+1} \leftarrow S_t \cup \{v\}$	$S_{t+1} \leftarrow S_t \setminus \{v\}$



Gibbs Sampler



Gibbs Sampler



Mixing time

Does the Markov chain converge?

Mixing time

Does the Markov chain converge?

Total variation distance

$$d(t) = d_{\text{tv}}\left(\mathbb{P}_{S_t}, \pi\right)$$

Mixing time

Does the Markov chain converge?

Total variation distance

$$d(t) = \max\{d_{\text{tv}}\left(\mathbb{P}_{S_t}, \pi\right) \mid S_0 \in \Omega\}$$

Mixing time

Does the Markov chain converge?

Total variation distance

$$d(t) = \max\{d_{\text{tv}}\left(\mathbb{P}_{S_t}, \pi\right) \mid S_0 \in \Omega\}$$

Under mild assumptions (ergodicity), $d(t) \xrightarrow{t \to \infty} 0$

Mixing time

Does the Markov chain converge?

Total variation distance

$$d(t) = \max\{d_{\text{tv}}\,(\mathbb{P}_{S_t}, \pi) \mid S_0 \in \Omega\}$$

Under mild assumptions (ergodicity), $d(t) \xrightarrow{t \to \infty} 0$

How long does it take to get "close enough" to π?

Mixing time

Does the Markov chain converge?

Total variation distance

$$d(t) = \text{max}\{d_{\text{tv}}\left(\mathbb{P}_{S_t}, \pi\right) \mid S_0 \in \Omega\}$$

Under mild assumptions (ergodicity), $d(t) \xrightarrow{t \to \infty} 0$

How long does it take to get "close enough" to π?

Mixing time $\quad t_{\text{mix}}(\epsilon) = \text{min}\,\{t \mid d(t) \leq \epsilon\}$

Goal

- Mixing times for general PSMs are exponential in $|V| = n$

Goal

- Mixing times for general PSMs are exponential in $|V| = n$
- Exponential even for pairwise models [Jerrum and Sinclair, '93]

Goal

- Mixing times for general PSMs are exponential in $|V| = n$
- Exponential even for pairwise models [Jerrum and Sinclair, '93]

> We establish sufficient conditions for sub-exponential mixing of the Gibbs sampler on PSMs.

Polynomial-time Mixing

F is modular if $\quad F(A) + F(B) \quad = \quad F(A \cup B) + F(A \cap B)$

Polynomial-time Mixing

sub- $\geq$

F is modular if $\quad F(A) + F(B) = F(A \cup B) + F(A \cap B)$

Polynomial-time Mixing

sub- $\geq$

F is modular if $F(A) + F(B) = F(A \cup B) + F(A \cap B)$

super- $\leq$

Polynomial-time Mixing

$$F \text{ is } \begin{array}{c}\text{sub-}\\ \text{modular}\\ \text{super-}\end{array} \text{ if } \quad F(A) + F(B) \begin{array}{c}\geq\\ =\\ \leq\end{array} F(A \cup B) + F(A \cap B)$$

"Distance" from modularity

$$\big|F(A) + F(B) - F(A \cup B) - F(A \cap B)\big|$$

Polynomial-time Mixing

$$F \text{ is } \begin{matrix} \text{sub-} \\ \text{modular} \\ \text{super-} \end{matrix} \text{ if } \quad F(A) + F(B) \begin{matrix} \geq \\ = \\ \leq \end{matrix} F(A \cup B) + F(A \cap B)$$

"Distance" from modularity

$$\zeta_F := \max_{A,B \subseteq V} \left| F(A) + F(B) - F(A \cup B) - F(A \cap B) \right|$$

Polynomial-time Mixing

submodular	constant	modular	normalized monotone submodular
$F(S)$ $=$	c $+$	$\sum_{v \in S} m_v$ $+$	$f(S)$

$$F(S) = c + \sum_{v \in S} m_v + f(S)$$

Polynomial-time Mixing

submodular | constant | modular | normalized monotone submodular

$$F(S) = c + \sum_{v \in S} m_v + f(S)$$

$$\exp(F(S)) \propto \prod_{v \in S} \exp(m_v) \times \exp(f(S))$$

PSM | product | interactions

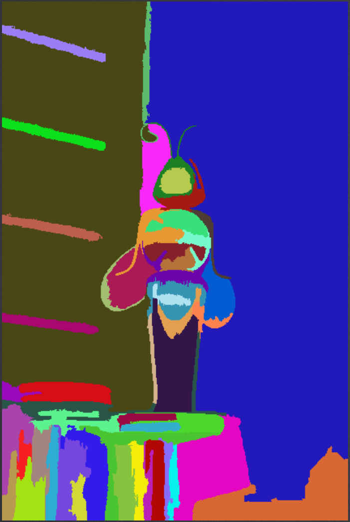
Polynomial-time Mixing

Theorem 1

For any submodular or supermodular set function F, the mixing time of the Gibbs sampler is bounded by

$$t_{\text{mix}}(\epsilon) = \mathcal{O}\left(n^2 \exp(\zeta_f) \log \epsilon^{-1}\right).$$

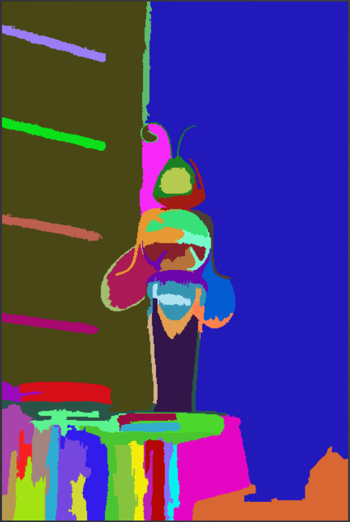
Polynomial-time Mixing



$$F_i(S) = \phi\left(|S \cap V_i|\right)$$

$$F(S) = \sum_{i=1}^{L} F_i(S)$$

Polynomial-time Mixing



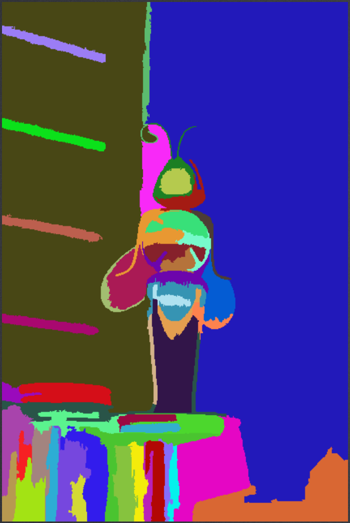
$$F_i(S) = \phi\left(|S \cap V_i|\right)$$

$$F(S) = \sum_{i=1}^{L} F_i(S)$$

Easy to show that

$$\zeta_f \leq L\phi_{\text{max}}$$

Polynomial-time Mixing



$$F_i(S) = \phi\left(|S \cap V_i|\right)$$

$$F(S) = \sum_{i=1}^{L} F_i(S)$$

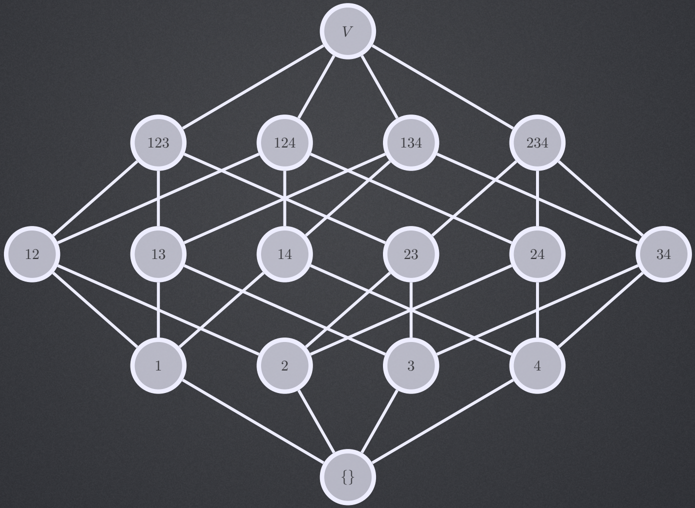
Easy to show that

$$\zeta_f \leq L\phi_{\text{max}}$$

$$\boxed{|V_i| \approx 10^5 \;\; \text{vs.} \;\; L \approx 50}$$

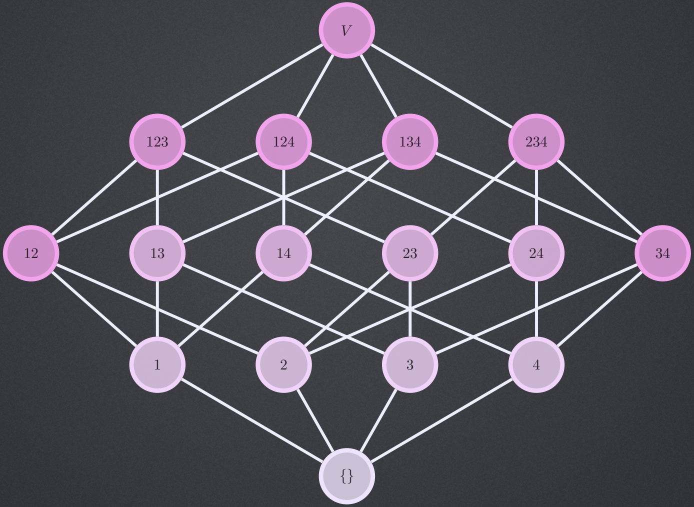
Proof Outline

Method of canonical paths [Sinclair, '92]



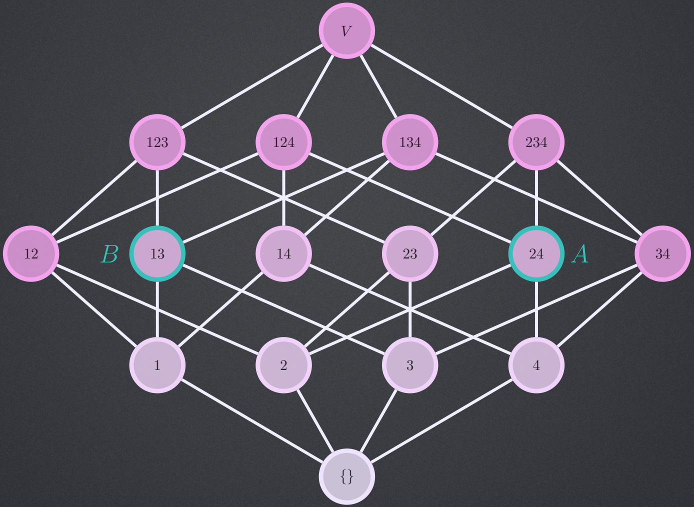
Proof Outline

For each $A, B \subseteq V$, need to route $p(A)p(B)$ amount of flow



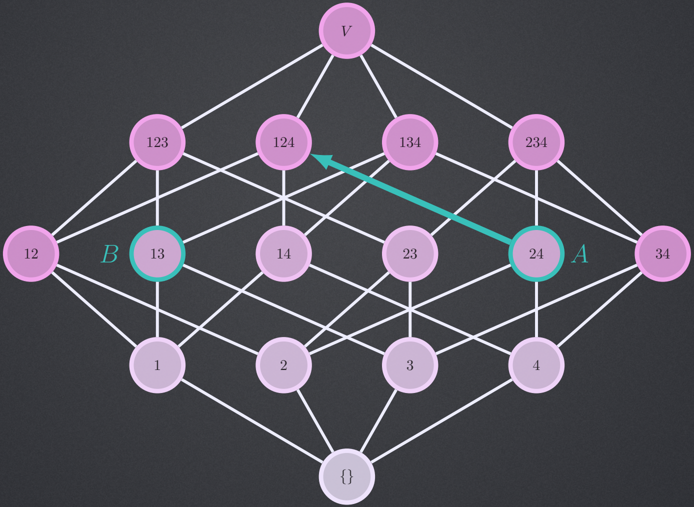
Proof Outline

Construct a "canonical path" between each $A, B \subseteq V$



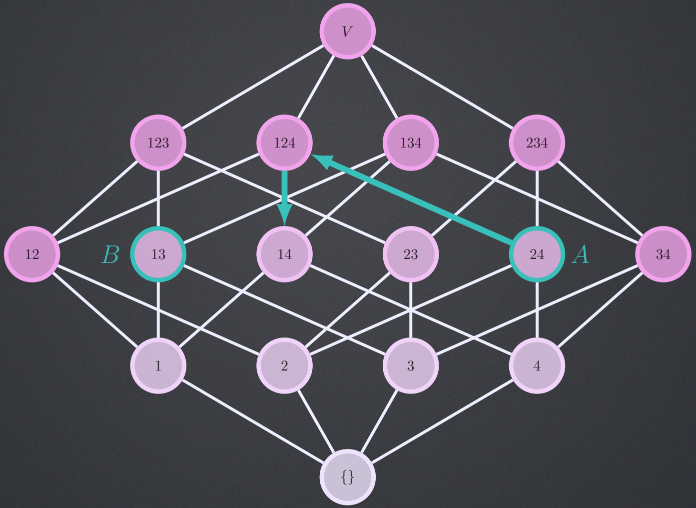
Proof Outline

Construct a "canonical path" between each $A, B \subseteq V$



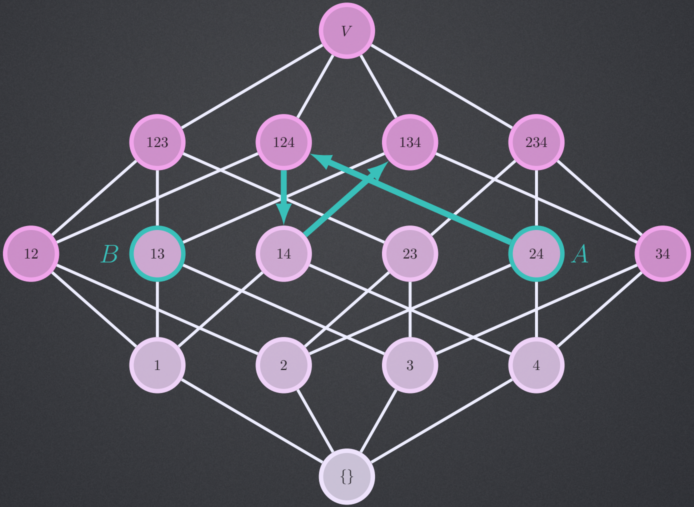
Proof Outline

Construct a “canonical path” between each $A, B \subseteq V$



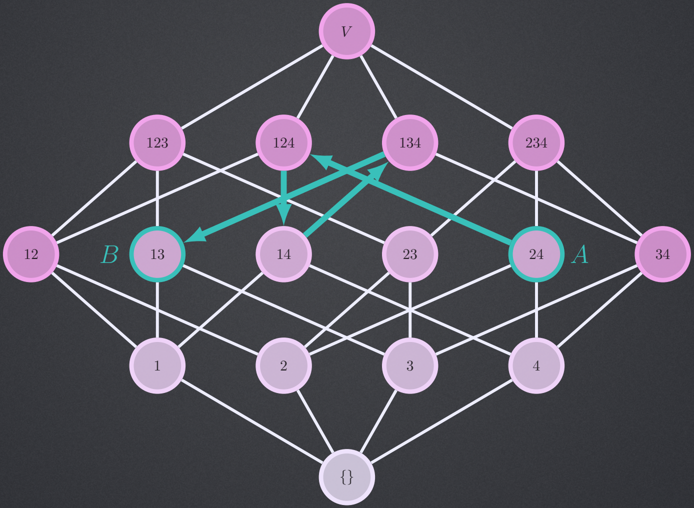
Proof Outline

Construct a "canonical path" between each $A, B \subseteq V$



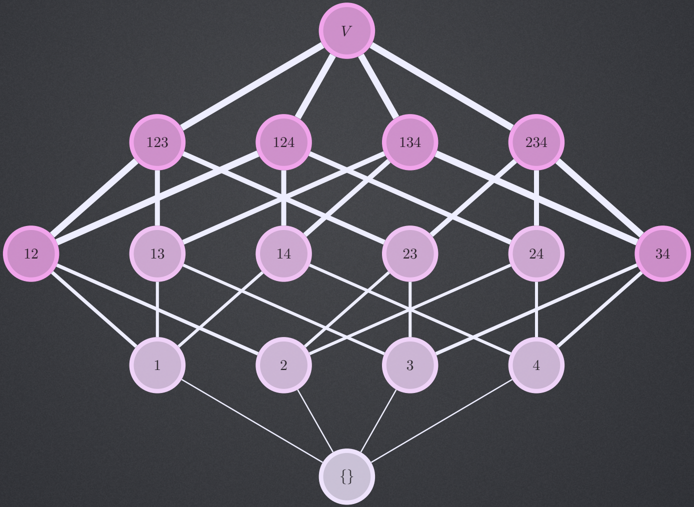
Proof Outline

Construct a “canonical path” between each $A, B \subseteq V$



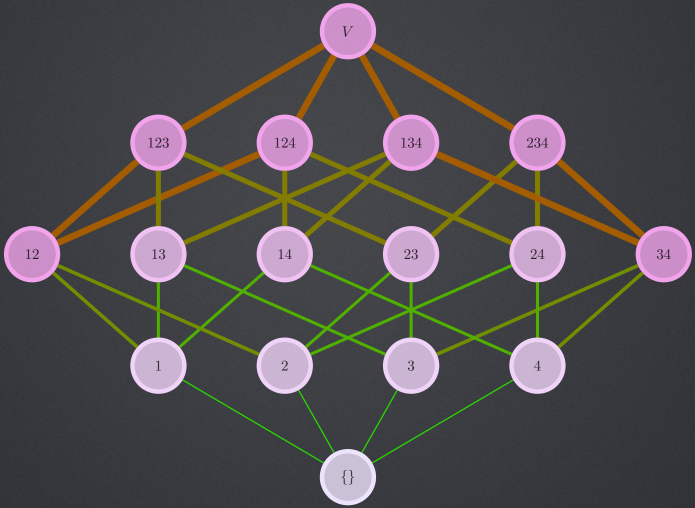
Proof Outline

Capacity of an edge $\sim$ transition probability of Gibbs sampler



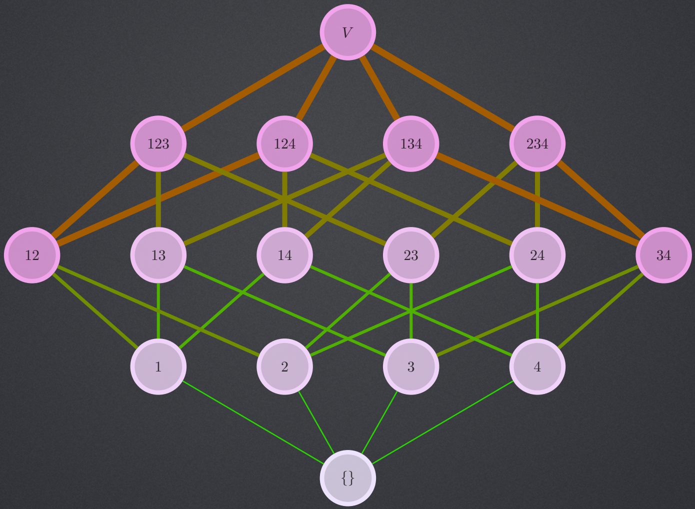
Proof Outline

Congestion of an edge $\sim$ (total flow through edge) / capacity



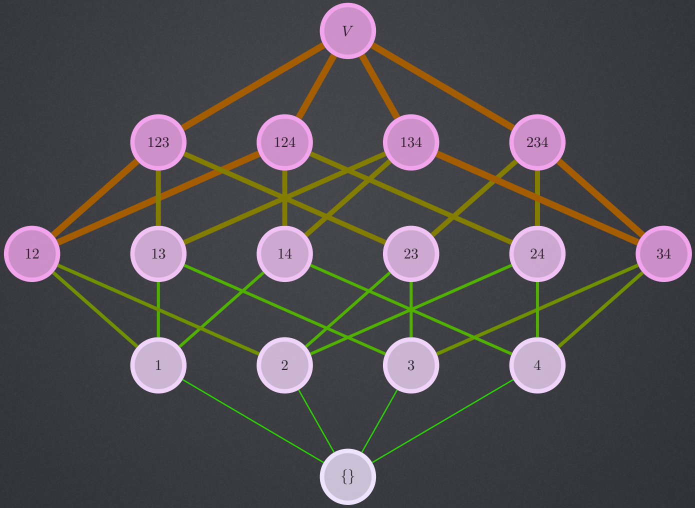
Proof Outline

$t_{\mathsf{mix}}(\epsilon) = \mathcal{O}\left(\max\{\text{congestion}\} \log \epsilon^{-1}\right)$ [Sinclair, '92]



Proof Outline

We bound the maximum congestion of a PSM using ζ_f



Fast Mixing

Theorem 2

For any submodular or supermodular set function F, if $\gamma_f < 1$, the mixing time of the Gibbs sampler is bounded by

$$t_{\text{mix}}(\epsilon) \leq \frac{1}{1-\gamma_f} n \left(\log n + \log \epsilon^{-1}\right).$$

- γ_f = "maximum total influence"

Fast Mixing

Theorem 2

For any submodular or supermodular set function F, if $\gamma_f < 1$, the mixing time of the Gibbs sampler is bounded by

$$t_{\text{mix}}(\epsilon) \leq \frac{1}{1-\gamma_f} n \left(\log n + \log \epsilon^{-1}\right).$$

- γ_f = "maximum total influence"
- Simple way to bound γ_f, if $f(S) = \sum_i f_i(S)$

Fast Mixing

Theorem 2

For any submodular or supermodular set function F, if $\gamma_f < 1$, the mixing time of the Gibbs sampler is bounded by

$$t_{\text{mix}}(\epsilon) \leq \frac{1}{1-\gamma_f} n \left(\log n + \log \epsilon^{-1}\right).$$

- γ_f = "maximum total influence"
- Simple way to bound γ_f, if $f(S) = \sum_i f_i(S)$
- Closely related to Dobrushin uniqueness conditions, and influence matrix norms [Dyer et al., '09]

Fast Mixing

Theorem 2

For any submodular or supermodular set function F, if $\gamma_f < 1$, the mixing time of the Gibbs sampler is bounded by

$$t_{\text{mix}}(\epsilon) \leq \frac{1}{1-\gamma_f} n \left(\log n + \log \epsilon^{-1}\right).$$

- γ_f = "maximum total influence"
- Simple way to bound γ_f, if $f(S) = \sum_i f_i(S)$
- Closely related to Dobrushin uniqueness conditions, and influence matrix norms [Dyer et al., '09]
- Similar theorem by [Rebeschini and Karbasi, '15]

Evaluation

Compare against variational approach [Djolonga and Krause, '14]

log-supermodular	log-supermodular	log-submodular
pairwise MRF	higher-order MRF	facility location

Evaluation

Compare against variational approach [Djolonga and Krause, '14]

log-supermodular	log-supermodular	log-submodular
pairwise MRF	higher-order MRF	facility location

Evaluation

Compare against variational approach [Djolonga and Krause, '14]

log-supermodular

pairwise MRF

log-supermodular

higher-order MRF

log-submodular

facility location

- Compute $p(v \mid S)$

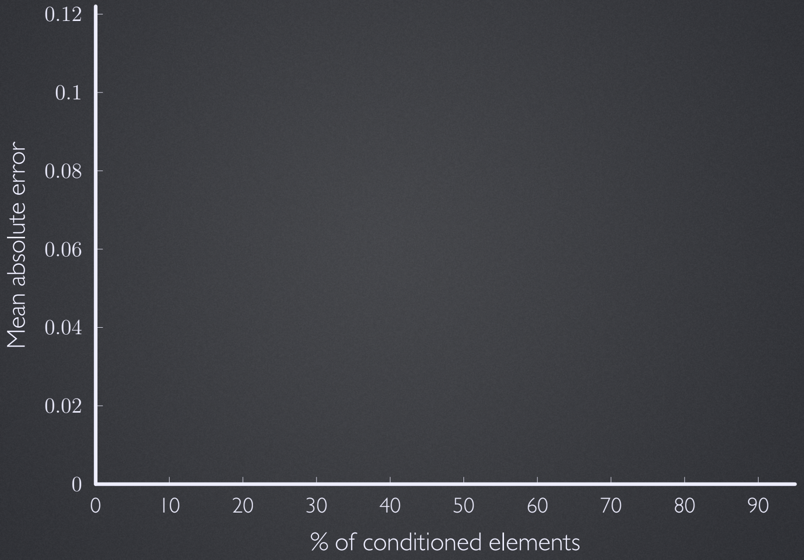
Evaluation

Compare against variational approach [Djolonga and Krause, '14]

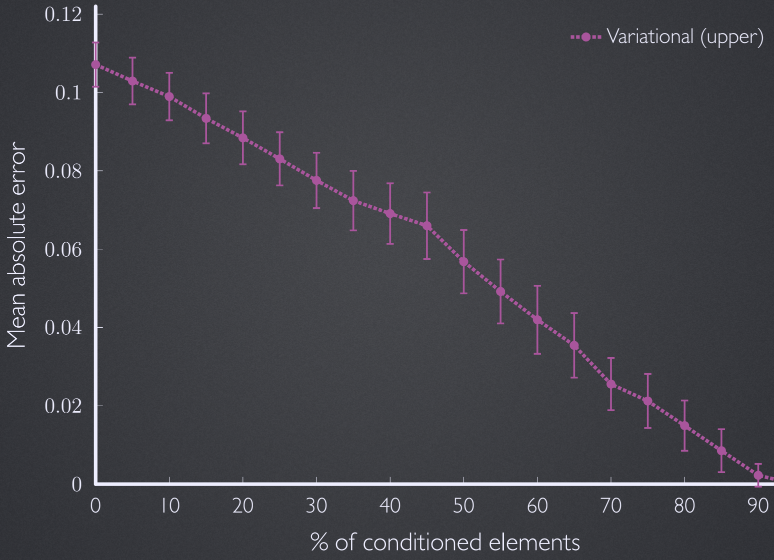
log-supermodular	log-supermodular	log-submodular
pairwise MRF	higher-order MRF	facility location

- Compute $p(v \mid S)$
- $|V| = 20 \quad \longrightarrow$ compare to exact marginals

Evaluation

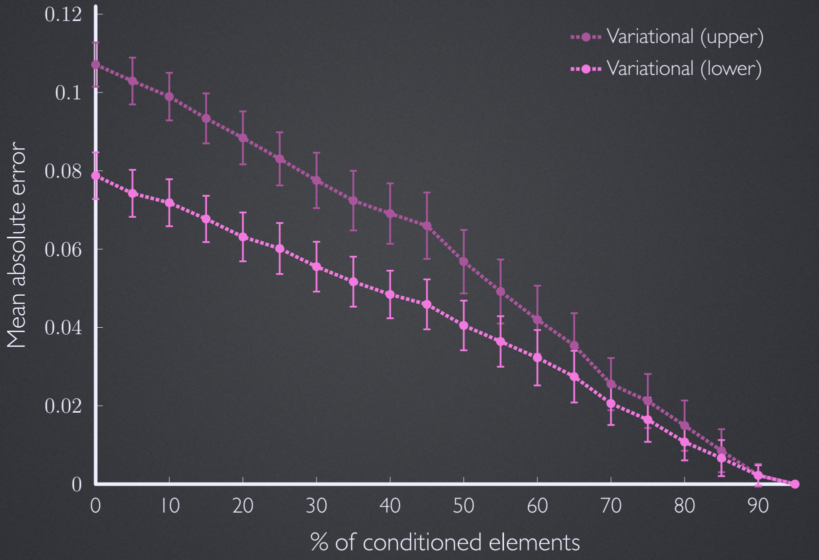

0.12
0.1
0.08
0.06
0.04
0.02
0
Mean absolute error
0
10
20
30
40
50
60
70
80
90
% of conditioned elements

Evaluation



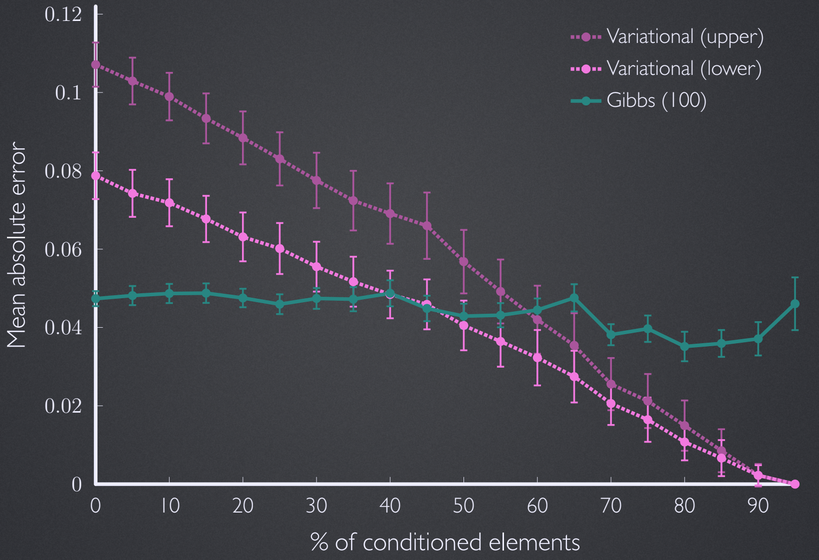
Variational (upper)
Mean absolute error
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50
60
70
80
90
% of conditioned elements

Evaluation

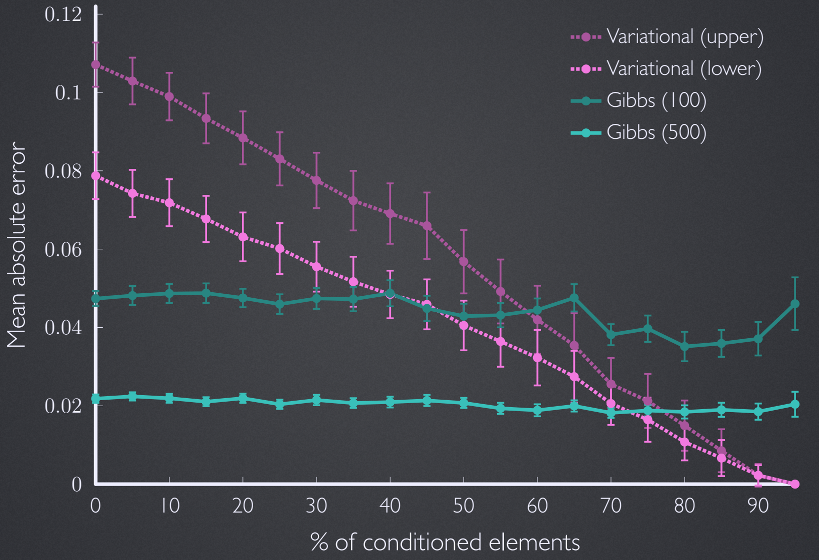


Variational (upper)
Variational (lower)
Mean absolute error
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50
60
70
80
90
% of conditioned elements

Evaluation

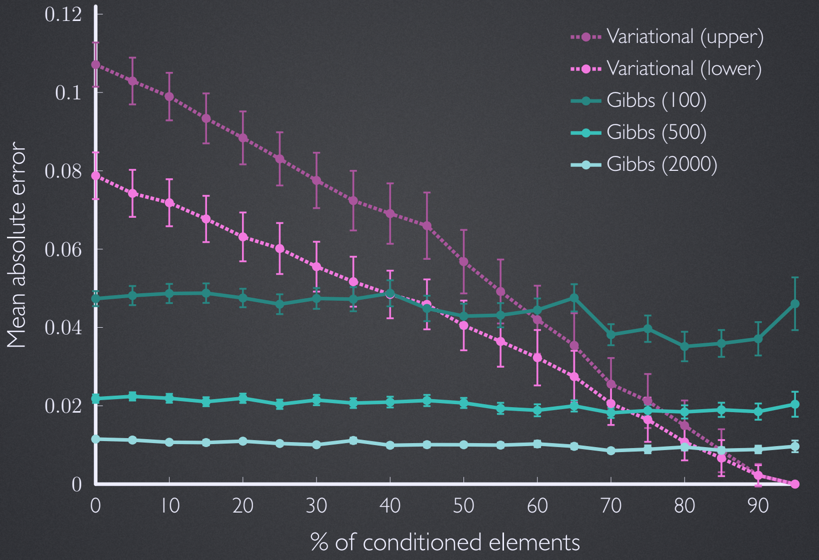

Variational (upper)
Variational (lower)
Gibbs (100)
Mean absolute error
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50
60
70
80
90
% of conditioned elements

Evaluation

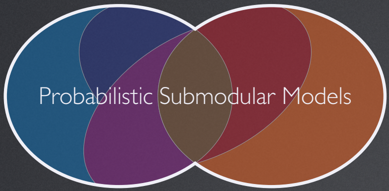


Variational (upper)
Variational (lower)
Gibbs (100)
Gibbs (500)
Mean absolute error
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50
60
70
80
90
% of conditioned elements

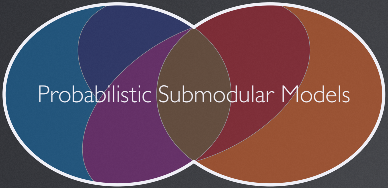
Evaluation

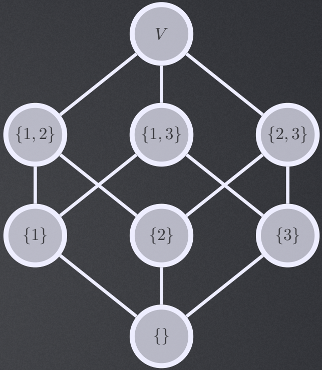

Variational (upper)
Variational (lower)
Gibbs (100)
Gibbs (500)
Gibbs (2000)
0.12
0.1
0.08
0.06
0.04
0.02
0
Mean absolute error
0
10
20
30
40
50
60
70
80
90
% of conditioned elements

Conclusion

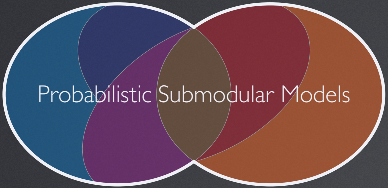


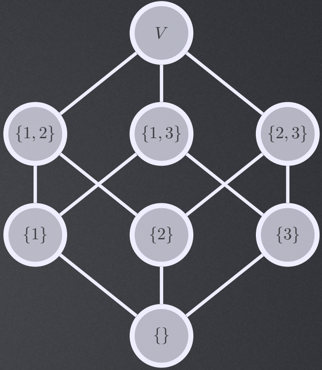
Conclusion





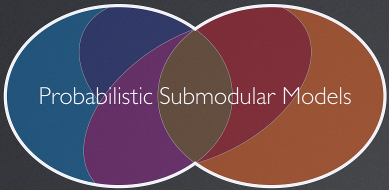
Conclusion

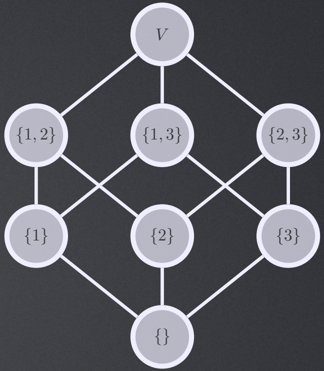




- Identify higher-order models amenable to efficient inference

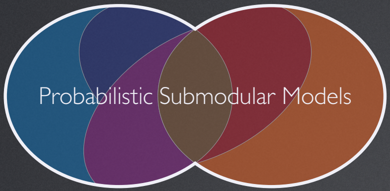
Conclusion

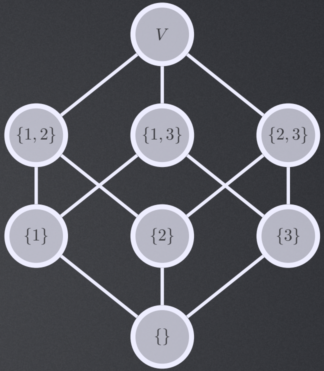




- Identify higher-order models amenable to efficient inference
- First indications that sub-/supermodularity can lead to faster mixing

Conclusion





- Identify higher-order models amenable to efficient inference
- First indications that sub-/supermodularity can lead to faster mixing

Poster #70

Backup I

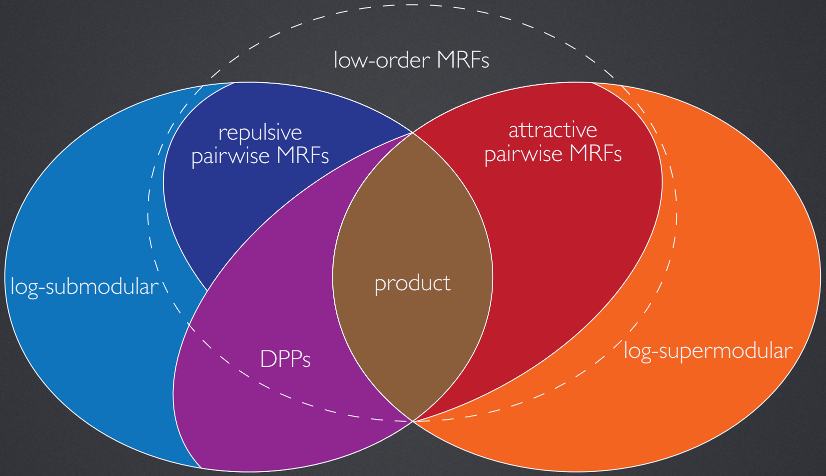
Start at S_0

For $t = 1, 2, \ldots$

- Select random $v \in V$
- $\Delta \leftarrow F(S_t \cup \{v\}) - F(S_t \setminus \{v\})$
- $p_{\text{add}} \leftarrow e^{\Delta} / \left(1 + e^{\Delta}\right)$
- Flip biased coin

p_{add}	$1 - p_{\text{add}}$
$S_{t+1} \leftarrow S_t \cup \{v\}$	$S_{t+1} \leftarrow S_t \setminus \{v\}$

Backup II



low-order MRFs
repulsive
pairwise MRFs
attractive
pairwise MRFs
log-submodular
product
DPPs
log-supermodular

Backup III

Theorem 1

For any set function F, the mixing time of the Gibbs sampler is bounded by

$$t_{\text{mix}}(\epsilon) = \mathcal{O}\left(n^2 \exp(2\zeta_F) \log \epsilon^{-1}\right).$$

For any submodular or supermodular set function F, the mixing time of the Gibbs sampler is bounded by

$$t_{\text{mix}}(\epsilon) = \mathcal{O}\left(n^2 \exp(\zeta_f) \log \epsilon^{-1}\right).$$