Fastest Mixing Markov Chain on a Graph

Stephen Boyd (Elec. Engr.)
Persi Diaconis (Stat. & Math)
Lin Xiao (Aero/Astro.)

Stanford University

MS&E 314, Semidefinite Programming
Stanford, 3/11/2003
Markov chain on a graph

• connected undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

$$\mathcal{V} = \{1, \ldots, n\}, \quad \mathcal{E} = \{(i, j) \mid i \text{ and } j \text{ connected}\}$$

we’ll assume each vertex has self-loop, i.e., $(i, i) \in \mathcal{E}$

• define Markov chain on vertices $X(t) \in \{1, \ldots, n\}$, with transition probabilities

$$P_{ij} = \text{Prob}(X(t + 1) = j \mid X(t) = i)$$

- each edge $(i, j) \in \mathcal{E}$ labeled with transition probability $P_{ij}$
- we’ll take $P_{ij} = 0$ for $(i, j) \notin \mathcal{E}$, and $P_{ij} = P_{ji}$

• $P$ must satisfy $P_{ij} \geq 0$, $P1 = 1$, $P = P^T$, $P_{ij} = 0$ for $(i, j) \notin \mathcal{E}$
example:

\[
P_{ii} = 1 - \sum_{j \neq i} P_{ij}
\]
Stationary distribution

- let $\pi_i(t) = \text{Prob}(X(t) = i)$, then $\pi(t) = (\pi_1(t), \ldots, \pi_n(t))$ is the probability distribution at time $t$

$$\pi(t + 1)^T = \pi(t)^T P \quad \pi(t)^T = \pi(0)^T P^t$$

- stationary distribution $\pi_{st}$ satisfies

$$\pi_{st}^T P = \pi_{st}^T$$

the global balance equation

- since $P = P^T$ and $P1 = 1$, uniform distribution $\pi_{st} = 1/n$ is stationary

$$\lim_{t \to \infty} \|\pi(t) - 1/n\| = 0$$
Mixing rate

• since $P = P^T$, all eigenvalues are real; can order as

$$1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq -1$$

• asymptotic rate of convergence to stationary distribution determined by second largest (in magnitude) eigenvalue, the mixing rate

$$\mu(P) = \max_{i=2,\ldots,n} |\lambda_i| = \max\{\lambda_2(P), -\lambda_n(P)\}$$

• distribution of $X(t)$ approaches uniform as $\mu^t$ (if $\mu < 1$), e.g.,

$$\sup_{\pi(0)} \|\pi(t) - \mathbf{1}/n\|_1 \leq \sqrt{n}\mu^t$$

the smaller $\mu$ is, the faster the Markov chain mixes
Mixing time

• how long to wait for the Markov chain to be close to stationary?

• define mixing time

\[ \tau = \frac{1}{\log(1/\mu)} \]

gives (asymptotic) time for norm of error to decrease by factor 1/e
Fastest mixing Markov chain problem

fastest mixing Markov chain (FMMC) problem:

\[
\text{minimize } \mu(P) \\
\text{subject to } P \geq 0, \quad P1 = 1, \quad P = P^T \\
P_{ij} = 0, \quad (i, j) \notin \mathcal{E}
\]

- optimization variable is \( P \); problem data is graph
- can add other constraints

another interpretation: find fastest mixing symmetric Markov chain with fixed sparsity pattern (\( i.e. \), allowed transitions)
Background

• *Markov chain Monte Carlo simulation*, with applications in statistics, physics, chemistry, biology, computer science . . .
  
  – random sampling of a huge state space with a specified distribution
  – construct a Markov chain that converges asymptotically to the desired distribution
  – simulate the Markov chain until close to stationary, then use states of the chain as random samples

• efficiency of simulation determined by mixing rate

• previous work: bound the mixing rate with various techniques, and derive heuristics to obtain faster mixing chains

• this talk: find the fastest mixing Markov chain (and the mixing rate)

• limited by the size of practical problems
Two common suboptimal schemes

let $d_i$ be degree of vertex $i$, i.e., number of edges connected to vertex $i$ (not counting self-loops)

- maximum degree chain: with $d_{\text{max}} = \max_{i \in V} d_i$

\[
P_{ij}^{\text{md}} = \frac{1}{d_{\text{max}}}, \quad i \neq j, \ (i, j) \in \mathcal{E}
\]

- Metropolis-Hastings chain

\[
P_{ij}^{\text{mh}} = \frac{1}{\max\{d_i, d_j\}}, \quad i \neq j, \ (i, j) \in \mathcal{E}
\]

diagonal entries determined by $P_{ii} = 1 - \sum_{j \neq i} P_{ij}$
A simple example

- maximum degree and Metropolis-Hastings
  \[ \mu^{\text{md}} = \mu^{\text{mh}} = \frac{2}{3} \]

- can we do better? yes!
  \[ \mu^* = \frac{3}{7} \]
  is, in fact, optimal for FMMC

- can we always find the best? how difficult is it?
  how suboptimal is maximum degree or Metropolis-Hastings?
Outline

• convex optimization & SDP formulation of FMMC

• examples

• Lagrange dual of FMMC and optimality conditions

• exploit structure in interior-point methods

• subgradient method

• extension to reversible Markov chains
Convexity of mixing rate

$\mu(P)$ is **convex function** of $P$

- variational characterization of $\mu(P)$:

\[
\mu(P) = \max \{ \lambda_2(P), -\lambda_n(P) \} \\
\lambda_2(P) = \sup \{ v^T P v \mid \| v \|_2 \leq 1, \ 1^T v = 0 \} \\
\lambda_n(P) = \sup \{ -v^T P v \mid \| v \|_2 \leq 1, \ 1^T v = 0 \}
\]

- $\mu(P)$ is spectral norm of $P$ on $1^\perp = \{ v \mid 1^T v = 0 \}$:

\[
\mu(P) = \left\| (I - (1/n)11^T) P (I - (1/n)11^T) \right\|_2 = \left\| P - (1/n)11^T \right\|_2
\]

- for $X = X^T$, $\lambda_1(X) + \lambda_2(X)$ and $-\lambda_n(X)$ are convex; here $\lambda_1 = 1$, so max\{\lambda_2(X), -\lambda_n(X)\} is convex
Convex optimization formulation of FMMC

minimize \( \mu(P) = \|P - \frac{1}{n}11^T\|_2 \)

subject to \( P \geq 0, \quad P1 = 1, \quad P = P^T \)
\( P_{ij} = 0, \quad (i, j) \notin \mathcal{E} \)

• convex optimization problem

• nondifferentiable objective function, linear constraints

• hence, can solve efficiently; have duality theory, . . .
SDP formulation of FMMC

introducing a scalar variable $s$ to bound the norm of $P - (1/n)11^T$

\[
\begin{align*}
\text{minimize} & \quad s \\
\text{subject to} & \quad -sI \preceq P - (1/n)11^T \preceq sI \\
& \quad P \succeq 0, \quad P1 = 1, \quad P = P^T \\
& \quad P_{ij} = 0, \quad (i, j) \notin \mathcal{E}
\end{align*}
\]

a semidefinite program (SDP) in variables $P, s$

- $X \preceq Y$ means $Y - X$ is positive semidefinite
- $P \succeq 0$ denotes elementwise inequality
Extensions

can add other convex constraints on the transition probabilities

**fastest local degree chain**: require probability on edge to be function of degrees of vertices:

\[
P_{i,j}^{ld} = \phi(d_i, d_j), \quad i \neq j, \ (i, j) \in E
\]

- diagonal entries determined by \( P_{ii} = 1 - \sum_{j \neq i} P_{i,j} \)
- includes Metropolis-Hastings as special case \( P_{i,j} = 1 / \max\{d_i, d_j\} \)
- for convex/SDP formulation, add linear equality constraints

\[
P_{i,j} = P_{k,l} \text{ whenever } d_i = d_k < d_j = d_l
\]
Outline

- convex optimization & SDP formulation of FMMC
- examples
- Lagrange dual of FMMC and optimality conditions
- exploit structure in interior-point methods
- subgradient method
- extension to reversible Markov chains
Small example (a)

\[1/2 \quad 1/2 \quad 1/2\]

\[\mu^\text{md} = \mu^\text{mh} = \mu^\text{ld} = \mu^* = \lambda_2 = -\lambda_n = \sqrt{2}/2\]
Small example (b)

\[ \mu^{\text{md}} = \lambda_2 = \frac{2}{3} \]

\[ \mu^{\text{mh}} = \lambda_2 = \frac{2}{3} \]

\[ \mu^{\text{id}} = \mu^* = \lambda_2 = \frac{7}{11} \]
Small example (c)

\[
\begin{align*}
\mu^{\text{md}} &= \mu^{\text{mh}} = -\lambda_n = 2/3 \\
\mu^{\text{ld}} &= \mu^* = \lambda_2 = -\lambda_n = 3/7
\end{align*}
\]
Small example (d)

\[ \mu_{mh} = \frac{7}{12} \]

\[ \mu_{md} = \mu^* = \frac{1}{4} \]
A random family with 50 vertices

- randomly generate symmetric matrix $A \in \mathbb{R}^{50 \times 50}$, where $A_{ij}$, for $i \leq j$, are i.i.d. uniformly distributed on interval $[0, 1]$
- for each threshold $c \in [0, 1]$, place an edge between $i$ and $j$ if $A_{ij} \leq c$
- let $c = 0.1, 0.2, \ldots, 0.9$, obtain a monotone family of graphs

![Graphs with mixing rate and mixing time vs number of edges]
Eigenvalue distribution

for the graph with $c = 0.2$. The dashed lines indicate $\pm \mu$ for each chain.
Outline

• convex optimization & SDP formulation of FMMC

• examples

• Lagrange dual of FMMC and optimality conditions

• exploit structure in interior-point methods

• subgradient method

• extension to reversible Markov chains
Dual of FMMC problem

primal FMMC:

minimize \( \mu(P) = \|P - (1/n)11^T\|_2 \)
subject to \( P \geq 0, \quad P1 = 1, \quad P = P^T \)
\( P_{ij} = 0, \quad (i, j) \notin \mathcal{E} \)

dual FMMC (with variables \( Y, z \)):

maximize \( 1^Tz \)
subject to \( Y1 = 0, \quad Y = Y^T \)
\( \|Y\|_* = \sum_{i=1}^n |\lambda_i(Y)| \leq 1 \)
\( (z_i + z_j)/2 \leq Y_{ij}, \quad (i, j) \in \mathcal{E} \)

(\( \| \cdot \|_* \) is indeed the dual of the spectral norm, \textit{nuclear norm})
Weak duality

if $P$ primal feasible, and $Y$, $z$ dual feasible, then $1^T z \leq \mu(P)$

quick proof:

\[
\text{Tr} \ Y \left( P - (1/n)11^T \right) \leq \|Y\|_* \|P - (1/n)11^T\|_2 \\
\leq \|P - (1/n)11^T\|_2 \\
= \mu(P)
\]

\[
\text{Tr} \ Y \left( P - (1/n)11^T \right) = \text{Tr} \ YP = \sum_{i,j} Y_{ij} P_{ij} \\
\geq \sum_{i,j} (1/2)(z_i + z_j) P_{ij} \\
= (1/2)(z^T P 1 + 1^T P z) \\
= 1^T z
\]
Strong duality

- primal and dual FMMC problems are solvable, and have same optimal value

- there are primal feasible $P^*$, and dual feasible $Y^*$, $z^*$ with

$$\|P^* - (1/n)11^T\|_2 = 1^T z^*$$
Optimality conditions

• primal feasibility

\[ P \geq 0, \quad P = P^T, \quad P \mathbf{1} = \mathbf{1}, \quad P_{ij} = 0 \text{ for } (i, j) \notin \mathcal{E} \]

• dual feasibility

\[ Y = Y^T, \quad Y \mathbf{1} = 0, \quad \|Y\|_* \leq 1, \quad (z_i + z_j)/2 \leq Y_{ij} \text{ for } (i, j) \in \mathcal{E} \]

• complementary slackness

\[
((z_i + z_j)/2 - Y_{ij}) P_{ij} = 0 \\
Y = Y_+ - Y_-, \quad Y_+ = Y_+^T \geq 0, \quad Y_- = Y_-^T \geq 0 \\
\text{Tr} \ Y_+ + \text{Tr} \ Y_- = 1 \\
PY_+ = \mu(P)Y_+, \quad PY_- = -\mu(P)Y_-
\]
Outline

- convex optimization & SDP formulation of FMMC
- examples
- Lagrange dual of FMMC and optimality conditions
- exploit structure in interior-point methods
- subgradient method
- extension to reversible Markov chains
The incidence matrix

- label edges (not including self-loops) by \( l = 1, \ldots, m \); and define vertex-edge incidence matrix \( A \in \mathbb{R}^{n \times m} \)

\[
A_{il} = \begin{cases} 
1 & \text{if edge } l \text{ starts from vertex } i \\
-1 & \text{if edge } l \text{ ends at vertex } i \\
0 & \text{otherwise.}
\end{cases}
\]

- example

- the columns of \( A \)

\[
a_l = e_i - e_j, \quad l \sim (i, j)
\]
An alternative representation

- introduce $p \in \mathbb{R}^m$, with $p_l$ being transition probability on edge $l$, then

$$P(p) = I - A \text{diag}(p) A^T = I - \sum_{l=1}^{m} p_l a_l a_l^T$$

$$p \geq 0, \quad |A|p \leq 1$$

$|A|$: elementwise absolute value; edge directions make no difference

- FMMC problem in terms of new variable $p \in \mathbb{R}^m$

minimize $$\|I - A \text{diag}(p) A^T - (1/n)11^T\|_2$$

subject to $$p \geq 0, \quad |A|p \leq 1$$
The centering problem

- SDP formulation of FMMC

\[
\begin{align*}
\text{minimize} & \quad s \\
\text{subject to} & \quad -sI \preceq I - A \text{diag}(p) A^T - (1/n)11^T \preceq sI \\
& \quad p \geq 0, \quad |A|p \leq 1
\end{align*}
\]

- the centering problem with logarithmic barrier functions

\[
\begin{align*}
\text{minimize} & \quad ts - \log \det (sI - (I - A \text{diag}(p) A^T) + (1/n)11^T) \\
& \quad - \log \det (sI + (I - A \text{diag}(p) A^T) - (1/n)11^T) \\
& \quad - \sum_{l=1}^m \log p_l - \sum_{i=1}^n \log (1 - (|A|p)_i)
\end{align*}
\]

as \( t \) becomes large enough, get good approximation of FMMC
Exploit sparse structure

- use Newton’s method to solve the centering problem
- frequent computing of $(sI - P + 11^T/n)^{-1}$ and $(sI + P - 11^T/n)^{-1}$
- structure: sparse + rank-one
- use Sherman-Woodbury-Morrison formula, e.g.,

\[
(sI - P + 11^T/n)^{-1} = (sI - P)^{-1} - \frac{(sI - P)^{-1}11^T(sI - P)^{-1}}{n + 11^T(sI - P)^{-1}1}
\]

can efficiently compute $(sI - P)^{-1}$ by sparse factorization
Assembly gradient and Hessian

- Newton step $v$ computed by solving $Hv = -g$

- for convenience, define two matrices (note $P = I - \sum_{l=1}^{m} p_l a_l a_l^T$)

$$U = \left( sI + P - (1/n)11^T \right)^{-1}$$

$$V = \left( sI - P + (1/n)11^T \right)^{-1}$$

- gradient $g = (g_0, g_1, \ldots, g_m)$, subscript 0 indicates variable $s$

$$g_0 = t - \text{Tr} \ U - \text{Tr} \ V$$

$$g_l = \text{Tr}(U a_l a_l^T) - \text{Tr}(V a_l a_l^T) = a_l^T U a_l - a_l^T V a_l$$

$$= (U_{ii} + U_{jj} - 2U_{ij}) - (V_{ii} + V_{jj} - 2V_{ij})$$
• Hessian $H \in \mathbb{R}^{(m+1)\times(m+1)}$

\[
H_{00} = \text{Tr} U^2 + \text{Tr} V^2
\]

\[
H_{0l} = -\text{Tr}(Ua_l a_l^T U) + \text{Tr}(V a_l a_l^T V)
\]

\[
= -a_l^T U^2 a_l + a_l^T V^2 a_l
\]

\[
= -((U^2)_{ii} + (U^2)_{jj} - 2(U^2)_{ij}) + ((V^2)_{ii} + (V^2)_{jj} - 2(V^2)_{ij})
\]

\[
H_{ll'} = \text{Tr}(Ua_l a_l^T U a_{l'} a_{l'}^T) + \text{Tr}(V a_l a_l^T V a_{l'} a_{l'}^T)
\]

\[
= (a_l U a_{l'})^2 + (a_l V a_{l'})^2
\]

\[
= (U_{ii'} + U_{jj'} - U_{ij'} - U_{i'j})^2 + (V_{ii'} + V_{jj'} - V_{ij'} - V_{i'j})^2
\]

edgest denoted by $l \sim (i, j), \ l' \sim (i', j')$
Outline

- convex optimization & SDP formulation of FMMC
- examples
- Lagrange dual of FMMC and optimality conditions
- exploit structure in interior-point methods
- subgradient method
- extension to reversible Markov chains
Solution methods

• for small FMMC problems, up to 1000 variables: standard SDP solvers

• can exploit structure to gain efficiency, e.g., sparsity in $P$

• large problems up to 100000 edges: subgradient method
Subgradient of $\mu$

- mixing rate $\mu(P) = \|P - (1/n)11^T\|_2$ nondifferentiable, convex

- $G = G^T$ is a subgradient of $\mu$ at $P$ if for all $\tilde{P} = \tilde{P}^T$ and $\tilde{P}1 = 1$

$$\mu(\tilde{P}) \geq \mu(P) + \text{Tr } G(\tilde{P} - P)$$
• for example, if \( \mu(P) = \lambda_2(P) \) and \( u \) is associated unit eigenvector, then

\[
\mu(P) = \lambda_2(P) = u^T Pu
\]

\[
\mu(\tilde{P}) \geq \lambda_2(\tilde{P}) \geq u^T \tilde{P} u
\]

subtracting the both sides of equality from inequality, we have

\[
\mu(\tilde{P}) \geq \mu(P) + u^T (\tilde{P} - P)u = \mu(P) + \text{Tr}(uu^T)(\tilde{P} - P)
\]

so \( G = uu^T \) is a subgradient of \( \mu \) at \( P \)

• if \( \mu(P) = -\lambda_n(P) \) and \( v \) is associated unit eigenvector, then a subgradient is given by \( G = -vv^T \)
Subdifferential of $\mu$

The subdifferential $\partial \mu$ at $P$ is the set of subgradients

$$\partial \lambda^*(P) = \text{Co}(\{vv^T | Pv = \mu v, \|v\|_2 = 1\}) \cup \{-vv^T | Pv = -\mu v, \|v\|_2 = 1\}$$

$$= \{Y | Y = Y_+ - Y_-, Y_+ = Y_+^T \geq 0, Y_- = Y_-^T \geq 0, \text{Tr } Y_+ + \text{Tr } Y_- = 1, PY_+ = \mu Y_+, PY_- = -\mu Y_- \}$$

related to optimality conditions (KKT)
Subgradient of \( \mu \) as function of \( p \)

- mixing rate \( \mu(p) = \| I - \sum_{l=1}^{m} p_l a_l a_l^T - (1/n)11^T \|_2 \)

- if \( G \) is a subgradient of \( \mu(P) \), then a subgradient of \( \mu(p) \) is

\[
g(p) = (-a_1^T G a_1, \ldots, -a_m^T G a_m)
\]

if \( \mu(P) = \lambda_2(P) \) and \( u \) is associated unit eigenvector, then \( G = uu^T \)

\[
g_l(p) = -(u_i - u_j)^2, \quad l \sim (i, j), \quad l = 1, \ldots, m
\]

if \( \mu(P) = -\lambda_n(P) \) and \( v \) is associated unit eigenvector, then \( G = -vv^T \)

\[
g_l(p) = (v_i - v_j)^2, \quad l \sim (i, j), \quad l = 1, \ldots, m
\]

- for large sparse matrix, can compute \( \lambda_2, \lambda_n \) and associated eigenvectors very efficiently by Lanczos methods
Subgradient method

given a feasible $p^{(1)}$ at $k = 1$ (e.g., maximum-degree or M-H chain)

repeat

1. Compute a subgradient $g^{(k)}$ of $\mu$ at $p^{(k)}$, and set

$$p^{(k+1)} = p^{(k)} - \alpha_k g^{(k)} / \|g^{(k)}\|$$

2. approximately project $p^{(k+1)}$ onto $\{p \mid p \geq 0, \ |A|p \leq 1\}$

3. $k := k + 1$

• step lengths satisfy the diminishing rule:

$$\alpha_k \geq 0, \quad \lim_{k \to \infty} \alpha_k = 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty$$
A large example using subgradient method

random graph with 10000 vertices and 100000 edges; $\alpha_k = 1/\sqrt{k}$

starting point: Metropolis-Hastings chain (with $\mu = 0.73$)
Eigenvalue distribution

Metropolis-Hastings chain $k = 0$

subgradient method $k = 500$

Distribution of the 101 largest and 100 smallest eigenvalues
Outline

- convex optimization & SDP formulation of FMMC
- examples
- Lagrange dual of FMMC and optimality conditions
- exploit structure in interior-point methods
- subgradient method
- extension to reversible Markov chains
Extension: fastest mixing to nonuniform distribution

• we are given desired equilibrium distribution \( \pi = (\pi_1, \ldots, \pi_n) \)

• we consider \( P \) with same sparsity pattern as graph, but not symmetric

• we do require **reversible** chain:

\[
\pi_i P_{ij} = \pi_j P_{ji}, \quad i, j = 1, \ldots, n
\]

the *detailed balance equation*

• let \( \Pi = \text{diag}(\pi) \), then detailed balance is equivalent to

\[
\Pi P = P^T \Pi
\]

• the matrix \( \Pi^{-1/2} P \Pi^{1/2} \) is symmetric, with same eigenvalues as \( P \)
• eigenvector of $\Pi^{-1/2} P \Pi^{1/2}$ associated with maximum eigenvalue (which is one) is

$$q = (\sqrt{\pi_1}, \ldots, \sqrt{\pi_n})$$

• asymptotic rate of convergence of distribution to $\pi$ determined by

$$\mu(P) = \left\| \Pi^{-1/2} P \Pi^{1/2} - qq^T \right\|_2$$

which is convex in $P$

• SDP formulation of fastest mixing reversible Markov chain

$$\begin{align*}
\text{minimize} & \quad s \\
\text{subject to} & \quad -sI \preceq \Pi^{-1/2} P \Pi^{1/2} - qq^T \preceq sI \\
& \quad P \succeq 0, \quad P1 = 1, \quad \Pi P = P^T \Pi \\
& \quad P_{ij} = 0, \quad (i, j) \notin \mathcal{E}
\end{align*}$$
Summary

FMMC problem (and many variations) are convex problems, in fact SDPs

- can solve modest problems exactly and easily
- can solve larger problems via subgradient method

Current research

- exploit symmetry: reduce number of variables and matrix dimensions
- fast distributed averaging, fast resource allocation over networks