The Complexity of Estimating Convergence Time

Nayantara Bhatnagar (University of Delaware)
work with
Andrej Bogdanov (Chinese University of Hong Kong)
Elchanan Mossel (Univ. of Pennsylvania and UC Berkeley)

January 26, 2016

Convergence diagnostics

- MCMC used in Bayesian inference, computational physics and chemistry, image processing, phylogeny ...
- Eventually the chain will converge to the desired target distribution.
- May or may not have bounds on the mixing time. Bounds may not be practical.
- How to tell whether the chain is close to converged?
- In practice many visual, statistical tests are used - convergence diagnostics.

Definitions

Probability measures μ and ν on finite Ω. The **total variation distance** between μ and ν is

$$\|\mu - \nu\|_{tv} := \max_{A \subset \Omega} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|$$

Markov chain M on Ω with transition matrix P and stationary distribution π.

$$d(t) := \max_{x,y \in \Omega} \|P^t(x, \cdot) - P^t(y, \cdot)\|_{tv}.$$

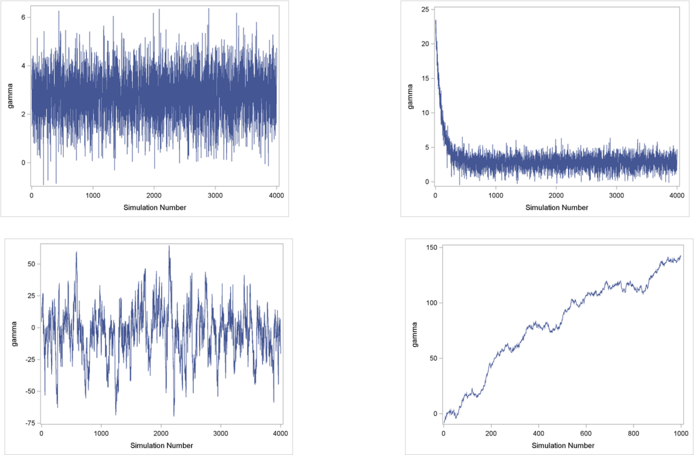
The ε-**mixing time** is

$$\tau(\varepsilon) := \inf\{t : d(t) \leq \varepsilon\}$$

The ε-**mixing time started at** x is

$$\tau_x(\varepsilon) := \inf\{t : \|P^t(x, \cdot) - \pi\|_{tv} \leq \varepsilon\}$$

Traceplot



SAS/STAT(R) 9.22 User's Guide - Assesing Markov Chain Convergence

Statistical tests

Name	Description	Interpretation of the Test
Gelman-Rubin	Uses parallel chains with dispersed initial values to test whether they all converge to the same target distribution. Failure could indicate the presence of a multi-mode posterior distribution (different chains converge to different local modes) or the need to run a longer chain (burn-in is yet to be completed).	One-sided test based on a variance ratio test statistic. Large $\widehat{R}_c$ values indicate rejection.
Geweke	Tests whether the mean estimates have converged by comparing means from the early and latter part of the Markov chain.	Two-sided test based on a z-score statistic. Large absolute z values indicate rejection.
Heidelberger-Welch (stationarity test)	Tests whether the Markov chain is a covariance (or weakly) stationary process. Failure could indicate that a longer Markov chain is needed.	One-sided test based on a Cramer–von Mises statistic. Small p-values indicate rejection.
Heidelberger-Welch (half-width test)	Reports whether the sample size is adequate to meet the required accuracy for the mean estimate. Failure could indicate that a longer Markov chain is needed.	If a relative half-width statistic is greater than a predetermined accuracy measure, this indicates rejection.
Raftery-Lewis	Evaluates the accuracy of the estimated (desired) percentiles by reporting the number of samples needed to reach the desired accuracy of the percentiles. Failure could indicate that a longer Markov chain is needed.	If the total samples needed are fewer than the Markov chain sample, this indicates rejection.
autocorrelation	Measures dependency among Markov chain samples.	High correlations between long lags indicate poor mixing.
effective sample size	Relates to autocorrelation; measures mixing of the Markov chain.	Large discrepancy between the effective sample size and the simulation sample size indicates poor mixing.

SAS/STAT(R) 9.22 User's Guide - Assesing Markov Chain Convergence

[Cowles-Carlin '96] Review of 13 diagnostics and scenarios where each can fail.

Complexity theoretic framework for diagnostic algorithm

MC is a "rule" for determining next state.

Circuit $C : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}^n$ **specifies** P if

$$\mathbb{P}(C(x,r) = y) = P(x,y)$$

Diagnostic algorithm D decides if at time t:

- Chain within $1/4$ tv-distance of π: $\tau(1/4) \leq t$.
- Chain at least $1/4$ tv-distance from π: $\tau(1/4) > t$.

Exact distance at time t.

Complexity theoretic framework for diagnostic algorithm

MC is a "rule" for determining next state.

Circuit $C : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}^n$ **specifies** P if

$$\mathbb{P}(C(x,r) = y) = P(x,y)$$

Diagnostic algorithm D decides at time t:

- `mixed`: Chain within $1/8$ in tv-distance of π: $\tau(1/8) \leq t$.
- `not mixed`: Chain at least $1/2$ in tv-distance from π: $\tau(1/2) > t$.

Allow a gap in approximation to tv-distance.

Complexity theoretic framework for diagnostic algorithm

MC is a "rule" for determining next state.

Circuit $C : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}^n$ **specifies** P if

$$\mathbb{P}(C(x,r) = y) = P(x,y)$$

Diagnostic algorithm D decides:

- `mixed`: At time t, chain within $1/8$ in tv-distance of π: $\tau(1/8) \leq t$.
- `not mixed`: At time $ct, c \geq 1$, chain at least $1/2$ in tv-distance from π: $\tau(1/2) > ct$.

Allow a gap in approximation to tv-distance as well as time.

Diagnostic algorithm formulations

$\textsc{TestCon}_{c,\delta}$	Input:	C specifies P on $\Omega \subset \{0,1\}^n$, $x \in \Omega$, $t \in \mathbb{N}$.
	Promise:	P is ergodic.
	YES:	$\tau_x(1/4 - \delta) \le t$.
	NO:	$\tau_x(1/4 + \delta) > ct$.
$\textsc{PolyTestCon}_{c,\delta}$	Input:	$(C, 1^t, 1^{t_{max}})$.
	Promise:	P is ergodic and $\tau(1/4) \le t_{max}$.
	YES:	$\tau(1/4 - \delta) \le t$.
	NO:	$\tau(1/4 + \delta) > ct$.
$\textsc{PolyTestConInit}_{c,\delta}$	Input:	$(C, x, 1^t, 1^{t_{max}})$.
	Promise:	P is ergodic and $\tau(1/4) \le t_{max}$.
	YES:	$\tau_x(1/4 - \delta) \le t$.
	NO:	$\tau_x(1/4 + \delta) > ct$.

Testing convergence in a general case

$\textsc{TestCon}_{c,\delta}$	Input:	C specifies P on $\Omega \subset \{0,1\}^n$, $x \in \Omega$, $t \in \mathbb{N}$.
	Promise:	P is ergodic.
	YES:	$\tau_x(1/4 - \delta) \leq t$.
	NO:	$\tau_x(1/4 + \delta) > ct$.

(PSPACE: set of all decision problems that can be solved by a Turing machine using space polynomial in the input.)

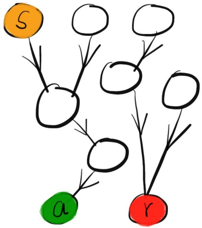
Theorem 1 (B-Bogdanov-Mossel '11). Let $1 \leq c \leq \exp\left(n^{O(1)}\right)$. Then,

- For $\exp\left(-n^{O(1)}\right) < \delta \leq 1/4$, $\textsc{TestCon}_{c,\delta}$ is in PSPACE.
- For $0 \leq \delta < 1/4$, $\textsc{TestCon}_{c,\delta}$ is PSPACE-hard.

$\textsc{TestCon}_{c,\delta}$ is PSPACE-hard

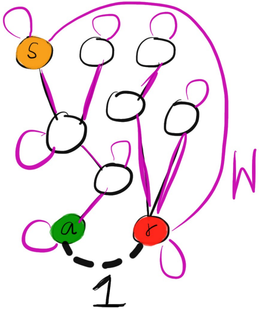
Reduction from a PSPACE complete problem A to $\textsc{TestCon}_{c,\delta}$.

Computation graph G of Turing machine T_A



$\longrightarrow$

(reversible) MC on vertices of G

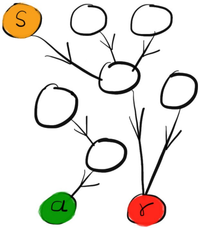


In the YES case, s and a are in the same component.

$\textsc{TestCon}_{c,\delta}$ is PSPACE-hard

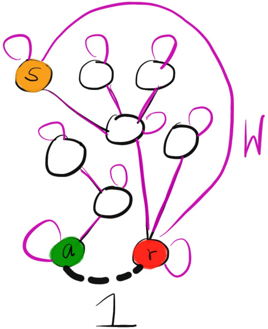
Reduction from a PSPACE complete problem A to $\textsc{TestCon}_{c,\delta}$.

Computation graph G of Turing machine T_A



$\longrightarrow$

(reversible) MC on vertices of G

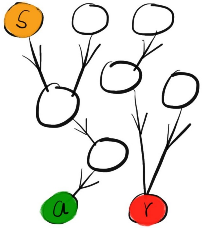


In the NO case, s and a are not in the same component.

Note: W must be chosen so that the reduction is polynomial in the input to A.

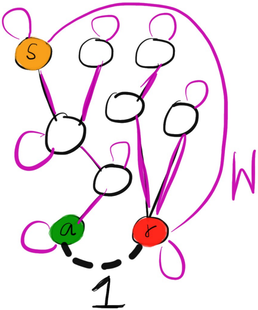
$\text{TESTCON}_{c,\delta}$ is PSPACE-hard

Computation graph G of Turing machine T_A



$\longrightarrow$

(reversible) MC on vertices of G



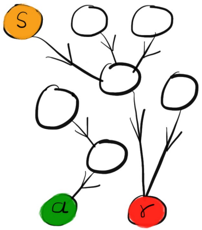
YES case: Each state of MC has const. degree $\leq D$.

$$\pi(x) = \frac{\sum\limits_{y \sim x} w_{xy}}{\sum\limits_{e \in E} w_e} \geq \frac{1}{D2^n}, \qquad \Phi \geq \frac{W}{D2^n W} = \frac{1}{D2^n}$$

$$\tau(\varepsilon) \leq \frac{2}{\Phi^2} \log\left(\frac{2}{\pi_{min}\varepsilon}\right) \leq \frac{10D^3 2^{3n}}{\varepsilon}.$$

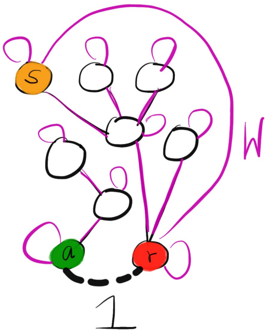
$\textsc{TestCon}_{c,\delta}$ is PSPACE-hard

State diagram of Turing machine M_A



$\longrightarrow$

(reversible) MC on states of M_A



NO case: MCs X_t started at s, Y_t started at a.

$$d(t) \geq \mathbb{P}(\forall t' \leq t, X_{t'} \notin cmp(a)) - \mathbb{P}(\exists t' \leq t \text{ s.t. } Y_{t'} \in cmp(s)) \geq 1 - \frac{2t}{W}$$

So,

$$\tau(1/4 + \delta) \geq \tau(1/2) \geq \frac{W}{4}.$$

Set $W = \frac{1000cD^3 2^{3n}}{1-4\delta}, t = \frac{10D^3 2^{3n}}{1-4\delta}, x = s$.

Testing convergence with polynomial mixing bound

$\textsc{PolyTestCon}_{c,\delta}$	Input:	$(C, 1^t, 1^{t_{max}})$.
	Promise:	P is ergodic and $\tau(1/4) \leq t_{max}$.
	YES:	$\tau(1/4 - \delta) \leq t$.
	NO:	$\tau(1/4 + \delta) > ct$.

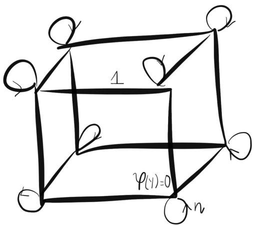
Theorem 2 (B-Bogdanov-Mossel '11).

- For $0 \leq \delta < 1/4$, $c < \frac{3/4-\delta}{2}\sqrt{t_{max}/t^2 n^3}$, $\textsc{PolyTestCon}_{c,\delta}$ is coNP-hard.
- For $0 < \delta \leq 1/4$, $\textsc{PolyTestCon}_{c,\delta}$ is in coAM.

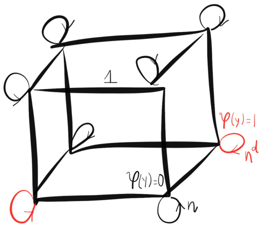
co-NP hardness of $\textsc{PolyTestCon}_{c,\delta}$

By reduction from UNSAT. Input is Ψ a CNF formula on n variables.

$\textsc{PolyTestCon}_{c,\delta}$ instance $(C, 1^t, 1^{t_{max}})$:



$\tau(1/4 - \delta) \leq C(\delta) n \log(n)$



$\tau(1/4 + \delta) > \frac{1}{2} n^{d-1}(3/4 - \delta) > cC(\delta) n \log(n)$

By a lower bound on conductance, $t_{max} \leq 32n^{2d+1}$.

Set $t = C(\delta) n \log(n)$.

Testing convergence given polynomial mixing and initial state

$\textsc{PolyTestConInit}_{c,\delta}$	Input:	$(C, x, 1^t, 1^{t_{max}})$.
	Promise:	P is ergodic and $\tau(1/4) \leq t_{max}$.
	YES:	$\tau_x(1/4 - \delta) \leq t$.
	NO:	$\tau_x(1/4 + \delta) > ct$.

Theorem 3 (B-Bogdanov-Mossel '11).

- For $0.11602 < \delta \leq 1/4$ and $c \geq 1$, $\textsc{PolyTestConInit}_{c,\delta} \in$ SZK.
- For $0 \leq \delta \leq 1/4$ and $c \leq \frac{2}{1+4\delta} t_{max}/t$, $\mathrm{PTCS}_{c,\delta}$ is SZK-hard.
- For $0 < \delta \leq 1/4$, $\mathrm{PTCS}_{c,\delta} \in$ AM$\cap$coAM.

(SZK : Statistical Zero Knowledge)

Proof Systems

Proof system for a language $L \subset \{0,1\}^n$ and a verification algorithm V with

- Completeness: If $x \in L$, there is a proof π so $V(x, \pi) = \texttt{accept}$.
- Soundness: If $x \notin L$, for all π^*, $V(x, \pi^*) = \texttt{reject}$.
- Efficiency: $V(x, \pi)$ runs in time polynomial in $|x|$.

NP is defined this way.

How much knowledge does one gain from verifying a proof?

Zero knowledge proofs

[Goldwasser-Micali-Rackoff '89] Prover P convinces verifier V of an assertion. V learns nothing but the truth of the assertion.

Interaction $(P, V)(x)$ between P and V with polynomial messages exchanged, and private coin tosses.

- Completeness: If $x \in L$, V accepts in $(P, V)(x)$ w. p. $\geq 2/3$.
- Soundness: If $x \notin L$, for "any" P^*, V accepts in $(P^*, V)(x)$ w.p. $\leq 1/3$.
- Efficiency: V runs in time polynomial in $|x|$.

Zero knowledge: The verifier could have simulated the entire interaction.

Statistical zero knowledge

"SZK": Class of languages for which there is an interaction statistically indistinguishable from the simulator with ZK.

Canonical hard problem:

$\textsc{StatDiff}_{\mathbf{s},\mathbf{c}}$	Input:	Circuits $C, C' : \{0,1\}^n \to \{0,1\}^n$ of dist. μ_1, μ_2 on $\{0,1\}^n$.
	YES:	$\|\mu_1 - \mu_2\|_{tv} \geq \mathbf{c}$.
	NO:	$\|\mu_1 - \mu_2\|_{tv} < \mathbf{s}$.

[Sahai-Vadhan '97] Let $\mathbf{0} \leq c, \mathbf{s} \leq 1$.

- For $\mathbf{c}^2 > s$, $\textsc{StatDiff}_{\mathbf{s},\mathbf{c}}$ is in SZK.
- $\textsc{StatDiff}_{\mathbf{s},\mathbf{c}}$ is SZK-hard.

SZK contains problems believed to be hard (e.g. $\textsc{GraphNonIso}$) , but cannot contain *NP*-complete problems.

SZK-hardness for $\textsc{PolyTestConInit}_{c,\delta}$

By reduction from $\textsc{StatDiff}_{\mathbf{s},\mathbf{c}}$.

$\textsc{StatDiff}_{\mathbf{s},\mathbf{c}}$	Input:	Circuits $C, C' : \{0,1\}^n \to \{0,1\}^n$ with μ_1, μ_2 on $\{0,1\}^n$.
	YES:	$\|\mu_1 - \mu_2\|_{tv} \geq \mathbf{c}$.
	NO:	$\|\mu_1 - \mu_2\|_{tv} < \mathbf{s}$.

(C, C'): instance of $\textsc{StatDiff}_{\mathbf{s},\mathbf{c}}$ with $\mathbf{c} = 1, \mathbf{s} = 1/4 - \delta$.

Construct an instance of $\textsc{PolyTestConInit}_{c,\delta}$.

MC (Y_t, Z_t) on $[M] \times \{0,1\}^n$.

- Choose Z_{t+1}:
  - If $Y_t = 1$, choose $Z_{t+1} \sim \mu_1$.
  - If $Y_t = 2$, choose $Z_{t+1} \sim \mu_2$.
  - Otherwise, set $Z_{t+1} = Z_t$.
- Choose Y_{t+1} uniformly from $[M]$.

$$\pi = U_{[M]} \times \frac{\mu_1 + \mu_2}{2}$$

SZK-hardness for $\textsc{PolyTestConInit}_{c,\delta}$

Let $x = (1, 0^n)$

$$\|P^t(x,\cdot) - \pi\|_{tv} = \frac{1}{2}\left(\frac{m-2}{m}\right)^{t-1} \|\mu_1 - \mu_2\|_{tv}$$

YES case: For $t \geq 1, M \geq 3$

$$\|P^t(x,\cdot) - \pi\|_{tv} < \frac{1}{2}\mathbf{s} < \frac{1}{4} - \delta$$

NO case: If $ct < \frac{M}{4} \ln\left(\frac{2}{1+4\delta}\right)$,

$$\|P^t(x,\cdot) - \pi\|_{tv} \geq \frac{1}{2}\left(\frac{M-2}{M}\right)^{ct-1} \mathbf{c} > \frac{1}{4} + \delta$$

In both cases $\tau(1/4) \leq M$.

Set $t_{max} = M, t = 1$.

Conclusions

- Efficient algorithms are not believed to exist for PSPACE-complete, coNP-complete or SZK-complete problems.
- Diagnostic algorithms do not exist for large classes of MCMC algorithms, unless there are efficient algorithms for PSPACE or coNP or SZK.
- (Woodard) Hardness for diagnosing convergence from a given state when π is known upto a global constant?
- Hardness for Gibbs samplers (conditional distribution of each variable can be sampled)?