Computable Exponential Bounds for Markov Chains and MCMC Simulation

Ioannis Kontoyiannis
Athens Univ of Econ & Business

joint work with
S.P. Meyn, L.A. Lastras-Montaño

Probability Seminar, Columbia University, December 2007
Outline

1. Nonasymptotic Bounds for Markov Chains
   Motivation: Markov Chain Monte Carlo

2. A General Information-Theoretic Bound
   Csiszár’s Lemma and Jensen’s inequality

3. Large Deviations Bounds: Analysis & Optimization
   Doeblin chains
   An (MCMC) example of the Gibbs sampler
   Geometrically ergodic chains
     \[ \sim \] Controlling averages and excursions
   A general MCMC sampling criterion

4. The i.i.d. case: A geometrical explanation
Motivation

A Common Task

Calculate the expectation $E_\pi(F) = \sum_{x \in S} \pi(x)F(x)$ of a given $F : S \to \mathbb{R}$.

In many cases, the distribution $\pi = (\pi(x) ; x \in S)$ is known explicitly but it’s **impossible** to calculate its values in practice.

Typical in Bayesian stat, statistical mechanics, networks, image processing, . . .

Markov Chain Monte Carlo

It is often simple to construct an ergodic Markov chain $\{X_1, X_2, \ldots\}$ with stationary distribution $\pi$.

In that case, we estimate $E_\pi(F)$ by the partial sums $\frac{1}{n} \sum_{i=1}^{n} F(X_i)$.

Problem

*How long a simulation sample $n$ do we need for an accurate estimate?*
The Setting: Deviation Bounds for Markov Chains

We have

Ergodic Markov chain \( \{X_1, X_2, \ldots \} \), discrete state-space \( S \) [for simplicity]

Transition kernel \( P(x, y) = \Pr\{X_{n+1} = y | X_n = x\} \), initial condition \( x_1 \in S \)

Stationary distribution \( \pi = (\pi(x) ; x \in S) \)

Goal

Find explicit, computable, nonasymptotic bounds on

\[
\Pr\left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq E_\pi(F) + \epsilon \right\}
\]

\( \sim \) In MCMC, this leads to precise performance guarantees and sampling criteria (or stopping rules)

\( \sim \) Similar questions appear in numerous other applications
A General Information-Theoretic Bound

Let

\[ H(P \| Q) = \sum_{x \in S} P(x) \log \frac{P(x)}{Q(x)} = \text{relative entropy} \]

\[ \| P - Q \| = \sum_{x \in S} |P(x) - Q(x)| = 2 \times [\text{total variation distance}] \]
A General Information-Theoretic Bound

Let
\[ H(P\|Q) = \sum_{x \in S} P(x) \log \frac{P(x)}{Q(x)} = \text{relative entropy} \]
\[ \|P - Q\| = \sum_{x \in S} |P(x) - Q(x)| = 2 \times [\text{total variation distance}] \]

**Theorem 1**

For any Markov chain \( \{X_n\} \), any function \( F : S \to \mathbb{R} \) bounded above, any \( c > 0 \) and any initial condition \( X_1 = x_1 \), we have

\[
\log \Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq c \right\} \leq -(n - 1) H(W\|W^1 \times P)
\]

for some bivariate distribution \( W = (W(x, y)) \) on \( S \times S \) with marginals \( W^1 \) and \( W^2 \) that satisfy

\[
\|W^1 - W^2\| \leq \frac{2}{n - 1} \quad \text{and} \quad E_{W^1}(F') \geq c - \frac{\sup_x F(x)}{n - 1}
\]

and \( W^1 \times P \) denotes the bivariate distr \( (W^1 \times P)(x, y) = W^1(x)P(x, y) \).
Interpretation

**Our result**

To *use* the above bound, we need to look at

$$\log \Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq c \right\} \leq -(n-1) \inf_{W} H(W\|W^1 \times P)$$

over all $W$ s.t.

$$\|W^1 - W^2\| \leq \frac{2}{n-1} \quad \text{and} \quad E_{W^1}(F) \geq c - \frac{\sup_x F(x)}{n-1}$$
Interpretation

Our result

To use the above bound, we need to look at
\[
\log \Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq c \right\} \leq -(n - 1) \inf_{W} H(W\|W^1 \times P)
\]
over all \(W\) s.t.
\[
\|W^1 - W^2\| \leq \frac{2}{n - 1} \quad \text{and} \quad E_{W^1}(F) \geq c - \frac{\sup_{x} F(x)}{n - 1}
\]

Donsker and Varadhan’s classic result

For a very restricted class of chains, asymptotically in \(n\)
\[
\log \Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq c \right\} \approx -n \inf_{W} H(W\|W^1 \times P)
\]
over all \(W\) s.t. \(W^1 = W^2\) and \(E_{W^1}(F) \geq c\)
Remarks

Theorem 1 offers an elementary yet general explanation of Donsker and Varadhan’s exponent and their upper bound.
The result and proof are *outrageously* general and simple.
Remarks

Theorem 1 offers an elementary yet general explanation of Donsker and Varadhan’s exponent and their upper bound.

The result and proof are outrageously general and simple.

Proof.

Step 1. Csiszár’s Lemma. Let $p$ be an arbitrary probability measure on any probability space, and $E$ any event with $p(E) > 0$. Let $p|_E$ denote the corresponding conditional measure. Then:

$$\log p(E) = -H(p|_E||p)$$
Remarks

\[\sim\] Theorem 1 offers an elementary yet general explanation of Donsker and Varadhan’s exponent and their upper bound

\[\sim\] The result and proof are outrageously general and simple

Proof.

Step 1. Csiszár’s Lemma. Let \( p \) be an arbitrary probability measure on any probability space, and \( E \) any event with \( p(E) > 0 \). Let \( p|_E \) denote the corresponding conditional measure. Then:

\[
\log p(E) = -H(p|_E \| p)
\]

With \( p = \) distribution of \((X_1, X_2, \ldots, X_n)\)
and \( E = \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq c \right\} \):

\[
\log \Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq c \right\} = -H(p|_E \| p)
\]
Proof cont’d

Step II.

Write \( p|_E \) as a product of conditionals and \( p \) as a product of \textit{bivariate} conditionals.

Expanding the log in \( H(p|_E \| p) \) (“chain rule”)

transforms this relative entropy between \( n \)-dimensional distributions into a sum of relative entropies between bivariate ones.

\[
\log \Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq c \right\} = - \sum_{i=1}^{n-1} H(p^{i,i+1} \| p^i \times P)
\]
Proof cont’d

Step II.

Write $p|_E$ as a product of conditionals and $p$ as a product of \textit{bivariate} conditionals

Expanding the log in $H(p|_E\|p)$ ("chain rule")

transforms this relative entropy between \textit{n}-dimensional distributions into a sum of relative entropies between \textit{bivariate} ones

$$\log \Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq c \right\} = - \sum_{i=1}^{n-1} H(p^{i,i+1}\| p^i \times P)$$

Step III.

Use convexity (Jensen) to simplify and combine into

$$\log \Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq c \right\} \leq -(n-1) H(W\| W^i \times P)$$

Check $W$ has the required properties \hfill \Box
The “Nicest” Chains

Doeblin chains

**Defn**  A Markov chain \( \{X_n\} \) on a general alphabet is called a *Doeblin* chain iff it converges to equilibrium exponentially fast, uniformly in the initial condition \( X_1 = x_1 \), i.e., iff

\[
\sup_{x \in S} \sum_{y \in S} |P^n(x, y) - \pi(y)| \to 0 \quad \text{exponentially fast}
\]

**Equivalent characterization**  There exists a number of steps \( m \), a probability measure \( \rho \), and \( \alpha > 0 \), such that:

\[
\Pr\{X_m \in E \mid X_1 = x_1\} \geq \alpha \rho(E) \quad \text{for all } x_1, E
\]

\[\sim\] Doeblin chains *don’t* satisfy the Donsker-Varadhan conditions

\[\sim\] They *don’t even* satisfy the usual large deviations principle!
Theorem 2

For any Doeblin chain \( \{X_n\} \), any bounded function \( F : S \to \mathbb{R} \), any \( \epsilon > 0 \), and any initial condition \( X_1 = x_1 \), we have

\[
\log \Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq E_\pi(F) + \epsilon \right\} \leq -(n - 1) \frac{1}{2} \left[ \left( \frac{\alpha}{m F_{\text{max}}} \right) \epsilon - \frac{3}{n - 1} \right]^2
\]

where \( F_{\text{max}} = \sup_x |F(x)| \)

\( \sim \) In the case of i.i.d. \( \{X_n\} \), Theorem 3 essentially reduces to Hoeffding’s bound, which is tight in that case

\( \sim \) In the general case, this is the best bound known to date, improving [Glynn & Ormoneit 2002] by a factor of 2 in the exponent
Note

$$\log \Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq E_{\pi}(F) + \epsilon \right\} \leq -(n - 1) \frac{1}{2} \left[ \left( \frac{\alpha}{m F_{\text{max}}} \right) \epsilon - \frac{3}{n - 1} \right]^2$$

$\sim$ Bound only depends on $F$ via its maximum

$\sim$ Explicit exponent, quadratic in $\epsilon$

$\sim$ Bound only depends on the chain via $\alpha, m$

$\sim$ Good convergence estimates $\Rightarrow$ good bounds on $\alpha, m$

$\Rightarrow$ better exponents
Proof outline

**Step I.** From Theorem 1 we get

\[
\log \Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq E_\pi(F) + \epsilon \right\} \leq -(n - 1)H(W\|W^1 \times P)
\]

for an appropriate \( W \)

**Step II.** Using Pinsker’s and then Jensen’s inequality we bound

\[
H(W\|W^1 \times P) \geq \frac{1}{2} \left[ \sum_{x,y} W^1(x)|P(x,y) - W(y|x)| \right]^2 \quad (*)
\]

**Step III. Lemma.** For any row vector \( v \) with \( \sum_x v(x) = 0 \), we have

\[
\|v(I - P)\| \geq \frac{\alpha}{m} \|v\|
\]

**Step IV.** Get bounds on the dual of a LP related to (*) \( \square \)
Extend to Geometrically Ergodic Chains?

→ In many applications, we are interested in *unbounded* functions $F$

→ Most chains found in applications (like MCMC) are *not Doeblin*, but geometrically ergodic

**Defn**  A Markov chain $\{X_n\}$ is **geometrically ergodic** iff it converges to equilibrium exponentially fast, *not necessarily uniformly in the initial condition*

→ The most general class for which exponential bounds might hold

→ Same bounds *cannot* hold exactly as before

→ But: There *is* a different exponential bound in this case

→ The following example motivates its form . . .
A Hard Example for the Gibbs Sampler: The Witch’s Hat

**Setting:** Use (randomized) Gibbs sampler to compute average of $F(x, y) = e^{5x} + e^{5y}$ w.r.t. the “witch’s hat distr” with $\epsilon = \frac{1}{251}$
A Hard Example for the Gibbs Sampler: The Witch’s Hat

Setting: Use (randomized) Gibbs sampler to compute average of $F(x, y) = e^{5x} + e^{5y}$ w.r.t. the “witch’s hat distr” with $\epsilon = \frac{1}{251}$

Problem: Estimates very sensitive to the rare visits to the “brim”
A Hard Example for the Gibbs Sampler: The Witch’s Hat

**Setting:** Use (randomized) Gibbs sampler to compute average of $F(x, y) = e^{5x} + e^{5y}$ w.r.t. the “witch’s hat distr” with $\epsilon = \frac{1}{251}$

**Problem:** Estimates very sensitive to the rare visits to the “brim”

**Idea:** Consider the new function

$$U(x) = F(x) - E\left[F(X_2) | X_1 = x\right]$$

and note that $E_\pi(U) = 0$

[Cf. Henderson (1997)]
A Sampling Criterion for this Gibbs Sampler

Idea: Together with the averages of $F$ also compute the averages of $U$
A Sampling Criterion for this Gibbs Sampler

**Idea:** Together with the averages of $F$ also compute the averages of $U$

We know: $E_{\pi}(U) = 0$

**Sampling Criterion:**
Sample the $F$-averages **only** when the $U$-averages are between $\pm u$ for some small $u > 0$
More Simulation Results from the Witch’s Hat

Averages of $F$ and sampling times (purple)

Averages of $U$ and time spent at the peak (red)
More Simulation Results from the Witch’s Hat

Averages of $F$ and sampling times (purple)

Averages of $U$ and time spent at the peak (red)
Generally: Geometrically Ergodic Chains

Defn  A Markov chain \( \{X_n\} \) is **geometrically ergodic** iff it converges to equilibrium exponentially fast, *not necessarily uniformly in the initial condition*.

Equivalent characterization  There exists a function \( V : S \to \mathbb{R} \), a finite set \( S_0 \subset S \), and positive constants \( b, \delta \), such that:

\[
E[V(X_2) \mid X_1 = x] - V(x) \leq -\delta V(x) + b \mathbb{I}_{S_0}(x) \quad \text{for all } x
\]

Bounds  
Suppose the function of interest \( F : S \to \mathbb{R} \) is possibly **unbounded**, but with \( \| F^2 \|_V := \sup_x \frac{F(x)^2}{V(x)} < \infty \)

Define a **screening function** \( U(x) = V(x) - E[V(X_2) \mid X_1 = x] \)
Theorem 3

For any geometrically ergodic chain \( \{X_n\} \), any function \( F : S \to \mathbb{R} \) as above, any \( \epsilon, u > 0 \), and any initial condition \( X_1 = x_1 \):

\[
\log \Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq E_\pi(F) + \epsilon \land \left| \frac{1}{n} \sum_{i=1}^{n} U(X_i) \right| \leq u \land X_n \in S_0 \right\}
\]
An Exponential Bound for Geometrically Ergodic Chains

**Theorem 3**

For any geometrically ergodic chain \( \{X_n\} \), any function \( F : S \rightarrow \mathbb{R} \) as above, any \( \epsilon, u > 0 \), and any initial condition \( X_1 = x_1 \):

\[
\log \Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq E\pi(F') + \epsilon \quad \& \quad \left| \frac{1}{n} \sum_{i=1}^{n} U(X_i) \right| \leq u \quad \& \quad X_n \in S_0 \right\} 
\leq -(n - 1) \frac{1}{2} \left[ \left( \frac{\delta}{8\xi \|F^2\| \nu} \right) \left( \frac{\epsilon - \frac{F_{\text{max},0}}{n-1}}{u + b + \frac{U_{\text{max},0}}{n-1}} \right)^2 - \frac{2}{n - 1} \right]^2
\]

where \( F_{\text{max},0} = \max_{x \in S_0} |F(x)| \), \( U_{\text{max},0} = \max_{x \in S_0} |U(x)| \) and \( \xi \) is the “convergence parameter” of the chain.
General Sampling Criterion for Geometrically Ergodic Chains

**Note:** Apart from the fact that the above bound is explicitly computable, it naturally leads us to formulate the following sampling criterion.

**Given:** A geometrically ergodic chain \( \{X_n\} \)
- Its parameters \( V, b, \delta, S_0 \)
- A function \( F \) s.t. \( F^2 \leq CV \)

**Set:** The screening function \( U(x) := V(x) - E[V(X_2)|X_1 = x] \)
- A “small” threshold \( u > 0 \)

**Sampling Criterion:** Sample the results of the chain only at times \( n \) when \( X_n \in S_0 \) and \( |\frac{1}{n} \sum_{i=1}^{n} U(X_i)| \leq u \)

**Explanation:** Control averages and excursions
Comments on the Sampling Criterion

~ Geometric ergodicity in general easy to verify

~ Many choices for \( V(x) \), and \( V \approx F \) often works

~ To apply the sampling criterion, the screening function

\[
U(x) = V(x) - E[V(X_2)|X_1 = x]
\]

needs to be analytically computable

~ Easily so for the Gibbs sampler,
   some versions of the Metropolis algorithm . . .
Comments on Theorem 3

Why is the exponent in Theorem 3 of $O(\epsilon^2)$ and not $O(\epsilon^4)$?

Proof outline similar to one for Doeblin case

Theorem 3 applies even to cases where

\[
\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} F(X_i) \geq E_\pi(F) + \epsilon\right\}
\]

decays sub-exponentially (e.g., discrete $M/M/1$ queue)

How is it that the addition of two non-rare events

\[
\left\{\left|\frac{1}{n}\sum_{i=1}^{n} U(X_i)\right| \leq u\right\} \cap \left\{X_n \in S_0\right\}
\]

makes the probability exponentially small?!

Specialize to the i.i.d. case for an explanation . . .
An “i.i.d. version” of Theorem 3

**Setting:** Estimate $E_P(F)$ where $F$ is “heavy tailed” from i.i.d. samples $X_1, X_2, \ldots \sim P$

Suppose we have a $U$ with known $E_P(U) = 0$, s.t.

$U$ “dominates” $F$: $\operatorname{ess\ sup}[F(X) - \beta U(X)] < \infty$, for all $\beta > 0$

Assume $E_P(F^2), E_P(U^2)$ both finite
An “i.i.d. version” of Theorem 3

Setting: Estimate $E_P(F)$ where $F$ is “heavy tailed” from i.i.d. samples $X_1, X_2, \ldots \sim P$
Suppose we have a $U$ with known $E_P(U) = 0$, s.t.
$U$ “dominates” $F$: $\text{ess sup}[F(X) - \beta U(X)] < \infty$, for all $\beta > 0$
Assume $E_P(F^2), E_P(U^2)$ both finite

Theorem 4

(i) The “standard” error prob is subexponential: $\forall \epsilon > 0$:
$$\lim_{n \to \infty} -\frac{1}{n} \log \text{Pr} \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq E_P(F) + \epsilon \right\} = 0$$
An “i.i.d. version” of Theorem 3

**Setting:** Estimate $E_P(F)$ where $F$ is “heavy tailed” from i.i.d. samples $X_1, X_2, \ldots \sim P$

Suppose we have a $U$ with known $E_P(U) = 0$, s.t.

$U$ “dominates” $F$: $\text{ess sup}[F(X) - \beta U(X)] < \infty$, for all $\beta > 0$

Assume $E_P(F^2), E_P(U^2)$ both finite

**Theorem 4**

(i) The “standard” error prob is subexponential: $\forall \epsilon > 0$:

$$
\lim_{n \to \infty} -\frac{1}{n} \log Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq E_P(F) + \epsilon \right\} = 0
$$

(ii) The “screening” error prob is exponential: $\forall \epsilon, u > 0$:

$$
\lim_{n \to \infty} -\frac{1}{n} \log Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq E_P(F) + \epsilon \& \left| \frac{1}{n} \sum_{i=1}^{n} U(X_i) \right| \leq u \right\} > 0
$$
(i) $\Pr\{\text{standard error}\} \approx \exp\left\{-n \inf_{Q \in \Sigma} H(Q \| P)\right\}$

where $\Sigma = \{Q : E_Q(F) \geq E_P(F) + \epsilon\}$ and the infimum is $= 0$
Geometrical Explanation of Theorem 4

(i) \( \Pr\{\text{standard error}\} \approx \exp\left\{ -n \inf_{Q \in \Sigma} H(Q\|P) \right\} \)

where \( \Sigma = \{Q : E_Q(F) \geq E_P(F) + \epsilon\} \) and the infimum is \( = 0 \)

(ii) \( \Pr\{\text{screening error}\} \approx \exp\left\{ -n \inf_{Q \in E} H(Q\|P) \right\} = \exp\left\{ -n H(Q^*\|P) \right\} \)

where \( E = \{Q : E_Q(F) \geq E_P(F) + \epsilon, \ |E_Q(U)| < u\} \)

and the infimum is \( > 0 \)
(iii) The “screening” error prob satisfies:

Let $K > 0$ arbitrary. Then $\forall \epsilon > 0$, $0 < u \leq K\epsilon$

$$\log \Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} F(X_i) \geq E_P(F) + \epsilon \quad \& \quad \left| \frac{1}{n} \sum_{i=1}^{n} U(X_i) \right| \leq u \right\}$$

$$\leq -\frac{n}{2} \left[ \frac{M}{M^2 + (1 + \frac{1}{2K})^2} \right]^2 \epsilon^2$$

where $M = \text{ess sup} \left[ F(X) - \frac{1}{2K} U(X) \right]$
Theorem 4: A Heavy-Tailed Simulation Example

\[ \frac{1}{k} \sum_{i=1}^{k} X_i^{3/4} \]

Sampling times
Concluding Remarks

Information-Theoretic Methods
Convexity, elementary properties
Strikingly effective in a brutally technical area...

Markov Chain Bounds
Doeblin chains
Geometrically ergodic chains
Functional analysis and optimization
A new sampling criterion
Further applications in MCMC...
Simulating a Simple Queue in Discrete Time

Consider: The chain $X_{n+1} = [X_n - S_{n+1}]_+ + A_{n+1}$ where:

\begin{align*}
\{A_n\} \text{ i.i.d. } &\sim (1 + \kappa)\alpha \cdot \text{Bern}\left(\frac{1}{1+\kappa}\right) \quad \text{and} \quad \{S_n\} \text{ i.i.d. } \sim 2\mu \cdot \text{Bern}\left(\frac{1}{2}\right) \\
\text{the load } \rho = \frac{E(A_k)}{E(S_n)} = \frac{\alpha}{\mu} \text{ is heavy, } \rho \approx 1, \quad \text{and} \quad F(x) = x
\end{align*}
Simulating a Simple Queue in Discrete Time

Consider: The chain $X_{n+1} = [X_n - S_{n+1}]_+ + A_{n+1}$ where:

- $\{A_n\}$ i.i.d. $\sim (1 + \kappa)\alpha \cdot \text{Bern}(\frac{1}{1+\kappa})$ and $\{S_n\}$ i.i.d. $\sim 2\mu \cdot \text{Bern}(\frac{1}{2})$
- The load $\rho = \frac{\mathbb{E}(A_k)}{\mathbb{E}(S_n)} = \frac{\alpha}{\mu}$ is heavy, $\rho \approx 1$, and $F(x) = x$

Then: $\{X_n\}$ is geometrically ergodic with $V(x) = e^{\epsilon x}$

$U(x) = V(x) - E[V(X_2)|X_1 = x]$ is an easily computable quadratic

No exponential error bound can be proved on the error probability!
Simulating a Simple Queue in Discrete Time

Consider: The chain $X_{n+1} = [X_n - S_{n+1}]_+ + A_{n+1}$ where:

$\{A_n\}$ i.i.d. $\sim (1 + \kappa)\alpha \cdot \text{Bern}(\frac{1}{1+\kappa})$ and $\{S_n\}$ i.i.d. $\sim 2\mu \cdot \text{Bern}(\frac{1}{2})$

the load $\rho = \frac{E(A_k)}{E(S_n)} = \frac{\alpha}{\mu}$ is heavy, $\rho \approx 1$, and $F(x) = x$

Then: $\{X_n\}$ is geometrically ergodic with $V(x) = e^{\epsilon x}$

$U(x) = V(x) - E[V(X_2)|X_1 = x]$ is an easily computable quadratic

No exponential error bound can be proved on the error probability!