

Quantitative Bounds for Markov Chain
Convergence: Total Variation and
Wasserstein distances

A. Deniz Sezer
University of Calgary
(joint work with Neal Madras)

October 15, 2008

X : an ergodic Markov Chain with state space S .

π : Stationary distribution of X .

Let d be a metric on the space of probability distributions s.t.
 $d(P_x^n, \pi) \rightarrow 0$.

Goal: To find an explicit function $g(x, n)$ such that

$$d(P_x^n, \pi) \leq g(x, n)$$

Motivation: MCMC

In many applications it is desirable to sample from a given probability distribution π on a state space S .

Example: Bayesian estimation

θ : parameter vector

Y : data

Distribution of Y : $f(Y|\theta) \times p(\theta)d\theta$

$p(\theta)d\theta$: prior distribution of θ

Goal: To compute $\bar{\theta} = E(\theta|X)$.

(Bayesian analogue of sample mean).

To do this we can sample $\theta_1, \dots, \theta_n$ from the the posterior distribution, $p(\theta|X)$, of θ and then compute their average.

What is an MCMC algorithm?

$(X_n)_{n \geq 1}$: a Markov Chain with state space S , with stationary distribution π .

Instead of sampling from π , run X_n long enough.

A quantitative bound gives an exact estimate for the number of steps we need to run this Markov chain to reach the desired accuracy.

Why Markov Chain Monte Carlo?

Gibbs sampling:

$\theta = (\theta_1, \dots, \theta_n)$. Suppose we can compute

$$p_i(\theta_i | \{\theta_1, \dots, \theta_n\} - \{\theta_i\}, Y).$$

Set $X_0 = \theta_0$, then iteratively update each component of X_0 , according to p_i . Repeat. This gives a Markov chain whose stationary distribution is

$$p(\theta | X).$$

Advantage: Computation of p_i involves one dimensional integrals only as opposed to $p(\theta | X)$.

Metropolis-Hastings Algorithm: We need an auxiliary Markov chain Z_n with transition density $q(x, y)$. Set $Y_0 = Z_0$. We generate Z_1 , and independently toss a coin with tails probability

$$p = \min\left(1, \frac{\pi(Z_1) q(Z_1, Y_0)}{\pi(Y_0) q(Y_0, Z_1)}\right).$$

If the coin comes up tails we set $Y_1 = Z_1$, otherwise we set $Y_1 = Y_0$ and repeat.

Y_n is a Markov Chain with stationary distribution π .

Advantage: Does not require full knowledge of π . It is enough if we only know

$$\pi \sim \nu$$

where ν is a measure but not a probability measure.

This is ideal for Bayesian estimation, because

$$p(\theta|X) \sim p(\theta)p(X|\theta)$$

which can be easily computed.

Goal: To find an explicit function $g(x, n)$ such that

$$d(P_x^n, \pi) \leq g(x, n)$$

Choice of d :

Total Variation (TV) metric (denoted d_{TV})

$$d_{TV}(\mu_1, \mu_2) = \sup\{|\mu_1(A) - \mu_2(A)| : A \subset \mathcal{X}\} \quad (1)$$

Coupling representation of $d_{TV}(\mu_1, \mu_2)$:

$$d_{TV}(\mu_1, \mu_2) = \inf(\Pr\{X_1 \neq X_2\} : X_1 \sim \mu_1, X_2 \sim \mu_2).$$

Convergence in total variation distance

Geometric ergodicity:

$$d_{TV}(P_x^n \pi) \leq C_x \rho^n$$

Uniform ergodicity:

$$d_{TV}(P_x^n, \pi) \leq C \rho^n \text{ (C does not depend on } x\text{.)}$$

How to find ρ ?

Operator theory approach:

$$\pi\text{-reversibility: } P(x, dy)\pi(dx) = P(y, dx)\pi(dy)$$

makes P a self-adjoint operator acting on $L^2(\pi)$.

Theorem: For a reversible Markov Chain, (under mild conditions) P_x^n converges to π in total variation distance if and only if P has an $L^2(\pi)$ spectral gap, ρ . Moreover, $\exists C_x < \infty$ such that

$$d_{TV}(P_x^n, \pi) \leq C_x(1 - \rho)^n$$

- Not practical in continuous state spaces.

Coupling approach:

Doebelin's condition holds if there exists a probability measure ν and $0 < \epsilon < 1$ such that

$$P(x, dy) \geq \epsilon \nu(dy).$$

Theorem: If Doebelin's condition holds then,

$$\|P_x^n - \pi\|_{\text{var}} \leq (1 - \epsilon)^n$$

Minorization and drift conditions (Rosenthal)

It is possible to get similar bounds using coupling when Doeblin's condition holds only on a set C , if a drift function to C exists. More precisely, one needs

$$P(x, dy) \geq \epsilon \nu(dy) \text{ for } x \in C,$$

and a function $V > 1$ and $\alpha > 1$ s.t. $E(V(Y_{n+1})|Y_{n+1} = y) < V(y)/\alpha$ whenever $y \in C^c$.

Alternative metric: Wasserstein metric (denoted W).

$$W(\mu_1, \mu_2) = \inf \left\{ \int_{\mathcal{X}} \int_{\mathcal{X}} \rho(x, y) M(dx, dy) : M \in \text{Joint}(\mu_1, \mu_2) \right\}$$

Coupling representation:

$$W(\mu_1, \mu_2) = \inf(E(\rho(X_1, X_2)) : X_1 \sim \mu_1, X_2 \sim \mu_2).$$

Our strategy:

First get a quantitative bound for convergence in the Wasserstein metric.

Random dynamical systems gives us an elegant framework to do this.

Second, find a relationship between total variation distance and the Wasserstein distance

An iterated function system is a sequence of random maps of the form

$$F_n(x) = f_1 \circ f_2 \circ \dots \circ f_n(x)$$

or

$$\tilde{F}_n(x) = f_n \circ f_{n-1} \circ \dots \circ f_1(x)$$

where f_1, f_2, \dots are independent and identically distributed (i.i.d.) random maps.

- $\{\tilde{F}_n(x) : n \geq 1\}$ is a Markov chain.
- $\{F_n(x) : n \geq 1\}$ is not a Markov chain, but under certain conditions converges pointwise to a random variable, X_∞ , independent of the starting point x . (In this case the system is called attractive)

Suppose X_∞ exists. Then

$$W(P^n(x, \cdot), \pi) \leq E[\rho(F_n(x), X_\infty)].$$

So one can try to get a bound for $E[\rho(F_n(x), X_\infty)]$.

One condition that guarantees attractivity is strong contractivity.
That is,

$$E[\log \text{Lip}f] < 0.$$

This condition is a generalization of the stronger condition that

$$\rho(f(x), f(y)) < r\rho(x, y)$$

a.s. all f .

D. Steinsaltz's local contractivity condition and convergence theorem

Definition: An iterated function system is locally contractive if there exists a drift function $\phi : \mathcal{X} \mapsto [1, \infty)$ and $r \in (0, 1)$ such that

$$G_n(x) := E[D_x F_n] \leq \phi(x)r^n$$

where $D_x f := \limsup_{y \rightarrow x} \frac{\rho(f(x), f(y))}{\rho(x, y)}$.

Theorem: (D. Steinsaltz) If an iterated function system is locally contractive with a drift function ϕ and if

$$C_x := E[\rho(f(x), x) \sup_{0 \leq t \leq 1} \{\phi(x + t(f(x) - x))\}] < \infty$$

then

$$W(F_n(x), F_\infty(y)) \leq E\rho(F_n(x), F_\infty(y)) \leq \left(\frac{C_x}{1-r} + \rho(x, y)\Phi(x; y) \right) r^n$$

where

$$\Phi(x, y) = \sup_{0 \leq t \leq 1} \{\phi(x + t(y - x))\}.$$

Growth condition: A continuous function $\phi : \mathcal{X} \mapsto [1, \infty)$ is a drift function if

$$r := \sup_x E \left[\frac{\phi(f(x))}{\phi(x)} D_x f \right] < 1.$$

Proof: Let

$$\mathcal{L}(g)(x) \doteq E[g(f(x))D_x f].$$

Then,

$$G_n(x) = \mathcal{L}^n(1)(x),$$

Note that the growth condition is equivalent to

$$\mathcal{L}\phi \leq r\phi.$$

(We call such a ϕ an r -sub-eigenfunction for \mathcal{L} .) So, if $\phi > 1$ and is an r -sub-eigenfunction then

$$\mathcal{L}^n 1 \leq \mathcal{L}^n \phi \leq r^n \phi.$$

How to apply the local contractivity convergence theorem: Finding a drift function:

The first strategy: Find a linear operator $\tilde{\mathcal{L}}$ s.t.

(1) $\tilde{\mathcal{L}} \geq \mathcal{L}$,

(2) $\tilde{\mathcal{L}}$ is simpler to manage.

One kind of operator that we can manage is the following:

Let $\{A_i\}_{i=1}^n$ be a finite partition of the state space S and

$$\tilde{\mathcal{L}}\phi(x) = b(x) \sum_{i=1}^n \mathbf{1}_{A_i}(x) \int_S \phi(s) \mu_i(ds)$$

where $b(x)$ is a non-negative function and each μ_i is a finite measure on \mathcal{S} .

Theorem: In order for $\tilde{\mathcal{L}}$ to have an r -sub-eigenfunction it is necessary and sufficient that the matrix

$$Q(i, j) = \int_{A_i} b(x) \mu_j(dx)$$

has an r -sub-eigenvector $p = (p_1, p_2, \dots, p_n)$, i.e (a non-negative p s.t. $pQ \leq rp$). Moreover, if p is an r -sub-eigenvector for Q then the function

$$\phi(x) = \sum_{i=1}^n p_i 1_{A_i}(x) b(x)$$

(and any positive multiple of it) is an r -sub-eigenfunction for $\tilde{\mathcal{L}}$.

The second strategy: Suppose

$$\mathcal{L}(\phi)(x) = \int \phi(y)K(x, dy)$$

and let $\tilde{\mathcal{L}}$ be related to \mathcal{L} as follows:

$$\tilde{\mathcal{L}} = \frac{1}{h(x)} \int \phi(y)h(y)K(x, dy)$$

where h is a strictly positive function.

- If ϕ is an r -sub-eigenfunction for \mathcal{L} then $\frac{\phi}{h}$ is an r -sub-eigenfunction for $\tilde{\mathcal{L}}$.

From Wasserstein Distance to Total Variation Distance

One-shot coupling: Let $\{f_t\}$ be a sequence of i.i.d. random maps.

Let S_0 and \tilde{S}_0 be two r.v.'s, and let

$$S_t = f_t(S_{t-1}) \text{ and } \tilde{S}_t = f_t(\tilde{S}_{t-1}) \text{ for } t = 1, \dots, n-1.$$

Suppose at time n , we can find two copies of f_t , f_n and \hat{f}_n , so that with high probability we have “one-shot coupling”, i.e.

$$S_n = f_n(S_{n-1}) = \hat{f}_n(\hat{S}_{n-1}) = \hat{S}_n.$$

Theorem: (a) Assume that there is a constant A such that

$$\int_{\chi} |p(x, z) - p(y, z)| \lambda(dz) \leq A \rho(x, y) \text{ for all } x, y \in \chi.$$

Then

$$d_{TV}(\mu P^n, \pi) \leq \frac{A}{2} W(\mu P^{n-1}, \pi) \text{ for all } n \geq 1.$$

(b) Assume the following conditions hold:

i) There exists $h \geq 0$ on χ such that

$$\int_{\chi} |p(x, z) - p(y, z)| \lambda(dz) \leq \frac{\rho(x, y)}{\max\{h(x), h(y)\}} \quad \text{for all } x, y \in \chi,$$

(ii) There are $B > 0$ and $q > 0$ such that

$$\pi(\{y : h(y) < \epsilon\}) \leq B\epsilon^q \quad \text{for all sufficiently small positive } \epsilon,$$

(iii) $\lim_{n \rightarrow \infty} W(\mu P^n, \pi) = 0$.

Then there exists a number \tilde{C} , independent of n , such that

$$d_{TV}(\mu P^n, \pi) \leq \tilde{C} [W(\mu P^{n-1}, \pi)]^{q/(1+q)} \quad \text{for all sufficiently large } n.$$

Example: Normal Gibbs Sample

A random sample $Y := Y_1, \dots, Y_J$ and two random variables θ and σ ,

$$Y_i | \theta, \sigma \sim \text{i.i.d } N(\theta, \sigma^2)$$

$$\theta \sim N(\xi, K^{-1})$$

The prior distribution of $S = \sigma^{-2} \sim \Gamma(\alpha, \beta)$.

We are interested in bounding the rate of convergence of the Gibbs sampler of the posterior distribution of this model.

Without loss of generality, we can assume that ξ is zero and that K is either 0 or 1.

$$\bar{Y} = \frac{1}{J} \sum_{j=1}^J Y_j \quad \text{and} \quad \Sigma_0 = \beta + \frac{1}{2} \sum_{j=1}^J (Y_j - \bar{Y})^2$$

$K=1$: If

$$r \doteq \frac{1}{\sqrt{2\pi}} \left(2|\bar{Y}| \log \left(\frac{J(\alpha + \frac{J}{2})}{\Sigma_0} + 1 \right) + \left(1 - \frac{1}{\sqrt{\frac{J(\alpha + \frac{J}{2})}{\Sigma_0} + 1}} \right) \right) < 1,$$

then

$$W(P_x^n, \pi) \leq \frac{C_{\epsilon, x}}{1 - r_\epsilon} r_\epsilon^n$$

where $r_\epsilon = r + \epsilon A$

$$A = \frac{(|\bar{Y}| + 1)(\alpha + \frac{J}{2})J\sqrt{2\pi}}{\Sigma_0^2}$$

$$C_{\epsilon, x} = \frac{J(2|\bar{Y}| + 1)}{\epsilon 2\sqrt{2\pi}} \left(|x| + \frac{\alpha + J/2}{\Sigma_0} \right).$$

$K = 0$. Let

$$r_\epsilon = \frac{1}{2\alpha + J - 2} + \epsilon \frac{2\alpha + J}{\Sigma_0^2}.$$

Then

$$W(P_x^n, \pi) \leq \frac{C_{\epsilon, x} r_\epsilon^n}{1 - r_\epsilon}$$

where

$$C_{\epsilon, x} \leq \frac{1}{\epsilon} \max\left(\frac{1}{x^2}, \epsilon\right) \left(x + \frac{\alpha + J/2}{\Sigma_0}\right).$$

Theorem: Let μ be an arbitrary initial probability distribution on $(0, \infty)$. Then

$$d_{TV}(\mu P_1^n, \pi_1) \leq \frac{J}{2} \left(1 + \frac{|\bar{Y}|}{\sqrt{2\pi}} \right) W(\mu P_1^{n-1}, \pi_1) \text{ for } n = 1, 2, \dots$$

and there is a constant \tilde{C} (independent of n) such that

$$d_{TV}(\mu P_0^n, \pi_0) \leq \tilde{C} W(\mu P_0^{n-1}, \pi_0)^w \text{ for all sufficiently large } n$$

where $w = (2\alpha + J - 1)/(2\alpha + J + 1)$.