Quasi-Reliable Estimates of Effective Sample Size

Robert Skeel

Purdue University / Arizona State University

July 20, 2016

with Youhan Fang and Yudong Cao

Evaluating Performance of MCMC Methods

Statistician Charles Geyer (1992):

> *It would enforce a salutary discipline if the gold standard for comparison of Markov chain Monte Carlo schemes were asymptotic variance . . . it is easier to invent methods than to understand exactly what their strengths and weaknesses are . . . .*

Physicist Alan Sokal (1997):

> *the better algorithm is the one that has the smaller autocorrelation time, when time is measured in units of CPU time*

The Cost of an Independent Sample

$$\tau \times \text{cost per step} = \frac{N}{ESS} \times \text{cost per step}$$

where τ is the *integrated autocorrelation time*,
$N =$ sample size, and
ESS is *effective sample size*.

The Cost of an Independent Sample

$$\tau \times \text{cost per step} = \frac{N}{ESS} \times \text{cost per step}$$

where τ is the *integrated autocorrelation time*,
$N =$ sample size, and
ESS is *effective sample size*.

> However, not generally accepted for molecular simulation, because very often ... $ESS = 0$.

The Cost of an Independent Sample

$$\tau \times \text{cost per step} = \frac{N}{ESS} \times \text{cost per step}$$

where τ is the *integrated autocorrelation time*,
N = sample size, and
ESS is *effective sample size*.

However, not generally accepted for molecular simulation,
because very often ... $ESS = 0$.

Nonetheless, for *comparing* methods;
one can simply use small representative problems.

Evaluating Accuracy of MCMC Estimates

Alan Sokal again:

> *It is important to emphasize that unless further information is given—namely, the autocorrelation time of the algorithm—*. . . statements [of sample size] have no value whatsoever*.*

What about problems for which even $ESS = 1$ is infeasible?

What to Do With Intractable Problems

1. Do "exploration" instead of sampling.

2. Change the problem:
   - 2.1 Use a simpler model (Daniel Zuckerman).
     A coarsened model with error bars is better
     than a refined model without error bars
     (because then the limitations are more explicit).

   - 2.2 Artificially limit the extent of state space,
     and adjust conclusions accordingly.

Reliability and Quasi-reliability

Reliability is impractical in general due to metastability.

Here
quasi-reliability means ensuring thorough sampling
of that part of state space that has already been explored,
to minimize the risk of missing an opening
to an unexplored part of state space.

More concretely, it means
thorough sampling of those modes that have already been visited,
to minimize the risk of missing an opening to an unvisited mode.

Outline

Problem Statement

Given probability density $\rho_{\mathrm{q}}(\mathbf{q})$, $\mathbf{q} \in \mathbb{R}^d$,
known only up to a multiplicative factor,
compute observables $\mathbb{E}[u(\mathbf{Q})]$,
which are expectations for specified functions $u(\mathbf{q})$.

Note use of upper case for random variables.

Markov Chains

$$\mathbf{Q}_0 \to \mathbf{Q}_1 \to \cdots \to \mathbf{Q}_{N-1}$$

Assume ergodic with stationary density $\rho_{\mathrm{q}}(\mathbf{q})$.
Also, assume $\mathbf{Q}_0 \sim \rho_{\mathrm{q}}(\mathbf{q})$.

To estimate an observable, use

$$\mathbb{E}[u(\mathbf{Q})] \approx \overline{U} = \frac{1}{N} \sum_{n=0}^{N-1} u(\mathbf{Q}_n),$$

but use just one realization

$$\mathbf{q}_0 \to \mathbf{q}_1 \to \cdots \to \mathbf{q}_{N-1}.$$

Variance of the Estimated Mean

$$\mathrm{Var}[\overline{U}] = \frac{1}{N}\mathrm{Var}[u(\mathbf{Q})]\left(1 + 2\sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)\frac{C(k)}{C(0)}\right)$$

where the covariances

$$C(k) = \mathbb{E}[(u(\mathbf{Q}_0) - \mu)(u(\mathbf{Q}_k) - \mu)], \quad \text{with } \mu = \mathbb{E}[u(\mathbf{Q})]).$$

In the limit $N \to \infty$,

$$\mathrm{Var}[\overline{U}] = \frac{1}{N}\mathrm{Var}[u(\mathbf{Q})]\tau + o(\frac{1}{N})$$

where τ is the integrated autocorrelation time

$$\tau = 1 + 2\sum_{k=1}^{+\infty}\frac{C(k)}{C(0)}.$$

Estimates of CoVariances

$$C_N(k) = \frac{1}{N} \sum_{n=0}^{N-k-1} (u(\mathbf{q}_n) - \overline{u})(u(\mathbf{q}_{n+k}) - \overline{u}).$$

Priestley (1981, pp.323–324)

MCMC in Extended State Space

Some samplers
augment state variables $\mathbf{q}$ with auxiliary variables $\mathbf{p}$,
extend the p.d.f. to $\rho(\mathbf{q}, \mathbf{p})$ so that

$$\int \rho(\mathbf{q}, \mathbf{p}) \mathrm{d}\mathbf{p} = \rho_{\mathrm{q}}(\mathbf{q}),$$

and make moves in extended state space

$$\mathbf{z}_0 \to \mathbf{z}_1 \to \cdots \to \mathbf{z}_{N-1}$$

where $\mathbf{z} = (\mathbf{q}, \mathbf{p})$.

Forward Transfer Operator

Associated with the MCMC propagator is an operator $\mathcal{F}$, which maps a relative density $u_{n-1} = \rho_{n-1}/\rho$ for $\mathbf{Z}_{n-1}$ to a relative density $u_n = \rho_n/\rho$ for $\mathbf{Z}_n$:

$$u_n = \mathcal{F} u_{n-1} \quad \text{where } \mathcal{F} u_{n-1}(\mathbf{z}) = \frac{1}{\rho(\mathbf{z})} \int \rho(\mathbf{z}|\mathbf{z}') u_{n-1}(\mathbf{z}') \rho(\mathbf{z}') \mathrm{d}\mathbf{z}'$$

and $\rho(\mathbf{z}|\mathbf{z}')$ is the transition probablity for the chain.

Note that $u \equiv 1$ is an eigenfunction of $\mathcal{F}$ for eigenvalue $\lambda_1 = 1$.

Error in Probability Density

starting from $\rho_0(\mathbf{q}) = \delta(\mathbf{q} - \mathbf{q}_0)$ is proportional to $1/|1 - \lambda_2|$, where λ_2 is the nonunit eigenvalue of $\mathcal{F}$ nearest 1.

However, the spectral gap is not the relevant quantity—in general.

Error in Estimate of τ

An estimate of τ has a standard deviation $\approx \sqrt{M/N}\tau$
where M is the number of terms taken in formula for τ.
Therefore, use instead

$$\tau \approx 1 + 2 \sum_{k=1}^{N-1} w(k) \frac{C(k)}{C(0)}$$

where $w(k)$ is a decreasing function known as a *lag window*.

The tiny program `acor` of Jonathan Goodman (2009)
uses a lag window that is 1 from 0 to roughly 1.4τ
and decreases linearly from 1.4τ to 2.4τ.
It requires the number of samples N to exceed 35τ, roughly.

Outline

Preliminaries

Thorough Sampling

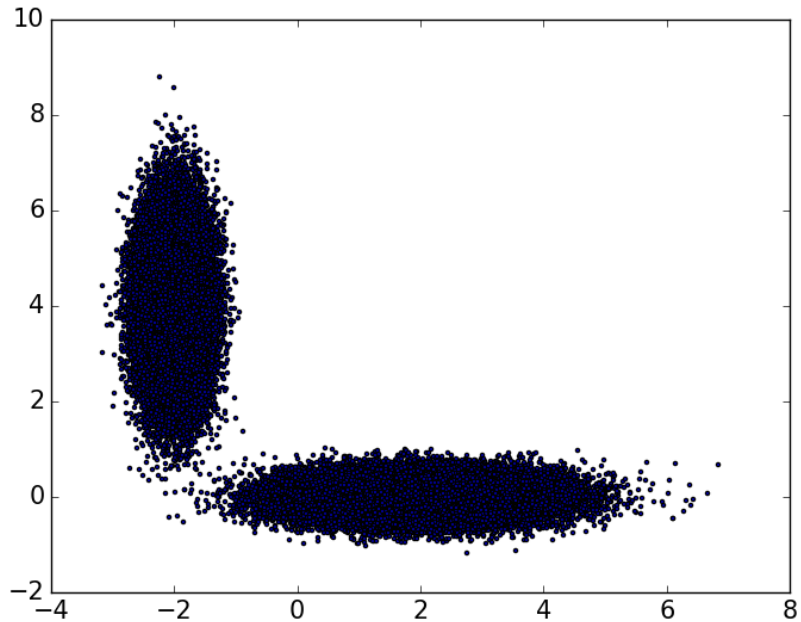
Coping with Statistical Error

Reversible Samplers

Weakness of Existing Approach

Mixture of Gaussians

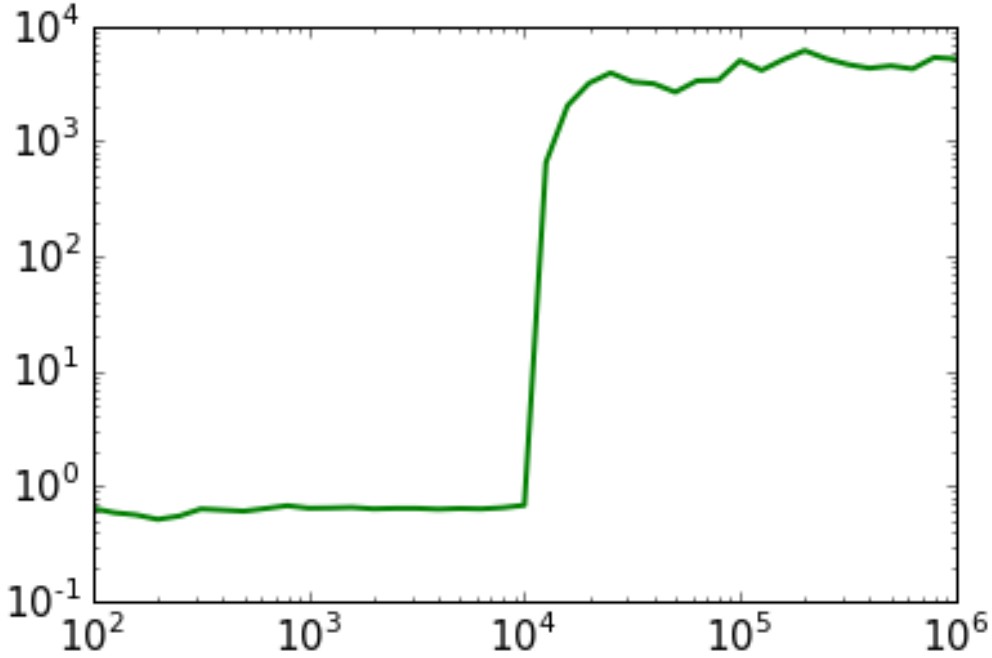
$$-\log \rho(q_1, q_2) =$$
$$\frac{1}{2}(16(q_1 + 2)^2 + (q_2 - 4)^2) + \frac{1}{2}((q_1 - 2)^2 + 16q_2^2) + \text{const}$$



Observable function: $u(q_1, q_2) = q_1$.

Estimate of τ as N Increases

Sampler: Euler-Maruyama Brownian dynamics (w/o rejections).



Need to ensure thorough sampling.

Definition of Thorough Sampling:

for any subset A of state space,
an estimate $\overline{1_A}$ of $\mathbb{E}[1_A(\mathbf{Q})] = \mathsf{Pr}(\mathbf{Q} \in A)$ should satisfy

$$\mathrm{Var}[\overline{1_A} - \mathbb{E}[1_A(\mathbf{Q})]] \leq \frac{1}{4} tol^2.$$

Definition of Thorough Sampling:

for any subset A of state space,
an estimate $\overline{1_A}$ of $\mathbb{E}[1_A(\mathbf{Q})] = \Pr(\mathbf{Q} \in A)$ should satisfy

$$\mathrm{Var}[\overline{1_A} - \mathbb{E}[1_A(\mathbf{Q})]] \leq \frac{1}{4} tol^2.$$

Since

$$\mathrm{Var}[\overline{1_A} - \mathbb{E}[1_A(\mathbf{Q})]] \approx \tau_A \frac{1}{N} \mathrm{Var}[1_A(\mathbf{Q})] \leq \frac{1}{4N} \tau_A,$$

it is enough to have

$$\frac{1}{4N} \tau_A \leq \frac{1}{4} tol^2 \quad \text{for all } A.$$

Symmetries

There may be permutations P such that
$\rho_{\mathrm{q}}(P\mathbf{q}) = \rho_{\mathrm{q}}(\mathbf{q})$ and
$u(P\mathbf{q}) = u(\mathbf{q})$ for all interesting u.

Then, consider only A for which $1_A(P\mathbf{q}) = 1_A(\mathbf{q})$.

Symmetries

There may be permutations P such that
$\rho_{\mathrm{q}}(P\mathbf{q}) = \rho_{\mathrm{q}}(\mathbf{q})$ and
$u(P\mathbf{q}) = u(\mathbf{q})$ for all interesting u.

Then, consider only A for which $1_A(P\mathbf{q}) = 1_A(\mathbf{q})$.

For simplicity, instead of only indicator functions,
consider all functions in

$$W = \{u = u(\mathbf{q}) \mid \mathbb{E}[u(\mathbf{Q})] = 0, u(P\mathbf{q}) = u(\mathbf{q}) \text{ for symmetries } P\}.$$

For More Explicit Notation

Introduce the inner product

$$\langle v, u \rangle = \int \overline{v(\mathbf{z})} u(\mathbf{z}) \rho(\mathbf{z}) \, \mathrm{d}\mathbf{z}.$$

Note that $\langle 1, u \rangle = \mathbb{E}[u(\mathbf{Z})]$.

One can show by induction that

$$\mathbb{E}[v(\mathbf{Z}_0) u(\mathbf{Z}_k)] = \langle \mathcal{F}^k v, u \rangle.$$

and, in particular, $C(k) = \langle \mathcal{F}^k u, u \rangle$.

Maximum Autocorrelation Time

Define thorough sampling using

$$\tau_{\max} = \max_{u \in W} \left(1 + 2 \sum_{k=1}^{+\infty} \frac{\langle \mathcal{F}^k u, u \rangle}{\langle u, u \rangle} \right).$$

Spatial Discretization

Consider $u(\mathbf{q}) = \mathbf{a}^{\mathsf{T}}\mathbf{u}(\mathbf{q})$
where $u_i \in W$ are given and a_i are unknown.
Then

$$C(k) = \langle \mathcal{F}^k u, u \rangle = \mathbf{a}^{\mathsf{T}} C_k \mathbf{a}$$

where

$$C_k = \langle \mathcal{F}^k \mathbf{u}, \mathbf{u}^{\mathsf{T}} \rangle = \mathbb{E}[\mathbf{u}(\mathbf{Q}_0)\mathbf{u}(\mathbf{Q}_k)^{\mathsf{T}}]$$

and

$$\tau_{\max} \approx \max_{\mathbf{a}} \frac{\mathbf{a}^{\mathsf{T}} K \mathbf{a}}{\mathbf{a}^{\mathsf{T}} C_0 \mathbf{a}} \quad \text{where } K = C_0 + 2\sum_{k=1}^{+\infty} C_k$$

—an eigenvalue problem.

For $\mathbf{u}$, suggest $u_i(\mathbf{q}) = q_i$.

Cross Covariance Matrices

C_0 is symmetric positive definite.

The eigenvalue problem is well conditioned if C_k is symmetric.

This depends on the sampler.

Reversible MCMC Samplers

satisfy detailed balance:

$$\rho(\mathbf{z}'|\mathbf{z})\,\rho(\mathbf{z}) = \rho(\mathbf{z}|\mathbf{z}')\,\rho(\mathbf{z}')$$

e.g.,

- Brownian sampler, one step of the Euler-Maruyama integrator for Brownian dynamics with or without a MRRTT (Metropolis-Rosenbluth-**Rosenbluth**-Teller-Teller) acceptance test,
- hybrid Monte Carlo,
- a generalized hybrid Monte Carlo step, followed by a momenta flip,
- several steps with a reversible Langevin integrator, followed by a momenta flip.

Modified Reversible MCMC Samplers

The momenta flip is counterproductive:
Reversing the momenta flip seems to improve mixing.
A modified reversible propagator couples the reversible substep $\mathbf{z}_{n-1} \to \bar{\mathbf{z}}_n$ with a substep,

$$\mathbf{z}_n = R(\bar{\mathbf{z}}_n),$$

where R satisfies $R \circ R = \mathrm{id}$ and $\rho \circ R = \rho$.

For a momenta flip,
$R(\mathbf{q}, \mathbf{p}) = (\mathbf{q}, -\mathbf{p})$.

An Empirical Comparison

Cancès, Legoll & Stoltz (2007)

Discrepancy for a pair of torsion angles in alkanes.

dimension	36	36	87	114
Langevin dynamics	0.034	0.016	0.014	0.017
HMC	0.039	0.012	0.026	0.021
Brownian dynamics	0.079	0.023	0.040	0.031
Nosé-Hoover chains	0.103	0.046	0.029	0.035
Brownian sampler	0.104	0.034	0.048	0.061

1 million samples for column 1, 10 million for others.

Discretized Langevin Dynamics

Let $F(\mathbf{q}) = -\nabla_{\mathbf{q}} \log \rho(\mathbf{q}, \mathbf{p})$.

O: $\mathbf{p} := \sqrt{1 - \gamma \Delta t}\mathbf{p} + \sqrt{\gamma \Delta t}\mathcal{N}(0, I)$

B: $\mathbf{p} := \mathbf{p} + \frac{1}{2}\Delta t F(\mathbf{q})$

A: $\mathbf{q} := \mathbf{q} + \Delta t \mathbf{p}$

B: $\mathbf{p} := \mathbf{p} + \frac{1}{2}\Delta t F(\mathbf{q})$

O: $\mathbf{p} := \sqrt{1 - \gamma \Delta t}\mathbf{p} + \sqrt{\gamma \Delta t}\mathcal{N}(0, I)$

The dynamics has a *precise* stationary density.
It differs from the desired density by $\mathcal{O}(\Delta t^2)$.

Leimkuhler, Matthews & Stoltz (2015)

Such error is much smaller than statistical error.
(The Euler-Maruyama integrator is a special case of this.)

Symmetry of Cross Covariance Matrices

For a modified reversible propagator, its forward transfer operator is

$$\mathcal{F} = \mathcal{R}\bar{\mathcal{F}},$$

where $\bar{\mathcal{F}}$ is the operator for the reversible substep
and $\mathcal{R}$ is the operator for the substep $\mathbf{z}_n = R(\bar{\mathbf{z}}_n)$.

It can be shown that

$$\langle \bar{\mathcal{F}} g, f \rangle = \langle g, \bar{\mathcal{F}} f \rangle \quad \text{and} \quad \langle \mathcal{R} g, f \rangle = \langle g, \mathcal{R} f \rangle.$$

Additionally, assume $\mathcal{R}\mathbf{u}(\mathbf{q}) = \mathbf{u}(\mathbf{q})$.
Then matrices C_k can be shown to be symmetric
modulo sampling error.

Outline

Preliminaries

Thorough Sampling

Coping with Statistical Error

Reversible Samplers

Simplest Neural Network

Fit model

$$u(\mathbf{q}; x) = q_3 \tanh(q_1 x + q_2) + q_4$$

to data $y_i \approx u(\mathbf{q}; x_i)$:

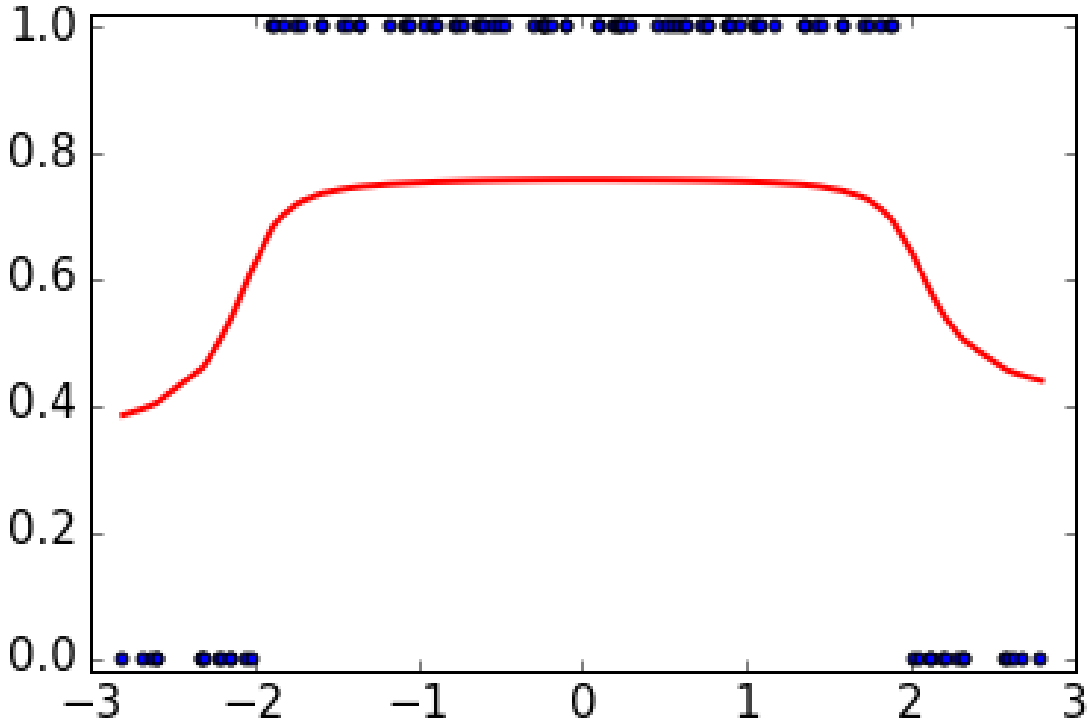
$$-\log \rho(\mathbf{q}) = \frac{1}{2}\beta \sum_{i=1}^{100} (y_i - u(\mathbf{q}; x_i))^2 + \frac{1}{2}\alpha \|\mathbf{q}\|^2 + \text{const}$$

where $\beta = 1.5$ and $\alpha = 0.01$. A Bayesian approach might use the mean

$$\mathbb{E}[u(\mathbf{Q}; x)]$$

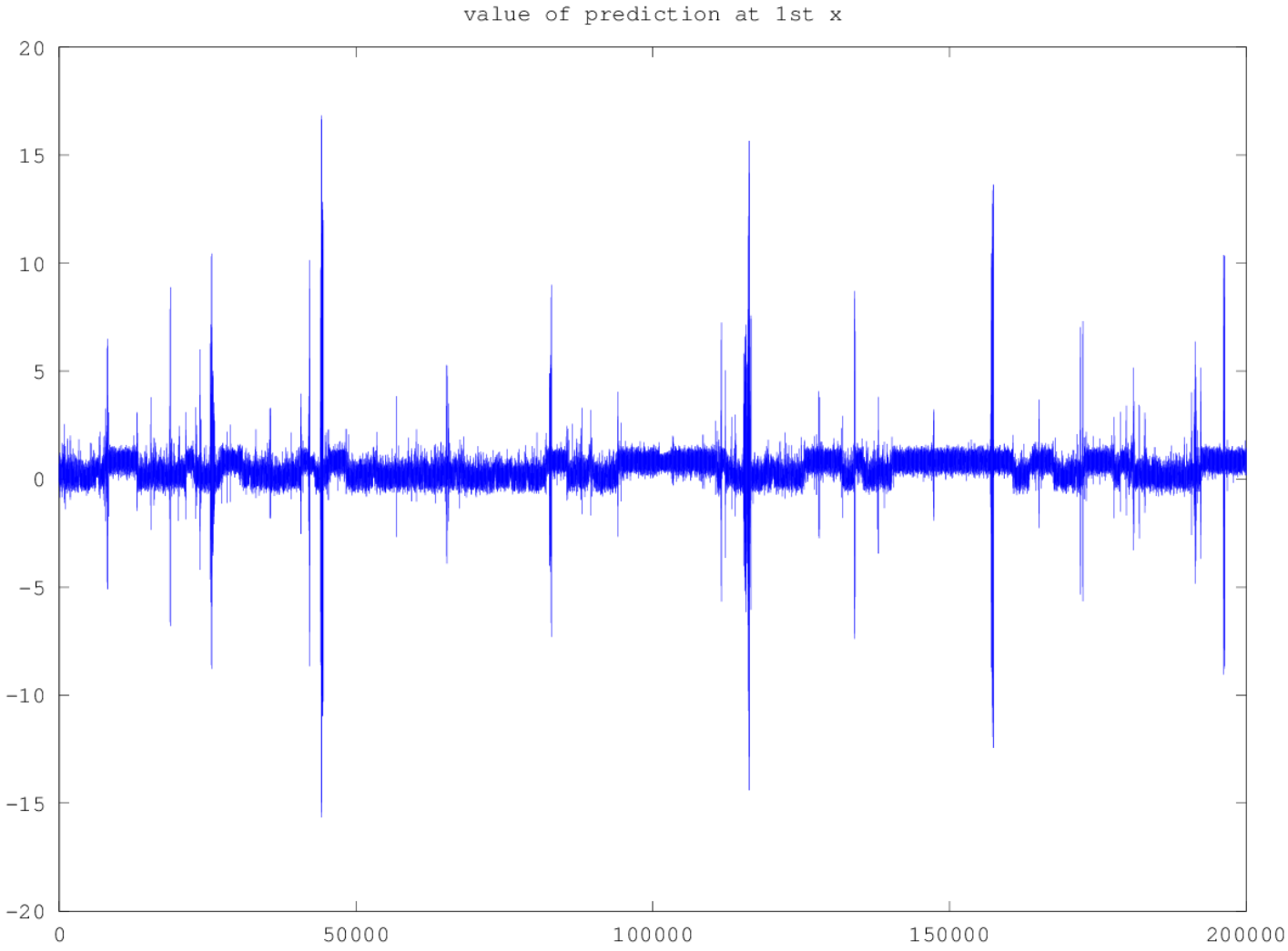
as the approximation.
Sampler: Euler-Maruyama Brownian dynamics w/o rejections.

Data and Predicted Function

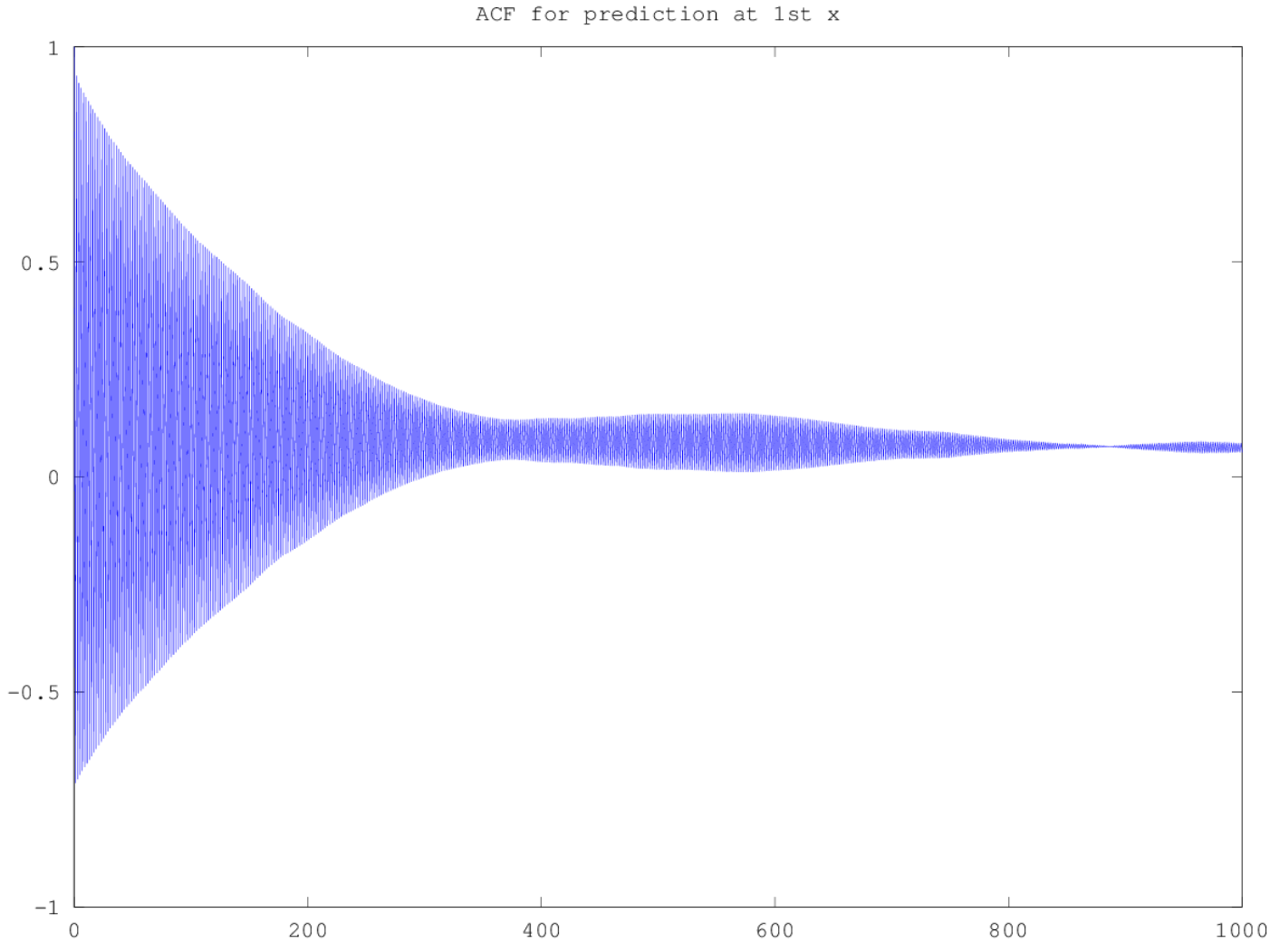


The given data is symmetric about $x = 0$;
lack of symmetry because $N = 1\,000\,000$ only.

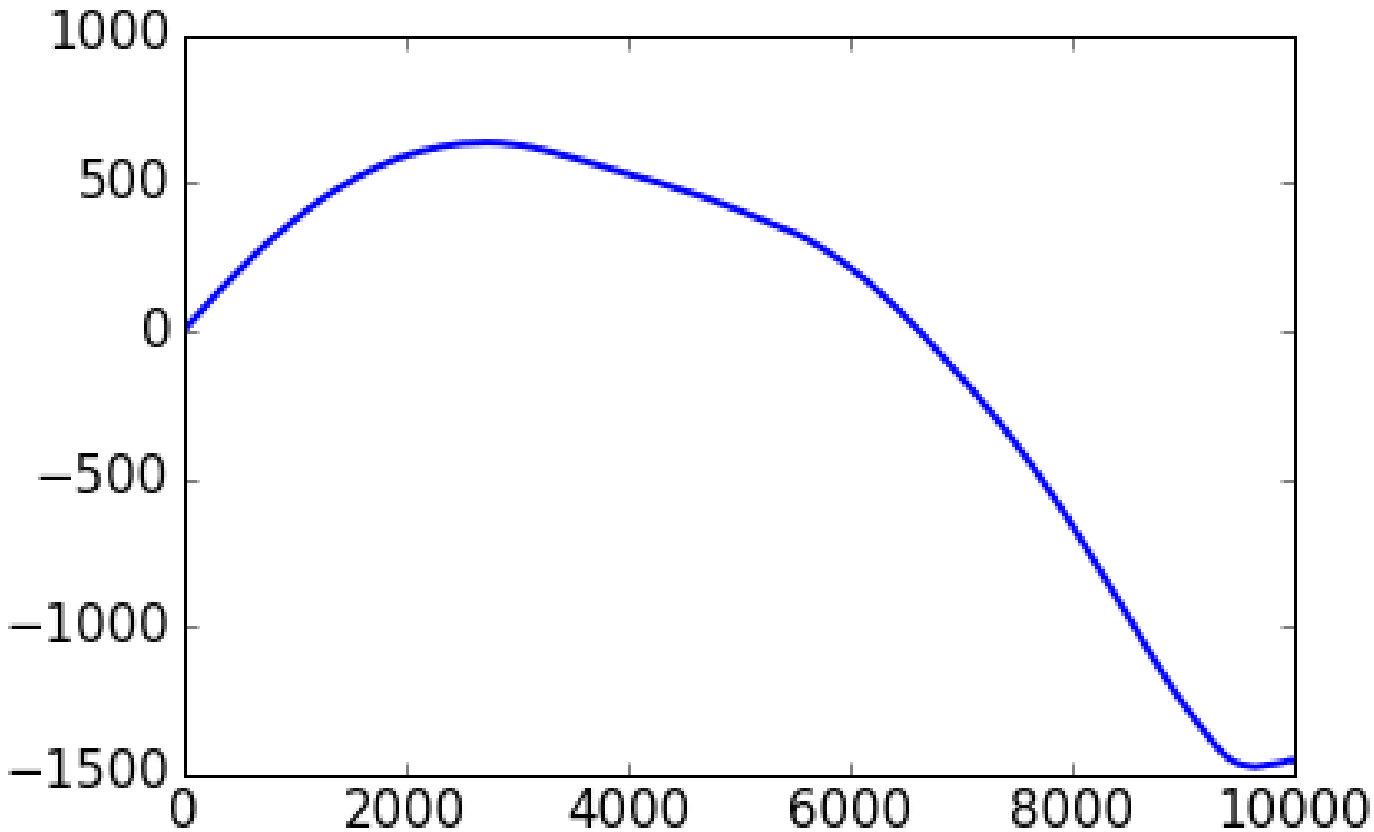
Time series for $u(\mathbf{q}; x_1)$



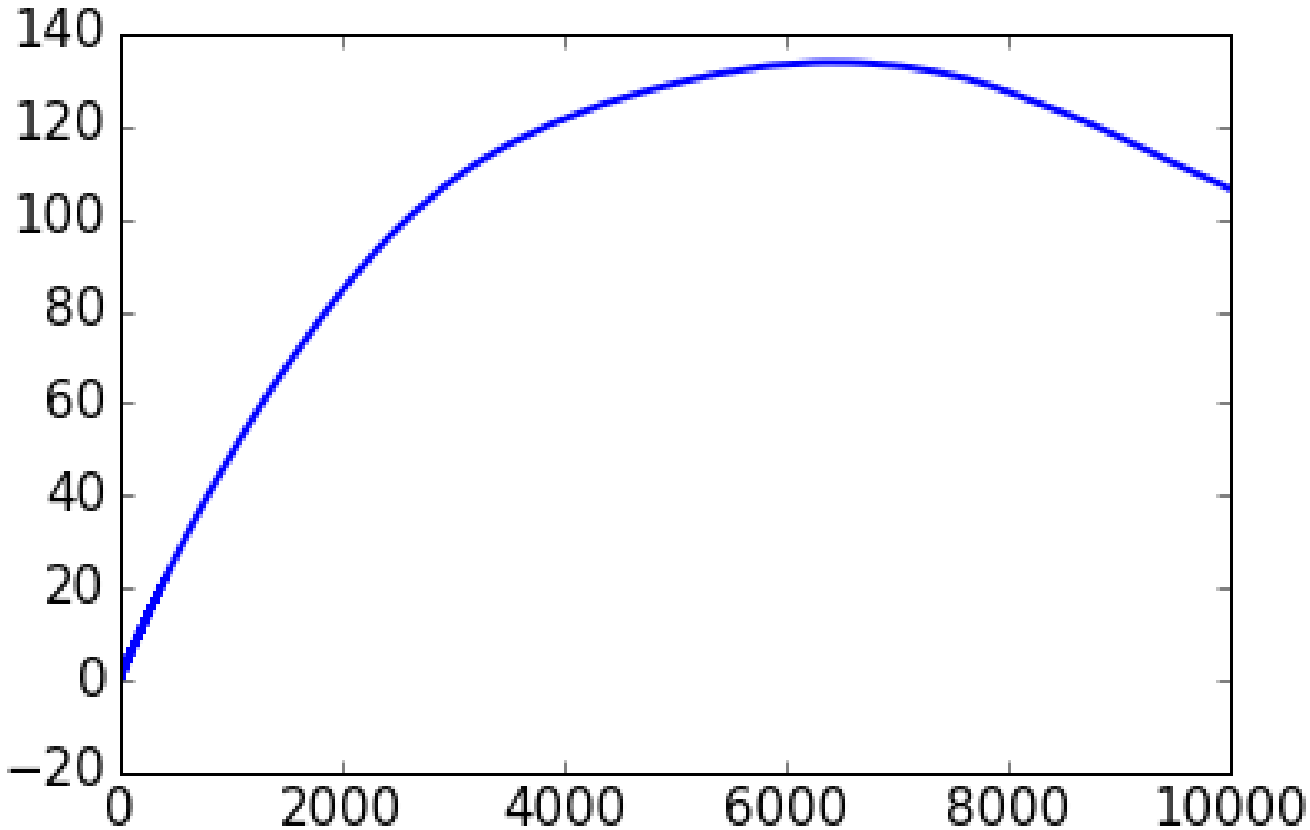
Autcorrelation Function (ACF) for $u(\mathbf{q}; x_1)$



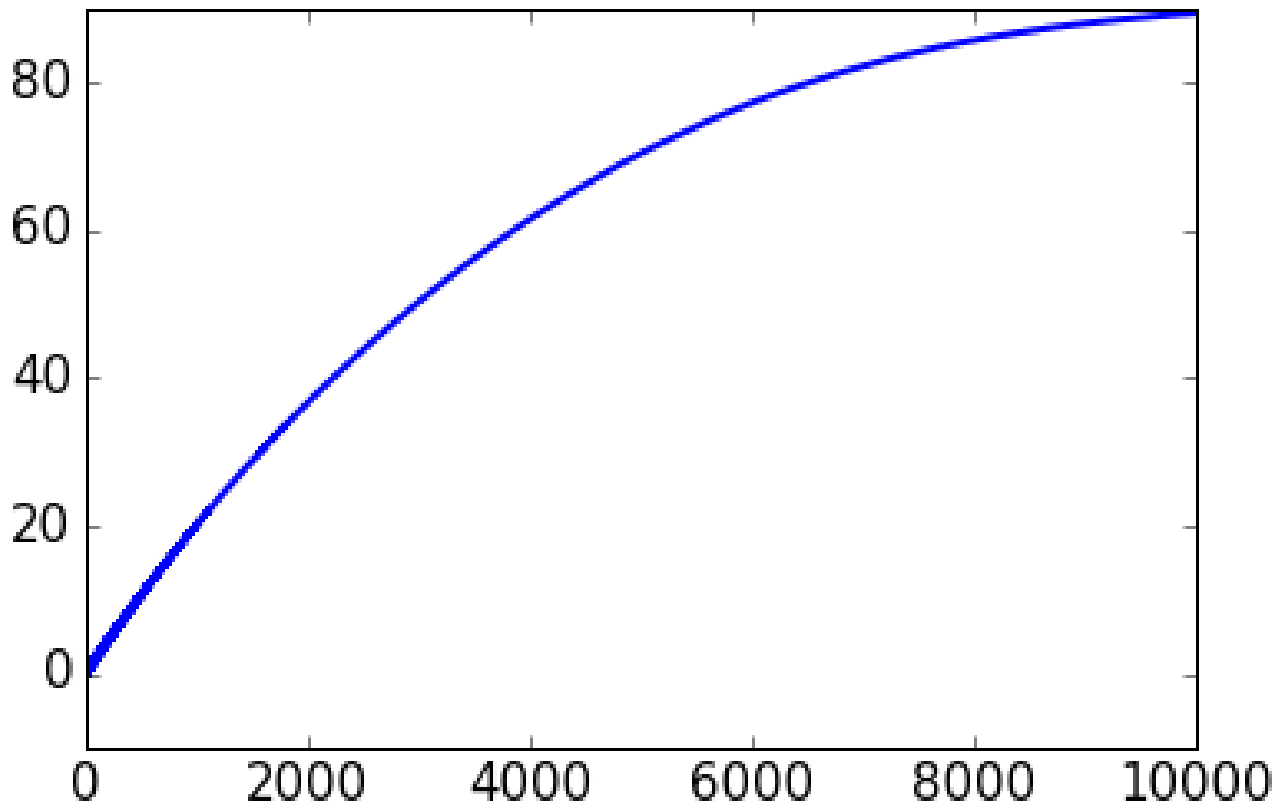
Integrated ACF for $N = 10\,000$



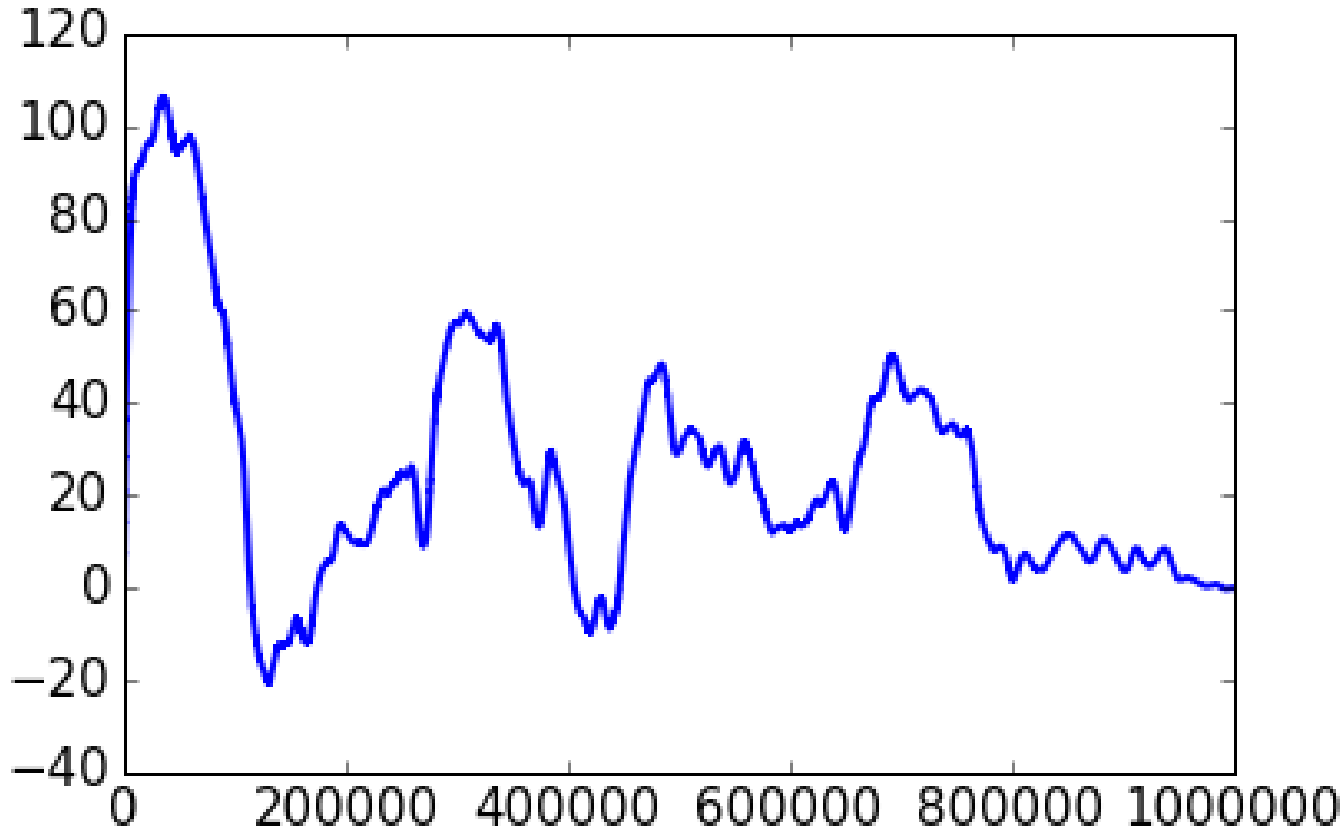
Integrated ACF for $N = 100\,000$



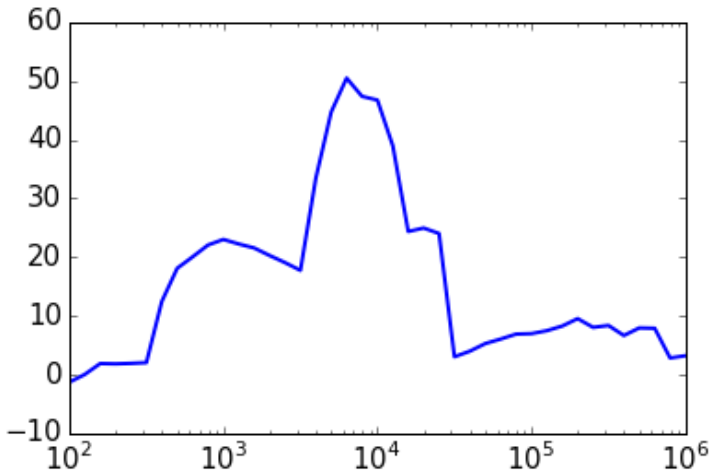
Integrated ACF for $N = 1\,000\,000$



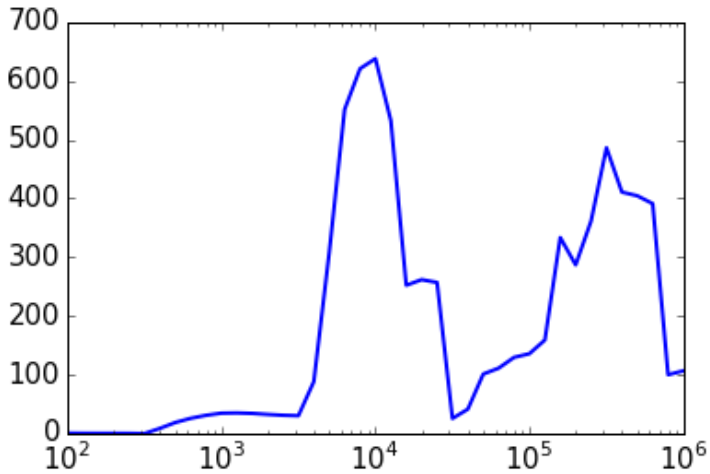
ACF for $N = 1\,000\,000$ integrated 100 times further



Estimated τ as N Increases



`acor`



our version

Outline

Preliminaries

Thorough Sampling

Coping with Statistical Error

Reversible Samplers

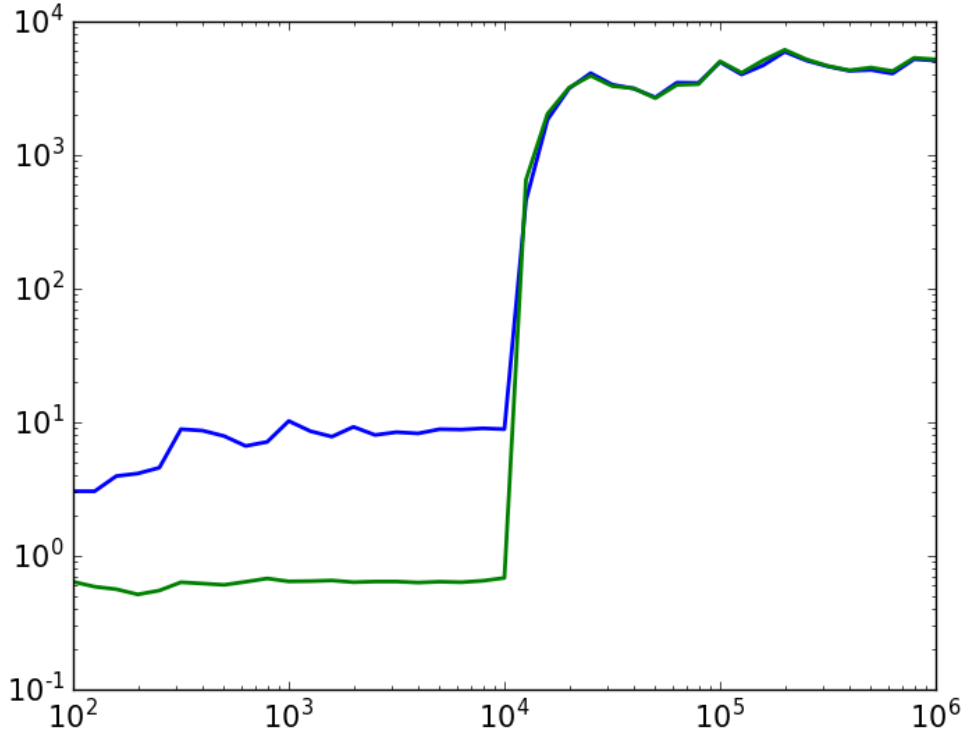
Reversible Samplers

If $\tau_{\max}$ were the maximum over *all* $u(\boldsymbol{z})$, then

$$\tau_{\max} = \frac{1+\lambda_2}{1-\lambda_2}.$$

Mixture of Gaussians, Again

Estimated τ and $\tau_{\max}$ as N increases:



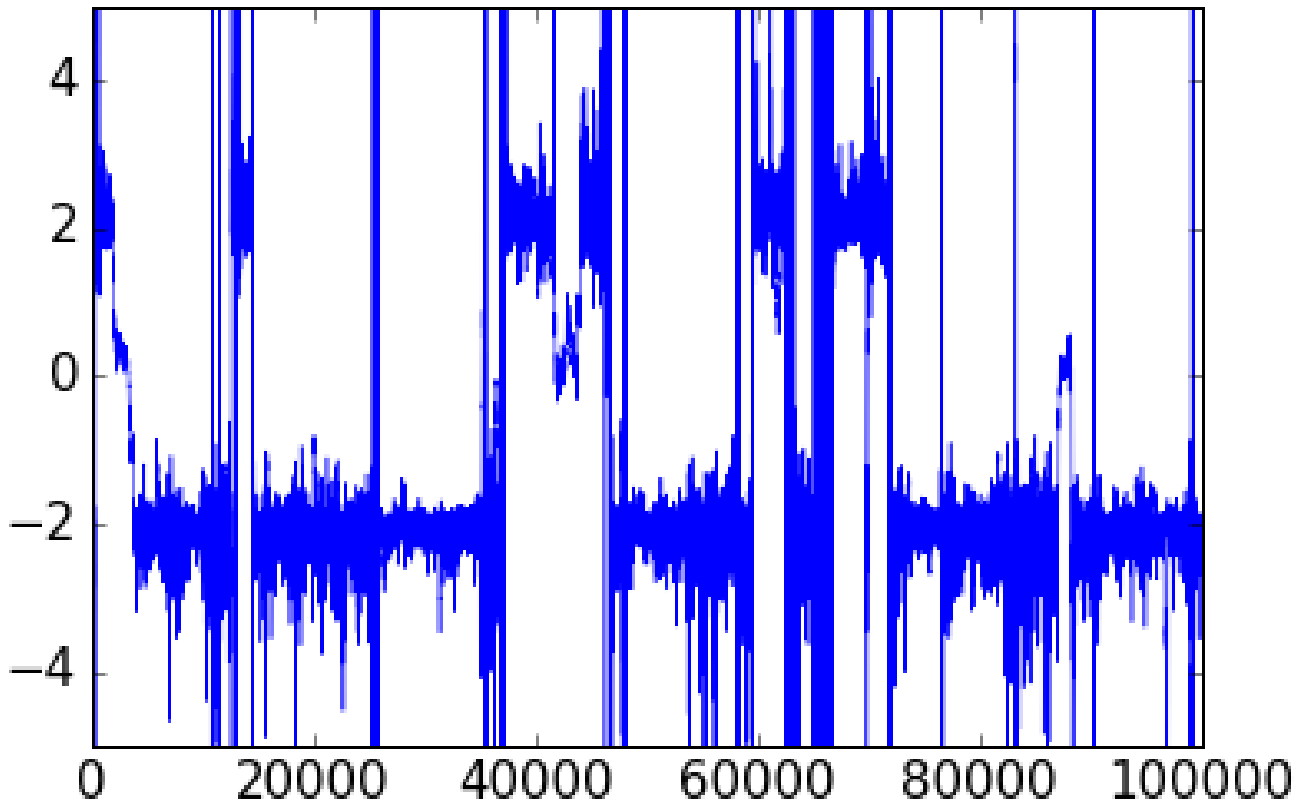
Simplest Neural Network, Again

Fitting model

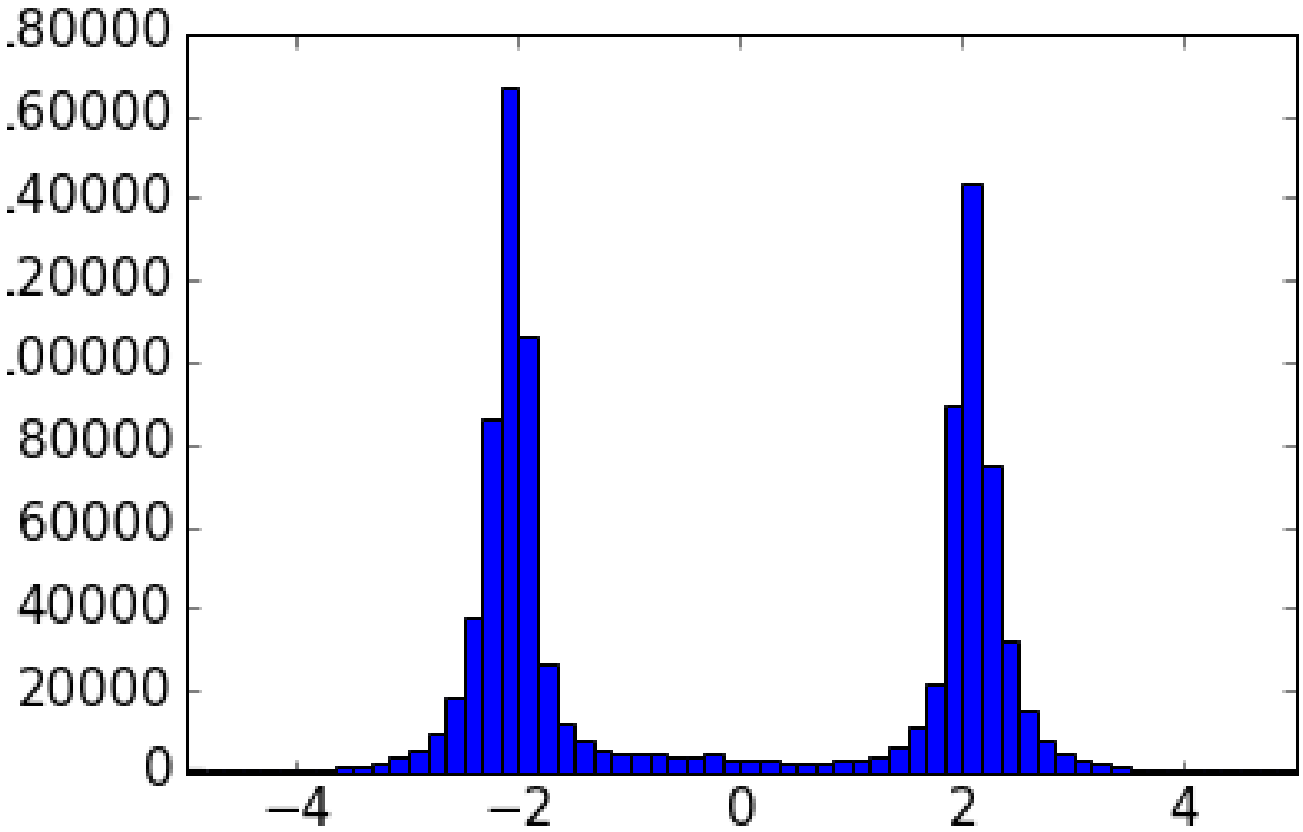
$$u(\mathbf{q}; x) = q_3 \tanh(q_1 x + q_2) + q_4$$

to data $y_i \approx u(\mathbf{q}; x_i)$.

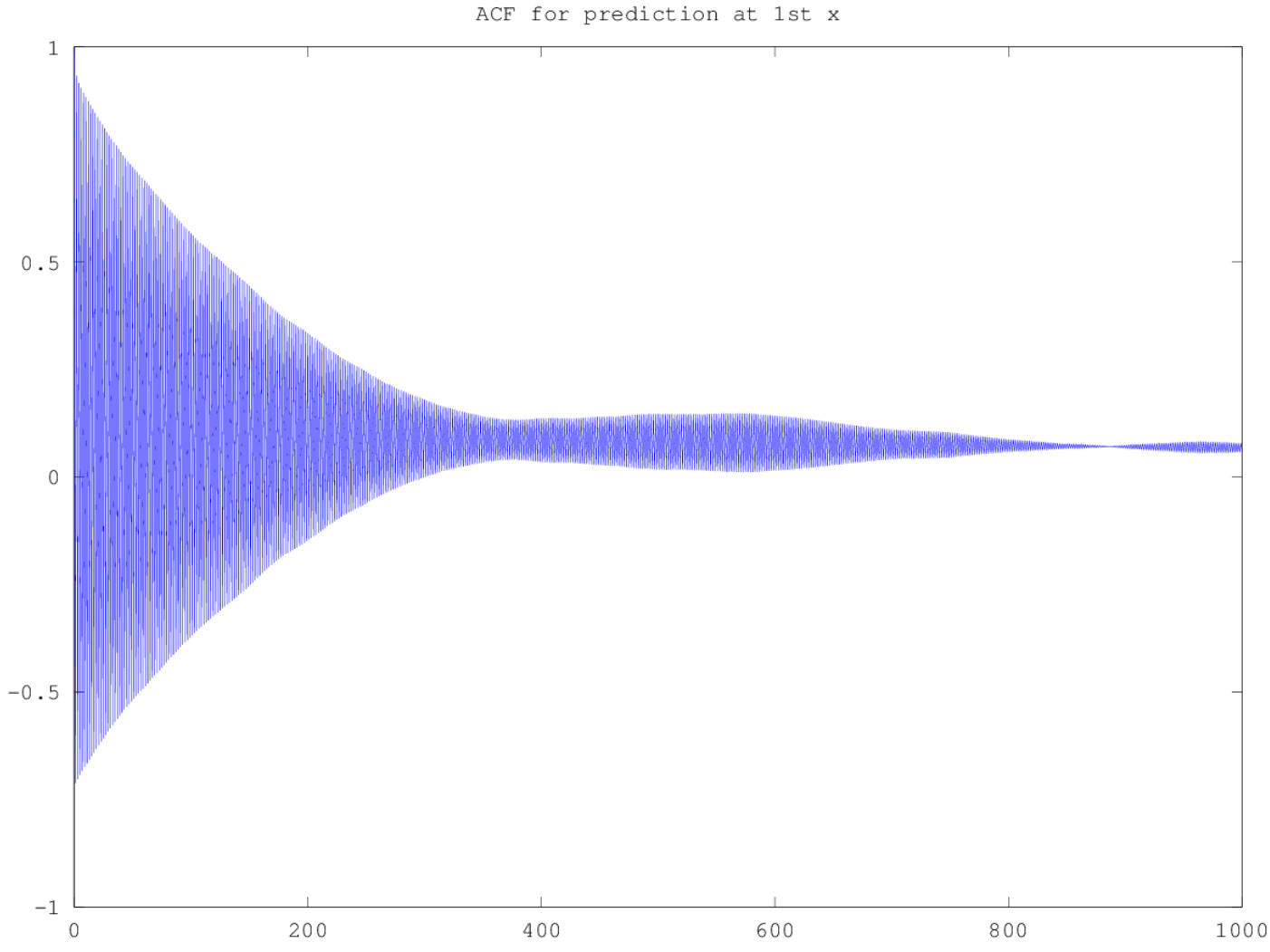
Time Series for Switch Location $-q_2/q_1$



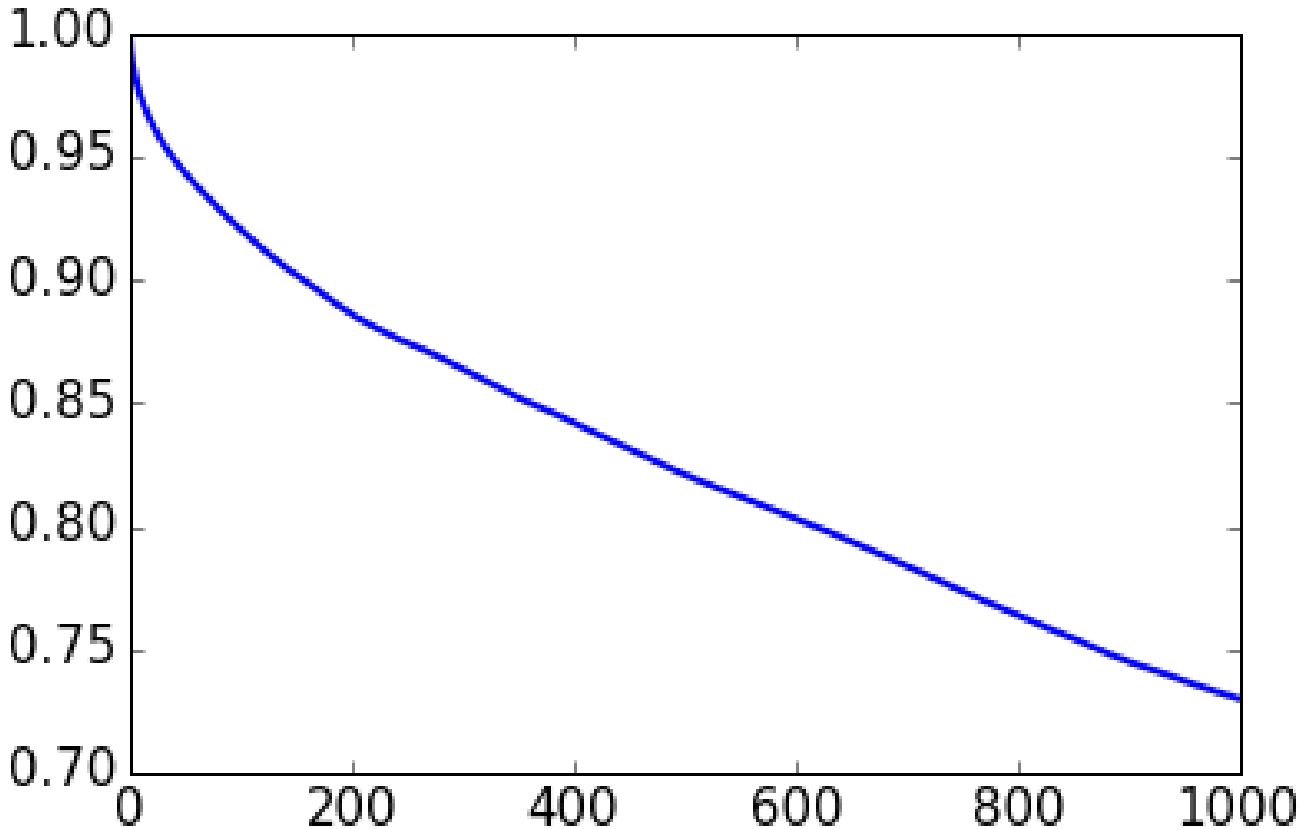
Histogram for Switch Location $-q_2/q_1$



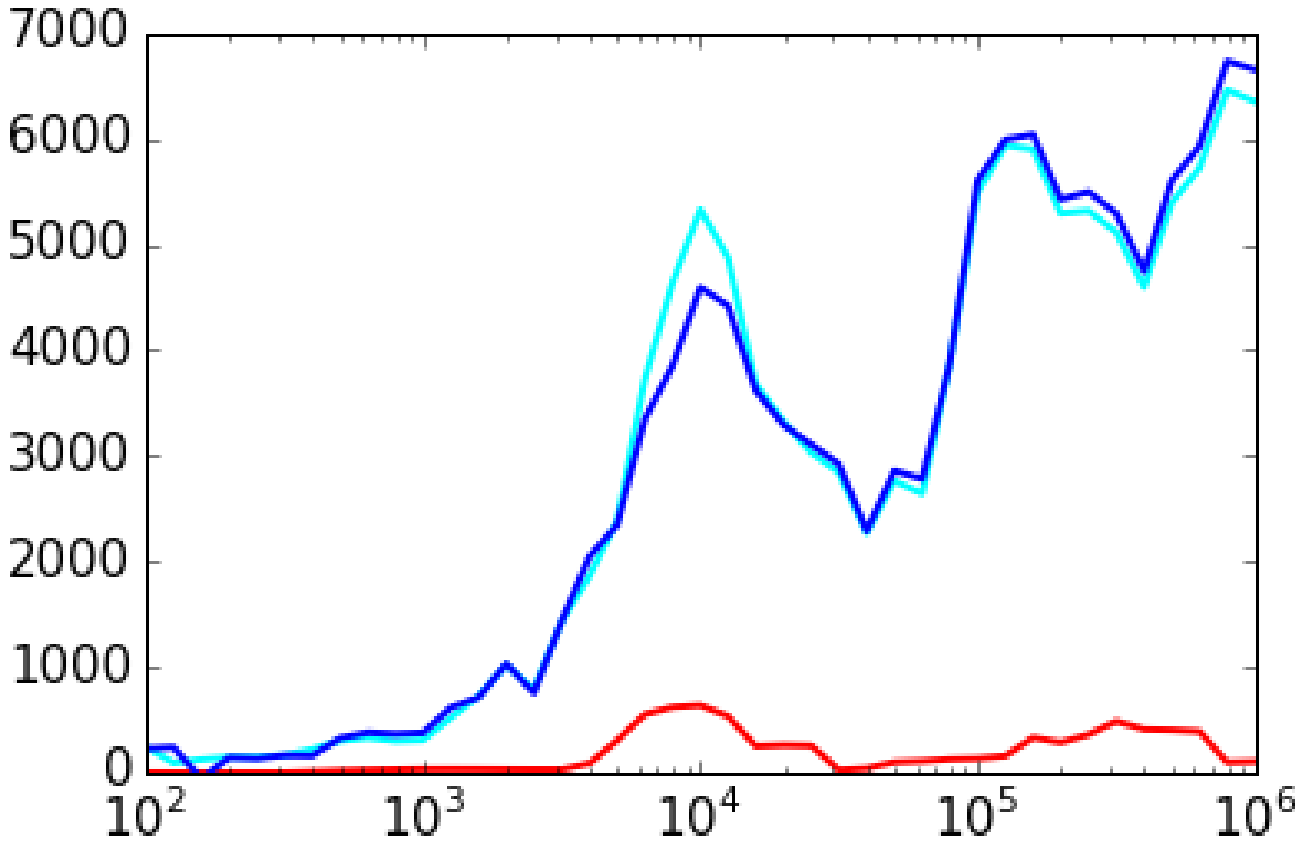
Autcorrelation Function (ACF) for $u(\mathbf{q}; x_1)$



ACF for Switch Location $-q_2/q_1$



Estimated τ and $\tau_{\max}$ as N Increases



Summarizing

Let τ denote the integrated autocorrelation time.

- Estimating τ is considered essential.
- Greater reliability is possible by trying to estimate the worst case τ.
- Better estimates of τ are needed.

Thank you