Gradient IS and Unadjusted Langevin for IS

Ingmar Schuster, Alain Durmus & Christian Robert
Université Paris Dauphine, Telecom ParisTech, Paris Dauphine

April 19, 2016

Section 1

Outline

Section 2

Introduction

Monte Carlo techniques

Monte Carlo techniques

- approximate integral/expected value H of some function h wrt density π
- full uncertainty information about H can be estimated
- straight forward when we can draw samples from π directly
- else
  - Markov Chain Monte Carlo (MCMC): construct Markov Chain to sample approximately from π after chain reaches stationary regime
  - Importance Sampling and Sequential Monte Carlo (IS/SMC): draw from an auxiliary distribution q instead of π and reweight the samples to account for change of distribution

Adaptive MCMC

- adaptive proposals in MH/MCMC
- will be ergodic wrt target π in the general case under some conditions such as (Roberts and Rosenthal, 2007)
  - all Markov kernels used are ergodic wrt π
  - diminishing adaptation, e.g. adaptation probability p_t at iteration t satisfies $p_t \to 0$ as $t \to \infty$
- we still can have $\sum_{t=1}^{N} p_t \to \infty$ as $N \to \infty$, i.e. infinite overall adaptation (cf. Andrieu and Thoms, 2008)
- conditions are sufficient but not necessary (other schemes might still be ergodic)
- simplest case: fit covariance matrix C to previous samples, used in Gaussian proposal for next point (Haario et al., 2001)

Importance Sampling estimators

- Importance Sampling samples from proposal q

$$
\begin{aligned}
H = & \int \pi(x)h(x)\mathrm{d}x = & \int \frac{\pi(x)}{q(x)}h(x)q(x)\mathrm{d}x \\
= & \mathbb{E}_q\left(w(x)h(x)\right) \approx & \frac{1}{w_\Sigma(\mathbf{X})}\sum_{X\in\mathbf{X}} w(X)h(X)
\end{aligned}
$$

where $\mathbf{X}$ iid from q, $w(X) = \pi(X)/q(X)$ is the unnormalized importance weight and $w_\Sigma(\mathbf{X}) = \sum_{X\in\mathbf{X}} w(X)$.

Adaptive Importance Sampling: Population Monte Carlo

$$\begin{aligned}\mathbb{E}\left(w_t(X)h(X)\right) &= \iint \frac{\pi(x)}{q_t(x)} h(x)\ \mathrm{d}q_t(x)\ \mathrm{d}g(q_t) \\ &= \iint \pi(x)h(x)\ \mathrm{d}x\ \mathrm{d}g(q_t) \\ &= \int H\ \mathrm{d}g(q_t) = H\end{aligned}$$

- Population Monte Carlo (PMC; Cappé et al., 2004)
  - improve proposal distributions q_t over iterations indexed by t by adapting to samples from previous iterations
  - special case of SMC which is very simple coming from an MCMC background

Population Monte Carlo Cappé et al. (2004)

Input: initial proposal density q_0, unnormalized density π, population size p, sample size m
Output: lists P, W of m samples and weights
Initialize $P = List()$
Initialize $W = List()$
while $len(P) \leq m$ **do**
  construct proposal distribution q_t
  generate set of p samples $\mathbf{X}_t$ from q_t and append it to P
    for all $X \in \mathbf{X}_t$ append weights $\pi(X)/q_t(X)$ to W
end while

- proposals q_t must not degenerate to distributions with thinner tails than π
- diminishing adaptation not a requirement, unlike in the adaptive MCMC theorems for the general case
- can adapt any which way we like without need to proof ergodicity
- estimate of evidence of our model (marginal likelihood, normalizing constant) at any time

Section 3

Gradient Importance Sampling

- Gradient IS updates a (scaled) covariance fit C_{t+1} with posterior samples from iteration t
- Proposal distribution at iteration $t+1$ is mixture

$$q_t(\cdot) = \sum_{X \in \mathbf{X}'_t} \mathcal{N}\left(\cdot | X + \delta C_{t+1} \nabla \log \pi(X), C_{t+1}\right)$$

  from which we draw $p = |\mathbf{X}'_t| = |\mathbf{X}_t| = \cdots = |\mathbf{X}_1|$ samples
- where $0 \leq \delta \leq 0.5$ is a parameter
- mixture components are closely related to discretized Langevin diffusion
- strong variance reductions through stratification when sampling from q_t

- Covariance C_t updated similar to Haario et al. (2001)
- define initial covariance matrix C_0 and iteration threshold t_0 at which adaptation starts, then

$$C_t = \begin{cases} C_0 & t \leq t_0 \\ s_d(\mathrm{Cov}(X_0, \ldots, X_{t-1}) + \epsilon I) & t > t_0 \end{cases}$$

- and the online fitting is achieved by the recursion

$$C_{t+1} = \frac{t-1}{t} C_t + \frac{s_d}{t} \left(t \bar{X}_{t-1} \bar{X}_{t-1}^T - (t+1) \bar{X}_t \bar{X}_t^T + X_t X_t^T + \epsilon I \right)$$

- where $\bar{X}_{t'}$ is the mean at time t'

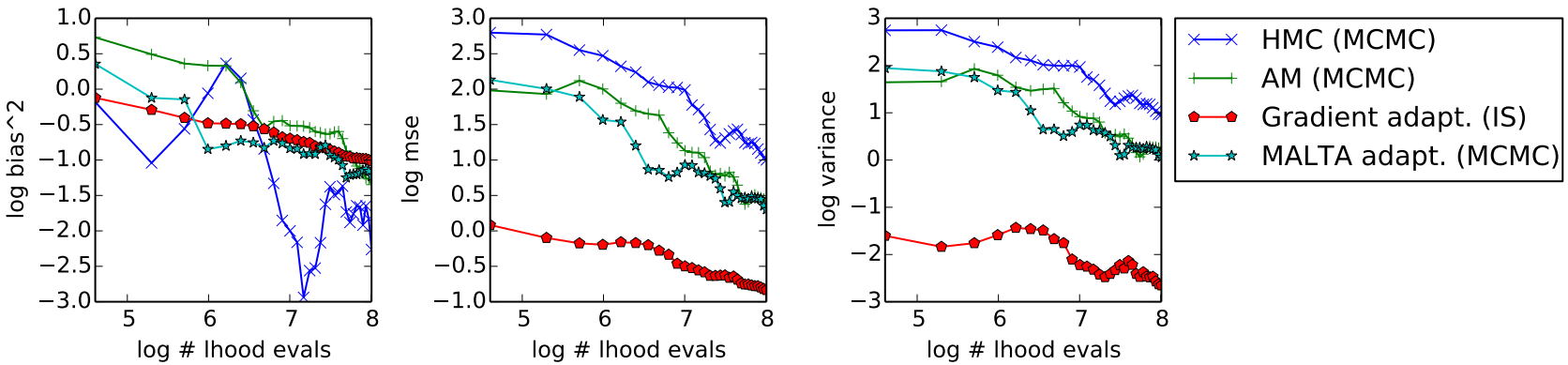
Section 4

Evaluation

Algorithms

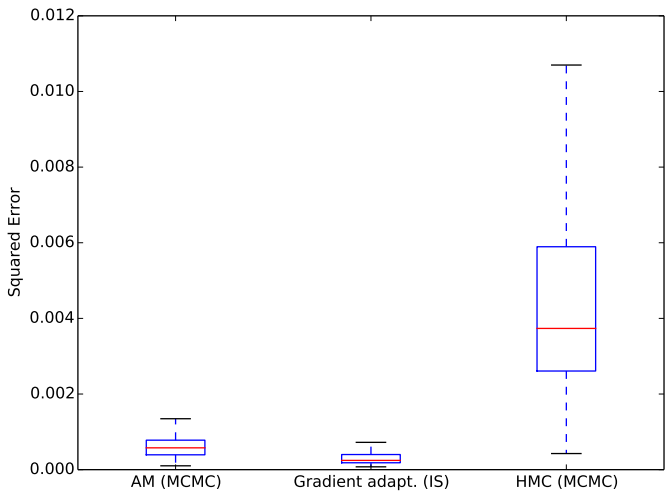
- In Gradient IS experiments, particle system of size $p = 10$ (!) is used
- Compare to
  - HMC (parameters tuned to give optimal acceptance rate of 0.65)
  - adaptive Random Walk (adapt. RW; Haario et al., 2001)
  - adaptive Metropolis adjusted Langevin with truncation (adaptive MALTA; Atchadé, 2006)
- for covariance fit $s_d = 1$, $\epsilon = 2$ and $t_0 = 10$ used for all adaptive algorithms
- Drift factor δ tuned to give lowest variance across simulations for adaptive algorithms

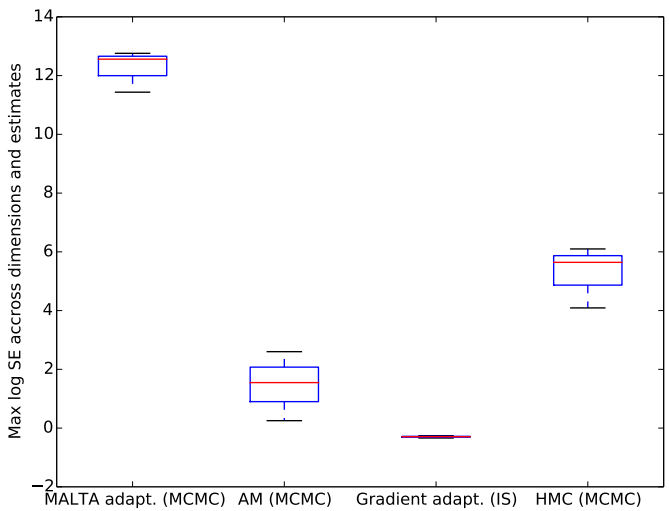
Strongly twisted 2D Banana



- GRIS shows clearest gains here
- trace plot of first coordinate (next slide) suggests that Adaptive Metropolis steps are not accepted very often
- Especially performance of adaptive MALTA is surprising, given that it adapts almost every parameter

Logistic Regression





- Logistic Regression Problem with German Credit dataset (Hoffman and Gelman, 2014)
  - 24 predictor variables, 1 classification outcome (creditworthy or not)
  - 1000 data points
- Especially MALTA performance is surprising, given it adapts almost every parameter

Section 5

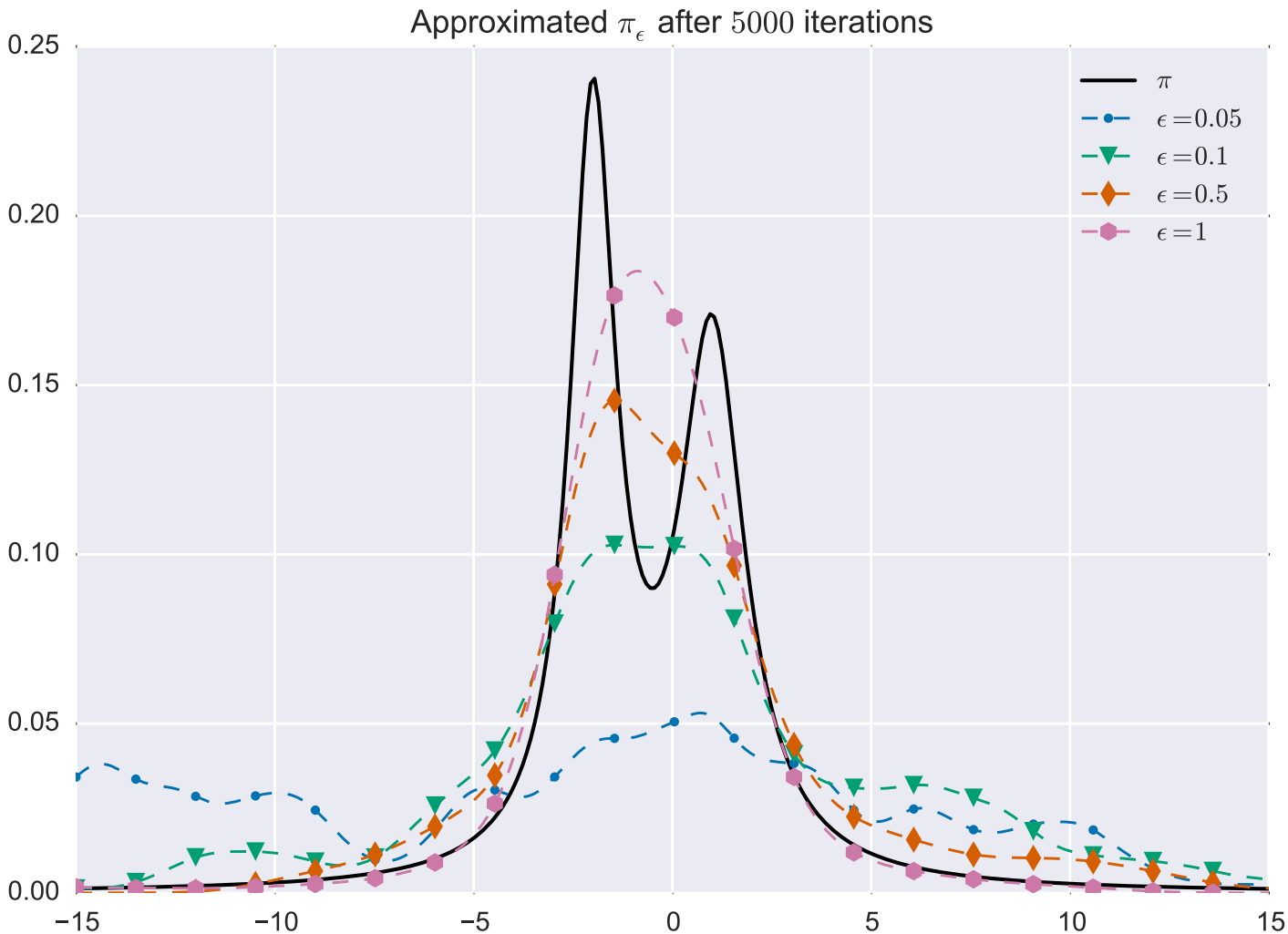
Unadjusted Langevin for IS

Unadjusted Langevin

- Taking a step back
- Unadjusted Langevin Algorithm (ULA) samples from some distribution π_ϵ close to π by using proposals

$$X_{t+1} \sim \mathcal{N}\left(\cdot | X_t + \epsilon \nabla \log \pi(X_t), \sqrt{2\epsilon} I\right) = \kappa_\epsilon(\cdot | X_t)$$

- under some Forster-Lyapunov conditions κ_ϵ is a Markov Kernel with unique invariant distribution π_ϵ (Durmus and Moulines, 2016)
- Thus we can approximate π_ϵ as $\pi_\epsilon(\cdot) = \int \pi_\epsilon(x)\kappa_\epsilon(\cdot|x)\mathrm{d}x \approx \sum_t \kappa_\epsilon(\cdot|X_t) = \hat{\pi}_\epsilon(\cdot)$ where the X_t are samples form the ULA chain


Approximated π_ϵ after 5000 iterations
0.25
0.20
0.15
0.10
0.05
0.00
−15
−10
−5
0
5
10
15
π
$\epsilon = 0.05$
$\epsilon = 0.1$
$\epsilon = 0.5$
$\epsilon = 1$

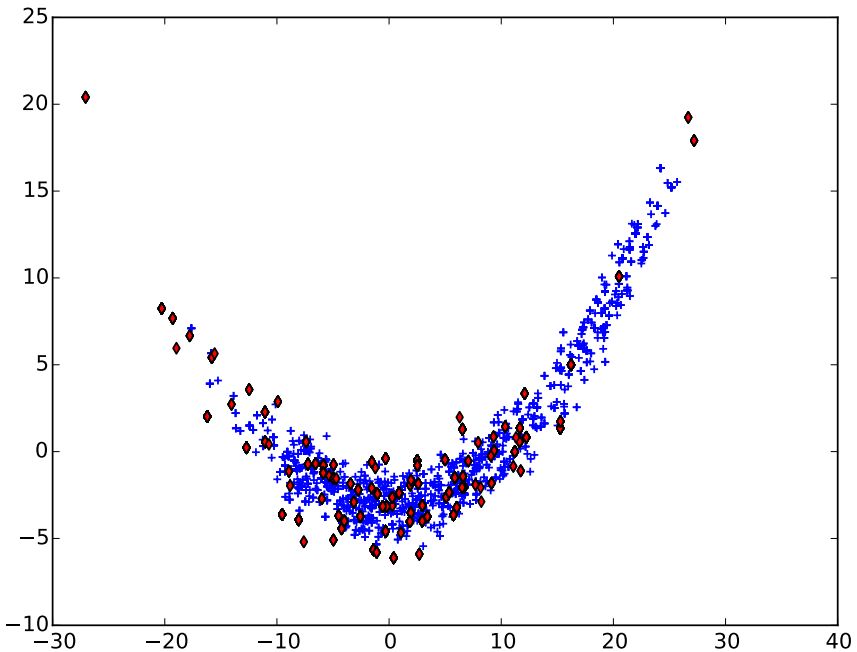
Unadjusted Langevin with RB Importance Weighting

- enables Rao-Blackwellized weights

$$w(X_t) = \frac{\pi(X_t)}{\hat{\pi}_\epsilon(X_t)}$$

- RB introduces asymptotically vanishing bias (weighting X_t with its dependend RVs $\kappa_\epsilon(\cdot|X_t), \kappa_\epsilon(\cdot|X_{t+1}), \dots$)
- experimentally strong gains over MALA
- bias can be avoided with two parallel ULA chains

100D Banana



- ULA-IS (red diamonds) vs. MALA (blue crosses)
- improves RMSE of estimate of first four moments on average by factor of 1.6

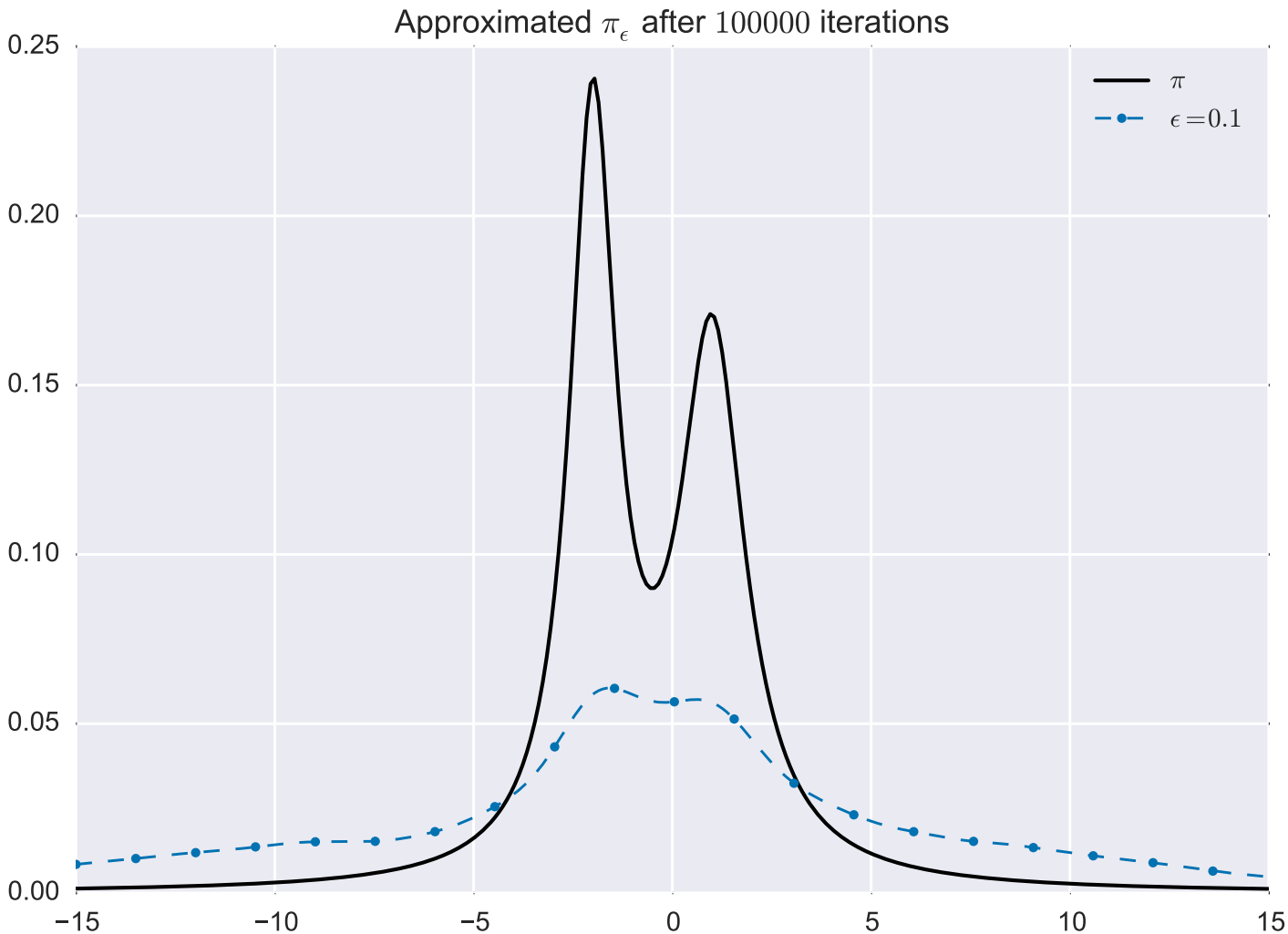
Section 6

Discussion

Contributions

- Gradient IS
  - First gradient informed sampling algorithm based on an iterated importance sampling scheme
  - Simple to implement (as compared to adaptive MALTA and HMC)
  - Few and very easily tuned parameters
- ULA IS
  - derived elementary approximation of π_ϵ
  - enables Rao-Blackwellized weights to reduce variance
  - might improve over MALA (doesn't get stuck)
  - might be combined with Romberg scheme to reduce discretization bias

Thanks!


Approximated π_ϵ after 100000 iterations
π
$\epsilon = 0.1$
0.25
0.20
0.15
0.10
0.05
0.00
−15
−10
−5
0
5
10
15

Literature I

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, 18(November):343–373.

Atchadé, Y. F. (2006). An adaptive version for the metropolis adjusted Langevin algorithm with a truncated drift. *Methodology and Computing in Applied Probability*, 8(March):235–254.

Cappé, O., Guillin, a., Marin, J. M., and Robert, C. P. (2004). Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929.

Durmus, A. and Moulines, E. (2016). Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm.

Haario, H., Saksman, E., and Tamminen, J. (2001). An Adaptive Metropolis Algorithm. *Bernoulli*, 7(2):223–242.

Literature II

Hoffman, M. and Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1351–1381.

Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, 44(2):458–475.