Sample Adaptive MCMC

Michael H. Zhu (*Stanford University*)

Abstract

- Sample Adaptive MCMC (SA-MCMC) is a MCMC method based on a reversible Markov chain for $\pi^{\otimes N}$ which
  - uses an adaptive proposal distribution based on the current state of N points
  - uses a sequential substitution procedure with one new likelihood evaluation per iteration and at most one updated point each iteration
- The SA-MCMC proposal distribution automatically adapts within its parametric family to best approximate the target distribution.
  - In contrast to many existing MCMC methods, SA-MCMC does not require any tuning of the proposal distribution.
  - SA-MCMC only requires specifying the initial state of N points, which can often be chosen *a priori*, thereby automating the entire sampling procedure with no tuning required.
- Experimental results demonstrate the fast adaptation and effective sampling of SA-MCMC.

Background

- For MCMC methods like Metropolis-Hastings, the choice of the proposal distribution $q(\cdot|\theta^{(k)})$ is important in practice for effective sampling from the target distribution.
- A suboptimal choice for the scale or shape of the proposal can lead to inefficient sampling, yet the design of an optimal proposal distribution is challenging when the properties of the target distribution are unknown, especially in high-dimensional spaces.
- Adaptive MCMC methods such as Adaptive Metropolis continually adapt the proposal distribution based on the entire history of past states.
  - However, the method is no longer based on a valid Markov chain, so the usual MCMC convergence theorems do not apply and the validity of the sampler and an ergodic theorem must be proved for each specific algorithm under certain technical assumptions.
- We use the following notation for Sample Adaptive MCMC:
  - Target distribution $p(\theta)$
  - Initialization distribution $q_0(\cdot)$
  - Proposal distribution $q(\cdot|\mu(S),\Sigma(S))$
  - N points in the state
  - κ burn-in iterations, K estimation iterations

Simulation examples

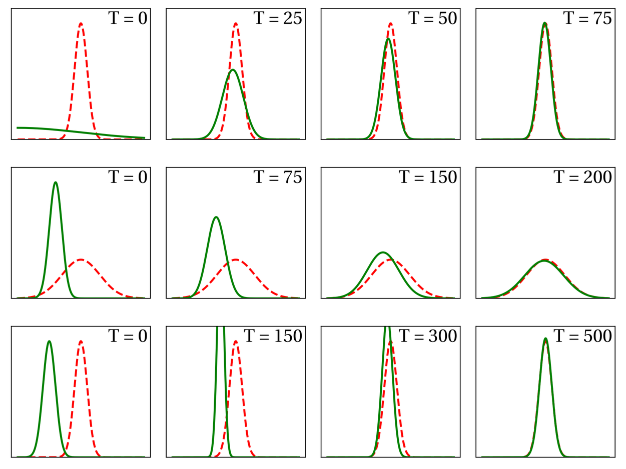

Figure: Adaptation of the SA-MCMC proposal distribution (green) to three target distributions (red).

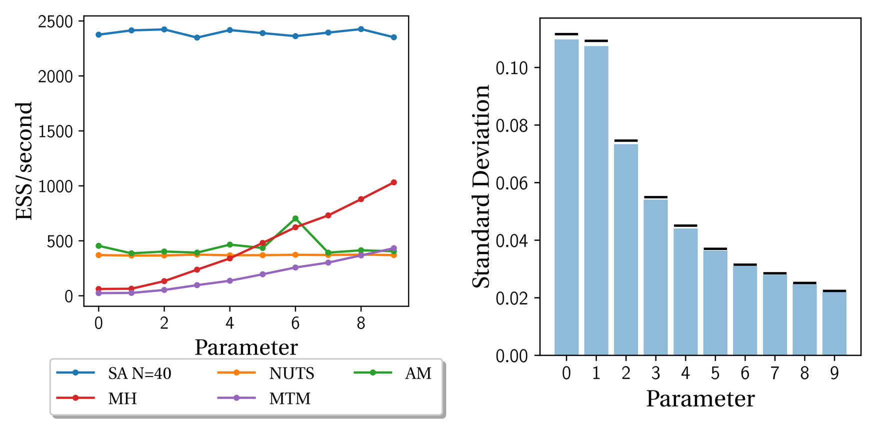

Figure: Bayesian linear regression. (***left***) Comparison of ESS/second for each parameter. (***right***) Standard deviation of the SA proposal distribution (blue bar), averaged over iterations, for each parameter compared with the ground truth posterior standard deviation (black line).

Algorithm

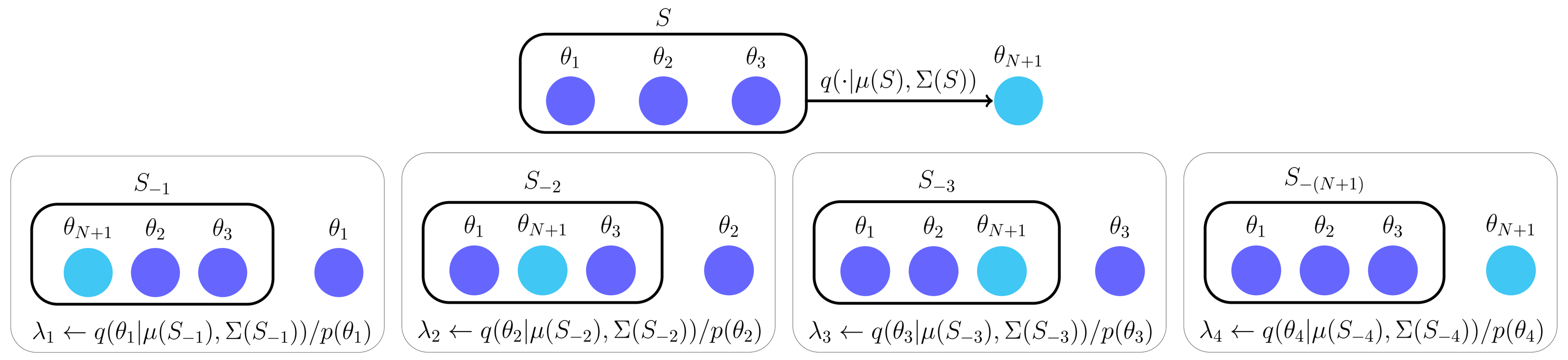

Figure: Illustration of one iteration of SA-MCMC for $N=3$. After the proposed point $\theta_{N+1} \sim q(\cdot|\mu(S),\Sigma(S))$ is sampled, the sets $S_{-1},\dots,S_{-(N+1)}$ are used to calculate the substitution probabilities $\lambda_1,\dots,\lambda_{(N+1)}$. One of the sets $S_{-1},\dots,S_{-(N+1)}$ is chosen to be the next state with probability proportional to λ_n.

Algorithm 1 Sample Adaptive MCMC

Require: $p(\theta)$, $q_0(\cdot)$, $q(\cdot|\mu(S),\Sigma(S))$, N, κ, K

```
Initialize S^(0) ← (θ_1,...,θ_N) where θ_n ~ q_0(·) for n = 1,...,N
for k = 1 to κ + K do
  Let S = (θ_1,...,θ_N) ← S^(k−1)
  Sample θ_{N+1} ~ q(·|μ(S),Σ(S))
  Let S_{−n} ← (S with θ_n replaced by θ_{N+1}) for n = 1,...,N. Let S_{−(N+1)} ← S.
  Let λ_n ← q(θ_n|μ(S_{−n}),Σ(S_{−n}))/p(θ_n) for n = 1,...,N+1
  Sample j ~ J with P[J = n] = λ_n / Σ_{i=1}^{N+1} λ_i,   1 ≤ n ≤ N+1
  Let S^(k) ← S_{−j}
end for
Return ∪_{k=κ+1,...,κ+K} S^(k)
```

Theory

- We prove the ergodicity of SA-MCMC under general conditions on the target distribution (the same assumptions as for Metropolis-Hastings) and a family of proposal distributions with diagonal covariance matrices.
- We prove the uniform ergodicity of SA-MCMC under the assumption that $q(\theta|\gamma)/\pi(\theta)$ is bounded above and below.

Experimental setup

- We use the following experimental setup:
  - For Metropolis-Hastings (MH), we use an isotropic normal distribution as the proposal distribution, $q_{\mathrm{MH}}(\cdot|\theta) = \mathcal{N}\left(\theta, \sigma_{q,\mathrm{MH}}^2 \mathbb{I}\right)$, with scale parameter $\sigma_{q,\mathrm{MH}}$, and initialize $\theta^{(0)} \sim q_{0,\mathrm{MH}}(\cdot) = \mathcal{N}\left(0, \sigma_{q,\mathrm{MH}}^2 \mathbb{I}\right)$.
  - For Adaptive Metropolis (AM), we use the optimal MH proposal distribution during the burn-in (non-adaptive) phase and then use the proposal distribution $q_{\mathrm{AM}}(\cdot|\theta^{(1)},\dots,\theta^{(k-1)}) = \mathcal{N}\left(\theta^{(k-1)}, s_{\mathrm{AM}}^2 \Sigma^{(k-1)}\right)$ at iteration k with scale parameter s_{AM} and sample covariance matrix $\Sigma^{(k-1)}$ of the past samples $(\theta^{(1)},\dots,\theta^{(k-1)})$.
  - For Multiple-Try Metropolis (MTM), we use the optimal MH proposal distribution with 3 tries.
  - For SA-MCMC (SA), we use $q_{0,\mathrm{SA}}(\cdot) = \mathcal{N}\left(0, \sigma_{q_0,\mathrm{SA}}^2 \mathbb{I}\right)$ with scale parameter $\sigma_{q_0,\mathrm{SA}}$ as the distribution for initializing the N starting points. For the proposal distribution, when using the full covariance matrix, we use the Gaussian family $q(\cdot|\mu(S),\Sigma(S)) = \mathcal{N}(\cdot|\mu(S),\Sigma(S))$.

Experimental results for Bayesian logistic regression

Table: Comparison of ESS/second for Bayesian logistic regression on (***top***) 11-dim MNIST 7s vs 9s using 10 features computed with PCA (***bottom***) 7-dim adult census income

	MH	MTM	AM (diag)	AM (full)	SA (diag)	SA (full)	NUTS
min(ESS)/s	13	5	17	37	23	**278**	54
median(ESS)/s	21	9	23	38	52	**290**	105
s/chain	733	3651	734	742	782	1112	1160
Hyperparameters	q=.02	q=.02 M=3	q=.02 s=.6	q=.02 s=.7	q_0=1 N=40	q_0=1 N=150	Stan
Acceptance rate	23%	48%	24%	26%	75%	98.9%	—
min(ESS)/s	1.4	0.6	13	16	67	**151**	40
median(ESS)/s	17	7	15	17	89	**158**	49
s/chain	2198	10951	2205	2217	2283	2509	2989
Hyperparameters	q=.016	q=.016 M=3	q=.016 s=.8	q=.016 s=.85	q_0=1 N=40	q_0=1 N=150	Stan
Acceptance rate	26%	52%	21%	24%	89%	99.2%	—

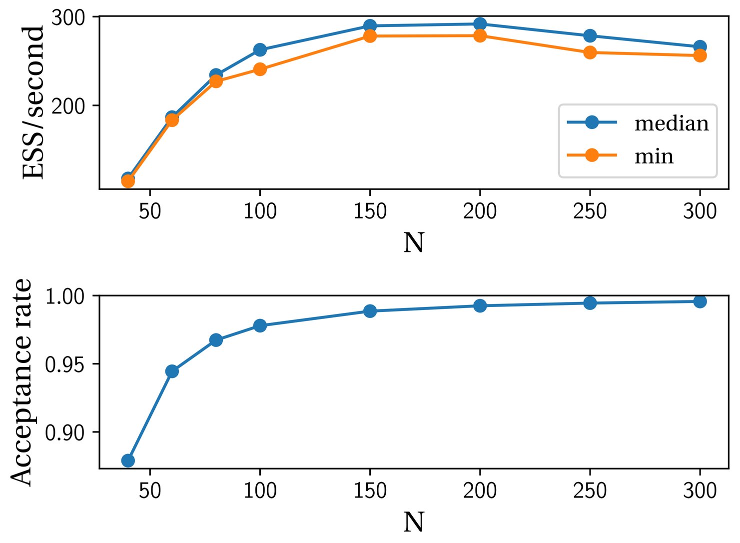

Figure: Plot of ESS/s and acceptance rate for SA-MCMC (full) versus N on MNIST.

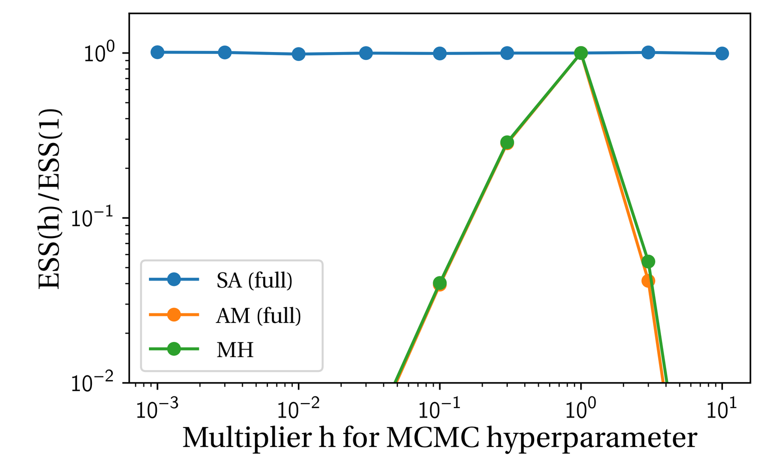

Figure: Impact of MCMC hyperparameter on ESS for MNIST. The ratio ESS(h)/ESS(1) measures the drop in ESS using $0.02h$ for q in MH, $0.7h$ for s in AM, and $1h$ for q_0 in SA.

Table: Comparison of ESS/second for Bayesian logistic regression on (***left***) 55-dim cover type (***right***) 51-dim MiniBooNE between AM (full), SA (full), and NUTS with a dense mass matrix.

	Cover type			MiniBooNE		
	AM	SA	NUTS	AM	SA	NUTS
min(ESS)/s	0.075	**2.34**	0.099	0.31	**3.35**	0.023
median(ESS)/s	0.078	**2.81**	0.114	0.38	**6.59**	0.039
s/chain	52,469	65,537	25,143	28,178	26,627	33,584
s/chain (burn-in)	4,770	5,958	16,980	1,342	2,421	19,051
s/chain (estimation)	47,699	59,579	8,163	26,836	24,206	14,533
# iter. (burn-in)	100,000	100,000	500	100,000	100,000	500
# iter. (estimation)	1,000,000	1,000,000	2,000	2,000,000	1,000,000	2,000
Hyperparameters	q=.004 s=.32	q_0=1 N=1,000	Stan (dense)	q=.007 s=.33	q_0=1 N=1000	Stan (dense)
Acceptance rate	25.1%	99.3%	—	25.7%	90.5%	—