SMOOTHED LANGEVIN PROPOSALS IN METROPOLIS-HASTINGS ALGORITHMS

ØIVIND SKARE[1], FRED ESPEN BENTH[1,2] AND ARNOLDO FRIGESSI[1]

Abstract. The Metropolis Adjusted Langevin Algorithm (MALA) samples from complex multivariate densities π. The proposal density is based on a discretized version of a Langevin diffusion, and is well defined only for continuously differentiable densities π. We propose a modified MALA algorithm when this condition is not fulfilled or when π has very rapid variations. The algorithm is illustrated on the Strauss model, for which two different classes of smoothing are proposed. In these examples smoothing gives advantages in terms of reduced asymptotic variance.

1. Introduction

Langevin diffusions are stochastic differential equations of the form $d\boldsymbol{L}_t = \frac{1}{2}\nabla \log \pi(\boldsymbol{L}_t)dt + d\boldsymbol{B}_t$, where $\boldsymbol{B}_t$ is Brownian motion on $\mathbb{R}^n$, $\pi(\boldsymbol{x})$ is a density (with respect to Lebesgue) and $\boldsymbol{x} \in \mathbb{R}^n$. If $\nabla \log \pi(\boldsymbol{x})$ is continuously differentiable and for some real values N, a and $b < \infty$ it holds that

$$(\nabla \log \pi(\boldsymbol{x}))^T \boldsymbol{x} \leq a|\boldsymbol{x}| + b, \quad |\boldsymbol{x}| > N$$

then the diffusion will have π as a stationary distribution. Under suitable assumptions $\boldsymbol{L}_t$ converges geometrically fast. For example, Roberts & Tweedie (1996) show that for the exponential class $\pi(x) \propto \exp(-\gamma|x|^\beta)$, this happens if $\beta \geq 1$.

Discretizations of Langevin diffusion have been used in Metropolis-Hastings (MH) algorithms as proposals in order to increase the convergence speed, as in many cases (although not always) the discretized diffusion process is approximately stationary. The Metropolis Adjusted Langevin Algorithm (MALA) is obtained by using the Euler discretization of the Langevin diffusion as the proposal in the MH algorithm. Roberts & Rosenthal (1995) have compared the MALA algorithm with the corresponding random walk MH algorithm when π is the multivariate normal distribution: MALA performs longer steps, and has a higher acceptance rate than the random walk MH algorithm. An adjustment of MALA is MALTA (Metropolis Adjusted Langevin Truncated Algorithm) with truncated drift term, which has more robust geometric ergodicity properties.

Key words and phrases. Langevin diffusions, Markov chain Monte Carlo, Metropolis-Hastings algorithm, Strauss model.

[1] Norwegian Computing Center, P.O. Box 114 Blindern, N-0314 Oslo, Norway. The first author has been funded by NORFA and the Research Council of Norway. The research is supported by the EU-TMR project on Spatial and Computational Statistics (ERB-FMRX-CT960095), the ESF programme on Highly Structured Stochastic Systems and the Strategic Institute Programme of the Norwegian Research Council n. 121144/420.

[2] MaPhySto - Centre for Mathematical Physics and Stochastics, University of Aarhus, Ny Munkegade, DK-8000 Aarhus C, Denmark. MaPhySto is funded by a grant from the Danish National Research Foundation.

What could be done if $\pi(\boldsymbol{x})$ is discontinuous or not differentiable at some points? The Langevin diffusion is then not defined at these points, and the conditions for $\boldsymbol{L}_t$ to have stationary distribution π are not fulfilled. The main idea of this paper is then to use the Langevin proposal, $d\boldsymbol{L}_t^\alpha = \frac{1}{2}\nabla \log \pi^\alpha(\boldsymbol{L}_t^\alpha)dt + d\boldsymbol{B}_t$, where π^α is a smoothed approximation of the target distribution π, and then accept with respect to the original target distribution π in the MH algorithm. The smoothing of π makes it possible to use gradient information of π in the proposals to better guide the state vector towards the modes of π. This may not be possible in the case of discontinuous π: for example, if the target distribution has a step discontinuity and is otherwise flat, the Langevin diffusion is zero a.e. One might expect that continuous but steep target distributions could give rise to slow Langevin derived MH algorithms. Therefore, we also investigate the possible effect of using over-smoothed targets in the Langevin proposal.

The approach of smoothing discontinuous $\pi(\boldsymbol{x})$ is illustrated on the classical model of Strauss (Strauss 1975), for which two different ways of smoothing are proposed. Let $\boldsymbol{x} = (\boldsymbol{x}_1, \dots \boldsymbol{x}_n)$ and $\boldsymbol{x}_i \in [0,1]^s$ and let π be the Strauss model with a fixed number of points, n, which has density

$$\pi(\boldsymbol{x}) = \Pi_{i=1}^n \Pi_{j=i+1}^n h_\gamma(d_{ij}) \ , \ \ d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|,$$

where the repulsion function $h_\gamma(d_{ij})$ is defined by

$$\begin{aligned} h_\gamma(d_{ij}) &= \gamma + (1-\gamma)h(d_{ij}) \\ h(d_{ij}) &= I_{[r,R]}(d_{ij}) \end{aligned}$$

for $\gamma \in [0,1]$ and $I_{[a,b]}(d)$ is equal to 1 if $a \leq d \leq b$, zero otherwise. We will use either the Euclidean distance on $\mathbb{R}^s$ or the distance defined on the torus $[0,1]^s$ given by

$$d_{ij} = \sqrt{\sum_{l=1}^{s} (\min(1 - |x_{i,l} - x_{j,l}|, |x_{i,l} - x_{j,l}|))^2}.$$

We take R to be the maximal value of d_{ij}, i.e. $\sqrt{s}$ for the non-torus and $\frac{1}{2}\sqrt{s}$ for the torus geometry.

2. Smoothed MALTA Algorithm for the Strauss model

The proposal step in the MH algorithm will be generated by a discretized version of the Langevin diffusion

$$d\boldsymbol{L}_t = \frac{1}{2}\nabla \log \pi(\boldsymbol{L}_t)dt + d\boldsymbol{B}_t. \tag{1}$$

where π_L is usually chosen to be equal to $\pi(\boldsymbol{x})$. However this choice of π_L is not always useful. In our example, we will obtain an almost everywhere zero drift, since $\nabla \log(\pi(\boldsymbol{x})) = 0 \quad$ a.e. The MALA algorithm will thus coincide with the random walk MH. Instead of choosing the target distribution π in (1), we will use a smooth approximation π^α of π. We consider the parametrised family of densities

$$\pi^\alpha(\boldsymbol{x}) = \Pi_{i=1}^n \Pi_{j=i+1}^n h_{\alpha,\gamma}(d_{ij})$$

with

$$h_{\alpha,\gamma}(d_{ij}) = \gamma + (1-\gamma)h_\alpha(d_{ij}),$$

where h_α is a smoothed version of h, depending on a smoothing parameter α. We choose α to be the angle of the tangent of h_α in r. Different choices of h_α could be conceived. We propose here two different classes. First define h_α as an exponential S-shaped approximation of h

$$h_\alpha(d) = \frac{1}{1+e^{-k(\alpha)f(d)}},\ d \in [0, R]$$

where the function $f(d)$ is given by

$$f(d) = \frac{R-r}{R-d} - \frac{r}{d},$$

and the function $k(\cdot)$ is chosen as follows: The slope in r of $h_\alpha(d)$ is $h'_\alpha(r) = \tan(\alpha)$. An easy calculation shows that

$$h'_\alpha(d) = k(\alpha)f'(d)h^2(d)e^{-k(\alpha)f(d)}.$$

Since $f(r) = 0$ and $h_\alpha(r) = \frac{1}{2}$ we get

$$h'_\alpha(r) = \frac{1}{4}k(\alpha)f'(r).$$

But

$$f'(r) = \frac{1}{r} + \frac{1}{R-r} = \frac{R}{r(R-r)},$$

and thus

$$\frac{1}{4}k(\alpha)\cdot\frac{R}{r(R-r)} = \tan(\alpha) \Longrightarrow k(\alpha) = \frac{4}{R}\tan(\alpha)r(R-r).$$

An alternative choice for h_α is the arctangent smoother

$$h_\alpha(d) = \frac{1}{2}\left(1 + \frac{2}{\pi}\arctan(k(\alpha)(x-r))\right),$$

with

$$h'_\alpha(d) = \frac{1-\gamma}{\pi}\frac{k(\alpha)}{1+k(\alpha)^2(d-r)^2}.$$

Here,

$$h'_\alpha(r) = \frac{1}{\pi}k(\alpha)$$

so that we obtain

$$k(\alpha) = \pi\tan(\alpha).$$

Our idea is to tune the smoothing parameter α to optimise the MH algorithm in terms of speed of convergence. First we spell out details of the algorithm. At step $k+1$, the proposal used in the MH algorithm is the Euler discretization of the Langevin diffusion process (1) defined for $\pi^\alpha(\boldsymbol{x})$ given by

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + \frac{1}{2}\nabla\log\pi^\alpha(\boldsymbol{X}_k)\ \delta + \sqrt{\delta}\ \boldsymbol{\varepsilon}_k$$

where $\boldsymbol{\varepsilon}_k \sim \mathcal{N}(0, I_{n\times s})$ and $I_{n\times s}$ is the identity matrix of dimension $n \times s$.

We now compute the drift term in the diffusion process $b(\boldsymbol{x}) = \frac{1}{2}\nabla \log \pi^{\alpha}(\boldsymbol{x})\ \delta$, where

$$[b(\boldsymbol{x})]_i = [\nabla \log \pi^{\alpha}(\boldsymbol{x})]_i = \frac{d}{d\boldsymbol{x}_i} \sum_{k=1}^{n} \sum_{j=k+1}^{n} \log(h_{\alpha,\gamma}(d_{kj}))$$

$$= \sum_{j \neq i}^{n} \frac{(1-\gamma)\frac{d}{dd_{ij}} h_{\alpha}(d_{ij})}{h_{\alpha,\gamma}(d_{ij})} \frac{d}{d\boldsymbol{x}_i} d_{ij} = \sum_{j \neq i}^{n} b_0(d_{ij}) \frac{d}{d\boldsymbol{x}_i} d_{ij}.$$

For the exponential smoother we have

$$b_0(d) = \frac{(1-\gamma) h_{\alpha}^2(d) k(\alpha) \left(\frac{R-r}{(R-d)^2} + \frac{r}{d^2}\right) e^{-k(\alpha) f(d)}}{h_{\alpha,\gamma}(d)},$$

while for the arctangent smoother we get

$$b_0(d) = \frac{(1-\gamma) k(\alpha)}{h_{\alpha,\gamma}(d) \pi (1 + k(\alpha)^2 (d-r)^2)}.$$

Next we have to distinguish between the torus and non-torus geometry. For the non-torus case the derivatives of d_{ij} are

$$\frac{d}{d\boldsymbol{x}_i} d_{ij} = \frac{d}{d\boldsymbol{x}_i} \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T (\boldsymbol{x}_i - \boldsymbol{x}_j)} = \frac{(\boldsymbol{x}_i - \boldsymbol{x}_j)}{d_{ij}}.$$

The torus case on the other hand has derivatives

$$[\frac{d}{d\boldsymbol{x}_i} d_{ij}]_l = \frac{1}{d_{ij}} \min(1 - |x_{i,l} - x_{j,l}|, |x_{i,l} - x_{j,l}|) \begin{cases} \operatorname{sign}(x_{i,l} - x_{j,l}) & |x_{i,l} - x_{j,l}| < 0.5, \\ -\operatorname{sign}(x_{i,l} - x_{j,l}) & |x_{i,l} - x_{j,l}| \geq 0.5. \end{cases}$$

As mentioned in the introduction, Roberts & Rosenthal (1995) suggested an adjustment of MALA called MALTA. In this algorithm the drift term is replaced by the truncated drift term

$$b_t(\boldsymbol{x}) = (t_r \sqrt{\delta}) \wedge b(\boldsymbol{x}),$$

where t_r is a given truncation parameter.

In Figures 1 and 2 we plot the drift term for the two different smoothing functions and various values of α. The case $\alpha = 0$ corresponds to the random walk proposal (zero drift everywhere); for $\alpha = 90$, the drift term is zero in all points $d \neq r$, while it is not defined for $d = r$. We will later see that the optimal values of α will be for α close to 90°. Here, the drift term has the largest values in a small region centred in r. This means that points separated approximately by r are most strongly repulsed.

3. Criteria of convergence: Asymptotic variance

Suppose that we want to estimate $m = E_{\pi}\{g(\boldsymbol{X})\}$ for some integrable function $g(\boldsymbol{x})$. Under certain regularity conditions the Central Limit Theorem for Markov chains holds (see e.g. Kipnis & Varadhan (1986)), so that as $T \to \infty$,

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T} g(\boldsymbol{X}^t) - m\right) \to \mathcal{N}(0, \tau^2)$$

in law, where the asymptotic variance τ^2 is given by $\tau^2 = \gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k$ and, under stationarity, $\gamma_k = \operatorname{Cov}(g(\boldsymbol{X}_i), g(\boldsymbol{X}_{i+k}))$ for all i.

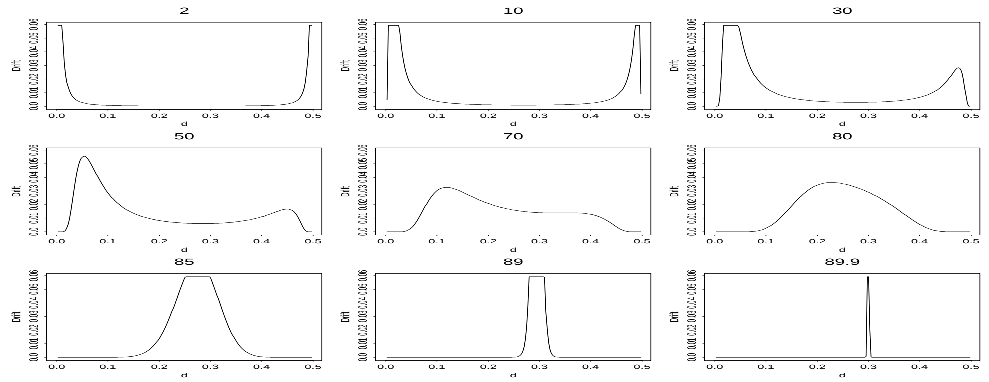

FIGURE 1. The truncated drift term for various values of the smoothing parameter α in the torus case for the exponential smoother having $\gamma = 0.1$, $r = 0.3$, $\delta = 0.00625$ and $t_r = 1.5$.

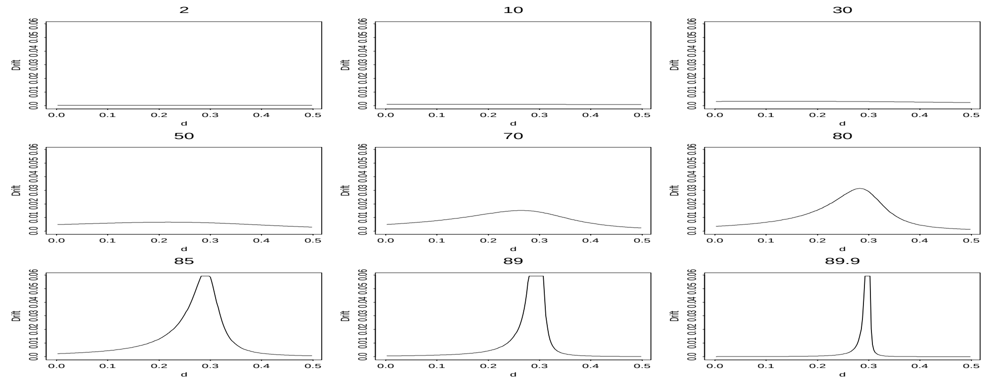

FIGURE 2. The truncated drift term for various values of the smoothing parameter α in the torus case for the arctangent smoother having $\gamma = 0.1$, $r = 0.3$, $\delta = 0.00625$ and $t_r = 1.5$.

The asymptotic variance will be used as the criteria of convergence and for comparison between different algorithms. It can be estimated by the initial positive sequence estimator (Geyer 1992)

$$\hat{\tau}^2 = \hat{\gamma}_0 + 2\hat{\gamma}_1 + \sum_{k=1}^{M} \hat{\Lambda}_k,$$

where $\hat{\Lambda}_k = \hat{\gamma}_{2k} + \hat{\gamma}_{2k+1}$ and M is the largest integer such that $\hat{\Lambda}_k$ are strictly positive for $k = 1, \ldots, M$. We shall compare various algorithms in terms of $\hat{\tau}^2$, preferring those for which this is smaller. We shall tune α to minimize $\hat{\tau}^2$.

4. Simulation study

The function g is chosen to be the number of pairwise overlaps,

$$g(\boldsymbol{X}) = \sum_{i=1}^{n} \sum_{j>i}^{n} 1_{[r,R]}(d_{ij}),$$

which is the sufficient statistic of the conditional Strauss model. The asymptotic variance is computed for different values of α with the other parameters fixed. The optimal α which minimizes the asymptotic variance is found. If the smoothed MALTA algorithm is to be useful, the optimal value of α should be different from $0°$ which corresponds to a random walk proposal. Furthermore the associated $\hat{\tau}_\alpha^2$ should be significantly smaller than $\hat{\tau}_{0°}^2$, where $\hat{\tau}_\alpha$ denotes the estimated asymptotic variance when the smoothed MALTA algorithm with smoothing α is used.

There are a number of other parameters to vary, including the model parameters n, s, r and γ and the algorithm parameters δ and t_r. We fix t_r to 1.5. Simulation studies indicate that the results are not too sensitive to variations in t_r. However, a too high value of t_r could cause the state vector to get stuck in certain point configurations. The parameters r and γ determine the degree of model complexity: a high value of r and a low value of γ makes it hard to sample from π. We have chosen to fix $\gamma = 0.1$ and then adjust r. The other parameters are determined in the following way. First, fix $n \in \{2, 3, \ldots ,\}$ and $s \in \{1, 2\}$. We choose the value of r, so that a reasonable difficulty in sampling is obtained; so let

$$r = \begin{cases} r_0(1/n)^{(1/s)} & \text{for the torus geometry,} \\ r_0(1/(n-1))^{(1/s)} & \text{for the non-torus geometry.} \end{cases}$$

$r_0 = 1$ gives approximately the largest feasible value of r in a hard core model with n points in $[0, 1]^s$. A larger value of r would make it very hard to place the n points. The factor r_0 must be adjusted somewhat as a function of n and s to keep the same difficulty in sampling; e.g for n equal to 3 and 10, we choose r_0 to be 0.9 and 0.6 respectively. The model complexity can be measured by the acceptance probability of the algorithm. For δ we choose an optimal δ_{opt} by simulations over a range of δ values and $\alpha = 0°$. So δ_{opt} is chosen as the δ minimizing $\tau_{0°}$. The layout of the experiments is given in Table 1. The $P[accept]$ column contains the average acceptance probabilities of a rejection sampler with uniform proposals.

Experiment	n	s	Torus	δ_{opt}	r	$P[accept]$
1	3	1	N	0.0062	0.450	0.037
2	3	1	Y	0.0062	0.300	0.057
3	3	2	N	0.0125	0.636	0.065
4	3	2	Y	0.0250	0.520	0.012
5	5	1	N	0.0039	0.200	0.007
6	5	1	Y	0.0020	0.160	0.012
7	5	2	N	0.0039	0.400	0.016
8	5	2	Y	0.0039	0.358	0.004
9	10	1	N	0.0005	0.067	0.0009
10	10	1	Y	0.0005	0.060	0.002

Table 1. The ten experiments.

Experiment	1	2	3	4	5	6	7	8	9	10
$\hat{\alpha}_{opt}$	70°	70°	50°	80°	50°	80°	85°	80°	89°	80°
$\hat{\tau}_{opt}$	4.16	3.52	31.8	3.38	21.7	11.3	58.3	18.1	104.8	50.3
$\hat{\tau}_{0^\circ}$	4.87	4.03	35.6	4.03	23.6	12.9	61.8	22.7	115.6	55.4

Table 2. The optimal values for α and τ for the exponential smoother.

The results from the 10 simulation experiments is given in Figure 3 for the exponential smoother and in Figure 4 for the arctangent smoother. The plots in the figures summarise the result of 10 independent experiments, each with 2×10^5 iterations after a burn-in of 2×10^3 iterations. Each chain produces estimates of $\hat{\tau}_\alpha$ for a range of different values of α. The average value of these ten estimates is indicated by a dot, while estimated 95% confidence intervals for τ_α are indicated by the vertical lines through each dot.

A reduction of asymptotic variance with respect to τ_{0° is seen in all the 10 experiments. The optimal value of α varies, but in the torus case it seems to be stable around 70°-80°, see Table 2. The torus model with exponential smoother gives the largest reduction in asymptotic variance, about 15%. The arctangent smoother gives quite similar results, however the reduction in asymptotic variance seems to be smaller. This indicates that the smoothed Langevin proposal is useful in a MALTA setting.

In the next example we consider the smoothing of a continuous but steep target distribution π. We take the distribution π_α for α near 90° as target distribution in the torus model with the exponential smoother. In Figure 5, we see that for target π_{89° the optimal α was around 70° and there is a reduction in asymptotic variance of about 20%. Again this is an indication that smoothing helps.

Finally, a larger simulation study is performed, using the estimated parameters obtained for the Spanish town example (Ripley 1988). Here, $n = 69$, $s = 2$ and γ and r are estimated to 0.5 and 0.0875 respectively. The experiment consists of a total of 45 chains of length 2×10^5 iterations after a burn-in of 2×10^3 iterations. The torus geometry and the exponential smoother are considered. See Figure 6 for the simulation results for different values of α. The optimum α_{opt} seems here to be around $\alpha = 89^\circ$. This indicates that the optimal value of α gets closer to 90° as $n \to \infty$ or $r \to 0$.

5. Concluding remarks

In the non-torus case the point pattern proposed by the smoothed MALTA algorithm is often rejected because some points are, due to repulsion, “forced” outside the legal region $[0, 1]^s$. We expect therefore the smoothed MALTA algortihm to perform better on the torus geometry. The simulation results in Figures 3 and 4 confirm this. The simulations indicate a reduction of approximately 15% of the asymptotic variance for the torus case when the exponential smoother is used with respect to random walk MH. The optimal α for the torus case and the exponential smoother seems to be stable around 70°-80° for small n, and seems to increase (Spanish towns example) closer to 90° for larger n. The performance of the smoothed MALTA algorithm as compared with random walk MH (corresponding to the $\alpha = 0^\circ$ case), may depend further on the other model and algorithm parameters. A larger simulation study is needed to see the dependence of α_{opt} on different model parameters and on t_r.

The value δ_{opt}, chosen to minimize $\tau_{\alpha=0^\circ}$, may not minimize τ_α for $\alpha \neq 0^\circ$. Our results may therefore be improved by finding a better δ. However, a simulation study of experiment 2 showed that the optimal δ did not differ for $\alpha = 0^\circ$ and $\alpha = 70^\circ$.

We conclude with a very important remark: When comparing the smoothed MALTA algorithm with the more simpler MH algorithm with random walk proposal, we do not take into account that the computation time of each iteration is larger. A look-up table for the drift term, initialised at the beginning of the simulation program, does however reduce the additional computation time considerably.

Acknowledgements: We thank Jesper Møller for helpful comments.

References

Geyer, C. J. (1992), 'Practical Markov chain Monte Carlo', *Statistical Science* **7**, 473–483.

Kipnis, C. & Varadhan, S. R. S. (1986), 'Central Limit Theorem for additive functionals of reversible Markov processes and applications to simple exclusions', *Comm. Math. Phys.* **104**, 1–19.

Ripley, B. D. (1988), *Statistical Inference for Spatial Processes*, Cambridge University Press, Cambridge.

Roberts, G. O. & Rosenthal, J. S. (1995), Optimal scaling of discrete approximations to Langevin diffusions, Research Report 95-11, University of Cambridge.

Roberts, G. O. & Tweedie, R. L. (1996), 'Exponential convergence of Langevin diffusions and their discrete approximations', *Bernoulli* **2**, 341–364.

Strauss, D. J. (1975), 'A model for clustering', *Biometrika* **62**, 467–475.

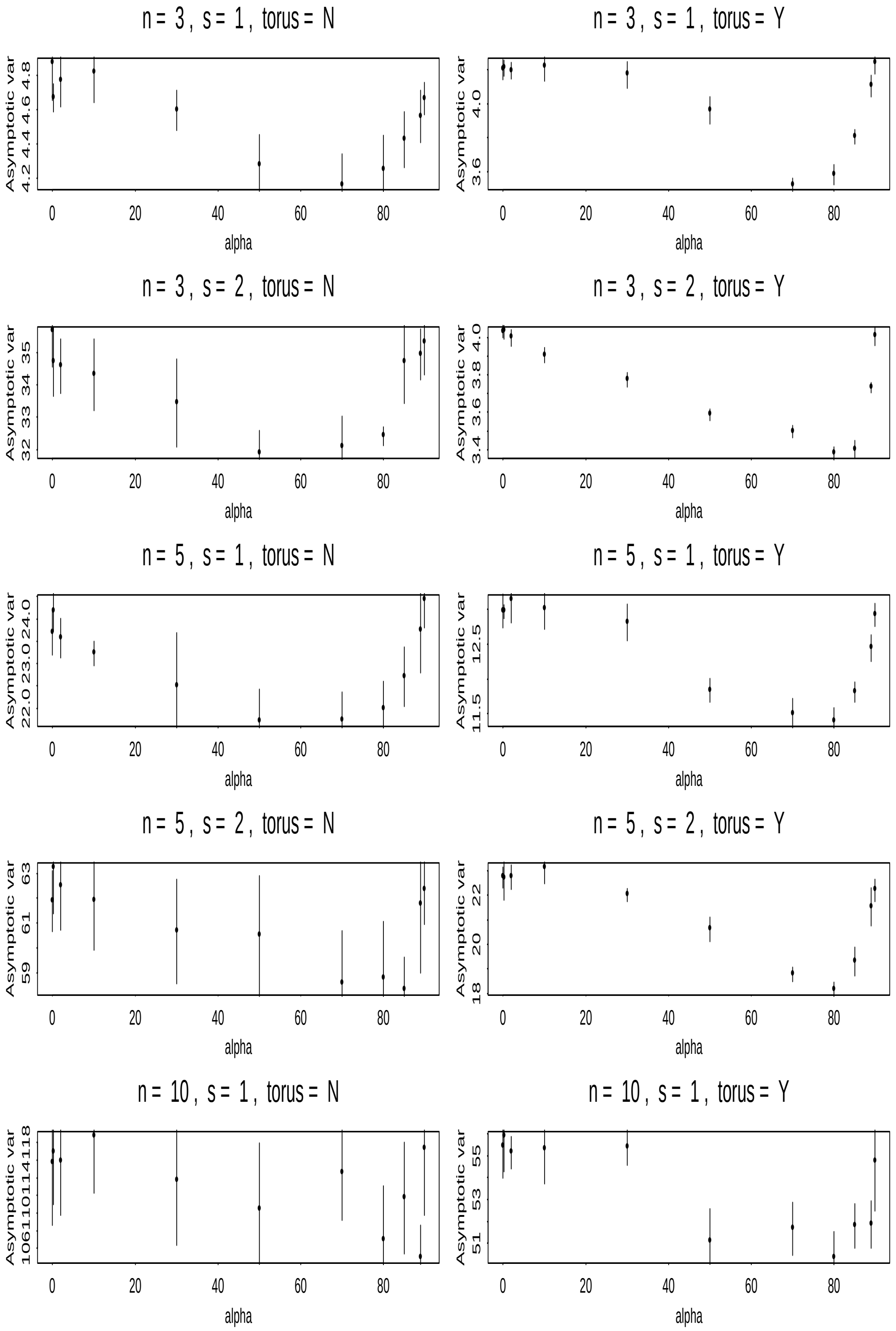

FIGURE 3. Asymptotic variances $\hat{\tau}_\alpha$ for the ten experiments corresponding to Table 1 using the exponential smoother for various values of α.

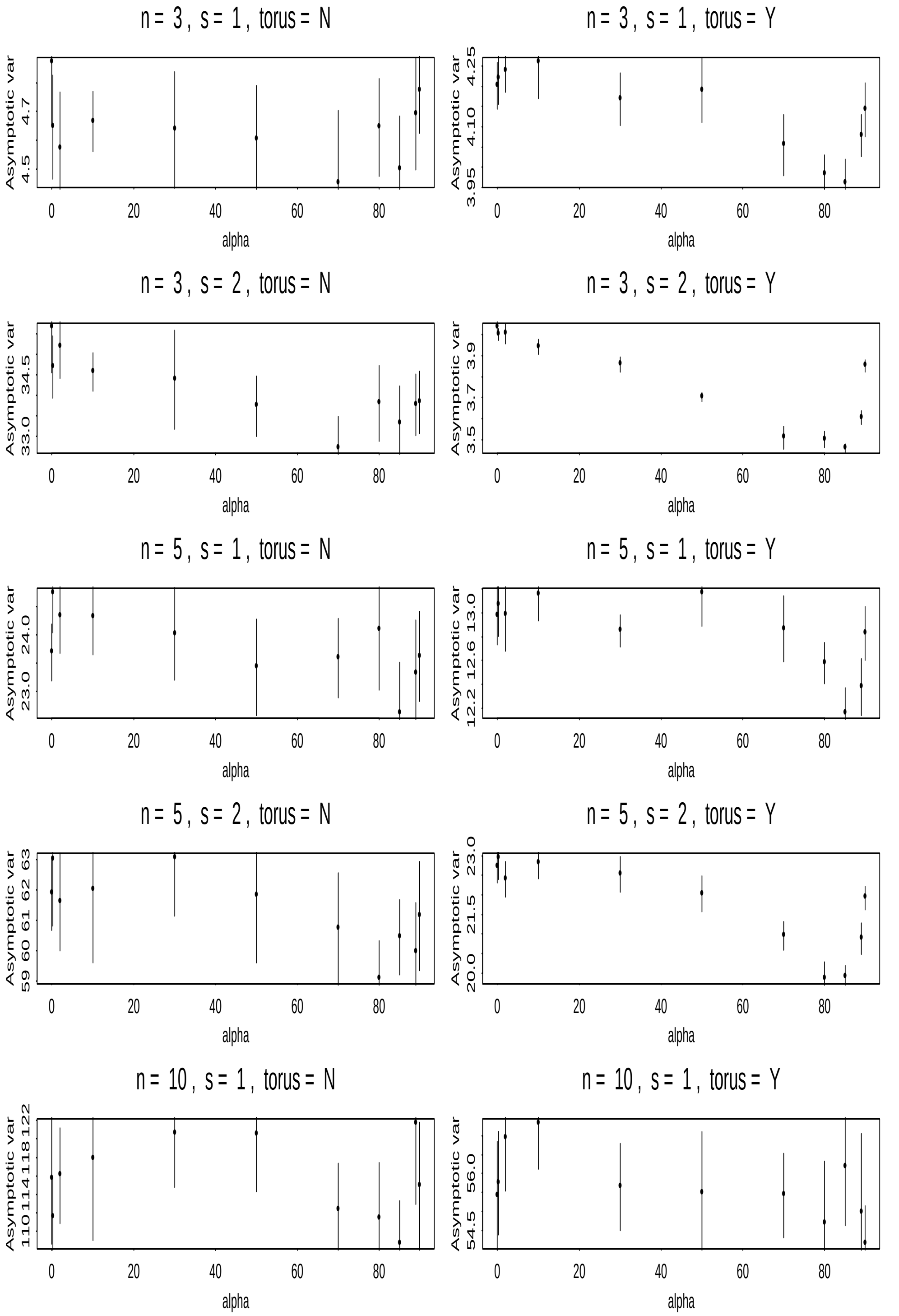

FIGURE 4. Asymptotic variances $\hat{\tau}_\alpha$ for the ten experiments corresponding to Table 1 using the arctangent smoother for various values of α.

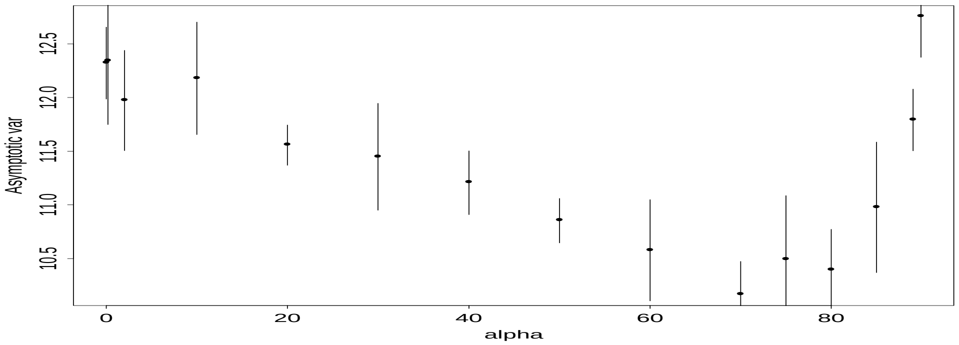

FIGURE 5. Experiment with continuous target π_{89°. The parameters correspond otherwise to experiment 6 in Table 1.

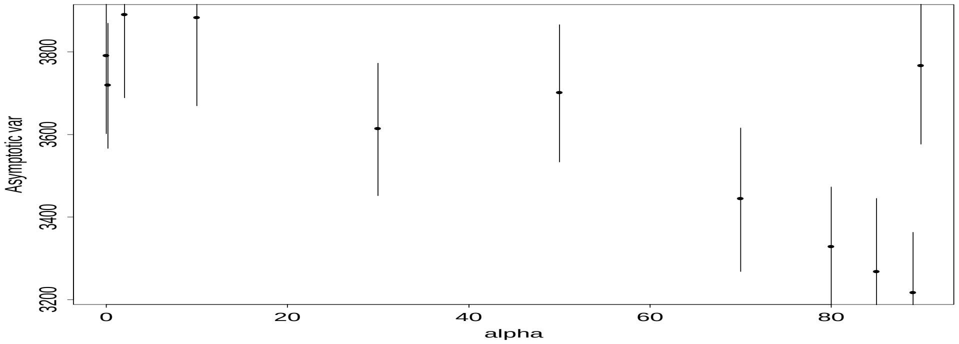

FIGURE 6. Experiment corresponding to the Spanish town example with $n = 69$, $s = 2$, $\gamma = 0.5$ and $r = 0.0875$.