Adaptive Monte Carlo Maximum Likelihood

Błażej Miasojedow, Wojciech Niemiro, Jan Palczewski
and Wojciech Rejchel

Abstract We consider Monte Carlo approximations to the maximum likelihood estimator in models with intractable norming constants. This paper deals with adaptive Monte Carlo algorithms, which adjust control parameters in the course of simulation. We examine asymptotics of adaptive importance sampling and a new algorithm, which uses resampling and MCMC. This algorithm is designed to reduce problems with degeneracy of importance weights. Our analysis is based on martingale limit theorems. We also describe how adaptive maximization algorithms of Newton-Raphson type can be combined with the resampling techniques. The paper includes results of a small scale simulation study in which we compare the performance of adaptive and non-adaptive Monte Carlo maximum likelihood algorithms.

Keywords Maximum likelihood · Importance sampling · Adaptation · MCMC · Resampling

B. Miasojedow · W. Niemiro (✉)
Faculty of Mathematics, Informatics and Mechanics,, University of Warsaw,
Warsaw, Poland
e-mail: W.Niemiro@mimuw.edu.pl

B. Miasojedow
e-mail: W.Miasojedow@mimuw.edu.pl

J. Palczewski
School of Mathematics, University of Leeds, Leeds, UK
e-mail: J.Palczewski@leeds.ac.uk

W. Niemiro · W. Rejchel
Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Toruń, Poland
e-mail: wrejchel@gmail.com

1 Introduction

Maximum likelihood (ML) is a well-known and often used method in estimation of parameters in statistical models. However, for many complex models exact calculation of such estimators is very difficult or impossible. Such problems arise if considered densities are known only up to intractable norming constants, for instance in Markov random fields or spatial statistics. The wide range of applications of models with unknown norming constants is discussed e.g. in [10]. Methods proposed to overcome the problems with computing ML estimates in such models include, among others, maximum pseudolikelihood [2], "coding method" [9] and Monte Carlo maximum likelihood (MCML) [5, 9, 15, 17]. In our paper we focus on MCML.

  In influential papers [4, 5] the authors prove consistency and asymptotic normality of MCML estimators. To improve the performance of MCML, one can adjust control parameters in the course of simulation. This leads to adaptive MCML algorithms. We generalize the results of the last mentioned papers first to an adaptive version of importance sampling and then to a more complicated adaptive algorithm which uses resampling and Markov chain Monte Carlo (MCMC) [7]. Our analysis is asymptotic and it is based on the martingale structure of the estimates. The main motivating examples are the autologistic model (with or without covariates) and its applications to spatial statistics as described e.g. in [9] and the autonormal model [11].

2 Adaptive Importance Sampling

Denote by $f_\theta, \theta \in \Theta$, a family of unnormalized densities on space $\mathcal{Y}$. A dominating measure with respect to which these densities are defined is denoted for simplicity by dy. Let y_{obs} be an observation. We intend to find the maximizer $\theta_\star$ of the log-likelihood

$$\ell(\theta) = \log f_\theta(y_{\text{obs}}) - \log c(\theta),$$

where $c(\theta)$ is the normalizing constant. We consider the situation where this constant,

$$c(\theta) = \int_{\mathcal{Y}} f_\theta(y)dy,$$

is *unknown and numerically intractable*. It is approximated with Monte Carlo simulation, resulting in

$$\ell_m(\theta) = \log f_\theta(y_{\text{obs}}) - \log \hat{c}_m(\theta), \tag{2.1}$$

where $\hat{c}_m(\theta)$ is a Monte Carlo (MC) estimate of $c(\theta)$. The classical importance sampling (IS) estimate is of the form

$$\hat{c}_m(\theta) = \frac{1}{m} \sum_{j=1}^{m} \frac{f_\theta(Y_j)}{h(Y_j)}, \qquad (2.2)$$

where $Y_1, \ldots, Y_m$ are i.i.d. samples from an instrumental density h. Clearly, an optimal choice of h depends on the maximizer $\theta_\star$ of ℓ, so we should be able to improve our initial guess about h while the simulation progresses. This is the idea behind *adaptive importance sampling* (AIS). A discussion on the choice of instrumental density is deferred to subsequent subsections. Let us describe an adaptive algorithm in the following form, suitable for further generalizations. Consider a parametric family $h_\psi, \psi \in \Psi$ of instrumental densities.

<div align="center">Algorithm AdapIS</div>

1. Set an initial value of $\psi_1, m = 1, \hat{c}_0(\theta) \equiv 0$.
2. Draw $Y_m \sim h_{\psi_m}$.
3. Update the approximation of $c(\theta)$:

$$\hat{c}_m(\theta) = \frac{m-1}{m} \hat{c}_{m-1}(\theta) + \frac{1}{m} \frac{f_\theta(Y_m)}{h_{\psi_m}(Y_m)}.$$

4. Update ψ: choose ψ_{m+1} based on the history of the simulation.
5. $m = m + 1$; go to 2.

At the output of this algorithm we obtain an AIS estimate

$$\hat{c}_m(\theta) = \frac{1}{m} \sum_{j=1}^{m} \frac{f_\theta(Y_j)}{h_{\psi_j}(Y_j)}. \qquad (2.3)$$

The samples Y_j are neither independent nor have the same distribution. However (2.3) has a nice *martingale* structure. If we put

$$\mathcal{F}_m = \sigma \left\{ Y_j, \psi_j : j \le m \right\}$$

then ψ_{m+1} is $\mathcal{F}_m$-measurable. The well-known property of unbiasedness of IS implies that

$$\mathbb{E} \left(\frac{f_\theta(Y_{m+1})}{h_{\psi_{m+1}}(Y_{m+1})} \middle| \mathcal{F}_m \right) = c(\theta). \qquad (2.4)$$

In other words, $f_\theta(Y_m)/h_{\psi_m}(Y_m) - c(\theta)$ are martingale differences (MGD), for every fixed θ.

2.1 Hypo-convergence of ℓ_m and Consistency of $\hat{\theta}_m$

In this subsection we make the following assumptions.

Assumption 1 For any $\theta \in \Theta$

$$\sup_{\psi} \int \frac{f_\theta(y)^2}{h_\psi(y)} \mathrm{d}y < \infty.$$

Assumption 2 The mapping $\theta \mapsto f_\theta(y)$ is continuous for each fixed y.

Assumption 1 implies that for any θ, there is a constant $M_\theta < \infty$ such that for all j,

$$\mathbb{E}\left(\left(\frac{f_\theta(Y_j)}{h_{\psi_j}(Y_j)} \right)^2 \Big| \mathcal{F}_{j-1} \right) \leq M_\theta, \quad \text{a.s.},$$

because $Y_j \sim h_{\psi_j}$. Note that Assumption 1 is trivially true if the mapping $y \mapsto f_\theta(y)/h_\psi(y)$ is uniformly bounded for $\theta \in \Theta$, $\psi \in \Psi$. Recall also that

$$m(\hat{c}_m(\theta) - c(\theta)) = \sum_{j=1}^{m} \left(\frac{f_\theta(Y_j)}{h_{\psi_j}(Y_j)} - c(\theta) \right)$$

is a zero-mean martingale. Under Assumption 1, for a fixed $\theta \in \Theta$, we have $\hat{c}_m(\theta) \to c(\theta)$ a.s. by the SLLN for martingales (see Theorem A.2, Appendix A), so $\ell_m(\theta) \to \ell(\theta)$ a.s. This is, however, insufficient to guarantee the convergence of maximum likelihood estimates $\hat{\theta}_m$ (maximizers of ℓ_m) to $\theta_\star$. Under our assumptions we can prove hypo-convergence of the log-likelihood approximations.

Definition 1 A sequence of functions g_m epi-converges to g if for any x we have

$$\sup_{B \in N(x)} \limsup_{m \to \infty} \inf_{y \in B} g_m(y) \leq g(x),$$

$$\sup_{B \in N(x)} \liminf_{m \to \infty} \inf_{y \in B} g_m(y) \geq g(x),$$

where $N(x)$ is a family of all (open) neighbourhoods of x.

A sequence of functions g_m hypo-converges to g if $(-g_m)$ epi-converges to $(-g)$.

An equivalent definition of epi-convergence is in the following theorem:

Theorem 1 ([14, Proposition 7.2]) g_m *epi-converges to* g *iff at every point* x

$$\limsup_{m \to \infty} g_m(x_m) \leq g(x), \quad \text{for some sequence } x_m \to x,$$

$$\liminf_{m \to \infty} g_m(x_m) \geq g(x), \quad \text{for every sequence } x_m \to x.$$

wrejchel@gmail.com

As a corollary to this theorem comes the proposition that will be used to prove convergence of $\hat{\theta}_m$, the maximizer of ℓ_m, to $\theta_\star$ (see, also, [1, Theorem 1.10]).

Proposition 1 *Assume that g_m epi-converges to g, $x_m \to x$ and $g_m(x_m) - \inf g_m \to$ 0. Then $g(x) = \inf_y g(y) = \lim_{m\to\infty} g_m(x_m)$.*

Proof (We will use Theorem 1 many times.) Let y_m be a sequence converging to x and such that $\limsup_{m\to\infty} g_m(y_m) \leq g(x)$ (such sequence y_m exists). This implies that $\limsup_{m\to\infty} \inf g_m \leq g(x)$. On the other hand, $g(x) \leq \liminf_{m\to\infty} g_m(x_m) = \liminf_{m\to\infty} \inf g_m$, where the equality follows from the second assumption on x_m. Summarizing, $g(x) = \lim_{m\to\infty} \inf g_m = \lim_{m\to\infty} g_m(x_m)$. In particular, $\inf g \leq \lim_{m\to\infty} \inf g_m$.

Take any $\varepsilon > 0$ and let x_ε be such that $g(x_\varepsilon) \leq \inf g + \varepsilon$. There exists a sequence y_m converging to x_ε such that $g(x_\varepsilon) \geq \limsup_{m\to\infty} g_m(y_m) \geq \limsup_{m\to\infty} \inf g_m$, hence $\lim_{m\to\infty} \inf g_m \leq \inf g + \varepsilon$. By arbitrariness of $\varepsilon > 0$ we obtain $\lim_{m\to\infty} \inf g_m \leq \inf g$. This completes the proof.

Theorem 2 *If Assumptions 1 and 2 are satisfied, then ℓ_m hypo-converges to ℓ almost surely.*

Proof The proof is similar to the proof of Theorem 1 in [4]. We have to prove that $\hat{c}_m$ epi-converges to c. Fix $\theta \in \Theta$.

Step 1: For any $B \in N(\theta)$, we have

$$\liminf_{m\to\infty} \inf_{\varphi\in B} \hat{c}_m(\varphi) \geq \int \inf_{\varphi\in B} f_\varphi(y)\mathrm{d}y =: \underline{c}(B). \qquad (2.5)$$

Indeed,

$$\inf_{\varphi\in B} \hat{c}_m(\phi) = \inf_{\varphi\in B} \frac{1}{m} \sum_{j=1}^{m} \frac{f_\varphi(Y_j)}{h_{\psi_j}(Y_j)} \geq \frac{1}{m} \sum_{j=1}^{m} \inf_{\varphi\in B} \frac{f_\varphi(Y_j)}{h_{\psi_j}(Y_j)}$$

$$= \frac{1}{m} \sum_{j=1}^{m} \left(\inf_{\varphi\in B} \frac{f_\varphi(Y_j)}{h_{\psi_j}(Y_j)} - \underline{c}(B) \right) + \underline{c}(B).$$

The sum is that of martingale differences, so assuming that there is $M < \infty$ such that

$$\sup_j \mathbb{E}\left(\left(\inf_{\varphi\in B} \frac{f_\varphi(Y_j)}{h_{\psi_j}(Y_j)} - \underline{c}(B) \right)^2 \middle| \mathcal{F}_{j-1} \right) \leq M$$

wrejchel@gmail.com

the SLLN implies (2.5). We have the following estimates:

$$
\mathbb{E}\left(\left(\inf_{\varphi \in B} \frac{f_\varphi(Y_j)}{h_{\psi_j}(Y_j)} - \underline{c}(B)\right)^2 \middle| \mathcal{F}_{j-1}\right) = \mathrm{Var}\left(\inf_{\varphi \in B} \frac{f_\varphi(Y_j)}{h_{\psi_j}(Y_j)} \middle| \mathcal{F}_{j-1}\right)
$$

$$
\leq \mathbb{E}\left(\left(\inf_{\varphi \in B} \frac{f_\varphi(Y_j)}{h_{\psi_j}(Y_j)}\right)^2 \middle| \mathcal{F}_{j-1}\right) \leq \mathbb{E}\left(\left(\frac{f_\theta(Y_j)}{h_{\psi_j}(Y_j)}\right)^2 \middle| \mathcal{F}_{j-1}\right) \leq M_\theta,
$$

where the last inequality is by Assumption 1.

Step 2: We shall prove that $\sup_{B \in N(\theta)} \liminf_{m \to \infty} \inf_{\varphi \in B} \hat{c}_m(\phi) \geq c(\theta)$.

The left-hand side is bounded from below by $\sup_{B \in N(\theta)} \underline{c}(B)$. Further, we have

$$
\sup_{B \in N(\theta)} \underline{c}(B) \geq \lim_{\delta \downarrow 0} \underline{c}(B(\theta, \delta)) = \int \lim_{\delta \downarrow 0} \inf_{\varphi \in B(\theta, \delta)} f_\varphi(y)\mathrm{d}y = \int f_\theta(y)\mathrm{d}y = c(\theta),
$$

where the first equality follows from the dominated convergence theorem (the dominator is f_θ) and the last—from the Assumption 2.

Step 3: We have

$$
\sup_{B \in N(\theta)} \limsup_{m \to \infty} \inf_{\varphi \in B} \hat{c}_m(\varphi) \leq \sup_{B \in N(\theta)} \inf_{\varphi \in B} \limsup_{m \to \infty} \hat{c}_m(\varphi) = \sup_{B \in N(\theta)} \inf_{\varphi \in B} c(\varphi) \leq c(\theta).
$$

Hence, $\sup_{B \in N(\theta)} \limsup_{m \to \infty} \inf_{\varphi \in B} \hat{c}_m(\varphi) \leq c(\theta)$.

Note that almost sure convergence in the next Proposition corresponds to the randomness introduced by AdapIS and y_{obs} is fixed throughout this paper.

Proposition 2 *If Assumptions 1 and 2 hold, $\theta_\star$ is the unique maximizer of ℓ and sequence $(\hat{\theta}_m)$ (where $\hat{\theta}_m$ maximizes ℓ_m) is almost surely bounded then $\hat{\theta}_m \to \theta_\star$ almost surely.*

Proof As we have already mentioned, by SLLN for martingales, $\ell_m(\theta) \to \ell(\theta)$, pointwise. Hypo-convergence of ℓ_m to ℓ implies, by Proposition 1, that the maximizers of ℓ_m have accumulation points that are the maximizers of ℓ. If ℓ has a unique maximizer $\theta_\star$ then any convergent subsequence of $\hat{\theta}_m$, maximizers of ℓ_m, converges to $\theta_\star$. The conclusion follows immediately.

Of course, it is not easy to show boundedness of $\hat{\theta}_m$ in concrete examples. In the next section we will prove consistency of $\hat{\theta}_m$ in models where log-likelihoods and their estimates are concave.

2.2 Central Limit Theorem for Adaptive Importance Sampling

Let $\hat{\theta}_m$ be a maximizer of ℓ_m, i.e. the AIS estimate of the likelihood given by (2.1) with (2.3). We assume that $\theta_\star$ is a unique maximizer of ℓ. For asymptotic normality of $\hat{\theta}_m$, we will need the following assumptions.

Assumption 3 First and second order derivatives of f_θ with respect to θ (denoted by ∇f_θ and $\nabla^2 f_\theta$) exist in a neighbourhood of $\theta_\star$ and we have

$$\nabla c(\theta) = \int \nabla f_\theta(y)\mathrm{d}y, \qquad \nabla^2 c(\theta) = \int \nabla^2 f_\theta(y)\mathrm{d}y.$$

Assumption 4 $\hat\theta_m = \theta_\star + O_\mathrm{p}(1/\sqrt{m})$.

Assumption 5 Matrix $D = \nabla^2 \ell(\theta_\star)$ is negative definite.

Assumption 6 For every y, function $\psi \mapsto h_\psi(y)$ is continuous and $h_\psi(y) > 0$.

Assumption 7 For some $\psi_\star$ we have $\psi_m \to \psi_\star$ almost surely.

Assumption 8 There exists a nonnegative function g such that $\int g(y)\mathrm{d}y < \infty$ and the inequalities

$$\sup_\psi \frac{f_{\theta_\star}(y)^{2+\alpha}}{h_\psi(y)^{1+\alpha}} \le g(y), \quad \sup_\psi \frac{|\nabla f_{\theta_\star}(y)|^{2+\alpha}}{h_\psi(y)^{1+\alpha}} \le g(y),$$

$$\sup_\psi \frac{\|\nabla^2 f_{\theta_\star}(y)\|^{1+\alpha}}{h_\psi(y)^\alpha} \le g(y)$$

are fulfilled for some $\alpha > 0$ and also for $\alpha = 0$.

Assumption 9 Functions $\nabla^2 \ell_m(\theta)$ are asymptotically stochastically equicontionuous at $\theta_\star$, i.e. for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\limsup_{m\to\infty} \mathbb{P}\left(\sup_{|\theta - \theta_\star| \le \delta} \|\nabla^2 \ell_m(\theta) - \nabla^2 \ell_m(\theta_\star)\| > \varepsilon \right) = 0.$$

Let us begin with some comments on these assumptions and note simple facts which follow from them. Assumption 3 is a standard regularity condition. It implies that a martingale property similar to (2.4) holds also for the gradients and hessians:

$$\mathbb{E}\left(\frac{\nabla f_\theta(Y_{m+1})}{h_{\psi_{m+1}}(Y_{m+1})} \Big| \mathcal{F}_m \right) = \nabla c(\theta), \quad \mathbb{E}\left(\frac{\nabla^2 f_\theta(Y_{m+1})}{h_{\psi_{m+1}}(Y_{m+1})} \Big| \mathcal{F}_m \right) = \nabla^2 c(\theta). \quad (2.6)$$

Assumption 4 stipulates square root consistency of $\hat\theta_m$. It is automatically fulfilled if ℓ_m is concave, in particular for exponential families. Assumption 7 combined with 6 is a "diminishing adaptation" condition. It may be ensured by an appropriately specifying step 4 of `AdapIS`. The next assumptions are not easy to verify in general, but they are satisfied for exponential families on finite spaces, in particular for our "motivating example": autologistic model. Let us also note that our Assumption 9 plays a similar role to Assumption (f) in [4, Theorem 7].

Assumption 8 together with (2.4) and (2.6) allows us to apply SLLN for martingales in a form given in Theorem A.2, Appendix A. Indeed, $f_{\theta_\star}(Y_m)/h_{\psi_m}(Y_m) - c(\theta_\star)$, $\nabla f_{\theta_\star}(Y_m)/h_{\psi_m}(Y_m) - \nabla c(\theta_\star)$ and $\nabla^2 f_{\theta_\star}(Y_m)/h_{\psi_m}(Y_m) - \nabla^2 c(\theta_\star)$ are MGDs with bounded moments of order $1 + \alpha > 1$. It follows that, almost surely,

$$\hat{c}_m(\theta_\star) \to c(\theta_\star), \qquad \nabla\hat{c}_m(\theta_\star) \to \nabla c(\theta_\star), \qquad \nabla^2\hat{c}_m(\theta_\star) \to \nabla^2 c(\theta_\star). \qquad (2.7)$$

Now we are in a position to state the main result of this section.

Theorem 3 *If Assumptions 3–9 hold then*

$$\sqrt{m}\left(\hat{\theta}_m - \theta_\star\right) \to \mathcal{N}(0, D^{-1}VD^{-1}) \quad \text{in distribution,}$$

where $D = \nabla^2\ell(\theta_\star)$ and

$$V = \frac{1}{c(\theta_\star)^2}\mathrm{VAR}_{Y\sim h_{\psi_\star}}\left[\frac{\nabla f_{\theta_\star}(Y)}{h_{\psi_\star}(Y)} - \frac{\nabla c(\theta_\star)}{c(\theta_\star)}\frac{f_{\theta_\star}(Y)}{h_{\psi_\star}(Y)}\right],$$

where $\psi_\star$ is defined in Assumption 7.

Proof It is well-known (see [12, Theorem VII.5]) that we need to prove

$$\sqrt{m}\nabla\ell_m(\theta_\star) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V) \qquad (2.8)$$

and that for every $M > 0$, the following holds:

$$\sup_{|\theta-\theta_\star|\leq M/\sqrt{m}} m\left|\ell_m(\theta) - \ell_m(\theta_\star)\right.$$
$$\left. - (\theta-\theta_\star)^\top\nabla\ell_m(\theta_\star) - \frac{1}{2}(\theta-\theta_\star)^\top D(\theta-\theta_\star)\right| \xrightarrow{\mathrm{p}} 0. \qquad (2.9)$$

First we show (2.8). Since $\nabla\ell_m(\theta) = \nabla f_\theta(y_{\mathrm{obs}})/f_\theta(y_{\mathrm{obs}}) - \nabla\hat{c}_m(\theta)/\hat{c}_m(\theta)$ and $\nabla\ell(\theta_\star) = \nabla f_{\theta_\star}(y_{\mathrm{obs}})/f_{\theta_\star}(y_{\mathrm{obs}}) - \nabla c(\theta_\star)/c(\theta_\star) = 0$, we obtain that

$$\nabla\ell_m(\theta_\star) = \frac{\nabla c(\theta_\star)}{c(\theta_\star)} - \frac{\nabla\hat{c}_m(\theta_\star)}{\hat{c}_m(\theta_\star)} = \frac{\dfrac{\nabla c(\theta_\star)}{c(\theta_\star)}\hat{c}_m(\theta_\star) - \nabla\hat{c}_m(\theta_\star)}{\hat{c}_m(\theta_\star)}. \qquad (2.10)$$

The denominator in the above expression converges to $c(\theta_\star)$ in probability, by (2.7). In view of Slutski's theorem, to prove (2.8) it is enough to show asymptotic normality of the numerator. We can write

$$\frac{\nabla c(\theta_\star)}{c(\theta_\star)}\hat{c}_m(\theta_\star) - \nabla\hat{c}_m(\theta_\star) = -\frac{1}{m}\sum_{j=1}^{m}\xi_j,$$

where we use the notation

$$\xi_j = \frac{\nabla f_{\theta_\star}(Y_j)}{h_{\psi_j}(Y_j)} - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} \frac{f_{\theta_\star}(Y_j)}{h_{\psi_j}(Y_j)}.$$

Now note that ξ_j are martingale differences by (2.4) and (2.6). Moreover,

$$\mathbb{E}\left(\xi_j \xi_j^T | \mathcal{F}_{j-1}\right) = \int \left(\frac{\nabla f_{\theta_\star}(y)}{h_{\psi_j}(y)} - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} \frac{f_{\theta_\star}(y)}{h_{\psi_j}(y)}\right)$$
$$\left(\frac{\nabla f_{\theta_\star}(y)}{h_{\psi_j}(y)} - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} \frac{f_{\theta_\star}(y)}{h_{\psi_j}(y)}\right)^\top h_{\psi_j}(y) \mathrm{d}y,$$

so Assumptions 6 and 7 via dominated convergence and Assumption 8 (with $\alpha = 0$ in the exponent) entail

$$\mathbb{E}\left(\xi_j \xi_j^T | \mathcal{F}_{j-1}\right) \xrightarrow{\text{a.s.}} c(\theta_\star)^2 V.$$

Now we use Assumption 8 (with $\alpha > 0$ in the exponent) to infer the Lyapunov-type condition

$$\mathbb{E}\left(|\xi_j|^{2+\alpha} | \mathcal{F}_{j-1}\right) \le \text{const} \cdot \int g(y) \mathrm{d}y < \infty.$$

The last two displayed formulas are sufficient for a martingale CLT (Theorem A.1, Appendix A). We conclude that

$$\frac{1}{\sqrt{m}} \sum_{j=1}^{m} \xi_j \xi_j^\top \xrightarrow{\text{d}} \mathcal{N}(0, c(\theta_\star)^2 V),$$

hence the proof of (2.8) is complete.

Now we proceed to a proof of (2.9). By Taylor expansion,

$$\ell_m(\theta) - \ell_m(\theta_\star) - (\theta - \theta_\star)^\top \nabla \ell_m(\theta_\star) = \frac{1}{2}(\theta - \theta_\star)^\top \nabla^2 \ell_m(\tilde{\theta})(\theta - \theta_\star)$$

for some $\tilde{\theta} \in [\theta, \theta_\star]$. Consequently, the LHS of (2.9) is

$$\le \sup_{\substack{|\theta - \theta_\star| \le M/\sqrt{m} \\ \tilde{\theta} \in [\theta, \theta_\star]}} m \left| \frac{1}{2}(\theta - \theta_\star)^\top \left(\nabla^2 \ell_m(\tilde{\theta}) - \nabla^2 \ell(\theta_\star)\right)(\theta - \theta_\star) \right|$$

$$\le \frac{M^2}{2} \sup_{|\theta - \theta_\star| \le M/\sqrt{m}} \left\| \nabla^2 \ell_m(\theta) - \nabla^2 \ell(\theta_\star) \right\|$$

$$\le \frac{M^2}{2} \sup_{|\theta - \theta_\star| \le M/\sqrt{m}} \left\| \nabla^2 \ell_m(\theta) - \nabla^2 \ell_m(\theta_\star) \right\| + \frac{M^2}{2} \left\| \nabla^2 \ell_m(\theta_\star) - \nabla^2 \ell(\theta_\star) \right\|.$$

wrejchel@gmail.com

The first term above goes to zero in probability by Assumption 9. The second term also goes to zero because

$$
\begin{aligned}
\nabla^2 \ell_m(\theta_\star) - \nabla^2 \ell(\theta_\star) &= \nabla^2 \log c(\theta_\star) - \nabla^2 \log \hat{c}_m(\theta_\star) \\
&= \frac{\nabla^2 c(\theta_\star)}{c(\theta_\star)} - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} \frac{\nabla c(\theta_\star)^\top}{c(\theta_\star)} - \frac{\nabla^2 \hat{c}_m(\theta_\star)}{\hat{c}_m(\theta_\star)} + \frac{\nabla \hat{c}_m(\theta_\star)}{\hat{c}_m(\theta_\star)} \frac{\nabla \hat{c}_m(\theta_\star)^\top}{\hat{c}_m(\theta_\star)} \xrightarrow{\text{p}} 0,
\end{aligned}
$$

in view of (2.7). Therefore (2.9) holds and the proof is complete.

2.3 Optimal Importance Distribution

We advocate adaptation to improve the choice of instrumental distribution h. But which h is the best? If we use (non-adaptive) importance sampling with instrumental distribution h then the maximizer $\hat{\theta}_m$ of the MC likelihood approximation has asymptotic normal distribution, namely $\sqrt{m}\left(\hat{\theta}_m - \theta_\star\right) \xrightarrow{\text{d}} \mathcal{N}(0, D^{-1} V D^{-1}), (m \to \infty)$ with

$$
V = \frac{1}{c(\theta_\star)^2} \mathrm{VAR}_{Y \sim h} \left[\frac{\nabla f_{\theta_\star}(Y)}{h(Y)} - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} \frac{f_{\theta_\star}(Y)}{h(Y)} \right].
$$

This fact is well-known [4] and is a special case of Theorem 3. Since the asymptotic distribution is multidimensional its dispersion can be measured in various ways, e.g., though the determinant, the maximum eigenvalue or the trace of the covariance matrix. We examine the trace which equals to the asymptotic mean square error of the MCML approximation (the asymptotic bias is nil). Notice that

$$
c(\theta_\star)^2 V = \mathrm{VAR}_{Y \sim h} \frac{\eta(Y)}{h(Y)} = \mathbb{E}_{Y \sim h} \frac{\eta(Y)\eta(Y)^\top}{h(Y)^2} = \int \frac{\eta(y)\eta(y)^\top}{h(y)} \mathrm{d}y,
$$

where

$$
\eta(y) = \nabla f_{\theta_\star}(y) - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} f_{\theta_\star}(y).
$$

Since $\mathrm{tr}\left[D^{-1}\eta(y)\eta(y)^\top D^{-1} \right] = (D^{-1}\eta(y))^\top D^{-1}\eta(y) = |D^{-1}\eta(y)|^2$, the minimization of $\mathrm{tr}(D^{-1} V D^{-1})$ is equivalent to

$$
\int \frac{|D^{-1}\eta(y)|^2}{h(y)} \mathrm{d}y \to \min,
$$

subject to $h \geq 0$ and $\int h = 1$. By Schwarz inequality we have

$$\int \frac{|D^{-1}\eta(y)|^2}{h(y)}\mathrm{d}y = \int \left(\frac{|D^{-1}\eta(y)|}{\sqrt{h(y)}} \right)^2 \mathrm{d}y \int \left(\sqrt{h(y)} \right)^2 \mathrm{d}y$$

$$\geq \left(\int \frac{|D^{-1}\eta(y)|}{\sqrt{h(y)}} \sqrt{h(y)}\mathrm{d}y \right)^2 = \left(\int |D^{-1}\eta(y)|\mathrm{d}y \right)^2,$$

with equality only for $h(y) \propto |D^{-1}\eta(y)|$. The optimum choice of h is therefore

$$h_\star(y) \propto \left| D^{-1}\left(\nabla f_{\theta_\star}(y) - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} f_{\theta_\star}(y) \right) \right|. \qquad (2.11)$$

Unfortunately, this optimality result is chiefly of theoretical importance, because it is not clear how to sample from $h_\star$ and how to compute the norming constant for this distribution. This might well be even more difficult than computing $c(\theta)$.

The following example shows some intuitive meaning of (2.11). It is a purely "toy example" because the simple analitical formulas exist for $c(\theta)$ and $\theta_\star$ while MC is considered only for illustration.

Example 1 Consider a binomial distribution on $\mathcal{Y} = \{0, 1, \ldots, n\}$ given by $\pi_\theta(y) = \binom{n}{y}p^y(1-p)^{n-y}$. Parametrize the model with the log-odds-ratio $\theta = \log p/(1-p)$, absorb the $\binom{n}{y}$ factor into the measure $\mathrm{d}y$ to get the standard exponential family form with

$$f_\theta(y) = \mathrm{e}^{\theta y} \quad \text{and} \quad c(\theta) = \sum_{y=0}^{n} \binom{n}{y}\mathrm{e}^{\theta y} = (1 + \mathrm{e}^\theta)^n.$$

Taking into account the facts that $\nabla c(\theta_\star)/c(\theta_\star) = y_{\mathrm{obs}}$ and $\nabla f_\theta(y) = y\mathrm{e}^{\theta y}$ we obtain that (2.11) becomes $h_\star(y) \propto |y - y_{\mathrm{obs}}|\mathrm{e}^{\theta y}$ (factor D^{-1} is a scalar so can be omitted). In other words, the optimum instrumental distribution for AIS MCML, expressed in terms of $p = \mathrm{e}^\theta/(1 + \mathrm{e}^\theta)$ is

$$\mathbb{P}_{Y \sim h_\star}(Y = y) \propto \binom{n}{y}|y - y_{\mathrm{obs}}|p^y(1 - p)^{n-y}.$$

3 Generalized Adaptive Scheme

Importance sampling, even in its adaptive version (AIS), suffers from the degeneracy of weights. To compute the importance weights $f_\theta(Y_m)/h_{\psi_m}(Y_m)$ we have to know norming constants for every h_{ψ_m} (or at least their ratios). This requirement severly restricts our choice of the family of instrumental densities h_ψ. Available instrumental densities are far from $h_\star$ and far from $f_\theta/c(\theta)$. Consequently the weights tend to

degenerate (most of them are practically zero, while a few are very large). This effectively makes AIS in its basic form impractical. To obtain practically applicable algorithms, we can generalize AIS as follows. In the same situation as in Sect. 2, instead of the AIS estimate given by (2.3), we consider a more general Monte Carlo estimate of $c(\theta)$ of the form

$$\hat{c}_m(\theta) = \frac{1}{m} \sum_{j=1}^{m} \hat{d}(\theta, \psi_j),\qquad\qquad(3.1)$$

where the summands $\hat{d}(\theta, \psi_j)$ are computed by an MC method to be specified later. For now let us just assume that this method depends on a control parameter ψ which may change at each step. A general adaptive algorithm is the following:

<div align="center">Algorithm AdapMCML</div>

1. Set an initial value of $\psi_1, m = 1, \hat{c}_0(\theta) \equiv 0$.
2. Compute an "incremental estimate" $\hat{d}(\theta, \psi_m)$.
3. Update the approximation of $\hat{c}_m(\theta)$:

$$\hat{c}_m(\theta) = \frac{m-1}{m}\hat{c}_{m-1}(\theta) + \frac{1}{m}\hat{d}(\theta, \psi_m).$$

4. Update ψ: choose ψ_{m+1} based on the history of the simulation.
5. $m = m + 1$; go to 2.

AdapIS in Sect. 2 is a special case of AdapMCML which is obtained by letting $\hat{d}(\theta, \psi_m) = f_\theta(Y_m)/h_{\psi_m}(Y_m)$.

3.1 Variance Reduction Via Resampling and MCMC

The key property of the AIS exploited in Sect. 2 is the martingale structure implied by (2.4) and (2.6). The main asymptotic results generalize if *given ψ, the estimates of $c(\theta)$ and its derivatives are conditionally unbiased*. We propose an algorithm for computing $\hat{d}$ in (3.1) which has the unbiasedness property and is more efficient than simple AIS. To some extent it is a remedy for the problem of weight degeneracy and reduces the variance of Monte Carlo approximations. As before, consider a family of "instrumental densities" h_ψ. Assume they are properly normalized ($\int h_\psi = 1$) and the control parameter ψ belongs the same space as the parameter of interest θ ($\Psi = \Theta$). Further assume that for every ψ we have at our disposal a Markov kernel P_ψ on $\mathcal{Y}$ which preserves distribution $\pi_\psi = f_\psi/c(\psi)$, i.e. $f_\psi(y)\mathrm{d}y = \int f_\psi(y')P_\psi(y', \mathrm{d}y)\mathrm{d}y'$. Let us fix ψ. This is a setup in which we can apply the following importance sampling-resampling algorithm ISReMC:

<div align="center">Algorithm ISReMC</div>

1. Sample $Y_1, \ldots, Y_l \sim h_\psi$.
2. Compute the importance weights $W_i = w(Y_i) = \dfrac{f_\psi(Y_i)}{h_\psi(Y_i)}$ and put
   $W_\bullet = \sum_{i=1}^l W_i$.
3. Sample $Y_1^\star, \ldots, Y_r^\star \sim \sum_{i=1}^l \delta_{Y_i}(\cdot) W_i / W_\bullet$ [Discrete distribution with mass $W_i / W_\bullet$ at point Y_i].
4. For $k = 1, \ldots, r$ generate a Markov chain trajectory, starting from $Y_k^\star$ and using kernel P_ψ:

$$Y_k^\star = Y_k^0, Y_k^1, \ldots, Y_k^s, Y_k^{s+1}, \ldots, Y_k^{s+n}.$$

Compute $\hat{d}(\theta, \psi)$ given by

$$\hat{d}(\theta, \psi) = \frac{W_\bullet}{l} \frac{1}{r} \sum_{k=1}^{r} \frac{1}{n} \sum_{u=s+1}^{s+n} \frac{f_\theta(Y_k^u)}{f_\psi(Y_k^u)}. \tag{3.2}$$

This algorithm combines the idea of resampling (borrowed from sequential MC; steps 2 and 3) with computing ergodic averages in multistart MCMC (step 4; notice that s is a burn-in and n is the actual used sample size for a single MCMC run, repeated r times). More details about ISReMC are in [7]. In our context it is sufficient to note the following key property of this algorithm.

Lemma 1 *If $\hat{d}(\theta, \psi)$ is the output of ISReMC then for every θ and every ψ,*

$$\mathbb{E}\hat{d}(\theta, \psi) = c(\theta).$$

If Assumption 3 holds then also

$$\mathbb{E}\nabla\hat{d}(\theta, \psi) = \nabla c(\theta), \qquad \mathbb{E}\nabla^2\hat{d}(\theta, \psi) = \nabla^2 c(\theta).$$

Proof We can express function $c(\theta)$ and its derivatives as "unnormalized expectations" with respect to the probability distribution with density $\pi_\psi = f_\psi / c(\psi)$:

$$c(\theta) = \mathbb{E}_{Y \sim \pi_\psi} \frac{f_\theta(Y)}{f_\psi(Y)} c(\psi),$$

$$\nabla c(\theta) = \mathbb{E}_{Y \sim \pi_\psi} \frac{\nabla f_\theta(Y)}{f_\psi(Y)} c(\psi), \qquad \nabla^2 c(\theta) = \mathbb{E}_{Y \sim \pi_\psi} \frac{\nabla^2 f_\theta(Y)}{f_\psi(Y)} c(\psi).$$

Let us focus on $\mathbb{E}\hat{d}(\theta, \psi)$. Write

$$a(y) = \mathbb{E}\left(\frac{1}{n} \sum_{u=s+1}^{s+n} \frac{f_\theta(Y^u)}{f_\psi(Y^u)} \middle| Y^0 = y\right) \tag{3.3}$$

for the expectation of a *single* MCMC estimate started at $Y^0 = y$. Kernel P_ψ preserves π_ψ by assumption, therefore $\mathbb{E}_{Y \sim \pi_\psi} a(Y) = \mathbb{E}_{Y \sim \pi_\psi} f_\theta(Y)/f_\psi(Y) = c(\theta)/c(\psi)$. Put differently, $\int a(y) f_\psi(y) \mathrm{d}y = c(\theta)$.

We make a simple observation that

$$\mathbb{E}\left(\hat{d}(\theta, \psi) \middle| Y_1, \ldots, Y_l, Y_1^\star, \ldots, Y_r^\star\right) = \frac{W_\bullet}{l} \frac{1}{r} \sum_{k=1}^{r} a(Y_k^\star).$$

This conditional expectation takes into account only randomness of the MCMC estimate in step 4 of the algorithm. Now we consecutively "drop the conditions":

$$\mathbb{E}\left(\hat{d}(\theta, \psi) \middle| Y_1, \ldots, Y_l\right) = \frac{W_\bullet}{l} \sum_{i=1}^{l} a(Y_i) \frac{W_i}{W_\bullet} = \frac{1}{l} \sum_{i=1}^{l} a(Y_i) W_i.$$

The expectation above takes into account the randomness of the resampling in step 3. Finally, since $Y_i \sim h_\psi$ in step 1, we have

$$\mathbb{E}\hat{d}(\theta, \psi) = \mathbb{E}a(Y_i) W_i = \mathbb{E}_{Y \sim h_\psi} a(Y) \frac{f_\psi(Y)}{h_\psi(Y)}$$

$$= \int a(y) f_\psi(y) \mathrm{d}y = c(\theta).$$

This ends the proof for $\hat{d}$. Exactly the same argument applies to $\nabla \hat{d}$ and $\nabla^2 \hat{d}$.

We can embed the unbiased estimators produced by `ISReMC` in our general adaptive scheme `AdapMCML`. At each step m of the adaptive algorithm, we have a new control parameter ψ_m. We generate a sample from h_{ψ_m}, compute weights, resample and run MCMC using ψ_m. Note that the whole sampling scheme at stage m (including computation of weights) depends on ψ_m but not on θ. In the adaptive algorithm random variable ψ_{m+1} is $\mathcal{F}_m$ measurable, where $\mathcal{F}_m$ is the history of simulation up to stage m. Therefore the sequence of incremental estimates $\hat{d}(\theta, \psi_m)$ satisfies, for every $\theta \in \Theta$,

$$\mathbb{E}(\hat{d}(\theta, \psi_{m+1}) | \mathcal{F}_m) = c(\theta). \tag{3.4}$$

Moreover, first and second derivatives exist and

$$\mathbb{E}(\nabla \hat{d}(\theta, \psi_{m+1})|\mathcal{F}_m) = \nabla c(\theta), \qquad \mathbb{E}(\nabla^2 \hat{d}(\theta, \psi_{m+1})|\mathcal{F}_m) = \nabla^2 c(\theta). \qquad (3.5)$$

Formulas (3.4) and (3.5) are analogues of (2.4) and (2.6).

3.2 Asymptotics of Adaptive MCML

In this subsection we restrict our considerations to *exponential families on finite spaces*. This will allow us to prove main results without formulating complicated technical assumptions (integrability conditions analoguous to Assumption 8 would be cumbersome and difficult to verify). Some models with important applications, such as *autologistic* one, satisfy the assumptions below.

Assumption 10 Let

$$f_\theta(y) = \exp[\theta^\top t(y)],$$

where $t(y) \in \mathbb{R}^d$ is the vector of sufficient statistics and $\theta \in \Theta = \mathbb{R}^d$. Assume that y belongs to a finite space $\mathcal{Y}$.

Now, since $\mathcal{Y}$ is finite (although possibly very large),

$$c(\theta) = \sum_y \exp[\theta^\top t(y)].$$

Note that Assumption 3 is automatically satisfied.

Assumption 11 Control parameters ψ belong to a compact set $\Psi \subset \mathbb{R}^d$.

We consider algorithm `AdapMCML` with incremental estimates $\hat{d}$ produced by `ISReMC`. The likelihood ratio in (3.2) and its derivatives assume the following form:

$$\begin{aligned}
\frac{f_\theta(Y)}{f_\psi(Y)} &= \exp[(\theta - \psi)^\top t(Y)], \\
\frac{\nabla f_\theta(Y)}{f_\psi(Y)} &= t(Y) \exp[(\theta - \psi)^\top t(Y)], \\
\frac{\nabla^2 f_\theta(Y)}{f_\psi(Y)} &= t(Y)t(Y)^\top \exp[(\theta - \psi)^\top t(Y)]
\end{aligned} \qquad (3.6)$$

(the derivatives are with respect to θ, with ψ fixed). Assumptions 11 and 10 together with Assumption 6 imply that $\hat{d}(\theta, \psi_j)$ are uniformly bounded, if θ belongs to a compact set. Indeed, the importance weights W_i in (3.2) are uniformly bounded

by Assumptions 6 and 11. Formula (3.6) shows that the ratios $f_\theta(y)/f_{\psi_j}(y) = \exp[(\theta - \psi_j)^\top t(y)]$ are also uniformly bounded for ψ_j and θ belonging to bounded sets. Since the statistics $t(y)$ are bounded, the same argument shows that $\nabla \hat{d}(\theta, \psi_j)$ and $\nabla^2 \hat{d}(\theta, \psi_j)$ are uniformly bounded, too.

For exponential families, $\log c(\theta)$ and $\log \hat{c}_m(\theta)$ are convex functions. It is a well known property of exponential family that $\nabla^2 \log c(\theta) = \text{VAR}_{Y \sim \pi_\theta} t(Y)$ and thus it is a nonnegative definite matrix. A closer look at $\hat{c}_m(\theta)$ reveals that $\nabla^2 \log \hat{c}_m(\theta)$ is also a variance-covariance matrix with respect to some discrete distribution. Indeed, it is enough to note that $\hat{c}_m(\theta)$ is of the form

$$\hat{c}_m(\theta) = \sum_{j,k,u} \exp[\theta^\top t_{j,k,u}] a_{j,k,u},$$

for some $t_{j,k,u} \in \mathbb{R}^d$ and $a_{j,k,u} > 0$ (although if `ISReMC` within `AdapMCML` is used to produce $\hat{c}_m(\theta)$ then $t_{j,k,u}$ and $a_{j,k,u}$ are quite complicated random variables depending on ψ_j).

Let $\hat{\theta}_m$ be a maximizer of $\ell_m(\theta) = \theta^\top t(y_{\text{obs}}) - \log \hat{c}_m(\theta)$ and assume that $\theta_\star$ is the unique maximizer of $\ell(\theta) = \theta^\top t(y_{\text{obs}}) - \log c(\theta)$.

Proposition 3 *If Assumptions 6, 10 and 11 hold, then $\hat{\theta}_m \to \theta_\star$ almost surely.*

Proof Boundedness of $\hat{d}(\theta, \psi_m)$ for a fixed θ together with (3.4) implies that $\hat{d}(\theta, \psi_m) - c(\theta)$ is a bounded sequence of martingale differences. It satisfies the assumptions of SLLN for martingales in Appendix A. Therefore $\hat{c}_m(\theta) \to c(\theta)$. Consequently, we also have $\ell_m(\theta) \to \ell(\theta)$, pointwise. Pointwise convergence of convex functions implies uniform convergence on compact sets [13, Theorem 10.8]. The conclusion follows immediately.

Theorem 4 *If Assumptions 5–7, 10 and 11 hold, then*

$$\sqrt{m}(\hat{\theta}_m - \theta_\star) \to \mathcal{N}(0, D^{-1} V D^{-1}) \text{ in distribution,}$$

where $D = \nabla^2 \ell(\theta_\star)$ and

$$V = \frac{1}{c(\theta_\star)^2} \text{VAR} \left[\nabla \hat{d}(\theta_\star, \psi_\star) - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} \hat{d}(\theta_\star, \psi_\star) \right],$$

where $\hat{d}(\theta_\star, \psi_\star)$ is a result of the IS/Resampling algorithm `ISReMC`, described in the previous subsection, with $\psi = \psi_\star$ and $\theta = \theta_\star$.

Note that $\hat{d}(\theta_\star, \psi_\star)$ is a purely imaginary object, being a result of an algorithm initialized at a "limiting instrumental parameter" $\psi_\star$ and evaluated at the "true MLE" $\theta_\star$, both unknown. It is introduced only to concisely describe the variance/covariance matrix V. Note also that $\nabla c(\theta_\star)/c(\theta_\star)$ is equal to $t(y_{\text{obs}})$, the *observed* value of the sufficient statistic.

Proof (of Theorem 4) The proof is similar to that of Theorem 3, so we will not repeat all the details. The key argument is again based on SLLN and CLT for martingales (see Appendix A). In the present situation we have more complicated estimators $\hat{d}(\theta, \psi_j)$ than in Theorem 3. They are now given by (3.2). On the other hand, we work under the assumption that f_θ is an exponential family on a finite state space $\mathcal{Y}$. This implies that conditions (3.4) and (3.5) are fulfilled and the martingale differences therein are uniformly bounded (for any fixed θ and also for θ running through a compact set). Concavity of $\ell_m(\theta)$ and $\ell(\theta)$ further simplifies the argumentation.

As in the proof of Theorem 3, we claim that (2.8) and (2.9) hold. The first of these conditions, (2.8), is justified exactly in the same way: by applying the CLT to the numerator and SLLN to the denominator of (2.10). Now, we consider martingale differences given by

$$\xi_j = \nabla \hat{d}(\theta_\star, \psi_j) - \frac{\nabla c(\theta_\star)}{c(\theta_\star)} \hat{d}(\theta_\star, \psi_j).$$

It follows from the discussion preceding the theorem that ξ_j are uniformly bounded, so the CLT can be applied. Similarly, SLLN can be applied to $\hat{d}(\theta_\star, \psi_j) - c(\theta_\star)$.

Assumption (9) holds because third order derivatives of $\log \hat{c}_m(\theta)$ are uniformly bounded in the neighbourhood of $\theta_\star$. This allows us to infer condition (2.9) in the same way as in the proof of Theorem 3.

Note also that we do not need Assumption 4. To deduce the conclusion of the theorem from (2.9) we have to know that $\hat{\theta}_m$ is square-root consistent. But this follows from the facts that $\ell_m(\theta)$ is concave and the maximizer of the quadratic function $-(\theta - \theta_\star)^\top \nabla \ell_m(\theta_\star) - \frac{1}{2}(\theta - \theta_\star)^\top D(\theta - \theta_\star)$ in (2.9) is square-root consistent by (2.8).

4 Simulation Results

In a series of small scale simulation experiments, we compare two algorithms. The first one, used as a "Benchmark" is a non-adaptive MCML. The other is `AdapMCML` which uses `ISReMC` estimators, as described in Sect. 3. Synthetic data used in our study are generated from autologistic model, described below. Both algorithms use Gibbs Sampler (GS) as an MCMC subroutine and both use Newton-Raphson iterations to maximize MC log-likelihood approximations.

4.1 Non-adaptive and Adaptive Newton-Raphson-Type Algorithms

Well-known Newton-Raphson (NR) method in our context updates points θ_m approximating maximum of the log-likelihood as follows:

$$\theta_{m+1} = \theta_m + \nabla^2 \ell_m(\theta_m)^{-1} \nabla \ell_m(\theta_m),$$

where ℓ_m is given by (2.1).

Non-adaptive algorithms are obtained when some fixed value of the "instrumental parameter" is used to produce MC samples. Below we recall a basic version of such an algorithm, proposed be Geyer [4] and examined e.g. in [9]. If we consider an exponenial family given by Assumption 10, then $\ell_m(\theta) = \theta^\top t(y_{\text{obs}}) - \log \hat{c}_m(\theta)$. Let ψ be fixed and $Y_0, Y_1, \ldots, Y_s, \ldots, Y_{s+m}$ be samples approximately drawn from distribution $\pi_\psi \propto f_\psi$. In practice an MCMC method is applied to produce such samples, s stands for a burn-in. In all our experiments the MCMC method is a deterministic scan Gibbs Sampler (GS). Now, we let

$$\hat{c}_m(\theta) \propto \frac{1}{m} \sum_{u=s+1}^{s+m} \exp[(\theta - \psi)^\top t(Y_u)].$$

Consequently, if $\omega_u(\theta) = \exp[(\theta - \psi)^\top t(Y_u)]$ and $\omega_\bullet(\theta) = \sum_{u=s+1}^{s+m} \omega_u(\theta)$, then the derivatives of the log-likelihood are expressed via weighted moments,

$$\nabla \ell_m(\theta) = t(y_{\text{obs}}) - \overline{t(Y)}, \quad \overline{t(Y)} = \frac{1}{\omega_\bullet(\theta)} \sum_{u=s+1}^{s+m} \omega_u(\theta) t(Y_u),$$

$$\nabla^2 \ell_m(\theta) = -\frac{1}{\omega_\bullet(\theta)} \sum_{u=s+1}^{s+m} \omega_u(\theta)(t(Y_u - \overline{t(Y)})(t(Y_u) - \overline{t(Y)})^\top.$$

The adaptive algorithm uses $\hat{c}_m(\theta)$ given by (3.1), with summands $\hat{d}(\theta, \psi_j)$ computed by `ISReMC`, exactly as described in Sect. 3. The MCMC method imbedded in `ISReMC` is GS, the same as in the non-adaptive algorithm. Importance sampling distribution h_ψ in steps 1 and 2 of `ISReMC` is pseudo-likelihood, described by formula (4.1) in the next subsection. Computation of ψ_{m+1} in step 4 of `AdapMCML` uses one NR iteration: $\psi_{m+1} = \psi_m + \nabla^2 \ell_m(\psi_m)^{-1} \nabla \ell_m(\psi_m)$, where ℓ_m is given by (2.1) with $\hat{c}_m$ produced by `AdapMCML`.

4.2 Methodology of Simulations

For our experiments we have chosen the autologistic model, one of chief motivating examples for MCML. It is given by a probability distribution on $\mathcal{Y} = \{0, 1\}^{d \times d}$ proportional to

$$f_\theta(y) = \exp\left(\theta_0 \sum_r y^{(r)} + \theta_1 \sum_{r \sim s} y^{(r)} y^{(s)}\right),$$

Table 1 Sufficient statistics, maximum log-likelihood and MPL estimate for example with $d = 10$

Statistic T	ML $\theta_\star$	Log-Lik $\ell(\theta_\star)$	MPL $\hat{\theta}$
$(59, 74)$	$(-1.21, 0.75)$	-15.889991	$(-1.07, 0.66)$

Table 2 Sufficient statistics, maximum log-likelihood and MPL estimate for example with $d = 15$

Statistic T	ML $\theta_\star$	Log-Lik $\ell(\theta_\star)$	MPL $\hat{\theta}$
$(142, 180)$	$(-0.46, 0.43)$	12.080011	$(-0.57, 0.54)$

where $r \sim s$ means that two points r and s in the $d \times d$ lattice are neighbours. The pseudo-likelihood h_ψ is given by

$$h_\psi(y) \propto \prod_r \exp\left(\theta_0 y^{(r)} + \theta_1 \sum_{s:r\sim s} y^{(r)} y^{(s)}_{\text{obs}}\right). \tag{4.1}$$

In our study we considered lattices of dimension $d = 10$ and $d = 15$. The values of sufficient statistics $T = \left(\sum_r y^{(r)}, \sum_{r\sim s} y^{(r)} y^{(s)}\right)$, exact ML estimators $\theta_\star$ and maxima of the log-likelihoods are in the Tables 1 and 2. We report results of several repeated runs of a "benchmark" non-adaptive algorithm and our adaptive algorithm. The initial points are (1) the maximum pseudo-likelihood (MPL) estimate, denoted by $\hat{\theta}$ (also included in the tables) and (2) point $(0, 0)$. Number of runs is 100 for $d = 10$ and 25 for $d = 15$. Below we describe the parameters and results of these simulations. Note that we have chosen parameters for both algorithms in such a way which allows for a "fair comparison", that is the amount of computations and number of required samples are similar for the benchmark and adaptive algorithms.

For $d = 10$: In benchmark MCML, we used 1000 burn-in and 39,000 collected realisations of the Gibbs sampler; then 20 iterations of Newton-Raphson were applied. `AdapMCML` had 20 iterations; parameters within `ISReMC` were $l = 1000$, $r = 1$, $s = 100$, $n = 900$.

The results are shown in Figs. 1 and 2.

For $d = 15$: In benchmark MCML, we used 10,000 burn-in and 290,000 collected realisations of the Gibbs sampler; then 10 iterations of Newton-Raphson were applied. `AdapMCML` had 10 iterations; parameters within `ISReMC` were $l = 10,000$, $r = 1$, $s = 1000$, $n = 19,000$.

The results are shown in Figs. 3 and 4. The benchmark algorithm started from 0 for $d = 15$ failed, so only the results for the adaptive algorithm are given in the right parts of Figs. 3 and 4.
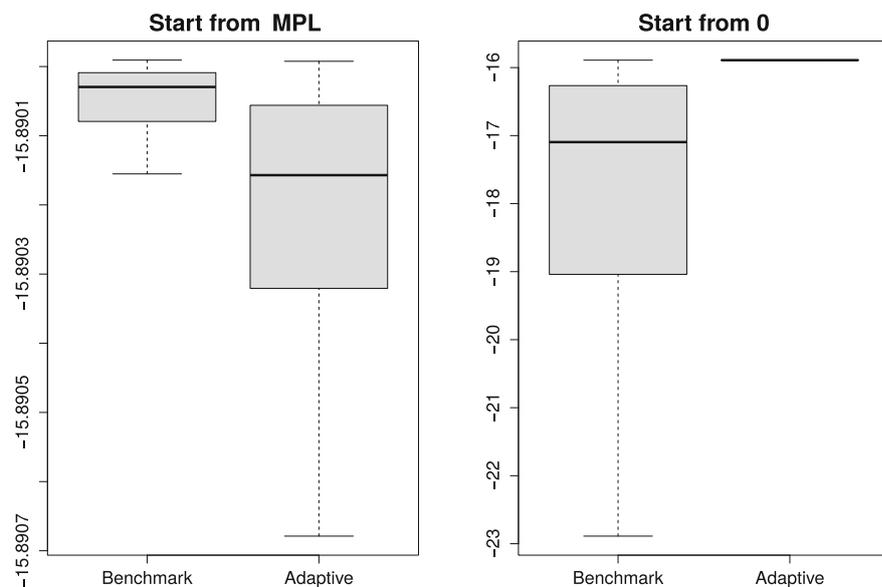
wrejchel@gmail.com

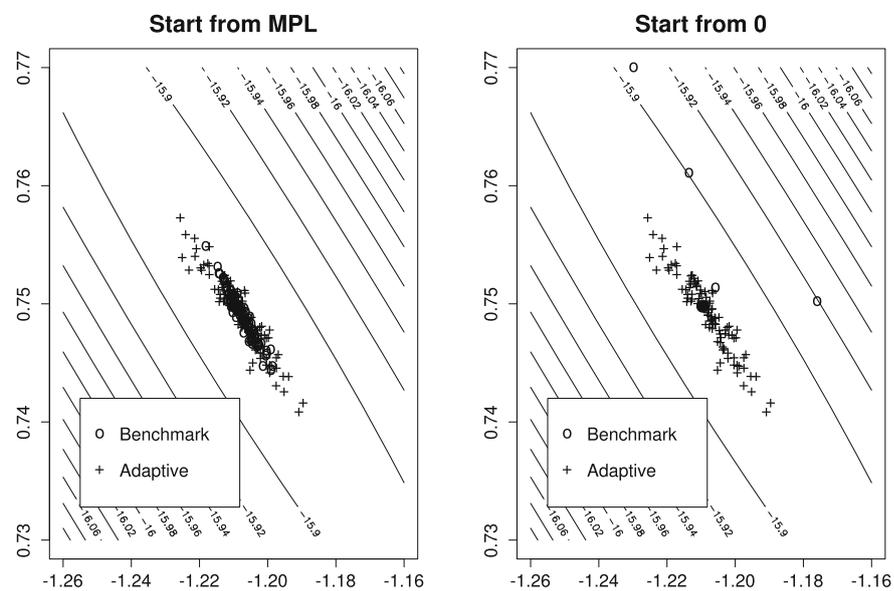**Fig. 1** Log-likelihood at the output of MCML algorithms; $d = 10$; 100 runs

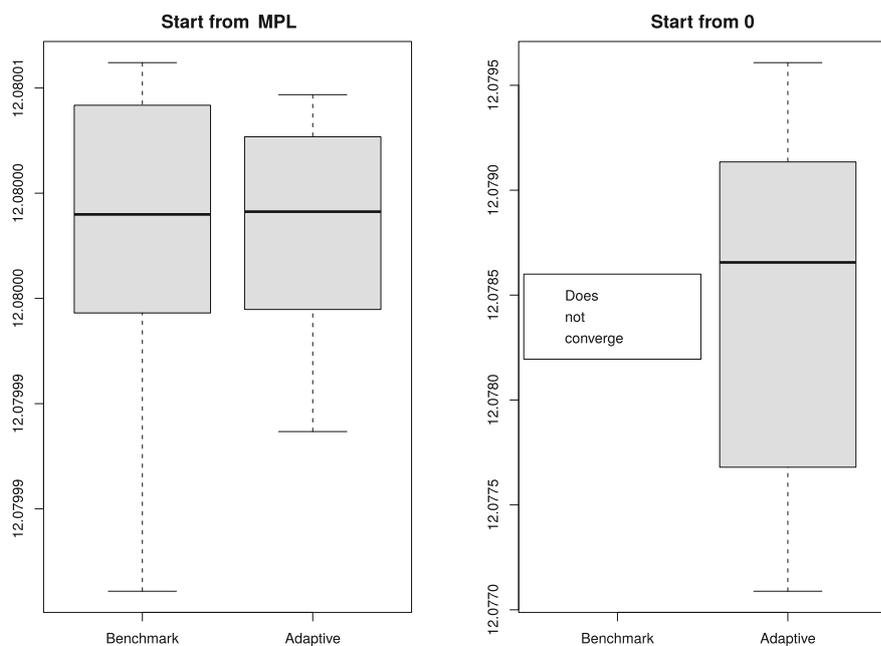**Fig. 2** Output of MCML algorithms; $d = 10$; 100 repetitions

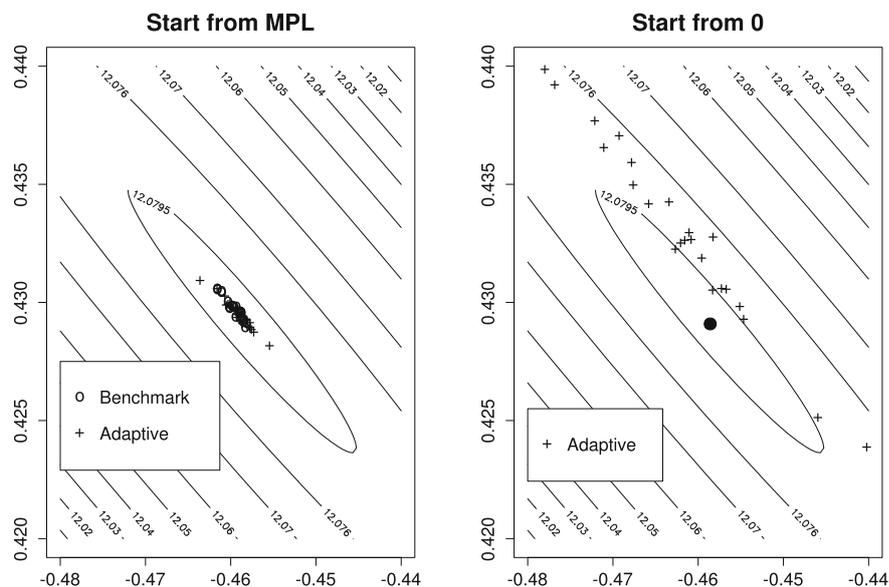**Fig. 3** Log-likelihood at the output of MCML algorithms; $d = 15$; 25 repetitions

Fig. 4 Output of MCML algorithms; $d = 15$; 25 repetitions

4.3 Conclusions

The results of our simulations allow to draw only some preliminary conclusions, because the range of experiments was limited. However, some general conclusions can be rather safely formulated. The performance of the benchmark, non-adaptive algorithm crucially depends on the choice of starting point. It yields quite satisfactory results, if started sufficiently close tho the maximum likelihood, for example from the maximum pseudo-likelihood estimate. Our adaptive algorithm is much more robust and stable in this respect. If started from a good initial point, it may give slightly worse results than the benchmark, but still is satisfactory (see Fig. 2). However, when the maximum pseudo-likelihood estimate is not that close to the maximum likelihood point, the adaptive algorithm yields an estimate with a lower variance (see Fig. 2). When started at a point distant from the maximum likelihood, such as 0, it works much better than a non-adaptive algorithm. Thus the algorithm proposed in our paper can be considered as more universal and robust alternative to a standard MCML estimator.

Finally let us remark that there are several possibilities of improving our adaptive algorithm. Some heuristically justified modifications seem to converge faster and be more stable than the basic version which we described. Modifications can exploit the idea of resampling in a different way and reweigh past samples in subsequent steps. Algorithms based on stochastic approximation, for example such as that proposed in [16], can probably be improved by using Newton-Raphson method instead of simple gradient descent. However, theoretical analysis of such modified algorithms becomes more difficult and rigorous theorems about them are not available yet. This is why we decided not to include these modified algorithms in this paper. Further research is needed to bridge a gap between practice and theory of MCML.

Appendix A: Martingale Limit Theorems

For completeness, we cite the following martingale central limit theorem (CLT):

Theorem A.1 ([8, Theorem 2.5]) *Let* $X_n = \xi_1 + \cdots + \xi_n$ *be a mean-zero (vector valued) martingale. If there exists a symmetric positive definite matrix V such that*

$$\frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\left(\xi_j \xi_j^T | \mathcal{F}_{j-1}\right) \xrightarrow{\text{p}} V, \tag{A.1}$$

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\Big(\xi_j\xi_j^T\mathbf{1}_{|\xi_j|>\varepsilon\sqrt{n}}\,|\mathcal{F}_{j-1}\Big)\xrightarrow{\mathrm{p}}0 \quad \textit{for each } \varepsilon>0, \qquad (\text{A.2})$$

then

$$\frac{X_n}{\sqrt{n}}\xrightarrow{\mathrm{d}}\mathcal{N}(0,V).$$

The Lindeberg condition (A.2) can be replaced by a stronger Lyapunov condition

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\Big(|\xi_j|^{2+\alpha}|\mathcal{F}_{j-1}\Big)\leq M \quad \text{for some } \alpha>0 \text{ and } M<\infty. \qquad (\text{A.3})$$

A simple consequence of [6, Theorem 2.18] (see also [3]) is the following strong law of large numbers (SLLN).

Theorem A.2 *Let $X_n=\xi_1+\cdots+\xi_n$ be a mean-zero martingale. If*

$$\sup_j \mathbb{E}\Big(|\xi_j|^{1+\alpha}|\mathcal{F}_{j-1}\Big)\leq M \quad \textit{for some } \alpha>0 \textit{ and } M<\infty$$

then

$$\frac{X_n}{n}\xrightarrow{\mathrm{a.s.}}0.$$

References

1. Attouch H (1984) Variational convergence of functions and operators. Pitman, Boston
2. Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc B 36:192–236
3. Chow YS (1967) On a strong law of large numbers for martingales. Ann Math Stat 38:610
4. Geyer CJ (1994) On the convergence of Monte Carlo maximum likelihood calculations. J R Stat Soc B 56:261–274
5. Geyer CJ, Thompson EA (1992) Constrained Monte Carlo maximum likelihood for dependent data. J R Stat Soc B 54:657–699
6. Hall P, Heyde CC (1980) Martinagale limit theory and its application. Academic Press, New York
7. Miasojedow B, Niemiro W (2014) Debiasing MCMC via importance sampling-resampling. In preparation
8. Helland IS (1982) Central limit theorems for martingales with discrete or continuous time. Scand J Stat 9:79–94
9. Huffer FW, Wu H (1998) Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species. Biometrics 54:509–524
10. Møller BJ, Pettitt AN, Reeves R, Berthelsen KK (2006) An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. Biometrika 93:451–458
11. Pettitt AN, Weir IS, Hart AG (2002) A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. Stat Comput 12:353–367

12. Pollard D (1984) Convergence of stochastic processes. Springer, New York
13. Rockafellar RT (1970) Convex analysis. Princeton University Press, Princeton
14. Rockafellar TJ, Wets RJ-B (2009) Variational analysis, 3rd edn. Springer, New York
15. Wu H, Huffer FW (1997) Modeling the distribution of plant species using the autologistic regression model. Environ Ecol Stat 4:49–64
16. Younes L (1988) Estimation and annealing for Gibbsian fields. Annales de l'I H P sec B 24(2):269–294
17. Zalewska M, Niemiro W, Samoliński B (2010) MCMC imputation in autologistic model. Monte Carlo Methods Appl 16:421–438