Asymptotics of maximum likelihood estimators based on Markov chain Monte Carlo methods

Błażej Miasojedow *, Wojciech Niemiro † and Wojciech Rejchel ‡

Abstract

In complex statistical models, in which exact computation of the likelihood is intractable, Monte Carlo methods can be applied to approximate maximum likelihood estimates. In this paper we consider approximation obtained via Markov chain Monte Carlo. We prove consistency and asymptotic normality of the resulting estimator, when both sample sizes (the initial and Monte Carlo one) tend to infinity. Our results can be applied to models with intractable normalizing constants and missing data models. We also investigate properties of estimators in numerical experiments.

Keywords: intractable normalizing constant, Markov chain, maximum likelihood estimation, missing data model, Monte Carlo method.

*Institute of Applied Mathematics and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland, B.Miasojedow@mimuw.edu.pl
†Institute of Applied Mathematics and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland, wniem@mimuw.edu.pl and Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Chopina 12/18, 87-100, Toruń, Poland wniem@mat.umk.torun.pl
‡corresponding author, Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Chopina 12/18, 87-100, Toruń, Poland, +48 566112943, 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 ML estimators is very difficult or impossible. Such problems arise in missing data models or if considered densities are known only up to intractable normalizing constants, for instance in Markov random fields or spatial statistics. In missing data models many Monte Carlo (MC) or Markov chain Monte Carlo (MCMC) methods have been proposed to approximate the observed likelihood [8, 10, 13]. There are also Monte Carlo methods for maximum likelihood that do not approximate the likelihood [19, 24, 26] and non-Monte Carlo methods dedicated to this problem [5]. The literature concerning the problem of the intractable normalizing constant is extensive as well. Among the proposed methods we should mention the maximum pseudolikelihood [2], Monte Carlo EM [7, 14], the Monte Carlo maximum likelihood (MCML) [4, 10] and the Markov chain Monte Carlo maximum likelihood (MCMCML) [9, 10]. In the current paper we focus on the MCMCML method.

In influential papers [9, 10] the authors prove consistency and asymptotic normality of MCML estimators under the assumption that the initial sample is fixed, and only the Monte Carlo sample size tends to infinity. Both sources of randomness (one due to the initial sample and the other due to Monte Carlo simulations) are considered in [4, 16, 23, 27]. The authors of the first mentioned paper apply the general importance sampling recipe. They show that for their scheme of simulations, the MC sample size has to grow exponentially fast with the initial sample to ensure consistency of the estimator. As the remedy for this problem they propose to use a preliminary estimator, which is consistent. Another possibility to overcome this problem is proposed in [23]. The log-likelihood is first decomposed into independent summands and then importance sampling is applied. Papers [4, 23] describe asymptotic properties of MCML estimators only for models with missing data, while [16] investigates models with intractable normalizing constants and explanatory variables. However, in [4, 16, 23] it is assumed that one can efficiently generate independent samples from a given distribution. In many practical problems it is impossible and Markov chain simulation is very useful in such cases. Our goal is to investigate asymptotic properties of estimators.
obtained in this way, i.e. MCMCML. The MCMCML approach has been applied successfully in practice [9, 11, 25, 27]. However, to the best of our knowledge, there is no full theoretical justification of it in the literature. Our paper fills this gap at least partially and can be viewed as a generalization of some of the results contained in [16, 23]. The methods used in these papers are extended and developed to work in the case, where the MC sample is a Markov chain.

Our main result is the asymptotic normality of MCMCML estimators. We also prove consistency under general assumptions. We focus on models with intractable normalizing constants, but the main result can be also applied to models with missing data. We prove our theorem using classical methods, but our argument is not standard, because we consider two sources of randomness (due to the initial sample and the MCMC sample). Moreover, the MCMC approximation, given in (2.3) below, is a sum of two expressions with a rather complicated dependence structure (the second sum is conditionally a functional of a Markov chain, but also depends on the initial sample). Even though we work in a more difficult scenario, we significantly simplify argument and weaken regularity assumptions comparing to [23]. It is discussed in detail in Section 3.

The asymptotic analysis provided in our paper describes the behaviour of MCMCML estimators both in terms of number of observations $n$ and of number of Monte Carlo samples $m$. The efficiency of computational scheme essentially depends on the choice of $m$. There is a trade-off between the computational cost and the quality of the final estimator. A reasonable requirement is that the MCMCML estimator should satisfy a $\sqrt{n}$-Central Limit Theorem (as the practically unavailable exact ML). On the other hand, we should avoid unnecessary computational burden. Our main result indicates that it is good to choose $m$ of the same order as $n$ (say $m/n \to a$ with $0 < a < \infty$). Note however, that this conclusion is true for the MCMCML scheme considered in our paper, and e.g. in [16, 23]. We assume that a single run of an MCMC algorithm of length $m$ is used to approximate the likelihood corresponding to all the observed data. There are also different MCML schemes proposed in the literature, which use several independent MCMC runs (or several independent i.i.d. samples) to estimate each of likelihood contributions, e.g. [3]. The overall computational cost of such schemes behave as $mn$ and the good choice is then to let $m = o(n)$, as in the last cited
paper.

The paper is organized as follows: in Section 2 we describe the models under consideration. In Section 3 we state the main results (Theorem 3.2, Corollary 3.3, Theorem 3.4) and show their applications. In the next part of the paper we study properties of estimators in numerical experiments and Section 5 contains conclusions. The proofs are given in Section 6.

## 2 Description of the models

In the paper we consider models with intractable normalizing constants and models with missing data. We focus on the former, but the latter can be investigated similarly.

### Models with intractable constants

We consider the following parametric model with covariates

\[
p(y|x, \theta) = \frac{1}{C(x, \theta)} f(y|x, \theta),
\]

where \(y \in \mathcal{Y} \subset \mathbb{R}^d\) is a response variable, \(x \in \mathcal{X} \subset \mathbb{R}^l\) is a covariate or “explanatory” variable, \(\theta \in \Theta \subset \mathbb{R}^p\) is a parameter describing the relation between \(y\) and \(x\). The normalizing constant,

\[
C(x, \theta) = \int f(y|x, \theta)dy,
\]

is difficult or intractable.

We assume that the data consist of \(n\) independent observations \((Y_1, X_1), \ldots, (Y_n, X_n)\), which come from a joint distribution with density \(g(y, x)\). It is not necessary to assume that \(g(y|x) = p(y|x, \theta_0)\) for some \(\theta_0\). The case when no such \(\theta_0\) exists, i.e. the model is misspecified, makes the considerations only
slightly more difficult. Thus, let us consider the following log-likelihood
\[ \ell_n(\theta) = \log p(Y_1, \ldots, Y_n | X_1, \ldots, X_n, \theta) \]
\[ = \sum_{i=1}^{n} \log f(Y_i | X_i, \theta) - \sum_{i=1}^{n} \log C(X_i, \theta). \]  

(2.1)

The first term in (2.1) is easy to compute, while the second one is approximated by Markov chain Monte Carlo. Let \( h(y) \) be an importance sampling (instrumental) distribution and note that
\[ C(x, \theta) = \int f(y|x, \theta) dy = \int \frac{f(y|x, \theta)}{h(y)} h(y) dy = \mathbb{E}_{Y \sim h} \frac{f(Y|x, \theta)}{h(Y)} \]
for fixed \( \theta, x \). Therefore, a natural approximation of the normalizing constant is
\[ C_m(x, \theta) = \frac{1}{m} \sum_{k=1}^{m} \frac{f(Y^k|x, \theta)}{h(Y^k)}, \]
where \( Y^1, \ldots, Y^m \) is a sample drawn from \( h \) or, which is more realistic and is considered in the current paper, \( Y^1, \ldots, Y^m \) is a Markov chain with \( h \) being a density of its stationary distribution. The MCMC sample is independent of the initial sample. From the Law of Large Numbers (LLN) for Markov chains (e.g. [15, Th. 17.0.1]) we have \( C_m(x, \theta) \to C(\theta) \), when \( m \to \infty \) and \( \theta, x \) are fixed. Thus, an MCMC approximation of the log-likelihood \( \ell_n(\theta) \) is
\[ \ell^m_n(\theta) = \sum_{i=1}^{n} \log f(Y_i | X_i, \theta) - \sum_{i=1}^{n} \log C_m(X_i, \theta). \]

(2.3)

The considered MCMCML approach is based on maximization of the function (2.3) instead of (2.1).

Let us note that the general Monte Carlo recipe can also lead to approximation schemes different from (2.3). For instance, we could generate \( n \) independent MCMC samples instead of one, i.e. \( Y^1_i, \ldots, Y^m_i \sim h_i, i = 1, \ldots, n \) and use the \( i \)-th sample to approximate \( C(x_i, \theta) \). Using this scenario one can obtain estimators with better convergence rates, but at the cost of increased computational complexity. Another scheme, proposed in [4], approximates the log-likelihood by
\[ \sum_{i=1}^{n} \log f(Y_i | X_i, \theta) - \log \frac{1}{m} \sum_{k=1}^{m} \prod_{i=1}^{n} \frac{f(Y^k_i | X_i, \theta)}{h_i(Y^k_i)}. \]
However, this scheme leads to estimators with unsatisfactory asymptotics, unless a preliminary estimator is used, as shown in [4]. Thus, we focus our attention only on (2.3).

**Models with missing data**

These models are introduced in the same way as in [23], but our notation is slightly different. We assume that $x$ is observed and $y$ is missing in the complete data $(x, y)$.

The joint density is denoted by $f(x, y|\theta)$, while the unavailable marginal density is $f(x|\theta) = \int f(x, y|\theta)dy$. Let the observed data $X_1, \ldots, X_n$ be i.i.d. from some density $g$. Obviously, a maximizer of the log-likelihood

$$\sum_{i=1}^{n} \log f(X_i|\theta)$$

cannot be calculated. We use a Markov chain $Y_1, \ldots, Y^m$ with the stationary distribution $h$, which is independent of $X_1, \ldots, X_n$, to approximate the unknown marginal density $f(x|\theta)$ by

$$\frac{1}{m} \sum_{k=1}^{m} \frac{f(x, Y^k|\theta)}{h(Y^k)}.$$

Therefore, the MCMCML approach is based on maximization of

$$\sum_{i=1}^{n} \log \left[ \frac{1}{m} \sum_{k=1}^{m} \frac{f(X_i, Y^k|\theta)}{h(Y^k)} \right],$$

which is equivalent to maximization of

$$\sum_{i=1}^{n} \log f(X_i|\theta) + \sum_{i=1}^{n} \log \frac{1}{m} \sum_{k=1}^{m} \frac{f(Y^k|X_i, \theta)}{h(Y^k)}. \hspace{1cm} (2.4)$$
3 Main results

In this section we state the main results of the paper and describe their applications. We examine in detail models with intractable normalizing constants and covariates. Similar results for missing data models are only briefly commented on, because they are straightforward modifications of our theorems. First we provide sufficient conditions for consistency of MCMCML estimators, then we proceed to our main focus, which is asymptotic normality of these estimators.

We need the following notations: the MCMC approximation (2.3) multiplied by \( \frac{1}{n} \) is denoted by \( \bar{\ell}_n^m(\theta) \) and decomposed as follows

\[
\bar{\ell}_n^m(\theta) = \bar{\ell}_n(\theta) - r_n^m(\theta),
\]

where

\[
\bar{\ell}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[ \log f(Y_i|X_i, \theta) - \log C(X_i, \theta) \right] = \frac{1}{n} \sum_{i=1}^{n} \log p(Y_i|X_i, \theta),
\]

\[
r_n^m(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[ \log C_m(X_i, \theta) - \log C(X_i, \theta) \right] = \frac{1}{n} \sum_{i=1}^{n} \log Z_m(X_i, \theta),
\]

where for fixed \( \theta, x \)

\[
Z_m(x, \theta) = \frac{1}{m} \sum_{k=1}^{m} p(Y^k|x, \theta^*) h(Y^k). \]

Let \( \theta^* \) be a maximizer of \( \ell(\theta) = \mathbb{E}_{Y,X \sim g} \log p(Y|X, \theta) \), i.e. the Kullback-Leibler projection.

Let symbols \( \nabla \) and \( \nabla^2 \) denote derivatives with respect to \( \theta \) and introduce the following notations:

\[
\Psi(y|x) = \frac{\nabla p(y|x, \theta^*)}{h(y)},
\]

\[
\Psi(y) = \mathbb{E}_{X \sim g} \Psi(y|X).
\]

Consistency of MCMCML estimators is relatively easy to establish for models with concave log-likelihoods (e.g. autologistic, auto-Poisson and autonormal models in [2]). In the general case it can be established using hypo-convergence of \( \bar{\ell}_n^m \). For the background on hypo-convergence (epi-convergence),
we refer to [1] or [10]. In the following theorem we investigate this type of convergence.

**3.2 Theorem.** Suppose the following assumptions are satisfied:

1. \( p(y|x, \theta) \) is continuous with respect to \( \theta \in \Theta \) for all \( y \in Y \) and \( x \in X \),
2. \( \sup_{x \in X} |Z_m(x, \theta) - 1| \to_{a.s.} 0, \ m \to \infty, \) for all \( \theta \in \Theta \),
3. for every \( \theta \in \Theta \) there exists a neighbourhood \( U_1 \) of \( \theta \) such that
   \[
   \sup_{x \in X} \left| \frac{1}{m} \sum_{k=1}^{m} \inf_{\gamma \in U_1} \frac{p(Y^k|x, \gamma)}{h(Y^k)} - \mathbb{E}_{Y \sim h} \inf_{\gamma \in U_1} \frac{p(Y|x, \gamma)}{h(Y)} \right| \to_{a.s.} 0, \ m \to \infty, 
   \]
4. (a) for every \( \theta \in \Theta \) there exists a neighbourhood \( U_2 \) of \( \theta \) such that
   \[
   \mathbb{E}_{(Y,X) \sim \rho} \left| \log \sup_{\gamma \in U_2} p(Y|X, \gamma) \right| < \infty,
   \]
   (b) for every \( \theta \in \Theta \) there exists a neighbourhood \( U_3 \) of \( \theta \) such that the expression \( \inf_{x \in X} \mathbb{E}_{Y \sim h} \inf_{\gamma \in U_3} \frac{p(Y|x, \gamma)}{h(Y)} \) is positive.

Then \( \bar{\ell}_m \) hypo-convexes to \( \ell \) almost surely.

Theorem 3.2 is an extension of [23, Theorem 2.2] to the case when the Monte Carlo sample is a Markov chain instead of i.i.d. random variables. Assumptions 1-3 of this theorem are analogs of conditions (2), (5), (4) in [23, Theorem 2.2], respectively. The last assumption in Theorem 3.2 is stronger than condition (3) in [23, Theorem 2.2]. However, this condition is not very restrictive. It is satisfied, for instance, in the autologistic model considered in Example 3.6.

Combining Theorem 3.2 with [1, Theorem 1.10] we obtain existence of a consistent sequence of MCMCML estimators as in [10, 23].

**3.3 Corollary.** Under the assumptions of Theorem 3.2, if \( \theta_* \) exists and is unique, then with probability one for sufficiently large \( n, m \) there exist local maximizers \( \hat{\theta}_n^m \) of the MCMC approximation (2.3) such that \( \hat{\theta}_n^m \to \theta_* \) as \( n, n \to \infty \).
Proof of Corollary 3.3. Let $K(\theta_*, r)$ be a compact ball in $\mathbb{R}^p$ with the centre $\theta_*$ and a radius $r > 0$. Function $\bar{\ell}_m$ is continuous, so there exists a point, denoted $\hat{\theta}_m^*$, at which the function attains the maximum on $K(\theta_*, r)$. Point $\theta_*$ is the unique minimizer of $\ell$ and with probability one, $\bar{\ell}_m$ hypo-converges to $\ell$ by Theorem 3.2. From [10, Proposition 1] and comments after this proposition it follows that $\hat{\theta}_m^*$ converges to $\theta_*$. This implies that for $n, m$ large enough $\hat{\theta}_m^*$ is in the interior of $K(\theta_*, r)$, so it is a local maximizer of $\bar{\ell}_m$.

Now we state the main result of the paper that concerns asymptotic normality of MCMCML estimators.

3.4 Theorem. Suppose the following assumptions are satisfied:

1. $(Y_k)_{k \geq 1}$ is a reversible and geometrically ergodic homogeneous Markov chain on the state space $Y$, which has stationary distribution with a density $h$ and initial distribution with density $q$ such that $\|q/h\|_\infty = \sup_y |q(y)/h(y)| < \infty$.
2. second partial derivatives of $p(y|x, \theta)$ with respect to $\theta \in \Theta$ exist and are continuous for all $y \in Y$ and $x \in X$, and may be passed under the integral sign in $\int p(y|x, \theta)dy$ for fixed $x \in X$,
3. there is an interior point $\theta_*$ of $\Theta$ such that $E_{(Y,X) \sim g} \nabla \log p(Y|X, \theta_*) = 0$, matrices
   \begin{align*}
   V &= \text{VAR}_{(Y,X) \sim g} \nabla \log p(Y|X, \theta_*),
   D &= E_{(Y,X) \sim g} \nabla^2 \log p(Y|X, \theta_*)
   \end{align*}
   are well defined and matrix $D$ is negative definite,
4. the expectation $E_{Y \sim h, X \sim g} |\Psi(Y|X)|^2$ is finite,
5. the MCMCML estimator $\hat{\theta}_m^*$ is a consistent estimator of $\theta_*$,
6. $\sup_{\theta \in U} \left| \nabla^2 \bar{\ell}_n(\theta) - \text{E}_{(Y,X) \sim g} \nabla^2 \log p(Y|X, \theta) \right| \to_p 0, \quad n \to \infty$, where $U = \{\theta : |\theta - \theta_*| \leq \delta\}$ is some neighbourhood of $\theta_*$ ($\delta > 0$).
7. \(a) \sup_{x \in X} |Z_m(x, \theta_*) - 1| \to_p 0, \quad m \to \infty, \)
(b) \( \sup_{x \in \mathcal{X}} |\nabla Z_m(x, \theta_*)| \to_p 0, \ m \to \infty, \)

(c) \( \sup_{x \in \mathcal{X}} |\nabla^2 Z_m(x, \theta)| \to_p 0, \ m \to \infty, \) for some neighbourhood \( U \) of \( \theta_* \).

Then
\[
\left( \frac{V}{n} + \frac{W}{m} \right)^{-\frac{1}{2}} D \left( \hat{\theta}_m^m - \theta_* \right) \to_d N(0, I), \quad n, m \to \infty,
\]

for
\[
W = \text{VAR}_h \bar{\Psi}(Y^1) + 2 \sum_{k=2}^{\infty} \text{COV}_h \left( \bar{\Psi}(Y^1), \bar{\Psi}(Y^k) \right),
\]
where \( \text{VAR}_h \) and \( \text{COV}_h \) denote the stationary covariance matrices.

In Theorem 3.4 we prove that the maximizer of (2.3) satisfies
\[
(3.5) \quad \hat{\theta}_m^m \sim_{\text{approx}} N \left( \theta_*, D^{-1} \left( \frac{V}{n} + \frac{W}{m} \right) D^{-1} \right).
\]

Formula (3.5) means that the estimator \( \hat{\theta}_m^m \) behaves like a normal vector with the mean \( \theta_* \), when both the initial sample size \( n \) and the Monte Carlo sample size \( m \) are large. Suppose that \( \hat{\theta}_n \) is a maximizer of \( \ell_n(\theta) \), which is a genuine maximum likelihood estimator, then
\[
\hat{\theta}_n \sim_{\text{approx}} N \left( \theta_*, \frac{1}{n} D^{-1} V D^{-1} \right).
\]

Note that the first component of the asymptotic variance in (3.5) is the same as the asymptotic variance of the maximum likelihood estimator \( \hat{\theta}_n \). The second component, \( D^{-1} W D^{-1} / m \), is due to Monte Carlo randomness. Furthermore, if \( m \) is large, then asymptotic behaviour of \( \hat{\theta}_m^m \) and \( \hat{\theta}_n \) is similar. Finally, if the model is correctly specified, that is \( g(y|x) = p(y|x, \theta_0) \) for some \( \theta_0 \), then \( \theta_* = \theta_0 \) and \( D = -V \).

Now we discuss the assumptions of Theorem 3.4. Assumption 1 relates to the MCMC sample and is not restrictive. Conditions 2-4 are standard regularity assumptions. Consistency of \( \hat{\theta}_m^m \) stated in Condition 5 is easily satisfied, when
the log-likelihood is concave. In the general case we can use Corollary 3.3. The key conditions in Theorem 3.4 are Assumptions 6 and 7. They stipulate uniform convergence of a sum of independent random variables (Assumption 6) and a sum along a trajectory of Markov chain (Assumption 7). We show that our conditions are satisfied in the widely-used autologistic model.

3.6 EXAMPLE. Consider the autologistic model with covariates defined as follows. The response variable is a binary vector \( y = (y(s), s = 1, \ldots, d) \in \{0, 1\}^d \). The vector of covariates or “explanatory” variables is \( x = (x(s), s = 1, \ldots, l) \in X \), where \( X \) is a compact subset of \( \mathbb{R}^l \). The parameter is \( \theta = (\beta, \alpha) \), where \( \beta = (\beta_{r,s}) \) and \( \alpha = (\alpha_{r,s}) \) are matrices of dimensions \( d \times d \) and \( d \times l \), respectively. For identifiability, assume \( \beta_{r,s} = \beta_{s,r} \). The probability distribution on \( Y = \{0, 1\}^d \) is defined as follows

\[
p(y|x, \beta, \alpha) := \frac{1}{C(x, \beta, \alpha)} \exp \left( \sum_{r=1}^{d} \sum_{s=1}^{d} \beta_{r,s} y(r)y(s) + \sum_{r=1}^{d} \sum_{s=1}^{l} \alpha_{r,s} y(r)x(s) \right) .
\]

Obviously, the normalizing constant \( C(x, \beta, \alpha) \) is intractable for large \( d \). We denote

\[
T(x, y) = (y(1)y(1), y(1)y(2), \ldots, y(d)y(d), y(1)x(1), \ldots, y(d)x(l)) .
\]

Then considering the pair of matrices \( (\beta, \alpha) \) as a vector denoted by \( \theta \) we can simply write

\[
p(y|x, \theta) = \frac{\exp(\theta^T T(x, y))}{C(x, \theta)} .
\]

Let us now show that the assumptions of Theorem 3.4 hold in the autologistic model. Assumption 1 is fulfilled if \( (Y^k)_{k \geq 1} \) is (e.g.) a Gibbs sampler on \( Y \) with a stationary discrete density \( h \) such that \( h(y) > 0 \) for all \( y \in \{0, 1\}^d \).

To check that Assumptions 2-4 are satisfied, we refer to some general properties of exponential families, c.f. [12]. Note that \( p(y|x, \theta) \) is, for every fixed \( x \), a regular exponential family of densities and the cumulant generating function \( -\log C(x, \theta) \) is a concave, analytic function of \( \theta \in \mathbb{R}^d \). Assumption 2 is trivially satisfied, since the integral with respect to \( y \) is just a finite sum.
To see that matrices $V$ and $D$ are well defined, it is enough to note that the derivatives,

$$
\nabla \log p(y|x, \theta) = -\frac{\nabla C(x, \theta)}{C(x, \theta)} + T(x, y),
$$

$$
\nabla^2 \log p(y|x, \theta) = -\frac{\nabla^2 C(x, \theta)}{C(x, \theta)} + \frac{\nabla C(x, \theta) \nabla C(x, \theta)^T}{C^2(x, \theta)}
$$

are continuous functions of $(x, y)$ on a compact space (for $\theta$ fixed). The same argument shows that Assumption 4 is satisfied.

We have $E_{(Y, X) \sim g} \nabla \log p(Y|X, \theta_*) = 0$, because we can exchange the order of operators `$E_{(Y, X) \sim g}$' and `$\nabla$' using [22, Th. 115]. If $\theta_*$ maximizes $E_{(Y, X) \sim g} \log p(Y|X, \theta)$, then the derivative at $\theta_*$ is zero.

More delicate is negative definiteness of $D$. Standard properties of exponential families imply that $D(x, \theta) := \nabla^2 \log p(y|x, \theta)$ is negative semidefinite for every fixed $x$. Negative semidefiniteness of $D = E_{X \sim g}D(X, \theta_*)$ follows immediately. However, $D(x, \theta)$ is not negative definite, because for fixed $x$ the parameters of the exponential family are not identifiable: $\theta = (\beta, \alpha)$ and $\theta' = (\beta', \alpha')$ define the same probability distribution on $Y$ iff $\beta_{r,r} + \sum_{s=1}^{l} \alpha_{r,s} x(s) = \beta'_{r,r} + \sum_{s=1}^{l} \alpha'_{r,s} x(s)$ and $\beta_{r,s} = \beta'_{r,s}$ for $r \neq s$. To ensure negative definiteness of $D$ we have to assume that the support of $X \sim g$ is not contained in any affine subspace of $\mathbb{R}^d$ of lower dimension. Indeed, by [12, Th. 2.4], $\theta^T D(x, \theta_*) \theta = 0$ iff for every $r$ we have $\beta_{r,r} + \sum_{s=1}^{l} \alpha_{r,s} x(s) = 0$ and $\beta_{r,s} = 0$ for $r \neq s$ ($\theta = (\beta, \alpha)$ arranged as a vector).

Condition 6 is a uniform LLN over $U = \{\theta : |\theta - \theta_*| \leq \delta\}$. Notice that $\nabla^2 \log p(y|x, \theta)$ is continuous in $\theta$ and bounded. Moreover, this function is matrix-valued, so it is enough to show uniform convergence for each component. Since $U$ is compact and samples $(Y_1, X_1), \ldots, (Y_n, X_n)$ are i.i.d., Condition 6 is implied by [6, Theorem 16(a)].

Uniform convergence in Conditions 7 (a)-(c) relates to the MC sample which is a Markov chain. In Conditions 7(b) and 7(c) we have vector- and matrix-valued functions, respectively, so again it is enough to prove uniform convergence for each component. Using compactness of $X, U$ and continuity of the function $T$ in $x$, we can prove Conditions 7 (a)-(c) similarly to Condition 6. Indeed, [6, Theorem 16(a)] can be extended to Markov chains, if we apply Strong Law of Large Numbers (SLLN) for Markov chains on the top of page
110 in [6] (in fact, this theorem can be extended to arbitrary sequence of random variables that admits SLLN).

Condition 5 can be verified as follows. Notice that for fixed \( \theta \) in a neighbourhood \( U \) of \( \theta \), we have

\[
\ell_n^m(\theta) \to_p E_{(Y,X) \sim \theta} \log p(Y|X, \theta) \quad \text{as} \quad n, m \to \infty.
\]

Indeed, from assumption 7 we obtain a uniform version over \( \theta \in U \) of assumption 7(a). Using it, we prove convergence \( r_n^m(\theta) \to_p 0 \), that implies (3.7). Moreover, we can express

\[
\ell_n^m(\theta) = \frac{1}{n} \sum_{i=1}^{n} \theta^T T(X_i, Y_i) - \frac{1}{n} \sum_{i=1}^{n} \log \left( \frac{1}{m} \sum_{k=1}^{m} \exp \left( \theta^T T(X_i, Y^k) \right) \right).
\]

The Hessian of \( \ell_n^m(\theta) \) is a weighted covariance matrix with positive weights that sum up to 1, so \( \ell_n^m \) is concave. This property and (3.7) implies convergence of maximizers in Condition 5 as in [18].

3.8 REMARK. Our Theorem 3.4 can be applied to the autonormal model and auto-Poisson model as well. These models have been introduced in [2] and we refer to this paper for definitions and motivations behind them. Note that in these two examples the observation space is not compact. Nevertheless, verification of the assumptions of Theorem 3.4 for these models is quite analogous as for the autologistic model in Example 3.6, because these assumptions do not include compactness of the observation space. Therefore, we omit details.

3.9 REMARK. In Example 3.6 we check uniform convergence over \( \mathcal{X} \) and \( \mathcal{X} \times U \) from Assumption 7 using [6, Theorem 16(a)]. To apply this fact we need that both sets are compact. Therefore, we assume that the covariate space \( \mathcal{X} \) is compact in Example 3.6.

3.10 REMARK. Theorem 3.4 can be also applied to models with missing data as described in Section 2. Indeed, using the close relation between MCMC approximations (3.1) and (2.4) we should just follow the proof of Theorem 3.4. Thus, under analogous regularity assumptions to Theorem 3.4 we obtain

\[
\left( \frac{V}{n} + \frac{W}{m} \right)^{-\frac{1}{2}} D \left( \hat{\theta}_n^m - \theta \right) \to_d \mathcal{N}(0, I), \quad n, m \to \infty,
\]
where $V = \text{VAR}_{X \sim \theta} \nabla \log f(X|\theta_*)$, $D = \mathbb{E}_{X \sim \theta} \nabla^2 \log f(X|\theta_*)$ and

$$W = \text{VAR}_h \bar{\Psi}(Y^1) + 2 \sum_{k=2}^{\infty} \text{COV}_h (\bar{\Psi}(Y^1), \bar{\Psi}(Y^k))$$

for $\bar{\Psi}(y) = \mathbb{E}_{X \sim \theta} \nabla f(y|X, \theta_*)/h(y)$. Therefore, we extend [23, Theorem 2.3] from the i.i.d. case to the Markov chain case. The only price we pay for this generalization is (mild) Condition 1. Moreover, we weaken their condition (6). Namely, we replace the requirement that the class is Donsker by the requirement that the class is Glivenko-Cantelli. This fact follows from argumentation used to obtain

$$\left( \frac{V}{n} + \frac{W}{m} \right)^{-\frac{1}{2}} \nabla \bar{\ell}_n^m(0) \to_d \mathcal{N}(0, I), \quad n, m \to \infty,$$

which is needed in the proof of Theorem 3.4 and its analog in the proof of [23, Theorem 2.3]. While [23] uses an advanced theory and complicated method based on weak convergence of stochastic processes and its properties (see [23, Lemma A.4]), we show in Section 6 that this analysis can be based on much simpler methods.

3.11 REMARK. The efficiency of the MCMCML algorithm depends on the choice of the instrumental density $h$ and the MCMC sample size $m$. Theorem 3.4 gives some general indications how to tackle both these issues. As we have already mentioned in Introduction, the asymptotics of $\hat{\theta}_n^m$ suggest that $m$ should be of the same order as $n$. The choice of $h$ should take into account two requirements. A well-known result for reversible Markov chains implies that $W \leq \frac{1+\varrho}{1-\varrho} \text{VAR}_h \bar{\Psi}(Y^1)$, where $1-\varrho$ is the spectral gap of the MC chain $Y^k$ (the inequality is understood in the sense of nonnegative definiteness; a one-dimensional version can be found in [21]). Therefore, on the one hand, we want to choose $h$ such that $\text{VAR}_h \bar{\Psi}(Y^1)$ be as small as possible. This issue is discussed in [17, Subsection 2.3]. On the other hand, we want to choose $h$ such that there exists an efficient MCMC algorithm targeting $h$, i.e. with a spectral gap $1-\varrho$ as big as possible. For a model without covariates there is a heuristic choice of $h$, which is a reasonable compromise between the two requirements: one can choose e.g. $h = p(\cdot|\hat{\theta})$, where $\hat{\theta}$ is some preliminary estimator of $\theta_*$ (this choice is advocated in [25, 11, 27]).
4 Numerical experiments

In our experiments we use the autologistic model with $d = 10$ and $l = 10$. This model is described in Example 3.6. We use two settings of parameters: the first one with relatively small values of parameters, namely $\beta := \beta_{r,s} = 0.01$ for all $r, s$. Moreover, $\alpha_{s,s} = \pm 1$ with signs chosen randomly and $\alpha_{r,s} = 0$ for $r \neq s$. In the second scenario we consider larger values of parameters: $\beta = 0.1$ and $\alpha_{s,s} = \pm 10$, where signs are again chosen randomly. In both settings we simulate $n = 100$ observations, with covariates $x \in \mathbb{R}^{10}$ drawn from the standard normal distribution $\mathcal{N}(0, I)$. In all experiments we use $p(y|x, \gamma \theta)$ as an instrumental distribution, where $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_{10})$ is a vector of columnwise means $\bar{x}_i$. Moreover, $\tilde{\theta}$ is chosen as the true values of unknown parameters or as the maximum pseudolikelihood estimator. Finally, the scaling factor $\gamma$ is chosen from the set $\{0.4, 0.6, 0.8, 1\}$.

We choose the MCMC sample size as $m = 100$ and replicate experiments 100 times per each setting. The boxplots of estimators of the correlation parameter $\beta$ and Q-Q plots of scaled estimation errors are presented on Fig. 1 and Fig. 2 for the first setting with small values of parameters. Results for the second setting are given on Fig. 3 and Fig. 4. We observe that choosing the maximum pseudolikelihood estimator in the instrumental distribution we introduce high variability of final estimators in all cases. It can be explained by the fact that we use the sample size $n = 100$ to estimate 11 parameters of the model and the maximum pseudolikelihood estimators do not work well in this case. When we use true values of parameters in the instrumental distribution, then the quality of estimators looks satisfactory for all scaling factors $\gamma$ in setting 1. However, in setting 2 we observe the growth of variance for scaling factors close to 1. This fact agrees with our theoretical results. Namely, the asymptotic variance of the estimator is a combination of the variance due to the data and the variance due to the MCMC procedure. It is well known, that the Gibbs sampler does not ,,mix“ well for for relatively large correlation parameters. It explains the behaviour of the estimator for $\beta = 0.1$.

Summarizing, the choice of the instrumental distribution in the MCMC procedure is crucial and has a significant impact on the quality of the estimator. The natural choice is to take the distribution corresponding to the parameter $\theta_\star$. The results for setting 1 confirms this approach. In practice $\theta_\star$ is
unknown, so we propose to use a preliminary estimator of $\theta^\star$, for instance the pseudolikelihood estimator. However, as we can observe in Fig. 1 and Fig. 2, it decreases the accuracy of the final estimator. In setting 2 the quality of estimators becomes worse, because we also meet the problems with slow mixing of Gibbs samplers.

5 Conclusions

We consider the problem of estimation in complex statistical models, when the likelihood cannot be computed efficiently. We study asymptotic properties of estimators based on Markov chain Monte Carlo approximation of the likelihood. Namely, we establish their consistency and asymptotic normality.

The choice of the instrumental distribution in the MCMC method plays an important role. We propose a heuristic approach based on a preliminary
estimator. However, this problem should be investigated thoroughly and can be a fruitful topic of future studies.

In the paper we consider random covariates. The case of deterministic covariates is also popular, therefore obtaining similar results for this scenario would be useful. However, it is not just a simple consequence of our results, because in the deterministic covariates case the initial sample \((Y_1, X_1), \ldots, (Y_n, X_n)\) consists of independent but non-identically distributed random vectors. Therefore, some tools used in the current paper (for instance, CLT or uniform LLN from [6]) will have to be replaced by their analogs working for non-identically distributed vectors. It is an interesting but involved problem, which will be a scope of a future paper.
6 Proofs and auxiliary results

This section contains proofs of Theorem 3.2, Theorem 3.4 and lemmas, which are used in these proofs.

Proof of Theorem 3.2

The proof is analogous to the proof of [23, Theorem 2.2] and is mainly based on three facts given in [23, Lemma A.1-A.3]. In Lemma 6.1 below we state only an analog of [23, Lemma A.2], because the remaining ones are not difficult to establish. Indeed, almost sure pointwise convergence of the MCMC approximation to the true log-likelihood \( \ell \) follows from Assumption 2. Upper semi-continuity of \( \ell \) can be obtained using Assumptions 1 and 4(a). □

6.1 Lemma. Suppose that Assumptions 3 and 4 of Theorem 3.2 are satisfied.
Figure 4: Setting 2 - Q-Q plots for scaled errors of estimators, i.e. \( \sqrt{n}(\hat{\beta} - \beta) \), for different instrumental distributions.

Then for arbitrary \( \theta \)

\[
\liminf_{(m,n) \to (\infty, \infty)} \inf_{\gamma \in B} -\bar{e}_n^m(\gamma) \geq \mathbb{E}_{X \sim g} \log \left[ \mathbb{E}_{Y \sim h} \inf_{\gamma \in B} \frac{p(Y \mid X, \gamma)}{h(Y)} \right] - \mathbb{E}_{(Y,X) \sim g} \log \sup_{\gamma \in B} p(Y \mid X, \gamma)
\]

with probability one for each subset \( B \) of \( U_1 \cap U_2 \cap U_3 \), where \( U_1, U_2, U_3 \) are introduced in Assumptions 3 and 4.

Proof. Fix \( \theta \) and \( B \) contained in neighbourhoods of \( \theta \) given in Assumptions 3 and 4 of Theorem 3.2. Notice that by Assumption 4 the right-hand side of the inequality in the lemma is greater than \(-\infty\).

By superadditivity of the infimum operation, subadditivity of the supremum
operation and monotonicity of the logarithm function we obtain

\[
\begin{align*}
\text{(6.4)} \quad \inf_{\gamma \in B} -\bar{\ell}_n^m(\gamma) & \geq \frac{1}{n} \sum_{i=1}^{n} \log \left[ \frac{1}{m} \sum_{k=1}^{m} \inf_{\gamma \in B} \frac{p(Y^k | X_i, \gamma)}{h(Y^k)} \right] \\
\text{(6.5)} & - \frac{1}{n} \sum_{i=1}^{n} \log \sup_{\gamma \in B} p(Y_i | X_i, \gamma)
\end{align*}
\]

Now we prove that the right-hand side of (6.4) converges almost surely to the right-hand side of (6.2) as well as (6.5) to (6.3). The latter convergence is simply implied by SLLN and Assumption 4(a). Therefore, we focus on the former one and write

\[
\frac{1}{n} \sum_{i=1}^{n} \log \frac{1}{m} \sum_{k=1}^{m} \inf_{\gamma \in B} \frac{p(Y^k | X_i, \gamma)}{h(Y^k)} - \mathbb{E}_{X \sim g} \log \mathbb{E}_{Y_i \sim h} \inf_{\gamma \in B} \frac{p(Y | X, \gamma)}{h(Y)}
\]

\[
\begin{align*}
\text{(6.6)} & = \frac{1}{n} \sum_{i=1}^{n} \left[ \log \frac{1}{m} \sum_{k=1}^{m} \inf_{\gamma \in B} \frac{p(Y^k | X_i, \gamma)}{h(Y^k)} - \log \mathbb{E}_{Y_i \sim h} \inf_{\gamma \in B} \frac{p(Y | X, \gamma)}{h(Y)} \right] \\
\text{(6.7)} & + \frac{1}{n} \sum_{i=1}^{n} \log \mathbb{E}_{Y_i \sim h} \inf_{\gamma \in B} \frac{p(Y | X, \gamma)}{h(Y)} - \mathbb{E}_{X \sim g} \log \mathbb{E}_{Y_i \sim h} \inf_{\gamma \in B} \frac{p(Y | X, \gamma)}{h(Y)}
\end{align*}
\]

To finish the proof we show that (6.6) tends to zero almost surely by Assumptions 3 and 4(b) of Theorem 3.2, while (6.7) tends to zero by SLLN and Assumption 4(b).

\[\square\]

**Proof of Theorem 3.4**

We start with the following lemma, which plays the key role in the proof of Theorem 3.4.

**6.8 Lemma.** Let \((Y^k)_{k \geq 1}\) be a reversible and geometrically ergodic homogeneous Markov chain with spectral gap \(1 - \rho\). Assume its stationary distribution \(\pi\) and initial distribution \(\nu\) are such that \(\|d\nu / d\pi\|_\infty = \sup_y |(d\nu / d\pi)(y)| < \infty\). Let \(\phi\) be a function such that \(\int \phi(y)\pi(dy) = 0\) and \(\int \phi(y)^2 \pi(dy) = \|\phi\|_\pi^2 < \infty\). For all \(k \leq l\) it holds that

\[
||\mathbb{E}[\phi(Y^k)\phi(Y^l)]|| \leq \left\| \frac{d\nu}{d\pi} \right\|_\infty \|\phi\|_\pi^2 \rho^{l-k}.
\]
Note that we will apply Lemma 6.8 in a situation where both the distributions \( \pi \) and \( \nu \) have densities \( h \) and \( q \), respectively. Then the Radon-Nikodym derivative \((d\nu/d\pi)(y)\) is simply \( q(y)/h(y) \). First we give a proof of the main result, then a proof of Lemma 6.8.

**Proof of Theorem 3.4.** Let us first describe the outline of the proof. The beginning of the proof is standard, namely we define

\[
D^m_n = \int_0^1 \nabla^2 \tilde{\ell}_n^m \left( \theta_* + s(\hat{\theta}_n^m - \theta_*) \right) \, ds.
\]

By Taylor expansion and the fact that \( \hat{\theta}_n^m \) maximizes \( \tilde{\ell}_n^m \) we obtain

\[-\nabla \tilde{\ell}_n^m(\theta_*) = D^m_n(\hat{\theta}_n^m - \theta_*).
\]

If we show that

(6.9) \[ D^m_n \to_p D, \quad n, m \to \infty \]

and

(6.10) \[ \left( \frac{V}{n} + \frac{W}{m} \right)^{-1} \nabla \tilde{\ell}_n^m(\theta_*) \to_d \mathcal{N}(0, I), \quad n, m \to \infty, \]

then the conclusion of the theorem will follow from (6.9), (6.10) and Slutsky’s theorem.

The argumentation to obtain (6.9) is the same as in the proof of [23, Lemma A.6]. The fact that the MC sample is a Markov chain instead of independent random variables does not play any role in it. Therefore, we omit the proof of (6.9) and focus on (6.10).

To show (6.10) we express \( \nabla \tilde{\ell}_n^m(\theta_*) \) as follows:

\[
\nabla \tilde{\ell}_n^m(\theta_*) = \nabla \tilde{\ell}_n(\theta_*) - \nabla r_n^m(\theta_*) = \nabla \tilde{\ell}_n(\theta_*) - \frac{1}{n} \sum_{i=1}^n \nabla \log Z_m(X_i, \theta_*)
\]

\[
= \nabla \tilde{\ell}_n(\theta_*) - \frac{1}{n} \sum_{i=1}^n \left[ \frac{\nabla Z_m(X_i, \theta_*)}{Z_m(X_i, \theta_*)} - \nabla Z_m(X_i, \theta_*) \right]
\]

(6.11) \[ - \frac{1}{n} \sum_{i=1}^n \left[ \nabla Z_m(X_i, \theta_*) - Z_m(\theta_*) \right] \]

(6.12) \[ - \frac{1}{n} \sum_{i=1}^n \left[ \nabla Z_m(X_i, \theta_*) - Z_m(\theta_*) \right] - \tilde{Z}_m(\theta_*), \]

21
where \( \bar{Z}_m(\theta) = \mathbb{E}_{X \sim g} \nabla Z_m(X, \theta) \). The first term of the above displayed equation depends only on the initial, i.i.d. sample and \( \sqrt{n} \nabla \bar{\ell}_n(\theta_*) \to_d \mathcal{N}(0, V) \) as \( n \to \infty \). This fact is well-known and follows from the CLT for i.i.d. variables. The last term depends only on the MCMC sample. Since

\[
\bar{Z}_m(\theta_*) = \frac{1}{m} \sum_{k=1}^{m} \bar{\Psi}(Y^k),
\]

we can apply the CLT for Markov chains to infer that \( \sqrt{m} \bar{Z}_m(\theta_*) \to_d \mathcal{N}(0, W) \) as \( m \to \infty \). We will later prove that both the middle terms, (6.11) and (6.12) are negligible in the sense that they are \( o_p(1/\sqrt{m}) \). Provided this is done, the rest of the proof is easy. We first assume that \( \frac{n}{n+m} \to a \) and consider three cases corresponding to rates at which \( n \) and \( m \) go to infinity: \( 0 < a < 1 \), \( a = 0 \) and \( a = 1 \). Once (6.10) is proved in these three special cases, the subsequence principle shows that it is valid in general (for \( n \to \infty \) and \( m \to \infty \) at arbitrary rates). We consider only the case \( 0 < a < 1 \), because argumentation for the others is similar. Since the Monte Carlo sample is independent of the observed one, we infer that

\[
\sqrt{n + m} \nabla \bar{\ell}_n(\theta_*) = \sqrt{n + m} \bar{Z}_m(\theta_*) + o_p(1),
\]

\[\to_d \mathcal{N}(0, V/a + W/(1 - a)).\]

To finish the proof, just note that

\[
\sqrt{n + m} \left( V/a + W/(1 - a) \right)^{-\frac{1}{2}} \left( V/n + W/m \right)^{\frac{1}{2}} \to I \quad n, m \to \infty.
\]

We are left with the task of bounding the terms (6.11) and (6.12). This is the difficult and novel part of the proof. Since the MC sample \( Y^1, \ldots, Y^m \) is a Markov chain, we will need Lemma 6.8. (Note that the behaviour of these terms would be much easier to examine, if we considered a simplified scenario of i.i.d. Monte Carlo, as in [16].)

We start with (6.12). We are to show that

\[
A_m^n = \frac{\sqrt{m}}{n} \sum_{i=1}^{n} \left[ \nabla Z_m(X_i, \theta_*) - \bar{Z}_m(\theta_*) \right] = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \left[ \Psi(Y^k|X_i) - \bar{\Psi}(Y^k) \right]
\]

\[\to_d \mathcal{N}(0, V/a + W/(1 - a)).\]
goes to 0 in probability, as \( m, n \to \infty \). This a vector-valued expression, but it is enough to bound its components separately. Let \( a^m_n, \psi(y|x) \) and \( \bar{\psi}(y) \) denote any single component of \( A^m_n, \Psi(y|x) \) and \( \bar{\Psi}(y) \), respectively. Write also \( \phi(y|x) = \psi(y|x) - \bar{\psi}(y) \). We will bound

\[
E(a^m_n)^2 = E\left[ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \phi(Y^k|X_i) \right]^2,
\]

where the symbol \( E \) refers to the expectation with respect to the both samples (the i.i.d. variables \( X_1, \ldots, X_n \) and Markov chain \( Y^1, \ldots, Y^m \) started at \( \nu \)). If we first fix \( Y^k \)s then the random variables \( \phi_m(X_i) = \sum_{k=1}^{m} \phi(Y^k|X_i)/\sqrt{m} \) are i.i.d. and centered. Therefore the expectation with respect to the \( X_i \)s in (6.14) is the variance of a mean of i.i.d. summands, equal to \( \frac{1}{n} \) times the variance of a single summand. Consequently,

\[
E(a^m_n)^2 = \frac{1}{n} E\left[ \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \phi(Y^k|X) \right]^2
\]

(the expectation with respect to \( X \sim g \) and the MC sample \( Y^k \)). Clearly,

\[
E(a^m_n)^2 = \frac{1}{nm} \sum_{k=1}^{m} E\phi^2(Y^k|X) + \frac{2}{nm} \sum_{1 \leq k < l \leq m} E\phi(Y^k|X)\phi(Y^l|X).
\]

Now, if we fix \( X \) and consider randomness in \( Y^k \)s, then \( \phi(Y^k|X) \) is a functional of the Markov chain. Since \( E_{Y \sim h} \phi(Y|x) = 0 \) for each \( x \), we are in a position to apply Lemma 6.8. Consequently, for \( k \leq l \) we have

\[
E_{X \sim g, \nu} \phi(Y^k|X)\phi(Y^l|X) \leq \left\| \frac{q}{h} \right\|_{\infty} \ E_{X \sim g, Y \sim h} \phi^2(Y|X)\rho^{l-k}.
\]

The norm and the expectation on the right-hand side of (6.16) are finite, by Assumptions 1 and 5 of the Theorem. Since \( m + 2 \sum_{1 \leq k < l \leq m} \rho^{l-k} \leq m(1 + \rho)/(1 - \rho) \), from (6.15) we obtain

\[
E(a^m_n)^2 \leq \frac{1}{n} \left\| \frac{q}{h} \right\|_{\infty} \ E_{X \sim g, Y \sim h} \phi^2(Y|X) \frac{1 + \rho}{1 - \rho}.
\]

Therefore \( E(a^m_n)^2 \to 0 \) as \( m \to \infty \) and \( n \to \infty \) (at an arbitrary rate). Consequently, \( a^m_n \to_p 0 \) and we have proved asymptotic negligibility of (6.12) (i.e. that this term is \( o_p(1/\sqrt{m}) \)).
The last step is bounding (6.11). We are to show that

$$B_n^m = \sqrt{m} \sum_{i=1}^{n} \left[ \nabla Z_m(X_i, \theta_\star) \right] - \nabla Z_m(X_i, \theta_\star)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - Z_m(X_i, \theta_\star))}{Z_m(X_i, \theta_\star)} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \psi(Y^k|X_i)$$

goes to 0 in probability. As in the previous part of the proof, we will consider a single component $b_n^m$ of vector $B_n^m$. By Cauchy-Schwarz inequality,

$$(6.17) \quad |b_n^m| \leq \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left[ \frac{Z_m(X_i, \theta_\star) - 1}{Z_m^2(X_i, \theta_\star)} \right]} \sqrt{\frac{1}{m} \sum_{i=1}^{n} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \psi(Y^k|X_i)}$$

By Assumption 7(a) we obtain that for arbitrary $\varepsilon > 0, \eta > 0$ and sufficiently large $m$ with probability at least $1-\eta$ for every $x \in \mathcal{X}$

$$1 - \varepsilon \leq Z_m(x, \theta_\star) \leq 1 + \varepsilon.$$ 

Therefore, the term under the first square root in (6.17) tends in probability to 0, because with probability at least $1-\eta$

$$\frac{1}{n} \sum_{i=1}^{n} \left[ \frac{Z_m(X_i, \theta_\star) - 1}{Z_m^2(X_i, \theta_\star)} \right] \leq \sup_{x \in \mathcal{X}} \left[ \frac{Z_m(x, \theta_\star) - 1}{Z_m^2(x, \theta_\star)} \right] \leq \frac{\varepsilon^2}{(1-\varepsilon)^2},$$

if $m$ is sufficiently large. To show that the second square root in (6.17) is bounded in probability we use Markov’s inequality and proceed similarly to bounding (6.15). We can apply Lemma 6.8, because $E_{Y \sim h} \psi(Y|x) = 0$ for each $x$ and we obtain

$$\mathbb{E} \frac{1}{n} \sum_{i=1}^{n} \left| \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \psi(Y^k|X_i) \right|^2 \leq \left\| \frac{g}{h} \right\|_{\infty} \mathbb{E}_{X \sim g, Y \sim h} \psi^2(Y|X) \frac{1 + \rho}{1 - \rho}.$$ 

It follows that $b_n^m \to_P 0$ and this ends the proof. \hfill \Box

Now we prove Lemma 6.8.
Proof of Lemma 6.8. We consider Hilbert space $L^2_{\pi}$ of functions $\phi: Y \to \mathbb{R}$ with finite norm $\int \phi(y)^2 \pi(dy) = \|\phi\|_{L^2_{\pi}}^2$. The transition kernel $P(y, \cdot)$ of a Markov chain $(Y^k)_{k \geq 1}$ is associated with linear operator $P$ defined by $P\phi(y) = \int_Y \phi(z) P(y, dz)$. We also define an operator $\Pi$ by $\Pi \phi(y) = \int_Y \phi(z) \pi(dz)$. We assume that the Markov chain is reversible and geometrically ergodic, which is equivalent to $\|P - \Pi\|_{L^2_{\pi}} = \rho$, where $\|\cdot\|_{L^2_{\pi}}$ is the operator norm and $1 - \rho > 0$ is the spectral gap [20].

Below, $E_\nu$ denotes the expectation with respect to the Markov chain with the initial distribution $\nu$. We start with the following observation:

$$
E_\nu[\phi(Y^k) \phi(Y^l)] = E_\nu[E(\phi(Y^k) \phi(Y^l)|Y^k)]
= E_\nu[\phi(Y^k) E(\phi(Y^l)|Y^k)]
= E_\nu[\phi(Y^k) P^{l-k} \phi(Y^k)].
$$

Using the fact that $\pi$ is a stationary distribution, we obtain

$$
|E_\nu[\phi(Y^k) P^{l-k} \phi(Y^k)]| = \left| \int \phi(y) P^{l-k} \phi(y) \int P^k(z, dy) \nu(dz) \right|
= \left| \int \phi(y) P^{l-k} \phi(y) \int P^k(z, dy) \frac{d\nu}{d\pi}(z) \pi(dz) \right|
\leq \left\| \frac{d\nu}{d\pi} \right\|_\infty \int |\phi(y) P^{l-k} \phi(y)| \int P^k(z, dy) \pi(dz)
= \left\| \frac{d\nu}{d\pi} \right\|_\infty \int |\phi(y) P^{l-k} \phi(y)| \pi(dy).
$$

Therefore, we obtain that

$$
|E_\nu[\phi(Y^k) \phi(Y^l)]| \leq \left\| \frac{d\nu}{d\pi} \right\|_\infty E_\pi |\phi(Y) P^{l-k} \phi(Y)|.
$$

Moreover, using the fact that $\Pi \phi = 0$, the Cauchy-Schwarz inequality, the definition of the operator norm and the property that $P^k - \Pi = (P - \Pi)^k$ for $k \geq 1$ we obtain

$$
E_\pi |\phi(Y) P^{l-k} \phi(Y)| = \int |\phi(y) (P^{l-k} - \Pi) \phi(y)| \pi(dy)
\leq \|\phi\|_{\pi} \|(P^{l-k} - \Pi) \phi\|_{\pi}
\leq \|\phi\|_{\pi}^2 \|P^{l-k} - \Pi\|_{L^2_{\pi}}
\leq \|\phi\|_{\pi}^2 \|P - \Pi\|_{L^2_{\pi}}^{l-k} = \|\phi\|_{\pi}^2 \rho^{l-k},
$$

25
which finishes the proof. □

Acknowledgments

We would like to thank the action editor and reviewers for their comments, that have improved the paper.

References


