Gerhard Winkler:

An Adaptive Gradient Algorithm for Maximum Likelihood Estimation in Imaging: A Tutorial

Sonderforschungsbereich 386, Paper 120 (1998)

Online unter: http://epub.ub.uni-muenchen.de/

An Adaptive Gradient Algorithm for Maximum Likelihood Estimation in Imaging: A Tutorial

Gerhard Winkler

Institute of Biomathematics and Biometry
GSF - National Research Center for Environment and Health
D-85764 Neuherberg, Germany
e-mail: gwinkler@gsf.de

Keywords: adaptive algorithm, stochastic approximation, stochastic gradient descent, MCMC methods, maximum likelihood, Gibbs fields, imaging

Abstract

Markov random fields serve as natural models for patterns or textures with random fluctuations at small scale. Given a general form of such fields each class of pattern corresponds to a collection of model parameters which critically determines the abilitity of algorithms to segment or classify. Statistical inference on parameters is based on (dependent) data given by a portion of patterns inside some observation window. Unfortunately, the corresponding maximum likelihood estimators are computationally intractable by classical methods. Until recently, they even were regarded as intractable at all. In recent years stochastic gradient algorithms for their computation were proposed and studied. An attractive class of such algorithms are those derived from adaptive algorithms, wellknown in engeneering for a long time.

We derive convergence theorems following closely the lines proposed by M. Métivier and P. Priouret (1987). This allows a transparent (albeit somewhat technical) treatment. The results are weaker than those obtained by L. Younes (1988).

1 Introduction

Markov random fields serve as flexible models in image analysis, speech recognition and many other fields. In particular, textures with random fluctuations at small scale are reasonably described by random fields. A large class of recursive neural networks can be reinterpreted in this framework as well.

Let a pattern be represented by a finite rectangular array $x = (x_s)_{s \in S}$ of 'greyvalues' or 'colours' $x_s \in G_s$ in 'pixels' $s \in S$ where all sets G_s and S are finite. A (finite) *random field* is a strictly positive probability measure Π on the

(finite) space $\mathbf{X} = \prod_{s\in S} G_s$ of all configurations x. Taking logarithms shows that Π is of the *Gibbsian form*

$$\Pi(x) = Z^{-1}\exp(-K(x)),\ Z = \sum_z \exp(-K(z)), \tag{1}$$

with an *energy function* K on $\mathbf{X}$. 'Modelling' a certain type of pattern or texture amounts to the choice of a random field, typical samples of which share sufficiently many statistical properties with samples from the real pattern. Hence the choice of K usually is based on statistical inference besides prior knowledge. A nonparametric approach is not feasible and we restrict attention to the (linear) parametric case. We consider families

$$\mathbf{\Pi} = \{\Pi(\cdot;\vartheta) : \vartheta \in \Theta\},$$

of random fields on $\mathbf{X}$, where $\Theta \subset \mathsf{R}^d$ is the parameter space and each distribution is a Gibbs field of the *exponential form*

$$\Pi(\cdot;\vartheta) = Z(\vartheta)^{-1}\exp(\langle\vartheta, H\rangle), \vartheta \in \Theta.$$

The energy is given by $K_\vartheta = -\langle\vartheta, H\rangle$ where $H = (H_1, \ldots, H_d)$ is a vector of functions on $\mathbf{X}$, $\vartheta \in \Theta \subset \mathsf{R}^d$, and $\langle\vartheta, H\rangle$ is the Euclidean inner product.

Given a sample $x \in \mathbf{X}$, a *maximum likelihood estimator* $\hat{\vartheta}(x)$ maximizes the *(log-) likelihood function*

$$L(x,\cdot) : \Theta \longrightarrow \mathsf{R},\ \vartheta \longmapsto \ln\Pi(x;\vartheta).$$

The covariance of H_i and H_j under $\Pi(\cdot;\vartheta)$ will be denoted by $\mathsf{cov}(H_i, H_j;\vartheta)$ and the corresponding covariance matrix by $\mathsf{cov}(H;\vartheta)$. Straightforward calculations give ([15], Prop. 13.2.1)

Proposition 1.1 *Let Θ be open. The likelihood function $\vartheta \mapsto L(x;\vartheta)$ is infinitely often continuously differentiable for every x. The gradient is given by*

$$\frac{\partial}{\partial\vartheta_i}L(x;\vartheta) = H_i(x) - \mathsf{E}(H_i;\vartheta) \tag{2}$$

and the Hessean matrix is given by

$$\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}L(x;\vartheta) = -\mathsf{cov}(H_i, H_j;\vartheta).$$

In particular, the likelihood function is concave.

This result tells us that direct computation of ML estimators in the present context is not possible. In fact, the expectation in the gradient is a sum over $\mathbf{X}$ which may have cardinality of order $256^{256\times 256}$. Because of the mentioned

misgivings, ML estimators on large spaces until recently were thought to be computationally intractable. Therefore, J. Besag in [2], [3] suggested the coding and the pseudo-likelihood method where the full likelihood function is replaced by 'pseudo-likelihoods' based on conditional probabilities only. These estimators are computationally feasible in many cases (cf. [15] and also [9]).

In the last decade, accompanied by the development of learning algorithms for Neural Networks and encouraged by the increase of computer power, recursive algorithms for the computation or at least apprioximation of maximum likelihood estimators themselves were studied. Many of them are related to basic gradient ascent (which is ill-famed for poor convergence). More sophisticated methods from numerical analysis violate the requirement of 'locality' which basically means that the updates can be computed component by component from the respective preceding components. In this paper, we study the asymptotics of adaptive algorithms which we hasten to define now.

We want to compute maximum likelihood estimators for Gibbs fields, i.e. maximize the likelihood function $W = L(x;\cdot)$ for a fixed sample $x \in \mathbf{X}$. The starting point is steepest ascent

$$\vartheta_{(k+1)} = \vartheta_{(k)} + \gamma_{k+1}\nabla W(\vartheta_{(k)}) = \vartheta_{(k)} + \gamma_{k+1}\Big(H(x) - \mathsf{E}(H;\vartheta_{(k)})\Big) \tag{3}$$

with *gains* γ_k (possibly varying in time). Given $\vartheta_{(0)} \in \mathsf{R}^d$, the *adaptive algorithm* is recursively defined by

$$\begin{aligned} \vartheta_{(k+1)} &= \vartheta_{(k)} + \gamma_{k+1}\left(H(x) - H(\xi_{k+1})\right) \\ \mathsf{P}\left(\xi_{k+1} = z \,|\xi_k = x\right) &= P_k(x,z) \end{aligned} \tag{4}$$

where P_k is the Markov kernel of one sweep of the Gibbs sampler for $\Pi(\cdot;\vartheta_{(k)})$. The Gibbs sampler is a Markov process on $\mathbf{X}$ which via a law of large numbers gives estimates of the expectations appearing in (3). It will be defined below. Note that $(\xi_k, \vartheta_{(k)})_{k\geq 0}$ is a Markov process taking values in $\mathbf{X} \times \mathsf{R}^d$ (and living on a suitable probability space $(\Omega, \mathcal{F}, \mathsf{P})$). We shall be mainly interested in the marginal process $(\vartheta_{(k)})_{k\geq 0}$.

Let us briefly comment on the philosophy behind. Consider the ordinary differential equation (ODE)

$$\dot\theta(t) = \nabla W(\theta(t)),\ t \geq 0,\ \theta(0) = \vartheta_{(0)}, \tag{5}$$

where $\dot\theta = d\theta/dt$. Under mild assumptions on W, each of these initial value problems has a unique solution $(\theta(t))_{t\geq 0}$ and $\theta(t) \to \vartheta_*$ as $t \to \infty$ (cf. Proposition 6.1). Hence a process $\left(\vartheta_{(k)}\right)$ converges to ϑ_* if it stays near a solution. Steepest ascent can be interpreted as an Euler method for the discrete approximation of solutions of the ODE. Similarly, the paths of (4) will be compared to the solutions of (5).

Coupling the Gibbs sampler and an ascent algorithm like in (4) amounts to adaptive algorithms which play an important role in fields of engineering like system identification, signal modelling, adaptive filtering and others. They received considerable interest in recent years and were studied by MÉTIVIER and PRIOURET (1987), [13], in a general framework. The circle of such ideas is illustrated, surveyed and extended in the monograph BENVENISTE, MÉTIVIER and PRIOURET (1990), [1], an extended English version of the French predecessor from (1987). The monograph FREIDLIN and WENTZELL (1984), [6], had considerable influence on the development of the theory.

The theory of adaptive algorithms was applied to ML estimation in imaging by L. YOUNES (1988) - (1989), [16], [17], [18]. In some respects, the theory gets simpler in this setting due to the boundedness of the energy function K. On the other hand, some assumptions from the general theory are not met and therefore additional estimates are required. YOUNES (1988), [17], proves almost sure convergence developing a heavy technical machinery. We decided to steer a middle course: we shall follow the lines of MÉTIVIER, PRIOURET (1987), [13], closely in order not to obscure the main ideas by too many technical details. On the other hand the results will be weaker than YOUNES'.

The reader we have in mind should be acquainted with general probability spaces and conditional expectations. He or she should also have met discrete-time, continuous-space Markov processes and martingales. Concerning martingale theory, part of the six pages [14], pp. 42-47, is sufficient. For more background information the reader may consult [15] and [7].

2 The Gibbs Sampler

To complete the definition of the algorithm (4) the Gibbs sampler is introduced now. Let Π be a Gibbs field of the form (1) and consider the Markov chain recursively defined by the rules:

1. Enumerate the sites in S, i.e. let $S = \{1, \dots, |S|\}$.
2. Choose an initial configuration $x^{(0)} \in \mathbf{X}$.
3. Given $x^{(k)}$, pick a greyvalue $y_1 \in G_1$ at random from $\Pi(X_1 = \cdot | X_j = x_j^{(k)}, j \neq 1)$. Given the configuration *updated in the last pixel* repeat this step for $s = 2, \dots, |S|$. Now a 'sweep' is finished with the result $x^{(k+1)}$.

The symbol $|S|$ denotes the number of elements in S, the enumeration of S is called a *deterministic visiting scheme*; the projections $\mathbf{X} \to G_j$, $x \mapsto x_j$ are denoted by X_j and $\Pi(A|B)$ is the conditional probability of A given B. Formally, the Gibbs sampler is a homogeneous Markov chain with transition probability

$$P(x, y) = \Pi_1 \dots \Pi_{|S|},$$

with the pixelwise transitions, called *local characteristics*, given by

$$\Pi_k(x,y) = \begin{cases} \Pi(X_k = y_k | X_j = x_j, j \neq k) & \text{if } y_{S\setminus\{k\}} = x_{S\setminus\{k\}} \\ \\ 0 & \text{otherwise.} \end{cases}$$

The conditional probabilites are easily computed:

$$\Pi(X_k = y_k | X_j = x_j, j \neq k) = Z_k^{-1} \exp(-K(y_k x_{S\setminus\{k\}}), \;\; Z_k = \sum_{z_k} \exp(K(z_k x_{S\setminus\{k\}}).$$

Here we adopted the notation from [15]: the symbol $y_k x_{S\setminus\{k\}}$ denotes the configuration in $\mathbf{X}$ with k^{th} component y_k and which equals x off k.

Let ξ_i denote the random configuration $x^{(i)}$ after the i^{th} sweep. The laws $\mathsf{P} \circ \xi_i^{-1}$ of the variables ξ_i approximate the unique invariant (and even reversible) distribution Π and and the process obeys the law of large numbers ([15], Thm. 5.1.4):

Theorem 2.1 *The Gibbs sampler fulfills*

$$\mathsf{P}(\xi_i = x) \longrightarrow \Pi(x), \; i \to \infty, \;\; \textit{for every } x \in \mathbf{X}.$$

and for every function f on $\mathbf{X}$ and every $\varepsilon > 0$, there is a constant c such that

$$\mathsf{P}\left(\left| \frac{1}{n} \sum_{i=0}^{n-1} f(\xi_i) - \mathsf{E}(f;\Pi) \right| > \varepsilon \right) \leq \frac{c}{n\varepsilon^2} \exp(|S|\Delta). \tag{6}$$

The constant Δ is the *maximal local oscillation* of the energy function K of Π given by

$$\Delta = \max\{|K(x) - K(y)| : x_{S\setminus\{s\}} = y_{S\setminus\{s\}}, s \in S\}$$

and $c = 13\|f\|^2$ for the L^1-norm $\|f\| = \sum_x |f(x)|$. Random visiting schemes are in use as well (and sometimes even preferable) but in the discussion below only deterministic ones will appear. The *maximal d-dimensional oscillation*

$$\bar{\Delta} = \max\left\{ \|H(x) - H(y)\|_2 : x, y \in \mathbf{X} \right\}. \tag{7}$$

will be needed too.

3 Main Results

The main results will be stated and discussed in this section. The proofs will be given later. Throughout the discussion it will be assumed that

$$1 \geq \gamma_1 \geq \gamma_2 \geq \ldots > 0, \;\; \sum_{k=1}^{\infty} \gamma_k = \infty. \tag{8}$$

This includes the case of constant gain $\gamma_k = \gamma$. Only Gibbs samplers with *deterministic visiting scheme*, as introduced above, will be adopted. Concerning the ODE (5) the assumptions will be:

Hypothesis 3.1 *(1)* $\Theta = \mathsf{R}^d$,
(2) $\mathsf{cov}(H;\vartheta)$ *is positive definite for each* $\vartheta \in \mathsf{R}^d$,
(3) the function W *attains its (unique) maximum at* $\vartheta_* \in \mathsf{R}^d$.

The last two assumptions are fulfilled with high probability, if the family of distributions in question is identifyable and the sample x is taken on a sufficiently large observation window S, by recent consistency results ([4], for a summary see [15] and [7]).

For a finite time horizon $T > 0$ let

$$n(T) = \min\{n \geq 0 : \gamma_1 + \ldots + \gamma_{n+1} \geq T\}.$$

We shall use the notation $t_n = \sum_{k=1}^{n} \gamma_k$. Let $(\vartheta_{(k)})$ and $(\theta(t))$ be given by (4) and (5), respectively. A weak approximation theorem can be stated as follows:

Theorem 3.2 *There are constants* C, D *and* L *such that for every* $T > 0$ *and* $\varepsilon > 0$

$$\begin{aligned} \mathsf{P}\left(\sup_{m \leq n(T)} \left\|\vartheta_{(m)} - \theta(t_m)\right\|_2 \geq \varepsilon\right) &\leq \frac{C}{\varepsilon^2}(1+T)\left(1+\mathrm{e}^{DT}\right)\mathrm{e}^{2LT}\sum_{k=1}^{n(T)}\gamma_k^2 \\ &\leq \frac{C}{\varepsilon^2}T(1+T)\left(1+\mathrm{e}^{DT}\right)\mathrm{e}^{2LT}\gamma_1. \end{aligned}$$

Theorem 3.2 generalizes results in Derevetskii and Fradkov (1974) for independent ξ_k. The dependent case was studied first in Ljung (1977) - (1978). Better bounds can be obtained tracking the constants more carefully than we shall do (cf. for example (27)).

Remark. The bound on the right hand side tends to 0 as γ_1 tends to 0. The constants depend continuously on $\|\vartheta_{(0)}\|_2$ by (19) below. Hence there are common constants for all $\vartheta_{(0)}$ in a given compact set $Q \subset \mathsf{R}^d$. Assume now $\sum_{k=1}^{\infty}\gamma_k^2 < \infty$ and suppose that at time r the algorithm is restarted in Q. The theorem applied to the process $\left(\xi_{r+k}, \vartheta_{(r+k)}\right)_{k\geq 0}$ with gains γ_{r+k} shows that the approximation gets better and better as r tends to infinity since $\sum_{k\geq r}\gamma_k^2$ tends to 0. Let now Ω_Q denote the set of those $\omega \in \Omega$ for which the path $\left(\vartheta_{(k)}(\omega)\right)_{k\geq 0}$ returns to Q again and again. The above observation can be used to prove almost sure convergence to ϑ_* on Ω_Q.

In the present case $\theta(t) \to \vartheta_*$ for every solution of the ODE (5). We give a more precise quantitative estimate. Let us introduce some notation before. Let $\lambda(\vartheta_*) < 0$ be the largest eigenvalue of $\mathsf{cov}(H;\vartheta_*)$,

$$M_3 = \sup\{|\partial_i\partial_j\partial_k W(\vartheta)| : i,j,k,= 1,\ldots,d; \vartheta \in \mathsf{R}^d, \|\vartheta - \vartheta_*\|_2 \leq 1\},$$

$$r = \min\left\{1, \frac{|\lambda(\vartheta_*)|}{2M_3 n^{5/2}}\right\}.$$

and finally

$$\gamma_\rho = \inf\{\|\nabla W(\vartheta)\|_2^2 : \|\vartheta - \vartheta_*\|_2^2 \geq \rho\}.$$

Then

Lemma 3.3 *Each initial value problem (5) has a unique solution* $(\theta(t))_{t\geq 0}$ *and* $\theta(t) \to \vartheta_*$ *as* $t \to \infty$. *Moreover,*

$$\|\theta(t) - \vartheta_*\|_2 \leq r \exp\left(-\frac{|\lambda(\vartheta_*)|}{2}(t-\tau)\right) \quad \textit{for } t \geq \tau$$

where $\tau = |W(\vartheta_0) - W(\vartheta_*)|/\gamma_r$.

All this is proved in the Appendix. By these results the following makes sense.

Corollary 3.4 *Given* $\varepsilon > 0$ *choose* $T > 0$ *such that* $\|\theta(T) - \vartheta_*\|_2 \leq \varepsilon$. *Then*

$$\mathsf{P}\left(\left\|\vartheta_{(n(T))} - \vartheta_*\right\|_2 > 2\varepsilon\right) \leq C(\gamma_1, T)$$

where $C(\gamma_1, T) \to 0$ *as* $\gamma_1 \to 0$.

For special choices of gains almost sure convergence holds.

Theorem 3.5 (YOUNES **(1988)**) *Let* $\gamma_k = (uk)^{-1}$, $u > 2\sigma\bar{\Delta}^2$. *Then*

$$\vartheta_{(n)} \longrightarrow \vartheta_* \ \mathsf{P} - \textit{almost everywhere.}$$

As already mentioned, the proof of this strong result is fairly technical and will not be given here. Nevertheless, the main idea is similar to that presented below.

The following informal argument might nourish the hope that the program can be carried out: We know that the Gibbs sampler as a time-homogeneous Markov process obeys the law of large numbers, i.e means w.r.t. $\Pi(\cdot;\vartheta)$ can be approximated by means in time (cf. Theorem 2.1). Let $\gamma > 0$ be a constant gain. Choose $r > 0$ and $n > 0$. Then

$$\begin{aligned}
\vartheta_{(r+n)} &= \vartheta_{(r)} + \gamma \sum_{k=0}^{n-1} \left(H(x) - H(\xi_{r+k+1})\right) \\
&= \vartheta_{(n)} + (n\gamma)\left(H(x) - \frac{1}{n}\sum_{k=0}^{n-1} H\left(\xi_{r+k+1}\right)\right) \\
&\approx \vartheta_{(r)} + (n\gamma)\nabla W(\vartheta_{(r)}).
\end{aligned}$$

The approximation holds if $(\xi_{r+k})_{k=0}^{n-1}$ is approximately stationary and n is large. For the former, the parameters $\vartheta_{(r+k)}$ should vary rather slowly which is the case for small gain and small n. The proper balance between the requirements 'n small' and 'n large' is one of the main problems in the proof.

4 Error Decomposition and L^2-Estimates

Algorithm (4) is interpreted as steepest ascent perturbed by additive noise. The noise term is decomposed into a sum of convenient terms by means of the Poisson equation and L^2-estimates of the single terms are derived. These are the basic ingredients for the proof of the main result.

4.1 Error Decomposition

Since (4) can be written in the form

$$\vartheta_{(k+1)} = \vartheta_{(k)} + \gamma_{k+1} \left(\nabla W \left(\vartheta_{(k)} \right) + \mathsf{E} \left(H; \vartheta_{(k)} \right) - H \left(\xi_{k+1} \right) \right) \tag{9}$$

it amounts to gradient ascent perturbed by (non-Gaussian, non-white) noise

$$g_{(k)} = g \left(\xi_{k+1}; \vartheta_{(k)} \right), \quad g(x;\vartheta) = \mathsf{E}(H;\vartheta) - H(x).$$

Control of the error term is based on a clever decomposition. It will be shown that g can be written in the form

$$g(\cdot;\vartheta) = f_\vartheta - P_\vartheta f_\vartheta \tag{10}$$

where P_ϑ is the Markov kernel of one sweep of the Gibbs sampler for $\Pi_\vartheta = \Pi(\cdot;\vartheta)$ and the maps f_ϑ fulfill $(I - P_\vartheta) f_\vartheta = g_\vartheta$ (ϑ will be written as a subscript if convenient). By (10) the error takes the form

$$g_{(k)} = f_{(k)} \left(\xi_{k+1} \right) - P_k f_{(k)} \left(\xi_{k+1} \right)$$

where $f_{(k)} = f_{\vartheta_{(k)}}$ etc.. The cumulated error in $\vartheta_{(k+1)}$ can be decomposed into four terms:

$$\begin{aligned}
\mathcal{E}_{(n)} &= \sum_{k=0}^{n-1} \gamma_{k+1} \left(\mathsf{E} \left(H; \vartheta_{(k)} \right) - H \left(\xi_{k+1} \right) \right) \\
&= \sum_{k=0}^{n-1} \gamma_{k+1} \left(f_{(k)} \left(\xi_{k+1} \right) - P_k f_{(k)} \left(\xi_k \right) \right) \\
&+ \sum_{k=0}^{n-1} \gamma_{k+1} \left(P_k f_{(k)} \left(\xi_k \right) - P_{k-1} f_{(k-1)} \left(\xi_k \right) \right) \\
&+ \sum_{k=0}^{n-1} \left(\gamma_{k+1} - \gamma_k \right) P_{k-1} f_{(k-1)} \left(\xi_k \right) \\
&+ \gamma_1 P_0 f_{(0)} \left(\xi_0 \right) - \gamma_n P_{n-1} f_{(n-1)} \left(\xi_{n-1} \right).
\end{aligned} \tag{11}$$

These terms will be estimated separately in L^2. Before, the decomposition is justified. In the following proof and many estimates below the *contraction coefficient* $c(P)$ of a Markov kernel P on a finite space $\mathbf{X}$ will be used. It is given by

$$c(P) = \max_{x,y} \| P(x,\cdot) - P(y,\cdot) \| \tag{12}$$

where for probability measures μ and ν the *total variation* of their difference is

$$\|\mu - \nu\| = \sum_{x \in \mathbf{X}} |\mu(x) - \nu(x)|.$$

It fulfills the basic inequalities

$$\|\mu P - \nu P\| \leq \|\mu - \nu\| \cdot c(P), \;\; c(PQ) \leq c(P)c(Q) \tag{13}$$

(for all basic facts concerning contraction coefficients cf. [15]. Section 4.2). Now we can prove

Lemma 4.1 *Let Π be a random field, P a Markov kernel and g a function on $\mathbf{X}$. Suppose that $c(P) < 1$, $\Pi P = \Pi$ and $\mathsf{E}(g; \Pi) = 0$. Then there is a function f on $\mathbf{X}$ which solves the* Poisson equation

$$(I - P)f = g.$$

This is a standard result from the potential theory of Markov chains.

Proof. Define formally the *potential kernel* of P by $G = \sum_{k \geq 0} P^k$ and set

$$f(x) = Gg(x). \tag{14}$$

Plainly, $G = I + PG$, and if the infinite series (14) exists,

$$f - Pf = Gg - PGg = Gg - (Gg - g) = g$$

as desired. Due to the assumptions,

$$\begin{aligned} |Gg(x)| &= \left| \sum_{k \geq 0} P^k(x, \cdot)g - \Pi P^k g \right| \\ &\leq \sum_{k \geq 0} \left| \Pi \left(P^k(x, \cdot) - P^k \right) g \right| \\ &\leq 2\|g\|_\infty \sum_{k \geq 0} c(P)^k. \end{aligned} \tag{15}$$

Since $c(P) < 1$ the last series converges which completes the proof.

4.2 Preliminary Estimates

In order to derive the announced L^2-estimates, some preliminary estimates are needed. The d-dimensional oscillation $\bar{\Delta}$ was introduced in (7).

The first estimates are obvious:

$$\left\| \vartheta_{(k+1)} - \vartheta_{(k)} \right\|_2 \leq \bar{\Delta}\gamma_{k+1} \tag{16}$$

$$\|\vartheta_{(n)}\|_2 \leq \|\vartheta_{(0)}\|_2 + \bar{\Delta} \sum_{k=1}^{n} \gamma_k \tag{17}$$

$$\|g_\vartheta\|_2 \leq \bar{\Delta} \tag{18}$$

It is easily seen that $c(P_\vartheta) \le 1-\exp(\sigma\bar{\Delta}\|\vartheta\|_2)$ ([15], (5.4)), and hence by (17),

$$(1-c(P_k))^{-1} \le \exp\left(\sigma\bar{\Delta}\|\vartheta_{(0)}\|_2\right)\exp\left(\sigma\bar{\Delta}^2 t_k\right) \le C\exp(Dt_k). \tag{19}$$

The following estimates are less obvious.

Lemma 4.2 *There are constants C and D such that*

$$\left\|f_{(k)}\right\|_2 \le C\exp(Dt_k) \tag{20}$$

$$\left\|P_{k+1}f_{(k+1)} - P_k f_{(k)}\right\|_2 \le C\exp(Dt_{k+1})\left\|\vartheta_{(k+1)} - \vartheta_{(k)}\right\|_2 \tag{21}$$

Proof. By (15) and (18),

$$\|f_\vartheta\|_2 \le 2\bar{\Delta}(1-c(P_\vartheta))^{-1}.$$

Hence (19) implies

$$\|f_{(k)}\|_2 \le 2\bar{\Delta}C\exp(Dt_k).$$

The proof of (21) is technical and lengthy and hence will be postponed to Section 5, Lemma 5.3. For the moment, take it for granted.

Let us finally note the simple but useful relation

$$\left\|\sum_{j=1}^p a_j\phi_j\right\|_2^2 \le \left(\sum_{j=1}^p a_j\right)\sum_{j=1}^p a_j\,\|\phi_j\|_2^2 \tag{22}$$

for $a_j \ge 0$ and $\phi_j \in \mathsf{R}^d$. If all a_j vanish there is nothing to show; otherwise it amounts to a modified definition of convexity.

4.3 L^2-Estimates

L^2-estimates for the four sums in (11) will be derived now. The first one is most interesting. Set

$$S_{(n)} = \sum_{k=0}^{n-1}\gamma_{k+1}\left(f_{(k)}\left(\xi_{k+1}\right) - P_k f_{(k)}\left(\xi_k\right)\right).$$

Lemma 4.3 *There are constants C and D such that*

$$\mathsf{E}\left(\max_{m\le n}\left\|S_{(m)}\right)\|_2^2\right) \le C\sum_{k=0}^{n-1}\exp(Dt_k)\gamma_{k+1}^2.$$

Proof. First we shall show that $S = \left(S_{(n)}\right)_{n\ge 0}$ is a martingale. To this end, let $\mathcal{F}_n$ denote the σ-field generated by $\xi_1,\ldots,\xi_n$. Note that $\vartheta_{(1)},\ldots,\vartheta_{(n)}$ are $\mathcal{F}_n$-measurable as well. By construction of the process,

$$\mathsf{E}\left(f_{(k)}\left(\xi_{k+1}\right)|\mathcal{F}_k\right) = P_k f_{(k)}\left(\xi_k\right).$$

Hence the term in $S_{(n)}$ with index $k = n-1$ vanishes conditioned on $\mathcal{F}_n$. The other summands are $\mathcal{F}_n$-measurable and hence invariant under conditioning. This proves the martingale property

$$\mathsf{E}\left(S_{(n)}\,|\mathcal{F}_n\right) = S_{(n-1)}.$$

By Jensen's inequality,

$$\begin{aligned}
&\mathsf{E}\left(\left\|f_{(k)}\left(\xi_{k+1}\right)\right\|_2^2\right) = \mathsf{E}\left(\mathsf{E}\left(\left\|f_{(k)}\left(\xi_{k+1}\right)\right\|_2^2|\mathcal{F}_k\right)\right)\\
\geq\ &\mathsf{E}\left(\mathsf{E}\left(\left\|f_{(k)}\left(\xi_{k+1}\right)\right\|_2|\mathcal{F}_k\right)^2\right) = \mathsf{E}\left(\left\|P_k f_{(k)}\left(\xi_k\right)\right\|_2^2\right).
\end{aligned}$$

By orthogonality of increments ([14], Lemma 3.1.1),

$$\mathsf{E}\left(\left|S_{(n)}\right|^2\right) = \mathsf{E}\left(\sum_{k=0}^{n-1}\gamma_{k+1}^2\left\|f_{(k)}(\xi_{k+1}) - P_k f_{(k)}(\xi_k)\right\|_2^2\right).$$

By (22), the previous estimate, and (20), one may proceed with

$$\mathsf{E}\left(\|S_{(n)}\|_2^2\right) \leq 2\mathsf{E}\left(\sum_{k=0}^{n-1}\gamma_{k+1}^2\|f_{(k)}(\xi_{(k+1)}\|_2^2\right) \leq 2C\sum_{k=0}^{n-1}\gamma_{k+1}^2\exp(2Dt_k).$$

The uniform estimate in $m \leq n$ finally follows from Doob's L^2-inequality ([14], Lemma 3.1.4):

$$\mathsf{E}\left(\max_{m\leq n}\|S_{(m)}\|_2^2\right) = 4\mathsf{E}\left(\|S_{(n)}\|_2^2\right).$$

This completes the proof.

The remaining three estimates are straightforward.

Lemma 4.4 *There are constants C and D such that*

$$\begin{aligned}
a_k = \left\|\gamma_{k+1}\left(P_k f_{(k)}(\xi_k) - P_{k-1}f_{k-1}(\xi_k)\right)\right\|_2 &\leq C\exp(Dt_k)\gamma_k^2\\
b_k = \left\|(\gamma_{k+1}-\gamma_k)P_{k-1}f_{(k-1)}(\xi_k)\right\|_2 &\leq C\exp(Dt_{k-1})(\gamma_{k+1}-\gamma_k)\\
c_k = \left\|\gamma_1 P_0 f_{(0)}(\xi_0) - \gamma_n P_{n-1}f_{(n-1)}(\xi_{n-1})\right\|_2 &\leq C\exp(2Dt_{n-1})\left(\gamma_n^2-\gamma_1^2\right)
\end{aligned}$$

Proof. By (21) and (17),

$$\begin{aligned}
&\left\|\gamma_{k+1}(P_k f_{(k)}(\xi_k) - P_{k-1}f_{(k-1)}(\xi_k)\right\|_2\\
\leq\ &\gamma_{k+1}C\|\vartheta_{(k)} - \vartheta_{(k-1)}\|_2\exp(Dt_k) \leq C'\gamma_k^2\exp(Dt_k)
\end{aligned}$$

which proves the first estimate. The second one follows from (20). The third one is implied by (20) and (22) with $p = 2$, $a_j = 1$.

Now we can put things together to derive the L^2-estimate of the total error.

Proposition 4.5 *There are constants C and D such that*

$$\mathsf{E}\left(\max_{m\le n}\|\mathcal{E}_{(m)}\|_2^2\right) \le C\exp(Dt_{n-1})\left(1+\sum_{k=1}^{n}\gamma_k^2\right)\left(\sum_{k=1}^{n}\gamma_k^2\right).$$

Proof. We shall use (11) and Lemmata 4.3 and 4.4. By (22) for $p=4$ and $a_j=1$,

$$\begin{aligned}&\mathsf{E}\left(\max_{m\le n}\|\mathcal{E}_{(m)}\|_2^2\right)\\ \le\ & 4\left(\mathsf{E}\left(\max_{m\le n}\|S_{(m)}\|_2^2\right)+\left(\sum_{k=1}^{n-1}a_k\right)^2+\left(\sum_{k=1}^{n-1}b_k\right)^2+c_n\right).\end{aligned}$$

Plainly,

$$\begin{aligned}\left(\sum_{k=1}^{n-1}a_k\right)^2 &\le C\left(\sum_{k=1}^{n-1}\gamma_k^2\exp(Dt_k)\right)^2 \le \exp(2Dt_{n-1})\left(\sum_{k=1}^{n-1}\gamma_k^2\right)\\ \left(\sum_{k=1}^{n-1}b_k\right)^2 &\le C\left(\sum_{k=1}^{n-1}(\gamma_{k+1}-\gamma_k)\exp(Dt_{k-1})\right)^2\\ &\le C\exp(2Dt_{n-2})(\gamma_n-\gamma_1)^2 \le C\exp(2Dt_{n-2})\gamma_1^2.\end{aligned}$$

Summation now gives the desired result.

For a finite time horizon the estimate boils down to

Corollary 4.6 *There are constants C and D such that for every $T>0$,*

$$\mathsf{E}\left(\max_{m\le n(T)}\|\mathcal{E}_{(m)}\|_2^2\right) \le C\mathrm{e}^{DT}(1+\gamma_1 T)\sum_{k=1}^{n(T)}\gamma_k^2.$$

5 Proof of the Approximation Theorem

We complete now the proof of Theorem 3.2 and append the missing estimates. The main tool is the following discrete Gronwall lemma.

Lemma 5.1 *If the real sequence $(b_k)_{k\ge 0}$ satisfies*

$$b_0=0,\ b_r \le C+D\sum_{k=1}^{r}\gamma_k b_{k-1}\ \ for\ \ r=1,\dots,n,$$

with $C,D\ge 0$ then

$$b_n \le C\exp\left(D\sum_{k=1}^{n}\gamma_k\right).$$

Proof. If C or D vanishes there is nothing to show. Hence we may assume that $D = 1$ (otherwise we modify the γ_k). First we show

$$1 + \sum_{k=1}^{r} \gamma_k \exp\left(\sum_{j=1}^{k-1} \gamma_j\right) \leq \exp\left(\sum_{k=1}^{r} \gamma_k\right), \ \ r = 1, \ldots, n. \tag{23}$$

For $r = 1$ this boils down to $1+\gamma_1 \leq \exp(\gamma_1)$ which plainly is true. The induction step reads

$$\begin{aligned}
\exp\left(\sum_{k=1}^{r+1} \gamma_k\right) &= \exp\left(\sum_{k=1}^{r} \gamma_k\right) \exp(\gamma_{r+1}) \\
&\geq \exp\left(\sum_{k=1}^{r} \gamma_k\right) + \gamma_{r+1} \exp\left(\sum_{k=1}^{r} \gamma_k\right) \\
&\geq 1 + \left(\sum_{k=1}^{r} \gamma_k \exp\left(\sum_{j=1}^{k-1} \gamma_j\right)\right) + \gamma_{r+1} \exp\left(\sum_{k=1}^{r} \gamma_k\right) \\
&= 1 + \left(\sum_{k=1}^{r+1} \gamma_k \exp\left(\sum_{j=1}^{k-1} \gamma_j\right)\right).
\end{aligned}$$

The first inequality follows from $\exp(x) \geq 1 - x$ and the second one from the induction hypothesis.

Since $b_1 \leq C$ the assertion holds for $r = 1$. If it holds for all $k \leq r$ then using (23), the assumption and the induction hypothesis

$$b_{r+1} \leq C\left(1 + \sum_{k=1}^{r+1} \gamma_k \exp\left(\sum_{j=1}^{k-1} \gamma_j\right)\right) \leq C \exp\left(\sum_{k=1}^{r+1} \gamma_k\right)$$

and the desired inequality is verified.

Solutions of the ODE (5) and steepest ascent will be compared now. Since

$$|(H_i - \mathsf{E}(H_i))| \leq \bar{\Delta}$$

the following estimates hold:

$$\|\nabla W(\vartheta)\|_2 \leq \bar{\Delta}, \ \ |\mathsf{cov}_\vartheta(H_i, H_j)| \leq \bar{\Delta}^2. \tag{24}$$

For a smooth map $t \mapsto \vartheta(t)$ the chain rule reads

$$\frac{d}{dt}\frac{\partial}{\partial \vartheta_i} W(\vartheta(t)) = \sum_{j=1}^{d} \frac{\partial^2}{\partial \vartheta_i \partial \vartheta_j} W(\vartheta(t)) \frac{d}{dt}\vartheta_j(t). \tag{25}$$

This will be used to prove:

Lemma 5.2 *There are a constant C and maps $\alpha_{(k)}$ such that*

$$\begin{aligned}\theta(t_{k+1}) - \theta(t_k) &= \gamma_{k+1}\nabla W(\theta(t_k)) + \alpha_{(k)} \\ \|\alpha_{(k)}\|_2 &\leq C\gamma_{k+1}^2.\end{aligned}$$

Proof. If $\theta(t)$ solves (5) then

$$\theta(t_{k+1}) - \theta(t_k) = \int_{t_k}^{t_{k+1}} \nabla W(\theta(t))\, dt. \tag{26}$$

By the mean value theorem

$$\frac{\partial}{\partial\vartheta_i}W(\theta(t)) = \frac{\partial}{\partial\vartheta_i}W(\theta(t_k)) + \frac{d}{dt}\frac{\partial}{\partial\vartheta_i}W(\theta(s_i(t)))(t - t_k)$$

for some $s_i(t) \in [t_k, t]$. By (25) applied to $t \mapsto \theta(t)$ and since $\dot\theta(t) = \nabla W(\theta(t))$ the identity holds with

$$\alpha_{(k),i} = \int \sum_{j=1}^{d} \frac{\partial^2}{\partial_i\partial_j}W(\theta(s_i(t)))\nabla W(s_i(t))(t - t_k)\, dt.$$

Since $|t - t_k| \leq \gamma_{k+1}$ if $t_k \leq t \leq_{k+1}$ and by (24),

$$|\alpha_{(k),i}| \leq d\bar\Delta^3\gamma_{k+1}^2$$

which implies the inequality.

These and the previous L^2-estimates are combined with the Gronwall lemma to prove the main result.

Proof of Theorem 3.2. Let $(\theta_t)_{t\geq 0}$ be a solution of (5). By (9) and Lemma 5.2,

$$\begin{aligned}\vartheta_{(k)} &= \vartheta_{(k-1)} + \gamma_k\nabla W(\vartheta_{(k-1)}) + \gamma_k g_{(k-1)} \\ \theta(t_k) &= \theta(t_{(k-1)}) + \gamma_k\nabla W(\theta(t_{k-1})) + \alpha_{(k-1)}.\end{aligned}$$

Hence

$$\vartheta_{(n)} - \theta(t_n) = \sum_{k=0}^{n-1}\gamma_{k+1}(\nabla W(\vartheta_{(k)}) - \nabla W(\theta(t_k))) + \mathcal{E}_{(n)} - \sum_{k=0}^{n-1}\alpha_{(k)}.$$

By Lemmata 5.2 and 5.1,

$$\begin{aligned}\|\vartheta_{(n)} - \theta(t_n)\|_2 &\leq L\sum_{k=0}^{n-1}\gamma_{k+1}\|\vartheta_{(k)} - \theta(t_k)\|_2 + \|\mathcal{E}_{(n)}\|_2 + C\sum_{k=0}^{n-1}\gamma_{k+1}^2 \\ &\leq \left(\|\mathcal{E}_{(n)}\|_2 + C\sum_{k=0}^{n-1}\gamma_{k+1}^2\right)\exp\left(L\sum_{k=1}^{n}\gamma_k\right).\end{aligned}$$

If $t_n \le T$ then

$$\|\vartheta_{(n)} - \theta(t_n)\|_2^2 \le 2\left(\|\mathcal{E}_{(n)}\|_2^2 + C^2\left(\sum_{k=0}^{n-1}\gamma_{k+1}^2\right)^2\right)\exp(2LT).$$

By (22),

$$\left(\sum_{k=1}^{n(T)}\gamma_k^2\right)^2 \le \sum_{k=1}^{n(T)}\gamma_k \sum_{k=1}^{n(T)}\gamma_k^3 \le \gamma_1 T\sum_{k=1}^{n(T)}\gamma_k^2.$$

Together with Proposition 4.5 this implies

$$\mathsf{E}\left(\max_{m\le n(T)}\|\vartheta_{(m)} - \theta(t_m)\|_2^2\right) \le \{2C\exp(DT)(1+\gamma_1 T)+2C^2\gamma_1 T\}\exp(2LT)\sum_{k=1}^{n(T)}\gamma_k^2.$$

Since $\gamma_1 \le 1$ the last quantity is dominated by an expression of the form

$$C(1+T)(1+\exp(DT))\exp(2LT)\sum_{k=1}^{n(T)}\gamma_k^2. \tag{27}$$

Application of Markov's inequality now completes the proof.

Finally, the estimate (21) is verified.

Lemma 5.3 *There are constants C and D such that*

$$\begin{aligned}\left|\frac{\partial}{\partial\vartheta_j}f_{\vartheta,i}\right| &\le C\exp(2D\|\vartheta\|_2)\\ \|P_{k+1}f_{(k+1)} - P_k f_{(k)}\|_2 &\le C\exp(2Dt_{k+1})\|\vartheta_{(k+1)} - \vartheta_{(k)}\|_2.\end{aligned}$$

Proof. Existence of partial derivatives and the first estimate will be proved simultaneously. By (14), $f_\vartheta = \sum_{k\ge 0}P_\vartheta^k g_\vartheta$. Since

$$\mathsf{E}_\vartheta(g_\vartheta) = 0,\ \Pi_\vartheta P_\vartheta^k g_\vartheta, \tag{28}$$

one has

$$f_\vartheta(x) = \sum_{k\ge 0}\sum_{x',y}\Pi_\vartheta(x')\left(P_\vartheta^k(x,y) - P_\vartheta^k(x',y)\right)g_i(y) =: \sum_{k\ge 0}S_\vartheta^{(k)}(x).$$

Since ϑ and k will be fixed for a while, they will be dropped.

(1) The chain rule gives

$$\begin{aligned}\frac{\partial}{\partial\vartheta_j}S_i^{(k)} &= \sum_{x',y}\frac{\partial}{\partial\vartheta_j}\Pi(x')\left(P^k(x,y) - P^k(x',y)\right)g_i(y)\\ &+ \sum_{x',y}\Pi(x')\left(\frac{\partial}{\partial\vartheta_j}P^k(x,y) - \frac{\partial}{\partial\vartheta_j}P^k(x',y)\right)g_i(y)\\ &+ \sum_{x',y}\Pi(x')\left(P^k(x,y) - P^k(x',y)\right)\frac{\partial}{\partial\vartheta_j}g_i(y)\\ &= : R^{(1)} + R^{(2)} + R^{(3)}.\end{aligned} \tag{29}$$

It is easy to see that

$$\left|R^{(1)}\right| + \left|R^{(3)}\right| \leq 2(\bar{\Delta}^2 + L)c(P)^k; \tag{30}$$

in fact,

$$\frac{\partial}{\partial \vartheta_j}\Pi(x') = \Pi(x')\ln\Pi(x')$$

and hence by Proposition 1.1,

$$\left|\frac{\partial}{\partial \vartheta_j}\Pi(x')\right| \leq \bar{\Delta}$$

which proves the estimate of $R^{(1)}$. Concerning $R^{(3)}$ use Proposition 1.1 and (24).

(2) Estimating $R^{(2)}$ is cumbersome and tricky. First the partial derivatives have to be computed. For $s \in S$ and $x, y \in \mathbf{X}$ the local characteristic is given by

$$\Pi_{\{s\}}(x,z) = \mathbf{1}_{x_{S\setminus\{s\}}=z_{S\setminus\{s\}}}(z)\Pi(z_s|z_{S\setminus\{s\}}).$$

Proposition 1.1 for conditional Gibbs fields yields

$$\begin{aligned}\frac{\partial}{\partial \vartheta_j}\Pi\left(z_s|z_{S\setminus\{s\}}\right) &= \left(\frac{\partial}{\partial \vartheta_j}\ln\Pi(z_s|z_{S\setminus\{s\}})\right)\Pi(z_s|z_{S\setminus\{s\}} \\ &= (H_j(z) - \mathsf{E}(H_j|z_{S\setminus\{s\}}))\Pi_s(z_s|z_{S\setminus\{s\}}) \\ &= :\phi_s(z)\Pi(z_s|z_{S\setminus\{s\}}) \end{aligned} \tag{31}$$

which implies

$$\frac{\partial}{\partial \vartheta_j}\Pi_{\{s\}}(x,z) = \phi_s(z)\Pi_{\{s\}}(x,z).$$

In one sweep $z \in \mathbf{X}$ is reached from x with positive probability along precisely one path $z_0 = x, z_1, \ldots, z_\sigma = z$ and hence

$$P(x,z) = \prod_{s=1}^{\sigma}\Pi_{\{s\}}(z_{s-1}, z_s)$$

where it is tacitly assumed that $S = \{1, \ldots, \sigma\}$ and that the sites are visited in increasing order. By the product rule

$$\frac{\partial}{\partial \vartheta_j}P(x,z) = P(x,z)\sum_{s=1}^{\sigma}\phi_s(z_s) =: P(x,z)\phi(x,z)$$

where

$$|\phi(x,z)| \leq \sigma\bar{\Delta}. \tag{32}$$

Repeating this argument for

$$P^k(x,y) = \sum_{z_1,\ldots,z_{k-1}} P(x,z_1)\cdot\ldots\cdot P(z_{k-1},y),\ k \geq 1,$$

gives

$$\frac{\partial}{\partial\vartheta_j}P^k(x,y)=\sum_{z_1,\dots,z_{k-1}}\sum_{r=1}^{k}\phi(z_{r-1},z_r)P(x,z_1)\cdot\ldots\cdot P(z_{\sigma-1},y)$$

where $z_0=x$ and $z_\sigma=y$. Rearranging summation gives

$$\frac{\partial}{\partial\vartheta_j}P^k(x,y)=\sum_{r=1}^{k}\sum_{u,v}\phi(u,v)P(u,v)P^{r-1}(x,u)P^{k-r}(v,y). \tag{33}$$

(3) Plugging (33) into $R^{(2)}$ results in

$$R^{(2)}=\sum_{r=1}^{k}\sum_{x',u,v}\Pi(x')\left(P^{r-1}(x,u)-P^{r-1}(x',u)\right)P(u,v)\phi(u,v)\cdot\sum_{y}P^{r-k}(v,y)g_i(y).$$

By (28) the last sum $\sum_y\dots$ can be replaced by

$$\sum_{w,y}\Pi(w)\left(P^{k-r}(v,y)-P^{k-r}(w,y)\right)g_i(y)$$

which can be estimated by $2\|g_i\|_\infty c(P)^{k-1}$. Similarly, the remaining term is bounded by $2\sigma\bar{\Delta}kc(P)^{r-1}$. In summary,

$$\left|R^{(2)}\right|\le Ckc(P)^{k-1}. \tag{34}$$

Note finally that $R^{(2)}=0$ if $k=0$.

(4) Putting (29), (30) and (34) together yields

$$\begin{aligned}\left|\sum_{k=0}^{\infty}\frac{\partial}{\partial\vartheta_j}S^{(k)}_{\vartheta,i}\right| &\le C\left(\sum_{k=0}^{\infty}c(P_\vartheta)^k+\sum_{k=0}^{\infty}(k+1)c(P_\vartheta)^k\right)\\ &= C\left((1-c(P_\vartheta))^{-1}+(1-c(P_\vartheta))^{-2}\right)\\ &\le 2C\left(1-c(P_\vartheta)\right)^{-2}\le 2C\exp\left(2D\|\vartheta\|_2\right)\end{aligned}$$

where (19) was used in the last line. This shows that derivatives of partial sums converge uniformly on every compact subset of R^d; hence differentiation and summation may be interchanged and the first inequality holds.

(5) For the second inequality, use the triangle inequality

$$\begin{aligned}&\left\|P_{n+1}f_{(n+1)}(y)-P_nf_{(n)}(y)\right\|\\ \le\ &\left\|\left(P_{n+1}-P_n\right)f_{(n+1)}(y)\right\|_2+\left\|P_n\left(f_{(n+1)}-f_{(n)}\right)(y)\right\|_2\\ =\ &: A+B.\end{aligned}$$

Setting $\psi(s)=\vartheta_{(n)}+s\left(\vartheta_{(n+1)}-\vartheta_{(n)}\right)$, (32) implies

$$\begin{aligned}&|P_{n+1}(x,y)-P_n(x,y)|=\left|\int_0^1\frac{d}{dt}P_{\psi(s)}(x,y)\,ds\right|\\ =\ &\left|\int_0^1\langle\vartheta_{(n+1)}-\vartheta_{(n)},\nabla P_{\psi(s)}(x,y)\rangle\,ds\right|\le\left\|\vartheta_{(n+1)}-\vartheta_{(n)}\right\|_2\sqrt{d}\sigma\bar{\Delta}.\end{aligned}$$

Hence by (20),

$$A \leq C \exp(Dt_n) \left\| \vartheta_{(n+1)} - \vartheta_{(n)} \right\|_2 .$$

By (17), both $\vartheta_{(n)}$ and $\vartheta_{(n+1)}$ are contained in a ball B of radius $C' + \bar{\Delta} t_{n+1}$ around $0 \in \mathsf{R}^d$ and by convexity $\psi(s)$, $s \in [0,1]$ as well. Hence the first inequality implies

$$\begin{aligned} B &\leq \left\| \left(f_{(n+1)} - f_{(n)} \right)(y) \right\|_2 = \left| \int_0^1 \langle \vartheta_{(n+1)} - \vartheta_{(n)}, \nabla f_{\psi(s)} \, ds \right| \\ &\leq \left\| \vartheta_{(n+1)} - \vartheta_{(n)} \right\|_2 C \exp(2Dt_{n+1}) . \end{aligned}$$

The estimates of A and B imply the second inequality.

6 Appendix: How Close is $\theta(t)$ to ϑ_*?

It is shown now that each of the initial value problems (5) has a unique solution and that each solution converges to the (unique) maximum likelihood estimator. Moreover estimates are given for the time one has to wait until a solution enters a given neighbourhood of ϑ_*. This yields an estimate of the finite time horizon T in Theorem 3.2 needed to guarantee a prescribed precision of the approximation of ϑ_* by the algorithm (4). This appendix is included for convenience of the reader only since all arguments are standard.

Let $\| \cdot \|_M$ denote the matrix norm.

Proposition 6.1 *Each initial value problem (5) has a unique solution $\theta(t), t \geq 0$, and $\theta(t) \to \vartheta_*$ as $t \to \infty$.*

Proof. By Proposition 1.1, the estimates (24) and (25) applied to the map $\vartheta(t) = \vartheta + t(\vartheta' - \vartheta)$, one has the estimate

$$\left\| \frac{d}{dt} \nabla W(\vartheta(t)) \right\|_2 \leq \| \mathsf{cov}(H; \vartheta(t)) \|_M \| \dot{\vartheta}(t) \|_2 \leq d \bar{\Delta}^2 \| \dot{\vartheta}(t) \|_2 . \tag{35}$$

Hence

$$\| \nabla W(\vartheta') - \nabla W(\vartheta) \|_2 \leq \int_0^1 \left\| \frac{d}{dt} \nabla W(\vartheta(t)) \right\|_2 dt \leq d \bar{\Delta}^2 \| \vartheta' - \vartheta \|_2$$

and thus the right hand side of (5) is Lipschitz continuous. In particular, each initial value problem (5) has a unique solution $\theta(t), t \geq 0$. Solutions of (5) converge to ϑ_* as $t \to \infty$ which follows from elementary stability theory: We have

$$\frac{d}{dt} W \circ \theta(t) = -|\nabla W(\theta(t))|^2$$

by the chain rule and (5). Hence W is a global Ljapunov function for the gradient system (5) and, moreover, ϑ_* is the only critical point. This completes the proof.

A rough estimate for the distance between the solution of the differential equation (5) and the limit ϑ_* will be derived now. It will first be stated in a general form. Consider the gradient system

$$\dot{\theta}(t) = -\nabla V(\theta(t)),\ \theta(0) = \vartheta_0. \tag{36}$$

Throughout the discussion the following hypothesis will be assumed:

Hypothesis 6.2 *(1)* $V \in \mathcal{C}^3(\mathsf{R}^n, \mathsf{R})$, $V(0) = 0$, $V(\vartheta) > 0$ *if* $\vartheta \neq 0$.
(2) The Hessean matrix C *of* V *is positive definite at* $\vartheta = 0$.
(3) For each $\rho > 0$, $\gamma_\rho = \inf\{\|\nabla V(\vartheta)\|_2^2 : \|\vartheta\|_2^2 \geq \rho\} > 0$.

Note that $0 \in \mathsf{R}^d$ now plays the role of ϑ_* in the previous discussion. Hence the maximal solution θ of (36) is defined on $[0, \infty)$ and $\theta(t) \to 0$ as $t \to \infty$. Let

$$M_3 = \sup\left\{|\partial_i \partial_j \partial_k V(\vartheta)| : i, j, k = 1, \ldots, d; \vartheta \in \mathsf{R}^d, \|\vartheta\|_2 \leq 1\right\}$$

and let $\lambda(\vartheta)$ denote the smallest eigenvalue of $C(\vartheta)$. By (2) in Hypothesis 6.2 one has $\lambda(0) > 0$. Finally set

$$r = \min\left\{1, \frac{\lambda(0)}{2M_3 n^{5/2}}\right\}.$$

The following result yields the desired estimate for the distance between $\theta(t)$ and the optimal ϑ_*.

Theorem 6.3 *Assume that the above hypothesis hold. Let* $(\theta(t))$ *be the solution of (36) and*

$$\tau(\vartheta_0) = \frac{V(\vartheta_0)}{\gamma_r}.$$

Then the following holds:
(i) If $\|\theta(\tau_0)\|_2 \leq r$ *for some* $\tau_0 \in \mathsf{R}$, *then*

$$\|\theta(t)\|_2 \leq r \cdot \exp\left(-\frac{\lambda(0)}{2}(t - \tau_0)\right) \quad \textit{for } t \geq \tau_0.$$

(ii) In fact,

$$\|\theta(\tau(\vartheta_0))\|_2 \leq r.$$

The following is straightforward. For a $d \times m$-matrix A let $\|A\|_\infty = \max_{i,j} |A_{i,j}|$.

Lemma 6.4 *For any* $d \times m$*-matrix* A *and any orthogonal* $d \times d$*-matrix* T,

$$\|TA\|_\infty \leq \sqrt{d} \cdot \|A\|_\infty,\ \ \|AT\|_\infty \leq \sqrt{d} \cdot \|A\|_\infty.$$

The next Lemma is of Gerschgorin type. Let $\sigma(M)$ denote the set of eigenvalues (in the complex plane) of a complex $d \times d$-matrix M.

Lemma 6.5 *Let M be a complex $d \times d$-matrix. Then*

$$\sigma(M) \subset \bigcup_{i=1}^{d} M_{ii} + B(r_i)$$

where $B(r_i)$ is the closed ball in the complex plain with radius r_i and $r_i = \sum_{j \neq i} |M_{ij}|$.

Proof. Let λ be an eigenvalue of M with eigenvector $v \neq 0$. Let further i denote the index with $|v_i| = \max_j |v_j|$. Then $\sum_j M_{ij} v_j = \lambda v_i$ implies

$$|M_{ii} - \lambda||v_i| = \left| \sum_{j \neq i} M_{ij} v_j \right| \leq r_i |v_i|$$

and hence $|M_{ii} - \lambda| \leq r_i$. This completes the proof.

Next the Hessean matrix is estimated.

Lemma 6.6 *For $v, w \in \mathsf{R}^d$, $\|v\|_2 \leq 1$, $\|w\|_2 \leq 1$,*

$$|C_{ij}(v) - C_{ij}(w)| \leq \sqrt{n} M_3 \|v - w\|_2 \text{ for all } i, j.$$

Proof. This follows from the elementary computation

$$\left| \int_0^1 \langle \nabla C_{ij}(tv + (1-t)w), v - w \rangle \, dt \right| \leq \sup_{\|z\|_2 \leq 1} \|\nabla C_{ij}(z)\|_2 \|v - w\|_2$$

$$= \sup_{\|z\|_2 \leq 1} \sqrt{\sum_{k=1}^{d} (\partial_k \partial_i \partial_j W(z))^2} \|v - w\|_2 \leq \sqrt{n} M_3 \|v - w\|_2.$$

The last Lemma estimates eigenvalues.

Lemma 6.7 *Suppose that $\|\vartheta\|_2 \leq r$. Then*

$$\lambda(\vartheta) \geq \lambda(0)/2.$$

Proof. Let $\|\vartheta\|_2 \leq r$. By Lemma 6.6,

$$C(\vartheta) = C(0) + \tilde{A}, \ \|\tilde{A}\|_\infty \leq \sqrt{d} \cdot M_3 \cdot r.$$

There is an orthogonal matrix T such that $T^* C(0) T = D$, where T^* denotes the transpose of T and D is the diagonal matrix with the eigenvalues $\lambda_1, \ldots, \lambda_d$ as diagonal elements; we may assume $\lambda_1 = \lambda(0) > 0$. Further, $T^* C(\vartheta) T = D + T^* \tilde{A} T =: D + A$ and by Lemma 6.4,

$$\|A\|_\infty \leq \sqrt{d} \cdot \sqrt{d} \|\tilde{A}\|_\infty \leq n^{3/2} M_3 r.$$

Since $\sigma(C(\vartheta)) = \sigma(T^*C(\vartheta)T)$, Lemma 6.5 gives

$$\lambda(\vartheta) \in \bigcup_{i=1}^{n} (\lambda_i + A_{ii} + B(r_i)),$$

where $r_i = \sum_{j \neq i} |A_{ij}| \leq (d-1)\|A\|_\infty$. By the definition of r,

$$\begin{aligned} & \lambda(\vartheta) \geq \lambda_1 - |A_{11} - (d-1)\|A\|_\infty \geq \lambda_1 - d\|A\|_\infty \\ = \quad & \lambda(0) - n\|A\|_\infty \geq \lambda(0) - d^{5/2} M_3 r = \lambda(0)/2 \end{aligned}$$

which had to be shown.

Now the preparations for the proof of the Theorem are complete.

Proof of Theorem 6.3. For $\|v\| \leq r$ Lemma 6.7 yields the estimate

$$\begin{aligned} \langle v, \nabla V(v) \rangle &= \left\langle v, \int_0^1 C(tv)\, dt \cdot v \right\rangle = \int_0^1 \langle v, C(tv)v \rangle\, dt \\ &\geq \int \lambda(tv) \|v\|_2^2\, dt \geq \frac{\lambda(0)}{2} \|v\|_2^2. \end{aligned}$$

For the proof of *(i)* assume that $\|\theta(\tau_0)\|_2 \leq r$. The set

$$I = \{ s \in [\tau_0, \infty) : \|\theta(t)\| \leq r \text{ for all } t \in [\tau_0, s] \}$$

is a closed interval. We are going to prove $I = [\tau_0, \infty)$. If $\theta(\tau_0) = 0$ then the solution stays there and the assertion clearly holds. Otherwise $\theta(s) \neq 0$ for some $s \geq \tau_0$. Let $U(t) = \|\theta(t)\|_2^2$ for $t \geq \tau_0$. If $s \in I$ then

$$\begin{aligned} & \dot{U}(s) = \langle \theta(s), \dot{\theta}(s) \rangle = -\langle \theta(s), \nabla V(\theta(s)) \rangle \\ \leq \quad & -\frac{\lambda(0)}{2} \|\theta(s)\|_2^2 = -\lambda(0) U(s) < 0. \end{aligned}$$

Hence I is also open in $[\tau_0, \infty)$. Since $\tau_0 \in I$ this implies $I = [\tau_0, \infty)$. The estimate for $\dot{U}$ implies

$$U(s) \leq U(\tau_0) \exp(-\lambda(0)(s - \tau_0)) \text{ for } s \geq \tau_0,$$

and hence

$$\begin{aligned} & \|\theta(s)\|_2 \leq \|\theta(\tau_0) \exp\left(-\frac{\lambda(0)}{2}(s - \tau_0) \right) \\ \leq \quad & r \exp\left(-\frac{\lambda(0)}{2}(s - \tau_0) \right) \text{ for } s \geq \tau_0 \end{aligned}$$

which proves *(i)*.

For the proof of *(ii)* assume $\|\theta(\tau(\vartheta_0))\|_2 > r$. Then $\|\theta(s)\|_2 > r$ on $[0, \tau(\vartheta_0)]$ by part *(i)*. By the very definition of γ_r and (36),

$$\frac{d}{ds} V(\theta(s)) = -\|\nabla V(\theta(s))\|_2^2 \leq -\gamma_r$$

for $s \in [0, \tau(\vartheta_0)]$ and hence

$$V(\theta(\tau(\vartheta_0))) \leq V(\vartheta_0) - \gamma_r \tau(\vartheta_0) = 0$$

which contradicts $\theta(\tau(\vartheta_0)) \neq 0$ and the first hypothesis. We conclude $\|\theta(\tau(\vartheta_0))\|_2 \leq r$ and the proof is complete.

We thank B. Lani-Wayda, who pointed out to us the estimates in the Appendix.

References

[1] Benveniste A., Métivier M. and Priouret P. (1990): *Adaptive Algorithms and Stochastic Approximations.* Springer Verlag: Berlin Heidelberg New York London Paris Tokyo HongKong Barcelona

[2] Besag J. (1974): Spatial Interaction and the Statistical Analysis of Lattice Systems (with discussion). *J. of the Royal Statist. Soc.*, **B**, **36**, 192–236

[3] Besag J. (1977): Efficiency of Pseudolikelihood for Simple Gaussian Field. *Biometrika* **64**, 616–619

[4] Comets F. (1992): On Consistency of a Class of Estimators for Exponential Families of Markov random fields on the Lattice. *The Ann. of. Statist.* **20**, 455–486

[5] Derevitzkii D.P. and Fradkov A.L. (1974): Two Models for Analyzing the Dynamics of Adaption Algorithms. *Automation and Remote Control* **35** 1, 59-67

[6] Freidlin M.I. and Wentzell A.D. (1984): *Random Perturbations of Dynamical Systems.* Springer Verlag: Berlin Heidelberg New York

[7] X. Guyon (1995): *Random Fields on a Network.* Springer Verlag: New York, Berlin

[8] Guyon X. and Künsch H.R. (1992): Asymptotic Comparison of Estimators in the Ising Model. In: *Stochastic Models, Statistical methods and Algorithms in Image Analysis*, P. Barone, A. Frigessi, M. Piccioni, eds. Lecture Notes in Statistics 74, Springer Verlag, 177-198

[9] Jensen J.L. and Künsch H.R. (1993): On Asymptotic Normality of Pseudo Likelihood Estimates for Pairwise Interaction Processes. To appear in *Ann. Inst. Statist. Math.*

[10] Ljung L. (1977a): On Positive Real Transfer Functions and the Convergence of some Recursions. *IEEE Trans. on Automatic Control* **AC-22** 4, 539-551

[11] Ljung L. (1977b): Analysis of Recursive Stochastic Algorithms. *IEEE Trans. on Automatic Control* **AC-22** 4, 551-575

[12] Ljung L. (1978): Convergence of an Adaptive Filter Algorithm. *Int. J. Control* **27** 5, 673-693

[13] Métivier M. and Priouret P. (1987): Théorèmes de Convergence Presque Sure pour une Classe d'Algorithmes Stochastique à Pas dÉcroissant. *Probab. Th. Rel. Fields* **74**, 403–428

[14] Weizsäcker, H.v. and Winkler G. (1991): *Stochastic Integrals.* Friedr. Vieweg & Sohn: Braunschweig/Wiesbaden

[15] Winkler G. (1995): *Image Analysis, Random Fields and Dynamic Monte Carlo Methods. An Introduction to Mathematical Aspects.* Springer-Verlag: Berlin Heidelberg New York

[16] Younes L. (1988a): Estimation pour Champs de Gibbs et Application au Traitement d'Images. Université Paris Sud Thesis

[17] Younes L. (1988b): Estimation and Annealing for Gibbsian Fields. *Ann. Inst. Henri Poincare* **24**, No. 2, 269–294

[18] Younes L. (1989): Parametric Inference for Imperfectly Observed Gibbsian Fields. *Prob. Th. Rel. Fields* **82**, 625–645