Fast simulated annealing in ^{d} and an application to maximum likelihood estimation

Sylvain Rubenthaler, Tobias Rydén, Magnus Wiktorsson

Fast simulated annealing in $\mathbb{R}^d$ and an application to maximum likelihood estimation in state-space models

Sylvain RUBENTHALER[a],
Tobias RYDÉN[b] and Magnus WIKTORSSON[b]

12 September 2006

Abstract

Using classical simulated annealing to maximise a function ψ defined on a subset of $\mathbb{R}^d$, the probability $\mathbb{P}(\psi(\theta_n) \leq \psi_{\max} - \varepsilon)$ tends to zero at a logarithmic rate as n increases; here θ_n is the state in the n-th stage of the simulated annealing algorithm and $\psi_{\max}$ is the maximal value of ψ. We propose a modified scheme for which this probability is of order $n^{-1/3} \log n$, and hence vanishes at an algebraic rate. To obtain this faster rate, the exponentially decaying acceptance probability of classical simulated annealing is replaced by a more heavy-tailed function, and the system is cooled faster. We also show how the algorithm may be applied to functions that cannot be computed exactly but only approximated, and give an example of maximising the log-likelihood function for a state-space model.

Keywords : Central limit and other weak theorems, Computational methods in Markov chains, Sequential estimation, Markov processes with continuous parameter, Monte Carlo methods, Stochastic programming.

MSC : 60F05, 60J22, 60J25, 62L12, 65C05, 82C80, 90C15.

1 Introduction

Simulated annealing is a simulation-based approach to the problem of optimising a function. In the present paper we will be concerned with a real-valued function, ψ say, defined on a subset Θ of $\mathbb{R}^d$, and our aim is to *maximise* ψ. Thus we assume that ψ is bounded and that its supremum is attained at least at one point. Simulated annealing is designed to find the global maximum of ψ, even if ψ has local maxima. It has been extensively studied, see for instance Del Moral and Miclo (1999), Catoni (1999) and Cot and Catoni (1998) among many others, and Bartoli and Del Moral (2001) for an elementary introduction to

[a]Université de Nice - Sophia Antipolis, Laboratoire Dieudonné, Parc Valrose, 06108 Nice Cédex 02, France, rubentha@math.unice.fr

[b]Centre for Mathematical Sciences, Lund University, Box 118, 221 00 Lund, Sweden, tobias.ryden@matstat.lu.se or magnus.wiktorsson@matstat.lu.se. Both authors were supported by grants from the Swedish National Research Council.

the subject. The classical simulated annealing algorithm departs from a Markov transition kernel, which we denote by $K(\cdot,\cdot)$, on Θ, and a positive sequence $(\beta_n)_{n\geq 0}$ increasing to infinity. The sequence (β_n) is often referred to as an (inverse) *cooling schedule*, because $1/\beta_n$ is often interpreted as a temperature; this terminology originates from statistical physics. Then, starting from an initial point $\theta_0 \in \Theta$, a sequence $(\theta_n)_{n\geq 0}$ is constructed recursively as follows.

(a1) In stage n, given the current state θ_n, sample a new proposed position Z from $K(\theta_n,\cdot)$.

(a2) Set $\theta_{n+1} = Z$ with probability

$$\exp(-\beta_n(\psi(\theta_n) - \psi(Z))_+)$$

and $\theta_{n+1} = \theta_n$ otherwise.

Here $(\cdot)_+$ is the positive part. We notice that if $\psi(Z) \geq \psi(\theta_n)$, then the proposed new state Z is accepted with probability one. A proposal Z at which ψ is smaller than at the current θ_n may be accepted, but this becomes increasingly unlikely for large n since $\beta_n \to \infty$.

The basic idea of simulated annealing is as follows. The update rule above corresponds to a Markov transition kernel, K_β say, on Θ; cf. (2.6) below. Under additional assumptions including that K is positive recurrent and reversible with respect to its stationary distribution, γ say,

$$\gamma(dx)\, K(x,dy) = \gamma(dy)\, K(y,dx),$$

one can prove that for fixed β, the stationary distribution of K_β is absolutely continuous with respect to γ with Radon-Nikodym derivative proportional to $\exp\{\beta\psi(x)\}$ (cf. Catoni, 1999, Proposition 1.2, or Bartoli and Del Moral, 2001, p. 64). This indicates that as β increases, this stationary distribution becomes increasingly concentrated around the maxima of ψ. Now, in the beginning of the simulation scheme β_n is small (the temperature is high), and the particle θ_n is allowed to explore the space Θ rather freely. When the temperature cools down (β_n gets large), the particle is more and more lured to the regions where ψ is large and should in the limit end up at a maximum point of ψ.

Obviously, the kernel K and the sequence (β_n) are important design parameters of the algorithm. A typical choice for (β_n) is a logarithmic increase; $\beta_n = \beta_0 \log(n+e)$ for some $\beta_0 > 0$. We note that with this cooling schedule, the acceptance probability in (a2) above becomes

$$(n+e)^{-\beta_0(\psi(\theta_n)-\psi(Z))_+}. \tag{1.1}$$

Under additional regularity assumptions one can prove that for β_0 small enough and if ψ has a single global maximum, it holds that for all $\varepsilon > 0$,

$$\mathbb{P}(\psi(\theta_n) \leq \psi_{\max} - \varepsilon) \to 0 \quad \text{as } n \to \infty, \tag{1.2}$$

where $\psi_{\max} = \sup_{x\in\Theta} \psi(x)$. How fast is this convergence? In many works on simulated annealing the space Θ is assumed finite, and one may then let $\varepsilon \to 0$ and thus study $\mathbb{P}(\psi(\theta_n) < \psi_{\max})$. Typically this probability tends to zero at

an algebraic rate, see for instance Gielis and Maes (1999, Eq. (22)) (take f as the indicator function of non-optimal states) and references in this paper. For a continuous Θ the situation is different. If $\Theta \subset \mathbb{R}^d$ one can show (see Appendix A) that the rate of convergence in (1.2) is only logarithmic. Alternatively, one can prove that there are numbers C and C' such that for any $\varepsilon > 0$,

$$\mathbb{P}(\psi(\theta_n) \leq \psi_{\max} - \varepsilon) \leq C n^{-C'\varepsilon} \tag{1.3}$$

Thus, the algebraic rate becomes infinitely slow as $\varepsilon \to 0$. Locatelli (2001) proposed a refinement of the annealing scheme that reaches non-vanishing algebraic rates, but it requires knowledge of $\psi_{\max}$ which is an assumption we do not want to make.

In the present paper we propose a modified simulated annealing scheme such that for any $\varepsilon > 0$ there is a number C_ε such that

$$\mathbb{P}\left(\psi(\theta_n) \leq \psi_{\max} - \varepsilon\right) \leq C_\varepsilon n^{-1/3}(1 + \log n). \tag{1.4}$$

We will then say that the rate of convergence is $1/3$, up to a logarithmic term.

2 Description of the new simulated annealing scheme

Just as in classical simulated annealing, the proposed scheme departs from a Markov transition kernel K and a cooling schedule (β_n). The difference lies in that the exponential function of the classical algorithm's update is replaced by a different function, and that the cooling schedule is altered. More precisely, we let $g : \mathbb{R}^+ \to \mathbb{R}^+$ be a C^∞-function such that $g(0) = 1$, g is non-decreasing and $g(t) \to \infty$ as $t \to \infty$. We set $f = 1/g$ and suppose that f is convex and such that $\sup_{t \geq 0} |tf'(t)| < \infty$. Then the algorithm looks as follows.

(b1) In stage n, given the current state θ_n, sample a new proposed position Z from $K(\theta_n, \cdot)$.

(b2) Set $\theta_{n+1} = Z$ with probability

$$f(\beta_n(\psi(\theta_n) - \psi(Z))_+)$$

and $\theta_{n+1} = \theta_n$ otherwise.

In classical simulated annealing $g(t) = e^t$. In Section 3 we advocate the particular choice $g(t) = 1 + t/\tau$ for some $\tau > 0$, and thus $f(t) \sim \tau/t$ as $t \to \infty$. Compared to $f(t) \sim \exp(-t)$, this allows the algorithm to be 'more bold' in exploring regions far away from the current state. On the other hand we will let β_n be of order n^α with $\alpha = 1/3$, so that this sequence increases much faster than logarithmically. Together, these conditions imply (1.4). We also remark that with g as above and $\beta_n = n^{1/3}$, the acceptance probability in (b2) becomes

$$\frac{1}{1 + \frac{n^{1/3}}{\tau}(\psi(\theta_n) - \psi(Z))_+}, \tag{2.5}$$

which should be compared to (1.1); we see that (2.5) decays much slower as $\psi(\theta_n) - \psi(Z) \to \infty$, and thus again that the new algorithm is less likely to reject proposals with function values far below the current one.

Modifications of the acceptance function $f(t) = \exp(-t)$ of classical simulated annealing to speed up convergence rates have been discussed extensively in the statistical physics literature, and is there often referred to as 'fast simulated annealing'. The acceptance function $f(t) = 1/(1 + t/\tau)$ introduced above is similar to functions used in such papers; for instance, it corresponds to $\lambda = 1$ in Eq. (28) of Gielis and Maes (1999), and to $q_A = 2$ in Eq. (5) of Tsallis and Stariolo (1996). None of these authors obtained rate of convergence results for these schemes however. Tsallis and Stariolo (1996, Example 3) did obtain a convergence rate for $f(t) = 1/(1+t)^2$ and showed that this rate is indeed faster than for classical simulated annealing; the result however assumes that $\psi_{\max}$ is known and these authors worked exclusively on a finite set Θ.

We now return to the algorithm and define, for any $\beta > 0$ and $x, y \in \Theta$,

$$a_\beta(x, y) = f(\beta(\psi(x) - \psi(y))_+).$$

One step of the above algorithm is then described by a Markov transition kernel K_β defined as

$$K_\beta(x, dy) = a_\beta(x, y)K(x, dy) + \left(1 - \int_\Theta a_\beta(x, z)\, K(x, dz)\right) \delta_x(dy). \qquad (2.6)$$

Thus, assuming that the initial point θ_0 is random and drawn from some probability distribution η_0 on Θ, the sequence $(\theta_n)_{n\geq 0}$ is an inhomogeneous Markov chain with initial law η_0 and transition kernels $(K_{\beta_n})_{n\geq 0}$; more precisely, for any n, θ_{n+1} has conditional distribution $K_{\beta_n}(\theta_n, \cdot)$.

We will suppose that Θ is equipped with its Borel σ-field $\mathcal{B}(\Theta)$, and we will also assume that the Markov transition kernel K satisfies the following condition.

Hypothesis 1. *There exists $\varepsilon_K > 0$ and a probability measure λ on $(\Theta, \mathcal{B}(\Theta))$ such that*

$$\forall (x, A) \in \Theta \times \mathcal{B}(\Theta) :\ \varepsilon_K \lambda(A) \leq K(x, A) \leq \frac{1}{\varepsilon_K}\lambda(A).$$

Of course, Hypothesis 1 is easier to fulfil if Θ is compact or bounded.

Regarding the function ψ, we also make some assumptions. Put, for any $\varepsilon > 0$ and $a < b$,

$$U^{\varepsilon,a,b} = \{x \in \Theta : \psi_{\max} - b - \varepsilon < \psi(x) \leq \psi_{\max} - a - \varepsilon\}.$$

Hypothesis 2. *The oscillations of ψ are bounded, that is,*

$$\operatorname{osc}(\psi) := \sup_{x,y\in\Theta} |\psi(x) - \psi(y)| < \infty.$$

Hypothesis 3. *Either one of the following two assumptions holds true.*

(i) For all $\varepsilon > 0$ small enough there are numbers $C_0(\varepsilon) > 0$ and $\varepsilon' > 0$ such that for all $\delta > 0$,

$$\lambda(U^{\varepsilon,\delta i,\delta(i+1)}) \leq C_0\delta \quad or \quad i\delta \geq \varepsilon'.$$

(ii) The function ψ has a single global maximum, $\theta_{\max}$ say, located in the interior of Θ (which is thus non-empty). The probability measure λ is absolutely continuous with respect to Lebesgue measure and its density is locally bounded. The function ψ is C^3 in $\{x : \psi(x) > \psi_{\max} - \varepsilon''\}$, which is a neighbourhood of $\theta_{\max}$ (for some $\varepsilon'' > 0$), and the quadratic form $\psi''(\theta_{\max})$ is negative definite.

The attentive reader will notice that one could replace the assumption of a unique maximum by an assumption that there are a finite number of maxima, and that one could replace (ii) above by some more sophisticated assumptions on the derivatives of ψ. This requires a higher level of technicality but the whole proof would contain the same ideas and this is why we write the assumptions in this way.

3 Rate of convergence

Throughout the remainder of the paper we take $\beta_n = n^\alpha \vee 1$ for some $0 < \alpha < 1$. The choice of this particular sequence will be explained in Remark 3.8. We denote by $(\theta_n)_{n\geq 0}$ the sequence produced by the annealing scheme for this cooling schedule. The main result of the present section is the following.

Theorem 3.1. *Let $M = \operatorname{osc}(\psi) \vee \operatorname{osc}(\psi)^\alpha$ and suppose that $g(t^\alpha)/t \to 0$ as $t \to \infty$. Then for all $\varepsilon > 0$ small enough (if Hypothesis 3(i) holds) or $0 < \varepsilon \leq \varepsilon''$ (if Hypothesis 3(ii) holds), there exists a $C_\varepsilon > 0$ depending on ε such that for all n,*

$$\mathbb{P}(\psi(\theta_n) \leq \psi_{\max} - \varepsilon) \leq C_\varepsilon \left(\frac{g(M(n+1)^\alpha)^2}{n} + \frac{1}{n^\alpha} \left(1 + \int_0^{1+n^\alpha \operatorname{osc}(\psi)} f(t)\, dt \right) \right) + f(\varepsilon' n^\alpha)$$

under Hypothesis 3(i), or

$$\mathbb{P}(\psi(\theta_n) \leq \psi_{\max} - \varepsilon) \leq C_\varepsilon \left(\frac{g(M(n+1)^\alpha)^2}{n} + \frac{1}{n^\alpha} \int_0^{n^\alpha \operatorname{osc}(\psi)^2} f(t)\, dt \right) + f((\varepsilon'' - \varepsilon) n^\alpha)$$

under Hypothesis 3(ii).

Corollary 3.2. *Choosing $\alpha = 1/3$ and $g(t) = 1 + t/\tau$, where $\tau > 0$ is arbitrary, the bounds of Theorem 3.1 are $C_\varepsilon C n^{-1/3}(1 + \log n)$.*

Remark 3.3. *If we want to have terms of the same order in the bounds of Theorem 3.1, we see that $g(Mn^\alpha)^2/n$ and $f(\varepsilon' n^\alpha)$ (or $f((\varepsilon'' - \varepsilon)n^\alpha)$, depending on the case) should be of the same order. Thus $f(t)$ should be of order $t^{-1/3}$ as $t \to \infty$. With this choice all terms in the bound have the same order, and so there is something optimal to it. With our inequalities, it does not seem possible to have a better rate.*

Remark 3.4. *In Corollary 3.2 there is a parameter $\tau > 0$ which can be chosen arbitrarily. This parameter plays the role of a temperature like in classical simulated annealing and can be tuned by the user to optimise convergence. On the contrary to classical simulated annealing there is, theoretically, no restriction on τ.*

Before going into the proof of these results, we will proceed through some technical lemmas. First we however give some additional notation. The total variation distance $\|\mu-\nu\|_{\mathrm{TV}}$ between two probability measures μ and ν is defined as $\sup_A |\mu(A) - \nu(A)|$, where the supremum is taken over the σ-field on which the measures are defined. The set of probability measures on $(\Theta, \mathcal{B}(\Theta))$ will be denoted by $\mathcal{P}(\Theta)$.

Lemma 3.5. *For any $\beta > 0$ it holds that*

$$\forall (x, A) \in \Theta \times \mathcal{B}(\Theta) : \; K_\beta(x, A) \geq \varepsilon_K f(\beta \operatorname{osc}(\psi))\, \lambda(A).$$

Corollary 3.6. *The preceding lemma and Dobrushin's theorem (see Dobrushin, 1956, or Del Moral and Guionnet, 2001) imply that for any $\beta > 0$ and any probability measures μ and ν on Θ,*

$$\|\mu K_\beta - \nu K_\beta\|_{\mathrm{TV}} \leq (1 - \varepsilon_K f(\beta \operatorname{osc}(\psi)))\|\mu - \nu\|_{\mathrm{TV}}.$$

Proof of Lemma 3.5. Take $\beta > 0$ and $(x, A) \in \Theta \times \mathcal{B}(\Theta)$. Then

$$\begin{aligned} K_\beta(x, A) &\geq \int_A a_\beta(x, y)\, K(x, dy) \\ &\geq \int_A f(\beta \operatorname{osc}(\psi))\varepsilon_K\, \lambda(dy) \\ &= \varepsilon_K f(\beta \operatorname{osc}(\psi))\lambda(A). \end{aligned}$$

□

The above corollary implies that for any μ, the sequence $(\mu K_\beta^n)_{n\geq 0}$ is a Cauchy sequence in total variation norm. Thus there exists a total variation limit (cf. Lindvall, 2002, p. 232), which we denote by μ_β. This probability measure is invariant for K_β, and it does not depend on the particular choice of the initial distribution μ. It is hence the unique invariant distribution of K_β.

The convergence of simulated annealing hinges on the fact that the law of θ_n, which we denote by η_n, is close to μ_{β_n}, and that for large β_n the measure μ_{β_n} is concentrated on the regions where ψ is large. This concentration is the subject of the next lemma. We set

$$U^\varepsilon = \{x \in \Theta : \psi(x) > \psi_{\max} - \varepsilon\},$$

and $U^{\varepsilon,c}$ is its complement in Θ.

Lemma 3.7. *For all $\beta > 0$ and $\varepsilon > 0$ small enough (if Hypothesis 3(i) holds) or $0 < \varepsilon \leq \varepsilon''$ (if Hypothesis 3(ii) holds), there is a constant C_ε depending on ε*

such that

$$\mu_\beta(U^{\varepsilon,c}) \le \frac{C_\varepsilon}{\beta}\left(1+\int_0^{1+\beta\,\mathrm{osc}(\psi)} f(t)\,dt\right) + f(\beta\varepsilon') \quad \textit{under Hypothesis 3(i),}$$

$$\mu_\beta(U^{\varepsilon,c}) \le \frac{C_\varepsilon}{\beta}\int_0^{\beta\,\mathrm{osc}(\psi)^2} f(t)\,dt + f(\beta(\varepsilon''-\varepsilon)) \qquad \textit{under Hypothesis 3(ii).}$$

Proof. Fix $\beta > 0$ and ε in the appropriate range. We have

$$\begin{aligned}\mu_\beta(U^{\varepsilon,c}) &= \mu_\beta K_\beta(U^{\varepsilon,c})\\ &= \iint_{x\in U^{\varepsilon,c},y\in U^{\varepsilon,c}} \mu_\beta(dx)\,K_\beta(x,dy)\\ &\qquad + \iint_{x\in U^{\varepsilon},y\in U^{\varepsilon,c}} \mu_\beta(dx)\,K_\beta(x,dy).\end{aligned}$$

For $x \in U^{\varepsilon,c}$ and $y \in U^\varepsilon$, $K_\beta(x,dy) = K(x,dy)$. Thus the first integral above can be bounded as

$$\begin{aligned}\iint_{x\in U^{\varepsilon,c},y\in U^{\varepsilon,c}} \mu_\beta(dx)\,K_\beta(x,dy) &= \int_{x\in U^{\varepsilon,c}} \mu_\beta(dx)\left(1-\int_{y\in U^\varepsilon} K(x,dy)\right)\\ &\le \int_{x\in U^{\varepsilon,c}} \mu_\beta(dx)\left(1-\int_{y\in U^\varepsilon} \varepsilon_K\,\lambda(dy)\right)\\ &= (1-\varepsilon_K\lambda(U^\varepsilon))\mu_\beta(U^{\varepsilon,c}).\end{aligned}$$

Similarly, for the second integral,

$$\begin{aligned}&\iint_{x\in U^{\varepsilon},y\in U^{\varepsilon,c}} \mu_\beta(dx)\,K_\beta(x,dy)\\ &\quad= \iint_{x\in U^{\varepsilon},y\in U^{\varepsilon,c}} \mu_\beta(dx)a_\beta(x,y)\,K(x,dy)\\ &\quad\le \iint_{x\in U^{\varepsilon},y\in U^{\varepsilon,c}} \mu_\beta(dx)a_\beta(x,y)\frac{1}{\varepsilon_K}\,\lambda(dy)\\ &\quad\le \frac{1}{\varepsilon_K}\int_{y\in U^{\varepsilon,c}} f(\beta(\psi_{\max}-\varepsilon-\psi(y)))\,\lambda(dy),\end{aligned}$$

so that

$$\mu_\beta(U^{\varepsilon,c}) \le \frac{1}{\varepsilon_K^2\lambda(U^\varepsilon)}\int_{y\in U^{\varepsilon,c}} f(\beta(\psi_{\max}-\varepsilon-\psi(y)))\,\lambda(dy).$$

To finish the proof we will now bound the above integral as in the statement of the lemma. If Hypothesis 3(i) holds, take $\delta = 1/\beta$ and proceed as

$$\begin{aligned}\int_{y\in U^{\varepsilon,c}} f(\beta(\psi_{\max}-\varepsilon-\psi(y)))\,\lambda(dy) &\le \sum_{i=0}^{\mathrm{osc}(\psi)/\delta} f(i)\lambda(U^{\varepsilon,\delta i,\delta(i+1)})\\ &\le \sum_{i=0}^{\mathrm{osc}(\psi)/\delta} C_0(\varepsilon)\delta f(i) + \sum_{i\ge\varepsilon'/\delta} f(i)\lambda(U^{\varepsilon,\delta i,\delta(i+1)})\\ &\le C_0(\varepsilon)\delta\left(f(0)+\int_0^{1+\mathrm{osc}(\psi)/\delta} f(t)\,dt\right) + f\left(\frac{\varepsilon'}{\delta}\right).\end{aligned}$$

If Hypothesis 3(ii) holds we employ Morse's lemma (see e.g. Berger and Gostiaux, 1988, Theorem 4.2.12) to make a change of variables in $\{x : \psi(x) > \psi_{\max} - \varepsilon''\}$ such that with some bounded function ξ (that only depends on ψ),

$$\begin{aligned}
&\int_{y\in U^{\varepsilon,c}} f(\beta(\psi_{\max} - \varepsilon - \psi(y)))\,\lambda(dy) \\
&\qquad \le \int_{\psi(y)>\psi_{\max}-\varepsilon'', y\in U^{\varepsilon,c}} f(\beta(\psi_{\max} - \varepsilon - \psi(y)))\,\lambda(dy) \\
&\qquad\quad + \int_{\psi(y)\le\psi_{\max}-\varepsilon''} f(\beta(\varepsilon'' - \varepsilon))\,\lambda(dy) \\
&\qquad \le \int_{\sqrt{\varepsilon}}^{\operatorname{osc}(\psi)} f(\beta(t^2-\varepsilon))\xi(t)\,dt + f(\beta(\varepsilon''-\varepsilon)) \\
&\qquad \le \frac{\|\xi\|_\infty}{2\sqrt{\varepsilon}\beta}\int_0^{\beta\operatorname{osc}(\psi)^2} f(u)\,du + f(\beta(\varepsilon''-\varepsilon)),
\end{aligned}$$

after a change of variable $u = \beta(t^2 - \varepsilon)$. □

Remark 3.8. *In the following we will show that η_n is close to μ_{β_n}. Using Lemma 3.7 to bound $\mu_{\beta_n}(U^{\varepsilon,c})$, we obtain a bound larger than $1/\beta_n$. We would like to compare $\eta_n(U^{\varepsilon,c})$ to a power of n, so it is natural at this point to take, for some $\alpha > 0$,*

$$\beta_n = n^\alpha \vee 1.$$

For technical reasons appearing in the proof of Theorem 3.1, we need to take $\alpha < 1$.

The law η_n approaches μ_{β_n} which becomes increasingly concentrated on regions where ψ is large, but at the same time β_n is changing. The following lemma serves us to bound the distance between μ_β and $\mu_{\beta'}$.

Lemma 3.9. *With $C = (1/\varepsilon_K)\sup_{t\ge 0}|tf'(t)|$ it holds that for any $\beta' > \beta > 0$,*

$$\|\mu_\beta - \mu_{\beta'}\|_{\mathrm{TV}} \le C g(\beta\operatorname{osc}(\psi))\left(\frac{\beta'}{\beta} - 1\right).$$

Proof. We have, using Corollary 3.6,

$$\begin{aligned}
\|\mu_\beta - \mu_{\beta'}\|_{\mathrm{TV}} &\le \|\mu_\beta K_\beta - \mu_{\beta'}K_\beta\|_{\mathrm{TV}} + \|\mu_{\beta'}K_\beta - \mu_{\beta'}K_{\beta'}\|_{\mathrm{TV}} \\
&\le (1 - \varepsilon_K f(\beta\operatorname{osc}(\psi)))\|\mu_\beta - \mu_{\beta'}\|_{\mathrm{TV}} + \sup_{\mu\in\mathcal{P}(\Theta)}\|\mu K_\beta - \mu K_{\beta'}\|_{\mathrm{TV}}.
\end{aligned}$$

Pick $\mu \in \mathcal{P}(\Theta)$. We may construct two coupled samples from μK_β and $\mu K_{\beta'}$ respectively by first sampling x from μ, then sampling y from $K(x,\cdot)$, sampling U from the uniform distribution on $(0,1)$ and finally accepting the proposal y if $U \le \alpha_\beta(x,y)$ or $U \le \alpha_{\beta'}(x,y)$ respectively. Similarly to Appendix B we may

then conclude that

$$\begin{aligned}\|\mu K_\beta - \mu K_{\beta'}\|_{\mathrm{TV}} &\le \sup_{x,y\in\Theta} |a_\beta(x,y) - a_{\beta'}(x,y)| \\ &\le \sup_{0\le u\le \mathrm{osc}(\psi)} |f(\beta u) - f(\beta' u)| \\ &\le \sup_{0\le u\le \mathrm{osc}(\psi)} (|f'(\xi u)|\beta u)\left(\frac{\beta'}{\beta} - 1\right),\end{aligned}$$

where $\xi = \xi(u)$ is a point between β and β'. Since f is assumed convex and non-increasing, and hence $|f'|$ non-increasing, it holds that $|f'(\xi u)| \le |f'(\beta u)|$. We thus arrive at the bound

$$\|\mu_\beta - \mu_{\beta'}\|_{\mathrm{TV}} \le \frac{1}{\varepsilon_K f(\beta\,\mathrm{osc}(\psi))} \sup_{0\le u\le \mathrm{osc}(\psi)} (|f'(\beta u)|\beta u)\left(\frac{\beta'}{\beta} - 1\right).$$

Since $|tf'(t)|$ is assumed bounded, the proof is complete. □

Proof of Theorem 3.1. Set $\Delta_n = \|\eta_n - \mu_{\beta_n}\|_{\mathrm{TV}}$. If we can prove the inequality

$$\Delta_n \le C\frac{g(M(n+1)^\alpha)^2}{n}, \tag{3.7}$$

the result will follow from Lemma 3.7 and, in case of Hypothesis 3(i), the bound $g \ge 1$.

In order to prove (3.7) the assumption $g(x^\alpha)/x \to 0$ as $x \to \infty$ will be instrumental. We start by deriving a recursive bound for Δ_n. By Corollary 3.6 and Lemma 3.9 we have, for all n,

$$\begin{aligned}\Delta_{n+1} &\le \|\eta_n K_{\beta_n} - \mu_{\beta_n} K_{\beta_n}\|_{\mathrm{TV}} + \|\mu_{\beta_n} - \mu_{\beta_{n+1}}\|_{\mathrm{TV}} \\ &\le (1 - \varepsilon_K f(\beta_n\,\mathrm{osc}(\psi)))\|\eta_n - \mu_{\beta_n}\|_{\mathrm{TV}} + C\left(\frac{\beta_{n+1}}{\beta_n} - 1\right) g(n^\alpha\,\mathrm{osc}(\psi)) \\ &\le (1 - \varepsilon_K f(\beta_n\,\mathrm{osc}(\psi)))\Delta_n + C\frac{g(n^\alpha\,\mathrm{osc}(\psi))}{n+1}.\end{aligned}$$

Iterating this recursion yields

$$\begin{aligned}\Delta_{n+1} \le \sum_{q=1}^{n} \prod_{k=q+1}^{n} (1 - \varepsilon_K f(\beta_k\,\mathrm{osc}(\psi))) \times C\frac{g(q^\alpha\,\mathrm{osc}(\psi))}{q+1} \\ + \prod_{k=1}^{n}(1 - \varepsilon_K f(\beta_k\,\mathrm{osc}(\psi))) \times \|\eta_1 - \mu_{\beta_1}\|_{\mathrm{TV}},\end{aligned}$$

where an empty product (when $q = n$) is interpreted as unity.

Define F such that $F'(x) = f(x^\alpha)$. Then for $1 \le q \le n-1$,

$$\begin{aligned}\log \prod_{k=q+1}^{n} (1 - \varepsilon_K f(\beta_k\,\mathrm{osc}(\psi))) &\le -\sum_{k=q+1}^{n} \varepsilon_K f(\mathrm{osc}(\psi)k^\alpha) \\ &\le -\varepsilon_K \int_{q+1}^{n+1} f(\mathrm{osc}(\psi)x^\alpha)\,dx \\ &= -\frac{\varepsilon_K}{\mathrm{osc}(\psi)}(F(\mathrm{osc}(\psi)(n+1)) - F(\mathrm{osc}(\psi)(q+1))).\end{aligned}$$

For $q = n$ this is an equality. Putting $C_1 = \varepsilon_K / \operatorname{osc}(\psi)$ and $C_2 = \operatorname{osc}(\psi)$ we thus obtain

$$\begin{aligned}\Delta_{n+1} &\le e^{-C_1F(C_2(n+1))} \sum_{q=1}^{n} e^{C_1F(C_2(q+1))} C \frac{g(C_2 q^\alpha)}{q+1} \\ &+ 2e^{-C_1(F(C_2(n+1))-F(C_2))} \\ &\le C e^{-C_1F(C_2(n+1))} \int_1^{n+1} e^{C_1F(C_2(x+1))} \frac{g(C_2 x^\alpha)}{x}\, dx \\ &+ 2e^{-C_1(F(C_2(n+1))-F(C_2))}.\end{aligned}$$

Denote the integral on the right-hand side by I_{n+1}. First we notice that since $g \ge 1$, $I_{n+1} \to \infty$ as $n \to \infty$. Next we rewrite this integral as

$$I_{n+1} = \int_1^{n+1} e^{C_1F(C_2(x+1))} C_1C_2 f(C_2^\alpha (x+1)^\alpha) \times \frac{g(C_2^\alpha (x+1)^\alpha)}{C_1C_2} \frac{g(C_2 x^\alpha)}{x}\, dx,$$

where $C_1C_2 f(C_2^\alpha (x+1)^\alpha)$ is the derivate of the exponent. By partial integration, integrating the first factor of the integrand above and dropping all negative contributions (recall that g and g' are non-negative), we obtain the bound

$$\begin{aligned}I_{n+1} &\le \left[e^{C_1F(C_2(x+1))} \frac{g(C_2^\alpha (x+1)^\alpha)}{C_1C_2} \frac{g(C_2 x^\alpha)}{x} \right]_1^{n+1} \\ &+ \int_1^{n+1} e^{C_1F(C_2(x+1))} \frac{g(C_2^\alpha (x+1)^\alpha)}{C_1C_2} \frac{g(C_2 x^\alpha)}{x^2}\, dx. \qquad (3.8)\end{aligned}$$

Denote the integral on the right-hand side of (3.8) by I'_{n+1}. This integral is similar to I_{n+1}, the difference being that the integrand is multiplied by a constant times $g(C_2^\alpha (x+1)^\alpha)/x$. Since this ratio tends to zero as $x \to \infty$, and since $I_{n+1} \to \infty$ (as noted above), it holds that for any $0 < \kappa < 1$, $I'_{n+1} \le \kappa I_{n+1}$ for sufficiently large n. Hence

$$\begin{aligned}I_{n+1} &\le \frac{1}{1-\kappa} \left[e^{C_1F(C_2(x+1))} \frac{g(C_2^\alpha (x+1)^\alpha)}{C_1C_2} \frac{g(C_2 x^\alpha)}{x} \right]_1^{n+1} \\ &\le C e^{C_1F(C_2(n+2))} \frac{g(M(n+2)^\alpha)^2}{n+1}\end{aligned}$$

for sufficiently large n; recall that $M = C_2 \vee C_2^\alpha$. Summing up thus far, we have shown that

$$\begin{aligned}\Delta_{n+1} &\le C \left(e^{C_1(F(C_2(n+2))-F(C_2(n+1)))} \frac{g(M(n+2)^\alpha)^2}{n+1} + e^{-C_1F(C_2(n+1))} \right) \\ &\le C \left(\frac{g(M(n+2)^\alpha)^2}{n+1} + e^{-C_1F(C_2(n+1))} \right), \qquad (3.9)\end{aligned}$$

where the second inequality follows as F' is bounded.

Now take an arbitrary $m > 0$. Since $g(x^\alpha)/x \to 0$ there is an $x_m > 0$ such that $g(x^\alpha)/x \le 1/m$ for $x \ge x_m$, or, equivalently, $f(x^\alpha) \ge m/x$ for $x \ge x_m$. Integrating this inequality yields $F(x) - F(x_m) \ge m \log(x/x_m)$, so that

$$e^{-C_1F(x)+C_1F(x_m)} \le \left(\frac{x}{x_m} \right)^{-mC_1}$$

for $x \geq x_m$. Picking m such that $mC_1 = 1$ we see that as $n \to \infty$, the second term on the right-hand side of (3.9) is of smaller order than the first one. We conclude that

$$\Delta_{n+1} \leq C \frac{g(M(n+2)^\alpha)^2}{n+1},$$

which is (3.7). □

4 Simulated annealing on a function that cannot be computed exactly

In this section we assume that the function ψ to be maximised cannot be computed explicitly, but that we have available an approximation to it. This approximation, denoted by ψ^N, can be stochastic, based on Monte Carlo procedures; the next section shows such an example. The precision of the approximation, stochastic or not, is indexed by an integer-valued parameter N, and the larger the N, the better the approximation. This parameter can be, for instance, the number of replications in a Monte Carlo method. The following hypothesis makes precise the quality of the approximation.

Hypothesis 4. *For all $N \geq 1$ we can compute a deterministic or stochastic approximation ψ^N of ψ such that*

$$\mathbb{E}|\psi^N(x) - \psi(x)| \leq \frac{a_1}{\sqrt{N}} \quad \textit{for all } x \in \Theta,$$

and, almost surely,

$$|\psi^N(x) - \psi^N(y)| \leq 2\,\mathrm{osc}(\psi) \quad \textit{for all } x, y \in \Theta.$$

We suppose that this hypothesis holds true in all of the following. The attentive reader will notice that the second of the above assumptions can be replaced by the existence of a constant C such that, almost surely, $|\psi^N(x) - \psi^N(y)| \leq C$ for all $x, y \in \Theta$. In the case of approximation by a sample mean of i.i.d. summands, the first part of the hypothesis follows from the Marcinkiewicz-Zygmund inequality; see Appendix B for more details.

The sequence (β_n), the cooling schedule, is again chosen as

$$\beta_n = n^\alpha \vee 1, \tag{4.10}$$

although below we argue for the choice $\alpha = 1/4$ rather than $\alpha = 1/3$ as in the previous section. We will let the parameter N depend on the iteration number n as well, $N = N_n$, and we will assume that the increase is affine in n, meaning that $N_n = \lceil N_0 + N_1 n \rceil$ for some numbers $N_0 \geq 0$ and $N_1 > 0$ where $\lceil x \rceil$ denotes rounding x upwards to the nearest integer. We comment on other choices of $(N_n)_{n \geq 0}$ following the proof of Theorem 4.1 below.

We now formalise the simulated annealing procedure in this modified context. The procedure is again described as a random sequence, denoted by $(\bar{\theta}_n)_{n \geq 0}$, with $\bar{\theta}_0$ sampled from the law η_0 (as is θ_0). The function g is chosen as in Corollary 3.2, and $(\bar{\theta}_n)$ evolves as follows.

(c1) In stage n, given the current state $\bar{\theta}_n$, sample a new proposed position Z from $K(\bar{\theta}_n, \cdot)$.

(c2) Set $\bar{\theta}_{n+1} = Z$ with probability

$$f(\beta_n(\psi^{N_n}(\bar{\theta}_n) - \psi^{N_n}(Z))_+)$$

and $\bar{\theta}_{n+1} = \bar{\theta}_n$ otherwise.

This procedure requires some comments. In step (c2), ψ is approximated at two points, θ_n and Z. In the case of random approximations it is unimportant whether these two evaluations are independent or not, as we shall see below, but it is important that they are independent of approximations computed in previous steps (smaller n) of the algorithm. The reason for this is that, if such independence holds, the sequence $(\bar{\theta}_n)$ forms a Markov chain, and this Markov chain is the object of our study. Moreover, the additional randomness in step (c2) associated with the phrases 'sample a new proposed position...' and 'with probability...', typically obtained by drawing random numbers uniformly in $(0, 1)$, must be based on two mutually independent sequences of independent random numbers, also independent of the function approximations ψ^N; this is just as in the previous annealing schemes however.

In cases where the random function approximations ψ^N are such that they depend on random variables that are drawn once and for all and then stay fixed over n (sometimes called 'fixed randomness'), so that ψ^N is fixed at each point in Θ, we can, as long as N stays fixed too, apply the results of the previous section to the function ψ^N provided that it satisfies the regularity assumptions made there. Main questions are then rather whether these assumptions indeed are satisfied for ψ^N, and how well the maximum of ψ^N and its location approximate those of ψ.

We now return to the sequence $(\bar{\theta}_n)$. As noted above, this sequence is an (inhomogeneous) Markov chain. For any $\beta > 0$ and $N \geq 1$, we define the function

$$a_\beta^N(x, y) = f(\beta(\psi^N(x) - \psi^N(y))_+).$$

For fixed x and y this is indeed a random variable, the randomness coming from the evaluations $\psi^N(x)$ and $\psi^N(y)$. We write $\mathbb{E}_N$ for the expectation with respect to the random variables used to compute ψ^N at a point for some approximation index N, and $\mathbb{P}_N$ for the corresponding probability. The kernels $K_{\beta_n}^{N_n}$ of $(\bar{\theta})_{n\geq 0}$, defined by

$$K_{\beta_n}^{N_n}(x, A) = \mathbb{P}(\bar{\theta}_{n+1} \in A \,|\, \bar{\theta}_n = x)$$

for any $x \in \Theta$ and $A \in \mathcal{B}(\Theta)$, can then be expressed as

$$K_\beta^N(x, dy) = \mathbb{E}_N\left[a_\beta^N(x, y)K(x, dy) + \left(1 - \int_\Theta a_\beta^N(x, z)\, K(x, dz)\right)\delta_x(dy)\right]. \tag{4.11}$$

The final assumption we make before stating the main result of this section is the following.

Hypothesis 5. *There is a constant C_K such that for all $\beta > 0$ and $N' > N \geq 1$,*

$$\sup_{\mu\in\mathcal{P}(\Theta)} \|\mu K_\beta^N - \mu K_\beta^{N'}\|_{\mathrm{TV}} \leq C_K\beta\frac{N' - N}{N}. \tag{4.12}$$

In Appendix B we discuss this condition in detail for approximations obtained as sample means of i.i.d. random variables, and for approximations obtained using so-called particle filters. It turns out that Hypothesis 5 can often be verified through a coupling argument; that is, we couple the approximations ψ^N and $\psi^{N'}$ in a suitable way. We notice that by such an argument it also follows that provided Hypothesis 4 holds, one can bound the left-hand side of (4.12) by a constant times $\beta/\sqrt{N}$; the actual assumption above is thus stronger.

Theorem 4.1. *Assume that β_n is as in (4.10) with $\alpha < 1/2$ and that N_n increases linearly with n. Then under Hypotheses 1, 2, 3, 4 and 5, for all $\varepsilon > 0$ small enough (if Hypothesis 3(i) holds) or $0 < \varepsilon \leq \varepsilon''$ (if Hypothesis 3(ii) holds), there exists a constant C'_ε depending on ε such that*

$$\mathbb{P}(\psi(\bar{\theta}_n) \leq \psi_{\max} - \varepsilon) \leq C'_\varepsilon(n^{-\alpha}\log n \vee n^{3\alpha-1}) \tag{4.13}$$

for sufficiently large n.

Equating the two powers of this bounds leads to $\alpha = 1/4$ as the optimal choice, with corresponding rate of convergence $n^{-1/4}\log n$.

The proof of Theorem 4.1 is very similar to the proof of Theorem 3.1; before going into it, we will proceed through some technical lemmas. The following results can be shown exactly in the same manner as Lemma 3.5 and Corollary 3.6.

Lemma 4.2. *For any $\beta > 0$ and $N \geq 1$ it holds that*

$$\forall (x, A) \in \Theta \times \mathcal{B}(\Theta) : \; K_\beta^N(x, A) \geq \varepsilon_K \lambda(A) f(2\beta \operatorname{osc}(\psi)).$$

Corollary 4.3. *For all $\beta > 0$, $N \geq 1$ and any probability measures μ and ν on Θ,*

$$\|\mu K_\beta^N - \nu K_\beta^N\|_{\mathrm{TV}} \leq (1 - \varepsilon_K f(2\beta \operatorname{osc}(\psi)))\|\mu - \nu\|_{\mathrm{TV}}.$$

We point out, in particular, that these results hold true regardless of whether the two function approximations required for computing $a_\beta^N(x, y)$ are independent or not.

The results imply that for any $\beta > 0$ and $N \geq 1$, the kernel K_β^N has a unique stationary distribution, which we denote by μ_β^N. We will show that under certain conditions, μ_β^N is concentrated around the maximum of ψ.

Lemma 4.4. *For all $\beta > 0$, $\varepsilon > 0$ small enough (if Hypothesis 3(i) holds) or $0 < \varepsilon < \varepsilon''$ (if Hypothesis 3(ii) holds) and $N \geq 1$ such that $N \geq \beta^2$ and $N \geq (8a_1/\varepsilon)^2$, there is a constant C''_ε depending on ε but not on N such that*

$$\mu_\beta^N(U^{\varepsilon,c}) \leq C''_\varepsilon \frac{1 + \log\beta}{\beta}.$$

Proof. We proceed as in Lemma 3.7 and thus write

$$\begin{aligned}\mu_\beta^N(U^{\varepsilon,c}) &= \mu_\beta^N K_\beta^N(U^{\varepsilon,c}) \\ &= \iint_{x \in U^{\varepsilon,c}, y \in U^{\varepsilon,c}} \mu_\beta^N(dx)\, K_\beta^N(x, dy) + \iint_{x \in U^{\varepsilon}, y \in U^{\varepsilon,c}} \mu_\beta^N(dx)\, K_\beta^N(x, dy)\end{aligned}$$

and

$$\iint_{x\in U^{\varepsilon,c},y\in U^{\varepsilon,c}} \mu_\beta^N(dx)\, K_\beta^N(x,dy) = \int_{x\in U^{\varepsilon,c}} \mu_\beta^N(dx) \left(1 - \int_{y\in U^\varepsilon} K_\beta^N(x,dy)\right).$$

For $x \in U^{\varepsilon,c}$ it holds that

$$\begin{aligned}
\int_{y\in U^\varepsilon} K_\beta^N(x,dy) &= \mathbb{E}_N\left(\int_{y\in U^\varepsilon} f(\beta(\psi^N(x)-\psi^N(y))_+)\, K(x,dy)\right)\\
&\geq \int_{y\in U^\varepsilon} \varepsilon_K \mathbb{P}_N(\psi^N(y)\geq \psi^N(x))\, \lambda(dy)\\
&\geq \int_{y\in U^{\varepsilon/2}} \varepsilon_K[1-\mathbb{P}_N(\psi^N(x)-\psi(x)-\psi^N(y)+\psi(y)\geq \varepsilon/2)]\, \lambda(dy)\\
&\geq \int_{y\in U^{\varepsilon/2}} \varepsilon_K\left(1-\frac{4a_1}{\varepsilon\sqrt{N}}\right)\lambda(dy) \geq \frac{\varepsilon_K}{2}\lambda(U^{\varepsilon/2}),
\end{aligned}$$

where a_1 is in Hypothesis 4 and we used Markov's inequality and the assumption $N \geq (8a_1/\varepsilon)^2$. Hence

$$\int_{x\in U^{\varepsilon,c},y\in U^{\varepsilon,c}} \mu_\beta^N(dx)\, K_\beta^N(x,dy) \leq \mu_\beta^N(U^{\varepsilon,c})\left(1-\frac{\varepsilon_K}{2}\lambda(U^{\varepsilon/2})\right)$$

and

$$\mu_\beta^N(U^{\varepsilon,c}) \leq \frac{2}{\varepsilon_K \lambda(U^{\varepsilon/2})} \iint_{x\in U^\varepsilon,y\in U^{\varepsilon,c}} \mu_\beta^N(dx)\, K_\beta^N(x,dy).$$

The integral in this bound equals

$$\begin{aligned}
&\mathbb{E}_N\left(\iint_{x\in U^\varepsilon,y\in U^{\varepsilon,c}} \mu_\beta^N(dx) f(\beta(\psi^N(x)-\psi^N(y))_+)\, K(x,dy)\right)\\
&\leq \mathbb{E}_N\left(\iint_{x\in U^\varepsilon,y\in U^{\varepsilon,c}} \mu_\beta^N(dx)\frac{1}{\varepsilon_K} f(\beta(\psi^N(x)-\psi^N(y))_+)\, \lambda(dy)\right)\\
&= \frac{1}{\varepsilon_K}\iint_{x\in U^\varepsilon,y\in U^{\varepsilon,c}} \mu_\beta^N(dx) f(\beta(\psi(x)-\psi(y))_+)\\
&\qquad\qquad \times\mathbb{E}_N\left(\frac{f(\beta(\psi(x)-\psi(y))_+ + \beta R^N(x,y))}{f(\beta(\psi(x)-\psi(y)))_+}\right)\lambda(dy),
\end{aligned}$$

where $R^N(x,y) = (\psi^N(x)-\psi^N(y))_+ - (\psi(x)-\psi(y))_+$. Notice that the expression inside the final expectation is bounded by $g(\beta(\psi(x)-\psi(y))_+)$. Thus the expectation itself, inserting $g(t) = 1 + t/\tau$, may be bounded as

$$\begin{aligned}
&\mathbb{E}_N\left(\frac{1+\frac{\beta}{\tau}(\psi(x)-\psi(y))_+}{1+\frac{\beta}{\tau}(\psi(x)-\psi(y))_+ + \frac{\beta}{\tau}R^N(x,y)}\right)\\
&\leq 2\,\mathbb{P}_N\left(\left|\frac{\beta}{\tau}R^N(x,y)\right| \leq \frac{1+\frac{\beta}{\tau}(\psi(x)-\psi(y))_+)}{2}\right)\\
&+ \left(1+\frac{\beta}{\tau}(\psi(x)-\psi(y))_+\right)\mathbb{P}_N\left(\left|\frac{\beta}{\tau}R^N(x,y)\right| \geq \frac{1+\frac{\beta}{\tau}(\psi(x)-\psi(y))_+)}{2}\right)\\
&\leq 2 + \frac{2\beta}{\tau}\mathbb{E}_N|R^N(x,y)|.
\end{aligned}$$

Now notice that in the expression for $R^N(x,y)$, $\psi(x)-\psi(y)\geq 0$ for those x and y appearing in the integral. It is easy to check that for any real a and b such that $b\geq 0$, $|a_+-b|\leq |a-b|$. Hence $|R^N(x,y)|\leq |\psi^N(x)-\psi(x)|+|\psi^N(y)-\psi(y)|$, and the above expectation is thus bounded by $2+4\beta a_1/(\tau\sqrt{N})\leq 2+4a_1/\tau$, where we used the assumption $N\geq\beta^2$.

As in the proof of Lemma 3.7 we may conclude that

$$\iint_{x\in U^\varepsilon, y\in U^{\varepsilon,c}} \mu_\beta^N(dx)\, K_\beta^N(x,dy) \leq \frac{2(1+2a_1/\tau)}{\varepsilon_K}\int_{y\in U^{\varepsilon,c}} f(\beta(\psi_{\max}-\varepsilon-\psi(y)))\,\lambda(dy).$$

Summing up thus far, we have proved that

$$\mu_\beta^N(U^{\varepsilon,c})\leq \frac{4(1+2a_1/\tau)}{\varepsilon_K^2\lambda(U^{\varepsilon/2})}\int_{y\in U^{\varepsilon,c}} f(\beta(\psi_{\max}-\varepsilon-\psi(y)))\,\lambda(dy).$$

This integral can be bounded just as in the proof of Lemma 3.7, and with $f(t)=1/(1+t/\tau)$ these bounds are of order $C''_\varepsilon(1+\log\beta)/\beta$. □

We now formulate an analogue of Lemma 3.9.

Lemma 4.5. *For any $\beta'>\beta>0$,*

$$\|\mu_\beta^N-\mu_{\beta'}^N\|_{\mathrm{TV}}\leq \frac{1+2\beta\,\mathrm{osc}(\psi)/\tau}{4\varepsilon_K}\left(\frac{\beta'}{\beta}-1\right).$$

Proof. We have, using Corollary 4.3,

$$\begin{aligned}\|\mu_\beta^N-\mu_{\beta'}^N\|_{\mathrm{TV}} &\leq \|\mu_\beta^N K_\beta^N-\mu_{\beta'}^N K_\beta^N\|_{\mathrm{TV}}+\|\mu_{\beta'}^N K_\beta^N-\mu_{\beta'}^N K_{\beta'}^N\|_{\mathrm{TV}}\\ &\leq (1-\varepsilon_K f(2\beta\,\mathrm{osc}(\psi)))\|\mu_\beta^N-\mu_{\beta'}^N\|_{\mathrm{TV}}+\sup_{\mu\in\mathcal{P}(\Theta)}\|\mu K_\beta^N-\mu K_{\beta'}^N\|_{\mathrm{TV}}.\end{aligned}$$

Using (4.11), Hypothesis 4 and an argument as in the proof of Lemma 3.9, we find that for all $\mu\in\mathcal{P}(\Theta)$,

$$\begin{aligned}\|\mu K_\beta^N-\mu K_{\beta'}^N\|_{\mathrm{TV}} &\leq \sup_{x,y\in\Theta}\mathbb{E}_N|a_\beta^N(x,y)-a_{\beta'}^N(x,y)|\\ &\leq \sup_{0\leq u\leq 2\,\mathrm{osc}(\psi)}|f(\beta u)-f(\beta' u)|\\ &\leq \sup_{0\leq u\leq 2\,\mathrm{osc}(\psi)}(|f'(\beta u)|\beta u)\left(\frac{\beta'}{\beta}-1\right).\end{aligned}$$

With $g(t)=1+t/\tau$, the above supremum is bounded by $1/4$. Thus

$$\|\mu_\beta^N-\mu_{\beta'}^N\|_{\mathrm{TV}}\leq \frac{1}{\varepsilon_K f(2\beta\,\mathrm{osc}(\psi))}\frac{1}{4}\left(\frac{\beta'}{\beta}-1\right)=\frac{1+2\beta\,\mathrm{osc}(\psi)/T}{4\varepsilon_K}\left(\frac{\beta'}{\beta}-1\right).$$

□

Proof of Theorem 4.1. First we notice that given the assumptions, including $\alpha < 1/2$, Lemma 4.4 shows that $\mu_{\beta_n}^{N_n}(U^{\varepsilon,c})$ is bounded by $C''_\varepsilon n^{-\alpha}(1+\alpha\log n)$ for sufficiently large n. This term is the first part of the maximum in (4.13).

Next we denote by $\bar\eta_n$ the law of $\bar\theta_n$ and put $\Delta_n = \|\bar\eta_n - \mu_{\beta_n}^{N_n}\|_{\mathrm{TV}}$. Write

$$\begin{aligned}\Delta_{n+1} &\le \|\bar\eta_n K_{\beta_n}^{N_n} - \mu_{\beta_n}^{N_n} K_{\beta_n}^{N_n}\|_{\mathrm{TV}} + \|\mu_{\beta_n}^{N_n} - \mu_{\beta_{n+1}}^{N_n}\|_{\mathrm{TV}} + \|\mu_{\beta_{n+1}}^{N_n} - \mu_{\beta_{n+1}}^{N_{n+1}}\|_{\mathrm{TV}}\\ &\le (1-\varepsilon_K f(2\beta_n \operatorname{osc}(\psi)))\Delta_n + \frac{1+2\beta_n\operatorname{osc}(\psi)/\tau}{4\varepsilon_K}\left(\frac{\beta_{n+1}}{\beta_n}-1\right)\\ &\quad + \|\mu_{\beta_{n+1}}^{N_n} - \mu_{\beta_{n+1}}^{N_{n+1}}\|_{\mathrm{TV}}, \end{aligned} \tag{4.14}$$

where we used Corollary 4.3 and Lemma 4.5 to bound the first two terms. With our choice of β_n, the second term on the right-hand side is of order $n^{\alpha-1}$.

To bound the third term we proceed as in the proof of Lemma 4.5; use Corollary 4.3 to see that for any β, N and N',

$$\begin{aligned}\|\mu_\beta^N - \mu_\beta^{N'}\|_{\mathrm{TV}} &\le \|\mu_\beta^N K_\beta^N - \mu_\beta^{N'} K_\beta^N\|_{\mathrm{TV}} + \|\mu_\beta^{N'} K_\beta^N - \mu_\beta^{N'} K_\beta^{N'}\|_{\mathrm{TV}}\\ &\le (1-\varepsilon_K f(2\beta\operatorname{osc}(\psi)))\|\mu_\beta^N - \mu_\beta^{N'}\|_{\mathrm{TV}} + \sup_{\mu\in\mathcal{P}(\Theta)}\|\mu K_\beta^N - \mu K_\beta^{N'}\|_{\mathrm{TV}}\end{aligned}$$

to arrive at

$$\|\mu_\beta^N - \mu_\beta^{N'}\|_{\mathrm{TV}} \le \frac{1}{\varepsilon_K f(2\beta\operatorname{osc}(\psi)))}\sup_{\mu\in\mathcal{P}(\Theta)}\|\mu K_\beta^N - \mu K_\beta^{N'}\|_{\mathrm{TV}}.$$

Apply this bound with $\beta = \beta_{n+1}$, $N = N_n$ and $N' = N_{n+1}$ to see that the final term of (4.14) is bounded by a constant times $\beta_{n+1}^2(N_{n+1}-N_n)/N_n$ under Hypothesis 5; this ratio is of order $n^{2\alpha-1}$ given that N_n is assumed to be affine in n.

Summing up thus far, we have proved that

$$\Delta_{n+1} \le (1-\varepsilon_K f(2\beta_n\operatorname{osc}(\psi)))\Delta_n + \frac{C}{n^{1-2\alpha}}$$

for some constant C. Using this inequality we can show as in the proof of Theorem 3.1 that for all $\varepsilon > 0$, $\Delta_n \le C'_\varepsilon n^{3\alpha-1}$ for some constant C'_ε depending on ε. Indeed, in the proof of Theorem 3.1, replace the factor x in the denominator of the expression that forms the integrand in I_{n+1} by $x^{1-\alpha}$ and proceed from there. The term $C'_\varepsilon n^{3\alpha-1}$ is the second part of the maximum in (4.13). □

One may consider other ways of increasing N_n, for instance $N_n = \lceil N_0 + N_1 n^\delta\rceil$ for some $\delta > 0$. For $\delta > 1$ the expression $(N_{n+1}-N_n)/N_n$ is then still of order n^{-1} however, so there is no improvement in the proof of Theorem 4.1 compared to the case of affine increase. For $\delta < 1$ the above expression is of order $n^{-\delta}$, since the N_n are integer-valued. The bound corresponding to (4.13) then becomes of order $n^{-\alpha}\log n \vee n^{3\alpha-\delta}$, with the optimal α being $\delta/4$.

The above seems to suggest that the rate $n^{-1/3}\log n$ of Section 3 is unobtainable when the function ψ is approximated. This is not the case however, but it requires a slightly different approach to analysis than above, and also

typically a faster increase of N_n. In the proof of Theorem 4.1 we compared $\bar{\eta}_n$ to $\mu_{\beta_n}^{N_n}$. Consider instead comparing to μ_{β_n}, as in the proof of Theorem 3.1, and write

$$\begin{aligned}\bar{\eta}_{n+1} - \mu_{\beta_{n+1}} &= \bar{\eta}_n K_{\beta_n}^{N_n} - \mu_{\beta_n} K_{\beta_n}^{N_n} \\ &+ \mu_{\beta_n} K_{\beta_n}^{N_n} - \mu_{\beta_{n+1}} K_{\beta_n}^{N_n} \\ &+ \mu_{\beta_{n+1}} K_{\beta_n}^{N_n} - \mu_{\beta_{n+1}} K_{\beta_{n+1}}^{N_n} \\ &+ \mu_{\beta_{n+1}} K_{\beta_{n+1}}^{N_n} - \mu_{\beta_{n+1}} K_{\beta_{n+1}}.\end{aligned}$$

On the right-hand side the total variation norm of the first difference is bounded by $(1 - \varepsilon_K f(2\beta_n \operatorname{osc}(\psi)))\|\bar{\eta}_n - \mu_{\beta_n}\|_{\mathrm{TV}}$ (Corollary 4.3), and the norms of the remaining differences are bounded by terms of order $n^{\alpha-1}$ (Lemma 3.9), n^{-1} (use part of the proof of Lemma 4.5) and $n^\alpha/N_n^{1/2}$ respectively. To obtain the order $n^\alpha/N_n^{1/2}$ of the final term we can couple the kernels K_β and K_β^N in a way similar to that used in the first part of Appendix B, thus obtaining a bound on the total variation distance of order $\beta \sup_{x\in\Theta} \mathbb{E}_N|\psi_N(x) - \psi(x)|$; by Hypothesis 4 this expression is of order $\beta/N^{1/2}$. Thus we do not require Hypothesis 5 for this analysis.

We can now put $\Delta_n = \|\bar{\eta}_n - \mu_{\beta_n}\|_{\mathrm{TV}}$ and mimic the proof of Theorem 4.1. To obtain the rate of convergence $n^{-\alpha}\log n$, the norms of all differences on the right-hand side, except the first one, must be of order $n^{-2\alpha}$. This in turn requires taking $\alpha \leq 1/3$ and N_n of the order $n^{6\alpha}$. In particular this applies when $\alpha = 1/3$, so that this rate is obtainable but at the cost of quickly increasing N_n at rate n^2. We also notice that when $\alpha = 1/4$, to obtain the rate of convergence $n^{-1/4}\log n$ it is required to take N_n of order $n^{3/2}$, which is larger than the linear rate used in Theorem 4.1.

However, a more fair way to look at convergence rates is to express them in terms of the number of numerical operations performed. We assume that the computational cost of computing an approximation $\psi^N(x)$ is of order N; this is for instance the case for the Monte Carlo schemes discussed in Appendix B. With N_n being affine in n, the total computational cost up to stage n of the simulated annealing scheme is thus of order n^2. Denoting the total number of numerical operations performed by C, we then find that the convergence rate is of order $C^{-1/8}\log C$. If we rather use the second bound above, which requires N_n of order $n^{6\alpha}$, we see that the computational cost up to stage n is of order $n^{6\alpha+1}$ and the convergence rate is of order $C^{-\alpha/(6\alpha+1)}\log C$ for $0 < \alpha \leq 1/3$. The optimal α is $\alpha = 1/3$, with rate $C^{-1/9}\log C$. This is inferior to $C^{-1/8}\log C$, so that the decomposition of the proof of Theorem 4.1 is superior; it does require Hypothesis 5 however.

5 A numerical illustration

In this section we consider simulated annealing applied to the likelihood function of a state-space model as in Appendix B.2. Thus assume that we have an observed sequence $(y_t)_{1\leq t\leq T}$ from a state-space model $((S_t, Y_t))_{1\leq t\leq T}$, whose Markov transition kernel Q and conditional output densities $r(\cdot|s)$ both depend

on an unknown parameter (vector) θ which we wish to estimate using maximum likelihood.

The log-likelihood function that we aim to maximise is

$$\ell_T(\theta) = \sum_{t=1}^{T} \log p_\theta(y_t|y_{1:t-1}) = \sum_{t=1}^{T} \log \int r_\theta(y_t|s)\, \pi^\theta_{t|t-1}(ds),$$

where $p_\theta(y_t|y_{1:t-1})$ is the conditional density of Y_t given $Y_{1:t-1}$, and $\pi^\theta_{t|t-1}$ is the predictive distribution $\mathbb{P}_\theta(S_t \in \cdot \,|\, y_{1:t-1})$. As $\pi^\theta_{t|t-1}$ can in general not be computed we need to approximate the log-likelihood function, and one way to do that is through

$$\ell^N_T(\theta) = \sum_{t=1}^{T} \log \int r_\theta(y_t|s)\, \pi^{\theta,N}_{t|t-1}(ds),$$

where we take $\pi^{\theta,N}_{t|t-1}(ds)$ as the particle filter approximation of Appendix B.2.

The log-likelihood function is essentially a sum of functions of the form studied in Appendix B.2, except for the logarithmic transformation. Assuming however, as in Appendix B.2, that r_θ is uniformly bounded from below by some $\underline{r} > 0$, we find that each of the integrals above are bounded from below by $\underline{r}$. Moreover, using the inequality $|\log x - \log y| \le |x-y|/(x \wedge y)$, valid for all $x, y > 0$, we find that

$$|\ell^N_T(\theta) - \ell^{N'}_T(\theta)| \le \frac{1}{\underline{r}} \sum_{t=1}^{T} \left| \int r_\theta(y_t|s)\, \pi^{\theta,N}_{t|t-1}(ds) - \int r_\theta(y_t|s)\, \pi^{\theta,N'}_{t|t-1}(ds) \right|.$$

This bound involves sums of functions of the form studied in Appendix B.2 (take $h(s) = r_\theta(y_t|s)$), and we can proceed as there to show that Hypothesis 5 holds. A similar argument where we replace $\ell_T(\theta)$ by the exact likelihood and appeal to Theorem 7.4.4 of Del Moral (2004) shows that Hypothesis 4 holds.

5.1 Simulation study

We considered the benchmark model (Doucet *et al.*, 2001, Eqs. 8.3.4–8.3.5)

$$S_t = aS_{t-1} + b\frac{S_{t-1}}{1+S^2_{t-1}} + \gamma\cos(1.2t) + \sigma_v V_t, \tag{5.15}$$

$$Y_t = \frac{S^2_t}{20} + \sigma_w W_t, \tag{5.16}$$

where (S_t) is the unobserved Markov chain taking values in $\mathbb{R}$, (Y_t) is the observable process and (V_t) and (W_t) are mutually independent sequences of i.i.d. standard Gaussian random variables. We wish to estimate the five model parameters $\theta = (a, b, \gamma, \sigma_v, \sigma_w)$ given a sequence $(y_t)_{1\le t\le T}$ of observations, and we did so using the approximate maximum likelihood (ML) approach outlined above with the bootstrap particle filter, i.e. particle mutations following the system dynamics (5.15). We remark that the state space of the model above is not compact, so that the conditional densities $r_\theta(y|s)$ are not bounded from below

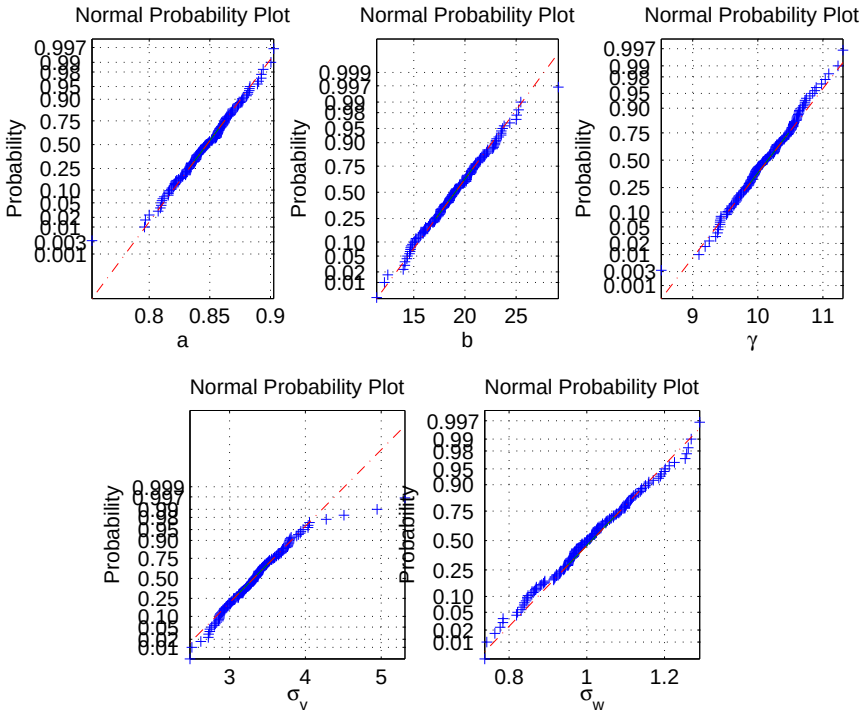

Figure 1: Normal probability plots of approximate ML estimates of parameters $(a, b, \gamma, \sigma_v, \sigma_w)$ in the model (5.15)–(5.16), obtained from 150 replications of 5,000 iterations of the simulated annealing scheme applied to the particle filter approximation of the log-likelihood.

in s. The model does thus not fulfil the technical conditions made above, but the results below are still an illustrative example of how the simulated annealing scheme performs in a particular case.

We simulated a single trajectory $(y_t)_{1\leq t\leq T}$ of length $T = 500$ with parameters $\theta^0 = (a^0, b^0, \gamma^0, \sigma_v^0, \sigma_w^0) = (0.9, 18, 10, \sqrt{10}, 1)$. In the simulated annealing scheme we let the inverse temperature be $\beta_n = 10n^{1/4}$, corresponding to $\tau = 1/10$ in Corollary 3.2, and let number of particles at step n be $N_n = n \vee 20$, a function which is affine for $n \geq 20$ (Theorem 4.1). The algorithm was run for 5,000 iterations in each of 150 independent replications. The parameter space Θ was taken as the five-dimensional hyper-rectangle $[0.45, 1.8] \times [9, 36] \times [5, 20] \times [0.316, 36] \times [0.5, 2]$. For K we used a Gaussian random walk proposal (on the log-scale for the standard deviations), where we constrained the random walk to Θ; any coordinate of the parameter proposed outside Θ was pulled back to the boundary. The incremental covariance of the kernel at step n was a diagonal matrix whose i-th diagonal element was the squared i-th side length of Θ divided by $\log(n+1)^2$. In each replication the initial point θ_0 was drawn uniformly on Θ.

After 5,000 iterations of the simulated annealing algorithm, the sample means and standard errors of the parameter estimates $\bar{\theta}_{5000}$ (over the 150 repli-

cations) were $(0.85, 19.1, 10.1, 3.4, 1.01)$ and $(0.024, 3.0, 0.46, 0.41, 0.11)$ respectively. These sample means are in good agreement with the true θ^0. Ideally we would like to compare to the ML estimates, which are however unavailable. Figure 1 shows that the estimates follow normal distributions with good accuracy, with the exception of σ_v. This of course is an empirical observation for which we have no theoretical support, as we have not discussed convergence in law of the differences $\theta_n - \theta_{\max}$ and $\bar{\theta}_n - \theta_{\max}$, suitably scaled, where $\theta_{\max}$ is the point where ψ is maximal.

References

Bartoli, N. and Del Moral, P. (2001) *Simulation et algorithmes stochastiques*. Cépaduès.

Berger, M. and Gostiaux, B. (1988) *Differential Geometry: Manifolds, Curves, and Surfaces*. New York: Springer-Verlag.

Catoni, O. (1999) Simulated annealing algorithms and Markov chains with rare transitions. In *Séminaire de Probabilités, XXXIII*, vol. 1709 of *Lecture Notes in Mathematics*, 69–119. Berlin: Springer.

Cot, C. and Catoni, O. (1998) Piecewise constant triangular cooling schedules for generalized simulated annealing algorithms. *Ann. Appl. Probab.*, **8**, 375–396.

Del Moral, P. (2004) *Feynman-Kac Formulae. Geneological and Interacting Particle Systems with Applications*. New York: Springer.

Del Moral, P. and Guionnet, A. (2001) On the stability of interacting processes with applications to filtering and genetic algorithms. *Ann. Inst. H. Poincaré Probab. Statist.*, **37**, 155–194.

Del Moral, P. and Miclo, L. (1999) On the convergence and applications of generalized simulated annealing. *SIAM J. Control Optim.*, **37**, 1222–1250.

Dobrushin, R. (1956) Central limit theorem for non-stationary Markov chains. I,II. *Theory of Probability and its Applications*, **1**, 65–80, 329–383.

Doucet, A., de Freitas, N. and Gordon, N. (eds.) (2001) *An Introduction to Sequential Monte Carlo Methods*. New York: Springer.

Gielis, G. and Maes, C. (1999) A simple approach to time-inhomogenoeus dynamics and applications to (fast) simulated annealing. *J. Phys.* A*: Math. Gen.*, **32**, 5389–5407.

Lindvall, T. (2002) *Lectures on the Coupling Method*. Mineola, NY: Dover Publications. Corrected reprint of the 1992 original.

Locatelli, M. (2001) Convergence and first hitting time of simulated annealing algorithms for continuous global optimization. *Math. Methods Oper. Res.*, **54**, 171–199.

Shiryaev, A. N. (1995) *Probability*. New York: Springer-Verlag, 2nd ed.

Tsallis, C. and Stariolo, D. A. (1996) Generalized simulated annealing. *Physica* A, **233**, 395–406.

A Rate of convergence of classical simulated annealing

In this section we prove the bound (1.3) and also, by studying a specific example, that this bound cannot be improved generally. We assume that Hypotheses 1–2 and Hypothesis 3(ii) hold. Since we now consider classical simulated annealing we have $f(t) = \exp(-t)$, and we take $\beta_n = \beta_0 \log(n+e)$ with $1/\beta_0 > \operatorname{osc}(\psi)$ (cf. Bartoli and Del Moral, 2001, Theorem 2.3.5). As in Section 3 we let η_n be the law of θ_n and denote by μ_β the invariant distribution of K_β.

Now write

$$\mathbb{P}(\psi(\theta_n) \leq \psi_{\max} - \varepsilon) = \eta_n(U^{\varepsilon,c}) = (\eta_n(U^{\varepsilon,c}) - \mu_{\beta_n}(U^{\varepsilon,c})) + \mu_{\beta_n}(U^{\varepsilon,c}).$$

We will show that the first term of this decomposition (the difference) tends to zero at algebraic rate, while the second term vanishes only logarithmically fast. Thus the left-hand side tends to zero at logarithmic rate too. In a specific example we will also show that the logarithmic rate for the second term, which in general is a bound, is in fact the exact rate; thus the logarithmic rate for the left-hand side cannot be improved generally. Here emerges an essential difference between classical simulated annealing and the new scheme analysed in Section 3. In both cases the total variation distance between the law η_n of θ_n and the invariant law μ_{β_n} vanishes at algebraic rate; $n^{\alpha-1}$ for classical simulated annealing (see below) and $n^{2\alpha-1}$ for the new scheme (Theorem 3.1). The rate at which μ_{β_n} concentrates around the maximum of ψ is much different however; this rate is algebraic too for the new scheme (Lemma 3.7), but only logarithmic (or algebraic with rate tending to zero) for the classical scheme.

We now proceed to the details. Put once again $\Delta_n = \|\eta_n - \mu_{\beta_n}\|_{\mathrm{TV}}$. We then have the recursion

$$\Delta_{n+1} \leq (1 - \varepsilon_K e^{-\beta_n \operatorname{osc}(\psi)})\Delta_n + (\beta_{n+1} - \beta_n)\operatorname{osc}(\psi);$$

see Bartoli and Del Moral (2001, Remark 3.3.13) and cf. the proof of Theorem 3.1. With the present choice of (β_n) we find $\beta_{n+1} - \beta_n \leq \beta_0/(n+1)$ and $\exp(-\beta_n \operatorname{osc}(\psi)) = (n+e)^{-\alpha}$ with $\alpha = \beta_0 \operatorname{osc}(\psi) < 1$. Iterating the above recursion yields

$$\begin{aligned}\Delta_{n+1} \leq \sum_{q=1}^{n} \prod_{k=q+1}^{n} \left(1 - \frac{\varepsilon_K}{(k+e)^\alpha}\right) \times \frac{\beta_0 \operatorname{osc}(\psi)}{q+1} \\ + \prod_{k=1}^{n} \left(1 - \frac{\varepsilon_K}{(k+e)^\alpha}\right) \times \|\eta_1 - \mu_{\beta_1}\|_{\mathrm{TV}},\end{aligned}$$

where an empty product (when $q = n$) is interpreted as unity. Bound the

product as

$$\begin{aligned}\log \prod_{k=q+1}^{n} \left(1 - \frac{\varepsilon_K}{(k+e)^\alpha}\right) &\leq -\sum_{k=q+1}^{n} \frac{\varepsilon_K}{(k+e)^\alpha} \\ &\leq -\varepsilon_K \int_{q+1}^{n+1} \frac{dx}{(x+e)^\alpha} \\ &= -C_1((n+e+1)^{1-\alpha} - (q+e+1)^{1-\alpha}),\end{aligned}$$

where $C_1 = \varepsilon_K/(1-\alpha)$. Thus, using $\beta_0 \operatorname{osc}(\psi) < 1$ again as well,

$$\begin{aligned}\Delta_{n+1} &\leq e^{-C_1(n+e+1)^{1-\alpha}} \sum_{q=1}^{n} e^{C_1(q+e+1)^{1-\alpha}} \frac{1}{q+1} \\ &+ e^{-C_1((n+e+1)^{1-\alpha} - (e+1)^{1-\alpha})} \|\eta_1 - \mu_{\beta_1}\|_{\mathrm{TV}} \\ &\leq e^{-C_1(n+e+1)^{1-\alpha}} \int_1^{n+1} e^{C_1(x+e+1)^{1-\alpha}} \frac{1}{x}\, dx + Ce^{-C_1(n+e+1)^{1-\alpha}}.\end{aligned}$$

By manipulating the integral on the right-hand side, I_{n+1} say, we can just as in the proof of Theorem 3.1 prove that

$$I_{n+1} \leq Ce^{C_1(n+e+2)^{1-\alpha}} \frac{1}{(n+1)^{1-\alpha}}.$$

Hence we obtain

$$\begin{aligned}\Delta_{n+1} &\leq Ce^{-C_1((n+e+1)^{1-\alpha} - (n+e+2)^{1-\alpha})} \frac{1}{(n+1)^{1-\alpha}} + Ce^{-C_1(n+e+1)^{1-\alpha}} \\ &\leq \frac{C}{(n+1)^{1-\alpha}}\end{aligned}$$

and thus $\Delta_n \leq C/n^{1-\alpha}$.

So far the difference between η_n and μ_{β_n}. We now turn to how concentrated μ_{β_n} is around the maximum of ψ. To start with we may employ Lemma 3.7, with f and β_n as above, to obtain

$$\mu_{\beta_n}(U^{\varepsilon,c}) \leq \frac{C_\varepsilon}{\beta_0 \log(n+e)} + (n+e)^{-\beta_0(\varepsilon'' - \varepsilon)};$$

a logarithmic rate in other words. We can also use the property mentioned in (Bartoli and Del Moral, 2001, p. 64), that μ_β equals $\exp(\beta\psi(x))\,\gamma(dx)$ up to a normalising constant with γ the invariant distribution of K, to obtain

$$\begin{aligned}\mu_\beta(U^{\varepsilon,c}) &= \frac{\displaystyle\int_{U^{\varepsilon,c}} e^{\beta\psi(y)}\, \gamma(dy)}{\displaystyle\int e^{\beta\psi(y)}\, \gamma(dy)} \\ &\leq \frac{\displaystyle\int_{U^{\varepsilon,c}} e^{\beta(\psi_{\max} - \varepsilon)}\, \gamma(dy)}{\displaystyle\int_{U^{\varepsilon/2}} e^{\beta(\psi_{\max} - \varepsilon/2)}\, \gamma(dy)} = \frac{e^{-\beta\varepsilon/2}}{\gamma(U^{\varepsilon/2})}. \qquad \text{(A.17)}\end{aligned}$$

Inserting β_n for β, it follows that

$$\mu_{\beta_n}(U^{\varepsilon,c}) \leq \frac{1}{\gamma(U^{\varepsilon/2})}(n+e)^{-(\beta_0/2)\varepsilon}.$$

This is the bound (1.3).

We now prove that this bound cannot be improved in general. Consider the example $\Theta = [-1/2, 1/2]$, $\psi(x) = -|x|$, $K(x, dy) = dy$. Thus K is an independence kernel that proposes uniformly on Θ. It is immediate that the invariant measure γ of K is Lebesgue measure on Θ, and that γ is K-reversible. Now $\mu_\beta(A)$ is proportional to $\int_A \exp(-\beta|y|)\,dy$, so that

$$\mu_\beta(U^{\varepsilon,c}) = \frac{\displaystyle\int_{\varepsilon<|y|\leq 1/2} e^{-\beta|y|}\,dy}{\displaystyle\int_\Theta e^{-\beta|y|}\,dy} = \frac{e^{-\beta\varepsilon} - e^{-\beta/2}}{1 - e^{-\beta/2}} \sim e^{-\beta\varepsilon} \quad \text{as } \beta \to \infty.$$

We can indeed, by an obvious modification of the argument above, adjust (A.17) into the bound $1/\gamma(U^{\varepsilon\delta}) \times e^{-\beta(1-\delta)\varepsilon}$, where $0 < \delta < 1$ is arbitrary. The rate of this bound thus can thus be made arbitrarily close to the exact rate of this example.

B Coupling function approximations

The purpose of this appendix is to illustrate how one may construct function approximations ψ^N that satisfy Hypothesis 5, and how the relatively 'high level' condition of this hypothesis can be guaranteed by more 'low level' assumptions.

Thus assume that we are given a probability measure μ on Θ, $\beta > 0$, and two approximation indices N and N'. We wish to bound $\|\mu K_\beta^N - \mu K_\beta^{N'}\|_{\mathrm{TV}} = \sup_A |\mu K_\beta^N(A) - \mu K_\beta^{N'}(A)|$, where the supremum is over $A \in \mathcal{B}(\Theta)$. We will accomplish this by constructing two coupled samples from μK_β^N and $\mu K_\beta^{N'}$ respectively as follows.

(i) Sample a point x from μ and then a point z from $K(x, \cdot)$.

(ii) Compute the function approximations $\psi^N(x)$, $\psi^N(z)$, $\psi^{N'}(x)$ and $\psi^{N'}(z)$. For the time being we do not specify exactly how this is done.

(iii) Sample a random number U from the uniform distribution on $(0, 1)$ and accept the proposal z if $U \leq f(\beta(\psi^N(x) - \psi^N(z))_+)$ or $U \leq f(\beta(\psi^{N'}(x) - \psi^{N'}(z))_+)$ respectively, for the two indices N and N'.

The samples μK_β^N and $\mu K_\beta^{N'}$ so constructed will be different only if the two decisions is step (iii) are different, so the probability of the former event is bounded by the probability of the latter one. To compute the probability that the decisions of step (iii) differ, we notice this event occurs if U falls in between the two function values used there, which, since U is uniform, happens with (conditional) probability

$$|f(\beta(\psi^N(x) - \psi^N(z))_+) - f(\beta(\psi^{N'}(x) - \psi^{N'}(z))_+)|.$$

Hence the probability of different decisions in step (iii) is bounded by

$$\sup_{x,z\in\Theta} \mathbb{E}|f(\beta(\psi^N(x)-\psi^N(z))_+) - f(\beta(\psi^{N'}(x)-\psi^{N'}(z))_+)|,$$

where the expectation is w.r.t. the function approximations ψ^N and $\psi^{N'}$. The difference of the function values can be bounded as

$$\beta|(\psi^N(x)-\psi^N(z))_+ - (\psi^{N'}(x)-\psi^{N'}(z))_+| \times f'(\zeta),$$

where ζ is point between the two function arguments. By the assumptions on f its derivative is necessarily bounded, and it is straightforward to check that for any real a and b, $|a_+ - b_+| \leq |a-b|$. Therefore the probability of different decisions in step (iii) is bounded by

$$\begin{aligned}\beta\|f'\|_\infty \sup_{x,z\in\Theta} \mathbb{E}|(\psi^N(x)-\psi^N(z)) - (\psi^{N'}(x)-\psi^{N'}(z))| \\ \leq 2\beta\|f'\|_\infty \sup_{x\in\Theta} \mathbb{E}|\psi^N(x)-\psi^{N'}(x)|.\end{aligned}$$

Thus, at this point we see that if the function approximations satisfy

$$\sup_{x\in\Theta} \mathbb{E}|\psi^N(x)-\psi^{N'}(x)| \leq C\frac{N'-N}{N} \tag{B.18}$$

for some constant C, Hypothesis 5 will hold.

Verifying (B.18) is, of course, a problem very much related to the specific construction of these approximations. In the following two subsections we will deal with two specific settings: i.i.d. sample means and particle filters.

B.1 Simple Monte Carlo sample means

Here we consider the possibly simplest of all approximation schemes: a sample mean of i.i.d. summands. Thus we assume that for a random variable ξ with some known distribution and some known function h, $\psi(x) = \mathbb{E}h(\xi;x)$ where the expectation is w.r.t. ξ, and that its approximation is

$$\psi^N(x) = \frac{1}{N}\sum_{i=1}^N h(\xi_i;x)$$

where the ξ_i are i.i.d. variables distributed as ξ. We note in passing that for this scheme the Marcinkiewicz-Zygmund inequality (Shiryaev, 1995, p. 498) with $p=1$ implies that Hypothesis 4 holds. Moreover, for $N' > N$,

$$\psi^N(x) - \psi^{N'}(x) = \left(\frac{1}{N}-\frac{1}{N'}\right)\sum_{i=1}^N h(\xi_i;x) - \frac{1}{N'}\sum_{i=N+1}^{N'} h(\xi_i;x)$$

and

$$\begin{aligned}\mathbb{E}|\psi^N(x)-\psi^{N'}(x)| &\leq \left(\frac{1}{N}-\frac{1}{N'}\right)N\mathbb{E}|h(\xi;x)| + \frac{1}{N'}(N'-N)\mathbb{E}|h(\xi;x)| \\ &= 2\,\mathbb{E}|h(\xi;x)|\,\frac{N'-N}{N'}.\end{aligned}$$

It is now immediate that if $\mathbb{E}|h(\xi;x)|$ is bounded in $x\in\Theta$, (B.18) holds.

B.2 Particle filter estimates

Consider a state-space model $((S_t, Y_t))_{t\geq 1}$, where (S_t) is an unobserved Markov chain on some general state space and (Y_t) is an observed sequence of random variables. The association between (S_t) and (Y_t) is local in the sense that (i) given (S_t), the Y-variables are conditionally independent, and (ii) given (S_t) and for any time index u, the conditional distribution of Y_u depends on S_u only.

We will denote the transition kernel of the Markov chain (S_t) by Q, and the conditional density of Y_t given $S_t = s$ by $r(\cdot|s)$. Both of these quantities are assumed to depend on some model parameters θ, which we indicate by writing Q_θ and r_θ respectively.

The function ψ we wish to approximate is $\psi(\theta) = \mathbb{E}_\theta[h(S_t) \,|\, y_{1:t-1}]$, that is, the expectation of some function h w.r.t. the so-called *predictive distribution* $\pi^\theta_{t|t-1}(\cdot) = \mathbb{P}_\theta(S_t \in \cdot \,|\, y_{1:t-1})$, where $t \geq 1$ is some time index, the notation $y_{1:t-1}$ is short for $y_1, y_2, \ldots, y_{t-1}$, and subindex '$t|t-1$' indicates that the distribution concerns the state at time t conditional on observed data up to time $t-1$.

The predictive distributions can, together with the so-called *filter distributions* $\pi^\theta_{t|t}(\cdot) = \mathbb{P}_\theta(S_t \in \cdot \,|\, y_{1:t})$, be computed recursively in time—at least in principle. The recursive formulae read

$$\pi^\theta_{t|t}(ds) = \frac{r_\theta(y|s)\, \pi^\theta_{t|t-1}(ds)}{\displaystyle\int r_\theta(y|s')\, \pi^\theta_{t|t-1}(ds')} \tag{B.19}$$

and

$$\pi^\theta_{t+1|t}(\cdot) = \int Q_\theta(s, \cdot)\, \pi^\theta_{t|t}(\cdot). \tag{B.20}$$

The first of these formulae is just Bayes' rule, and the second one means to propagate the filter through the state dynamics Q_θ.

In practice the above relations do no admit exact numerical solution except in two cases: when the state space of (S_t) is finite (so-called hidden Markov models; the integrals then turn into finite sums) and when the state-space model is linear with additive Gaussian noise (the solution then being provided by the Kalman filter). There are many ways to approximate these two recursions, and here we shall examine an approach referred to as *particle filters*. This section contains a full introduction neither to state-space models nor to particle filters, and we refer to Doucet *et al.* (2001) for a more complete coverage of both.

The basic idea of a particle filter is to approximate the filter and predictive distributions with the empirical distributions of a set of *particles*, whose positions are dynamically updated in time. There is not just one particle filter algorithm—the term rather refers to a framework for algorithms—and the particular algorithm we look at here is usually denoted the *bootstrap particle filter*. We now describe how this algorithm works; the parameter θ and population size N are fixed throughout.

Assume that at some time index t we have available a collection $(\xi^{\theta,N}_{t|t-1,i})_{1\leq i\leq N}$ of particles whose empirical distribution approximates $\pi^\theta_{t|t-1}$. The transformation (B.19) is approximated as follows.

(a) *Weighting.* Compute unnormalised weights $\tilde{w}_{t,i}^{\theta,N} = r_\theta(y_t|\xi_{t|t-1,i}^{\theta,N})$ and then normalised weights $w_{t,i}^{\theta,N} = \tilde{w}_{t,i}^{\theta,N} / \sum_j \tilde{w}_{t,j}^{\theta,N}$.

(b) *Resampling.* Create a sample $(\xi_{t|t,i}^{\theta,N})_{1\leq i\leq N}$ by sampling N times independently from $(\xi_{t|t-1,i}^{\theta,N})_{1\leq i\leq N}$ with weights $(w_{t,i}^{\theta,N})_{1\leq i\leq N}$.

The empirical distribution of the sample $(\xi_{t|t,i}^{\theta,N})_{1\leq i\leq N}$ obtained in the resampling step approximates $\pi_{t|t}^\theta$.

The transformation (B.20) is approximated as follows.

(c) *Mutation.* Create a sample $(\xi_{t+1|t,i}^{\theta,N})_{1\leq i\leq N}$ by independently sampling $\xi_{t+1|t,i}^{\theta,N}$ from $Q_\theta(\xi_{t|t,i}^{\theta,N}, \cdot)$.

The procedure is initialised at time $t = 0$ by letting $(\xi_{1|0,i}^{\theta,N})_{1\leq i\leq N}$ be an i.i.d. sample of size N from the initial distribution $P_\theta(S_1 \in \cdot)$ of the state process. This distribution may depend on θ but is otherwise assumed known.

The book by Del Moral (2004) is a thorough treatise of theoretical properties of particle filters, and in particular its Theorem 7.4.4 shows that Hypothesis 4 holds, provided that for each y_t, $r_\theta(y_t|s)$ is bounded in θ and s. We are here particularly interested in the particle approximations of the predictive distributions, and the update of these can be summarised as follows: compute the normalised weights $w_{t,i}^{\theta,N}$ and then sample for $1 \leq i \leq N$, independently, first an index j with probability $w_{t,j}^{\theta,N}$ and then $\xi_{t+1|t,i}^{\theta,N} \sim Q_\theta(\xi_{t|t-1,j}^{\theta,N})$.

We will now run, simultaneously, two particles filters of sizes $N' > N$ respectively. All other properties of the filers—data, parameters, dynamics—agree. The joint dynamics of the filters will be coupled in a way such that many particles of the two filters, at any given time index, coincide. Indeed, for each time index t we define a partition $J_t \cup J_t^c$ of $\{1, 2, \ldots, N'\}$ such that $\xi_{t|t-1,i}^{\theta,N} = \xi_{t|t-1,i}^{\theta,N'}$ for $i \in J_t$. The details of the coupling are as follows.

(i) *Initialisation.* Sample $(\xi_{1|0,i}^{\theta,N'})_{1\leq i\leq N'}$ independently from $\mathbb{P}_\theta(S_1 \in \cdot)$, let $\xi_{1|0,i}^{\theta,N} = \xi_{1|0,i}^{\theta,N'}$ for $1 \leq i \leq N$ and let $J_1 = \{1, 2, \ldots, N\}$, $J_1^c = \{N+1, N+2, \ldots, N'\}$.

(ii) *Recursion from t to $t+1$.* We have $\xi_{t|t-1,i}^{\theta,N} = \xi_{t|t-1,i}^{\theta,N'}$ for $i \in J_t$ and compute the weights $(w_{t,i}^{\theta,N})_{1\leq i\leq N}$ and $(w_{t,i}^{\theta,N'})_{1\leq i\leq N'}$.

When sampling the new particles, we couple the two filters in a way such that independently for each $1 \leq i \leq N$, one of the events below take place (index j has the same meaning as above):

- for $j \in J_t$,
  - $\xi_{t+1|t,i}^{\theta,N} = \xi_{t+1|t,i}^{\theta,N'} \sim Q_\theta(\xi_{t|t-1,j}^{\theta,N}, \cdot)$ with probability $w_{t,j}^{\theta,N} \wedge w_{t,j}^{\theta,N'}$;
  - $\xi_{t+1|t,i}^{\theta,N} \sim Q_\theta(\xi_{t|t-1,j}^{\theta,N}, \cdot)$ with probability $w_{t,j}^{\theta,N} - w_{t,j}^{\theta,N} \wedge w_{t,j}^{\theta,N'}$;
  - $\xi_{t+1|t,i}^{\theta,N'} \sim Q_\theta(\xi_{t|t-1,j}^{\theta,N'}, \cdot)$ with probability $w_{t,j}^{\theta,N'} - w_{t,j}^{\theta,N} \wedge w_{t,j}^{\theta,N'}$;

– for $j \in J_t^c \cap \{1, 2, \ldots, N\}$,
  – $\xi_{t+1|t,i}^{\theta,N} \sim Q_\theta(\xi_{t|t-1,j}^{\theta,N}, \cdot)$ with probability $w_{t,j}^{\theta,N}$;
  – $\xi_{t+1|t,i}^{\theta,N'} \sim Q_\theta(\xi_{t|t-1,j}^{\theta,N'}, \cdot)$ with probability $w_{t,j}^{\theta,N'}$.

Finally, for $N < i \leq N'$, $\xi_{t+1|t,i}^{\theta,N'} \sim Q_\theta(\xi_{t|t-1,j}^{\theta,N'}, .)$ with probability $w_{t,j}^{\theta,N'}$.

We let J_{t+1} be the set of indices $1 \leq i \leq N$ such that the first of the above events happened.

From this construction it is immediate that the distributions of the two filters are the same as if they had been run separately and independently in the usual manner. Let $\pi_{t|t-1}^{\theta,N}$ be the particle filter approximation to the predictive distribution at time index t;

$$\pi_{t|t-1}^{\theta,N}(A) = \frac{1}{N} \sum_{i=1}^{N} I_A(\xi_{t|t-1,i}^{\theta,N})$$

for all $A \in \mathcal{B}(\Theta)$, where I_A is the indicator function of A.

Proposition B.1. *Assume that observations $y_{1:T}$ are given and that there is a number $\underline{r} > 0$ such that $\underline{r} \leq r_\theta(y_t|s) \leq 1/\underline{r}$ for all $1 \leq t \leq T$, all s in the state space and all $\theta \in \Theta$. Then there are constants C_t for $1 \leq t \leq T$ such that for any integers $N' > N > 0$,*

$$\mathbb{E}\|\pi_{t|t-1}^{\theta,N} - \pi_{t|t-1}^{\theta,N'}\|_{\mathrm{TV}} \leq C_t \left(\frac{N' - N}{N}\right).$$

The constants C_t depend on $\underline{r}$, but otherwise the bound is uniform in θ. Therefore this result implies (B.18) for $\psi^N(\theta) = \int h(s)\, \pi_{t|t-1}^{\theta,N}(ds)$ whenever h is bounded on the state space of (S_t).

The requirement of a lower bound $\underline{r} > 0$ on r_θ, uniform in θ and s, will typically be satisfied only if both Θ and the state space of (S_t) are compact, or at least bounded. Boundedness of Θ is as good as implied by Hypothesis 1, whereas boundedness of the state space is a more serious limitation. Having said that we notice that this condition is recurring in the literature on particle filters, in particular when treating forgetting properties.

Proof of Proposition B.1. For any $A \in \mathcal{B}(\Theta)$,

$$\begin{aligned}
|\pi_{t|t-1}^{\theta,N}(A) - \pi_{t|t-1}^{\theta,N'}(A)| &\leq \left| \frac{1}{N} \sum_{i=1}^{N} I_A(\xi_{t|t-1}^{\theta,N}) - \frac{1}{N'} \sum_{i=1}^{N'} I_A(\xi_{t|t-1}^{\theta,N'}) \right| \\
&= \left| \sum_{i \in J_t}^{N} I_A(\xi_{t|t-1}^{\theta,N}) \left(\frac{1}{N} - \frac{1}{N'}\right) \right. \\
&\quad \left. + \frac{1}{N} \sum_{i \in J_t^c,\, i \leq N} I_A(\xi_{t|t-1}^{\theta,N}) + \frac{1}{N'} \sum_{i \in J_t^c} I_A(\xi_{t|t-1}^{\theta,N'}) \right| \\
&\leq \#J_t \left(\frac{1}{N} - \frac{1}{N'}\right) + \#J_t^c \left(\frac{1}{N} + \frac{1}{N'}\right) \\
&\leq 1 - \frac{N}{N'} + \#J_t^c \left(\frac{1}{N} + \frac{1}{N'}\right), \qquad \text{(B.21)}
\end{aligned}$$

where # denotes cardinality of a set and, in the last step, $\#J_t$ was bounded by N. We now seek to bound $\mathbb{E}(\#J_t^c)$.

Put $p_t = 1 - \#J_t/N'$ and define the σ-field $\mathcal{F}_t = \sigma(\xi^{\theta,N}_{t|t-1,i},\, 1 \le i \le N) \vee \sigma(\xi^{\theta,N'}_{t|t-1,i},\, 1 \le i \le N')$. Then conditionally on $\mathcal{F}_t$, $\#J_{t+1}$ is a binomial random variable with parameters N and $\sum_{i\in J_t}(w^{\theta,N}_{t,i} \wedge w^{\theta,N'}_{t,i})$. Using the definition of J_t and abbreviating $r_\theta(y_t|s)$ as $r_{t,\theta}(s)$, we find that

$$\begin{aligned}
\sum_{i\in J_t} w^{\theta,N}_{t,i} \wedge w^{\theta,N'}_{t,i} &= \frac{\sum_{i\in J_t} r_{\theta,t}(\xi^{\theta,N}_{t|t-1,i})}{\sum_{1\le i\le N} r_{\theta,t}(\xi^{\theta,N}_{t|t-1,i})} \left(1 \wedge \frac{\sum_{1\le i\le N} r_{\theta,t}(\xi^{\theta,N}_{t|t-1,i})}{\sum_{1\le i\le N'} r_{\theta,t}(\xi^{\theta,N'}_{t|t-1,i})}\right) \\
&= \left(\frac{1}{1 + \frac{\sum_{i\in J_t^c,\, i\le N} r_{\theta,t}(\xi^{\theta,N}_{t|t-1,i})}{\sum_{i\in J_t} r_{\theta,t}(\xi^{\theta,N}_{t|t-1,i})}}\right) \\
&\quad \times \left(1 \wedge \frac{1 + \frac{\sum_{i\in J_t^c,\, i\le N} r_{\theta,t}(\xi^{\theta,N}_{t|t-1,i})}{\sum_{i\in J_t} r_{\theta,t}(\xi^{\theta,N}_{t|t-1,i})}}{1 + \frac{\sum_{i\in J_t^c} r_{\theta,t}(\xi^{\theta,N'}_{t|t-1,i})}{\sum_{i\in J_t} r_{\theta,t}(\xi^{\theta,N}_{t|t-1,i})}}\right) \\
&\ge \frac{1}{1 + \underline{r}^{-2}\frac{p_t}{1-p_t}} \times \frac{1}{1 + \underline{r}^{-2}\frac{p_t}{1-p_t}} \\
&= \frac{1}{\left(1 + \underline{r}^{-2}\frac{p_t}{1-p_t}\right)^2} =: u(p_t).
\end{aligned}$$

We note that as u is convex and decreasing with $u(0) = 1$, there exists a constant $C_u > 0$ such that $u(p) \ge 1 - C_u p$.

The above-mentioned conditional binomial distribution of $\#J_t$ implies, together with the above inequality, that $\mathbb{E}(\#J_{t+1} \,|\, \mathcal{F}_t) \ge N u(p_t)$, and therefore

$$\begin{aligned}
\mathbb{E}(p_{t+1} \,|\, \mathcal{F}_t) &\le 1 - \frac{N u(p_t)}{N'} \\
&= 1 - u(p_t) + \frac{N' - N}{N'} u(p_t) \\
&\le 1 - u(p_t) + \frac{N' - N}{N'} \\
&=: v(p).
\end{aligned}$$

Applying this inequality recursively, it follows that $\mathbb{E}(p_t) \le v^{\circ t}(p_0)$, where superindex '$\circ t$' means t-fold function composition.

We notice that $p_0 = (N' - N)/N'$ and $v(p) \le C_u p + p_0 \le C_v(p + p_0)$ for some constant C_v, and by induction we find that there is a constant $C_{v,t}$ such that $v^{\circ t}(p) \le C_{v,t}(p + p_0)$. Thus $\mathbb{E}(\#J_t^c) = N'\mathbb{E}(p_t) \le 2N' C_{v,t} p_0 = 2C_{v,t}(N' - N)$. The proof is finished by inserting this bound into the right-hand side of (B.21), then taking the supremum over $A \in \mathcal{P}(\Theta)$ on the left-hand side and finally the expectation.

□