Using convolution to estimate the score function for intractable state transition models

Liang Dai and Thomas B. Schön

Department of Information Technology, Uppsala University, 751 05 Uppsala, Sweden. E-mail: {liang.dai, thomas.schon}@it.uu.se

Abstract

This note revisits a recent result for the score function estimation by making use of a basic property of the convolution operator. With a probabilistic view of the convolution property, the score function can be estimated as a function of the mean of the parameter posterior distribution, when a pseudo prior distribution for the parameter is introduced. In the note, the pseudo prior for the parameter is assumed to be Gaussian, which is a common choice in practice. In the end, a toy numerical experiment is implemented to show the efficacy of the estimator, where a particle MCMC method is used to sample from the parameter posterior distribution.

I. INTRODUCTION

In this note, we consider the problem of estimating the parameters $\theta$ in the following nonlinear state space model

$$x_{t+1}|x_t \sim f_\theta(x_{t+1}|x_t),$$

$$y_t|x_t \sim g_\theta(y_t|x_t),$$

where $x_t$ denotes the state and $y_t$ denotes the measurement. Furthermore, we assume that the system transition kernel $f_\theta(\cdot)$ is intractable, which will be explained shortly. This scenario can occur in many places, for example, when the state transition is given by nonlinear stochastic differential equations, and the observations are obtained at discrete time instances. In such
scenarios, it is often impossible to obtain a closed form expression for the transition kernel between the consecutive observation times. Note that although the expression for the transition kernel is intractable, we can still simulate the model to obtain samples of the states.

The Maximum (log-)Likelihood (ML) method is one attractive way for parameter estimation due to its large sample statistical efficiency [9]. To numerically solve the ML optimization problem, the gradient of the log-likelihood function, i.e. the score function, is often needed. When the state transition kernel is tractable, Fisher’s identity, together with a particle smoother can do the job [8]. However, for the intractable case, the task becomes more challenging. Recently, [3] introduced an approach for score function estimation in the intractable case by introducing a pseudo prior for the parameters. The score function can then be given as a function of the mean value of the parameter posterior distribution. Note that the method proposed in [3] is related to the work in [6], [7], which introduce the iterated filtering algorithm, a stochastic approximation based method for ML parameter estimation where the parameter is modeled as a random walk with diminishing variance.

The contribution of this note is that we provide an alternative derivation of the estimator in [3] by making use of a basic property of the convolution operator. Note that an idea close to the one presented here has been independently developed by the authors of [3], see their recent preprint on arxiv [4]. The main difference between this note and [4] lies in that the results in [4] directly utilize Stein’s identity, while this note starts from an elementary property of the convolution operator which can offer additional insights. Note that in both the current work and the results in [4], the pseudo prior distribution for the parameter is assumed to be Gaussian.

This note is structured as follows. In next section, the convolution operator and Stein’s identity are introduced and their connection is established. We will then illustrate how to leverage the connection to estimate the score function. After that, we briefly discuss the second order derivative estimation of the log likelihood function. Then, a numerical example is conducted to show the efficacy of the results. Finally, we conclude the paper with some open questions. In the letter, we will stick to the univariate case. The time indexes will also be dropped when no confusion occurs.

In the following, the double factorial $k!!$ denotes $k \times (k - 2) \cdots \times 3 \times 1$ when $k$ is odd, and $k \times (k - 2) \cdots \times 4 \times 2$ when $k$ is even.
II. THE CONVOLUTION OPERATOR AND STEIN’S IDENTITY

We first describe the convolution operator and one of its relevant properties for the later use. For two functions \( f(\alpha) : \mathbb{R} \to \mathbb{R} \) and \( g(\alpha) : \mathbb{R} \to \mathbb{R} \), the convolution of them is defined as

\[
(f \ast g)(\theta_0) = \int_{-\infty}^{+\infty} f(\theta)g(\theta_0 - \theta) \, d\theta = \int_{-\infty}^{+\infty} f(\theta_0 - \theta)g(\theta) \, d\theta.
\]

If both functions are differentiable and satisfying certain integrability conditions, the following basic property holds

\[
(f \ast g')(\theta_0) = (f' \ast g)(\theta_0),
\]

which can be written as

\[
\int_{-\infty}^{+\infty} f(\theta)g'(\theta_0 - \theta) \, d\theta = \int_{-\infty}^{+\infty} f'(\theta)g(\theta_0 - \theta) \, d\theta.
\]

Let \( g(\theta) = \mathcal{N}(\theta : 0, \tau^2) \), with a probabilistic view of (2), it gives that

\[
\tau^{-2} \mathbb{E}_\theta[(\theta - \theta_0)f(\theta)] = \mathbb{E}_\theta[f'(\theta)],
\]

where the expectation is taken with respect to \( \mathcal{N}(\theta : \theta_0, \tau^2) \).

Note that this is commonly referred to as Stein’s identity [10]. The integration by parts approach can also be used in establishing Stein’s identity. The benefit of the proposed approach is that it gives more intuition of the identity, by linking it to the well-known concept of convolution. Besides this, it can also be used to estimate the second (possibly higher) order derivative of the log-likelihood function, which will be discussed in Section IV. In next section, we will apply the property in Eq. (2) for the score function estimation.

III. SCORE FUNCTION ESTIMATION

A. Asymptotic analysis

Notice that the following result holds

\[
\lim_{\tau \to 0} \mathcal{N}(\theta : \theta_0, \tau^2) = \delta(\theta - \theta_0),
\]

where the \( \delta(\cdot) \) indicates the Dirac delta function. This allows us to conclude that the right hand side of Eq. (4) can approximate \( f'(\theta_0) \) well in the limit when \( \tau \) approaches zero. In turn, the left hand side is also a reasonable proxy for \( f'(\theta_0) \).
Assume that the likelihood of the data $y$ is given by $p(y|\theta)$, where $\theta$ is the system parameter.

Further, we denote

$$ l(\theta) = \log(p(y|\theta)) $$

and its first order derivative as

$$ l^{(1)}(\theta) = \frac{p'(y|\theta)}{p(y|\theta)}. $$

The task for us is to estimate $l^{(1)}(\theta)$ at $\theta_0$.

Notice that we can link the relation given in Eq. (4) to the dynamical system as follows. Let $f(\theta)$ in Eq. (4) be given by $f(\theta) = \frac{p(y|\theta)}{p(y)}$, we have that

$$ \tau^{-2} \mathbb{E}_{\theta_0,\tau} \left( (\theta - \theta_0) \frac{p(y|\theta)}{p(y)} \right) = \mathbb{E}_{\theta} \left( \frac{p'(y|\theta)}{p(y)} \right). \quad (6) $$

Based on Eq. (6), we make the following observations:

- Direct verification gives that the left hand side of Eq. (6) is the expectation of $(\theta - \theta_0)$ with respect to the posterior distribution $p(\theta|y)$ if we regard $\mathcal{N}(\theta : \theta_0, \tau^2)$ as the prior distribution of $\theta$. Later on, we will denote the left hand side of Eq. (6) by $\tau^{-2} \mathbb{E}_{\theta_0,\tau} (\theta - \theta_0)$, as in [3].
- Given Eq. (5), we have that

$$ p(y) = \int p(y|\theta)\mathcal{N}(\theta : \theta_0, \tau^2)d\theta \rightarrow p(y|\theta_0), $$

when $\tau \rightarrow 0$, which means that the right hand side of Eq. (6) will converge to the score function at $\theta_0$ when $\tau \rightarrow 0$. In turn, it implies that the left hand side of Eq. (6) will provide a reasonable estimate of the score function at $\theta_0$ when $\tau$ is small.

To summarize, this analysis offers an intuition of the feasibility in exploiting the property in Eq. (2) for the score stimulation, and the estimator is given by the quantity $\tau^{-2} \mathbb{E}_{\theta_0,\tau} (\theta - \theta_0)$.

Remark 1: Note that when an expression of the state transition kernel is not explicitly known, Fisher’s identity [8] is not applicable anymore. However, it is still possible to sample from the posterior distribution $p(\theta|y)$ by using particle MCMC methods [1], which provides a Monte Carlo approximation to the left hand side of Eq. (6), and in turn an approximation of the score function.

Remark 2: Note that if we let $g(\theta)$ be a probability density function which belongs to the exponential family, then by the convolution property in (2), a generalized Stein’s identity, similar to the one in Eq. (4), can also be established, see the results in [5]. With this generalization, an estimator of the score function estimation can be found analogously to what was done before.
Note that, in this case, the pseudo prior introduced for the parameter does not necessarily have to be Gaussian.

B. Convergence rate analysis

This section establishes the non-asymptotic analysis of the estimator. More specifically, we will establish the following result.

**Proposition 1:** Assume that \( p(y|\theta) \) is analytic and for given \( \theta_0 \) and \( \tau \), there exist \( C_1(\theta_0, \tau) \) and \( C_2(\theta_0, \tau) \) defined as

\[
C_1(\theta_0, \tau) = \frac{1}{p(y|\theta_0)} \left| \sum_{k \text{ is even}} \frac{p^{(k+1)}(y|\theta_0)\tau^{k-2}}{k!!} \right|
\]  

and

\[
C_2(\theta_0, \tau) = \left| \left( \sum_{k \text{ is even}} \frac{p^{(k)}(y|\theta_0)\tau^{k-2}}{k!!} \right) \int \frac{p'(y|\theta)N(\theta : \theta_0, \tau^2) d\theta}{p(y|\theta_0)p(y)} \right|.
\]  

Further, assume that the series \( \sum_{k=1}^{\infty} c_k(\theta) \) and \( \sum_{k=1}^{\infty} s_k(\theta) \) defined in Eq. (18) and Eq. (22) are uniform convergent series on \( \mathbb{R} \), then we have that

\[
|l^{(1)}(\theta_0) - \tau^{-2}E_{\theta_0,\tau}(\theta - \theta_0)| \leq C(\theta_0, \tau)\tau^2,
\]  

where \( C(\theta_0, \tau) = C_1(\theta_0, \tau) + C_2(\theta_0, \tau) \).

**proof 1:** By making use of (6) and that \( \mathbb{E}_\theta \left( \frac{p'(y|\theta_0)}{p(y|\theta_0)} \right) = \frac{p'(y|\theta_0)}{p(y|\theta_0)} \), the error in Eq. (9) can be decomposed according to

\[
|l^{(1)}(\theta_0) - \tau^{-2}E_{\theta_0,\tau}(\theta - \theta_0)| = \left| \frac{p'(y|\theta_0)}{p(y|\theta_0)} - \mathbb{E}_\theta \left( \frac{p'(y|\theta)}{p(y)} \right) \right| = \left| \mathbb{E}_\theta \left( \frac{p'(y|\theta_0)}{p(y|\theta_0)} - \frac{p'(y|\theta)}{p(y)} \right) \right| = \left| \mathbb{E}_\theta T_1(\theta) \right| + \left| \mathbb{E}_\theta T_2(\theta) \right|,
\]

where

\[
T_1(\theta) = \frac{p'(y|\theta_0)}{p(y|\theta_0)} - \frac{p'(y|\theta)}{p(y)}
\]

and

\[
T_2(\theta) = \frac{p'(y|\theta)}{p(y|\theta_0)} - \frac{p'(y|\theta)}{p(y)}.
\]

It now remains to bound the terms \( \left| \mathbb{E}_\theta T_1(\theta) \right| \) and \( \left| \mathbb{E}_\theta T_2(\theta) \right| \) separately. These bounds are derived in the appendix.
Remark 3: Note that although the convergence rates of the estimator in both the current work and [4] are the same, the derivation strategies are different. In [4], the authors rely on techniques from Bayesian asymptotic theory to directly analyze the convergence rate of the estimator, i.e. the left hand side of Eq. (4), while the current work tries to bound the right hand side of Eq. (4). This allows us to decompose the error into separate terms, where each term is relatively easy to bound.

IV. ESTIMATING HIGHER ORDER DERIVATIVES

Given Eq. (2), it also holds true that for \( l \geq 1 \), we have

\[
(f^{(l)}(\theta) * g(\theta))(\theta_0) = (f(\theta) * g^{(l)}(\theta))(\theta_0),
\]

where \( f^{(l)} \) denotes the \( l \)-th derivative of the function \( f \). This property makes it possible to generalize Stein’s identity to higher order cases and it also opens for the possibility of estimating higher order derivatives of the log-likelihood.

Next, we discuss the case when \( l = 2 \) by using the property given in Eq. (10). Let \( g(\theta) = \mathcal{N}(\theta : 0, \tau^2) \), and \( f(\theta) = \frac{p(y|\theta)}{p(y)} \), then according to Eq. (10), we have that

\[
-\tau^{-2} + \tau^{-4} \mathbb{E}_{\theta_0, \tau} (\theta - \theta_0)^2 = \mathbb{E}_{\theta} \left( \frac{p''(y|\theta)}{p(y)} \right),
\]

where \( \mathbb{E}_{\theta_0, \tau}(\cdot) \) is taken with respect to posterior distribution of \( \theta \), and \( \mathbb{E}_{\theta}(\cdot) \) is taken with respect to the pseudo prior distribution \( \mathcal{N}(\theta : \theta_0, \tau^2) \).

Furthermore, taking the square for both sides of Eq. (6), we have that

\[
\tau^{-4} \mathbb{E}_{\theta_0, \tau}^2 (\theta - \theta_0) = \mathbb{E}_{\theta}^2 \left( \frac{p'(y|\theta)}{p(y)} \right).
\]

Subtracting Eq. (12) from Eq. (11), we have that

\[
-\tau^{-2} + \tau^{-4} \nabla_{\theta_0, \tau}(\theta) = \mathbb{E}_{\theta} \left( \frac{p''(y|\theta)}{p(y)} \right) - \mathbb{E}_{\theta}^2 \left( \frac{p'(y|\theta)}{p(y)} \right),
\]

where \( \nabla_{\theta_0, \tau}(\theta) \) denotes the variance of \( \theta \) taken with respect to the posterior distribution of \( \theta \). Notice that when \( \tau \to 0 \), the right hand side of Eq. (13) will converge to

\[
\frac{p''(y|\theta_0)}{p(y|\theta_0)} - \left( \frac{p'(y|\theta_0)}{p(y|\theta_0)} \right)^2 = l^{(2)}(\theta) |_{\theta = \theta_0}.
\]

Hence, the left hand side of Eq. (13) is a reasonable proxy for \( l^{(2)}(\theta) |_{\theta = \theta_0} \), which can also be sampled using PMCMC methods analogously to what was done previously in the first order case.
V. A NUMERICAL ILLUSTRATION

In this section, we will test the estimator by applying it to a simple toy example. We consider the first order autoregressive system as follows

\[ x_{t+1} = \theta x_t + v_t, \]  
\[ y_t = x_t + w_t, \]

in which \( \{v_t\}_{t=1}^{T}, \{w_t\}_{t=1}^{T} \) and \( x_1 \) are all standard normal distributed random variables, \( T = 5 \), and the true \( \theta \) that was used in generating the data is given by \( \theta_0 = 0.8 \).

We want to estimate the score function at \( \theta_0 \) given the data \( \{y_t\}_{t=1}^{T} \). In the experiment, we make use of the Particle Metroplis-Hastings (PMH) sampler [1] to sample from the posterior distribution for the system parameter, that is \( p(\theta|y_{1:T}) \). Inside PMH, the bootstrap particle filter is used with \( N = 50 \) particles. The Markov chain is run for \( M = 30\,000 \) iterations, and the first 15\,000 samples are discarded to avoid the burn-in period. Note that, for such a linear Gaussian system, it is possible to evaluate the exact score function by making use of the so-called sensitivity derivatives advocated in [2]. This method is also used in this note to compute the exact value of the score function. The experimental results are reported in the Figure 1, please also find further details in the figure’s caption.

VI. CONCLUSION

In this note, we revisited the score function estimation problem by making use of a basic property of the convolution operator. We also illustrate the results with a toy example for which the true score function can be analytically computed. Empirically, we have observed that the numerical error will increase when the variance of the pseudo prior shrinks towards zero or becomes large. How to find the optimal value for the variance of the pseudo prior is unknown yet. Also, note that the connection between the convolution property and the score estimation is not specific for the state space models, hence it is interesting to study how the connection could be generalized to other applications.
Fig. 1. Boxplot of the average squared error for the estimates of the score function (at \( \theta_0 \)) when PMH is used to sample from the posterior distribution. The score value at \( \theta_0 \) obtained by the sensitivity analysis method is \(-1.1960\). The pseudo prior introduced is \( \mathcal{N}(\theta : \theta_0, \tau) \), where \( \tau \) ranges from 0.005 to 0.050 with the increment 0.005. For each \( \tau \), we run 20 independent simulations. Note that, as we can see from the plot, the best result is obtained when \( \tau \) is 0.02.

VII. APPENDIX

We first bound \( |\mathbb{E}_y T_1(\theta)| \) as follows.

\[
|\mathbb{E}_y T_1(\theta)| = \left| \int \left( \frac{p'(y|\theta_0)}{p(y|\theta_0)} - \frac{p'(y|\theta)}{p(y|\theta_0)} \right) \mathcal{N}(\theta : \theta_0, \tau^2) d\theta \right|
\]

\[
= \frac{1}{p(y|\theta_0)} \left| \int (p'(y|\theta_0) - p'(y|\theta)) \mathcal{N}(\theta : \theta_0, \tau^2) d\theta \right|
\]

\[
= \frac{1}{p(y|\theta_0)} \left| \int \left( \sum_{k=1}^{\infty} \frac{p^{(k+1)}(y|\theta_0)}{k!} (\theta - \theta_0)^k \right) \mathcal{N}(\theta : \theta_0, \tau^2) d\theta \right| 
\]

\[
= \frac{1}{p(y|\theta_0)} \left| \sum_{k=1}^{\infty} \frac{p^{(k+1)}(y|\theta_0)}{k!} \int (\theta - \theta_0)^k \mathcal{N}(\theta : \theta_0, \tau^2) d\theta \right|
\]

\[
= \frac{1}{p(y|\theta_0)} \left| \sum_{k \text{ is even}} \frac{p^{(k+1)}(y|\theta_0)\tau^k(k-1)!!}{k!} \right| 
\]

\[
= C_1(\theta_0, \tau)^2.
\]

In the previous derivations, in Eq. (16), we have made use of the Taylor expansion of the
function $p(y|\theta)$ around $\theta_0$. Furthermore, let
\begin{equation}
    c_k(\theta) = \frac{p^{(k+1)}(y|\theta_0)}{k!} \mathcal{N}(\theta : \theta_0, \tau^2)(\theta - \theta_0)^k,
\end{equation}
then by the uniform convergence assumption of $\sum_{k=1}^\infty c_k(\theta)$, the order of the integration and the infinite summation in Eq. (16) can be interchanged. In Eq. (17), we have applied the central moments of the normal distribution functions [11].

**Remark 4:** Notice that if we assume $C_1(\theta_0, \tau)$ to exist, then we have that $\lim_{\tau \to 0} C_1(\theta_0, \tau) = \frac{p^{(2)}(y|\theta_0)}{2p(y|\theta_0)}$.

Next, we proceed to establish a bound for $|E\theta T_2(\theta)|$. Denote
\[ \Delta = p(y) - p(y|\theta_0), \]
then we have that
\begin{align}
    |E\theta T_1(\theta)| &= \left| \int \frac{p'\left(y\left|\theta\right.\right)}{p(y|\theta_0)} - \frac{p'\left(y\left|\theta\right.\right)}{p(y)} \mathcal{N}(\theta : \theta_0, \tau^2) d\theta \right| \\
    &= \frac{\Delta}{p(y|\theta_0)p(y)} \left| \int p'\left(y\left|\theta\right.\right)\mathcal{N}(\theta : \theta_0, \tau^2) d\theta \right| \\
    &= |\Delta| \left| \int p'\left(y\left|\theta\right.\right)\mathcal{N}(\theta : \theta_0, \tau^2) d\theta \right| \frac{p(y|\theta_0)p(y)}{p(y|\theta_0)p(y)} \\
    &= |\Delta| \left| \int p'\left(y\left|\theta\right.\right)\mathcal{N}(\theta : \theta_0, \tau^2) d\theta \right| \\
    &= \left| \sum_{k \text{ is even}} \frac{p^{(k)}(y|\theta_0)}{k!} \tau^k \int (\theta - \theta_0)^k \mathcal{N}(\theta : \theta_0, \tau^2) d\theta \right|.
\end{align}

We will proceed by deriving a bound on $|\Delta|$. Let us start by noting that
\begin{align}
    |\Delta| &= |p(y) - p(y|\theta_0)| \\
    &= \left| \int (p(y|\theta) - p(y|\theta_0)) \mathcal{N}(\theta : \theta_0, \tau^2) d\theta \right| \\
    &= \left| \int \left( \sum_{k=1}^\infty \frac{p^{(k)}(y|\theta_0)}{k!} (\theta - \theta_0)^k \right) \mathcal{N}(\theta : \theta_0, \tau^2) d\theta \right| \\
    &= \left| \sum_{k=1}^\infty \frac{p^{(k)}(y|\theta_0)}{k!} \int (\theta - \theta_0)^k \mathcal{N}(\theta : \theta_0, \tau^2) d\theta \right| \\
    &= \left| \sum_{k \text{ is even}} \frac{p^{(k)}(y|\theta_0)}{k!!} \tau^k \right|.
\end{align}

Similar as before, due to the uniform convergence assumption on $\sum_{k=1}^\infty s_k(\theta)$, where
\begin{equation}
    s_k(\theta) = \frac{p^{(k)}(y|\theta_0)}{k!} \mathcal{N}(\theta : \theta_0, \tau^2)(\theta - \theta_0)^k,
\end{equation}
the order of integration and the infinite summation in Eq. (20) can be interchanged. By inserting Eq. (21) into Eq. (19), we obtain
\begin{equation}
    |E\theta T_1(\theta)| = C_2(\theta_0, \tau) \tau^2.
\end{equation}
Remark 5: Similar to the remark on $C_1(\theta_0, \tau)$, if we also assume that, for given $\theta_0$, $C_2(\theta_0, \tau)$ exists, then we have that
\[
\lim_{\tau \to 0} C_2(\theta_0, \tau) = \frac{p'(y|\theta_0)p^{(2)}(y|\theta_0)}{|p(y|\theta_0)|^2}.
\]
Combing all the previous results, we have that
\[
\left| \ell^{(1)}(\theta_0) - \tau^{-2}\mathbb{E}_{\theta_0,\tau}(\theta - \theta_0) \right| \leq C(\theta_0, \tau)\tau^2. \tag{24}
\]

REFERENCES