

Prof. Dr. Manfred Opper
Artificial Intelligence Group (retired)
Fakultät IV - Elektrotechnik und Informatik
TU Berlin

Gutachten zu:

Deterministic particle flows for stochastic nonlinear systems

vorgelegt von **Dimitra Despoina Maoutsa**

This thesis addresses three important computational problems in the field of stochastic dynamical systems which are described by stochastic differential equations (SDE). These are: efficient ways of simulating the temporal flow of the marginal densities, the optimal control of such systems and finally, the learning of the driving function of the SDE from an observed trajectory. These problems are of importance in many areas of research ranging from statistical physics over computational biology to machine learning.

In contrast to previous approaches to tackle these problems, Dimitra Maoutsa develops and investigates a novel framework which is based on ensembles of particles having an entirely *deterministic* dynamics which is constructed in such a way that marginal densities agree with those of the stochastic systems. The interaction between the particles is mediated by the gradient of the log-density of the particles, the so-called *score functions*, for which independently, empirical estimators have been developed in the field of machine learning in recent years.

The first chapter gives a short motivation and overview of the thesis, followed by a concise summary of the necessary mathematical background from stochastic analysis and machine learning. Chapter three is devoted to the role and practical estimation of score functions. This is a very active topic of research in machine learning and Dimitra Maoutsa presents a detailed overview of current approaches. The specific method developed for the thesis is based on kernel representations which need only first derivatives of kernels which leads to a reduced computational complexity. The chapter concludes with an empirical study of statistical properties (bias, variance) of the estimator. It would be interesting to further discuss the surprising increase of the variance with sample size (Fig. 3.9). How where the spatial averages obtained ?

The fourth chapter deals with a novel method for solving *Fokker-Planck equations* (FPE) which describe the temporal flow of the marginal density of SDE. After a detailed review of existing methods, the new deterministic approach is derived from a representation of the

FPE as a *Liouville equation* for a system of deterministic particles with an effective drift which is proportional to the score function. The estimator discussed in chapter three leads to practical algorithms which can outperform stochastic sampling for a fixed number of particles. The properties of the method and relations to other particle approaches as well as detailed numerical investigations are shown. It would be interesting to understand if the simulations leading to Fig.2 can be viewed as a realisation of the algorithm with linear and quadratic basis functions. What is the relation of the *Benamou–Brenier* approach mentioned on page 122 with the particle approach ?

Chapter five presents a generalisation of the new method to tackle a class of problems in optimal stochastic control for SDE, which have attracted considerable interest in statistical physics and machine learning as the *path–integral control* problem. The goal is to compute an extra additive force (the control) to the drift of the model, for which an expected cost functional (comprised of a specific control energy, path cost and end cost) is minimised. Dimitra Maoutsa derives a classical *Hamilton–Jacobi–Bellman* solution to this problem and a second derivation based on stochastic calculus which both result in a linear backward–Kolmogorov equation. She then reviews previous solution methods. The new approach is motivated by the role that the backward equation plays for smoothing problems in stochastic dynamics. It is shown how the control can be computed by combining a forward filtering equation with backward FPE. These equations are solved by the particle approach developed in the previous chapter. The particle reweighing in the forward filter requires an extra deterministic step involving an ensemble transform filter. A bit more of cross–referencing between chapters — Eq. (5.84) is obtained from (4.84) and (5.48) should be the same as (5.14)— would be helpful to follow the derivations. The method is studied and compared to a competing PICE approach on a variety of model systems, showing an overall good performance. About the discussion of results, it is not entirely clear to me why a small control energy is always beneficial. Shouldn’t one discuss the sum of control and path costs together ? Finally, a more detailed derivation of the population dynamics eq. (5.100) in an appendix would be beneficial for readers less familiar with mathematical biology.

Chapter six is devoted to an application of the control method developed in chapter five to the learning of drift functions from discrete time samples of the SDE dynamics. Dimitra Maoutsa presents a detailed review of drift estimation techniques. She shows how non–parametric estimators can be computed for the case of data which are sampled with high frequency. Their performance decreases considerably when the sampling time interval gets larger. In the latter case, one has to impute the unobserved SDE paths between observations and one can learn the drift by using a EM algorithm. The E–step of such an algorithm requires the computation of an expectation over SDE paths *conditioned* on the end points defined by observed data. Methods for sampling from these so–called SDE *bridges* can be derived from a control perspective which has been developed in the previous chapter five. A major problem arising from a direct application of these methods comes from the fact that end points are often highly improbable with respect to the uncontrolled dynamics. Hence, there are not enough particles from the forward filtering to give reliable estimates for the backward process. Dimitra Maoutsa develops a novel, geometric approach for dealing with

this problem. This geometric approach was developed by her alone without contribution of her supervisor. The main idea is to define a metric in the space of states of the SDE which is constructed from (samples of) the stationary density. From this one can compute geodesics between consecutive data points which can serve as proxies for typical paths. By adding an extra path cost term (which keeps paths close to the geodesics) to the optimal control problem, one obtains bridges of much better quality, which lead to an improved drift estimation in the EM algorithm. The performance and limitations of the new method is critically studied and discussed for two model systems. I think that this is a creative novel and promising idea. Its theoretical foundation needs more investigation: What is the metric in the large data limit ? How are geodesics related to the most likely paths in the *Onsager–Machlup* framework ?

Finally, the seventh chapter concludes with a summary of contributions and an outlook on future work.

This is a very good and successful thesis. It combines a broad variety of mathematical techniques (stochastic processes, optimal stochastic control, nonlinear filtering, path integrals, Riemannian geometry) with estimation techniques of machine learning to obtain new particle approaches for SDE and FPE. I expect that these approaches will stimulate further research in the field.

All in all, an impressive piece of work. Hence, I grade the thesis with:

Very Good (Sehr Gut).



(Professor Dr. Manfred Opper)

Prof. Dr.-Ing. Rolf Schuhmann
Fakultät IV – Elektrotechnik und Informatik
Technische Universität Berlin
Marchstraße 23
10587 Berlin

Fakultät 1 (MINT)
Mathematik, Informatik, Physik,
Elektro- und Informationstechnik
Lehrstuhl für Stochastik
und ihre Anwendungen

Prof. Dr. Carsten Hartmann

Konrad-Wachsmann-Allee 1
03046 Cottbus
Germany

T +49 (0)355 69 4150
F +49 (0)355 69 3595
E carsten.hartmann@b-tu.de

Cottbus, 15. März 2023

Gutachten zur Dissertation von Dimitra Despoina Maoutsa

Sehr geehrter Professor Schuhmann,

ich schicke Ihnen mein Gutachten zur Dissertation von Dimitra Maoutsa mit dem Titel *Deterministic particle flows for stochastic nonlinear systems*.

Background and overview

Sampling the stationary or non-stationary probability law of a diffusion process has been a standard problem in computational statistics for decades. The applications range from statistical mechanics to meteorology, systems biology, or material science, to name just a few. For stationary processes, one of the standard approaches to sample a probability distribution by means of a single long realisation of the process is Markov Chain Monte Carlo (MCMC). While MCMC is a powerful tool that allows for bias-free approximation of stationary (i.e. equilibrium) distributions, it cannot be applied to non-stationary or transient processes; moreover its convergence rate usually degrades in high-dimensions, unless clever MCMC proposals are used that require good a priori knowledge of the system. As an alternative, algorithms based on interacting particle systems that are based on either deterministic or stochastic flows that transform the target distribution into a simpler (e.g. Gaussian) distribution that can be easily sampled from have been recently proposed.

Dimitra Maoutsa's PhD thesis is concerned with deterministic transport algorithms for interacting particle systems (IPS). The key idea is (a) to reformulate the parabolic Fokker-Planck equation for the finite time marginal density of the process as a first-order transport equation of Liouville type and (b) to devise suitable approximations of the gradient of the log density (called "score") of the process rather than the (log) density as is commonly done.

Dimitra Maoutsa discusses various approximation procedures and estimators for the score function and applies them in the context of control and inference of diffusion processes. All theoretical derivations are accompanied and validated by extensive numerical experiments on both trivial toy examples and less trivial systems of practical relevance.

Some Specifics

The thesis is split into three main parts that are aligned with the keywords *simulation*, *control*, and *inference* in the thesis subtitle. After the introduction (Section 1) that contains a high-level overview of the key results of the thesis and a Section 2 that provides the mathematical tools relevant for this thesis (e.g. stochastic differential equations, divergence between probability measures, etc.), the first part (Sections 3 and 4) is concerned with score function approximations and IPS-based *simulation* of Fokker-Planck equations, the second part (Section 5) deals with the application of the IPS approach to the solution of so-called "linearly solvable" stochastic *control* problems, whereas the third part (Section

6) is devoted to an application to Bayesian *inference* with sparse observations. A summary of the key results and a brief prospect of future research are given in Section 7. The relevant Section 3–6 that contain the main results of the thesis are based on publications, each with Dimitra Maoutsa as first author who has been responsible for drafting, numerical implementation and tests, and writing.

The key idea outlined in Section 3 is to devise non-parametric estimators for the log gradient density based on a kernelisation of the score matching loss due to A. Hyvärinen (with an appropriate RKHS norm regularisation). By doing so, Dimitra Maoutsa can derive a closed-form expression for the score estimator that (a) does not require expensive computations of Hessian matrices of the density and (b) can be adapted so as to enforce sparsity of the kernel functions. The numerical tests show a good performance, similar to Stein’s method that also does not require the computation of Hessians. An interesting observation in Section 3.6 is that the bias in the gradient log density for a certain number $M \ll N$ of inducing points scales with $N^{-1/4}$ where N is the number of particles. While this behaviour is consistently observed over all the numerical examples, I could not find a possible explanation for this asymptotic behaviour of the bias. Moreover the authors describes that the variance of the estimator can saturated for $N \rightarrow \infty$ (see p. 52), which is another interesting observation that may be attributed to the interplay between M and N . The detailed study of the asymptotic properties of the estimator would be beyond the scope of the thesis, nevertheless the observations made here warrant further studies I suppose.

In Section 4, the non-parametric estimator is then embedded into an IPS framework to approximate both the time-dependent (“transient”) and stationary (“equilibrium”) solution of the Fokker-Planck equation (FPE). To this end, the FPE is recast as a Liouville-type continuity equation, which includes a velocity field that depends on the log gradient density and that can be estimated on the fly using the techniques outlined in Section 3. The solution of the Liouville equation is then simulated by sufficiently many interacting particles that follow the estimated deterministic velocity field, using the method of characteristics. (The velocity field is not really deterministic as it depends on the random samples that go into the gradient log estimator, but there is no extra diffusion that drives the particles.) The algorithmic idea is then tested with many different (e.g. linear and nonlinear, one- and multidimensional) examples and compared with a crude Monte-Carlo (CMC) method. In comparison with the CMC approximation The deterministic approach shows superior approximation of the modes of the distribution, consistently over the various examples, but less so of the tails of the distribution; this is not terribly surprising because the dynamics is non-diffusive, so the deterministic particles cannot access these small probability regions when are not present in gradient log estimator. (It would be interesting to compare this to the tail properties of Stein’s variational gradient descent algorithm that, interestingly enough, can be obtained from the approach developed in this thesis by changing the regularisation parameter as the author observes; see p. 91, first paragraph). The deterministic IPS needs far less particles than the stochastic CMC approximation, but it would have been interesting to compare the CMC and the deterministic IPS approach for a fixed computational budget, because I believe that there is a considerable computational overhead related to the gradient log estimation.

In Section 5, the IPS simulation algorithm is then applied to the solution of linear-quadratic stochastic optimal control problems on infinite time-horizon. For these problems, the dependence of the dynamics on the state can be nonlinear, but the control enters quadratically

into the cost function and affine-linearly in the equation. By design, these problems are linearly solvable, for the associated dynamic programming (also: HJB) equation can be transformed into a linear parabolic backward evolution equation by a logarithmic transformation. The linear equation has a probabilistic Feynman-Kac representation, so theoretically its solution could be computed by Monte Carlo. In practice, the CMC estimators may suffer from huge relative errors, and Dimitra Maoutsa has exploited a specific representation of the value function of the control problem in terms of a pair of adjoint linear evolution equations for a killed diffusion process that goes back to work by H. Kappen. (Here the killing rate is equal to the running cost of the underlying control problem.) As is shown, it is possible to formulate the problem of simulating the adjoint pair as a bridge sampling problem that boils down to solving a pair of forward-backward FPE. For the latter, the results from the previous section essentially carry over, and the Dimitra Maoutsa reports the results of extensive numerical experiments, including the controlled synchronisation of a 6-dimensional stochastic Kuramoto model as a nontrivial test case. For the latter, the author presents numerical results for the finite time horizon as well as for the infinite-time horizon case that can be dealt within the receding horizon or model predictive control framework. The numerical results are reported to be consistent with results obtained from CMC-based path integral algorithms. Even though the numerical results look convincing and the deterministic IPS method is reported to feature lower computational cost than path integral methods or iterative approaches based on forward-backward stochastic differential equations, the method is not plug-and-play; the author mentions several non-trivial modifications of the core algorithm that are necessary to guarantee the approximation fidelity for problems that involve both terminal and running costs, such as operator splitting, particle shifting or reweighting that may produce an additional computational overhead (see Section 5.7.1).

Finally, Section 6 is devoted to a Bayesian inference problem with sparse-in-time observations and unknown drift vector field. The key idea of what is called a “path augmentation scheme” is twofold: (a) assuming that the dynamics concentrates on a possibly low-dimensional (Riemannian) submanifold of the state space, Dimitra Maoutsa uses local covariance information of the observation data to estimate the Riemannian structure underlying the unobserved data, and then (b) employs the bridge sampling framework from the previous section to sample the Bayesian posterior under the geometric constraints imposed by the Riemannian structure. By the assumption that the unobserved components concentrate along the data manifold, the bridge sampling technique developed in Section 5 can be used to handle sparse data by filling the gaps between the filtered sparse state estimate. A feature of the path augmentation scheme is that the densely sampled paths can then be used to estimate the drift. By iterating state and drift estimation in an expectation-maximization like fashion, the initial estimates can then be refined until a prescribed tolerance is reached. The idea of the path augmentation approach is appealing and surely original, and it is supported by numerical experiments for the van-der-Pol oscillator and a stochastic Fitzhugh-Nagumo model. Nevertheless it is difficult to judge whether the assumptions of the data manifold and the path concentration property are really met for a system under consideration, which makes the scheme somewhat arbitrary (as the author admits; see p. 236, third paragraph). Moreover the description of the scheme is rather rudimentary, which is, however, hardly surprising since it comprises advanced techniques from very different mathematical fields, such as Bayesian statistics, control theory, Riemannian geometry, approximation theory, etc. that are not easily put into a coherent framework. As the discussion at the end of Section 6 shows, the author is aware of these difficulties and

list several possible issues and open problems, from lack of Markovianity to issues with non-trivial (e.g. state-dependent) diffusion coefficient, or problems to precisely estimating the dimension of the data manifold.

Critique

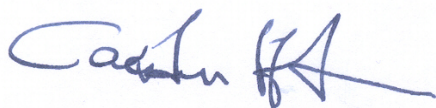
From a high-level perspective, the structure of the thesis is relatively clear, even though the three parts related to *simulation*, *control*, and *inference* are strongly interwoven. Since the chapters follow the outline of the author's publications, there are, however, some redundancies and repetitions here and there that makes it sometimes difficult to understand whether an argument or an equation from a previous Section is new or just reformulated. Yet, with few exceptions, Dimitra Maoutsa is very careful in formulating the underlying mathematical assumptions for an algorithm explicitly whenever this is possible. When this is not possible, the gaps are filled by carefully conducted numerical experiments. Therefore the theoretical considerations together with the numerical experiments provide a fairly complete picture of the pros and cons of the devised algorithms.

I have a few minor technical remarks (typos, definitions of symbols that are too far away from where the symbols appear first, improper references to equations, etc.), especially related to Section 2, that the author can take from my annotated copy of the thesis and that she can take into account when submitting the final version of her thesis. These minor remarks also comprise the aforementioned "few exceptions" regarding tacit assumptions that could be stated more clearly (e.g. in the discussion of the linear system in Section 4.8 where Dimitra Maoutsa tacitly assumes that the noise coefficient is a scalar multiple of the identity).

Conclusions

Dimitra Maoutsa's thesis is an impressive piece of work that devises efficient algorithms by cleverly combining tools from vastly different mathematical fields, such as statistics, stochastic analysis, machine learning, Riemannian geometry, and optimisation. Even though numerical test of challenging high-dimensional problems (say, beyond dimension 100) that could provide more detailed information about the computational efficiency of the schemes are lacking so far, the results in this thesis are promising and already pave the way to interesting applications.

The thesis contains original and novel results, and it is likely to make relevant contributions to the field. Dimitra Maoutsa has demonstrated that she is not only able to conduct independent research, but also that she can solve hard computational problems by easily combining fairly advanced mathematical and algorithmic tools. Overall, my critical remarks are nitpicking, and I recommend to accept this thesis with the mark: **sehr gut**



(Prof. Dr. Carsten Hartmann)
Chair of Stochastics and its Applications
BTU Cottbus-Senftenberg

Review of the PhD thesis manuscript “Deterministic particle flows for stochastic nonlinear systems” by Dimitra Despoina Maoutsa for the degree of Doctor at the TU Berlin.

Reviewer: Hilbert J. Kappen

The thesis addresses various important problems that occur in inference, learning and control. It is correctly observed that in all these problems the estimate of the gradient of the log density is of central importance. **I have read the thesis and find it very good (“sehr gut”).**

Chapter 2 contains a useful summary of fundamental methods that are relevant to the thesis, such as Ito scheme, Girsanov Thm, Radon-Nikodym derivative and the Wasserstein distances. Each of the chapters 3-6 contains a good overview of existing methods and motivates the contribution of that chapter. Each of these chapters contains an original contribution that is published in the scientific literature, except for chapter 6 which will be published on arxiv.

I find the application of the deterministic kernel-based approach to stochastic path integral control problems that is the subject of chapter 5 very original and an important novel contribution.

I briefly review each of the chapters, including a number of questions.

In chapter 3, the gradient of the log density is estimated from samples using a kernel. A sparse estimate is considered where the samples are replaced by (randomly selected) inducing points. The method is demonstrated on several artificial 1,2 and 3 dimensional problems. The method is compared with other methods on a challenging artificial two-dimensional problem.

Questions:

1. It is stated that the Stein method in Fig. 3.12 performs better than the other methods. However, this is not confirmed in fig. 3.13, where the Stein method has the largest error. Closer inspection of fig. 3.12 reveals that the Stein method performs perfectly wrong! It estimates zero gradient at locations where the gradient is large, and vice versa. This ‘inversion’ seems also the case, but less clear, for the proposed method in fig. 3.12. What is going on here?
2. The errors reported in fig. 3.13(left) are very large. What is the reason for comparing the normalized gradients fig. 3.13(right)?
3. Figs. 3.6-3.8 show the bias in the method for a 1-, 2- and 3-dimensional problem. In 1 and 3 dimension the bias is independent of the number of inducing points. In 2 dimensions, the bias increases with number of inducing points. One would naively expect that more inducing points introduces less bias. Why is this not observed?

Chapter 4 discusses simulation of the Fokker-Planck equation. It reviews common methods, such as the spectral decomposition, space discretization, and stochastic sampling methods. In Eq. 4.31 the FP equation is expressed as a gradient flow in probability space towards the stationary distribution. The objective that is minimized is the KL divergence of the instantaneous distribution towards the stationary distribution (Stein variational gradient descent). Using a kernel approximation, this becomes a deterministic particle dynamics Eq. 4.34. In 4.4 this idea is directly applied to the FP equation with no direct reference to the stationary distribution yielding a deterministic set of N coupled ODEs Eq. 4.47, which can be made more efficient by introducing inducing points.

Questions:

1. On pg. 76 it is mentioned that naive sampling methods suffer from the curse of dimensionality and suggest that this can be improved by considering deterministic particle methods. However, it seems to me that deterministic methods suffer essentially the same curse of dimensionality.
2. In section 4.7.2 a comparison is made between the Stein variational gradient descent and the new proposal. It is stated that the FP dynamics converges faster to equilibrium than the Stein dynamics. However, it seems to me that this depends on the type of FP dynamics and does not need to be true in general.
3. In fig. 4 and 5, it seems that the accuracy does not depend on the number of inducing points. Why is this the case? Why is the difference between S and D so large in figs. 4a,b and less so in figs. 5a,b?
4. The choice of the noise $D(x)=\sin^2(x)$ makes the bi-stable problem easier because the stationary distribution becomes unimodal. Why this easy choice? What would happen if $D(x)=\cos^2(x)$?
5. Is it a fair comparison to use same number of particles for S and D? D requires matrix inversion and is more costly. How does the gained efficiency of D versus S in terms of number of samples (fig. 4.13) compare to the additional cost of D vs. S?

Chapter 5 considers the stochastic optimal control problems, in particular the class of path integral control problems. The chapter formulates the optimal control in terms of the difference of two gradients of log probabilities Eq. 5.76. It is shown that both can be estimated using the deterministic particle scheme of the previous chapters.

The approach is tested numerically on several problems and compared to the previous PICE method. The results show that the new method is more efficient than PICE resulting in essentially the same solution (slightly larger control cost). This new method proposed in this chapter provides a truly original and important advance in the field of path integral control.

Questions:

1. For PI control problems it is well known that the sampling becomes increasingly harder for lower temperature (lower noise or lower control cost). How do the new deterministic methods perform in the low temperature regime?

Chapter 6 considers the problem of identifying a dynamical system from sparse observational data. In section 6.3, the densely observed case is considered. A previous method using a sparse Gaussian process (using inducing points) is reviewed and tested. section 6.5-6.7 considers the sparse case and proposes a new path augmentation framework that uses the kernel method of chapters 3 and 4 is applied to the Riemannian manifold of trajectories, also using the control methods of chapter 5. The method is numerically evaluated on several problems and shown to work very well.