Uncertainty Quantification for Molecular Models via Stochastic Gradient MCMC

Stephan Thaler [*§], **Julija Zavadlav** [* † ‡]

[*] *Professorship of Multiscale Modeling of Fluid Materials, TUM School of Engineering and Design, Technical University of Munich, Germany*
† *Munich Data Science Institute, Technical University of Munich, Germany*
§ *e-mail: stephan.thaler@tum.de*
‡ *e-mail: julija.zavadlav@tum.de*

Abstract: The quality of molecular dynamics (MD) simulations critically depends on the employed potential energy model. Accurate uncertainty quantification (UQ) of these models could increase trust in MD simulation predictions and promote progress in the field of active learning of neural network (NN) potentials. Bayesian methods promise reliable uncertainty estimates, but the high computational cost of training via classical Markov Chain Monte Carlo (MCMC) schemes has prevented their application to deep NN potentials. In this work, we propose stochastic gradient MCMC methods as a computationally efficient option for Bayesian UQ of MD potentials. The stochastic gradient Langevin dynamics method yields promising results for a tabulated coarse-grained water model and could thus be a feasible approach for NN potentials. Additionally, we illustrate the inherent limit of Bayesian UQ imposed by the functional form of the employed model.

Keywords: Bayesian inference, Uncertainty Quantification, Molecular dynamics

1. INTRODUCTION

Molecular dynamics (MD) simulations are the computational backbone of fields such as soft-matter physics and material science. The quality of MD simulations critically depends on the employed potential energy model defining particle interactions. Potentials are parametrized to match data from experiments (Thaler and Zavadlav (2021)) or high fidelity simulations. Given that experimental data and high fidelity simulations are expensive and only sparsely available, potentials are regularly applied outside the training domain. Hence, uncertainty quantification (UQ) of MD simulations is important to assess the trustworthiness of predictions (Zavadlav et al. (2019)).
An intriguing application of UQ in MD is active learning (Zhang et al. (2019)) of neural network (NN) potentials (Behler and Parrinello (2007)) from density functional theory (DFT) data. Active learning promises to minimize the number of expensive DFT simulations by quantifying the uncertainty of input states and iteratively augmenting the training data set only with states for which the NN potential is most uncertain. However, the efficiency of active learning hinges on the quality of UQ estimates. The common approach to UQ using NN ensembles (Hansen and Salamon (1990)) was found to be only marginally more informative than random selection of states (Kahle and Zipoli (2021)). While Bayesian NNs appear to yield more reliable uncertainty estimates, the high computational training cost of classical Markov Chain Monte Carlo (MCMC) schemes have prevented the application to real-world problems so far (Kahle and Zipoli (2021)).
In this work, we propose stochastic gradient MCMC methods (SG-MCMC) as a computationally efficient option for Bayesian UQ of MD potentials. Results for a tabulated coarse-grained (CG) model of water showcase reasonable uncertainty predictions.

2. METHODS

Bayesian UQ is centered around Bayes' theorem. The aim is to compute the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$ of model parameters $\boldsymbol{\theta}$ for a given data set $\mathcal{D}$ and model $\mathcal{M}$. MCMC is the gold-standard for approximating the posterior, which requires at least one evaluation of the likelihood $p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})$ and the prior $p(\boldsymbol{\theta}|\mathcal{M})$ for each update of $\boldsymbol{\theta}$. As computing the likelihood requires evaluation of the model for each data point in $\mathcal{D}$, training on large data sets with expensive models (e.g. NN potentials) quickly becomes infeasible. By contrast, SG-MCMC schemes evaluate the likelihood $p(\boldsymbol{\theta}|\tilde{\mathcal{D}},\mathcal{M})$ (and its gradient) only on a mini-batch $\tilde{\mathcal{D}} \subset \mathcal{D}$, allowing many updates of $\boldsymbol{\theta}$ per pass over $\mathcal{D}$ - analogous to stochastic gradient descent in maximum likelihood estimation. In the simplest case of the stochastic gradient Langevin dynamics method (Welling and Teh (2011)), learning rates λ_n at step n are required to converge to 0 such that generated samples of $\boldsymbol{\theta}$ are asymptotically unbiased, e.g. via a polynomial step size decay $\lambda_n = a(n+1)^{-\gamma}$, with decay rate γ and initial learning rate a . Hence, the increased computational efficiency comes at the cost of generating a biased estimate of $p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$ for a finite number of update steps.
To assess the quality of uncertainty estimates from SG-MCMC schemes, we learn a single-site CG water model parametrized by the control points of a cubic spline via

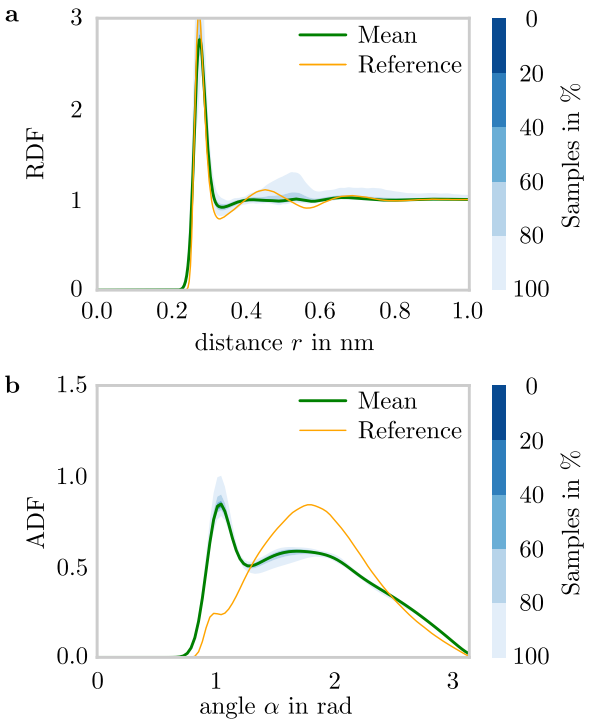

Fig. 1. Mean and credible interval of the predicted radial (RDF, **a**) and angular distribution function (ADF, **b**), together with the atomistic ground truth.

force-matching (Noid et al. (2008)). The ground-truth data consisting of 10^5 de-correlated states are obtained from a simulation of 905 water molecules in a cubic box of side length 3 nm with the atomistic SPC/FW (Wu et al. (2006)) water model at a temperature $T = 300$ K. We choose a uniform prior and assume a Gaussian likelihood with identity covariance matrix, where the variance σ^2 is treated as a learnable model parameter. The spline control points are initialized to the corresponding values of the potential of mean force (Reith et al. (2003)) and σ to 200 kJ / (mol nm). We approximate the posterior distribution via the stochastic gradient Langevin dynamics method (Welling and Teh (2011)) with the polynomial learning rate schedule ($a = 10^{-8}$, $\gamma = 0.33$). We train for 5 epochs with a mini-batch size of 5 and generate 1000 MCMC samples after the learning rate is reduced below $\alpha = 6 \cdot 10^{-10}$.

3. RESULTS

We evaluate the quality of the learned potential based on predicted observables by reference to the atomistic ground-truth. The mean predicted radial distribution function (RDF) deviates from the atomistic reference (Fig. 1 **a**), which is in line with tabulated 2-body potentials parametrized via maximum likelihood estimation (Scherer and Andrienko (2018)). The deviation results from the fact that the 2-body tabulated potential is a weak approximation to the distinct 3-body properties of water (Scherer and Andrienko (2018)). Importantly, the credible interval contains most of the ground truth RDF such that the extent of the deviation can be anticipated by practitioners. By contrast, the mean predicted angular distribution function (ADF) fails to reproduce the atomistic reference, but the narrow credible interval suggests high confidence in the incorrect prediction (Fig. 1 **b**).

Note that this UQ failure is not caused by a sub-optimal approximation of the true $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$ from the SG-MCMC scheme, but rather stems from the (implicit) conditioning of Bayes' theorem on the model $\mathcal{M}$: By definition, $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$ describes the posterior probability of all possible parameters $\boldsymbol{\theta}$ of $\mathcal{M}$. Effects that cannot be captured by any $\boldsymbol{\theta}$ cannot be represented in the uncertainty prediction. In this particular example, the ADF is predominantly determined by 3-body forces which cannot be represented by $\boldsymbol{\theta}$ in a 2-body potential. Hence, interpretation of results from Bayesian UQ critically depends on the employed $\mathcal{M}$.

4. CONCLUSION

Our results suggest that SG-MCMC methods could promote the application of UQ in MD simulations by reducing the computational burden of full-batch MCMC methods. However, further studies including investigation of the merits of more advanced SG-MCMC schemes, the number of necessary MCMC samples for reliable UQ results, as well as applying Bayesian UQ to highly expressive models such as NN potentials are required before obtained uncertainty estimates can be trusted in practice.

ACKNOWLEDGEMENTS

We thank Paul Fuchs for providing Fig. 1.

REFERENCES

Behler, J. and Parrinello, M. (2007). Generalized neural-network representation of high-dimensional potential-energy surfaces. *PRL*, 98(14), 146401.

Hansen, L.K. and Salamon, P. (1990). Neural network ensembles. *IEEE PAMI*, 12(10), 993–1001.

Kahle, L. and Zipoli, F. (2021). On the quality of uncertainty estimates from neural network potential ensembles. URL `arXiv:2108.05748`.

Noid, W., Chu, J.W., Ayton, G.S., Krishna, V., Izvekov, S., Voth, G.A., Das, A., and Andersen, H.C. (2008). The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.*, 128(24), 244114.

Reith, D., Pütz, M., and Müller-Plathe, F. (2003). Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.*, 24(13), 1624–1636.

Scherer, C. and Andrienko, D. (2018). Understanding three-body contributions to coarse-grained force fields. *Phys. Chem. Chem. Phys.*, 20(34), 22387–22394.

Thaler, S. and Zavadlav, J. (2021). Learning neural network potentials from experimental data via differentiable trajectory reweighting. *Nat. Commun.*, 12, 6884.

Welling, M. and Teh, Y.W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *ICML'11*.

Wu, Y., Tepper, H.L., and Voth, G.A. (2006). Flexible simple point-charge water model with improved liquid-state properties. *J. Chem. Phys.*, 124(2), 024503.

Zavadlav, J., Arampatzis, G., and Koumoutsakos, P. (2019). Bayesian selection for coarse-grained models of liquid water. *Sci. Rep.*, 9(1), 99.

Zhang, L., Lin, D.Y., Wang, H., Car, R., and Weinan, E. (2019). Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Mater.*, 3(2), 023804.