Sharpness-Aware Minimization for Robust Molecular Dynamics Simulations

Hikaru Ibayashi *
University of Southern California
ibayashi@usc.edu

Ken-ichi Nomura *
University of Southern California
knomura@usc.edu

Pankaj Rajak *
University of Southern California
rajak@usc.edu

Taufeq Mohammed *
University of Southern California
razakh@usc.edu

Ankit Mishra *
University of Southern California
ankitmis@usc.edu

Aravind Krishnamoorthy *
University of Southern California
kris658@usc.edu

Aiichiro Nakano *
University of Southern California
anakano@usc.edu

Abstract

Sharpness-aware minimization (SAM) is a novel regularization technique that takes advantage of not only the training error but also the landscape geometry of model parameters to improve model robustness. Although SAM has demonstrated the state-of-the-art (SOTA) performance in image classification, its applicability to physical system is yet to be examined. An ideal testbed is neural-network quantum molecular dynamics (NNQMD) simulations that accurately predict material properties, but the stability of their trajectories is severely limited by thermal noise. In this paper, we demonstrate for the first time that SAM regularizer achieves an order-of-magnitude reduction of the out-of-sample error in potential energy prediction using several SOTA models. Comparing NNQMD datasets with distinct structural characteristics, we found that SAM consistently reduces the out-of-sample error for a crystal dataset at high temperatures with enhanced thermal noise, thus proving the concept of SAM-enhanced robust NNQMD, while no clear trend was observed with an amorphous dataset. Our result suggests a possible correlation between materials structure and model parameter landscape.

1 Introduction

Molecular dynamics (MD) simulations are widely used to computationally study material properties by following the trajectories of constituent atoms. Accurate prediction of potential energy is essential for reliable MD simulations, but first-principles quantum mechanical (QM) calculations to obtain ground-truth potential energy are computationally prohibitive. Neural network quantum molecular dynamics (NNQMD) is revolutionizing MD simulations by predicting potential energy

*Collaboratory for Advanced Computing and Simulations, University of Southern California, Los Angeles, CA, U.S.

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

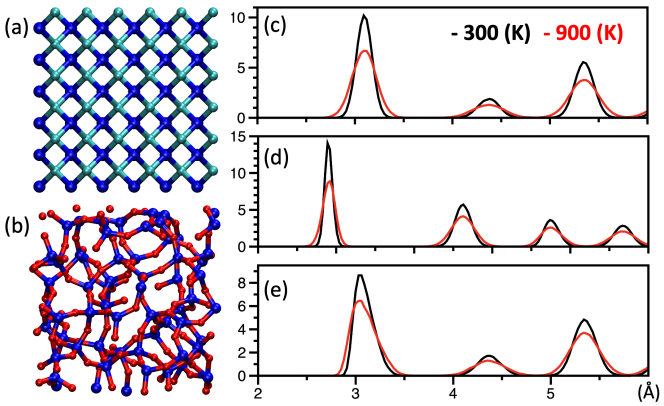

Figure 1: Snapshots of crystalline silicon carbide (a) and silicon dioxide glass (b), where spheres represent carbon (cyan), silicon (blue) and oxygen (red) atom positions and cylinders indicate chemical bonds. (c)-(e) Pair distribution functions for Si-Si, Si-C and C-C pairs at 300 K (black) and 900 K (red), respectively.

with QM accuracy using neural networks at a fraction of cost, thus allowing large spatiotemporal MD simulations [Pun et al., 2019, Mailoa et al., 2019].

Despite numerous successes, most NNQMD applications are limited to materials under gentle, near-equilibrium conditions [Jia et al., 2020], and NNQMD has rarely been applied to far-from-equilibrium processes to date. Specifically, at high temperature, the prediction error in atomic forces due to large thermal noise accumulates to produce unphysical behavior, which makes the simulation numerically unstable for larger systems at longer times, i.e., fidelity-scaling problem [Rajak et al., 2021]. While Rajak et al. [2020] attempted to alleviate the instability using active learning, their practical applicability is limited in terms of scalability and system dependence . It is an urgent task to find an alternative approach that is model- and system-agnostic without sacrificing the algorithmic scalability, to improve model generalizability for far-from-equilibrium NNQMD simulations.

We claim that *sharpness*, a recent notion in machine learning community, could provide a clue to solve the NNQMD instability problem. Sharpness of a neural network is defined as the sensitivity of the training loss against the weight parameters perturbation. Sharpness has gained attention particularly in recent years because an algorithm with sharpness regularization achieved the state-of-the-art (SOTA) performance in the image classification tasks [Foret et al., 2020]. More importantly for QMD applications, neural networks with regularized sharpness is shown to have strong robustness against the noise as well as its high generalization ability [Sun et al., 2020]. Provided that the simulation becomes unstable because of the combination of physical inaccuracy and numerical fragility, it is natural to apply a regularizer, SAM, that improves physical fidelity (i.e. generalizability of a model) and robustness.

In this paper, we examine the effect of sharpness regularization on the potential energy prediction performance using SOTA graph neural network (GNN). We start with a technical summary of MD simulation and sharpness regularization, followed by the main result that sharpness regularization significantly improves the prediction performance on multiple material datasets.[2] Our results also indicate a correlation between the sharpness of the weight parameters and the nature of the potential energy landscape.

2 Neural Network Quantum Molecular Dynamics

Molecular dynamics (MD) is an atomistic modeling method widely used in materials and chemical sciences. By numerically solving Newton's equations of motion, one obtains the complete history of atomic positions, $r^N = \{r_i \mid i = 1, \ldots, N\}$ (N is the number of atoms) as well as materials

[2]https://github.com/ibayashi-hikaru/Sharpness_MD

properties of interest. NNQMD is a newly emerged approximation scheme for MD simulation to statistically predict potential energy and thereby provide approximate trajectories of atoms [Pun et al., 2019, Mailoa et al., 2019].

Quantum Mechanical Simulation To generate training and test datasets, we use a reactive molecular dynamics (RMD) method based on ReaxFF interatomic potential [van Duin et al., 2001] and RXMD software [Nomura et al., 2020]. ReaxFF significantly reduces the computational cost compared to other QM-based approaches while reproducing chemical reactions and charge transfer at QM-level accuracy. We use two datasets with distinct structural characteristics, i.e., crystalline silicon carbide (SiC) and amorphous silicon dioxide glass (a-SiO_2), respectively shown in Figs. 1 (a) and (b). The two systems are thermalized at temperature 300 K for model training with and without SAM, and 900 K for the evaluation of out-of-sample error and the model generalizability against thermal noise. The glass structure of a-SiO_2 system was obtained by melt-quench method [Vashishta et al., 1990], resulting in a disordered network of chemical bonds.

To characterize the temperature effect on the atomic configuration, Figs. 1 (c) - (e) compare pair distribution functions of SiC thermalized at 300 K and 900 K. Due to the enhanced thermal motion by the elevated temperature of 900 K, the atomic positions deviate from training dataset, increasing the likelihood for a trained model to face unseen atomic configurations.

Representation of Atoms Toward the natural representation of atomic data, graph representation learned by GNN has been attracting great attention in materials and chemical science domains. In practice, GNNs have shown higher performance than the naïve approach based on static descriptors [Fung et al., 2021]. Several GNN models have been proposed for chemistry-related problems, mostly focusing on molecular systems [Fung et al., 2021, Chen et al., 2019, Schütt et al., 2017, Xie and Grossman, 2018]. GNN has also been used in material predictions involving periodic crystals, surfaces [Dunn et al., 2020, Louis et al., 2020, Park and Wolverton, 2020, Xie and Grossman, 2018], as well as MD simulations [Park et al., 2021].

Instability Issue and Related Works One of the key issues of NNQMD is its instability during long-time simulations, where accumulated error often causes a breakdown of simulation. An alleviation of the instability is the active learning approach, where a model is adaptively retrained on-the-fly by newly generated atomic configurations when the model prediction uncertainty exceeds a prescribed threshold [Vandermause et al., 2020]. However, this approach is not scalable because of its repeated training processes with a heavy cost. As an alternative approach, Rajak et al. [2021] proposed a physics-based regularization founded on statistical mechanics. While they have successfully performed a billion-atom NNQMD simulation of light-induced polarization dynamics in led titanate crystal, the proposed inductive bias is highly system-dependent, requiring a delicate tuning of the bias strength to prevent catastrophic failure while preserving QM-level accuracy. In contrast, the proposed SAM approach here is scalable and agnostic to a model or a system. More importantly, SAM minimizes sharpness, which, by definition, increases the model stability as we discuss in the next section.

3 Sharpness

Sharpness is defined to quantify the geometric property of the training loss surface with respect to weight parameters [Keskar et al., 2016]. Given a training data $S = \{(x_1, y_1), \ldots, (x_m, y_m))\}$, weight parameter w, and a loss function $\ell((x, y), w)$, the training loss and test loss of a neural network can be regarded as functions of w as follows.

$$\hat{L}(w) := \frac{1}{n} \sum_{i=0}^{n} \ell\left((x_i, y_i), w\right) \text{ and } L(w) := \mathrm{E}_{(x,y)}\left[\ell\left((x_i, y_i), w\right)\right] \tag{1}$$

Sharpness quantifies the sensitivity of $\hat{L}(w)$ against the perturbations on w, formally defined as follows.

$$\max_{\|\epsilon\|_2 \leq \rho} \hat{L}(w + \epsilon), \quad (\rho > 0 : \text{ hyperparameter }) \tag{2}$$

History of Sharpness Intriguingly, sharpness has its origin in physics. The whole sharpness research started with a finding in the research of complex systems that neural networks with high generalization ability tend to have flatter (less sharp) minima [Hochreiter and Schmidhuber, 1995]. After decades have passed and deep neural networks gained popularity, an increasing number of empirical results have suggested that a similar tendency also exists for deep neural networks [Keskar et al., 2016, Dziugaite and Roy, 2017, Sagun et al., 2017, Yao et al., 2018]. A large-scale experiment by Jiang et al. [2019] showed that sharpness defined in Eq. 2 has a particularly strong correlation with generalization among various measures. Inspired by those intriguing observations, sharpness has been implemented in practical optimization algorithms [Chaudhari et al., 2019], and one of those algorithms, SAM, has achieved the state-of-the-art performance in image classification tasks [Foret et al., 2020]. As a final remark, despite its remarkable empirical success, the machine learning community has not reached a theoretical understanding of sharpness. We refer the interested readers to some of the theoretical attempts to mathematically formalize the effectiveness of sharpness [Neyshabur et al., 2017, Kleinberg et al., 2018, He et al., 2019].

Sharpness Minimization Despite its theoretical strength, it is computationally nontrivial to minimize sharpness because of its high computational cost. Therefore, all existing approaches minimize sharpness indirectly by minimizing an approximate value [Chaudhari et al., 2019, Sun et al., 2020]. Similarly, SAM uses the following update rule to approximately minimize sharpness.

$$w = w - \eta \left(\nabla_w \hat{L}(w) + \nabla_{w'} \hat{L}\left(w'\right)\Big|_{w'=w+\frac{\rho}{\|\nabla \mathcal{L}(w)\|}\hat{L}(w)} \right) \quad (\eta > 0 : \text{ learning rate }) \tag{3}$$

This update consists only of first order derivative, which is computationally feasible A more theoretical explanation of how this update rule minimizes sharpness can be found in the Appendix A.

4 Experiments

To investigate the applicability of SAM to the MD problem, we have trained several GNN models (CGCNN [Xie and Grossman, 2018], SchNet [Schütt et al., 2017], MEGNet [Chen et al., 2019]) recently developed for materials applications using MatDeepLearn framework [Fung et al., 2021]. Model training is done with the 300 K dataset, and their generalizability is tested with the higher-temperature dataset at 900 K. We examine the effect of SAM with several values of the hyperparameter ρ in the loss function. As a baseline optimizer, we train each model with stochastic gradient decent (SGD) with decoupled weight decay (AdamW) [Loshchilov and Hutter, 2017] for 1000 epochs. Hyperparameters optimization is obtained using Ray Tune library. We have used NVIDIA GPU (V100) cluster nodes to perform model training, hyperparmaeter optimization, and test error evaluation.

Table 1: Out-of-sample errors baseline (without SAM) and with SAM (eV/atom)

	SiC			a-SiO_2		
	baseline	with SAM	weight (ρ)	baseline	with SAM	weight (ρ)
CGCNN	1.957	**0.008**	0.01	**0.065**	0.204	0.01
SchNet	0.025	**0.024**	0.02	0.071	**0.011**	0.02
MEGNet	19.17	**0.447**	0.05	**0.055**	0.169	0.01

We observe improved out-of-sample error for all models with the crystalline SiC dataset, ranging from 5 to 99.6% reduction compared to the baseline. Many of optimal test values were obtained with SAM weight ρ around 0.02, which is consistent with the value reported by Foret et al. [2020]. Unlike the SiC case, SAM does not consistently improve test performance for a-SiO_2 dataset. Glass landscape is known to be rough and consists of intrinsically sharp minima [Kushima et al., 2009]. It is remarkable to see that flat minima do not have high performance for some material data while they constantly show high performance on images [Jiang et al., 2019]. Provided that material science has the accumulation of knowledge of material properties, it is an interesting research question to reveal the relation between the property of energy landscape and loss landscape.

5 Conlusion

We have studied the effect of a novel SAM technique in materials dataset using SOTA GNN models. Two systems, SiC and a-SiO_2, have been examined, considering their distinct characteristics of chemical-bond network and associated potential energy landscape. The GNN models were optimized and trained using an ambient condition dataset. The effect of SAM in model generalizability was tested using a high-temperature dataset characterized by enhanced thermal noise. Our study suggests a promising avenue to realize materials models with enhanced generalizability and an insight into the future SAM algorithmic design taking advantage of the energy landscape of materials.

Broader Impact

SAM is a lightweight, model- and system-agnostic approach to improve the generalization performance in the physical sciences. Of those, SAM will be particularly beneficial for large-scale NNQMD simulation to study far-from-equilibrium material processes, where a carefully handcrafted model suffers from the fidelity-scaling problem. Our result encourages further applications of SAM to materials dataset, at the same time, suggests a possible route to future SAM algorithmic design.

Acknowledgments

This work was supported by the National Science Foundation, CyberTraining Program, Award OAC-2118061. Computations were performed at the Center for Advanced Research Computing of the University of Southern California.

References

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12): 124018, 2019.

Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.*, 31(9):3564–3572, May 2019.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Comput. Mater.*, 6(1):1–10, September 2020.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703. 11008*, March 2017.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-Aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010. 01412*, 2020.

Victor Fung, Jiaxin Zhang, Eric Juarez, and Bobby G Sumpter. Benchmarking graph neural networks for materials chemistry. *npj Comput. Mater.*, 7(1):1–8, June 2021.

Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. *arXiv preprint arXiv:1902. 00744*, February 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. In *Advances in Neural information Processing Systems*, pages 529–536, 1995.

Weile Jia, Han Wang, Mohan Chen, Denghui Lu, Lin Lin, Roberto Car, E Weinan, and Linfeng Zhang. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–14. ieeexplore.ieee.org, November 2020.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2019.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On Large-Batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609. 04836*, September 2016.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2698–2707, Stockholmsmässan, Stockholm Sweden, 2018. PMLR.

Akihiro Kushima, Xi Lin, Ju Li, Xiaofeng Qian, Jacob Eapen, John C Mauro, Phong Diep, and Sidney Yip. Computing the viscosity of supercooled liquids. II. silica and strong-fragile crossover behavior. *J. Chem. Phys.*, 131(16):164505, October 2009.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711. 05101*, November 2017.

Steph-Yves Louis, Yong Zhao, Alireza Nasiri, Xiran Wang, Yuqi Song, Fei Liu, and Jianjun Hu. Graph convolutional neural networks with global attention for improved materials property prediction. *Phys. Chem. Chem. Phys.*, 22(32):18141–18148, August 2020.

Jonathan P Mailoa, Mordechai Kornbluth, Simon Batzner, Georgy Samsonidze, Stephen T Lam, Jonathan Vandermause, Chris Ablitt, Nicola Molinari, and Boris Kozinsky. A fast neural network approach for direct covariant forces prediction in complex multi-element extended systems. *Nature Machine Intelligence*, 1(10):471–479, September 2019.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. *arXiv preprint arXiv:1706. 08947*, June 2017.

Ken-Ichi Nomura, Rajiv K Kalia, Aiichiro Nakano, Pankaj Rajak, and Priya Vashishta. RXMD: A scalable reactive molecular dynamics simulator for optimized time-to-solution. *SoftwareX*, 11: 100389, January 2020.

Cheol Woo Park and Chris Wolverton. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Phys. Rev. Materials*, 4(6):063801, June 2020.

Cheol Woo Park, Mordechai Kornbluth, Jonathan Vandermause, Chris Wolverton, Boris Kozinsky, and Jonathan P Mailoa. Accurate and scalable graph neural network force field and molecular dynamics with direct force architecture. *npj Comput. Mater.*, 7(1):1–9, May 2021.

GP Purja Pun, R Batra, R Ramprasad, and Y Mishin. Physically informed artificial neural networks for atomistic modeling of materials. *Nat. Commun.*, 10(1):1–10, 2019.

Pankaj Rajak, Kuang Liu, Aravind Krishnamoorthy, Rajiv K Kalia, Aiichiro Nakano, Ken-Ichi Nomura, Subodh C Tiwari, and Priya Vashishta. Neural network molecular dynamics at scale. In *2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 991–994. ieeexplore.ieee.org, May 2020.

Pankaj Rajak, Anikeya Aditya, Shogo Fukushima, Rajiv K Kalia, Thomas Linker, Kuang Liu, Ye Luo, Aiichiro Nakano, Ken-Ichi Nomura, Kohei Shimamura, Fuyuki Shimojo, and Priya Vashishta. Ex-NNQMD: Extreme-Scale neural network quantum molecular dynamics. In *2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 943–946. ieeexplore.ieee.org, June 2021.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of Over-Parametrized neural networks. *arXiv preprint arXiv:1706. 04454*, June 2017.

Kristof T Schütt, Pieter-Jan Kindermans, Huziel E Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *arXiv preprint arXiv:1706. 08566*, June 2017.

Xu Sun, Zhiyuan Zhang, Xuancheng Ren, Ruixuan Luo, and Liangyou Li. Exploring the vulnerability of deep neural networks: A study of parameter corruption. *arXiv preprint arXiv:2006. 05620*, June 2020.

Adri C T van Duin, Siddharth Dasgupta, Francois Lorant, and William A Goddard. ReaxFF: A reactive force field for hydrocarbons. *J. Phys. Chem. A*, 105(41):9396–9409, October 2001.

Jonathan Vandermause, Steven B Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *npj Comput. Mater.*, 6(1):1–11, March 2020.

P Vashishta, R K Kalia, J P Rino, and I Ebbsjö, I. Interaction potential for SiO2: A molecular-dynamics study of structural correlations. *Phys. Rev. B Condens. Matter*, 41(17):12197–12209, June 1990.

Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120(14):145301, April 2018.

Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. *arXiv preprint arXiv:1802. 08241*, February 2018.

A SAM minimizes sharpness

As an essential summary of [Foret et al., 2020], we briefly describe how Eq. 3 amounts to ends up minimizing sharpness. Consider the differentiation of sharpness Eq. 2, i.e,

$$\nabla \max_{\|\epsilon\|_2 \leq \rho} \hat{L}(w+\epsilon),$$

or equivalently the following form,

$$\nabla \hat{L}(w+\epsilon^*) \text{ with } \epsilon^* = \arg\max_{\|\epsilon\|_2 \leq \rho} \left(\hat{L}(w+\epsilon) \right)$$

By the first-order Taylor approximation, $\hat{L}(x, w+\epsilon)$ can be approximated to $\hat{L}(w) + \epsilon^T \nabla \hat{L}(w)$ and ϵ^* is reduced to the solution of the linear minimization problem.

$$\epsilon^* = \arg\max_{\|\epsilon\| \leq \rho} \hat{L}(w+\epsilon) \approx \arg\max_{\|\epsilon\| \leq \rho} \left(\hat{L}(w) + \epsilon^T \nabla \hat{L}(w) \right) = \arg\max_{\|\epsilon\| \leq \rho} \left(\epsilon^T \nabla \hat{L}(w) \right)$$

Thus, the approximated ϵ^* can be obtained as follows.

$$\epsilon^* \approx \rho \frac{\hat{L}(w)}{\|\nabla \hat{L}(w)\|}$$

With the approximated ϵ^* and omitting the second order derivative of as below,

$$\begin{aligned} \nabla \hat{L}(w+\epsilon^*) &= \nabla_w (w+\epsilon^*) \nabla_{w'} \hat{L}\left(w'\right) |_{w'=w+\epsilon^*} \\ &\approx \nabla_{w'} \hat{L}(w') |_{w'=w+\frac{\rho}{\|\nabla \hat{L}(w)\|} \hat{L}(w)} \end{aligned}$$

Eq. 3 is obtained, which essentially reduces sharpness (Eq. 2) on each step in approximation.