NANOSYSTEMS: PHYSICS, CHEMISTRY, MATHEMATICS, 2020, 11 (6), P. 642–650

Machine learning method for computation of optimal transitions in magnetic nanosystems

K. R. Bushuev[1], I. S. Lobanov[1,2]

[1]ITMO University, Kronverkskiy, 49, Saint Petersburg, 197101, Russia

[2]Saint Petersburg State University, Saint Petersburg, 198504, Russia

lobanov.igor@gmail.com

PACS 75.10.Hk, 75.75.-c, 82.20.Pm, 07.05.Mh **DOI 10.17586/2220-8054-2020-11-6-642-650**

Minimum energy path (MEP) is an important tool for computation of activation barriers and transition rates for magnetic systems. Recently, new methods for numeric computation of MEP were proposed based on conjugate gradient and L-BFGS methods [1] significantly improved convergence rate compared to nudged elastic band (NEB) method. Due to lack of strict mathematical theory for MEP optimization other more effective methods are expected to exist. In this article, we propose a machine learning based approach to search for MEP computation methods. We reformulate the NEB method as a differentiable transformation in the space of all paths parametrized by a family of metaparameters. Using rate of convergence as the loss function, we train NEB optimizer to find optimal metaparameters. This meta learning technique can be the basis for deriving new optimization methods for computing MEP and other non-classical optimization problems.

Keywords: Transition state, minimum energy path, machine learning, meta learning.

Received: 3 October 2020

Revised: 6 November 2020

1. Introduction

Magnetic systems are prominent candidates for the implementation of high-density storage and computing devices [2, 3]. The manufacturing of reliable magnetic devices is possible only using materials such that lifetime of metastable states carrying the information is long, but the states can be created and annihilated cheaply due to an external control such as spin current [4,5]. Since the desirable lifetime is many orders of magnitude larger than the period of the Larmor precession, direct simulation of stochastic dynamics is not suitable for the decay rate estimation [6–11]. State of the art transition rate estimation is based on transition rate theory [12–15]. The two most two challenging tasks of the lifetime estimation, both in harmonic transition state theory and in Langer's theory, are computation of determinants of Hessian matrix of energy (an analog of the partition function) and computation of the transition state itself [12, 16]. The transition state can be computed by an undirected search as an arbitrary first order saddle point on the energy surface using dimer method and similar approach [17–19]. The methods unfortunately do not return transition states of minimal energy, therefore the methods are of limited use for the activation barrier calculation. The activation barriers between two known metastable states are commonly computed by minimum energy path (MEP) based methods, such as nudged elastic band (NEB) [20–27]. Despite the high popularity of these methods, the mathematical theory of optimization methods for MEP evaluation is not well established, which is partially related to the multi-objective nature of the optimization, where the energy of each image on the path should be optimized as well as distribution of images along the path. Although higher order methods for MEP computation exist, such as conjugate gradient or L-BFGS [1], the steepest descent based method is still widely used. Due to lack of the mathematical theory, the optimization methods suffer from instability and poor choice of optimization parameters.

In recent years optimization theory, especially gradient based methods, gained acceptance due to the wide spread of machine learning methods. The right choice of the training method and meta parameters reduce training time and sometimes make problems tractable, which are not solvable by other means. Methods for tuning of optimization parameters for a specific class of problems belong to a subfield of machine learning called meta learning or learning to learn. In this article, we propose to use a meta learning approach to improve parameters of NEB method for MEP computation of magnetic systems.

There are several approaches to meta learning or in other words to optimization of the training procedure by machine learning techniques. In [28], genetic algorithms were used to speeding up the learning rate by changes of the learners policy based on "success-story algorithm". Meta-neural network approach in [29] demonstrated great potential to replace standard optimizer methods with neural network that was trained for solving optimization tasks. In [30], the automatic tuning of parametric learning rules is shown to lead to better generalization properties for the model. A new method for boosting up learner's average reward based on the inductive bias is proposed in [31] thus reducing the average number of interactions during training. A new type of neural network layers, called long short

term memory (LSTM), was introduced in [32]; the layer has a memory of events and improves adoption using historical data. Meta learning techniques were used to perform better generalization for training not only for single task but for groups of tasks with similar structure [33]. The meta learning approach can boost the training of neural network on large scale datasets [34]. In [35], it was shown that neural networks can be substituted for classical training algorithm in a wide range of applications. The widely used Adam training method [36] can be considered as kind of the meta learning, since it tunes its parameters according to the training history. The Adam method outperforms non-adaptive methods such as stochastic gradient descent and more primitive adaptive methods such as AdaGrad. A general strategy for training optimizers for black-box systems based on reinforcement learning was stated in [37, 38], the approach demonstrated good generalization ability. Studies in optimization of hyper parameters for neural networks with LSTM layers allow us to conclude that such architecture can boost up convergence rate in lot of application [39–42].

In the section 2 we recall basics of harmonic transition state theory for magnetic systems and give review of methods for MEP calculation focusing on NEB with improved tangent (IT) estimate. Then in the section 3 we modify IT-NEB to be suitable for machine learning, introduce criteria of comparison of discretized MEP, and implement a meta learning method to improve rate of convergence of the modified IT-NEB method.

2. Magnetic system transition state theory

In the Heisenberg model, states of magnetic system are defined by vectors of magnetic moments $\mathbf{M}_n$, where n is an atom of the lattice. Length μ_n of every vector is assumed to be fixed, since its relaxation time is much shorter than period of Larmor precession, therefore the state is convenient to express in terms of moments orientations $\mathbf{S}_n$, $\mathbf{M}_n = \mu_n \mathbf{S}_n$. When doing analysis of chiral states in magnetic systems, it is common to consider energy of the system given by the following quadratic form [8, 9]:

$$E[\mathbf{S}] = -\sum_{\langle n,k\rangle} J_{n,k}\mathbf{S}_n \cdot \mathbf{S}_k - \sum_{\langle n,k\rangle} \mathbf{D}_{n,k} \cdot (\mathbf{S}_n \times \mathbf{S}_k) - \sum_{n,m} K_m (\mathbf{K}_m \cdot \mathbf{S}_n)^2 - \sum_n \mathbf{H}_n \cdot \mathbf{S}_n,$$

where the summation is taken over pairs of interacting atoms $\langle n, k\rangle$, $J_{n,k}$ are Heisenberg exchange constants, $\mathbf{D}_{n,k}$ are Dzyaloshinskii-Moriya vectors, K_m and $\mathbf{K}_m$ are easy axis or easy plane anisotropy constant and vector respectively, $\mathbf{H}_n$ is the external magnetic field. Since the length of the direction vectors are restricted by constraints $\mathbf{S}_n^2 = 1$, special efforts should be undertaken to satisfy the constraints. The simplest way is to introduce spherical coordinates ϕ_n, θ_n: $\mathbf{S}_n = (\cos\phi_n \cos\theta_n, \sin\phi_n \cos\theta_n, \sin\theta_n)$. The spherical coordinates approach suffers from singularities near the poles $\theta_n = \pm\pi/2$, which leads to loss of precision in optimization and may ruin convergence; in this case, trigonometric functions are involved in the computation, thus reducing its speed. There are more advanced approaches that do not have these disadvantages, e.g. Cartesian coordinates based approach with Lagrange multipliers [16], an approach utilizing stereographic projection coordinates [43], rotation matrix based formulation [1]. In the article we will use spherical coordinates to facilitate idea of machine learning and avoid details of implementation of the advance methods.

In the thermal equilibrium with temperature T, the distribution of states is given by Boltzmann distribution with probability density: $\rho(\mathbf{S}) = Z^{-1}\exp(-E[\mathbf{S}]/(k_B T))$, where Z is the partition function. The magnetic system typically has many metastable states, such as ferromagnetic state, domain walls, skyrmions, skyrmionium, bag of skyrmions and other exotic particle-like states [43]. The states (especially domain wall and skyrmions) are proposed to be used as information carrier for magnetic data storage and processing devices [3–5]. To ensure operability of the devices, the metastable states should have long enough lifetimes, but they also should be easily annihilated and created in a controllable way to write the information. Stochastic dynamics of the magnetic systems can be simulated numerically integrating LandauLifshitzGilbert (LLG) equation [44]:

$$\frac{d\mathbf{S}_n}{dt} = -\gamma \mathbf{S}_n \times \mathcal{H}_n + \gamma\alpha \mathbf{S}_n \times \frac{d\mathbf{S}_n}{dt} + W(t), \quad \mathcal{H}_n = -\frac{E[\mathbf{S}]}{\partial \mathbf{S}_n} + \tau_n,$$

where γ is the gyromagnetic ratio, α is the damping constant, $W(t)$ is the thermal noise, and $\mathcal{H}$ is the effective magnetic field including torques τ due to spin currents. The direct simulation of the dynamics is not suitable for transition rate computation, since typically, the transitions are very rare events [8, 9]. Although there are attempts to solve the problem by path sampling [45], the widespread approach is transition state theory (TST). In TST, the probability to leave the state within a given time is estimated as the product of probabilities to be in the vicinity of the transition state and the probability to cross the dividing surface. Commonly, the transition rate is derived in the harmonic approximation giving rise to harmonic transition state theory [13] and Langers theory [6, 14]. In harmonic

transition state theory the transition rate is expressed as:

$$\kappa = \frac{\kappa^{dyn}\kappa^{ent}}{2\pi}\, e^{-\frac{\Delta E}{k_B T}}, \quad \kappa^{ent} = \sqrt{\frac{\det H^{MS}}{|\det H^{TS}|}},$$

where ΔE is the activation barrier, that is, the difference between the energy of the transition state (TS) and the metastable state (MS). The entropy prefactor is expressed in terms of determinants of the Hessian matrices H^{TS} and H^{MS} of energy at TS and MS, respectively. The dynamical prefactor is expressed in term of the eigenvector of Hessian corresponding to the negative eigenvalue (HTST, see details in [16]), or it equals to the positive eigenvalue of the matrix of the linearized LLG equation (Langer's theory). Computation of the determinant of the Hessians is the most challenging part in the formula, however recently a fast algorithm for computation of the determinants was proposed in [16] suitable for local interactions; long range dipole-dipole interaction probably can be taken into account by introducing a demagnetizing field [46]. Another important ingredient of the computation is search of TSs, which are first order saddle points on the energy surface. There are two main classes of methods for TS computation: undirected methods and minimum energy path (MEP) based methods. Undirected methods start from an initial state (commonly a metastable state) and follow minimum energy mode until the state is attracted to a TS, see e.g. dimer method [17] and review [18, 19]. Undirected methods are suitable, when resulting state of transition is not known.

For analysis of transitions between two given metastable states, MEP is the primary tool [20]. By definition MEP is a continuous path in the state space, connecting the metastable states, such that maximum of the energy on the path is the smallest among all possible paths. The maximum of the energy on the path is TS, clearly the TS is a first order saddle point of energy. According to the definition, the MEP is not unique, since essentially only the maximum is fixed. To restrict the set of MEP, it is common to assume that all point of the MEP obtain higher energy under variation of the path, hence all the points of the PATH have lowest possible energy. Since energy functional is generally quite complex, MEPs are computed by numerical methods. The path is discretized, introducing finite set of states (called images) along the path. The discretized path is called MEP, if for every image $S^k = (S^k_n) = (\theta^k_n, \phi^k_n)$ on the path projection of the effective field onto the orthogonal plane to the path equals zero [47]:

$$\mathcal{H}^k_\perp = \frac{\partial E[S^k]}{\partial S_n} - \left(\frac{\partial E[S^k]}{\partial S_n}\cdot\tau_n\right)\tau_n = 0 \;\; \forall n, \tag{1}$$

here τ_n is the unit tangent vector to the path. The two main problems of the definition are readily seen in the definition of the discrete MEP: (1) the energy is not necessarily monotone between neighbor images, hence energy barrier can be missed; (2) the tangent τ_n is not known explicitly and should be estimated numerically. To find the MEP numerically, an analog of the steepest descent method can be used, but special attention should be paid to avoid drift of images toward metastable states. The nudged elastic band method does iterations of the following form [47]:

$$S^k \mapsto S^k - \gamma\mathcal{H}^k_\perp + \mu e_k \tau_k, \quad e_k = \mathrm{dist}(S^{k+1}, S^k) - \mathrm{dist}(S^{k-1}, S^k), \tag{2}$$

where e_k are quasi-elastic forces, μ is the elasticity constant and γ is optimization step size. To ensure that TS will be among images on the path, so called climbing image modification of NEB is used, where for the image with largest energy, the component of the energy gradient parallel to the path is not eliminated but inverted [21, 24]:

$$\mathcal{H}^k_{\perp,\mathrm{CI}} = \frac{\partial E[S^k]}{\partial S_n} - 2\left(\frac{\partial E[S^k]}{\partial S_n}\cdot\tau_n\right)\tau_n = 0 \;\; \forall n,$$

forcing the image to move towards the TS. The optimization technique is successfully used for computation of activation barriers in the chemical reactions [22, 48, 49] as well as for magnetic systems [12, 50] and many others. A modification of the method for computation of only part of the MEP was proposed in [51], which is suitable for analysis MEP having TS much smaller than meta-stable states.

The naive tangent estimation τ_n using the central difference, unfortunately, makes NEB unstable, especially for a large number of images. The stability of method can be improved using direction to the neighbor image of the higher energy as an approximation to the tangent [24]. The maximum of energy along the path should be treated separately to avoid jumps when another image becomes maximum. According to [24], the tangent estimate can be chosen as follows:

$$\tau_n = \begin{cases} \tau^+_n, & \text{if } E_{n+1} > E_n > E_{n-1}, \\ \tau^-_n, & \text{if } E_{n+1} < E_n < E_{n-1}, \\ \tau^+_n \Delta E^{\max}_n + \tau^-_n \Delta E^{\min}_n, & \text{if } E_{n+1} > E_{n-1}, \\ \tau^+_n \Delta E^{\min}_n + \tau^-_n \Delta E^{\max}_n, & \text{if } E_{n+1} < E_{n-1}, \end{cases} \tag{3}$$

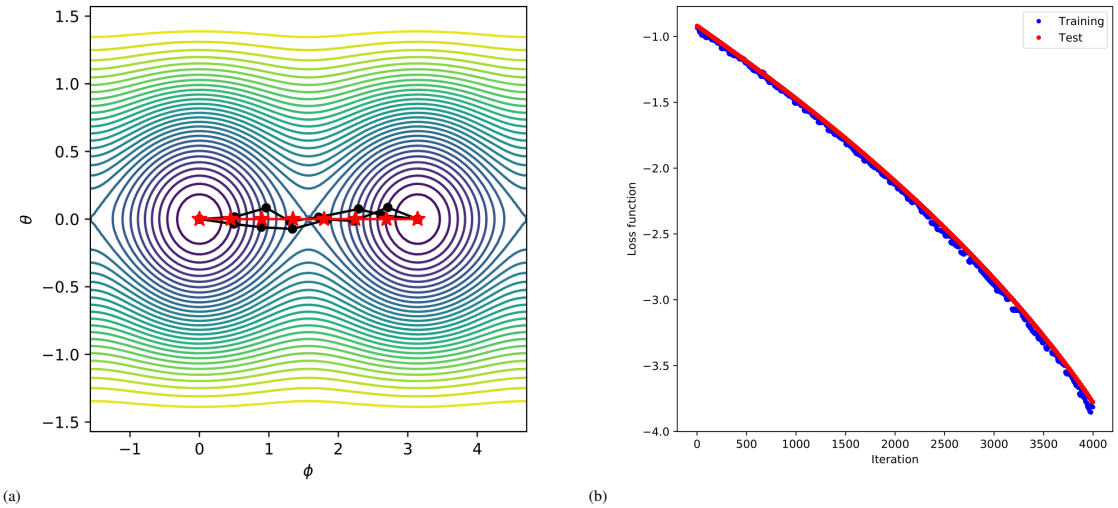

FIG. 1. (a) Energy surface for single spin system with easy axis $\mathbf{x}$ anisotropy $K_1 = 1$ and easy plane anisotropy $K_2 = -2$ with axis $\mathbf{z}$. MEP is shown by star markers. Initial approximation of the path is marked by dots. (b) Decay of loss L during training. Lower blue dots are loss on the training set. Upper red dots are loss on the test set. Training set consists of 100 paths, and is regenerated every 20 steepest descent iterations. The convergence rate was improved 20 times over the 200 epoch.

where the last two cases are applied if n is either maximum of minimum of energy on the path, $\tau_n^+ = S^{n+1} - S^n$, $\tau_n^- = S^n - S^{n-1}$ are right and left tangents at the image n, $E_n = E[S^n]$ is energy of the image n,

$$\Delta_n^{max} = \max(|E_{n+1} - E_n|, |E_{n-1} - E_n|), \quad \Delta_n^{min} = \min(|E_{n+1} - E_n|, |E_{n-1} - E_n|).$$

The choice of the elasticity parameter μ is somewhat arbitrary, and is said to not affect results significantly. However, in practice, the right choice of the elasticity parameter is crucial, since too large or too small value of μ ruins convergence, creating kinks on the path. Instead of elastic forces that push images toward their equidistant distribution, images can be redistributed every few optimization steps using linear interpolation of the path choosing uniform grid of local coordinates on the path, the method is known as the string method [25, 26]. The string method does not have arbitrary constants, but the redistribution step does not play well with higher order methods.

The NEB method can be used as a basis for gradient based method of second order. Interpreting equation (2) as steepest descend step $S \mapsto S - \gamma g$, where g is a quasi-gradient, one can apply conjugate gradient (CG) or L-BFGS method with g used instead of the gradient to obtain a higher order method. It seems that there is no potential function f such that $g = \partial f / \partial S$. Nevertheless both CG and L-BFGS methods demonstrate convergence, reducing optimization time in orders of magnitude [1].

Although methods for computation of the discrete MEP are widely used, their mathematical basis is far from being well established. While attempts are being made to apply variations of classical methods to MEP calculation, which is beyond the scope of the methods, the specialized methods for paths optimization are waiting to be discovered. The two features of the problem make discovery of the methods complicated: (1) the absence of a single functional for optimization (instead energy of each image is optimized separately) and (2) the need of simultaneous minimization of energy and equalization of images along the path. However, given a parametrized family of optimizers (implementing e.g. NEB method), the best optimizer can be found using machine learning techniques.

3. Training of optimization

To apply machine learning techniques for optimization of an algorithm, the algorithm must be expressed in terms of a parametrized family of transformations, which must be differentiable, as well as the loss function. The NEB method can be formulated as transformation M acting on paths, defined by the equation (2). We do not update the end points assuming that they are already relaxed to metastable states. Projected gradient defined by (1) and spring forces defined by (2) are differentiable function of its arguments, but improved tangent estimate given by (3) is not smooth. Here we introduce an analog of the improved tangent estimate from [24], but in a differentiable manner. We

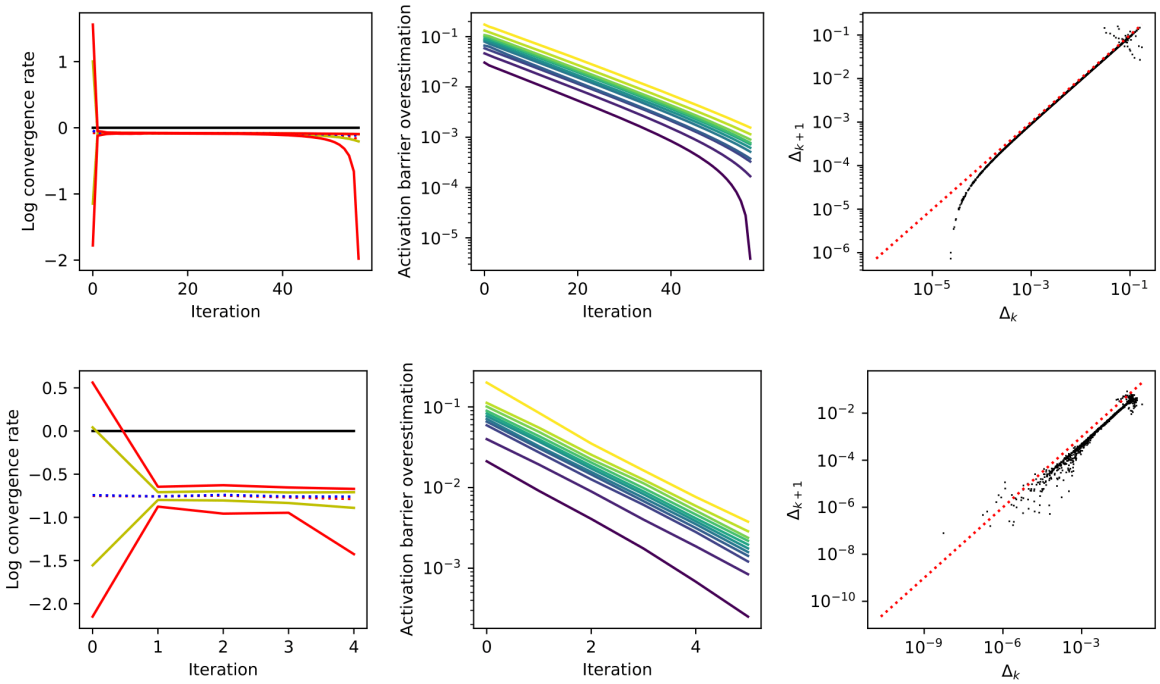

FIG. 2. Convergence of NEB for (upper row) initial set of parameters $\gamma = 0.01$, $\mu = 1$, $\beta = 3$; (bottom row) for optimized parameters $\gamma = 0.077$, $\mu = 1$, $\beta = 3$, estimated for 100 random paths. Left column: rate of convergence as function of the iteration number. Red lines represent minimum and maximum values, dotted line is the mean value, yellow lines are 10 and 90 percentile. Optimized parameters result in 20 times faster convergence. Center column: error of the activation barrier estimation as function of iteration number. Lines show all percentiles from 0 to 100 with step 10. Right column: activation barrier estimation error Δ_{k+1} on the next step as function of the error on the previous step Δ_k. Red dotted lines show sublinear convergence region. Optimized parameters give better convergence for large errors, however the optimization invalidates superlinear convergence in the limit of small errors, due to the definition of the loss depending only on finite number of iteration steps.

redefine the tangent estimate τ_n as a linear combination of left τ^- and right τ^+ tangents $\tau_n = a_n^+\tau_n^+ + a_n^-\tau_n^-$, where the coefficients a^+ and a^- are functions of the energy grows rate on the adjacent segments:

$$a_n^+ = W\left(\frac{\partial E[S^n]}{\partial S}\cdot\tau_n^-; \frac{\partial E[S^n]}{\partial S}\cdot\tau_n^+\right), \quad a_n^- = W\left(-\frac{\partial E[S^n]}{\partial S}\cdot\tau_n^+; -\frac{\partial E[S^n]}{\partial S}\cdot\tau_n^-\right).$$

We defined the weight W as a function of the gradient, which allows us to completely avoid energy computation during optimization, in contrast to the approach of [24]. The weight function W should select the right tangent if the energy increases at the image, be a smooth function and have symmetry with respect to path inversion; the assumptions are formalized as following natural conditions:

(1) (smoothness) $W(a;b)$ is an analytic function of a and b.
(2) (being a weight) $W(a;b) \geq 0$ and $W(a;b) + W(-b;-a) = 1$ for all a and b ().
(3) (stabilization of NEB) For positive a and b the right tangent is selected, that is $W(a;b) \approx 1$.

All the conditions are satisfied by the family of functions

$$W(a;b) = \frac{e^{\beta a} + e^{\beta b}}{e^{\beta a} + e^{-\beta a} + e^{\beta b} + e^{-\beta b}},$$

parametrized by $\beta > 0$. For $a, b \to 0$, the tangent estimate is close to the central finite difference, that is, $W(a;b) \approx 1/2$ regardless of signs of the arguments, which is different from the approach of [24]. However, the estimate coincides with the right tangent, $W \to 1$, as $a, b \to +\infty$. The parameter β controls sensitivity of W to the value of its arguments.

The NEB transformation M defined above depends on three parameters: step size γ, elasticity μ and smoothness of the tangent β. Whereas γ can be in theory estimated using e.g. Barzilai-Borwein method, the choice of μ and β

is unclear. We suggest to consider the mapping M as an artificial neural network with weights (γ, μ, β) and train it using machine leaning techniques to obtain the best convergence. According to definition of MEP as a continuous path, the maximum of energy along the path should be minimal among all the paths. Hence the optimality of the path is expressed by the value $Q[S] = \max_t E[S(t)]$ (smaller is better), where t runs all over the path and $S(t)$ is the continuous path obtain by linear interpolation of the discrete path S^k. We expect any optimization method for MEP computation to decrease Q-value doing its iterations. Suppose initial approximation to the MEP is given by path $S^{(1)}$, the path gradually optimized by the sequence of transformations $S^{(k+1)} = MS^{(k)}$. Denote by $\Delta^{(k)}$ the value of overestimation of the energy barrier: $\Delta^{(k)} = Q[S^{(k)}] - Q^0$, where $Q^0 = \min_S Q[S]$ is the true height of the activation barrier. The rate of convergence is given by the ratio $\rho^{(k)} = \Delta^{(k+1)}/\Delta^{(k)}$. It is desirable to have superlinearly convergent method, that is $\lim_{k\to\infty} \rho^{(k)} = 0$. In practice we do only finite number of optimization steps, therefore we want all $\rho^{(k)}$ to be as close to zero as possible. We define the loss function as mean of logarithms of the rates:

$$L[S^{(0)}] = \frac{1}{K} \sum_{k=1}^{K} \ln \rho^{(k)},$$

where only first K iterations of the method are considered. We said the parameters (γ, μ, β) are optimal, if they minimize loss function:

$$L = \operatorname*{argmin}_{\gamma,\mu\beta} \mathbb{E}(L[S^{(0)}]),$$

where the expectation value is taken for some distribution of paths near the MEP. The expectation value is estimated by arithmetic mean of $L[S]$ values for a set of random paths ${}^{(p)}S$:

$$L \approx \frac{1}{P} \sum_{p} L[{}^{(p)}S].$$

For computation of the maximum Q along the piecewise linear path defined by images S^k, we apply a combination of golden section search and successive parabolic interpolation on each line segment $S^k S^{k+1}$. We perform optimization of the parameters by steepest descent:

$$\gamma \mapsto \gamma - \nu \frac{\partial L}{\partial \gamma}, \quad \mu \mapsto \gamma - \nu \frac{\partial L}{\partial \mu}, \quad \beta \mapsto \gamma - \nu \frac{\partial L}{\partial \beta},$$

with the step size ν estimated by Barzilai-Borwein method.

The explicit form of derivative of L with respect to the parameters are quite complicated, therefore we use automatic differentiation to obtain the derivatives. The transform M was implemented in Python, and JAX library [52] was used to compute gradients and JIT compile functions. Due to extreme complexity of the gradient of the loss (we differentiate the algorithm for computation of maximum Q along the path and iterations of NEB method) and hardware limitations, we consider a simple case of single magnetic spin with easy axis anisotropy $K_1 > 0$ having the axis $\mathbf{K}_1 = \mathbf{x}$, and and easy plane anisotropy $K_2 < 0$ having the axis $\mathbf{K}_2 = \mathbf{z}$. Since NEB does not take into account exact form of the energy and number of degrees of freedom, the result should be qualitatively the same for general magnetic systems. Energy of the considered system has the following forms in the Cartesian and in the spherical coordinates:

$$E[\mathbf{S}] = -K_1 S_x^2 - K_2 S_z^2 = -K_1 \cos^2\theta \cos^2\phi - K_2 \sin^2\theta.$$

The system has two metastable states $(S_x, S_y, S_z) = (\pm 1, 0, 0)$ (or in polar coordinates $(\theta, \phi) = (0, 0)$ and $(0, \pi)$) having energy $-K^1$. The transition states are at the points $(S_x, S_y, S_z) = (0, \pm 1, 0)$; the energies of the both TS are the same and equal to $-K^2$. The MEP in the case lies in the Oxy plane avoiding problematic poles $\theta = \pm\pi/2$. Since we want NEB to reduce error in MEP estimation, and NEB has no proven convergence for arbitrary initial paths, we train and test our optimization method on the ideal MEP with position of points perturbed no more than 5% of the path length as shown in Fig. 1(a). Path consisting from 8 to 14 images were considered, demonstrating a qualitatively identical convergence pattern. We ran the optimization for the 200 epoch 20 iterations each in batches of 100 paths regenerated for each epoch. The loss function is computed using mean convergence rate estimation doing 5 iterations of the modified IT-NEB method. According to our tests, the initial NEB parameters can be chosen arbitrary, except for too large values of γ, where the NEB method diverges, and too small values of μ, when NEB is unstable. The loss function decrease history is shown in Fig. 1(b) for the initial parameters $\gamma = 0.01$, $\mu = 1$, $\beta = 3$. In the 200, epoch the convergence rate was improved 20 fold. Inspection of the gradient of loss function shows that the value of β is least significant for convergence rate. The value of μ affects the convergence only slightly, but it is important for stability of the method, which was not a subject of study of the current article.

Due to hardware restrictions, the considered loss function takes into account only rate of convergence of several first iterations of NEB method, therefore long time convergence of the method can be different. One of the reasons

to expect different behavior is change in the shape of the path, which is not covered by our ansatz on distribution of paths in the training set. Another possible issue in the real usage of NEB method is that initial approximation of the path can be very different from the actual MEP. We made benchmarks on the bended path doing 60 iteration of NEB with initial set of parameters and optimized parameters. the history of convergence of 100 random paths are shown in Fig. 2. The convergence of the NEB after training was also improved for bended paths. The optimized parameters give better mean convergence, at the same time, for some paths, the convergence has become worse, but is still better than worst case with the initial parameters. For large perturbations of path, the convergence of NEB is not monotonic, however for smaller error the convergence becomes linear. For moderate values of step size the convergence may become superlinear, unfortunately, the proposed training method does not improve Q-convergence.

4. Conclusion

In this article we demonstrated that existing software technologies such as automatic differentiation of arbitrary code, e.g. using JAX [52], can be used to implement meta learning to improve convergence rate of existing optimization methods such as NEB method by tuning meta parameters. The approach is most useful for the problems, such as MEP evaluation, which do not have an elaborated mathematical theory. Previously, the effectiveness of meta learning was shown for tuning of parameters of simple methods, such as gradient descent. In this article, we reveal that meta learning can be used to train much more complex methods, such as NEB, however, the complexity of the gradients of the loss function in the case imposes restrictions on the complexity of the method, i.e. number of iterations made by parts of the algorithm can not be too high.

The real power of this method can be revealed by training of parameter free higher order methods, which do not need tuning for every energy functional. The conjugate gradient (CG) and L-BFGS methods are examples of the methods, but their convergence for MEP computation is not proven. Meta learning techniques with neural networks capable of simulation of CG and L-BFGS iterations and more general transforms probably can be trained to obtain a higher order method specialized for MEP calculations.

One important issue with meta learning for optimization methods is choice of loss function, which can estimate the limit of convergence rate, when only finite (and rather small) number of iteration steps can be made. The related question is estimation of the stability of the method. In our research, we encountered large variation of the convergence rate on random samples, which make the choice of the loss function even harder.

Although we mentioned computation of lifetime of the magnetic metastable states as our motivational example, computation of MEP is an important tool for other problems with large numbers of degrees of freedom. MEP was introduced as a method the activation barriers estimation in chemical reactions, and the NEB method is still a valuable approach in the field [47, 48, 53, 54]. MEP for an artificial potential fields can be considered as an optimal path in motion planing problem for a mobile robot [55]. In all the subject the same meta learning technique can be applied, if higher order derivatives for the potential is accessible.

Acknowledgements

The study of classical optimization methods for transition state computation in the sections 1 and 2 was funded by Government of the Russian Federation (Grant 08-08). The development of meta learning algorithm for nudged elastic band method in the section 3 was funded by Russian Science Foundation (Grant 19-42-06302).

References

[1] Ivanov A.V., Dagbartsson D., Tranchida J., Uzdin V.M., Jnsson H. Efficient optimization method for finding minimum energy paths of magnetic transitions. *Journal of Physics: Condensed Matter*, 2020, **32**(34).

[2] Everschor-Sitte K., Masell J.,Reeve R.M., Klaui M. Perspective: Magnetic skyrmions - Overview of recent progress in an active research field. *J. Appl. Phys.*, 2018, **124**, P. 240901.

[3] Mittal S. A survey of techniques for architecting processor components using domain-wall memory. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 2016, **13**(2), P. 1–25.

[4] Parkin S.S.P., Hayashi M., Thomas L. Magnetic domain-wall racetrack memory. *Science*, 2008, **320**(5873), P. 190–194.

[5] Fert A., Cros V., Sampaio J. Skyrmions on the track. *Nature Nanotech*, 2013, **8**, P. 152–156.

[6] Schratzberger J., Lee J., Fuger M., Fidler J., Fiedler G., Schref T., Suess D. *Validation of the transition state theory with Langevin-dynamics simulations. Journal of Applied Physics*, 2010, **108**, P. 033915.

[7] Desplat L., Suess D., Kim J-V., Stamps R.L. Thermal stability of metastable magnetic skyrmions: Entropic narrowing and significance of internal eigenmodes. *Phys. Rev. B*, 2018, **98**, P. 134407.

[8] Bessarab P.F., Mller G.P., Lobanov I.S. et. al. Lifetime of racetrack skyrmions. *Sci. Rep.*, 2018, **8**, P. 3433.

[9] Potkina M.N., Lobanov I.S., Jnsson H., Uzdin V.M. Skyrmions in antiferromagnets: Thermal stability and the effect of external field and impurities. *Journal of Applied Physics*, 2020, **127**, P. 213906.

[10] Uzdin V.M., Potkina M.N., Lobanov I.S., Bessarab P.F., Jnsson H. The effect of confinement and defects on the thermal stability of skyrmions. *Physica B: Condensed Matter*, 2018, **549**, P. 6–9.

[11] Lobanov I.S., Jnsson H., Uzdin V.M. Mechanism and activation energy of magnetic skyrmion annihilation obtained from minimum energy path calculations. *Phys. Rev. B.*, 2016, **94**, P. 174418-2016.
[12] Bessarab P.F., Uzdin V.M., Jnsson H. Method for finding mechanism and activation energy of magnetic transitions, applied to skyrmion and antivortex annihilation. *Comp. Phys. Commun.*, 2015, **196**, P. 335–347.
[13] Bessarab P.F., Uzdin V.M., and Jnsson H. Harmonic transition-state theory of thermal spin transitions. *Phys. Rev. B*, 2012, **85**, P. 184409.
[14] Langer J.S. Statistical theory of the decay of metastable states. *Annals of Physics*, 1969, **54**(2), P. 258–275.
[15] Liashko S.Y., Lobanov I.S., Uzdin V.M., Jnsson H. Thermal stability of magnetic states in submicron magnetic islands. *Nanosystems: Physics, Chemistry, Mathematics*, 2017, **8**(5), P. 572–578.
[16] Lobanov I.S., Uzdin V.M. *The lifetime of big size topological chiral magnetic states. Estimation of the pre-exponential factor in the Arrhenius law.* 2020. arXiv:2008.06754
[17] Henkelman G., Jnsson H. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *J. Chem. Phys.*, 1999, **111**(15), P. 7010–7022.
[18] Olsen R.A., Kroes G.J., Henkelman G., Arnaldsson A., Jnsson H. Comparison of methods for finding saddle points without knowledge of the final states. *J. Chem. Phys.*, 2004, **121**(20), P. 9776–9792.
[19] Gutierrez M.P., Argaez C., Jnsson H. Improved minimum mode following method for finding first order saddle points. *J. Chem. Theo. Comput.*, 2017, **13**(1), P. 125–134.
[20] Henkelman G., Johannesson G., Jnsson H. Methods for finding saddle points and minimum energy paths. *Theoretical Methods in Condensed Phase Chemistry*, 2002, **5**, P. 269–302.
[21] Henkelman G., Uberuaga B.P., Jnsson H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.*, 2000, **113**(22), P. 9901–9904.
[22] Maragakis P., Andreev S.A., Brumer Y., Reichman D.R., Kaxiras E. Adaptive nudged elastic band approach for transition state calculation. *J. Chem. Phys.*, 2002, **117**, P. 4651–4658.
[23] Hoffmann M.,Mller G.P., Blgel S. Atomistic Perspective of Long Lifetimes of Small Skyrmions at Room Temperature. *Phys. Rev. Lett.*, 2020, **124**, P. 247201.
[24] Henkelman G., Jnsson H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.*, 2000, **113**(22), P. 9978–9985.
[25] Weinan E., Weiqing R., Vanden-Eijnden E. String method for the study of rare events. *Phys. Rev. B.*, 2002, **66**, P. 052301(4).
[26] Sheppard D., Terrell R., Henkelman G. Optimization methods for finding minimum energy paths. *J. Chem. Phys.*, 2008, **128**, P. 134106(10).
[27] Heistracher P., Abert C., Bruckner F., Vogler C., Suess D., GPU-Accelerated Atomistic Energy Barrier Calculations of Skyrmion Annihilations. *IEEE Transactions on Magnetics*, 2018, **54**(11), P. 1-5.
[28] Schmidhuber J. *Evolutionary Principles in Self-Referential Learning*. On Learning how to Learn: The Meta-Meta-Meta...-Hook. PhD thesis, Institut f. Informatik, Tech. Univ. Munich, 1987.
[29] Naik D.K., Mammone R.J. Meta-neural networks that learn by learning. In International Joint Conference on Neural Networks, *IEEE*, 1992, **1**, P. 437–442.
[30] Bengio S., Bengio Y., Cloutier J. On the search for new learning rules for ANNs. *Neural Processing Letters*, 1995, **2**(4), P. 26–30.
[31] Schmidhuber J., Zhao J., Wiering M. Shifting Inductive Bias with Success-Story Algorithm, Adaptive Levin Search, and Incremental Self-Improvement. *Machine Learning*, 1997, **28**, P. 105–130.
[32] Hochreiter S., Schmidhuber J. Long Short-Term Memory. *Neural Computation*, 1997.
[33] Thrun S., Pratt L. *Learning to learn.* Springer Science and Business Media, 1998.
[34] Younger A.S., Conwell P.R., Cotter N.E. Fixed-weight on-line learning. *Transactions on Neural Networks*, 1999, **10**(2), P. 272–283.
[35] Runarsson T.P., Jonsson M.T. Evolution and design of distributed learning rules. In IEEE Symposium on Combinations of Evolutionary Computation and Neural Networks, 2000, P. 59–63.
[36] Diederik P. Kingma, Jimmy Ba. Adam: A Method for Stochastic Optimization. 3rd International Conference for Learning Representations, San Diego, 2015.
[37] Andrychowicz M., Denil M., Gomez S., Hoffman M.W., Pfau D., Schaul T., Shillingford B., de Freitas N. *Learning to learn by gradient descent by gradient descent.* 2016, arXiv:1606.04474.
[38] Chen Y., Hoffman M.W., Gomez S.C., Denil M., Lillicrap T.P., Botvinick M., de Freitas N. *Learning to Learn without Gradient Descent by Gradient Descent.* 2016, arXiv:1611.03824.
[39] Wichrowska O., Maheswaranathan N., Hoffman M.W., Gomez S.C., Denil M., de Freitas N., Sohl-Dickstein J. *Learned Optimizers that Scale and Generalize.* 2017, arXiv:1703.04813.
[40] Hospedales T., Antoniou A., Micaelli P., Storkey A. *Meta-Learning in Neural Networks: A Survey.* 2020, arXiv:2004.05439.
[41] Carlon A., Espath L., Lopez R., Tempone R. *Multi-iteration stochastic optimizers.* 2020, arXiv:2011.01718
[42] Khromova K. *Optimization of the neural network training method.* Master dissertation. ITMO University, 2020.
[43] Rybakov F.N., Borisov A.B., Blgel S., Kiselev N.S. New Type of Stable Particlelike States in Chiral Magnets. *Phys. Rev. Lett.*, 2015, **115**, P. 117201.
[44] Mentink J.H., Tretyakov M.V., Fasolino A., Katsnelson M.I., Rasing Th. Stable and fast semi-implicit integration of the stochastic LandauLifshitz equation. *Journal of Physics: Condensed Matter*, 2010, **22**(17), P. 176001.
[45] Desplat L., Vogler C., Kim J.-V., Stamps R. L., Suess D. Path sampling for lifetimes of metastable magnetic skyrmions and direct comparison with Kramers' method. *Phys. Rev. B.*, 2020, **101**, P. 060403(R).
[46] Moskalenko M.A., Lobanov I.S., Uzdin V.M. Demagnetizing fields in chiral magnetic structures. *Nanosystems: Physics, Chemistry, Mathematics*, 2020, **11**(4), P. 401–407.
[47] Jonsson H., Mills G., Jacobsen K.W., Berne B.J., Ciccotti G., Coker D.F. *Nudged elastic band method for finding minimum energy paths of transitions, in Classical and Quantum Dynamics in Condensed Phase Simulations*. World Scientific, Singapore, 1998, P. 385–404.
[48] Henkelman G., Arnaldsson A., Jonsson H. Theoretical calculations of CH_4 and H_2 associative desorption from Ni(111): Could subsurface hydrogen play an important role? *J. Chem. Phys.*, 2006, **124**, P. 044706(9).
[49] Einarsdottir D.M., Arnaldsson A., Oskarsson F., Jonsson H. Path optimization with application to tunneling. *Lecture Notes in Computer Science*, 2012, **7134**, P. 45–55.

[50] Malottki S.V., Dupe B., Bessarab P.F., Delin A., Heinze S. Enhanced skyrmion stability due to exchange frustration. *Sci. Rep.*, 2017, **7**(10), P. 12299.

[51] Lobanov I.S., Potkina M.N., Jnsson H., Uzdin V.M. Truncated minimum energy path method for finding first order saddle points. *Nanosystems: Physics, Chemistry, Mathematics*, 2017, **8**(5), P. 586–595

[52] Bradbury J., Frostig R., Hawkins P., Johnson M.J., Leary C., Maclaurin D., Necula G., Paszke A., Vanderlas J., Wanderman-Milne S., Zhang Q. JAX: composable transformations of Python+NumPy programs. (`http://github.com/google/jax`), 2018.

[53] Mills G., Jonsson H., Schenter G.K. Reversible work based transition state theory: Application to H_2 dissociative adsorption. *Surf. Sci.*, 1995, **324**, P. 305–337.

[54] Zhu T., Li J., Samanta A., Kim H.G., Suresh S. Interfacial plasticity governs strain rate sensitivity and ductility in nanostructured metals. *PNAS*, 2007, **104**(9), P. 3031–3036.

[55] Fakoor M., Kosari A., Jafarzadeh M. Revision on fuzzy artificial potential field for humanoid robot path planning in unknown environment. *International Journal of Advanced Mechatronic Systems (IJAMECHS)*, 2015, **6**(4), P. 174–183