Abstract

Artificial neural networks (ANNs) are typically highly nonlinear systems which are finely tuned via the optimization of their associated, non-convex loss functions. Typically, the gradient of any such loss function fails to be dissipative making the use of widely-accepted (stochastic) gradient descent methods problematic. Adding a quadratic regularization term cannot always remedy this, in which case one needs to replace it with a higher order penalty term. However, the addition of such a term leads to the violation of the global Lipschitz continuity for the regularized gradient, which in turn makes gradient descent methods ineffective. This issue has been highlighted in the case of Euler discretizations (of which SGLD is an example) in Hutzenthaler et al. (2011), where it is proven that the difference of the exact solution of the corresponding stochastic differential equation (SDE) and of the numerical approximation at even a finite time point diverges to infinity in the strong mean square sense.

A natural way to address the above issue is to combine higher order regularization with taming techniques to improve the stability of any resulting algorithm. We offer a new learning algorithm based on an appropriately constructed variant of the popular stochastic gradient Langevin dynamics (SGLD), which is called tamed unadjusted stochastic Langevin algorithm (TUSLA). The roots of the TUSLA algorithm are based on the taming technology for diffusion processes with superlinear coefficients as developed in Sabanis (2013), Sabanis (2016) and used for MCMC algorithms in Sabanis and Zhang (2019),Brosse et al. (2019).

By proposing this new algorithm, the current article is the first, up to the authors' knowledge, to provide theoretical guarantees (for the discovery of approximate minimizers of empirical and population risks) for the use of SGLD algorithms for the fine tuning of neural networks. It also extends the current SGLD literature by dealing with non-convex optimization problems for a wide class of objective functions (locally Lipschitz gradients). We also provide a nonasymptotic analysis of the new algorithm's convergence properties in the context of non-convex learning problems with the use of ANNs. Thus, we provide finite-time guarantees for TUSLA to find approximate minimizers of both empirical and population risks. Numerical experiments are presented which confirm the theoretical findings and illustrate the need for the use of the new algorithm in comparison to vanilla SGLD within the framework of ANNs.

References

N. Brosse, A. Durmus, É. Moulines, and S. Sabanis. The tamed unadjusted Langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663, 2019.

M. Hutzenthaler, A. Jentzen, and P. E. Kloeden. Strong and weak divergence in finite time of euler's method for stochastic differential equations with non-globally lipschitz continuous coefficients. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 467(2130):1563–1576, 2011. ISSN 1364-5021.

S. Sabanis. A note on tamed euler approximations. *Electron. Commun. Probab.*, 18(47): 1–10, 2013.

S. Sabanis. Euler approximations with varying coefficients: the case of superlinearly growing diffusion coefficients. *Ann. Appl. Probab.*, 26(4):2083–2105, 2016.

S. Sabanis and Y. Zhang. Higher order Langevin Monte Carlo algorithm. *Electronic Journal of Statistics*, 13(2):3805–3850, 2019.