To interact or not? The convergence properties of interacting stochastic mirror descent

Anonymous authors

I. Introduction

In machine learning applications the optimization objective can be written in the general form,

$$x^* = \arg\min_{x\in\mathcal{X}}\{f(x)\}, \tag{1}$$

with $\mathcal{X} \subset \mathbb{R}^d$ be a closed convex constraint set and f a convex or, as in deep learning, a non-convex function. A good model can be characterized by i) convergence (close) to the optimum, ii) good generalization performance. For the first it is standard to use a decaying learning rate [7] and the latter can be quantified through e.g. the flatness of the minimum [4].

II. The algorithm

The conventional way of optimizing (1) is to run an instance of (stochastic) gradient descent. Another option is to run independent replicas of the algorithm and average the results. We will refer to each of these runs as a particle. The question we address in this work is whether it is beneficial to allow these particles *to interact* with each other. We will use the general framework of Stochastic Mirror Descent (SMD) [8], an efficient method used to solve both constrained and unconstrained problems. We propose the following continuous-time *interacting* stochastic mirror descent (ISMD) algorithm for estimating the minimizer x^*,

$$dz_t^i = -\nabla f(x_t^i)dt + \theta\sum_{j=1}^{N} A_{ij}(z_t^j - z_t^i)dt + \sigma dB_t^i,$$
$$x_t^i = \arg\min_{x\in\mathcal{X}} D_\Phi(x, z_t^i),$$

where each particle $i = 1, ..., N$ is driven by an independent Brownian motion B_t^i, $D_\Phi(x, y)$ is the Bregman divergence and Φ is the mirror map. The interesting feature here is that particles interact through $A = \{A_{ij}\}_{i,j=1}^N$, which is an $N \times N$ doubly-stochastic matrix representing the interaction weights and θ represents the *interaction strength* which plays a crucial role in obtaining convergence. We remark that the Brownian motion noise is an approximation of the noise in the function or gradient estimate (see e.g. [3]).

III. The benefits

A. Variance reduction

It is well-known that due to the noise in the optimization algorithm, the optimization will only converge to a neighborhood of the minimizer. Standard approaches for achieving better convergence include using a vanishing batch size [7] at the cost of slower convergence or increasing the batch size over time [2] also at the cost of an increase in computational cost. Our first result is that using ISMD we can achieve a similar variance reduction and converge closer to x^* [1]:

Result 1. *For a convex function f with a sufficiently high θ we have,*

$$\mathbb{E}[f(x^T) - f(x^*)] \leq \mathcal{O}\left(\frac{1}{T}\right) + C\frac{\sigma^2}{N}.$$

with x^T the time-average of the trajectory.

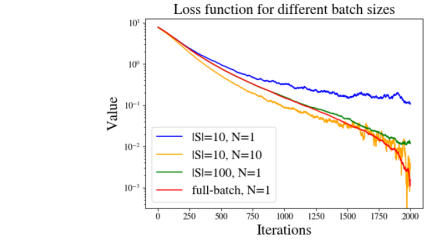
The above result shows that the distance to the optimum can be decreased by using more particles, i.e. increasing N. As shown in the top plot of Figure 1 we achieve comparable performance and convergence to the full batch gradient descent, when using a 'mini-batch' approach that uses N particles with a mini-batch size $|S|$ that is $1/N$ times the total number of data-points or summands in f.

B. Generalization

Averaging the trajectories of stochastic optimizers have shown to be beneficial in improving the generalization capability of the model by converging to flatter minima [5]. Here we show that interacting algorithms can achieve a similar result:

Result 2. *Let $\mathbf{z}$ be the vector of particles and $\mathcal{L}$ the Lagrangian of the interaction matrix. With interactions the algorithm optimizes $\sum_{i=1}^N f(x^i) + \frac{1}{2}\mathbf{z}^T\mathcal{L}\mathbf{z}$; without interactions it optimizes $\sum_{i=1}^N f(x^i)$. The additional term is a form of* regularizing *the objective function.*

Therefore, letting the N runs of the algorithm *interact*, besides the better convergence as described in the previous section, can also benefit the generalization capabilities of the model by smoothing the optimization objective [6]. This result can also be seen in the lower plots of Figure 1.



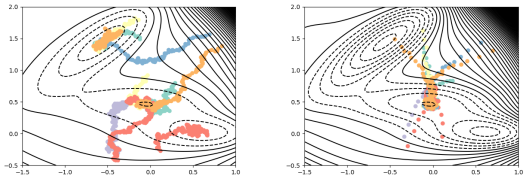


Fig. 1. Top: A linear model $f(x) = \frac{1}{m}\sum_{i=1}^m ||W_{i,\cdot}x - b_i||_2^2$. Using interacting particles allows to use a smaller batch size while still attaining convergence. Bottom: A non-convex function (Müller-Brown potential) with $N = 10$ using i.i.d. particles (left) and interacting particles with $\theta = 2$ (right). With interaction the particles converge to the flattest minimum (trace of Hessian in [-0.050,0.467] is 1702, compared to 4479 in [-0.558,1.442] and 3547 in [0.623,0.028])

IV. Conclusion

Our analysis showed that by controlling the interaction the variance of stochastic gradients could be reduced. At the same time interactions can help in converging to flatter minima which have been shown to generalize better. We believe interacting particles can be an effective optimization strategy.

REFERENCES

[1] A. BOROVYKH, P. PARPAS, N. KANTAS, AND G. PAVLIOTIS, *On stochastic mirror descent with interacting particles: convergence properties and variance reduction*, Submitted to Physica D., (2020).

[2] R. H. BYRD, G. M. CHIN, J. NOCEDAL, AND Y. WU, *Sample size selection in optimization methods for machine learning*, Mathematical programming, 134 (2012), pp. 127–155.

[3] P. CHAUDHARI, A. OBERMAN, S. OSHER, S. SOATTO, AND G. CARLIER, *Deep relaxation: partial differential equations for optimizing deep neural networks*, Research in the Mathematical Sciences, 5 (2018), p. 30.

[4] S. HOCHREITER AND J. SCHMIDHUBER, *Flat minima*, Neural Computation, 9 (1997), pp. 1–42.

[5] P. IZMAILOV, D. PODOPRIKHIN, T. GARIPOV, D. VETROV, AND A. G. WILSON, *Averaging weights leads to wider optima and better generalization*, arXiv preprint arXiv:1803.05407, (2018).

[6] N. KANTAS, P. PARPAS, AND G. A. PAVLIOTIS, *The sharp, the flat and the shallow: Can weakly interacting agents learn to escape bad minima?*, arXiv preprint arXiv:1905.04121, (2019).

[7] P. MERTIKOPOULOS AND M. STAUDIGL, *On the convergence of gradient-like flows with noisy gradient input*, SIAM Journal on Optimization, 28 (2018), pp. 163–197.

[8] A. S. NEMIROVSKY AND D. B. YUDIN, *Problem complexity and method efficiency in optimization.*, (1983).