A PDE APPROACH TO REGULARIZATION IN DEEP LEARNING

ADAM OBERMAN
JOINT WORK WITH CHAUDHARI, OSHER, SOATTO AND CARLIER

The fundamental tool in training deep neural networks is Stochastic Gradient Descent applied to the loss function, $f(x)$, which is high dimensional and nonconvex.

$$dx_t = -\nabla f(x_t)dt + \sqrt{\beta^{-1}}dW_t \tag{SGD}$$

There is a consensus in the field that some for of regularization of the loss function is needed, both for improving the generalization error and for accelerating the training time. However, there has been little progress in regularizing deep networks. Smoothing techniques, such a convolution, which are useful in low dimensions, are intractable due to the curse of dimensionality in the high dimensional setting.

Two recent algorithms have shown promise in this direction. The first, [ZCL15], used a mean field approach to perform SGD in parallel. The second, [CCS+16], replaced f in (SGD) with $f_\gamma(x)$, the *local entropy* of f, which is defined using notions from statistical physics [BBC+16].

We give a PDE interpretation of both algorithms [COO+17]. We show that, the algorithms, which can be interpreted as replacing (SGD) with a fast-slow system of SDEs, converge in the homogenization limit to the regularized evolution

$$dx_t = \nabla u(x, T-t) + \sqrt{\beta^{-1}}dW_t, \qquad 0 \le t \le T$$

where $u(x,t)$ is the solution of the viscous Hamilton-Jacobi PDE

$$u_t(x,t) + \frac{1}{2}|\nabla u(x,t)|^2 = \frac{\beta^{-1}}{2}\Delta u(x,t)$$

with initial data $u(x,0) = f(x)$.

We prove that, (for a slightly modified evolution) the expected value of the loss function is lower compared to (SGD).

The implementation is via the system of SDEs, which has a comparable computational cost to (SGD). Moreover, the parallel algorithm requires only passing the mean of the weights. Tools from optimal transportation [San17] are used to justify the fast convergence of the solution of the auxiliary problem.

In practice, this algorithm has significantly improved the training time (speed of convergence) for Deep Networks in high dimensions.

Current work is to improve the training in parallel. Making connections with Mean Field Games may lead to alternate interpretations or algorithmic improvements.

Date: June 28, 2017.

References

[BBC+16] Carlo Baldassi, Christian Borgs, Jennifer T Chayes, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes. *Proceedings of the National Academy of Sciences*, 113(48):E7655–E7662, 2016.

[CCS+16] 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, 2016. arXiv:arXiv:1611.01838.

[COO+17] Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto, and Guillame Carlier. Deep relaxation: partial differential equations for optimizing deep neural networks, 2017. arXiv:arXiv:1704.04932.

[San17] Filippo Santambrogio. {Euclidean, Metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 2017.

[ZCL15] Sixin Zhang, Anna E Choromanska, and Yann LeCun. Deep learning with elastic averaging sgd. In *Advances in Neural Information Processing Systems*, pages 685–693, 2015.