A PDE APPROACH TO REGULARIZATION IN DEEP LEARNING

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

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

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

In this talk we discuss a modification of (SGD) which significantly improves the training time as well as the generalization error [COO+17]. We also discuss a related algorithm also allows for effective training of DNNs in parallel [CBZ+17].

The algorithm is based on [CCS+16], which replaced f in (SGD) with $f_\gamma(x)$, the *local entropy* of f, which is defined using notions from statistical physics [BBC+16].

We show that the local entropy is the solution of a Hamilton-Jacobi equation.

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

where where T is a fixed time horizon, and $u(x,t)$ is the solution of initial value problem for the viscous Hamilton-Jacobi PDE

$$u_t(x,t) + \frac{1}{2}|\nabla u(x,t)|^2 = \frac{\beta^{-1}}{2}\Delta u(x,t), \qquad 0 \le t \le T$$

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

The gradient $\nabla u(x,t)$ can be computed using Langevin MCMC, by solving an auxiliary SGD equation.

Using the stochastic control interpretation of a slightly modified evolution, we prove that the expected value of the loss function is lower compared to (SGD).

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.

[CBZ+17] Pratik Chaudhari, Carlo Baldassi, Riccardo Zecchina, Stefano Soatto, and Ameet Talwalkar. Parle: parallelizing stochastic gradient descent, 2017. `arXiv:arXiv:1707.00424`.

[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`.

Date: August 27, 2017.