Implicit Regularization of SGD via Thermophoresis

Mingwei Wei[1], David J Schwab[2]

[1]Department of Physics and Astronomy, Northwestern University
[2]The Graduate Center, City University of New York

Abstract: A central ingredient in the impressive predictive performance of deep neural networks is optimization via stochastic gradient descent (SGD). Here we generalize the theory of thermophoresis from statistical mechanics and show that there exists an effective entropic force from SGD that pushes to reduce the gradient variance. We study this effect in detail in a simple two-layer model, where the thermophoretic force functions to decreases the weight norm and activation rate of the units. The strength of this effect is proportional to squared learning rate and inverse batch size, and is more effective during the early phase of training when the model's predictions are poor.

Thermophoresis in Physics

Thermophoresis, also known as the Soret effect, describes particle mass flow in response to both diffusion and a temperature gradient.

The resulting mass flow induced by diffusion and the temperature gradient is found to be

$$J = -D\nabla\rho - \rho D_T \nabla T$$

where D is the Einstein diffusion coefficient and D_T is defined as thermal diffusion coefficient.

Thermophoresis in General

We first define a kind of generalized random walk that has evolution equations for a particle state

$$\mathbf{q}_{t+1} = \mathbf{q}_t - \eta\gamma\mathbf{f}(\mathbf{q}_t, \xi)$$

where **f** is a vector function, γ and ξ are random variables, and η is a small number controlling the step size. It can be shown that there exists a mass flow

$$J = -\eta^2 \sqrt{\sum_{i\in U} g_i^2(\mathbf{q})} \sum_{i\in U} g_i(\mathbf{q})\partial_i\rho(\mathbf{q}) - \eta^2 \frac{\sum_{i,j\in U} g_i(\mathbf{q})g_j(\mathbf{q})\partial_j g_i(\mathbf{q})}{\sqrt{\sum_{i\in U} g_i^2(\mathbf{q})}}\rho(\mathbf{q}) + O(\eta^3)$$

where

$$g_i(\mathbf{q}) \coloneqq \sqrt{\int \gamma^2 f_i^2(\mathbf{q},\xi)d\mu(\gamma,\xi)}$$

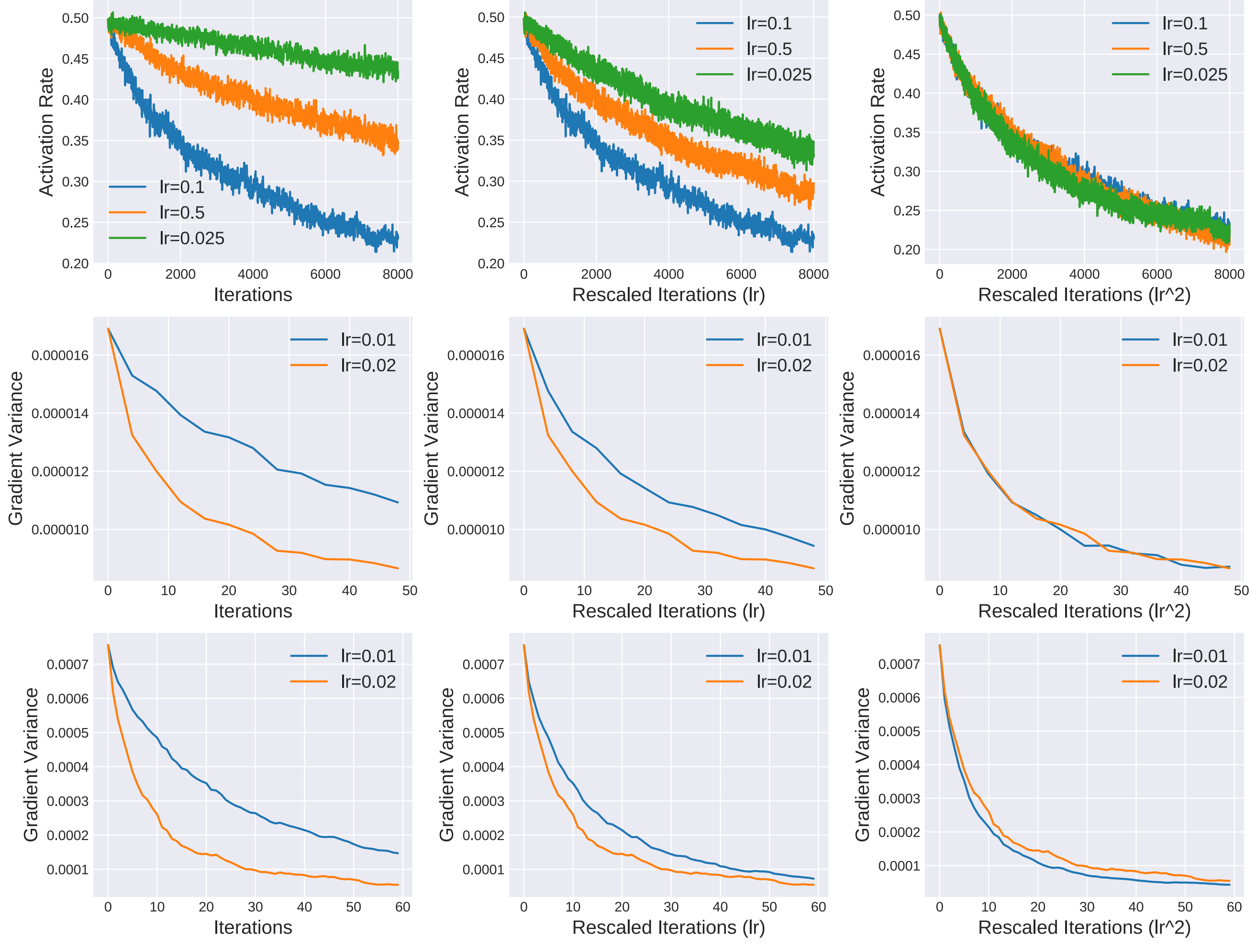
Thermophoresis in One Hidden Layer Networks

Consider the simple setting of one hidden layer networks:

$$f(\mathbf{x};\mathbf{V},\mathbf{W},\mathbf{b}) = \mathbf{V}\sigma(\mathbf{W}\mathbf{x}+\mathbf{b})$$

During training, the biases are pushed negative and the norm of V is suppressed during training, the effects of both of which are proportional to learning rate squared and inverse batch size 1/ | B | . There exists an effective force that pushes to decrease the model's activation rate and reduces the weight norm of the second layer.

The strength of this force scales as $F \propto \frac{\eta^2}{|B|}$.



DJS was partially supported by the Simons Foundation as an Investigator in the MMLS, the NSF through the Center for the Physics of Biological Function (PHY-1734030), and the NIH (R01EB026943-01).

Contact: dschwab@gc.cuny.edu