Winnowing with gradient descent

Manfred K. Warmuth MANFRED@GOOGLE.COM
Google Brain, Mountain View, CA 94043

Editor:

Abstract

We rewrite the non-negative weights w_i of multiplicative updates by ρ_i^2 and replace the multiplicative updates by a gradient descent step w.r.t. the new parameter ρ_i. We do this for the Winnow update, the Hedge update, and the unnnormalized and normalized exponentiated gradient (EG) updates for linear regression. When the original weights w_i are scaled to sum to one (as done for Hedge and normalized EG), then in the corresponding reparameterized update, the ρ_i parameters are now divided by $1/\|\boldsymbol{\rho}\|_2$ after the gradient descent step. We show that the reparameterizations closely track the original multiplicative updates by proving the same online regret bounds in each case (albeit in some cases with slightly worse constants). In particular, our work implies that the typical logarithmic dependence in the number of features of the multiplicative updates for sparse targets can also be achieved with the reparameterized gradient descent updates.

As a side, our work exhibits a simple linear 2-layer neural net that when trained with gradient descent can provable solve a certain sparse linear problem (known has the Hadamard problem) with exponentially fewer examples than any kernel method.

The talk will stay at a high level. In particular we will discuss the implication of these results on the power of neural nets versus kernel methods and on the influence of the training algorithm for neural nets. We conclude with a number of fundamental open problems.

Most of the presented work work was done jointly with Ehsan Amid.