Minibatch SGD without Replacement Drives Stable Convergence to Flat Minima

Minseok M. Kim[1], Yongjae Oh[1], Euijoon Kwon[2], Yunsik Choe[1],
Yongjoo Baek[1], and ***Youngkyoung Bae***[1]

[1]*Department of Physics and Astronomy, Seoul National University*

[2]*School of Physics, Korea Institute for Advanced Study*

While the implicit regularization of Stochastic Gradient Descent (SGD) is often attributed to the spatial covariance of gradient noise, the impact of its temporal correlations remains underexplored. Notably, recent studies have revealed that standard minibatch sampling without replacement inherently induces temporal anti-correlations in the gradient noise. To isolate and examine the effects of this anti-correlation in SGD, we introduce a tunable temporal anti-correlation into the gradient updates. By analyzing the effective dynamics with a finite learning rate in the over-parameterized regime, we prove that anti-correlated noise induces a Hessian-dependent implicit regularizer that drives optimization toward flat minima. Importantly, we identify a stability-regularization trade-off governed by the correlation time: while strong anti-correlation suppresses parameter fluctuations, a moderate level is required to maximize the bias toward flatness. We empirically validate our theory on matrix sensing and image classification tasks, demonstrating that controlling the temporal structure of the noise outperforms standard SGD.

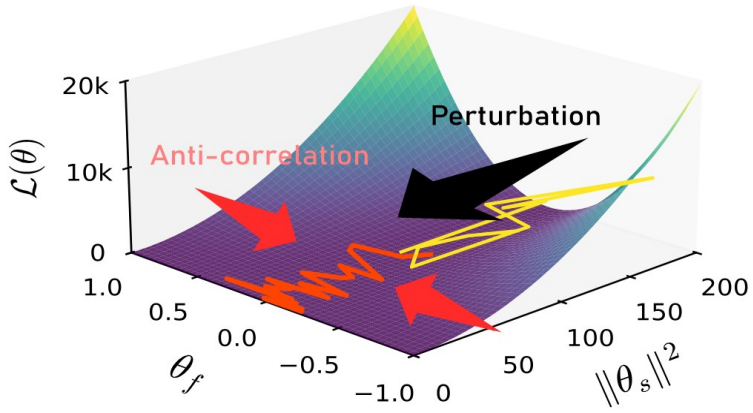

Figure 1. Schematics of the role of anti-correlated noise, in the SGD dynamics on the widening valley model (background loss surface). Red (yellow) line shows a path of parameters driven by anti-correlated (randomly perturbed) noise.