7 Role of variance in mini-batch gradients in finding better minima

Losses of deep neural network are usually highly non-convex containing a lot of local minima. A good optimizer should have the ability of escaping local and bad (i.e., sharp) minimizers to settle for a good/flat minimum. In SGD, the mini-batch gradient can be thought of as a noisy version of the full-batch gradient: $\nabla E_i(x) = \nabla E(x) + \eta_i$. So, when an optimizer is stuck in a valley having a bad local minima, the randomness in the noisy gradient $\nabla E_i(x)$ provides a possibility of **escaping** the valley (having a bad local minima). Very recently, this intuition has been mathematically formalized by Ibayashi & Imaizumi (2021). In their Theorem 2, the authors showed that the escape efficiency (reciprocal of mean exit time) of SGD is $\propto \exp(-\frac{B}{h}\Delta E \lambda_{max}^{-\frac{1}{2}})$, where B, h, ΔE and λ_{max} denote batch-size, learning rate, depth of minima and the largest eigenvalue of the Hessian, respectively. In short, a smaller batch-size (B) and a larger learning rate are crucial to escaping bad local minima.