From Stochastic Gradient-Descent to Ornstein-Uhlenbeck Process

When depicting the dynamics of Stochastic Gradient-Descent (SGD) on the information plane $(I(X;Z), I(Z;Y))$ over epochs, we observe a phase transition between:

1. *Drift Phase:* The layers increase the information on the labels, $I(Z;Y)$, while preserving the DPI order (lower layers have higher information), i.e., ERM.
2. *Diffusion Phase:* The layers information on the input, $I(X;Z)$, decreases and the layers "forget" irrelevant information until convergence, i.e., compression.

While the increase of $I(Z;Y)$ in the ERM phase is expected from the cross-entropy loss minimization, the surprising compression phase requires an explanation. There was no explicit regularization that could simplify the representations. The diffusion phase mostly adds random noise to the weights, and they evolve like Wiener processes, under the constraint of the training error. Such diffusion processes can be described by the Fokker-Planck equation, whose stationary distribution maximizes the entropy of the weights distribution, under the training error constraint. This maximizes $H(X|Z_i)$, or minimizes $I(X;Z_i) = H(X) - H(X|Z_i)$, because the input entropy $H(X)$ remains constant. This entropy maximization by additive noise, also known as *stochastic relaxation*, is constrained by the empirical error. By minimizing $I(X;Z_i)$ for each layer i, the diffusion phase leads to more compressed representations.

We can see the relation to diffusion explicitly when analyzing SGD as an Ornstein-Uhlenbeck process. Consider an objective function of the form $\mathcal{L}(\theta) = \sum_{n=1}^{N} l_n(\theta)$. Let $S(t)$ be a set of indices drawn uniformly at random from set $\{1, 2, \ldots, N\}$. We can form a stochastic estimate of the objective and a stochastic gradient,

$$\hat{L}(\theta) = \frac{N}{S} \sum_{n \in S} l_n(\theta), \tag{1}$$

$$\hat{g}_s(\theta) = \nabla_\theta \hat{L}(\theta). \tag{2}$$

In expectation, the stochastic gradient is the full gradient, $g(\theta) = \mathbb{E}(\hat{g}_s(\theta))$. The stochastic gradient is used for the update:

$$\theta(t+1) = \theta(t) - \hat{\epsilon}\hat{g}_S(\theta(t)). \tag{3}$$

To approximate SGD with a continuous time process, some assumptions are made:

1. Assume that the gradient noise $\hat{g}_S(\theta) - g(\theta)$ is Gaussian distributed.
2. Assume that the iterates $\theta(t)$ are constrained to a small enough region in parameter space that the sampling noise covariance of the stochastic gradients is constant.

3. Assume that the step size is small enough so that we can approximate the discrete-time Markov chain defined by SGD with a continuous-time Markov process.

Based on assumption 1, if S is big enough, then the central limit theorem should apply and we can write the stochastic gradient as

$$\hat{g}_S(\theta) \approx g(\theta) + \hat{\xi}_S(\theta), \quad \hat{\xi}_S(\theta) \sim \mathcal{N}(0, C(\theta)/S). \tag{4}$$

Decompose the covariance matrix C into $C = BB^T$, we introduce a rescaled noise covariance matrix $B_{\epsilon/S} = \sqrt{\epsilon/S}B$. And according to assumption 2 we can relax B to a constant irrelevant to θ. Then

$$\theta(t+1) - \theta(t) = -\epsilon g(\theta(t) + \sqrt{\epsilon}B_{\epsilon/S}W(t), \quad W(t) \sim \mathcal{N}(0, I). \tag{5}$$

In the continuous version is

$$d\theta(t) = -\nabla_\theta \mathcal{L}(\theta)dt + B_{\epsilon/S}dW(t). \tag{6}$$

As we want $\theta(t)$ to be close to the stationary point when $d\theta(t)$ is small, where the noise dominates the gradient. So we need a further assumption on the gradient, that is

4. Assume that the stationary distribution of the iterates is constrained to a region within which the objective is well approximated by a quadratic function.

Now we can write the $\nabla_\theta \mathcal{L}(\theta)$ explicitly as

$$d\theta(t) = -A(\theta(t) - \hat{\theta})dt + B_{\epsilon/S}dW(t). \tag{7}$$

The solution of SDE (7) is identified as a multidimensional Ornstein-Uhlenbeck process. The probability density function of the Ornstein-Uhlenbeck process satisfies the Fokker-Planck equation.