A Diffusion Theory for Deep Learning Dynamics: Stochastic Gradient Descent Exponentially Favors Flat Minima

Veprikov Andrey

Bayesian multimodeling
Department of Intelligent Systems, MIPT

November 2024

Stochastic Gradient Noise (SGN)

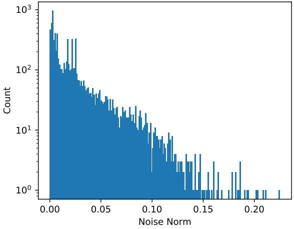
$$\theta_{t+1} = \theta_t - \eta \frac{\partial \hat{L}(\theta_t, x)}{\partial \theta_t} = \theta_t - \eta \frac{\partial L(\theta_t)}{\partial \theta_t} + \eta C(\theta_t)^{\frac{1}{2}} \zeta_t. \qquad (1)$$

- $\hat{L}(\theta)$ is the loss of one minibatch
- ζ_t is the noise variable
- $C(\theta)$ represents the gradient noise covariance matrix

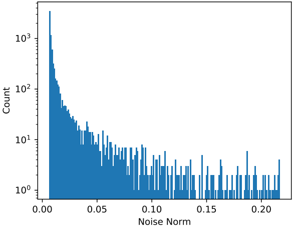
In the literature there are two main approaches of modeling SGN:

- Gaussian noise, $\zeta_t \sim \mathcal{N}(0, I)$.
- Lévy noise (stable variables)

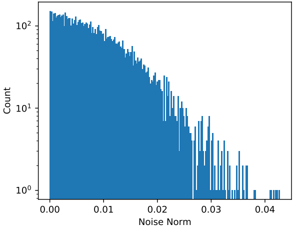
The Stochastic Gradient Noise Analysis



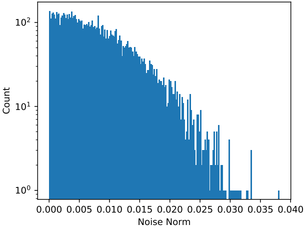
(a) “SGN"across parameters



(b) Lévy noise



(c) SGN across minibatches



(d) Gaussian noise

SGD Dynamics

The continuous-time ($\eta \to dt$) dynamics of SGD (1) is written as

$$d\theta = -\frac{\partial L(\theta)}{\partial \theta} dt + [2D(\theta)]^{\frac{1}{2}} dW_t,$$

where $dW_t \sim \mathcal{N}(0, Idt)$ and $D(\theta) = \frac{\eta}{2} C(\theta)$.

The associated Fokker-Planck Equation is written as

$$\frac{\partial P(\theta, t)}{\partial t} = \nabla \cdot [P(\theta, t) \nabla L(\theta)] + \nabla \cdot \nabla D(\theta) P(\theta, t). \tag{2}$$

In standard Stochastic Gradient Langevin Dynamics (SGLD), the injected gradient noise is fixed and isotropic Gaussian, $D = I$.

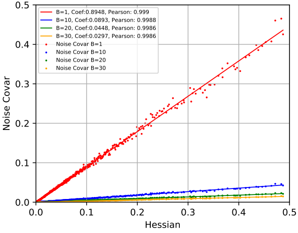
Formulating the SGN Covariance Matrix $C(\theta)$

$$
\begin{aligned}
C(\theta) &= \frac{1}{B}\left[\frac{1}{m}\sum_{j=1}^{m}\nabla L(\theta, x_j)\nabla L(\theta, x_j)^\top - \nabla L(\theta)\nabla L(\theta)^\top\right] \\
&\approx \frac{1}{Bm}\sum_{j=1}^{m}\nabla L(\theta, x_j)\nabla L(\theta, x_j)^\top \\
&= \frac{1}{B}\mathrm{FIM}(\theta) \approx \frac{1}{B}H(\theta).
\end{aligned}
$$

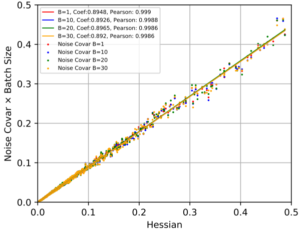
This approximately gives

$$
D(\theta) = \frac{\eta}{2}C(\theta) = \frac{\eta}{2B}H(\theta).
$$

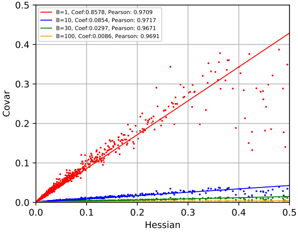
Empirical verification of $C(\theta) = H(\theta)/B$



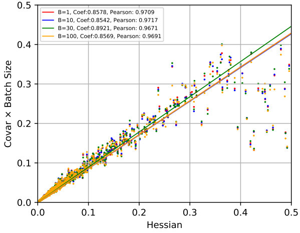
(a) Pretrained Model



(b) Pretrained Model



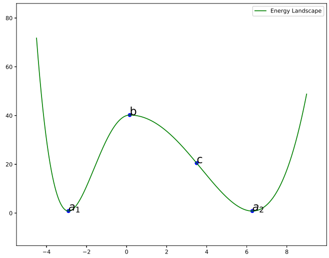
(c) Random Model



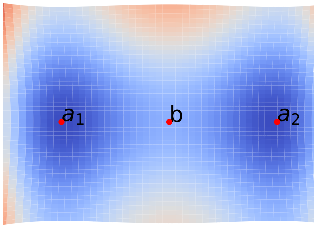
(d) Random Model

Kramers Escape Problem

- Sharp Valley a_1
- Flat Valley a_2
- Col b is the boundary between two valleys
- Col c locates outside of Valley a_1



(a) 1-Dimensional Escape



(b) High-Dimensional Escape

Definition of the mean escape time τ

We apply Gauss's Divergence Theorem to the Fokker-Planck Equation (2) resulting in

$$\nabla \cdot [P(\theta, t)\nabla L(\theta)] + \nabla \cdot \nabla D(\theta)P(\theta, t) = \frac{\partial P(\theta, t)}{\partial t} = -\nabla \cdot J(\theta, t),$$

where J is the probability current.

The mean escape time is expressed as

$$\tau = \frac{1}{\gamma} = \frac{P(\theta \in V_a)}{\int_{S_a} J \cdot dS}.$$

- $P(\theta \in V_a) = \int_{V_a} P(\theta)dV$ is the current probability inside Valley a
- J is the probability current produced by the probability source $P(\theta \in V_a)$

Assumptions

Assumption 1 (The Second Order Taylor Approximation)
The loss function around critical points $\theta^\star$ can be approximately written as

$$L(\theta) = L(\theta^\star) + g(\theta^\star)(\theta - \theta^\star) + \frac{1}{2}(\theta - \theta^\star)^\top H(\theta^\star)(\theta - \theta^\star).$$

Assumption 2 (Quasi-Equilibrium Approximation)
The system is in quasi-equilibrium near minima.

Assumption 3 (Low Temperature Approximation)
The gradient noise is small (low temperature).

Results for SGLD

Theorem 1 (SGLD Escapes Minima)

The loss function $L(\theta)$ is of class C^2 and n-dimensional. Only one most possible path exists between Valley a and the outside of Valley a. If Assumption 1, 2, and 3 hold, and the dynamics is governed by SGLD, then the mean escape time from Valley a to the outside of Valley a is

$$\tau = \frac{1}{\gamma} = 2\pi\sqrt{\frac{-\det(H_b)}{\det(H_a)}}\frac{1}{|H_{be}|}\exp\left(\frac{\Delta L}{D}\right).$$

- H_a and H_b are the Hessians of the loss function at the minimum a and the saddle point b
- $\Delta L = L(b) - L(a)$ is the loss barrier height
- e indicates the escape direction
- H_{be} is the eigenvalue of the Hessian H_b corresponding to the escape direction
- D is the diffusion coefficient, usually set to 1 in SGLD

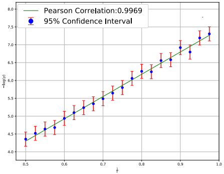
Results for SGD Diffusion

Theorem 2 (SGD Escapes Minima)

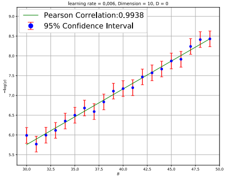
The loss function $L(\theta)$ is of class C^2 and n-dimensional. Only one most possible path exists between Valley a and the outside of Valley a. If Assumption 1, 2, and 3 hold, and the dynamics is governed by SGD, then the mean escape time from Valley a to the outside of Valley a is

$$\tau = 2\pi \frac{1}{|H_{be}|} \exp\left[\frac{2B\Delta L}{\eta} \left(\frac{s}{H_{ae}} + \frac{(1-s)}{|H_{be}|} \right) \right].$$

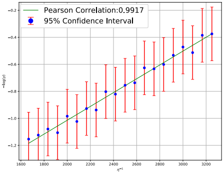
- $s \in (0, 1)$ is a path-dependent parameter
- H_{ae} and H_{be} are, respectively, the eigenvalues of the Hessians at the minimum a and the saddle point b corresponding to the escape direction e



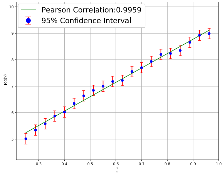
(a) $-\log(\gamma) = \mathcal{O}(\frac{1}{k})$



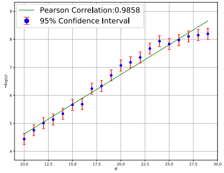
(b) $-\log(\gamma) = \mathcal{O}(B)$



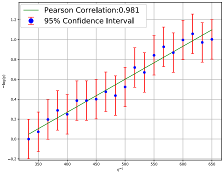
(c) $-\log(\gamma) = \mathcal{O}(\frac{1}{\eta})$



(d) $-\log(\gamma) = \mathcal{O}(\frac{1}{k})$



(e) $-\log(\gamma) = \mathcal{O}(B)$



(f) $-\log(\gamma) = \mathcal{O}(\frac{1}{\eta})$

Conclusion

- SGD favors flat minima exponentially more than sharp minima
- The ratio of the batch size and the learning rate exponentially matters
- Low dimensional diffusion
- High-order effects