Supplementary Materials:
The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects

Zhanxing Zhu [*1 2 3] **Jingfeng Wu** [*1] **Bing Yu** [1] **Lei Wu** [1] **Jinwen Ma** [1]

A. Derivations and Proofs for Main Paper

A.1. Derivation of Eq. (11) in main paper

Proof. The "mild smoothness assumptions" refers that $L(\theta_t) \in C^2$. Then the Ito's lemma holds (Øksendal, 2003). Thus,

$$
\begin{aligned}
&\mathrm{d}L(\theta_t) && (1)\\
=&\left(-\nabla L^T\nabla L + \frac{1}{2}\mathrm{Tr}\left(\Sigma_t^{\frac{1}{2}} H_t \Sigma_t^{\frac{1}{2}}\right)\right)\mathrm{d}t + \nabla L^T \Sigma_t^{\frac{1}{2}}\,\mathrm{d}W_t && (2)\\
=&\left(-\nabla L^T\nabla L + \frac{1}{2}\mathrm{Tr}\left(H_t\Sigma_t\right)\right)\mathrm{d}t + \nabla L^T \Sigma_t^{\frac{1}{2}}\,\mathrm{d}W_t. && (3)
\end{aligned}
$$

Taking expectation with respect to the distribution of θ_t, we have

$$
\mathrm{d}\mathbb{E}_{\theta_t} L(\theta_t) = \mathbb{E}\left(-\nabla L^T\nabla L + \frac{1}{2}\mathrm{Tr}(H_t\Sigma_t)\right)\mathrm{d}t, \quad (4)
$$

since the expectation of Brownian motion is zero.

Thus the solution of $\mathbb{E}_{\theta_t} L(\theta_t)$ is,

$$
\mathbb{E}L(\theta_t) = L(\theta_0) - \int_0^t \mathbb{E}\left(\nabla L^T\nabla L\right) + \int_0^t \frac{1}{2}\mathbb{E}\mathrm{Tr}(H_t\Sigma_t)\,\mathrm{d}t. \quad (5)
$$

□

A.2. Derivation of Eq. (13) in main paper

Proof. Without loss of generality, we assume that $L(\theta_0) = 0$.

For multivariate Ornstein-Uhlenbeck process, when $\theta_0 = 0$ is an constant, θ_t follows a multivariate Gaussian distribution (Øksendal, 2003).

For symmetric matrix A, let

$$
e^A := U^T \mathrm{diag}(e^{\lambda_1}, \dots, e^{\lambda_n}) U, \quad (6)
$$

where $\lambda_1, \dots, \lambda_n$ and U are the eigenvalues and eigenvector matrix of A.

Consider change of variables $\theta \to \phi(\theta, t) = e^{Ht}\theta_t$. Note that,

$$
\frac{\mathrm{d}e^{Ht}}{\mathrm{d}t} = He^{Ht}. \quad (7)
$$

Thus by applying Ito's lemma, we have

$$
\mathrm{d}\phi(\theta_t, t) = e^{Ht}\Sigma^{\frac{1}{2}}\,\mathrm{d}W_t, \quad (8)
$$

which we can integrate form 0 to t to obtain

$$
\theta_t = 0 + \int_0^t e^{H(s-t)}\Sigma^{\frac{1}{2}}\,\mathrm{d}W_s. \quad (9)
$$

The expectation of θ_t is zero. And by Ito's isometry (Øksendal, 2003), the covariance of θ_t is,

$$
\begin{aligned}
&\mathbb{E}\theta_t\theta_t^T && (10)\\
=&\mathbb{E}\left[\int_0^t e^{H(s-t)}\Sigma^{\frac{1}{2}}\,\mathrm{d}W_s\left(\int_0^t e^{H(r-t)}\Sigma^{\frac{1}{2}}\,\mathrm{d}W_r\right)^T\right] && (11)\\
=&\mathbb{E}\left[\int_0^t e^{H(s-t)}\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}e^{H(s-t)}\,\mathrm{d}s\right] && (12)\\
=&\mathbb{E}\left[\int_0^t e^{H(s-t)}\Sigma e^{H(s-t)}\,\mathrm{d}s\right] && (13)\\
=&\int_0^t e^{H(s-t)}\Sigma e^{H(s-t)}\,\mathrm{d}s. && (14)
\end{aligned}
$$

The last equation is because H and Σ are both constant.

[*] Equal contribution [1] School of Mathematical Sciences, Peking University, Beijing, China [2] Center for Data Science, Peking University, Beijing, China [3] Beijing Institute of Big Data Research, Beijing, China. Correspondence to: Zhanxing Zhu <zhanxing.zhu@pku.edu.cn>, Jingfeng Wu <pkuwjf@pku.edu.cn>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

Therefore

$$\begin{aligned}
\mathbb{E}L(\theta_t) &= \frac{1}{2}\mathbb{E}\text{Tr}\left(\theta_t^T H\theta_t\right) && (15)\\
&= \frac{1}{2}\text{Tr}\left(H\mathbb{E}\theta_t\theta_t^T\right) && (16)\\
&= \frac{1}{2}\int_0^t \text{Tr}\left(He^{H(s-t)}\Sigma e^{H(s-t)}\right)\mathrm{d}s && (17)\\
&= \frac{1}{2}\int_0^t \text{Tr}\left(e^{H(s-t)}H\Sigma e^{H(s-t)}\right)\mathrm{d}s && (18)\\
&= \frac{1}{2}\int_0^t \text{Tr}\left(e^{2H(s-t)}H\Sigma\right)\mathrm{d}s && (19)\\
&= \frac{1}{2}\text{Tr}\left(\frac{1}{2}H^{-1}\left(I-e^{-2Ht}\right)H\Sigma\right) && (20)\\
&= \frac{1}{4}\text{Tr}\left(\left(I-e^{-2Ht}\right)\Sigma\right). && (21)
\end{aligned}$$

Eq. (18) holds since H is symmetric. Further, by Taylor's expansion we have

$$\mathbb{E}L(\theta_t) = \frac{1}{4}\text{Tr}\left(\left(I-e^{-2Ht}\right)\Sigma\right) = \frac{t}{2}\text{Tr}(H\Sigma). \quad (22)$$

□

A.3. Proof of Proposition 1

Proof. $\text{Tr}(H\Sigma)$ can be decomposed as

$$\text{Tr}(H\Sigma) = \sum_{i=1}^{D}\lambda_i u_i^T\Sigma u_i. \quad (23)$$

Thus by the conditions of Proposition 1, we can bound $\text{Tr}(H\Sigma)$ as

$$\text{Tr}(H\Sigma) \geq u_1^T\Sigma u_1 \geq a\lambda_1\frac{\text{Tr}\Sigma}{\text{Tr}H}. \quad (24)$$

On the other hand,

$$\text{Tr}(H\bar{\Sigma}) = \frac{\text{Tr}\Sigma}{D}\text{Tr}H. \quad (25)$$

Thus,

$$\begin{aligned}
\frac{\text{Tr}(H\Sigma)}{\text{Tr}(H\bar{\Sigma})} \geq \frac{a\lambda_1 D}{(\text{Tr}H)^2} &\geq \frac{a\lambda_1 D}{\left(k\lambda_1+(D-k)D^{-d}\lambda_1\right)^2} && (26)\\
&= \mathcal{O}\left(aD^{2d-1}\right).
\end{aligned}$$

□

A.4. Proof of Proposition 2 in main paper

Proof. For simplicity, we define

$$\bar{f}(x;\theta) := \phi\circ f(x;\theta) \in [\delta, 1-\delta], \quad (27)$$

and

$$\ell(x,y;\theta) = \frac{1}{2}(\bar{f}(x;\theta)-y)^2. \quad (28)$$

Then the loss function becomes $L(\theta) = \mathbb{E}_{(x,y)}\ell(x,y;\theta)$.

Since both f and ϕ are piecewise linear, $\bar{f}(x;\theta)$ is also piece-wise linear with respect to θ. Thus the Hessian of $\bar{f}$ is zero almost everywhere.

We calculate the gradient and the Hessian of the loss:

$$\nabla_\theta L(\theta) = \mathbb{E}(\bar{f}(x;\theta)-y)\nabla_\theta\bar{f}(x;\theta); \quad (29)$$

$$\begin{aligned}
H(\theta) &= \nabla_\theta^2 L(\theta) && (30)\\
&= \mathbb{E}\nabla_\theta\bar{f}(x;\theta)\cdot\nabla_\theta\bar{f}(x;\theta)^T + \mathbb{E}(\bar{f}(x;\theta)-y)\nabla_\theta^2\bar{f}(x;\theta) && (31)\\
&= \mathbb{E}\nabla_\theta\bar{f}(x;\theta)\cdot\nabla_\theta\bar{f}(x;\theta)^T. \quad \text{almost everywhere.} && (32)
\end{aligned}$$

The last equation holds almost everywhere, since $\bar{f}(x;\theta)$ is piece-wise linear and its Hessian is zero almost everywhere.

On the other hand, the Fisher is

$$\begin{aligned}
F(\theta) &= \mathbb{E}\nabla_\theta\ell(x,y;\theta)\cdot\nabla_\theta\ell(x,y;\theta)^T && (33)\\
&= \mathbb{E}(\bar{f}(x;\theta)-y)^2\nabla_\theta\bar{f}(x;\theta)\cdot\nabla_\theta\bar{f}(x;\theta)^2. && (34)
\end{aligned}$$

(1) Note that $\bar{f}\in[\delta,1-\delta]$ and $y\in\{0,1\}$, thus

$$(\bar{f}(x;\theta)-y)^2 \geq \delta^2. \quad (35)$$

Therefore

$$F(\theta) \succeq \mathbb{E}\delta^2\nabla_\theta\bar{f}(x;\theta)\cdot\nabla_\theta\bar{f}(x;\theta)^2 = \delta^2 H(\theta), \quad (36)$$

holds almost everywhere.

(2) Around the minima where $\theta\in\{\theta : \left\|f(x;\theta)-y\right\| \leq \delta+\epsilon, \forall(x,y)\}$, we have

$$(\bar{f}(x;\theta)-y)^2 \leq (\delta+\epsilon)^2. \quad (37)$$

Therefore

$$F(\theta) \preceq \mathbb{E}(\delta+\epsilon)^2\nabla_\theta\bar{f}(x;\theta)\cdot\nabla_\theta\bar{f}(x;\theta)^2 = (\delta+\epsilon)^2 H(\theta), \quad (38)$$

holds almost everywhere around the minima. □

A.5. Proof of Proposition 3 in main paper

Proof. We only consider θ around the minima θ^* such that $\{\theta : \left\|\phi\circ f(x;\theta)-y\right\| \leq \delta+\epsilon, \forall(x,y)\}$. On the other hand by construction $\left\|\phi\circ f(x;\theta)-y\right\| \geq \delta$. Thus according to Proposition 2,

$$\delta^2 H(\theta) \preceq F(\theta) \preceq (\delta+\epsilon)^2 H(\theta) \quad (39)$$

holds almost everywhere.

Thus let $\lambda(\theta)$ and $u(\theta)$ being the maximal eigenvalue and its corresponding eigenvector of $H(\theta)$,

$$u(\theta)^T F(\theta) u(\theta) \geq \delta^2 u(\theta)^T H(\theta) u(\theta) = \delta^2 \lambda(\theta). \quad (40)$$

Since at the minimal θ^* the Hessian is not zero, thus there is a positive value $\lambda^* > 0$ such that $\lambda(\theta^*) > \lambda^* > 0$. Therefore by the continuity of $H(\theta)$, there are ϵ_1, δ_1, such that,

$$\lambda(\theta) > \lambda^* - \epsilon_1 > 0, \quad \forall \|\theta - \theta^*\| \leq \delta_1. \quad (41)$$

By Taylor's expansion,

$$\begin{aligned}\nabla L(\theta) &= \nabla L(\theta^*) + H(\theta^*)(\theta - \theta^*) + o(\theta - \theta^*) \\ &= H(\theta^*)(\theta - \theta^*) + o(\theta - \theta^*).\end{aligned} \quad (42)$$

Hence,

$$\left\|\nabla L(\theta)\right\|_2^2 \leq \left\|H(\theta^*)\right\|_2^2 \|\theta - \theta^*\|_2^2 + o\left(\|\theta - \theta^*\|_2^2\right). \quad (43)$$

Therefore, for all θ such that

$$\begin{aligned}\|\theta - \theta^*\|_2 &\leq \frac{\sqrt{\delta^2 \delta_2 (\lambda^* - \epsilon_1)}}{\left\|H(\theta^*)\right\|_2} && (44) \\ &\leq \frac{\sqrt{\delta^2 \delta_2 \lambda(\theta)}}{\left\|H(\theta^*)\right\|_2} && (45) \\ &\leq \frac{\sqrt{\delta_2 u(\theta)^T F(\theta) u(\theta)}}{\left\|H(\theta^*)\right\|_2}, && (46)\end{aligned}$$

we have

$$\left\|\nabla L(\theta)\right\|_2^2 \leq \delta_2 u(\theta)^T F(\theta) u(\theta) + o\left(\left|\delta_2 u(\theta)^T F(\theta) u(\theta)\right|\right). \quad (47)$$

On the other hand, by definition, the gradient covariance Σ and Fisher F has the following relationship,

$$\begin{aligned}\Sigma(\theta) &= \mathbb{E}(\nabla \ell(x, y; \theta) - \nabla L(\theta)) \cdot (\nabla \ell(x, y; \theta) - \nabla L(\theta))^T \\ &= \mathbb{E} \nabla \ell(x, y; \theta) \cdot \nabla \ell(x, y; \theta)^T - \nabla L(\theta) \nabla L(\theta)^T \\ &= F(\theta) - \nabla L(\theta) \nabla L(\theta)^T.\end{aligned} \quad (48)$$

Thus,

$$\begin{aligned}&\frac{u(\theta)^T \Sigma(\theta) u(\theta)}{\mathrm{Tr}\Sigma(\theta)} && (49) \\ =&\frac{u(\theta)^T F(\theta) u(\theta) - u(\theta)^T \nabla L(\theta) \nabla L(\theta)^T u(\theta)}{\mathrm{Tr}F(\theta) - \mathrm{Tr}(\nabla L(\theta) \nabla L(\theta)^T)} && (50) \\ =&\frac{u(\theta)^T F(\theta) u(\theta) - \left\|\nabla L(\theta)\right\|_2^2}{\mathrm{Tr}F(\theta) - \left\|\nabla L(\theta)\right\|_2^2} && (51) \\ \geq&\frac{u(\theta)^T F(\theta) u(\theta) - \left\|\nabla L(\theta)\right\|_2^2}{\mathrm{Tr}F(\theta)} && (52) \\ =&\frac{u(\theta)^T F(\theta) u(\theta)}{\mathrm{Tr}F(\theta)} \left(1 - \frac{\left\|\nabla L(\theta)\right\|_2^2}{u(\theta)^T F(\theta) u(\theta)}\right) && (53) \\ \geq&\frac{u(\theta)^T F(\theta) u(\theta)}{\mathrm{Tr}F(\theta)} \left(1 - \delta_2 - o\left(|\delta_2|\right)\right) && (54) \\ \geq&\frac{u(\theta)^T F(\theta) u(\theta)}{\mathrm{Tr}F(\theta)} \left(1 - 2\delta_2\right). && (55)\end{aligned}$$

Note that Eq. (39) indicates that

$$\begin{aligned}&\forall u, \quad u^T (F(\theta) - \delta^2 H(\theta)) u \geq 0 && (56) \\ &\text{and} \quad \mathrm{Tr}((\delta + \epsilon)^2 H(\theta) - F(\theta)) \geq 0. && (57)\end{aligned}$$

Thus

$$\begin{aligned}\frac{u(\theta)^T F(\theta) u(\theta)}{\mathrm{Tr}F(\theta)} &\geq \frac{\delta^2 u(\theta)^T H(\theta) u(\theta)}{(\delta + \epsilon)^2 \mathrm{Tr}H(\theta)} && (58) \\ &= \frac{\delta^2 \lambda(\theta)}{(\delta + \epsilon)^2 \mathrm{Tr}H(\theta)}. && (59)\end{aligned}$$

Therefore for all θ in the set of

$$\begin{aligned}&\left\{\left\|\phi \circ f(x; \theta) - y\right\| \leq \delta + \epsilon, \forall (x, y)\right\} \\ &\cap \left\{\|\theta - \theta^*\| \leq \delta^1\right\} \\ &\cap \left\{\|\theta - \theta^*\|_2 \leq \frac{\sqrt{\delta^2 \delta_2 (\lambda^* - \epsilon_1)}}{\left\|H(\theta^*)\right\|_2}\right\},\end{aligned} \quad (60)$$

we have

$$\begin{aligned}\frac{u(\theta)^T \Sigma(\theta) u(\theta)}{\mathrm{Tr}\Sigma(\theta)} &\geq \frac{u(\theta)^T F(\theta) u(\theta)}{\mathrm{Tr}F(\theta)} (1 - 2\delta_2) && (61) \\ &\geq \frac{(1 - 2\delta_2)\delta^2}{(\delta + \epsilon)^2} \frac{\lambda(\theta)}{\mathrm{Tr}H(\theta)}. && (62)\end{aligned}$$

□

B. About the non-convexity of the model in Proposition 2 in main paper

Suppose we only have one training data $\{x = (1, 1); y = 1\}$, and the threshold activation is

$$\phi(f) = \min\{\max\{f, 0.1\}, 0.9\}. \quad (63)$$

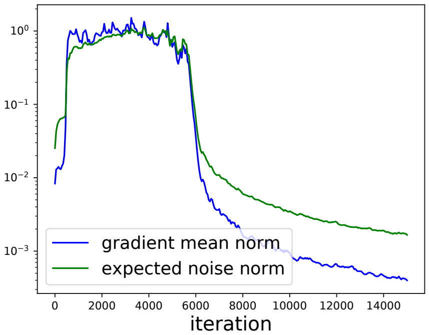

Figure 1. L_2 norm of gradient mean vs. the expected norm of noise during the training using SGD. The dataset and model are same as the experiments of FashionMNIST in main paper, or as in Section D.3

Thus the loss is

$$L(w_1, w_2) = (\phi(relu(w_1) - relu(w_2)) - 1)^2. \tag{64}$$

Hence

$$\begin{aligned} L(1,0) &= 0.01 \\ L(0,1) &= 0.81 \\ L(0.5,0.5) &= 0.81. \end{aligned} \tag{65}$$

Therefore

$$\frac{1}{2}L(1,0) + \frac{1}{2}L(0,1) < L(0.5,0.5), \tag{66}$$

which means that L is not convex.

It is also easy to see that L has multiple minima.

C. Additional experiments

C.1. Dominance of noise over gradient

Figure 1 shows the comparison of gradient mean and the expected norm of noise during training using SGD. The dataset and model are same as the experiments of FashionMNIST in main paper, or as in Section D.3. From Figure 1, we see that in the later stage of SGD optimization, the magnitude of noise indeed dominates that of gradient.

These experiments are implemented by TensorFlow 1.5.0.

C.2. The first 50 iterations of FashionMNIST experiments in main paper

Figure 2 shows the first 50 iterations of FashionMNIST experiments in main paper. We observe that SGD, GLD 1st eigvec(H), GLD Hessian and GLD leading successfully escape from the sharp minima found by GD, while GLD diag, GLD dynamic, GLD const and GD do not.

These experiments are implemented by TensorFlow 1.5.0.

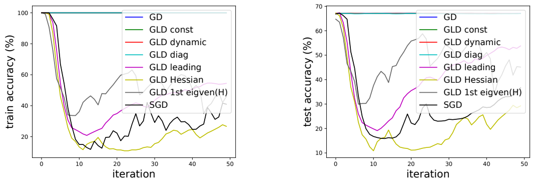

Figure 2. The fisrt 50 iterations of FashionMNIST experiments in main paper. Compared dynamics are initialized at θ^*_{GD} found by GD. The learning rate is same for all the compared methods, $\eta_t = 0.07$, and batch size $m = 20$. **Left**: Training accuracy versus iteration. **Right**: Test accuracy versus iteration.

D. Detailed setups for experiments in main paper

D.1. Two-dimensional toy example

Loss Surface The loss surface $L(w_1, w_2)$ is constructed by,

$$\begin{aligned} &s_1 = w_1 - 1 - x_1, \\ &s_2 = w_2 - 1 - x_2, \\ &\ell(w_1, w_2; x_1, x_2) = \min\{10(s_1 \cos\theta - s_2 \sin\theta)^2 \\ &\ + 100(s_1 \cos\theta + s_2 \sin\theta)^2, (w_1 - x_1 + 1)^2 + (w_2 - x_2 + 1)^2\}, \\ &L(w_1, w_2) = \frac{1}{N}\sum_{k=1}^{N} \ell(w_1, w_2; x_1^k, x_2^k), \end{aligned}$$

where

$$\begin{aligned} &\theta = \frac{1}{4}\pi, \\ &N = 100, \\ &x^k \sim \mathcal{N}(0, \Sigma), \quad \Sigma = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}. \end{aligned}$$

Note that Σ is the inverse of the Hessian of the quadric form generalizeing the sharp minima. And the 3-dimensional plot of the loss surface is shown in Figure 3.

Hyperparameters All learning rates are equal to 0.005. All dynamics concerned are tuned to share the same expected square norm, 0.01. The number of iteration during one run is 500.

These experiments are implemented by PyTorch $0.3.0$.

D.2. One hidden layer network

Hyperparameters The δ is set to be 0.001. The learning rate is 0.001. The optimizer is Adam for fast convergence, which does not affect our point on studying Tr$(H\Sigma)$.

The code is implemented in TensorFlow $1.9.0$.

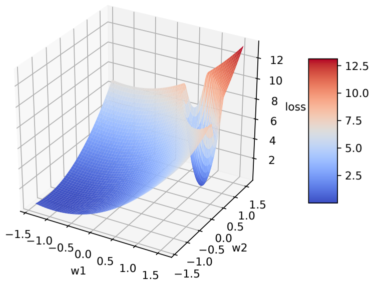

Figure 3. Constructed 2-dimensional surface in main paper.

D.3. FashionMNIST with corrupted labels

Dataset Our training set consists of $1,200$ examples randomly sampled from original FashionMNIST training set, and we further specify 200 of them with randomly wrong labels. The test set is same as the original FashionMNIST test set.

Model Network architecture:

$$\begin{aligned}\text{input} &\Rightarrow \text{conv1} \Rightarrow \text{max\_pool} \Rightarrow \text{ReLU} \Rightarrow \text{conv2}\\ &\Rightarrow \text{max\_pool} \Rightarrow \text{ReLU} \Rightarrow \text{fc1} \Rightarrow \text{ReLU}\\ &\Rightarrow \text{fc2} \Rightarrow \text{output}.\end{aligned}$$

Both two convolutional layers use 5×5 kernels with 10 channels and no padding. The number of hidden units between fully connected layers are 50. The total number of parameters of this network are $11,330$.

Training details

- **GD**: Learning rate $\eta = 0.1$. We tuned the learning rate (in diffusion stage) in a wide range of $\{0.5, 0.2, 0.15, 0.1, 0.09, 0.08, \dots, 0.01\}$ and no improvement on generalization.
- **GLD constant**: Learning rate $\eta = 0.07$, noise std $\sigma = 10^{-3}$. We tuned the noise std in range of $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ and no improvement on generalization.
- **GLD dynamic**: Learning rate $\eta = 0.07$.
- **GLD diagnoal**: Learning rate $\eta = 0.07$.
- **GLD leading**: Learning rate $\eta = 0.07$, number of leading eigenvalues $k = 20$, batchsize $m = 20$. We first randomly divide the training set into 60 mini batches containing 20 examples, and then use those minibatches to estimate covariance matrix.
- **GLD Hessian**: Learning rate $\eta = 0.07$, number of leading eigenvalues $= 20$, update frequence $f = 10$. Do to the limit of computational resources, we only update Hessian matrix every 10 iterations. But add Hessian generated noise every iteration. And to the same reason, we simplify set the coefficent of Hessian noise to $\sqrt{\mathrm{Tr}H/m\mathrm{Tr}\Sigma}$, to avoid extensively tuning of hyperparameter.
- **GLD 1st eigvec(H)**: Learning rate $\eta = 0.07$, as for GLD Hessian, and we set the coefficient of noise to $\sqrt{\lambda_1/m\mathrm{Tr}\Sigma}$, where λ_1 is the first eigenvalue of H.
- **SGD**: Learning rate $\eta = 0.07$, batchsize $m = 20$.

Estimation of Sharpness The sharpness are estimated by

$$\frac{1}{M}\sum_{j=1}^{M} L(\theta + \nu_j) - L(\theta), \quad \nu_j \sim \mathcal{N}(0, \delta^2 I), \tag{67}$$

with $M = 1,000$ and $\delta = 0.01$.

These experiments are implemented by TensorFlow 1.5.0.

D.4. SVHN and CIFAR-10

Dataset For SVHN experiments, we use $2,5000$ examples for training and $7,5000$ examples for test, to compromise with the computational burden of gradient descent. And for CIFAR-10 experiments, we use standard CIFAR-10 datasets. We do not use data augmentation since it could cause uncontrollable affects on analyzing SGD noise.

Model Standard VGG11 network without any regularizations including dropout, batch normalization, weight decay, etc. The total number of parameters of this network is $9,750,922$.

We choose VGG11 instead of ResNet because VGG11 achieves good generalization performance without using *Batch Normalization*, which has a subtle impact on SGD noise.

Training details Learning rates $\eta_t = 0.05$ are fixed for all optimizers, which is tuned for the best generalization performance of GD. The batch size of SGD is $m = 100$. The noise std of GLD constant is $\sigma = 10^{-3}$, which is tuned to best. Due to computational limitation, we only conduct experiments on GD, GLD const, GLD dynamic, GLD diag and SGD.

Estimation of Sharpness The sharpness are estimated by

$$\frac{1}{M}\sum_{j=1}^{M} L(\theta + \nu_j) - L(\theta), \quad \nu_j \sim \mathcal{N}(0, \delta^2 I), \tag{68}$$

with $M = 100$ and $\delta = 0.01$.

These experiments are implemented by PyTorch $1.0.0$.

References

Øksendal, B. Stochastic differential equations. In *Stochastic differential equations*, pp. 65–84. Springer, 2003.