CHAPTER 8. SGD STABILITY AND GENERALIZATION

Boao Kong **Kun Yuan**

November 7, 2023

1 Observation: SGD generalization than GD

SGD was first proposed to relieve the computational overhead suffered in GD. Surprisingly, it is often observed to generalize better than GD.

Here is an example. We trained a model to fit the FashionMINST. As is shown in figure 2 . When shifting the optimizer from GD to SGD at a point close to a global minimum, SGD escapes from that minimum and converges to another global minimum which generalizes better. The time it takes for the escape is very short compared to that required for SGD to converge [1].

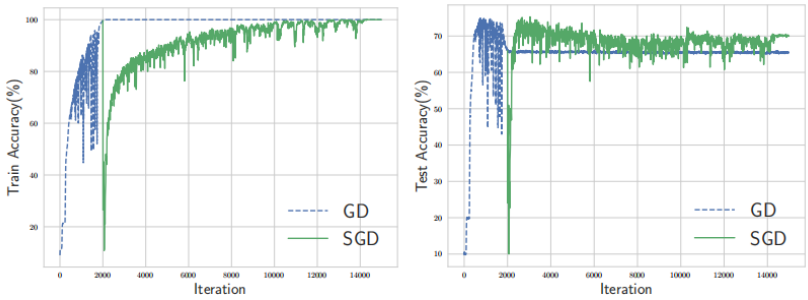

Figure 1: Fast escape phenomenon in fitting FashionMNIST.

From the result in the example above, it is essential to analyze the escape of SGD and its relationship with the generalization.

2 SGD=GD+noise

For $N > 0$, let $\{\xi_1, \xi_2, \cdots, \xi_N\}$ be the training set. Then consider the minimization of the following training error

$$F(x) = \frac{1}{N} \sum_{i=1}^{N} F(x, \xi_i) \tag{1}$$

optimized by SGD with batch size B:

$$x_{k+1} = x_k - \gamma g(x_k), \quad \forall\, k = 0, 1, 2, \cdots \tag{2}$$

where $g(x_k) = \frac{1}{B} \sum_{\xi \in B_K} \nabla F(x_k, \xi)$, B_k denotes a subset of the training set randomly selected a mini-batch of sample with size B which is independent of x_k. Then the term $g(x_k)$ provides an unbiased estimator of the full gradient $\nabla F(x_k)$. Assume the size of mini-batch B is large enough for the central limit theorem to hold, thus $g(x_k)$ follows a Gaussian distribution:

$$g(x_k) \sim \mathcal{N}\left(\nabla F(x_k), \Sigma^{\text{sgd}}(x_t)\right), \tag{3}$$

$$\Sigma^{\text{sgd}}(x_t) \approx \frac{1}{B}\left[\frac{1}{N}\sum_{i=1}^{N} \nabla F(x_k, \xi_i) \nabla F(x_k, \xi_i)^T - \nabla F(x_k) \nabla F(x_k)^T\right] \tag{4}$$

There we can rewrite Eq.(2) as,

$$x_{k+1} = x_t - \gamma \nabla F(x_t) + \gamma \varepsilon_k, \quad \varepsilon \sim \mathcal{N}\left(0, \Sigma^{\text{sgd}}(x_t)\right). \tag{5}$$

Eq. (5) consists of two parts, the first one is the iteration of GD, and the other is a Gaussian noise term ε_t. The noise may lead to favor of the solution, as it may drive the SGD iteration fast escape from the sharp minimum and converge to a flatter one [2].

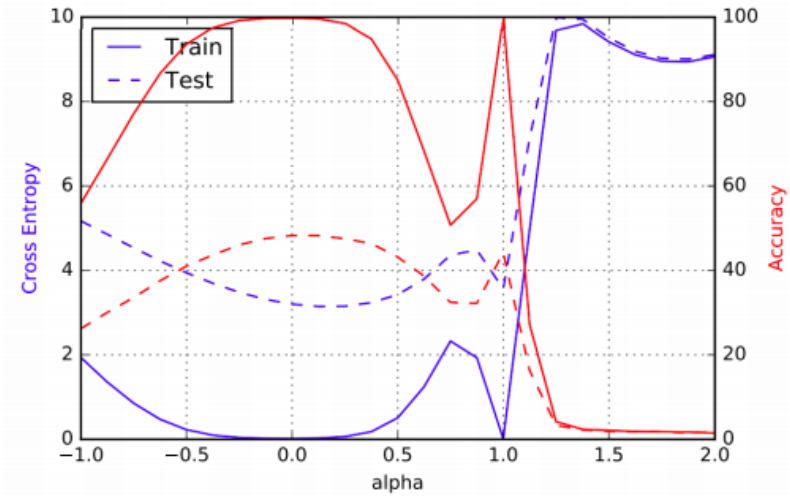

Figure 2: Fast escape phenomenon in fitting FashionMNIST.

Specifically, consider a Convolutional Neural Network optimized by SGO and GD. θ_{SGD} and θ_{GD} denotes the global minimum trained by SGD and GD, separately. Figure 2 shows the accuracy and training loss at $\theta(\alpha) = (1-\alpha)\theta_{\mathrm{SGD}}+\alpha\theta_{\mathrm{GD}}$. The global minimum θ_{SGD}, i.e. $\alpha = 0$, has a flatter landscape and a higher test accuracy than θ_{GD}, which corresponds to the case $\alpha = 1$. That result has announced the relationship between flatness and generalization. Next, we will present a theoretical understanding of the flatness and generalization of SGD through the theory of Linear Stability.

3 Linear Stability Analysis

For the stochastic dynamics (2), we firstly give the definition of fixed point.

Definition 3.1 (Fixed point)**.** We say x^* is a ***fixed point*** of stochastic dynamics (2), if for any training sample $\xi_i (i = 1, 2, \cdots, N)$, we have $\nabla F(x^*, \xi) = 0$.

It should be noted that this kind of fixed point does not always exist. However, for the *over-parametrized learning* problems, all the global minimum of the training loss $F(x)$ are fixed points. Note that if a fixed point is not stable, the noise of SGD will drive the optimizer move away. To formalize this, we introduced a kind of stability concept for the stochastic dynamics (2), i.e. linear stability:

Definition 3.2 (Linear stability)**.** Let x^* be a fixed point of stochastic dynamics (2). Consider the linearized dynamical system:

$$\hat{x}_{k+1} = \hat{x}_k - \gamma \nabla_x g(x_k)(\hat{x}_k - x^*). \tag{6}$$

We sat that x^* is ***linearly stable*** if there exists a constant C such that for all $k > 0$, we have:

$$\mathbb{E}[||\hat{x}_k||^2] \leq C||\hat{x}_0||^2 \tag{7}$$

Note that compared to the standard SGD iteration, the linearized dynamical system (6) replaces the gradient at x_k by its first-order approximation at the fixed point x^*. The linear stability defined above measures the instability by using the second-order moment of deviations. If x^* is not linearly stable, it is unlikely that x_k converges to x^*.

Let $H = \nabla^2 F(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla^2 F(x^*, \xi_i)$, it can be views of the sharpness of landscape of the training loss function. Then let $H_i = \nabla^2 F(x^*, \xi_i)$, and $\Sigma = \frac{1}{N}\sum_{i=1}^{N} H_i^2 - H^2$. Then we have the following sufficient condition of linear stable points of SGD:

Theorem 3.3 (Linear stability of SGD: A Sufficient Condition)**.** The global minimum x^* is linearly stable for SGD with learning rate γ and batch size B if the following condition is satisfied

$$\lambda_{\max}\left\{(I-\gamma H)^2+\frac{\gamma^2(N-B)}{B(N-1)}\Sigma\right\}\leq 1 \tag{8}$$

When x is 1-dim, this becomes a sufficient and necessary condition.

Proof. Without loss of generality, assume $x^*=0$. Then the linearized SGD is given by

$$x_{k+1}=x_k-\frac{\gamma}{B}\sum_{\xi\in B_k}\nabla^2 F(x^*,\xi)x_k \tag{9}$$

where B_k is a stochastic selected subset of the train set with the batch size B.

Hence, we have:

$$\mathbb{E}_{B_t}[x_{k+1}|x_k]=(I-\eta H)x_k \tag{10}$$

and

$$\mathbb{E}_{B_t}\left[||x_{k+1}||^2\middle|x_k\right]=\mathbb{E}_{B_t}\left[x_k^T\left(I-\frac{\gamma}{B}\sum_{\xi\in B_k}\nabla^2 F(x^*,\xi)\right)^2 x_k\middle|x_k\right] \tag{11}$$

$$=x_k^T\mathbb{E}_{B_k}\left[I-\frac{2\gamma}{B}\sum_{\xi\in B_k}\nabla^2 F(x^*,\xi)+\frac{\gamma^2}{B^2}\left(\sum_{\xi\in B_k}\nabla^2 F(x^*,\xi)\right)^2\right]x_k \tag{12}$$

$$=x_k^T\left[1-2\gamma H+\frac{\gamma^2}{B^2}\left(\frac{B(N-B)}{N(N-1)}\sum_{i=1}^N H_i^2+\frac{NB(B-1)}{N-1}H^2\right)\right]x_k \tag{13}$$

$$=x_k^T\left[(I-\gamma H)^2+\frac{\gamma^2(N-B)}{B(N-1)}\Sigma\right]x_k \tag{14}$$

where Eq.(13) is from the following equation:

$$\mathbb{E}_{B_t}\left(\sum_{\xi\in B_k}\nabla^2 F(x^*,\xi)\right)^2=\mathbb{E}_{B_t}\left(\sum_{\xi\in B_k}\nabla^2 F(x^*,\xi)^2+2\sum_{\xi,\zeta\in B_k}\nabla^2 F(x^*,\xi)\nabla^2 F(x^*,\zeta)\right) \tag{15}$$

$$=\frac{B}{N}\sum_{i=1}^N H_i^2+\frac{2B(B-1)}{N(N-1)}\sum_{i\neq j}H_iH_j \tag{16}$$

$$=\frac{B(N-B)}{N(N-1)}\sum_{i=1}^N H_i^2+\frac{B(B-1)}{N(N-1)}H^2 \tag{17}$$

Therefore, if we have

$$\lambda_{\max}\left\{(I-\gamma H)^2+\frac{\gamma^2(N-B)}{B(N-1)}\Sigma\right\}\leq 1 \tag{18}$$

we have as $t\to\infty$

$$\mathbb{E}||x_t||^2\leq\lambda_{\max}^t\left\{(I-\gamma H)^2+\frac{\gamma^2(N-B)}{B(N-1)}\Sigma\right\}\mathbb{E}||x_0||^2\leq\mathbb{E}||x_0||^2 \tag{19}$$

therefore x^* is linear stable.

If x is 1-dim, then H and Σ are sclars, and we have

$$\mathbb{E}x_{k+1}^2=\left[(I-\gamma H)^2+\frac{\gamma^2(N-B)}{B(N-1)}\Sigma\right]\mathbb{E}x_k^2 \tag{20}$$

$$=\left[(I-\gamma H)^2+\frac{\gamma^2(N-B)}{B(N-1)}\Sigma\right]^{t+1}\mathbb{E}x_0^2 \tag{21}$$

In this case, if

$$\left[(I-\gamma H)^2+\frac{\gamma^2(N-B)}{B(N-1)}\Sigma\right]>1 \tag{22}$$

then $\mathbb{E}x_t^2\to+\infty(t\to+\infty)$, which means that $x^*=0$ is not stable. □

Since GD can be viewed as a special case of SGD with full batch size $B=N$, from Theorem 3.3, we can give the following corollary of the case of GD:

Corollary 3.4 (Linear stability of GD: A Sufficient Condition)**.** The global minimum x^* is linearly stable for GD with learning rate γ and batch size B if the following condition is satisfied

$$||H||_2\leq\frac{2}{\gamma} \tag{23}$$

When x is 1-dim, this becomes a sufficient and necessary condition.

Proof. From Eq.(8) we have a sufficient condition of linear stability of GD:

$$\lambda_{\max}\left\{(I-\gamma H)^2\right\}\leq 1 \tag{24}$$

which means:

$$\max|\lambda_j(I-\gamma H)|\leq 1. \tag{25}$$

where $\lambda_j(A)$ denotes to the $j-$th eigenvalue of the matrix A. So Eq. (25) is equivalent to

$$||H||_2\leq\frac{2}{\gamma} \tag{26}$$

□

To understand the flatness around the linearly stable point, we consider the case when $d = 1$. In this case, H and Σ are both scalar, denoted as a and s Eq. (8) becomes the condition

$$(1 - \gamma a)^2 + \frac{\gamma^2(N - B)}{B(N - 1)} s^2 \tag{27}$$

Now assume that the learning rate γ is fixed. We use a and s as features of the global minima, to show how GD and SGD "selects" minima. From Eq. (27), we know that the global minima that GD can converge to satisfy $a \leq 2/\gamma$, and the global minima that SGD, as the second term of Eq. (27) does not equal to 0, so it leads to a more strict restrict of the flatness a. The stability regions are visualized in Figure 3 , which is called the sharpness-non-uniformity diagram [1].

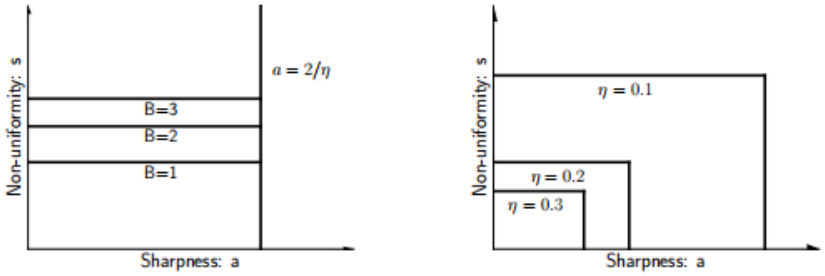

Figure 3: The sharpness-non-uniformity diagram.

From the sharpness-non-uniformity diagram we see that, when the learning rate is fixed, the set of global minima that are linearly stable for SGD is much smaller than that for GD. This means that compared to GD, SGD can filter out global minima with large non-uniformity.

3.1 A Necessary Condition with Squared Loss

In particular, if the loss function is given by

$$F(x, \xi_i) = \frac{1}{2}(h(x; u_i) - v_i)^2, \quad i = 1, 2, \cdots, N \tag{28}$$

where $\xi_i = (u_i, v_i)$ is a sample in training set. Then the following theorem 3.5 shows SGD may lead to an upper bound of $\mathrm{tr}(H)$ [3]:

Theorem 3.5 (Linear stability of SGD: A Necessary Condition)**.** Given the loss function as Eq. (28), the global minimum x^* is linearly stable for SGD with learning rate γ and batch size 1, then the following condition is satisfied

$$\mathrm{tr}(G(x^*)) \leq \frac{2}{\gamma} \tag{29}$$

where $G(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla F(x,\xi_i)\nabla F(x,\xi_i)^T$ be the associate empirical Fisher matrix, which equals to H at x^*.

Proof. For $i = 1, 2, \cdots, N$, let $g_i(x) = F(x, \xi_i)$, $e_i(x) = f(x; u_i) - v_i$, then the Fisher matrix G and Hessian matrix H can be rewritten as:

$$G(x) = \frac{1}{N}\sum_{i=1}^{N} g_i(x)g_i(x)^T \tag{30}$$

$$H(x) = \frac{1}{N}\sum_{i=1}^{N} g_i(x)g_i(x)^T + \frac{1}{N}\sum_{i=1}^{N} e_i(x)\nabla^2 F(x,(u_i,v_i)) \tag{31}$$

In over-parameterization setting, for all $i = 1, 2, \cdots, n$, we have $e_i(x) = f(x; u_i) - v_i = 0$ if x is a global minimum. Then $H(x^*) = G(x^*)$.

Let $\delta_k = x_k - x^*$, then the iteration of linearized SGD (6) with batch size 1 can be written as:

$$\delta_{k+1} = \left[I - \gamma H_{i_k}\right]\delta_k, \quad i_k \sim \mathrm{Unif}([N]) \tag{32}$$

where $H_{i_k} = \nabla^2 F(x^*, \xi_{i_k})$. Then,

$$\delta_{k+1}\delta_{k+1}^T = \left(I - \gamma H_{i_k}\right)\delta_k\delta_k^T\left(I - \gamma H_{i_k}\right) \tag{33}$$

$$= \delta_k\delta_k^T - \gamma\left(\delta_k\delta_k^T H_{i_k} + H_{i_k}\delta_k\delta_k^T\right) + \gamma^2 H_{i_k}\delta_k\delta_k^T H_{i_k} \tag{34}$$

Let $Q_k = \mathbb{E}[\delta_k\delta_k^T]$ be the deviation covariance matrix. Then taking the expectation of Eq. (34) gives:

$$Q_{k+1} = Q_k - \gamma(Q_k H + HQ_k) + \gamma^2\mathbb{E}\left[H_{i_k}\delta_k\delta_k^T H_{i_k}\right] \tag{35}$$

$$= (id - \gamma T_\gamma)Q_k \tag{36}$$

where $T_\gamma : \mathcal{S}_+ \to \mathcal{S}_+$ is given by

$$T_\gamma A = (HA + AH) - \gamma\mathbb{E}_{i\sim\mathrm{Unif[N]}}\left[H_i A H_i\right] \tag{37}$$

And if x^* is linearly stable, then there exist $C > 0$ such that

$$\left\|\mathbb{E}[\delta_k\delta_k^T]\right\|_F \leq C\left\|\mathbb{E}[\delta_0\delta_0^T]\right\|_F, \quad \forall k > 0 \tag{38}$$

By Eq. (36), to ensure Eq. (38), we need $T_\gamma \succeq 0$. This is equivalent to it holds that

$$\langle A, T_\gamma A\rangle = 2\mathrm{Tr}(AHA) - \gamma \mathbb{E}_{i\sim \mathrm{Unif[N]}}\left[\mathrm{Tr}(AH_i)AH_i\right] \geq 0, \quad \forall A \in \mathcal{S}_+. \tag{39}$$

Noting that

$$\mathbb{E}_{i\sim \mathrm{Unif[N]}}\left[\mathrm{Tr}(AH_i)AH_i\right] = \left(\mathbb{E}_{i\sim \mathrm{Unif[N]}}\left[g_i(x^*)Ag_i(x^*)^T\right]\right)^2 = \mathrm{Tr}^2(HA) \tag{40}$$

So, Eq. (39) implies:

$$2\mathrm{Tr}(HA^2) - \gamma \mathrm{Tr}^2(HA) \geq 0, \quad \forall A \in \mathcal{S}_+. \tag{41}$$

Taking $A = \mathrm{diag}(w_1, w_2, \cdots, w_p)$, we obtain

$$\frac{\left(\sum_j \lambda_j(H) w_j\right)^2}{\sum_j \lambda_j(H) w_j^2} \leq \frac{2}{\gamma} \tag{42}$$

Specifically, take $w_j = 1$ for $j = 1, 2, \cdots, n$ and then complete the proof. □

Remark. It should be stressed that the stability condition (42) is stronger than $\mathrm{Tr}(H) \leq 2/\gamma$. We only state the latter in the main text since it is more intuitive, has a clean relationship to sharpness, and is workable for further analysis in Section 4.

4 The Relationship between Sharpness and Generalization: A Case Study

In last section, we get the conclusion that the linear stability of SGD and GD lead to a restriction to $||H||_2$ and $\mathrm{Tr}(H)$, respectively. Here, we analyze the relationship between sharpness and generalization in a two-layer ReLU network. Consider the network: $h(x; u) = \sum_{j=1}^m a_j \sigma(w_j^T u)$, where $a_j \in \mathbb{R}$ and $w_j \in \mathbb{R}^d$, m denotes the network width and $\sigma(t) = \max\{t, 0\}$ is the ReLU function. The loss function is squared which is defined in Eq. (28). We also assume the input distribution to be $\rho = \mathrm{Unif}(\sqrt{d}\mathbb{S}^{d-1})$.

For $q > 0$, define the weighted l_2 norm:

$$||x||_{2,q} = \sum_j (||w_j||^2 + q a_j^2) \tag{43}$$

Then Theorem 4.1 can announce that both the 2-norm and trace of Hessian can build the equivalence to such $l_{2,q}$ norms:

Theorem 4.1. For any $\delta \in (0,1)$, let $N(d,\delta) = \inf\{n \in \mathbb{N} | d\log(n/\delta)/n \leq 1\}$.

- If $n \gtrsim N(d,\delta)$, then with the probably of $1-\delta$ we have

$$\mathrm{Tr}(G(x)) \sim ||x||_{2,d} \tag{44}$$

- If $n \gtrsim dN(d,\delta)$, then with the probably of $1-\delta$ we have

$$||G(x)||_2 \sim ||x||_{2,1} \tag{45}$$

For ReLU networks, it is well-known that the generalization gap can be controlled by the path norm $||x||_{\mathcal{P}} := \sum_j |a_j|||w_j||$, i.e.

$$\text{gen-gap}(x) \lesssim \frac{d}{N}||x||_{\mathcal{P}}^2 \tag{46}$$

Then, for the linearly stable global minimum x^*_{SGD}, the generalization gap can be informally bounded as follows:

$$\begin{aligned}\text{gen-gap}(x^*_{\mathrm{SGD}}) \lesssim \frac{d}{N}||x||_{\mathcal{P}}^2 = \frac{d}{N}\left(\sum_j |a_j|||w_j||\right)^2 = \frac{1}{4N}\left(2\sqrt{d}\sum_j |a_j|||w_j||\right)^2 \\ \lesssim \frac{1}{4N}\left(\sum_j |a_j|^2 + d||w_j||^2\right)^2 \lesssim \frac{1}{4N}\mathrm{Tr}(H)^2 \lesssim \frac{1}{N\gamma^2}\end{aligned} \tag{47}$$

For x^*_{GD}, we can take the same estimation:

$$\begin{aligned}\text{gen-gap}(x^*_{\mathrm{GD}}) \lesssim \frac{d}{N}||x||_{\mathcal{P}}^2 = \frac{d}{N}\left(\sum_j |a_j|||w_j||\right)^2 = \frac{d}{4N}\left(2\sum_j |a_j|||w_j||\right)^2 \\ \lesssim \frac{d}{4N}\left(\sum_j |a_j|^2 + ||w_j||^2\right)^2 \lesssim \frac{d}{4N}||H||_2^2 \lesssim \frac{d}{N\gamma^2}\end{aligned} \tag{48}$$

This result shows that the flatter minimum has a lower upper bound of the generalization gap. In addition, the linear stability of SGD provides a generalization guarantee independent to the model size. Thus, the size-independent nature of sharpness control strengthens the assurance of favorable generalization properties for the stable minima.

From (47) and (48), we can bound the test loss of SGD and GD, which is shown in Theorem 4.2:

Theorem 4.2. For SGD and GD with the same learning rate γ, denote by $\hat{x}_{sgd}$ and $\hat{x}_{gd}$ the linearly stable minimum of SGD and GD, respectively. Suppose $\sup_{x\in\mathcal{X}} |f^*(x)| \leq 1$. For any $\delta \in (0,1)$, if $n \gtrsim N(d,\delta)$, then the following holds with the probably of at least $1-\delta$:

$$f(\hat{x}_{sgd}) \lesssim \frac{B}{\gamma^2 N}, \quad f(\hat{x}_{gd}) \lesssim \frac{Bd}{\gamma^2 N} \tag{49}$$

where $f(x) = \mathbb{E}_{\xi\sim\mathcal{D}} F(x,\xi)$ is the test loss, $B = \log^3 n + \log(1/\delta)$.

5 Sharpness Aware Minimization (SAM)

Since the sharpness of loss landscape may harm the generalization, we can propose an approach to optimize Eq. (1) by seeking a minimum that has both low training loss and flatter landscape, i.e., find a parameter values whose entire neighborhoods have uniformly low training loss. Motivated by such an idea, we propose to select parameter values by solving the following ***Sharpness Aware Minimization*** (SAM) problem:

$$\min_x F^{SAM}(x) + \lambda ||x||_2^2 \tag{50}$$

where $F^{SAM}(x) = \max_{||\varepsilon||_2\leq\rho} F(x+\varepsilon)$ and $\rho > 0$. Here, as Eq. (51) shows, F^{SAM} consists of two parts, one is the traditional train loss, and the other measures how quickly the training loss can be increased by moving from x to a nearby parameter value, which can be seen as a metric of shapeness.

$$F^{SAM}(x) = F(x) + \max_{||\varepsilon||_2\leq\rho} [F(x+\varepsilon) - F(x)] \tag{51}$$

In practice, it is difficult to find the exact value of $F^{SAM}(x)$, but it can be effectively estimated by

$$\varepsilon^*(x) := \arg\max_{||\varepsilon||_2\leq\rho} F(x+\varepsilon) \approx \arg\max_{||\varepsilon||_2\leq\rho} \left[F(x) + \varepsilon^T \nabla_x F(x)\right] = \arg\max_{||\varepsilon||_2\leq\rho} \varepsilon^T \nabla_x F(x). \tag{52}$$

Then,

$$\varepsilon^*(x) \approx \rho \frac{\nabla_x F(x)}{\|\nabla_x F(x)\|_2}. \tag{53}$$

So the gradient of $F^{SAM}(x)$ can be estimated by

$$\begin{aligned} \nabla_x F^{SAM}(x) \approx \nabla_x F(x+\varepsilon^*(x)) &= \frac{d(x+\varepsilon^*(x))}{dx} \nabla_x F(x)|_{x+\varepsilon^*(x)} \\ &= \nabla_x F(x)|_{x+\varepsilon^*(x)} + \frac{d(\varepsilon^*(x))}{dx} \nabla_x F(x)|_{x+\varepsilon^*(x)} \end{aligned} \tag{54}$$

To further accelerate the computation, we drop the second-order terms and then obtain our

final gradient approximation:

$$\nabla_x F^{SAM}(x) \approx \nabla_x F(x)|_{x+\varepsilon^*(x)} \tag{55}$$

Finally, we have the iteration of SAM algorithm:

$$x_{k+1} = x_k - \gamma \cdot \frac{1}{B} \sum_{\xi \in B_t} \nabla F(x_k, \xi)|_{x_k + \varepsilon_k^*(x_k)} \tag{56}$$

where $\varepsilon_k^*(x_k) = \rho \dfrac{\nabla_x F(x_k, \xi)}{\|\nabla_x F(x_k, \xi)\|_2}$.

Figure 4 has shown the schematic of the SAM iteration. Instead of using the gradient at x_k, SAM uses the gradient at $x_k + \varepsilon^*(x_k)$ to update x_k, which may make the parameter having a flatter landscape. More detailed information on SAM can refer to the literature by Foret et al. [4].

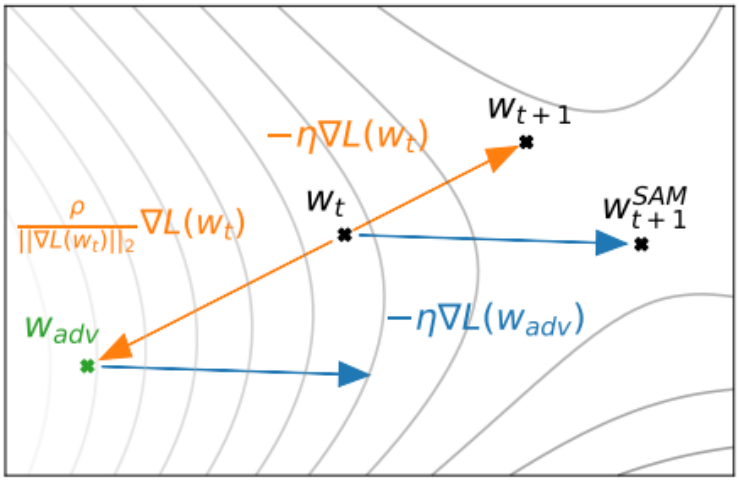

Figure 4: Schematic of the SAM parameter update.

References

[1] L. Wu, C. Ma *et al.*, "How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[2] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," *arXiv preprint arXiv:1609.04836*, 2016.

[3] L. Wu and W. J. Su, "The implicit regularization of dynamical stability in stochastic gradient descent," *arXiv preprint arXiv:2305.17490*, 2023.

[4] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *arXiv preprint arXiv:2010.01412*, 2020.