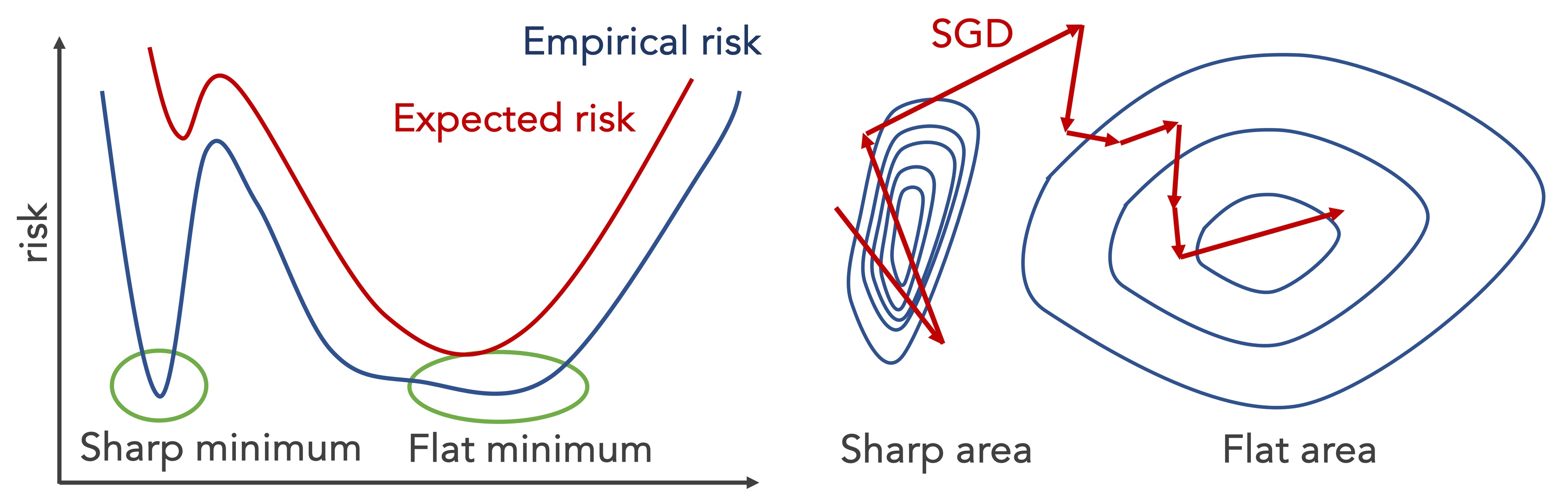
Why is Parameter Averaging Beneficial in SGD? An Objective Smoothing Perspective

Atsushi Nitanda [1,2] Ryuhei Kikuchi [2] Shugo Maeda [2] Denny Wu [3,4]

1 Centre for Frontier AI Research, CFAR
2 Kyutech, Kyushu Institute of Technology
3 NYU
4 FLATIRON INSTITUTE

1. Implicit Bias towards a Flat Minimum

Folklore: A flat minimum is better than sharp minima.
And stochastic gradient descent (SGD) prefers a flat minimum.



- [Kleinberg et al. (2018)] showed SGD approximately minimizes the smoothed objective convolved with the stochastic gradient noise.
- Averaged SGD (ASGD, SWA) also converges to a flat minimum.

We study the capability of ASGD to minimize the smoothed objective functions.

2. Alternative View of SGD and ASGD

Objective: $f : \mathbb{R}^d \to \mathbb{R}$: a nonconvex smooth function.

Stochastic Gradient Descent: for a random field $\epsilon_{t+1} : \mathbb{R}^d \to \mathbb{R}^d$

$$w_{t+1} = w_t - \eta(\nabla f(w_t) + \epsilon_{t+1}(w_t)).$$

An Alternative view of SGD [Kleinberg et al. (2018)]:
Through the change of variable $v_t = w_t - \eta \nabla f(w_t)$, SGD becomes

$$v_{t+1} = v_t - \eta \epsilon'_{t+1}(v_t) - \eta \nabla f(v_t - \eta \epsilon'_{t+1}(v_t)).$$

That is, SGD can be considered as the optimization method for

$$F(v) = \mathbb{E}[f(v - \eta \epsilon'(v))].$$

(However, note $\nabla F(v_t) \neq \mathbb{E}[\nabla f(v_t - \eta \epsilon'_{t+1}(v_t)]$.)

$F(v)$ is a smoothed objective that penalizes high curvature:

$$F(v) = f(v) + \frac{\eta^2}{2}\mathrm{Tr}(\nabla^2 f(v) \mathbb{E}[\epsilon'(v)\epsilon'(v)^\top]) + O(\eta^3).$$

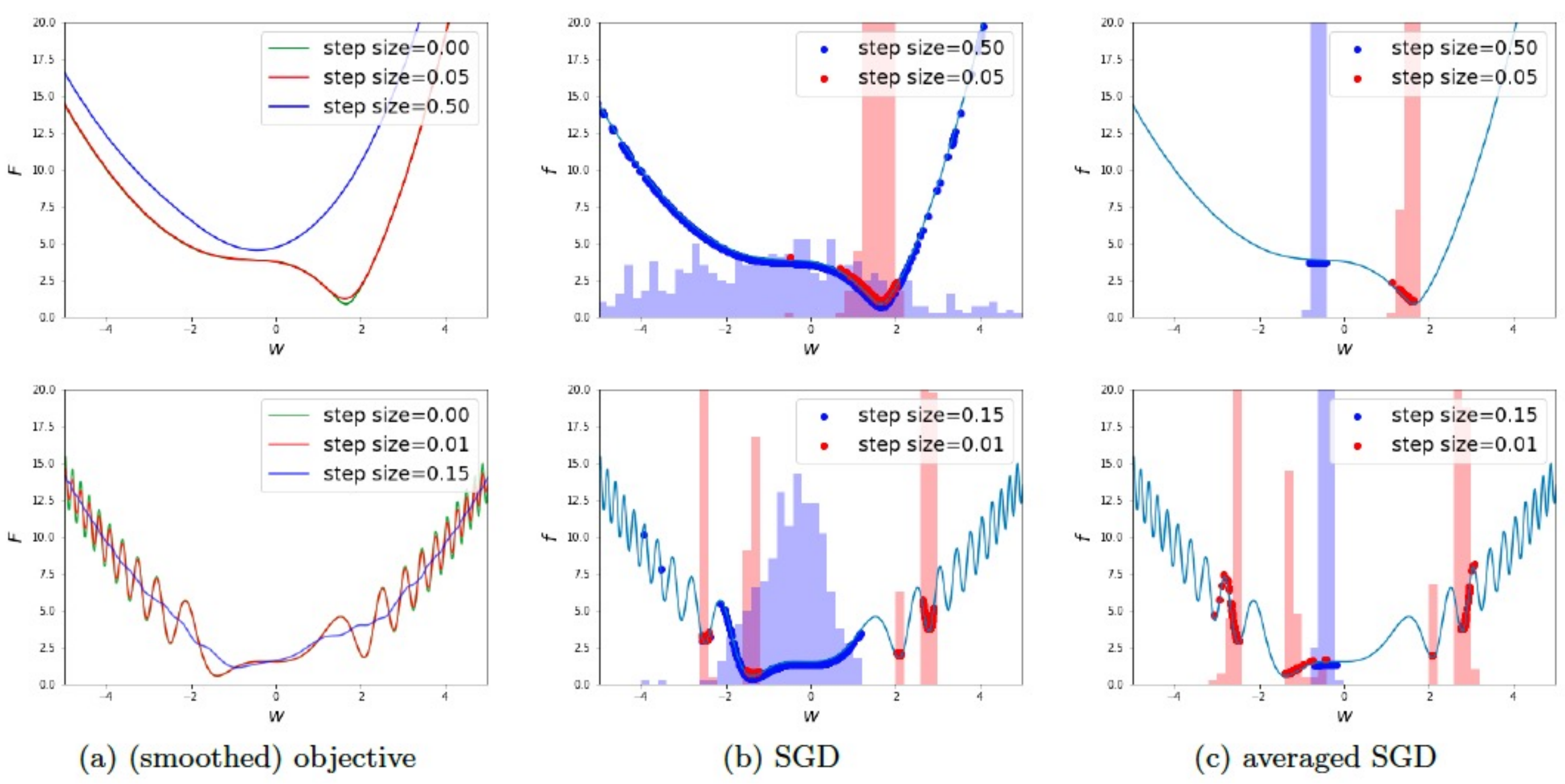


Fig. Alternative view of SGD and ASGD

ASGD is more stable than SGD to optimize the smoothed objective.

3. Main Results

Averaged SGD: Taking average of parameters during training:

$$\bar{v}_T = \frac{1}{T+1}\sum_{t=0}^{T} v_t, \quad \left(\bar{w}_{T+1} = \bar{v}_T + \frac{\eta}{T+1}\sum_{t=0}^{T}\epsilon_{t+1}(w_t) \xrightarrow{p} 0\right).$$

Assumption

(A1) $-L_d \preceq \nabla^2 f(w) \preceq L I_d$.
(A2) $\mathbb{E}[\epsilon_{t+1}(w)] = 0, \quad \mathbb{E}[\|\epsilon_{t+1}(w)\|^2] \leq \sigma_1^2, \quad \mathbb{E}[\|J_{\epsilon_{t+1}}(w)\|] \leq \sigma_2$.
(A3) $\nabla^2 F(v_*) \succeq \mu I_d$,
(A4) $\|\nabla F(v) - \nabla^2 F(v_*)(v - v_*)\| \leq M\|v - v_*\|^2$.
($v_* = \arg\min F(w)$)

Theorem: Running averaged SGD with $\eta \leq \frac{1}{2L}$ under (A1)-(A4), then

$$\lim_{T\to\infty} \mathbb{E}[\|\bar{v}_T - v_*\|] \leq \min\left\{ D_\infty, \frac{4\sigma_1\sigma_2\eta^{\frac{3}{2}}L^{\frac{1}{2}}}{\sqrt{3}\mu} + \frac{MD_\infty^2}{\mu}\right\},$$

$$\left(\text{SGD-error: } D_\infty = \lim_{T\to\infty} D_T = \sqrt{\tfrac{1}{T+1}\textstyle\sum_{t=0}^{T}\|v_t - v_*\|^2}\right)$$

This means nontrivial improvement by parameter averaging over SGD in the sense: $\lim_{T\to\infty}\mathbb{E}[\|\bar{v}_T - v_*\|] \ll D_\infty$ when

$$\frac{4\sigma_1\sigma_2\eta^{\frac{3}{2}}L^{\frac{1}{2}}}{\sqrt{3}\mu} \ll D_\infty \ll \frac{\mu}{M}.$$

Consider the case where μ, M are taken uniformly as $\eta \to 0$.

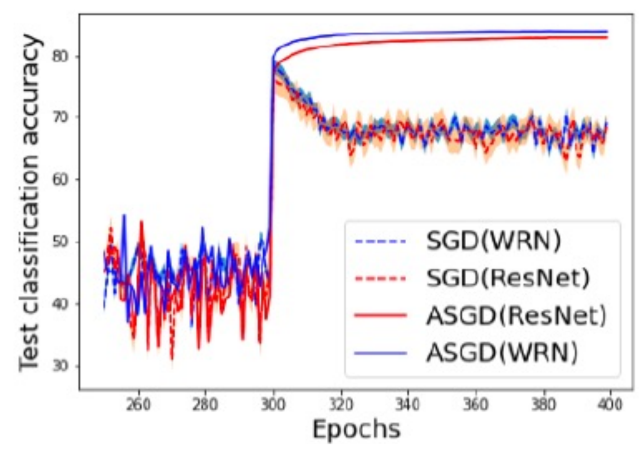
- Lower bound is satisfied for mildly small η because SGD oscillates $D_\infty \gtrsim \eta\sigma_3$ if $\mathbb{E}[\|\epsilon_{t+1}(w)\|^2] \geq \sigma_3^2$ under some conditions.
- Upper bound $D_\infty \ll O(1)$ means convergence to some extent.

Remark: η induces a trade-off between upper and lower bounds.

4. Experiments

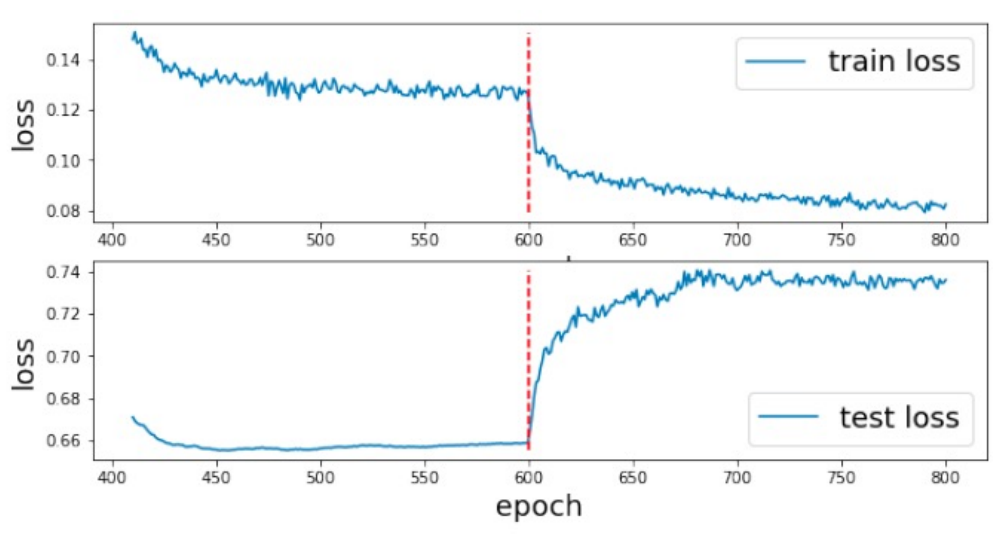
Test accuracies achieved by SGD and ASGD on CIFAR-10/100.

ASGD achieved high accuracy by using relatively large step-sizes.

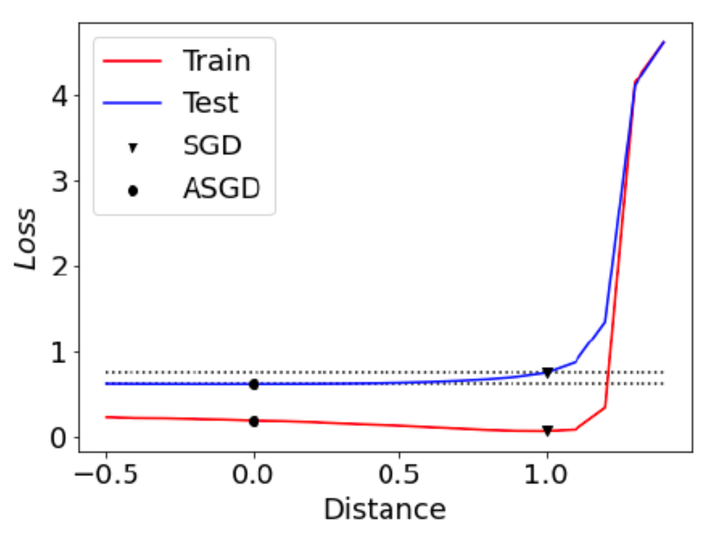


		CIFAR100				CIFAR10		
	η	ResNet-50	WRN-28-10	Pyramid	η	ResNet-50	WRN-28-10	Pyramid
SGD	s	80.83 (0.21)	81.81 (0.29)	81.43 (0.32)	s	95.95 (0.11)	96.85 (0.16)	96.41 (0.22)
Averaged SGD	s	82.13 (0.22)	83.13 (0.13)	84.23 (0.03)	s	96.58 (0.14)	97.24 (0.07)	97.07 (0.08)
	l	**82.87 (0.13)**	**84.23 (0.10)**	**85.12 (0.20)**	m	**96.89 (0.05)**	**97.44 (0.04)**	**97.28 (0.13)**
SAM	s	82.56 (0.14)	83.80 (0.27)	84.59 (0.24)	s	**96.34 (0.12)**	97.14 (0.05)	97.34 (0.03)
Averaged SAM	s	82.64 (0.12)	84.09 (0.30)	85.40 (0.12)	s	96.33 (0.10)	**97.21 (0.05)**	97.34 (0.03)
	l	**82.73 (0.28)**	**84.55 (0.17)**	**86.00 (0.04)**	m	96.31 (0.11)	97.20 (0.06)	**97.35 (0.06)**

A mildly large step size biases the convergent point in the final phase.



We run SGD for 200 epochs from a parameter obtained by the ASGD with 600 epochs. The red line is a change point of the methods. The learning rate for SGD is annealed to 0 from 0.02 used for the final phase of averaged SGD.



Sections of the train and test loss landscapes across the parameters obtained by ASGD and SGD. Losses form asymmetric valleys.