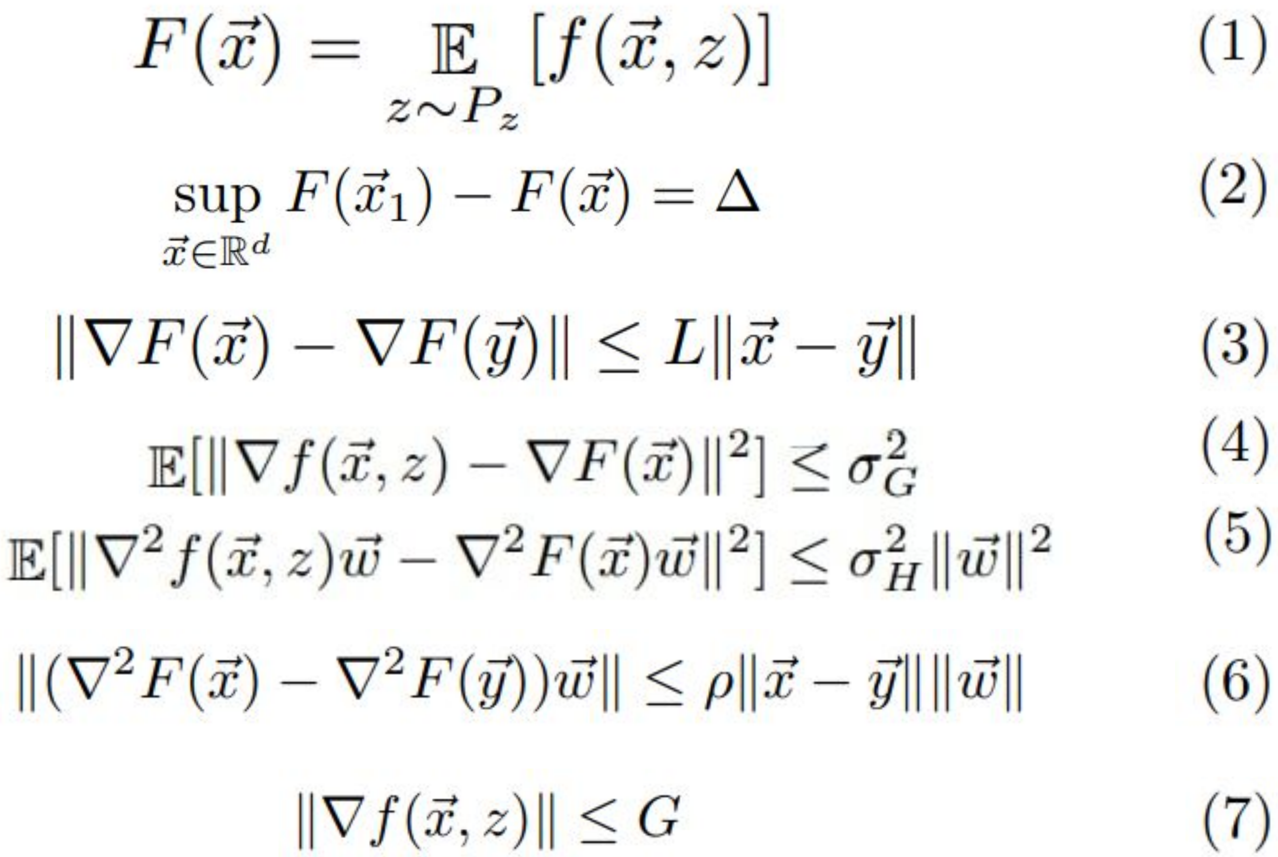
Better SGD using Second-order Momentum

Ruslan Iskhakov
Fyodor Noskov
Aziz Temirkhanov

Setting



$$F(\vec{x}) = \mathop{\mathbb{E}}_{z \sim P_z} [f(\vec{x}, z)] \tag{1}$$

$$\sup_{\vec{x} \in \mathbb{R}^d} F(\vec{x}_1) - F(\vec{x}) = \Delta \tag{2}$$

$$\|\nabla F(\vec{x}) - \nabla F(\vec{y})\| \leq L\|\vec{x} - \vec{y}\| \tag{3}$$

$$\mathbb{E}[\|\nabla f(\vec{x}, z) - \nabla F(\vec{x})\|^2] \leq \sigma_G^2 \tag{4}$$

$$\mathbb{E}[\|\nabla^2 f(\vec{x}, z)\vec{w} - \nabla^2 F(\vec{x})\vec{w}\|^2] \leq \sigma_H^2 \|\vec{w}\|^2 \tag{5}$$

$$\|(\nabla^2 F(\vec{x}) - \nabla^2 F(\vec{y}))\vec{w}\| \leq \rho\|\vec{x} - \vec{y}\| \|\vec{w}\| \tag{6}$$

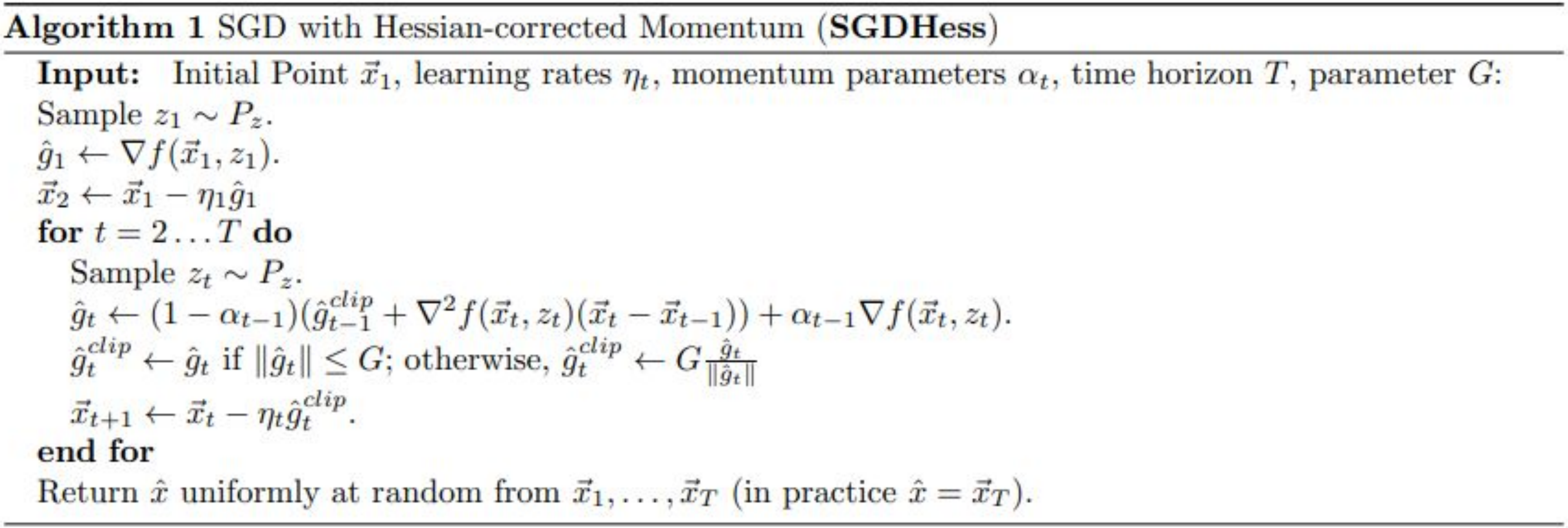
$$\|\nabla f(\vec{x}, z)\| \leq G \tag{7}$$

The standard SGD with momentum update

$$\hat{g}_t = (1 - \alpha)\hat{g}_{t-1} + \alpha \nabla f(\vec{x}_t, z_t)$$

$$\vec{x}_{t+1} = \vec{x}_t - \eta \hat{g}_t$$

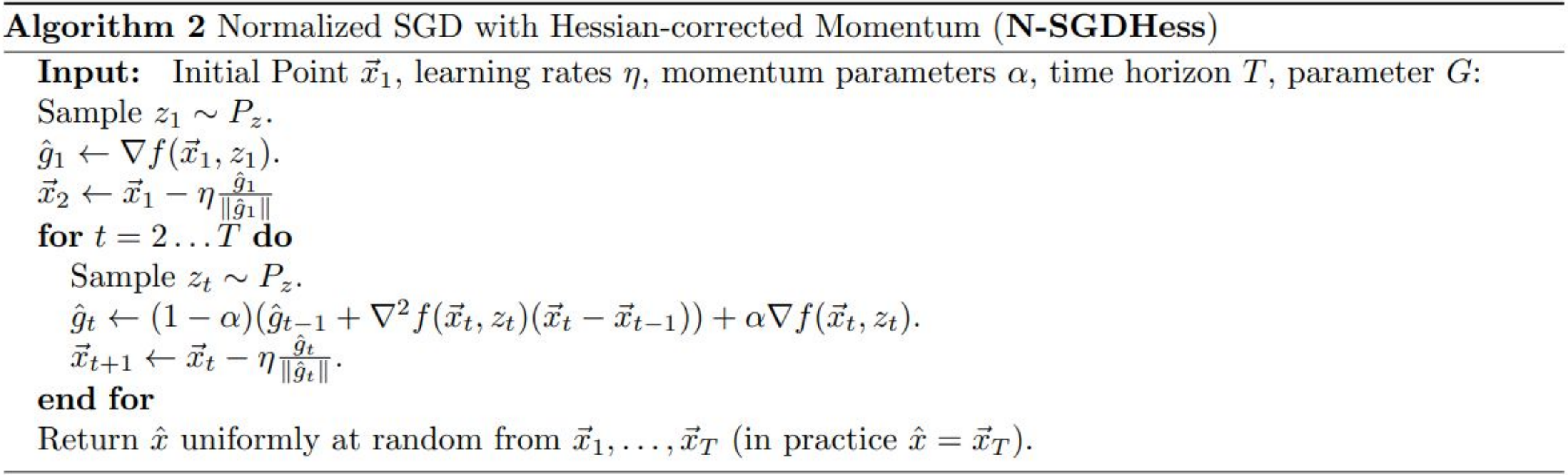
SGD with Hessian-corrected Momentum and clipping



Algorithm 1 SGD with Hessian-corrected Momentum (**SGDHess**)

```
Input: Initial Point $\vec{x}_1$, learning rates $\eta_t$, momentum parameters $\alpha_t$, time horizon $T$, parameter $G$:
Sample $z_1 \sim P_z$.
$\hat{g}_1 \leftarrow \nabla f(\vec{x}_1, z_1)$.
$\vec{x}_2 \leftarrow \vec{x}_1 - \eta_1 \hat{g}_1$
for $t = 2 \dots T$ do
    Sample $z_t \sim P_z$.
    $\hat{g}_t \leftarrow (1 - \alpha_{t-1})(\hat{g}_{t-1}^{clip} + \nabla^2 f(\vec{x}_t, z_t)(\vec{x}_t - \vec{x}_{t-1})) + \alpha_{t-1} \nabla f(\vec{x}_t, z_t)$.
    $\hat{g}_t^{clip} \leftarrow \hat{g}_t$ if $\|\hat{g}_t\| \leq G$; otherwise, $\hat{g}_t^{clip} \leftarrow G \frac{\hat{g}_t}{\|\hat{g}_t\|}$
    $\vec{x}_{t+1} \leftarrow \vec{x}_t - \eta_t \hat{g}_t^{clip}$.
end for
Return $\hat{x}$ uniformly at random from $\vec{x}_1, \dots, \vec{x}_T$ (in practice $\hat{x} = \vec{x}_T$).
```

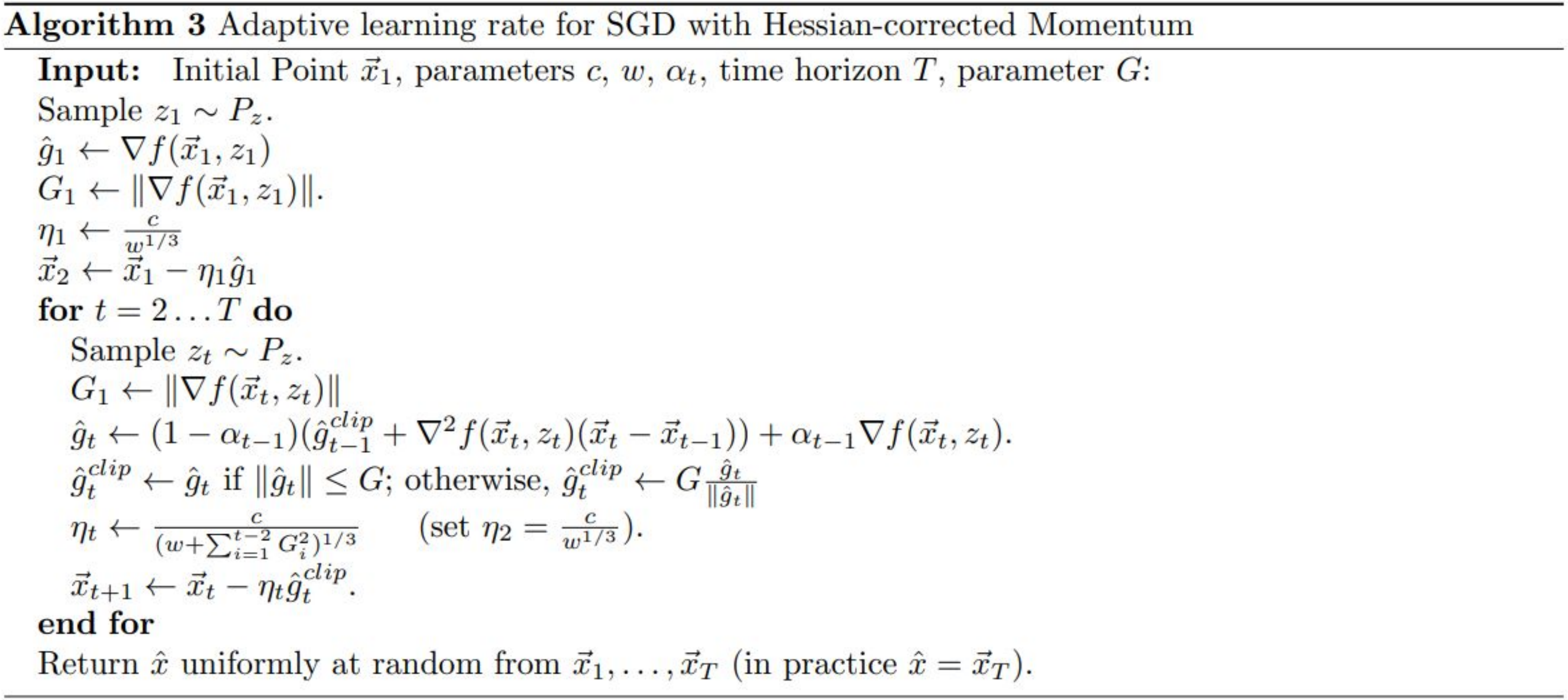
Normalized SGD with Hessian-corrected momentum



Algorithm 2 Normalized SGD with Hessian-corrected Momentum (**N-SGDHess**)

```
Input: Initial Point $\vec{x}_1$, learning rates $\eta$, momentum parameters $\alpha$, time horizon $T$, parameter $G$:
Sample $z_1 \sim P_z$.
$\hat{g}_1 \leftarrow \nabla f(\vec{x}_1, z_1)$.
$\vec{x}_2 \leftarrow \vec{x}_1 - \eta \frac{\hat{g}_1}{\|\hat{g}_1\|}$
for $t = 2 \dots T$ do
    Sample $z_t \sim P_z$.
    $\hat{g}_t \leftarrow (1-\alpha)(\hat{g}_{t-1} + \nabla^2 f(\vec{x}_t, z_t)(\vec{x}_t - \vec{x}_{t-1})) + \alpha \nabla f(\vec{x}_t, z_t)$.
    $\vec{x}_{t+1} \leftarrow \vec{x}_t - \eta \frac{\hat{g}_t}{\|\hat{g}_t\|}$.
end for
Return $\hat{x}$ uniformly at random from $\vec{x}_1, \dots, \vec{x}_T$ (in practice $\hat{x} = \vec{x}_T$).
```

Adaptive SGD with Hessian-corrected Momentum

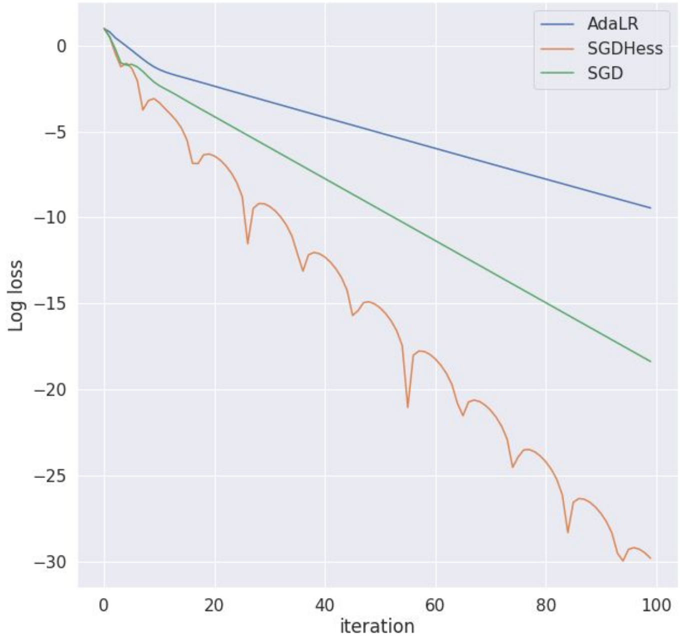


Algorithm 3 Adaptive learning rate for SGD with Hessian-corrected Momentum

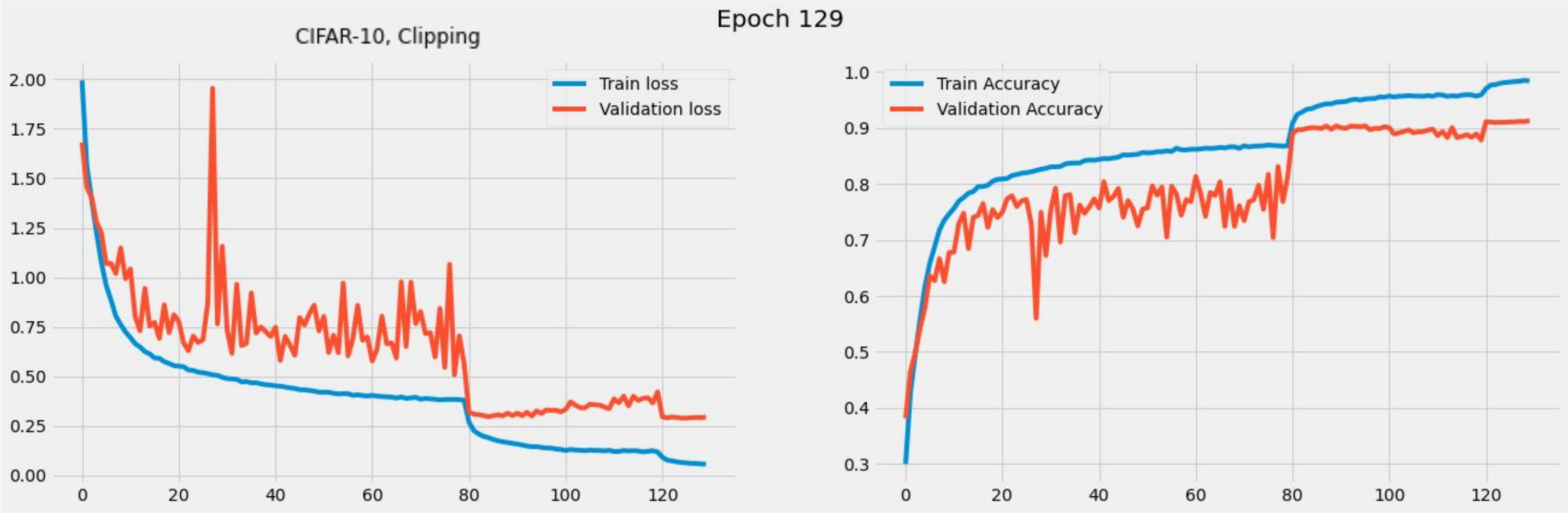
```
**Input:** Initial Point $\vec{x}_1$, parameters $c$, $w$, $\alpha_t$, time horizon $T$, parameter $G$:
Sample $z_1 \sim P_z$.
$\hat{g}_1 \leftarrow \nabla f(\vec{x}_1, z_1)$
$G_1 \leftarrow \|\nabla f(\vec{x}_1, z_1)\|$.
$\eta_1 \leftarrow \frac{c}{w^{1/3}}$
$\vec{x}_2 \leftarrow \vec{x}_1 - \eta_1 \hat{g}_1$
**for** $t = 2 \dots T$ **do**
    Sample $z_t \sim P_z$.
    $G_1 \leftarrow \|\nabla f(\vec{x}_t, z_t)\|$
    $\hat{g}_t \leftarrow (1 - \alpha_{t-1})(\hat{g}_{t-1}^{clip} + \nabla^2 f(\vec{x}_t, z_t)(\vec{x}_t - \vec{x}_{t-1})) + \alpha_{t-1} \nabla f(\vec{x}_t, z_t)$.
    $\hat{g}_t^{clip} \leftarrow \hat{g}_t$ if $\|\hat{g}_t\| \leq G$; otherwise, $\hat{g}_t^{clip} \leftarrow G \frac{\hat{g}_t}{\|\hat{g}_t\|}$
    $\eta_t \leftarrow \frac{c}{(w + \sum_{i=1}^{t-2} G_i^2)^{1/3}}$        (set $\eta_2 = \frac{c}{w^{1/3}}$).
    $\vec{x}_{t+1} \leftarrow \vec{x}_t - \eta_t \hat{g}_t^{clip}$.
**end for**
Return $\hat{x}$ uniformly at random from $\vec{x}_1, \dots, \vec{x}_T$ (in practice $\hat{x} = \vec{x}_T$).
```

Different algorithms on square function

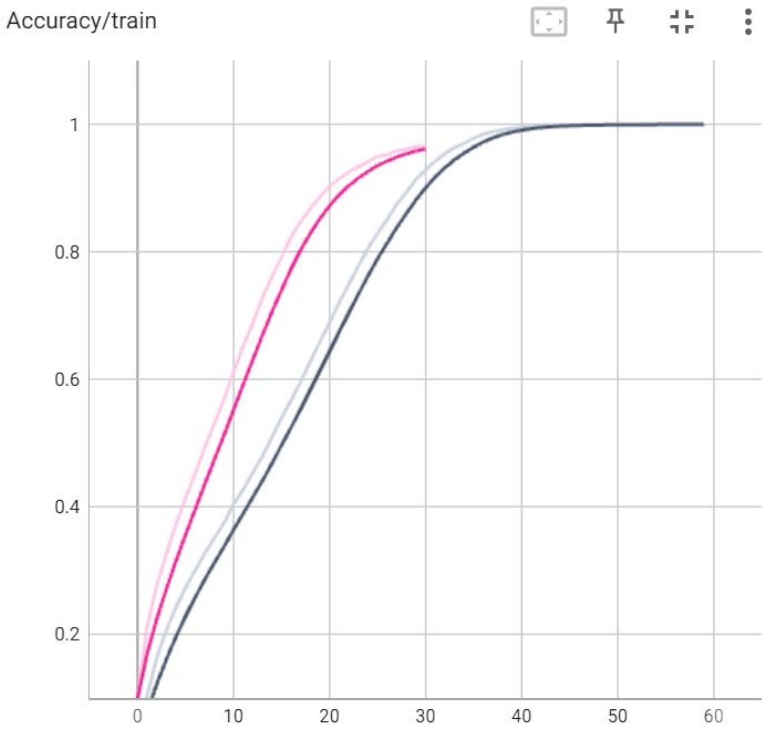
SGDHess after 100 iterations	1e-30
SGD after 100 iterations	1e-19
AdaHess after 100 iterations	1e-10

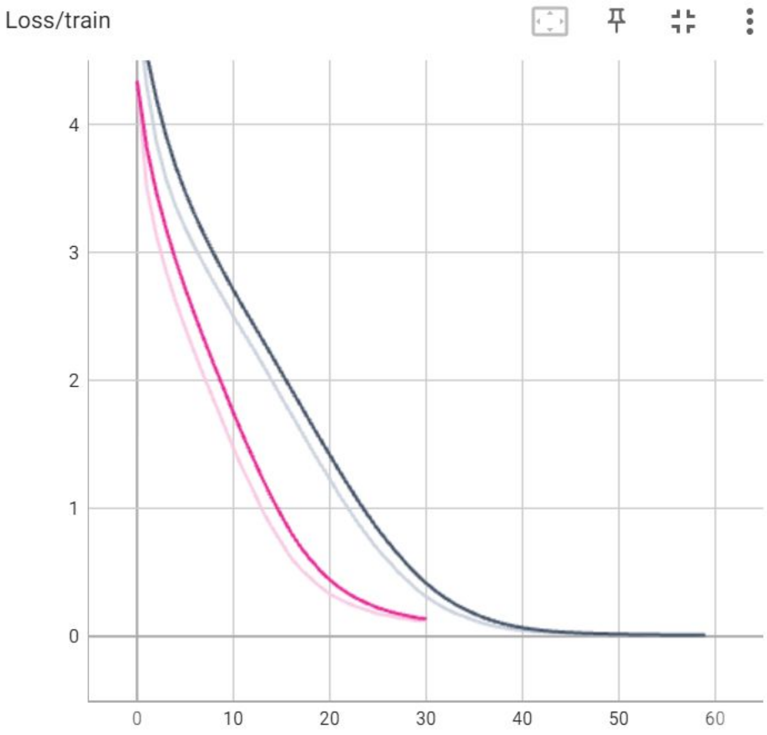


Experiments with Alg 1 (Clipping)

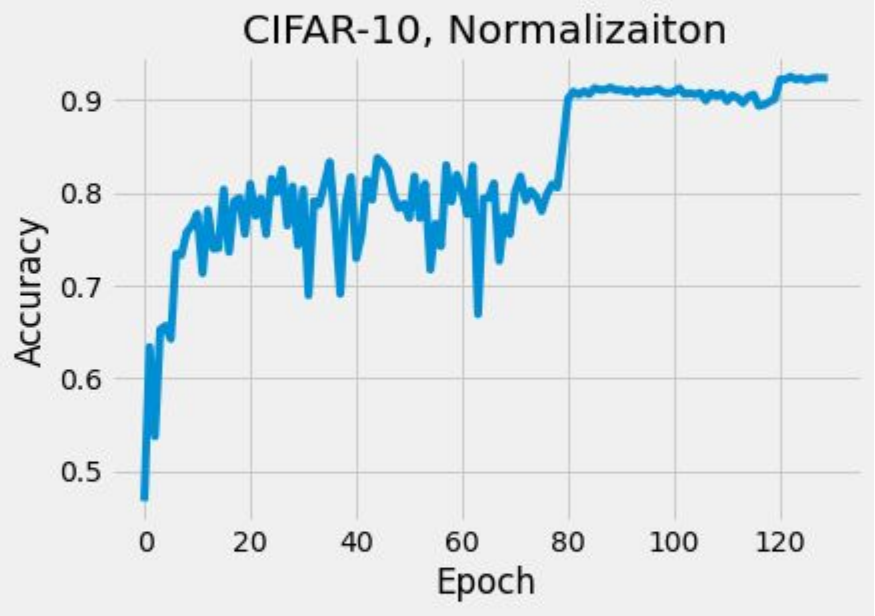


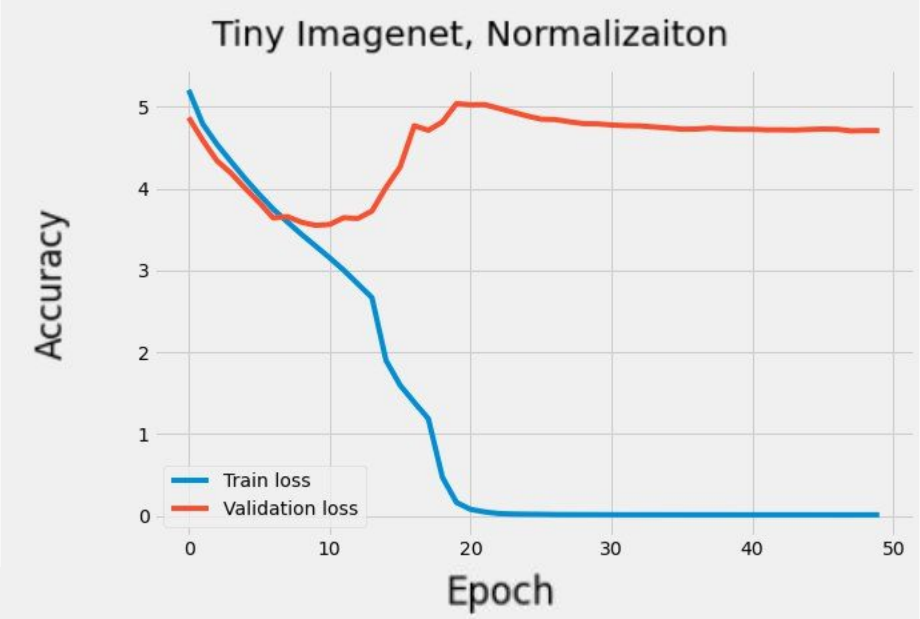
Tiny ImageNet





Experiments with Alg 2 (Normalization)

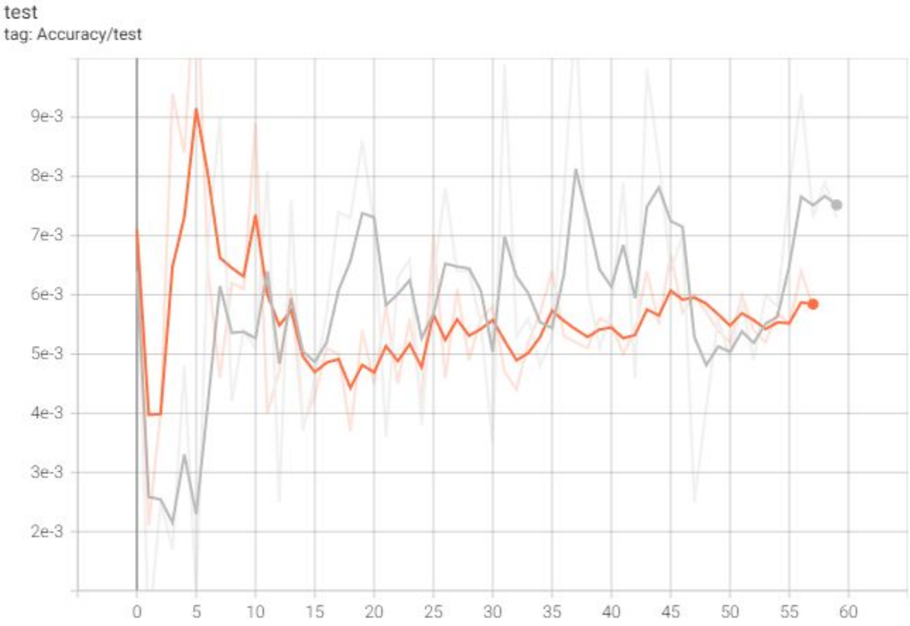


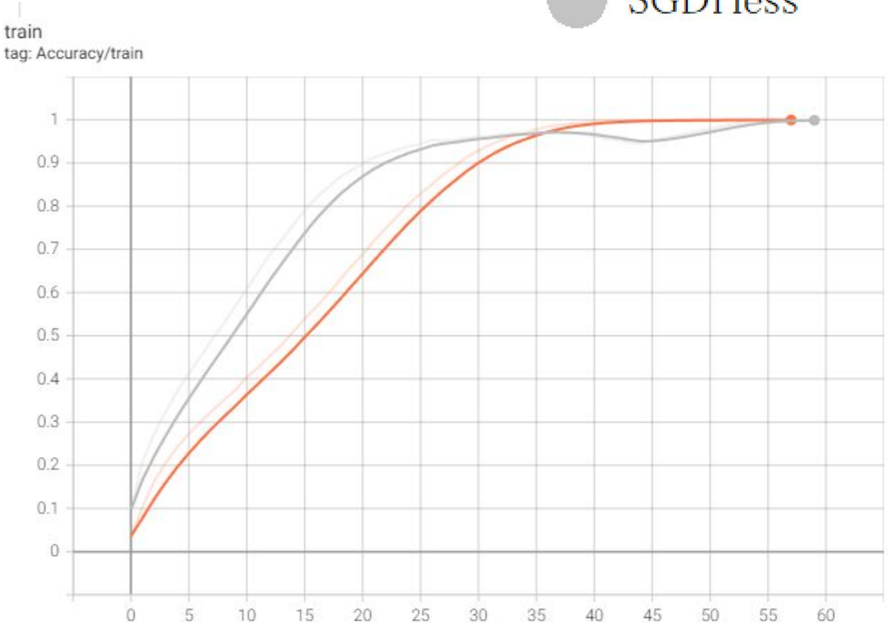


Experiments with Alg 2 (Normalization)

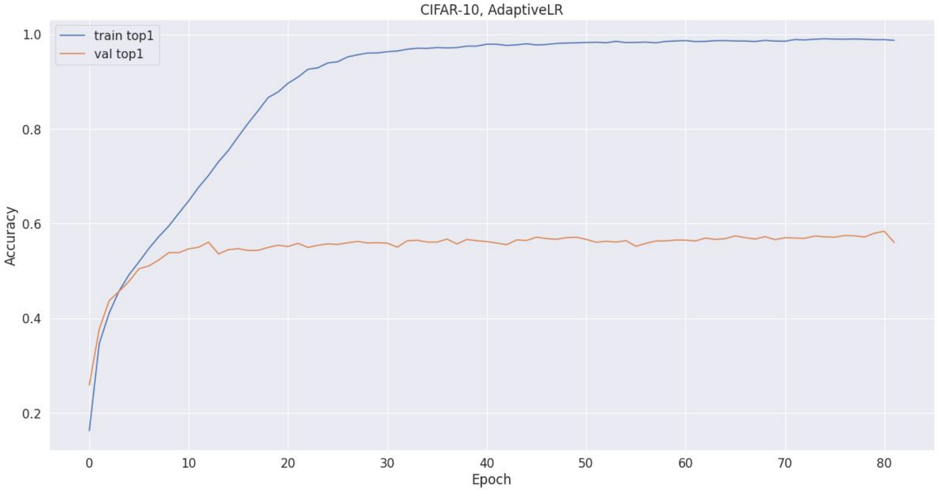
● SGD

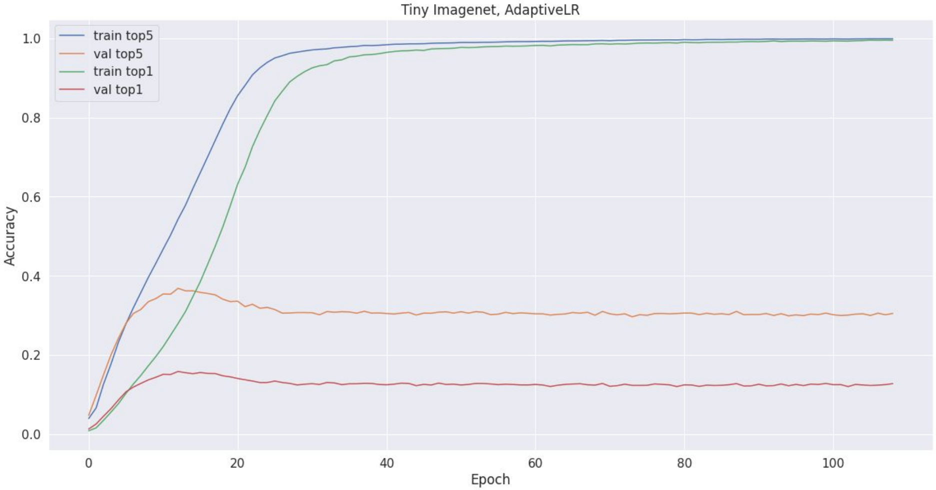
● SGDHess



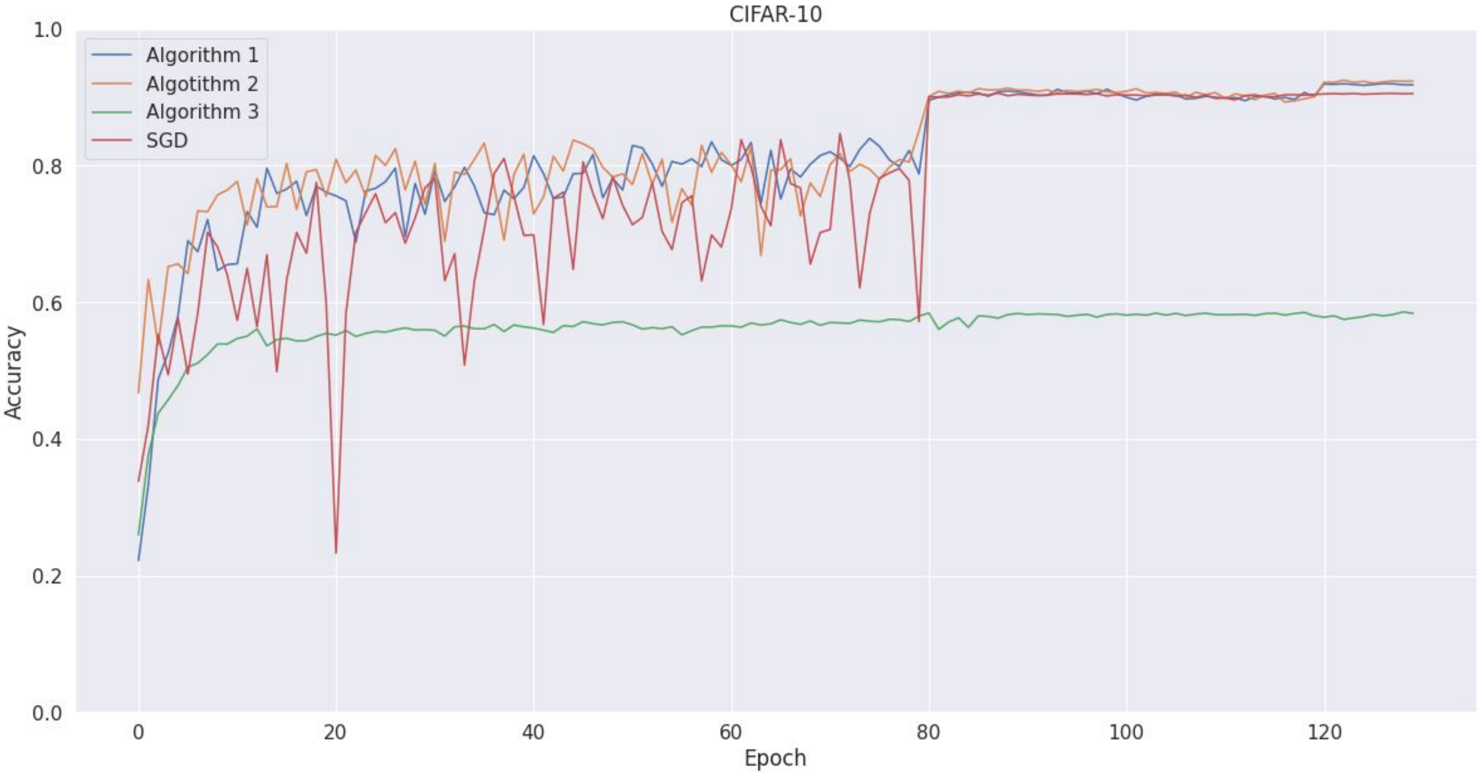


Experiments with Alg 3 (Adaptive LR)

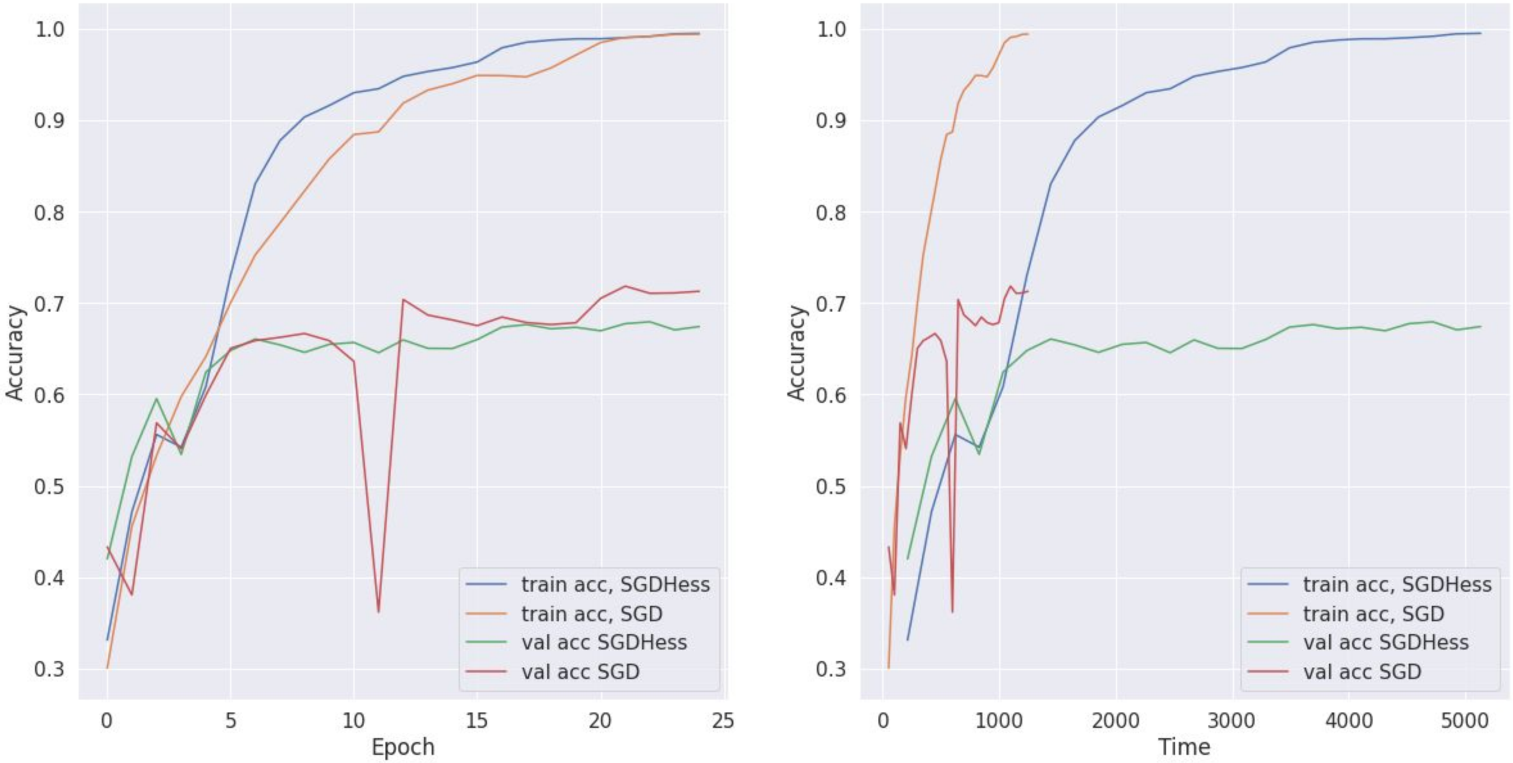

CIFAR-10, AdaptiveLR


Tiny Imagenet, AdaptiveLR

Comparison of algorithms



Time, resnet20 on CIFAR-10



Thank you for your attention!