HW 1 released, template

Notes on approximation released

Clarke Differential

Clarke Differential

Definition: Given $f : \mathbb{R}^d \to \mathbb{R}$, for every x, the Clark differential is defined as
$\partial f(x) \triangleq \text{conv}\left(\{s \in \mathbb{R}^d : \exists \{x_i\}_{i=1}^{\infty} \to x, \{\nabla f(x_i)\}_{i=1}^{\infty} \to s\}\right)$.
The elements in the subdifferential set are subgradients.

When does Clarke differential exists

Definition (Locally Lipschitz): $f : \mathbb{R}^d \to \mathbb{R}$ is locally Lipchitz if $\forall x \in \mathbb{R}^d$, there exists a neighborhood S of x, such that f is Lipchitz in S.

$\Rightarrow$ Clarke differential exists

Positive Homogeneity $\rightarrow$ motivate ReLU

Definition: $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is positive homogeneous of degree L if $f(\alpha x) = \alpha^L f(x)$ for any $\underbrace{\alpha \geq 0}_{\text{scalar}}$.

① ReLU : $\sigma(\alpha z) = \alpha \cdot \sigma(z)$

② monomials of degree L : $\prod_{i=1}^{d} x_i^{p_i}$, $\sum_{i=1}^{d} p_i = L$

$$\prod_{i=1}^{d} (\alpha x_i)^{p_i} = \alpha^{\sum_{i=1}^{d} p_i} \prod_{i=1}^{d} x_i^{p_i}$$

$$= \alpha^L \cdot \prod_{i=1}^{d} x_i^{p_i}$$

③ Norm : $\|\alpha x\| = \alpha \cdot \|x\|$

Positive Homogeneity

(4) Multi-layer ReLU

$$f(x, W_1, \dots, W_{H+1}) = W_{H+1}\,\sigma(W_H \cdots \sigma(W_1 x) \cdots)$$

for one-layer

$$f(x, W_1, \dots, \alpha W_h, \dots, W_{H+1}) = \alpha W_{H+1}\,\sigma(W_H \cdots \sigma(W_1 x) \cdots)$$

for all-layer

$$f(x, \alpha W_1, \dots, \alpha W_{H+1}) = \alpha^{H+1} f(W_H \cdots \sigma(W_1 x) \cdots)$$

$\Rightarrow$ $(H+1)$ - homogeneous function say $W_h \in \mathbb{R}^{m \times m}$

Positive Homogeneity

- independent of h /
- hold for ReLU

Fact: $\forall h = 1, \ldots, H+1$

$$\left\langle W_h, \frac{\partial f(x, W_1, \ldots, W_{H+1})}{\partial W_h} \right\rangle = \underline{f(x, W_1, \ldots, W_{H+1})}$$

A, B matrix: $\langle A, B \rangle = \sum_{ij} A_{ij} B_{ij}$

Pf: $A_h = \text{diag}(\sigma'(W_h \sigma(\cdots \sigma(W_1 x) \cdots)) \in \mathbb{R}^{m \times m}$

$(\sigma' = 0$ or $1) \Rightarrow$ pattern whether activation is on or off

$\underline{\sigma(z) = z \cdot \sigma'(z)}$: holds for ReLU

$$f(x, W_1, \ldots W_{H+1}) = W_{H+1} A_H W_H \cdots A_1 W_1 x$$

$$\frac{\partial f}{\partial W_h} = (W_{H+1} A_H \cdots W_{h+1} A_h)^T (A_{h-1} W_{h-1} \cdots W_1 x)^T$$

verify $\left\langle W_h, \frac{\partial f}{\partial W_h} \right\rangle = f(x, W_1, \ldots, W_{H+1})$

Positive Homogeneity and Clark Differential

Lemma: Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is Locally Lipschitz and L -positively homogeneous. For any $x \in \mathbb{R}^d$ and $s \in \partial f(x)$, we have $\langle s, x \rangle = Lf(x)$.

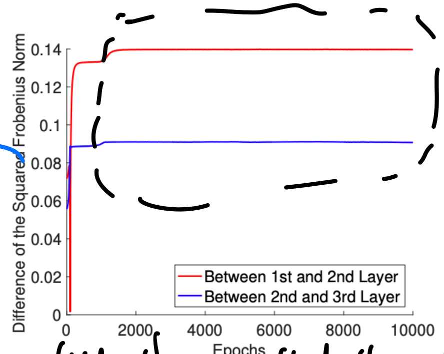
(HW 1)

Norm Preservation

$$f(x, W_1, W_2, W_3) = W_3\,\sigma(W_2\,\sigma(W_1 x))$$

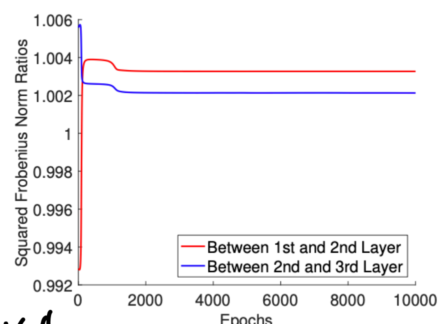
quadratic loss

$\|W_1\|_F \ll \|W_2\|_F \ll \|W_3\|_F^2$

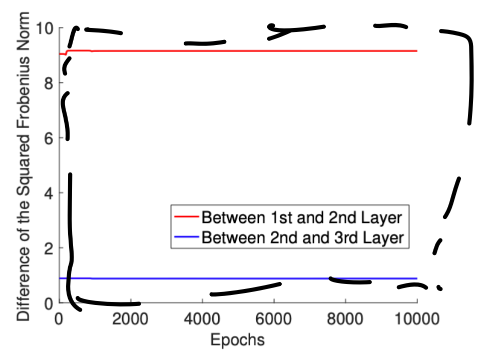


$\|W_1\|_F \approx \|W_2\|_F \approx \|W_3\|_F$

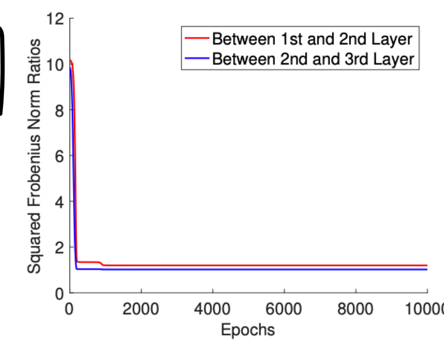
(a) Balanced initialization, squared norm differences.



(b) Balanced initialization, squared norm ratios.



(c) Unbalanced Initialization, squared norm differences.



(d) Unbalanced initialization, squared norm ratios.

— : $\|W_1\|_F^2 - \|W_2\|_F^2$

— : $\|W_2\|_F^2 - \|W_3\|_F^2$

Given A matrix, $\|A\|_F^2 = \sum_{i,j} A_{ij}^2$

Gradient flow and gradient inclusion

widely used in ML / opt theory

Discrete-time dynamics can be complex. Let's use continuous-time dynamics to simplify:

$$\text{Gradient flow: } x_{t+1} = x_t - \eta \nabla f(x_t) \Rightarrow \frac{dx(t)}{dt} = -\nabla f(x(t))$$

$$\text{Gradient inclusion: } \frac{dx(t)}{dt} \in -\partial f(x(t))$$

$$\frac{x_{t+1} - x_t}{\eta} = -\nabla f(x_t) \Longrightarrow \frac{dx(t)}{dt}$$

let $\eta \to 0$

$f(x, W_1, W_2, W_3) = W_3\,\sigma(W_2\,\sigma(W_1 x))$, $W_3 \times 10$, $W_2 / 10$

$W_2(t+1) = W_2(t) - \eta \frac{\partial f}{\partial W_2}$

Norm preservation by gradient inclusion

No assumption on loss

Theorem (Du, Hu, Lee '18) Suppose $\alpha > 0$, $f(x; (W_{H+1}, \ldots, \alpha W_i, \ldots, W_1)) = \alpha f(x, (W_{H+1}, \ldots, W_1))$, I.e., predictions are 1-homogeneous in each layer. Then for every pair of layers $(i, j) \in [H+1] \times [H+1]$, the gradient inclusion maintains: for all $t \geq 0$,

$$\frac{1}{2}\|W_i(t)\|_F^2 - \frac{1}{2}\|W_i(0)\|_F^2 = \frac{1}{2}\|W_j(t)\|_F^2 - \frac{1}{2}\|W_j(0)\|_F^2.$$

same for all layer

· if $\|W_i(0)\|_F^2$ small for all i

$\Rightarrow \|W_i(t)\|_F^2 \approx \|W_j(t)\|_F^2 \Rightarrow$ balance

Pf sketch: 1) $\frac{dW_i(t)}{dt}$ formula

2) $\frac{1}{2}\|W_i(t)\|_F^2 - \frac{1}{2}\|W_i(0)\|_F^2 = \int_0^t \frac{d}{d\tau}\frac{1}{2}\|W_i(\tau)\|_F^2\, d\tau$

$\Rightarrow$ independent of i

Optimization Methods for Deep Learning

Gradient descent for non-convex optimization

$\|A\|_2$: operator norm, largest absolute eigenvalue

Decsent Lemma: Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable, and $\|\nabla^2 f\|_2 \leq \beta$. Then setting the learning rate $\eta = 1/\beta$, and applying gradient descent, $x_{t+1} = x_t - \eta \nabla f(x_t)$, we have:

$f(x_t) \downarrow$

$$f(x_t) - f(x_{t+1}) \geq \frac{1}{2\beta} \|\nabla f(x_t)\|_2^2.$$

$f(x_t) \uparrow$

$\Rightarrow$ learning rate too large

Pf: by Taylor expansion & mean-value Theorem

$f(x+\Delta) = f(x) + \Delta^T \nabla f(x) + \frac{1}{2} \Delta^T \nabla^2 f(y) \Delta$ for some y

$\Delta^T \nabla^2 f(y) \Delta \leq \|\nabla^2 f(y)\|_2 \cdot \|\Delta\|_2^2 \leq \beta \|\Delta\|_2^2$

let $\Delta = -\eta \nabla f(x_t)$

$f(x_{t+1}) \leq f(x_t) - \eta \|\nabla f(x_t)\|_2^2 + \frac{1}{2} \beta \cdot \eta^2 \|\nabla f(x_t)\|_2^2$

$= f(x_t) - \frac{1}{2} \eta \|\nabla f(x_t)\|_2^2$

Converging to stationary points

approximate stationary point

Theorem: In $T = O(\frac{\beta}{\epsilon^2})$ iterations, we have $\|\nabla f(x)\|_2 \le \epsilon$.

Pf: $f(x_{t+1}) \le f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2$

Sum over $t = 0, \dots, T-1$

$$\sum_{t=1}^{T} f(x_t) \le \sum_{t=0}^{T-1} f(x_t) - \frac{\eta}{2}\sum_{t=0}^{T-1}\|\nabla f(x_t)\|_2^2$$

$$\Rightarrow \quad f(x_T) \le f(x_0) - \frac{\eta}{2}\sum_{t=0}^{T-1}\|\nabla f(x_t)\|_2^2$$

$$\Rightarrow \quad \frac{\eta}{2}\sum_{t=0}^{T-1}\|\nabla f(x_t)\|_2^2 \le f(x_0) - f(x_T)$$

$$\frac{\eta}{2} T \cdot \min_{0\le t\le T-1}\|\nabla f(x_t)\|_2^2 \le f(x_0) - \min_x f(x)$$

$$\Rightarrow \quad \min_{0\le t\le T-1}\|\nabla f(x_t)\|_2 \le \sqrt{\frac{2\beta(f(x_0) - \min_x f(x))}{(T-1)}} = \epsilon$$

scale T

Gradient Descent for Quadratic Functions

optimal $x=0$ $\lambda_{min}(A) > 0$

Problem: $\min_x \frac{1}{2}x^\top Ax$ with $A \in \mathbb{R}^{d\times d}$ being positive-definite.

Theorem: Let $\lambda_{\max}$ and $\lambda_{\min}$ be the largest and the smallest eigenvalues of A. If we set $\eta \leq \frac{1}{\lambda_{\max}}$, we have

$$\|x_t\|_2 \leq \left(1-\eta\lambda_{\min}\right)^t \|x_0\|_2$$

$$\begin{aligned}\|x_{t+1}\|_2 &= \|x_t - \eta A x_t\|_2 \\ &= \|(I-\eta A)x_t\|_2 \\ &\leq \|I-\eta A\|_2 \|x_t\|_2 \\ &\leq (1-\eta\lambda_{min})\|x_t\|_2 \\ &\leq (1-\eta\lambda_{min})^{t+1}\|x_0\|_2\end{aligned}$$

To make $\|x_t\|_2 \leq \varepsilon$

when $\eta = \frac{1}{\lambda_{max}}$

need $O\left(\frac{\lambda_{max}}{\lambda_{min}}\log\left(\frac{1}{\varepsilon}\right)\right)$ iters

$\kappa := \frac{\lambda_{max}}{\lambda_{min}}$ condition number

Momentum: Heavy-Ball Method (Polyak '64)

$\beta < 1$

doesn't work for general convex function

Problem: $\min_x f(x)$

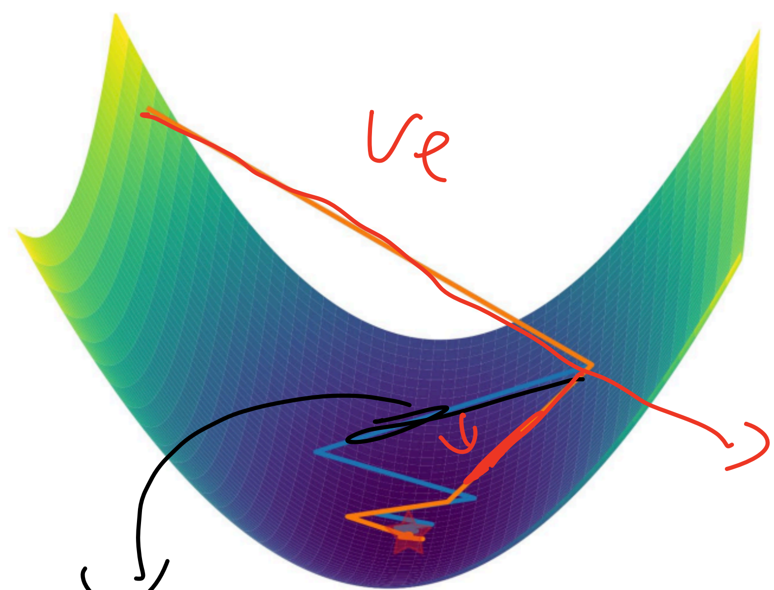
Method: $v_{t+1} = -\nabla f(x_t) + \beta v_t$

$$x_{t+1} = x_t + \eta v_{t+1}$$

For quadratic optimization provable improvement

$$O\left(\sqrt{\kappa} \cdot \log\left(\frac{1}{\varepsilon}\right)\right)$$

vs. $O\left(\kappa \cdot \log\left(\frac{1}{\varepsilon}\right)\right)$



$-\nabla f(x_t)$

Momentum: Nesterov Acceleration (Nesterov '89)

κ: $\frac{\text{smoothness}}{\text{S.C. parameter}}$

Problem: $\min_x f(x)$

lookahead

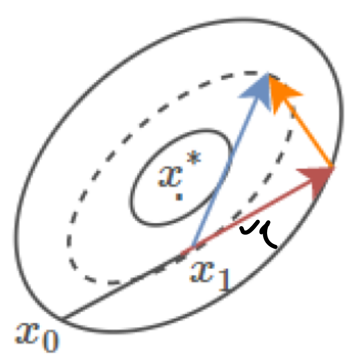
Method: $v_{t+1} = -\nabla f(x_t + \beta v_t) + \beta v_t$

$$x_{t+1} = x_t + \eta v_{t+1}$$

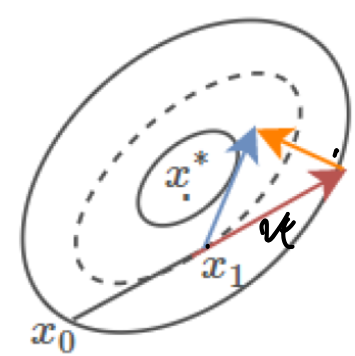
For general strongly convex function $O(\sqrt{\kappa}\log(\frac{1}{\epsilon}))$

· Continuous approximation

Polyak's Momentum



Nesterov Momentum



Newton's Method

2^{nd} order method

Newton's Method: $x_{t+1} = x_t - \eta(\underbrace{\nabla^2 f(x_t)})^{-1} \nabla f(x_t)$

- GD: $x_{t+1} = x_t - \eta \nabla f(x_t)$

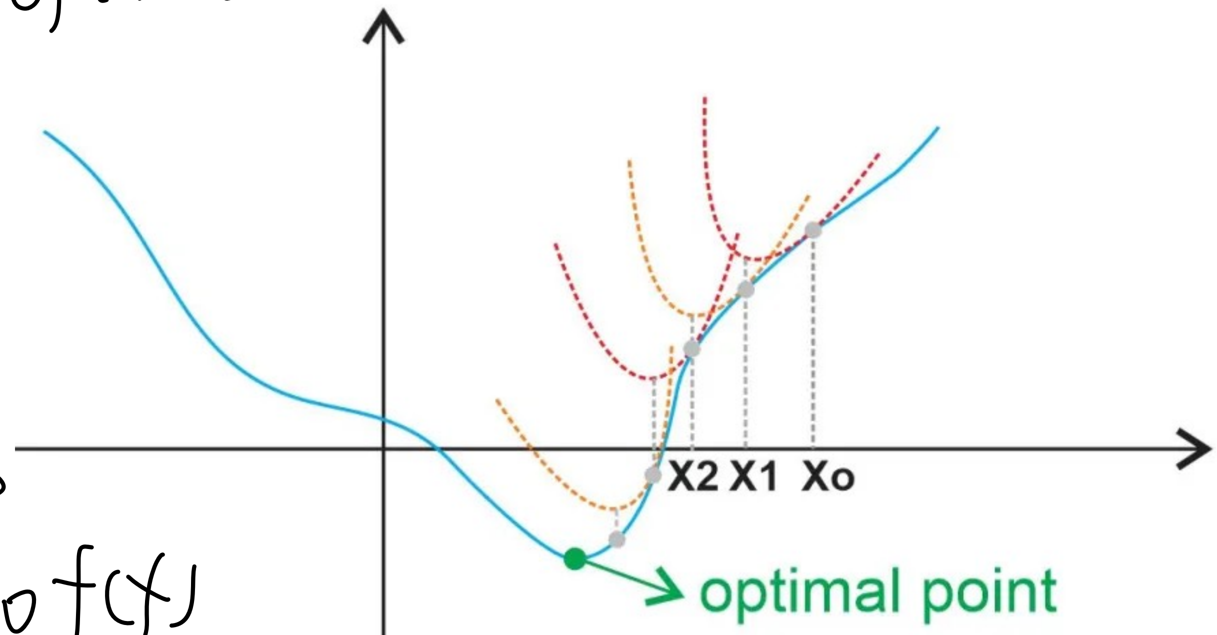
$\Leftrightarrow f(x+\Delta) \approx f(x) + \Delta^T \nabla f(x) + \frac{1}{2}\|\Delta\|_2^2$

$\Rightarrow \Delta = -\nabla f(x)$

- Newton

$f(x+\Delta) \approx f(x) + \Delta^T f(x) + \frac{1}{2}\Delta^T \nabla^2 f(x)\Delta$

$\Rightarrow \Delta = -(\nabla^2 f(x))^{-1} \nabla f(x)$



$O(\log\log(\frac{1}{\varepsilon}))$, no dependence K

- Problem: invert $\nabla^2 f(x) \Rightarrow O(d^3)$

AdaGrad (Duchi et al. '11)

Newton Method: $x_{t+1} = x_t - \eta(\nabla^2 f(x_t))^{-1} \nabla f(x_t)$

AdaGrad: separate learning rate for every parameter

pre-conditioner, G_t: diagonal

$$x_{t+1} = x_t - \eta\underbrace{(G_{t+1} + \epsilon I)^{-1}}\nabla f(x_t),\ (G_t)_{ii} = \sqrt{\sum_{j=1}^{t-1} \left(\nabla f(x_t)_i\right)^2}$$

- dynamic learning rate for each coordinate
- Default value works well

$\eta = 0.01$, $\varepsilon = 10^{-8}$

RMSProp (Hinton et al. ’12)

AdaGrad: separate learning rate for every parameter

$$x_{t+1} = x_t - \eta(G_{t+1} + \epsilon I)^{-1} \nabla f(x_t), \; (G_t)_{ii} = \sqrt{\sum_{j=1}^{t-1} \left(\nabla f(x_t)_i \right)^2}$$

RMSProp: exponential weighting of gradient norms

$$x_{t+1} = x_t - \eta(G_{t+1} + \epsilon I)^{-1/2} \nabla f(x_t),$$

$$(G_{t+1})_{ii} = \underbrace{\beta(G_t)_{ii} + (1-\beta)(\nabla f(x_t)_i)^2}_{\text{exponential average}}$$

$0 < \beta < 1$

AdaDelta (Zeiler '12)

RMSProp:

$$x_{t+1} = x_t - \eta(G_{t+1} + \epsilon I)^{-1/2}\nabla f(x_t),$$
$$(G_{t+1})_{ii} = \beta(G_t)_{ii} + (1-\beta)(\nabla f(x_t)_i)^2$$

$$\frac{\partial f}{\partial x} \Big/ \left(\frac{\partial^2 f}{\partial x^2}\right)^{1/2}$$

unit less

AdaDelta:

$$x_{t+1} = x_t - \eta \Delta x_t,$$
$$\Delta x_t = \sqrt{u_t + \epsilon} \cdot (G_{t+1} + \epsilon I)^{-1/2}\nabla f(x_t)$$
$$(G_{t+1})_{ii} = \rho(G_t)_{ii} + (1-\rho)(\nabla f(x_t)_i)^2,$$
$$u_{t+1} = \rho u_t + (1-\rho)\|\Delta x_t\|_2^2$$

unit observation

Newton: $\Delta \propto \frac{\partial}{\partial x} \Big/ \frac{\partial^2 f}{\partial x^2} \propto$ unit of x

$\Rightarrow$ approximate Newton

Adam (Kingma & Ba '14)

Momentum:

$$v_{t+1} = -\nabla f(x_t) + \beta v_t, \; x_{t+1} = x_t + \eta v_{t+1}$$

RMSProp: exponential weighting of gradient norms

$$x_{t+1} = x_t - \eta (G_{t+1} + \epsilon I)^{-1} \nabla f(x_t),$$
$$(G_t)_{ii} = \beta (G_t)_{ii} + (1-\beta)(\nabla f(x_t)_i)^2$$

Adam

$$v_{t+1} = \beta_1 v_t + (1-\beta_1) \nabla f(x_t)$$

momentum

$$(G_{t+1})_{ii} = \beta_2 (G_t)_{ii} + (1-\beta_2)(\nabla f(x_t)_i)^2$$
$$x_{t+1} = x_t - \eta (G_{t+1} + \epsilon I)^{-1/2} v_{t+1}$$

Default choice nowadays.

Are these actually useful

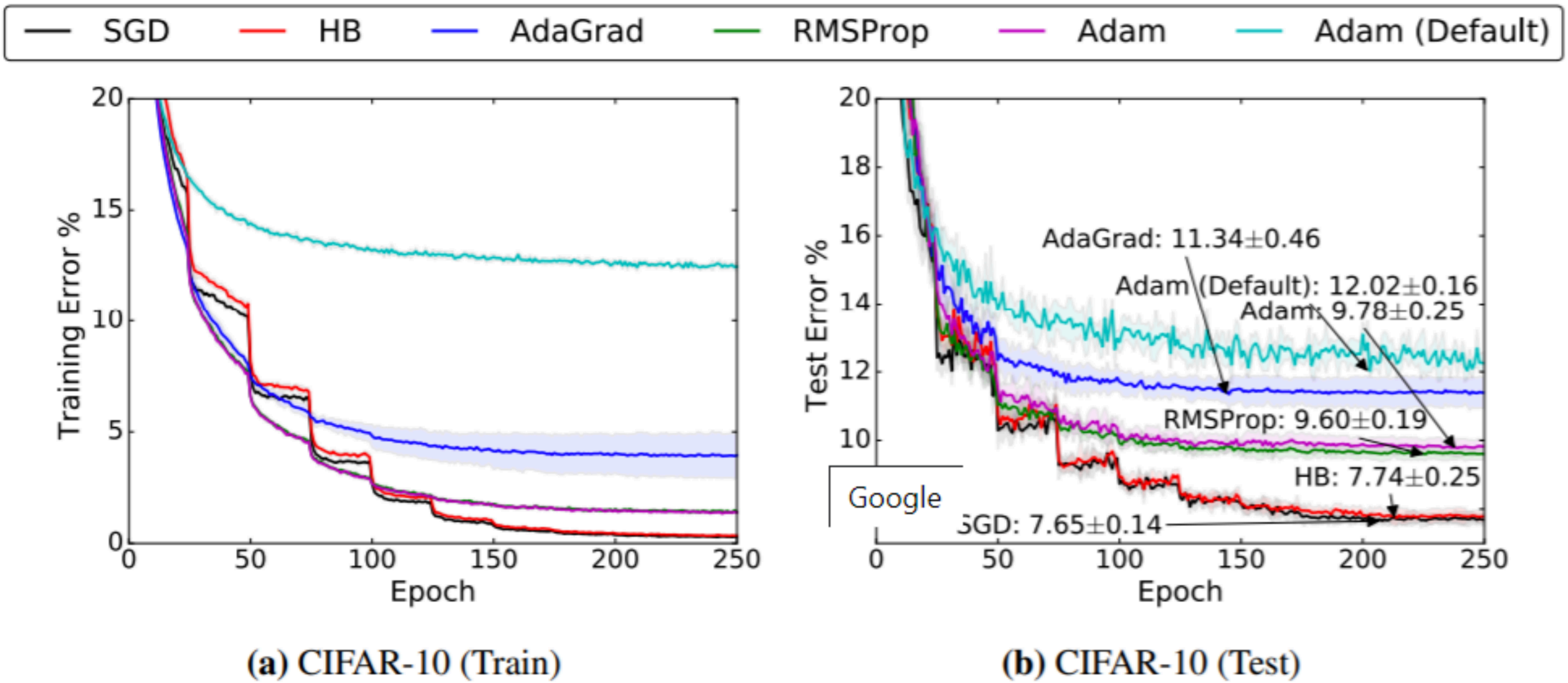


Figure 1: Training (left) and top-1 test error (right) on CIFAR-10. The annotations indicate where the best performance is attained for each method. The shading represents ± one standard deviation computed across five runs from random initial starting points. In all cases, adaptive methods are performing worse on both train and test than non-adaptive methods.

Wilson, Roelofs, Stern, Srebro, Recht '18