HW 1 Released

Simon OH → Thursday 10:30AM This Week

Clarke Differential

Subdifferential and Subgradient

Definition: Given $f : \mathbb{R}^d \to \mathbb{R}$, for every x, the subdifferential set is defined as $\partial_s f(x) \triangleq \{s \in \mathbb{R}^d : \forall x' \in \mathbb{R}^d, f(x') \geq f(x) + s^\top (x' - x)\}$. The elements in the subdifferential set are subgradients.

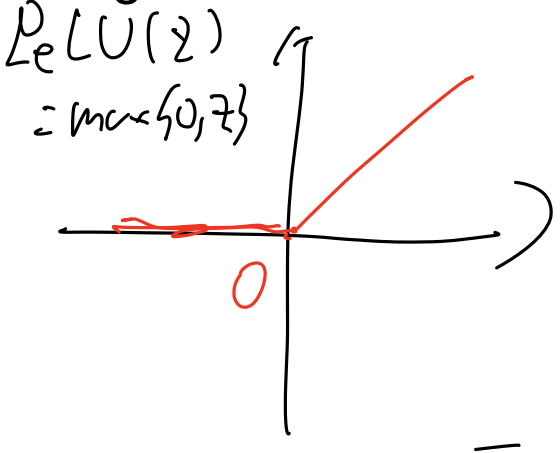
$\min f(w)$

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

Subgradient descent

$x_{t+1} = x_t - \eta g_t$

$g_t \in \partial_s f(x_t)$



$\partial_s f(0) = [0, 1]$

Subdifferential and Subgradient

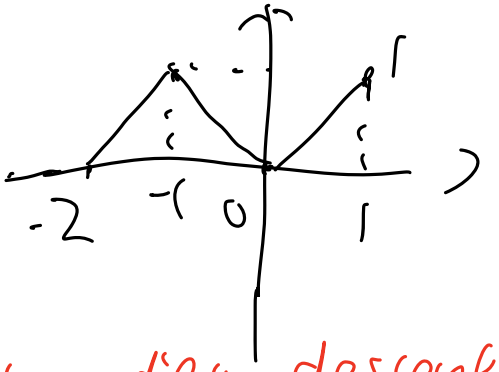
Definition: Given $f : \mathbb{R}^d \to \mathbb{R}$, for every x, the subdifferential set is defined as $\partial_s f(x) \triangleq \{s \in \mathbb{R}^d : \forall x' \in \mathbb{R}^d, f(x') \geq f(x) + s^\top (x' - x)\}$. The elements in the subdifferential set are subgradients.

- If f is convex $\rightarrow$ $\partial_s f$ exists everywhere
- If f is convex & differentiable
  $\partial_s f = \{\nabla f\}$
- $O\left(\frac{1}{\sqrt{T}}\right)$ in convex

Subdifferential is not enough

Definition: Given $f : \mathbb{R}^d \to \mathbb{R}$, for every x, the subdifferential set is defined as
$\partial_s f(x) \triangleq \{s \in \mathbb{R}^d : \forall x' \in \mathbb{R}^d, f(x') \geq f(x) + s^\top (x' - x)\}$. The elements in the subdifferential set are subgradients.

Problem: NN is not convex



subgradient descent is not well-defined

$x = -1$

we need s, $\forall x'$

$f(x') \geq f + s^\top (x' - (-1))$

let $x' = -2$

$0 \geq f + s(-1) \Rightarrow s \geq 1$

let $x' = 1$

$1 \geq f + 2s \Rightarrow s \leq 0$

Clarke Differential

$x_{t+1} = x_t - \eta g_t$

$g_t \in \partial f(x)$

Definition: Given $f : \mathbb{R}^d \to \mathbb{R}$, for every x, the Clark differential is defined as

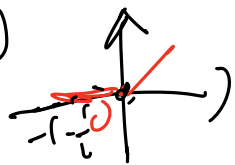
$$\partial f(x) \triangleq \text{conv}\left(\{s \in \mathbb{R}^d : \exists \{x_i\}_{i=1}^{\infty} \to x, \{\nabla f(x_i)\}_{i=1}^{\infty} \to s\}\right).$$

The elements in the subdifferential set are subgradients.

$$\text{conv}(S) = \left\{v : v = \sum_{i=1}^{n} \lambda_i u_i,\ u_i \in S,\ \lambda_i \geq 0,\ \sum_{i=1}^{n} \lambda_i = 1\right\}$$

example 1 $\partial \text{ReLU}(z)$

$= \max\{0, z\}$



$\{x_i\}, 1, -\frac{1}{2}, \ldots \to 0$

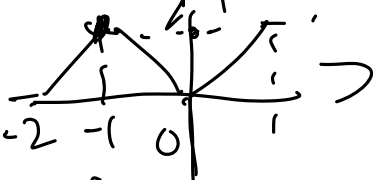
$\partial f(x_i) = 0$

$\{x_i\}, 1, \frac{1}{2}, \ldots \to 0$

$\partial f(x_i) = 1$

$\Rightarrow \partial f(0) = [0, 1]$

example 2:



$\{x_i\}: -2, -1.5, \ldots \to -1 \quad \partial f(x_i) = 1$

$\{x_i\}: 0, -\frac{1}{2}, \ldots \to -1 \quad \partial f(x_i) = -1$

$\Rightarrow \partial f(1) = [-1, 1]$

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.

local

$\forall x, x' \in S, \ |f(x) - f(x')| \leq L \cdot \|x - x'\|$

- If f is locally Lipshitz $\rightarrow \partial f$ exists everywhere
- If f is convex $\Rightarrow \partial f = \partial_s f$
- If f is differentiable $\Rightarrow \partial f = \{\nabla f\}$
- Satisfies chain rule

Positive Homogeneity

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

L- homogeneous

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

$G(z) = \max\{0, z\}$

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

$\prod_{i=1}^{d} (\alpha x_i)^{p_i} = \alpha^{\sum p_i} \cdot \prod x_i^{p_i} = \alpha^L \cdot \prod_i x_i^{p_i}$

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

Positive Homogeneity

$x \in \mathbb{R}^d$, $W_1 \in \mathbb{R}^{m \times d}$

$W_2, \dots W_H \in \mathbb{R}^{m \times m}$

$W_{H+1} \in \mathbb{R}^m$

(4) Multi-layer ReLU NN

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

1) for each layer: 1-homogeneous

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

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

2) for all layers

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

$\Rightarrow$ $(H+1)$-homogeneous

Positive Homogeneity

Fact: $\forall h$

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

independent of h

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

σ': 0 or 1 pattern whether the activation is on or not

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

$\sigma(z) = \sigma'(z) \cdot z$

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

$$\langle W_n, \frac{\partial f}{\partial W_n} \rangle = \langle W_n, (W_{i+1} A_{i+1} \cdots W_{n+1} A_n)^T (A_{n-1} \cdots W_1 X)^T \rangle$$

$A^T B^T = (BA)^T$

$$= tr\left((W_{i+1} A_{i+1} \cdots W_n)^T (A_{n-1} \cdots W_1 X)^T \right)$$

$tr(AB) = tr(BA)$

$$= tr\left((A_{n-1} \cdots W_1 X)^T (W_{i+1} \cdots A_n W_n)^T \right)$$

$$= tr\left(W_{i+1} A_{i+1} \cdots A_1 W_1 X \right)$$

$$= f(X, W_1, \cdots, W_{i+1})$$

□

Positive Homogeneity and Clark Differential

Clark Differential exists

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 $\langle s, x \rangle = Lf(x)$.

View $\chi = W_\eta$: f- homo

Norm Preservation

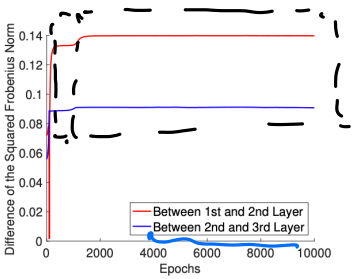
$f(X, W_1, W_2, W_3) = W_3\,\sigma(W_2\,\sigma(W_1 X))$, quadratic loss

$\|W_1\|_F^2$, $\|W_2\|_F^2$, $\|W_3\|_F^2$, $\|A\|_F^2 = \sum_{i,j} A_{ij}^2$

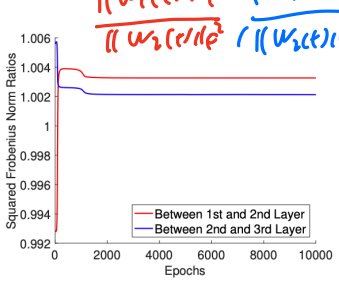
$W_3 \to 0$

$W_2 \to 0$

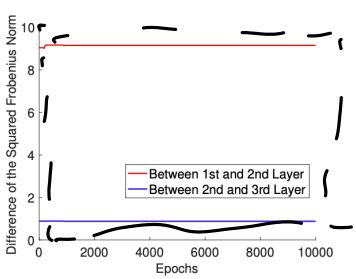
$W_2(t+1) = W_2(t) - \eta \frac{\partial f}{\partial W_2}$



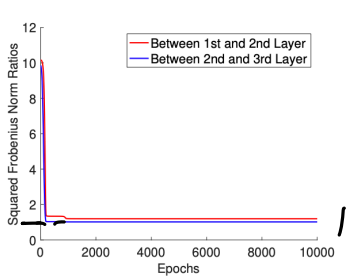
(a) Balanced initialization, squared norm differences.



(b) Balanced initialization, squared norm ratios.



(c) Unbalanced Initialization, squared norm differences.



(d) Unbalanced initialization, squared norm ratios.

$\|W_1(0)\|_F^2 \approx \|W_2(0)\|_F^2 \approx \|W_3(0)\|_F^2$,

- $\|W_1(t)\|_F^2 - \|W_2(t)\|_F^2$
- $\|W_2(t)\|_F^2 - \|W_3(t)\|_F^2$

$\|W_1(0)\|_F^2 \neq \|W_2(0)\|_F^2$ small

$\|W_1(t)\|_F^2 \uparrow \quad \|W_2(t)\|_F^2 \uparrow$

Gradient flow and gradient inclusion

Discrete-time dynamics can be complex. Let's use continuous-time dynamics to simplify: $\eta \to 0$

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

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

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

$dt \quad \eta \to 0$

Norm preservation by gradient inclusion

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 $(h, h') \in [H+1] \times [H+1]$, the gradient inclusion maintains: for all $t \geq 0$,

$$\frac{1}{2}\|W_h(t)\|_F^2 - \frac{1}{2}\|W_h(0)\|_F^2 = \frac{1}{2}\|W_{h'}(t)\|_F^2 - \frac{1}{2}\|W_{h'}(0)\|_F^2.$$

$\Leftarrow$) $\|W_h(t)\|_F^2 - \|W_h(0)\|_F^2 = \|W_{h'}(t)\|_F^2 - \|W_{h'}(0)\|_F^2$

- If $\|W_h(0)\|_F^2$ small $\longrightarrow$ $\|W_h(t)\|_F^2 \approx \|W_{h'}(t)\|_F^2$

Norm preservation by gradient inclusion

Pf: $loss = L(f(X, W_1 \dots W_{H+1}), y)$

$$\frac{dW_n(t)}{dt} = -L'(f(X, W_1, \dots W_{H+1}), y) \cdot \frac{\partial f(X, W_1, \dots, W_{H+1})}{\partial W_n}$$

$$\frac{1}{2}\|W_n(t)\|_F^2 - \frac{1}{2}\|W_n(0)\|_F^2$$

$$= \int_0^t \frac{d}{dt}\frac{1}{2}\|W_n(t)\|_F^2 \, dt$$

$$= \int_0^t \left\langle W_n, \frac{dW_n(t)}{dt} \right\rangle \cdot dt \quad \text{(Chain rule)}$$

$$= \int_0^t \left\langle W_n, \underbrace{-L'(f(X, W_1, \dots, W_{H+1}), y)}_{\text{scalar}} \cdot \underbrace{\frac{\partial f(X, \dots W_{H+1})}{\partial W_n}}_{(2)} \right\rangle$$

(1)

"subgradient" of h $= -\int_0^t L'(f(X, W_1, \dots, W_{H+1}), y) \cdot f(X, W_1, \dots W_{H+1}) \, dt$

Optimization Methods for Deep Learning

Gradient descent for non-convex optimization

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

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

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

对于 $f(x_t)$ 下降 LR too large

Pf: by Taylor expansion & mean-value theorem

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

$$\Delta^\top \nabla^2 f(y)\Delta \le \|\nabla^2 f(y)\|_2 \cdot \|\Delta\|_2^2 \le \beta\|\Delta\|_2^2 \quad \left(y = \lambda x + (1-\lambda)(x+\Delta),\ 0 \le \lambda \le 1\right)$$

Let $\Delta = -\eta \cdot \nabla f(x_t)$

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

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

Converging to stationary points

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

Gradient Descent for Quadratic Functions

Problem: $\min_x \frac{1}{2} x^\top A x$ 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

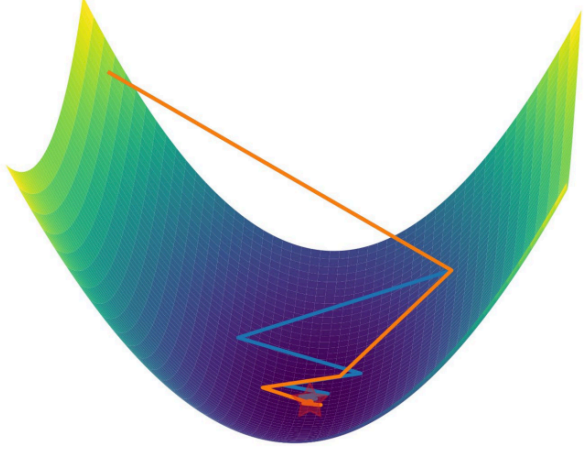
$$\|x_t\|_2 \leq \left(1 - \eta \lambda_{\min}\right)^t \|x_0\|_2$$

Momentum: Heavy-Ball Method (Polyak '64)

Problem: $\min_x f(x)$

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

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



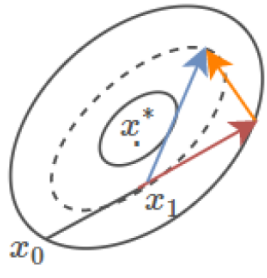
Momentum: Nesterov Acceleration (Nesterov ’89)

Problem: $\min_x f(x)$

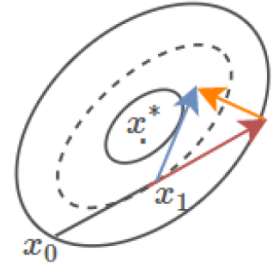
Method: $v_{t+1} = -\nabla f(x_t + \beta v_t) + \beta v_t$

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

Polyak's Momentum

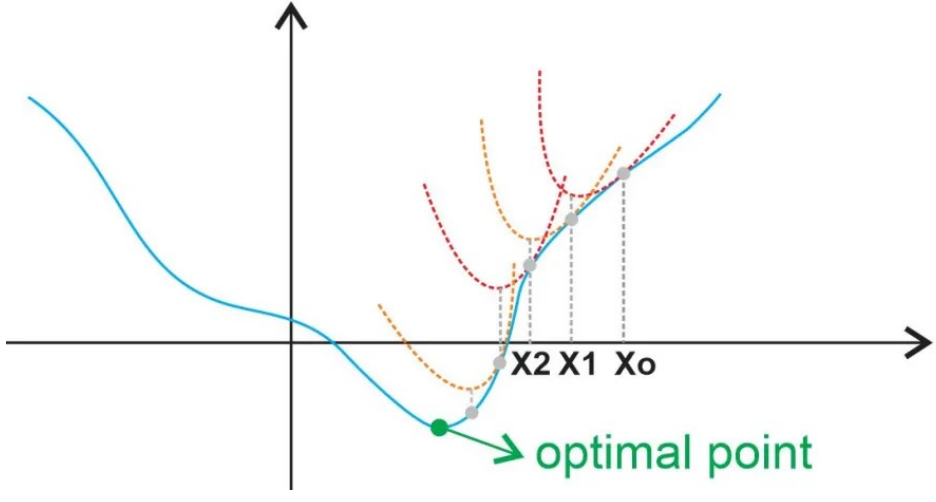


Nesterov Momentum



Newton's Method

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



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

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

RMSProp (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

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

AdaDelta (Zeiler '12)

RMSProp:

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

AdaDelta:

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

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

$$v_{t+1} = \beta_1 v_t + (1-\beta_1)\nabla f(x_t)$$
$$(G_{t+1})_{ii} = \beta_2(G_t)_{ii} + (1-\beta_2)(\nabla f(x_t)_i)^2$$
$$x_{t+1} = x_t - \eta(G_{t+1} + \epsilon I)^{-1/2} v_{t+1}$$

Default choice nowadays.

Other Optimizers

- AdamW
- NAdam
- RAdam
- GGT
- K-FAC
- …