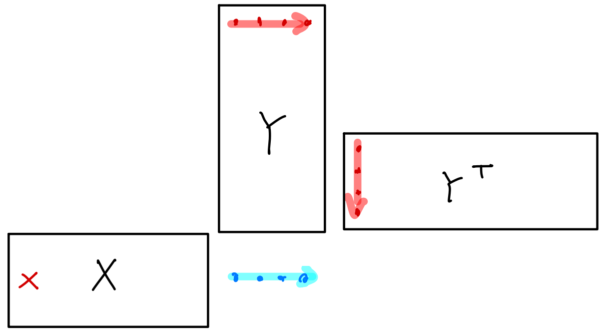
We know $\frac{\partial \mathcal{L}}{\partial W}$, $W = XY$

then ① $\frac{\partial \mathcal{L}}{\partial X} = \frac{\partial \mathcal{L}}{\partial W} \cdot Y^T$

② $\frac{\partial \mathcal{L}}{\partial Y} = X^T \frac{\partial \mathcal{L}}{\partial W}$



GD (Gradient Descent)

$$\min_{\theta} \mathcal{L}(\theta) \Rightarrow \boxed{\theta^{k+1} = \theta^{k} - \alpha \nabla \mathcal{L}(\theta^{k})}$$

Full Batch GD $\tilde{\nabla} f(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x)$ $\{1, 2, \ldots, N\}$

Stochastic " $\tilde{\nabla} f(x) = \nabla f_i(x)$ i

Mini Batch " $\tilde{\nabla} f(x) = \frac{1}{|K|}\sum_{k \in K} \nabla f_k(x)$ K

Sigmoid $\Rightarrow$ tanh $\Rightarrow$ ReLU

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$$\sigma'(z) = \sigma(z)(1-\sigma(z))$$

$$\tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

$$(\tanh)' = 1 - (\tanh)^2$$

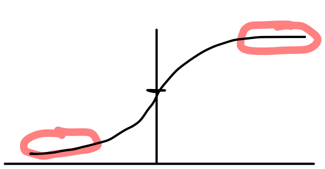
$$\mathrm{ReLU}(z) = \max(0, z)$$

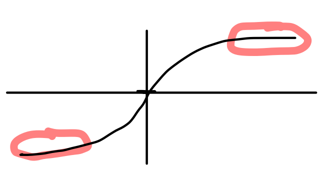
$$(\mathrm{ReLU})' = \begin{cases} 1 & z > 0 \\ 0 & \end{cases}$$

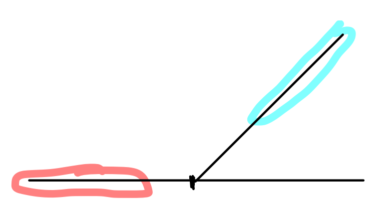
not zero centered.

zero centered.

not zero centered.







Xavier Init $\boxed{\sigma = \frac{1}{\sqrt{D_{in}}}}$

pf. let $y = w^T x$

$$= \sum_{i=1}^{D_{in}} w_i x_i$$

- w_i : i.i.d
- x, w independence.
- $E[w_i] = 0$, $Var(w_i) = \sigma^2$
- $E[x_i] = 0$, $Var[x_i] = V$

Our Goal : $\boxed{Var[y] = Var[x_i]}$

independent.

$$\iint x \cdot w \, p(x, w) \, dx \, dw$$

$$= \iint x \cdot w \, p(x) \cdot p(w) \, dx \cdot dw$$

- $Var[y] = E[y^2] - \cancel{E[y]}^2$
- $E[y] = \sum_i E[x_i w_i] = \sum_i E[x_i] E[w_i]$

$$= 0$$

- $\cancel{Var[y]} = E[y^2]$

$$= \sum_i E[w_i^2 x_i^2] + \sum_{i \neq k} E[\cancel{w_i}] E[\cancel{w_k}] E[x_i x_k]$$

$$= \sum_i \underbrace{E[w_i^2]}_{Var(w_i)} \underbrace{E[x_i^2]}_{Var(x_i)} = \sigma^2 \cancel{V} \cdot D_{in}$$

Note

$\sigma^2 \Uparrow$ $\sigma^2 \Downarrow$

$Var(y) \Uparrow$ $Var(y) \Downarrow$

결론 $\sigma = \frac{1}{\sqrt{D_{in}}}$

Xavier Init ♡ ReLU $\Rightarrow$ $\sigma = \sqrt{\frac{2}{D_{in}}}$

pf. • let $z = \sum_{i}^{D_{in}} w_i x_i$

• then $Var(z) = \sigma^2 V \cdot D_{in}$. $\Longrightarrow$ $z \sim N(0, \sigma^2 V \cdot D_{in})$

$\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad$ " δ

• let $h = \phi(z)$

• Our Goal : $Var(h) = Var(x_i)$

let $u = \frac{z^2}{2\delta}$

$dz = \frac{\delta}{z} du$

$$\cdot\ E[h] = \int_0^\infty z\, p(z)\, dz$$

$$= \int_0^\infty z \cdot \frac{1}{\sqrt{2\pi\delta}} \cdot \exp\left(-\frac{z^2}{2\delta}\right) dz.$$

$$= \int_0^\infty \frac{\cancel{z}}{\sqrt{2\pi\delta}} \cdot e^{-u} \cdot \frac{\delta}{\cancel{z}} du = \sqrt{\frac{\delta}{2\pi}}$$

$$\cdot\ E[h^2] = \int_0^\infty z^2 \cdot p(z)\, dz$$

$$= \frac{1}{2} \int_{-\infty}^\infty z^2 p(z)\, dz = \frac{1}{2} E[z^2]$$

$$= \frac{1}{2} Var[z^2] = \frac{1}{2}\delta$$

$$\cdot\ Var[h] = \frac{1}{2}\delta - \frac{\delta}{2\pi} \approx \boxed{\frac{1}{2}} Var[z]$$

"분산의 다양성이

1/2배가 된다"

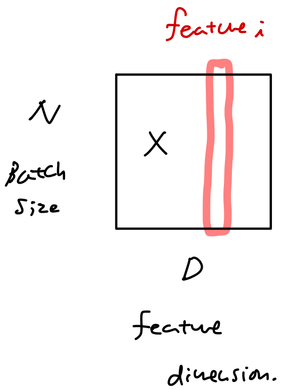
$$\cdot\ \frac{1}{2} \cdot \underbrace{\sigma^2 \cancel{V} \cdot D_{in}}_{\delta} = \cancel{V}$$

$$\sigma = \sqrt{\frac{2}{D_{in}}}$$

Batch Normalization

inference에서 $N=1$일 수도 있다!

$$\begin{cases} \mu^{run} \leftarrow m \cdot \mu^{run} + (1-m) \cdot \mu^{batch} \\ (\sigma^2)^{run} \leftarrow m \cdot (\sigma^2)^{run} + (1-m)(\sigma^2)^{batch} \end{cases}$$



- $\hat{x}_{ij} = \dfrac{x_{ij} - \mu_j}{\sqrt{\sigma_j^2 + \epsilon}}$

- $\hat{y}_{ij} = \gamma_j \hat{x}_{ij} + \beta_j$

학습한다.

Why BN works?

① BN ⇒ ICS ⇓

Internal Convergence Shift : shift in the mean/var of hidden activation during training

② BN ⇏ ICS ⇓

but BN ⇒ training ⇑

③ BN ⇒ smooth loss landscape. "gradient의 변화가 안정적"

FC 레이어에 Normalization 하면

cost = 0 (linear)

Gradient Descent! $\begin{cases} \text{minimize } f(x) \\ \text{update } x' = x - \eta \nabla f(x) \end{cases}$

- suppose $x_{t+1} = x_t + \eta v$
- then by Taylor Approximation

$$f(x_{t+1}) \approx f(x_t) + \langle \nabla f(x_t), \eta v \rangle$$

- minimizing $f(x_{t+1})$, $v = -\dfrac{\nabla f(x_t)}{\|\nabla f(x_t)\|}$
- $x_{t+1} = x_t - \eta \dfrac{\nabla f(x_t)}{\|f(x_t)\|}$

Lemma 3.1

- let $f : \mathbb{R}^d \to \mathbb{R}$ continuously differentiable func.
- let f β-smooth. $\Rightarrow$ $\forall x, y$ $|\nabla f(y) - \nabla f(x)| \le \beta \|x - y\|$

then $f(y) \le f(x) + \underbrace{\langle \nabla f(x), y - x \rangle}_{\text{접선}} + \underbrace{\frac{\beta}{2} \|y - x\|^2}_{\text{곡률선.}}$.

---

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

then $f(x_{t+1}) \le f(x_t) + \underbrace{\langle \nabla f(x_t), -\eta \nabla f(x_t) \rangle}_{-\eta \|\nabla f(x_t)\|^2} + \underbrace{\frac{\beta}{2} \|-\eta \nabla f(x_t)\|^2}_{\frac{\beta}{2}\eta^2 \|\nabla f(x_t)\|^2}$

$$= f(x_t) - \left[\eta - \frac{\beta}{2}\eta^2\right] \|\nabla f(x_t)\|^2$$

$\eta \ll 1$ then positive

Proof for 3.1

let $g(t) = f(x + t(y-x))$

then $g'(t) = \langle \nabla f(x + t(y-x)), y-x \rangle$

$g''(t) = (y-x)^T \nabla^2 f(x+t(y-x))(y-x) \leq \beta \|y-x\|^2$

$$\int_0^1 (1-s) g''(s)\, ds = \Big[(1-s) g'(s)\Big]_0^1 + \int_0^1 g'(s) ds$$

$$g(1) = g(0) + g'(0) + \int_0^1 (1-s) g''(s) ds$$

$$f(y) = f(x) + \langle \nabla f(x), y-x \rangle + \int_0^1 (1-s)(y-x)^T \nabla^2 f(x+s(y-x))(y-x)\, ds.$$

$$\leq \int_0^1 (1-s) \beta \|y-x\|^2 ds$$

$$= \frac{\beta}{2} \|y-x\|^2 .$$

let $\|\nabla f(y) - \nabla f(x)\| \le \beta \|x - y\|$

then $v^T \nabla^2 f(x) v \le \beta \|v\|^2$

let $\phi(t) = \nabla f(x + t(y-x))$

$\phi'(t) = \nabla^2 f(x + t(y-x))(y-x)$

then $\nabla f(y) - \nabla f(x) = \int_0^1 \phi'(t)\, dt$

$= \int_0^1 \nabla^2 f(x + t(y-x))\, dt \cdot (y-x)$

then $\langle \nabla f(y) - \nabla f(x), y-x \rangle \le \|\nabla f(y) - \nabla f(x)\| \|y-x\|$

$\le \beta \|y-x\|^2$

too small

$(y-x)^T \int_0^1 \nabla^2 f(x + t(\cancel{y}-x))\, dt\, (y-x) \le \beta \|y-x\|^2.$

$v^T \nabla^2 f(x) v \le \beta \|v\|^2.$

SGD

$$f(y) \le f(x) + \langle \nabla f(x), y-x \rangle + \frac{\beta}{2}\|y-x\|^2$$

let $x_{t+1} = x_t - \eta \tilde{\nabla} f(x_t)$ $\rightarrow$ $E_t[\tilde{\nabla} f(x_t)] = \nabla f(x_t)$

then $f(x_{t+1}) \le f(x_t) - \eta \langle \nabla f(x_t), \tilde{\nabla} f(x_t) \rangle + \frac{\beta}{2} \cdot \eta^2 \|\tilde{\nabla} f(x_t)\|^2$

$$E_t[f(x_{t+1})] \le f(x_t) - \eta \|\nabla f(x_t)\|^2 + \frac{\beta}{2}\eta^2 \underbrace{E_t[\|\tilde{\nabla} f(x_t)\|^2]}_{\le G}$$

$$E[\|\nabla f(x_t)\|^2] \le \frac{1}{\eta}\Big(E[f(x_t)] - E[f(x_{t+1})]\Big) + \frac{\beta}{2}\eta G$$

lower bound

$$\sum_{t=0}^{T-1} E[\|\nabla f(x_t)\|^2] \le \frac{1}{\eta}\left(f_0 - f^*\right) + \frac{\beta}{2}\eta G T$$

$$\min E[\|\nabla f(x_t)\|^2] \le \frac{1}{\eta T}[f_0 - f^*] + \frac{\beta}{2}\eta G$$

let $\eta = \frac{1}{\sqrt{T}}$ then $\min E[\|\nabla f(x_t)\|^2] = O\left(\frac{1}{\sqrt{T}}\right)$
