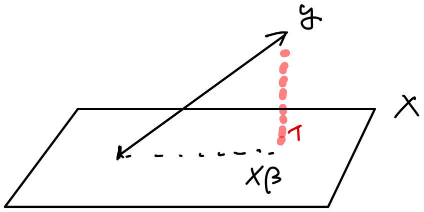


수선의 발.

$$X^T y = X^T X \beta$$

β를 어떻게 찾을까?

Least Square Estimate

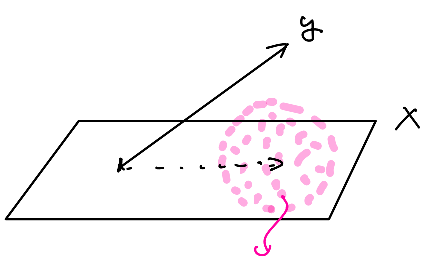
$$\nabla_\beta \, \beta^T A \beta = \nabla_\beta \sum_{i,j} \beta_i A_{ij} \beta_j = (A + A^T)\beta$$

$$
\begin{aligned}
(\text{loss}) &= \|y - X\beta\|^2 \\
&= (y - X\beta)^T (y - X\beta) \\
&= (y^T - \beta^T X^T)(y - X\beta) \\
&= y^T y - y^T X \beta - \beta^T X^T y + \beta^T X^T X \beta
\end{aligned}
$$

$$\nabla_\beta (\text{loss}) = -2 X^T y + 2 X^T X \beta = 0$$ (극값)

hence $X^T y = X^T X \beta$ 수선의 발

Maximum Likelyhood Estimate



가우시안 공

let $Y = X\beta + \varepsilon \sim N(X\beta, \sigma^2 I)$

then $p(Y) \Uparrow = \frac{1}{(2\pi)^{d/2} \cdot \sigma^d} \cdot \exp\left(\frac{-1}{2\sigma^2} \|y - X\beta\|^2 \Downarrow \right)$

likelyhood

squared
L2 Norm

$p(\theta)$: prior

$p(data|\theta)$: likelyhood

$p(data)$: evidence.

$$p(\theta|data) : \text{posterior} = \frac{p(data|\theta)\cdot p(\theta)}{\cancel{p(data)}}$$

Bernouli R.V "동전 던지기"

$$p(x_i|\theta) = \begin{cases} \theta & x_i = 1 \\ 1-\theta & x_i = 0 \end{cases}$$

$$= \theta^{x_i}(1-\theta)^{1-x_i}$$

---

likelyhood 의 관점.

$$p(x|\theta) = \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$$

$$\log p(x|\theta) = \Sigma x_i \log\theta + (n-\Sigma x_i)\log(1-\theta)$$

$$\nabla_\theta \log p(x|\theta) = \frac{\Sigma x_i}{\theta} - \frac{(n-\Sigma x_i)}{(1-\theta)}$$

say $\nabla_\theta \log p(x|\theta) = 0$ (MLE)

then .... $\theta = \frac{\Sigma x_i}{n}$

$$\text{Beta}(\alpha,\beta) = x^{\alpha-1}\cdot(1-x)^{\beta-1}\cdot\frac{\Gamma(\alpha+\beta)}{\cancel{\Gamma(\alpha)\cdot\Gamma(\beta)}}$$

where $0 \le x \le 1$

Note.

$\Gamma(n) = (n-1)!$

---

posterior 의 관점

실험결과 (H) 7 (T) 3

$\nabla_\theta \log p(\theta|X) = 0$ (MAP)

$$p(\theta|X) \propto \underbrace{1}_{p(\theta)}^{\text{Beta}(1,1)} \cdot \underbrace{\theta^7(1-\theta)^3}_{p(X|\theta)} = \overbrace{\theta^7(1-\theta)^3}^{\text{Beta}(8,4)}$$

$\approx$> Uniform Distribution 이니까

$\theta_{MAP} = \theta_{MLE}$ ★★

$= 0.7$

$$p(\theta|X) \propto \underbrace{\theta^1\cdot(1-\theta)^1}_{p(\theta)}^{\text{Beta}(2,2)} \underbrace{\theta^7(1-\theta)^3}_{p(X|\theta)} = \overbrace{\theta^8(1-\theta)^4}^{\text{Beta}(9,5)}$$

$\approx$> $\theta_{MAP} = 0.666\ldots$

엔트로피

- $h(x) = -\log p(x)$
- if x, y independent

$$h(x, y) = h(x) + h(y)$$

- $H(X) = E[-\log P(X)] \geq 0$
- $H(X, Y) = H(X) + H(Y|X)$

(축구) $H \Uparrow$ 독일 vs 스페인 $H \Downarrow$ 독일 vs 한국

KL Divergence

$D_{KL}(P \| Q)$

$$= \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$= \int p(x) \log \frac{p(x)}{q(x)} dx \geq 0$$

(x) symmetric

(x) triangular inequality

(pf)

$$\mathop{E}_{p(x)} \left[-\log \frac{q(x)}{p(x)} \right]$$

$$\geq -\log \mathop{E}_{p(x)} \left[\frac{q(x)}{p(x)} \right] \quad \text{Const}$$

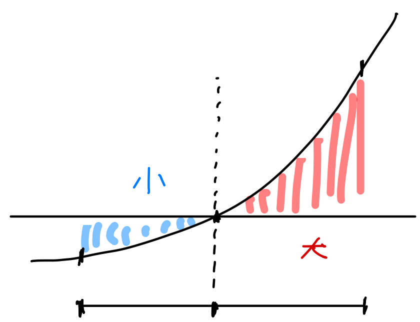
$$= -\log \sum_x \cancel{p(x)} \cdot \frac{q(x)}{\cancel{p(x)}} \geq 0 \quad \text{Const} = 1$$

Jensen's Inequality

f convex

then $E[f(x)] \geq f(E(x))$

$\Updownarrow$ $f(x) = \text{Const}$



Mutual Information

$$I(X;Y) = D_{KL}(P(x,y) \| p(x) \cdot p(y))$$

$$= \sum p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$= H(X) + H(Y) - H(X,Y)$$

$$= H(X) - H(X|Y)$$

Cross Entropy

- $H_p(q) = \mathop{E}_{p} [-\log q(x)]$

$$= -\sum_x p(x) \log q(x)$$

- $D_{KL}(P \| q) = \mathop{E}_{p} \left[\log \frac{p(x)}{q(x)} \right]$

$$= H_p(q) - \cancel{H(p)} \geq 0$$

fixed

overfitting

$$y = \underbrace{\begin{bmatrix} 1 & x^1 & \cdots & x^m \end{bmatrix}}_{\text{단일 변수에 의존}} \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}}_{\text{불필요하게 큰 차원}} \Rightarrow \text{overfitting}$$

Regularization

$$(loss) = \|Y - X\beta\|^2 + \underbrace{\lambda\|\beta\|^2}_{\text{Regularization}}$$

$$\nabla_\beta (loss) = \nabla_\beta \left(y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X \beta \right) + \nabla_\beta \lambda \beta^T \beta$$

$$= 2 \left[\underbrace{-X^T(Y - X\beta)}_{\text{데이터 예측}} + \underbrace{\lambda\beta}_{\text{단순한 모델}} \right] = 0$$

Kernel

$$K(x_1, \underbrace{x_2}_{\text{low dimension}}) = \phi(x_1)^T \underbrace{\phi(x_2)}_{\text{high dimension}}$$

$$(loss) = \|Y - \phi(X)^T\beta\|^2 + \lambda\|\beta\|^2$$

$$= y^T y - y^T \phi(X)^T \beta - \beta^T \phi(X) Y + \beta^T \phi(X)\phi(X)^T \beta + \lambda \beta^T \beta$$

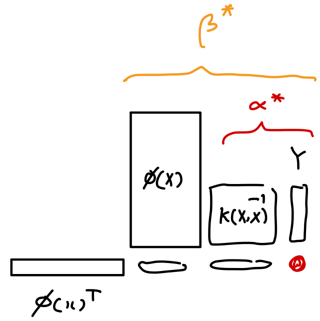
$$\nabla_\beta (loss) = 2 \left[-\phi(X)(Y - \phi(X)^T\beta) + \lambda\beta \right] = 0$$

$$\beta^* = (\phi(X)\phi(X)^T + \lambda I)^{-1} \phi(X) Y$$

$$= \phi(X) \underbrace{(\phi(X)^T\phi(X) + \lambda I)^{-1} Y}_{= \alpha^*}$$

$$f^*(x) = \phi(x)^T \beta^*$$

$$= K(x, X)\alpha^*$$

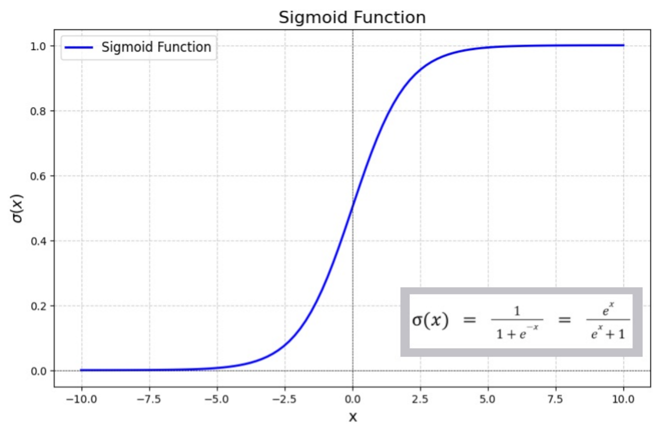


Sigmoid 임의의 값을 확률 형태로 변환해줌.

if probability $= \dfrac{e^z}{1+e^z}$ $\quad [0, 1]$

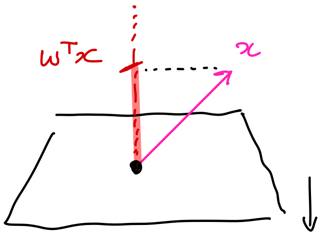
then odds $= e^z$ $\quad [0, \infty)$

$\log(\text{odds}) = z$ $\quad (-\infty, \infty)$



Note that $\sigma' = \sigma(1-\sigma)$

Logistic Regression (MLE)



아래로 당기면
신뢰도 상승

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

$$\nabla_w \sigma_i = \sigma_i(1-\sigma_i)\,\nabla_w(w^T x_i)$$
$$= \sigma_i(1-\sigma_i)\, x_i$$
$$\propto x$$

$$p(y_i \mid x_i, w) = \left(\sigma(w^T x_i)\right)^{y_i}\left(1-\sigma(w^T x_i)\right)^{1-y_i}$$

$$\mathcal{L} = \prod_{i=1}^{\sim} p_i$$

data

$$\log p_i = \left(y_i \log \sigma_i\right) + \left((1-y_i)\log(1-\sigma_i)\right)$$

$$\nabla_w(-\log p_i) = \left(y_i(\sigma_i - 1)\, x_i\right) + \left((1-y_i)\,\sigma_i x_i\right)$$

$$= \left(\begin{matrix} y_i \sigma_i x_i \\ - y_i x_i \end{matrix}\right) + \left(\begin{matrix} \sigma_i x_i \\ - y_i \sigma_i x_i \end{matrix}\right)$$

$$= (\sigma_i - y_i)\, x_i$$

예측 실제

리-표현변
위치

예측 실제

hence $$\nabla_w(-\log \mathcal{L}) = \sum_i \left(\sigma(w^T x_i) - y_i\right) x_i$$

$$\nabla_w \nabla_w(-\log p_i) = \left[\nabla_w(\sigma_i - y_i)\right] x_i^T$$
$$= \sigma_i' \, x_i \, x_i^T$$

$$\begin{array}{cccc} & X^T & S & X \\ \nabla_w\nabla_w(-\log \mathcal{L}) = & \begin{bmatrix} | & \\ x_1 & \cdots \\ | & \end{bmatrix} & \begin{bmatrix} \sigma_1' & & \\ & \ddots & \\ & & \sigma_n' \end{bmatrix} & \begin{bmatrix} - x_1 - \\ \ddots \end{bmatrix} \end{array}$$

≥ 0 ≥ 0

because $\sigma' = \sigma(1-\sigma) \geq 0$

for any v

$$v^T(X^T S X)\, v$$
$$= (Xv)^T S (Xv)$$
$$= \sum_i (\sigma_i')\left((Xv)_i\right)^2 \geq 0$$

≥ 0 ≥ 0

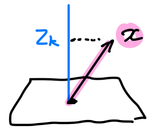
positive semi-definite. $\longleftrightarrow$ $-\log\mathcal{L}$ convex

$\langle v, Hv \rangle \geq 0$

有 global min

Multiclass classification

let $z_k = w_k^T x$



$$\nabla_{w_m}(w_k^T x) = \begin{cases} x & \text{if } m = k \\ 0 & \text{if } m \neq k. \end{cases}$$

$$= \delta_{mk} \cdot x$$

then $$p_k = \frac{e^{z_k}}{\sum_j e^{z_j}}$$

$$\nabla_{w_m}(p_k) = \frac{\sum_j e^{z_j}\left(e^{z_k} \cdot \overset{\delta_{mk} \cdot x}{(\nabla z_k)}\right) - e^{z_k}\left(\overset{e^{z_m} \cdot x}{\sum_j e^{z_j} \nabla z_j}\right)}{\left(\sum_j e^{z_j}\right)^2}$$

$$= p_k(\delta_{mk} - p_m)\, x$$

$$\nabla_{w_m}\left(-\log \begin{matrix}\text{likelyhood} \\ \text{of } x\end{matrix}\right) = \sum_{k=1}^{c} y_k \nabla(-\log p_k)$$

$$= \sum_{k=1}^{c} y_k \cdot \frac{1}{\cancel{p_k}} \cdot \cancel{p_k}(p_m - \delta_{mk})\, x$$

$$= (p_m - y_m)\, x$$

$$\nabla_{w_m}(-\log \mathcal{L}) = \sum_{i=1}^{N} (p_{im} - y_{im})\, x_i$$

예측　　실제

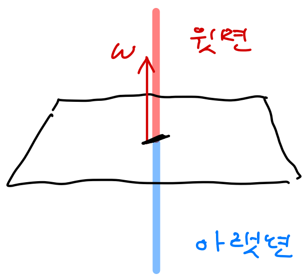
SVM 초평면 $w^T x + b = 0$

조건 ①

$\forall i \quad y_i (w^T x_i + b) \geq 0$

조건 ②

$\min_x |w^T x + b| = 1$

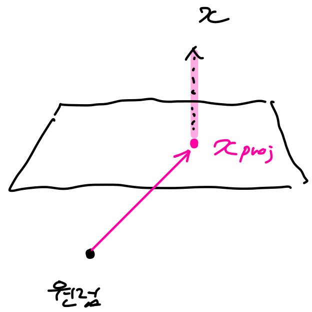


윗면 $y = 1$

$w^T x + b \geq 0$

아랫면 $w^T x + b \leq 0$

$y = -1$



$x = x_{proj} + d \frac{w}{\|w\|}$

$d = \frac{|w^T x + b|}{\|w\|}$

$= \frac{1}{\|w\|}$ (가장 가까운 점)

$\frac{1}{2}\|w\|^2$ $\qquad$ $y_i (w^T x_i + b) \geq 1$

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + \max_{\alpha_i \geq 0} \alpha_i (1 - y_i (w^T x_i + b)) = \begin{cases} 0 & y_i (w^T x_i + b) > 1 \\ \infty & \text{else} \end{cases}$$

$$\min_{w, b} \max_{\alpha_i \geq 0} \left[\frac{1}{2}\|w\|^2 + \sum_i \alpha_i (1 - y_i (w^T x_i + b)) \right] \geq \max_{\alpha_i \geq 0} \min_{w, b} [\quad " \quad]$$

$$= \frac{1}{2} w^T w + \sum_i \alpha_i - w^T \underbrace{\left(\sum_i \alpha_i y_i x_i \right)}_{w} - b \underbrace{\sum_i \alpha_i y_i}_{0}$$

maximize

$\sum_i \alpha_i - \frac{1}{2} w^T w$

with

$$\begin{cases} \nabla_w ② = w - \sum_i \alpha_i y_i x_i = 0 \\ \frac{\partial}{\partial b} ② = \sum_i \alpha_i y_i = 0 \end{cases}$$

say $i \in I$ whenever $\alpha_i > 0$

i.e $y_i(w^T x_i + b) = 1$

then $w = \sum_i \alpha_i y_i x_i \quad (i \in I)$

$b = y_i - w^T x_i \quad (i \in I)$

$$w^T w = \sum_i \alpha_i y_i (w^T x_i) \quad (i \in I)$$

$$= \sum_i \alpha_i (1 - y_i b) \quad \text{because} \quad y_i(w^T x_i + b) = 1$$

$$= \sum_i \alpha_i - b \cancel{\sum_i \alpha_i y_i}$$

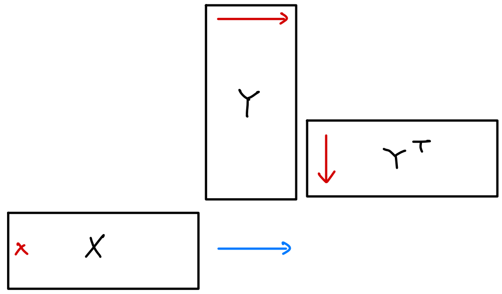
$$\frac{|x^T w + b|}{\|w\|} = \frac{1}{\sqrt{\sum \alpha_i}}$$

---

say $W = XY$

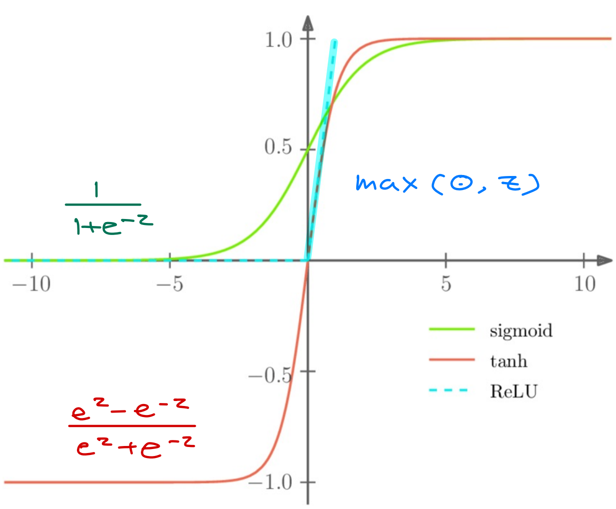
then $$\frac{\partial \mathcal{L}}{\partial X} = \frac{\partial \mathcal{L}}{\partial W} \cdot Y^T$$

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



Gradient Descent $\quad \theta_{k+1} = \theta_k - \alpha \nabla \underbrace{\mathcal{L}(\theta_k)}$ minimize

- Full Batch GD $\quad \tilde{\nabla} f(x) = \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x)$
- Stocastic GD $\quad \tilde{\nabla} f(x) = \nabla f_i(x)$
- Mini Batch GD



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

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

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

Note. σ 나 tanh는 (대충) linear

let $z = \sum_{i=1}^{D_{in}} w_i x_i$ s.t.

$$\begin{cases} w : \text{i.i.d}, \quad E[w_i] = 0 \quad Var[w_i] = \sigma^2 \\ x : \text{i.i.d} \quad \text{independent of } w \\ \quad E[x_i] \neq 0 \quad (\text{in general}) \end{cases}$$

let $S = E[x^2]$ denote the 2nd moment of the input.

---

we know that $E[z] = 0$

Note. independent

$$\iint xy\, p(x,y)\, dx\, dy = \iint xy\, p(x) p(y)\, dx\, dy$$

then $Var(z) = \sum_{i=1}^{D_{in}} Var(w_i x_i)$

$$= \sum_{i=1}^{D_{in}} E[w_i^2 x_i^2] - E[w_i x_i]^2 = D_{in} \sigma^2 S$$

$$E[w_i^2] E[x_i^2] \qquad E[w_i]^2 E[x_i]^2$$

$$\sigma^2 \quad S \qquad 0$$

in Xavier Init we want $E[z^2] = E[x^2]$

$$D_{in} \sigma^2 \cancel{S} = \cancel{S} \qquad \sigma = \frac{1}{\sqrt{D_{in}}}$$

---

now, let $h = ReLU(z)$

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

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

$$= \frac{1}{2} D_{in} \sigma^2 S$$

in He Init we want $E[h^2] = E[x^2]$

$$\frac{1}{2} D_{in} \sigma^2 \cancel{S} = \cancel{S} \qquad \sigma = \sqrt{\frac{2}{D_{in}}}$$

$$\begin{cases} \text{Xavier} \quad \sigma_{ref}^2 = \frac{1}{D_{in}} \\ \text{He} \quad \sigma_{ref}^2 = \frac{2}{D_{in}} \end{cases}$$

then $w \sim N(0, \sigma_{ref}^2)$

or $w \sim Uni(-\sqrt{3}\sigma_{ref}, \sqrt{3}\sigma_{ref})$

as variance of $Uni(-b, b)$ becomes

$$\int_{-b}^{b} x^2 \cdot \frac{1}{2b}\, dx$$

$$= \left[\frac{x^3}{6b} \right]_{-b}^{b} = \frac{b^3}{3}$$

learning rate decay

① at a few fixed points ?

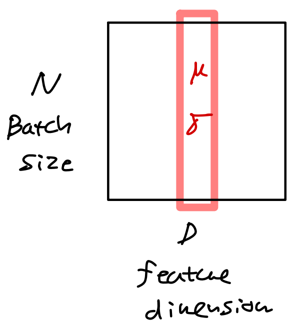
② cosine : $\frac{1}{2}\alpha_0\left(1+\cos\left(\frac{t\pi}{T}\right)\right)$

③ linear : $\alpha_0\left(1-\frac{t}{T}\right)$

④ linear sqrt : $\alpha_0/\sqrt{t}$

linear warmup : linearly increase learning steps at $t \approx 0$

Batch Normalization



$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sqrt{\sigma_j^2 + \varepsilon}}$$

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

(training)

$$\mu^{run} = m \cdot \mu^{run} + (1-m)\,\mu^{batch}$$

$$(\sigma^2)^{run} = m(\sigma^2)^{run} + (1-m)(\sigma^2)^{batch}$$

BN $\not\Rightarrow$ Internal Convergence Shift : shift in the mean/var of hidden activation

$\Rightarrow$ Smooth loss landscape : stable gradient.

GD minimizes $f(x)$

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

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

GD idea

set $v = -\dfrac{\nabla f(x_t)}{\| f(x_t) \|}$

Lemma 3.1

- $f : \mathbb{R}^d \to \mathbb{R}$ continuously diffable function.

- let f β-smooth : $|\nabla f(x) - \nabla f(y)| \leq \beta \|x - y\|$ $(\forall x, y)$

then $f(y) \leq f(x) + \underbrace{\langle f(x), y - x \rangle}_{\text{접선}} + \underbrace{\frac{\beta}{2} \|y - x\|^2}_{\text{[illegible]}}$

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

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

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

small step

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

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

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

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

lower bound

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

$$\min_t \Big(E \| \nabla f(x_t) \|^2 \Big) \leq \frac{1}{\eta T} (f_0 - f^*) + \frac{\beta}{2} \eta^2 G$$

let $\eta = \frac{1}{\sqrt{T}}$

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

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

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

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

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

$$= g(1) - g(0) - g'(0)$$

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

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

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

---

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

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

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

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

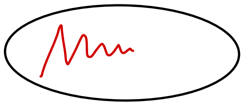
$$(y-x)^T \int_0^1 \nabla^2 f(x + t(\cancel{y}-x)) dt \, (y-x)$$

$$= \langle \nabla f(y) - \nabla f(x), y-x \rangle \leq \|\nabla f(y) - \nabla f(x)\| \, \|y-x\|$$

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

GD

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



강녹운동

+Momentum.

$$x_{t+1} = x_t + V_{t+1}$$

$$V_{t+1} = \rho V_t - \alpha \nabla f(x_t)$$

Nesterov Momentum

$$x_{t+1} = x_t + V_{t+1}$$

$$V_{t+1} = \rho V_t - \alpha \nabla f(x_t + \rho V_t)$$

$\| \nabla f(x) \|$. 가파름 정도 .

" 가파르면 느리게

완만하면 빠르게 "

AdaGrad

$$x = x - \alpha \frac{\nabla f(x_t)}{\sqrt{\sum (\nabla f(x_t))^2}}$$

$\rightarrow \infty$

RMSProp

$$x = x - \frac{\eta g_t}{\sqrt{E[g^2]_t}}$$

$$E[g^2]_t = \beta E[g^2]_{t-1} + (1-\beta) g_t^2$$

Adam

$$x_{t+1} = x - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t}}$$

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t$$

$m_0 = 0$

$v_0 = 0$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \qquad \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

if g const

then
$$\begin{cases} m_k = (1-\beta_1^k)\, g \\ v_k = (1-\beta_2^k)\, g^2 \end{cases}$$