# 10. Algorithmic Regularization [ Idea !!! ] = EARLY STOPPING

## 10.1 Early Stopping least-squares

- 우리가 관찰할 data $y_i = \langle x_i, \theta^* \rangle + z_i$ , $z_i \sim N(0, \sigma^2)$

- Goal : $\theta^*$를 잘 추정해보자. ≈ prediction risk를 줄이자.
- 이때 사용하는 estimator는 GD의 iterate $\theta_t$를 충분히 학습시키지 않고 early stopping 하면 regularization 효과 有

즉: GD를 완전히 수렴시키지 않고 적절한 $t$에서 멈추는 걸로 overfitting 방지 가능 = Algorithmic Regularization (Implicit Regularization)

### 10.1.2 Iterates of gradient descent

GD의 iteration 식을 수학적으로 전개해서 다음 식 유도 :

$$\theta_t = -\eta \sum_{k=0}^{t-1} V\Sigma^T (I - \eta \Sigma\Sigma^T)^k U^T r_0$$

전개한 식은 이후 risk decomposition에 사용

### 10.1.3 Risk of gradient descent iterates

$\theta_t$의 risk를 bias + variance로 나누자!

$$E[R(\theta_t)] = \underbrace{\sum_{i=1}^{d} (1 - \eta\sigma_i^2)^{2t}}_{\text{bias}^2} + \underbrace{\sigma^2 \sum_{i=1}^{d} \frac{(1-(1-\eta\sigma_i^2)^t)^2}{\sigma_i^2}}_{\text{variance}}$$

Insights : bias는 감소하고 variance는 증가함. ⇒ 전체 risk는 U자형 (U-shaped) 곡선을 따라감.

= 너무 작은 $t$는 underfitting , 너무 큰 $t$는 overfitting

→ 적절한 시점에 early stopping을 하는게 best risk.

### 10.1.4 Comparison to ridge regression estimator

일반 Ridge Regression 과 비교했을때?

- Ridge Regression Estimator

$$\hat{\theta}_{ridge} = V \cdot \mathrm{diag}\left(\frac{\sigma_i^2}{\sigma_i^2 + \lambda}\right) V^T \theta^* + V \cdot \mathrm{diag}\left(\frac{\sigma_i}{\sigma_i + \lambda}\right) U^T z$$

- Risk

$$E[R(\theta_t)] = \sum_{i=1}^{d} \left(1 - \frac{\sigma_i^2}{\sigma_i^2 + \lambda}\right)^2 + \sigma^2 \sum_{i=1}^{d} \left(\frac{\sigma_i}{\sigma_i^2 + \lambda}\right)^2$$

Ridge는 $\lambda$로 Regularization

GD 는 $t$로 Regularization

ONLY ) 모든 $\sigma_i$가 같을때만 동일 Risk 가짐

## 10.1 Early Stopping Least Squares

• Problem State

$y_i = \langle x_i, \theta^* \rangle + z_i$ , $z_i \sim N(0, \sigma^2)$ , linear model with noise 를 가정. 우리의 goal : to estimate well $\theta^*$

then model risk $\hat{h}(x) = \langle \hat{\theta}, x \rangle$ small

• suppose feature vector $x$ is Gaußian distributed ($x \sim N(0, I)$) then prediction risk is :

$$R(\hat{h}) = E[(\hat{h}(x) - y)^2] = E[(\langle \hat{\theta} - \theta^*, x \rangle - z)^2] = \|\hat{\theta} - \theta^*\|_2^2 + \sigma^2$$

Interpret) 만약 estimate $\hat{\theta}$ 가 true model $\theta^*$ 와 유사하면 risk가 낮아지고 "좋은 모델"이 됨.

AS AN ESTIMATOR )

we knew the iterates of gradient descent applied to the least-squares

$L(\theta) = \frac{1}{2} \| X\theta - y \|_2^2$

## 10.1.2 Iterates of gradient descent

GD의 Loss 부터 전개해보자.

① ②

$\nabla L(\theta) = X^T \underbrace{(X\theta - y)}_{r := X\theta - y \text{ redusial}} = X^T r \rightarrow r^{t+1} = (I - \eta XX^T) r^t$ & $r^t = (I - \eta XX^T)^t r^0$

① gradient descent Update $\theta^{t+1} = \theta^t - \eta \nabla L(\theta^t) = \theta^t - \eta X^T r^t$

$X\theta^{t+1} = X\theta^t - \eta XX^T r^t$

$r = X\theta - y$ $\rightarrow r^{t+1} = X\theta^{t+1} - y = (X\theta^t - \eta XX^T r^t) - y = r^t - \eta XX^T r^t$

$\rightarrow X\theta^t = r^t + y$

$\boxed{r^{t+1} = r^t - \eta XX^T r^t}$

②

$t = 0$ : $r^1 = (I - \eta XX^T) r^0$

$t = 1$ : $r^2 = (I - \eta XX^T) r^1 = (I - \eta XX^T)^t \cdot r^0$

$t = 2$ : $r^3 = (I - \eta XX^T) r^2 = (I - \eta XX^T)^t \cdot r^0$

⋮

$\boxed{r^t = (I - \eta XX^T)^t \cdot r^0}$

뒷장

• GD iterate 정의 : $\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle x_i, \theta\rangle)^2 = \frac{1}{n}\| y - X\theta \|_2^2$

• GD update 식 전개 :

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla \hat{R}(\theta_t) = \theta_t = \frac{2\eta}{t} X^T(y - X\theta_t)$$

에서 초기값 $\theta_0 = 0$ 이라 가정하고 반복식 전개하면 Closed-Form 나옴.

$$\theta_t - \theta_0 = (\theta_t - \theta_{t-1}) + (\theta_{t-1} - \theta_{t-2}) + \cdots + (\theta_t - \theta_0)$$
$$= -\eta \sum_{k=0}^{t-1} (X^TX)^k X^T r_0 \quad (\text{where } r_0 = y)$$

여기서 SVD 사용하면 $\theta_t = -\eta \sum_{k=0}^{t-1} (X^TX)^k X^T r_0$,

$X = U\Sigma V^T$, $y = X\theta^* + z$

$$X^TX = V\Sigma^2V^T, \quad (X^TX)^k = V\Sigma^{2k}V^T \qquad \theta_t = -\eta \sum_{k=0}^{t-1} (X^TX)^k X^T r_0 = -\eta \sum_{k=0}^{t-1} V\Sigma^{2k}V^T \cdot V\Sigma U^T \cdot y$$
$$= -\eta \sum_{k=0}^{t-1} V\Sigma^{2k+1}U^Ty = -\eta \cdot V \cdot \left(\sum_{k=0}^{t-1} \Sigma^{2k+1}\right) \cdot U^Ty$$

$\Sigma$는 diagonal matrix → $\Sigma^{2k+1} = \mathrm{diag}(\cdots)$ ⟹ 각 singular vector 방향 $v_i$ 에 대해 독립적으로 계산 가능

$$\theta_t = V \cdot \sum_{k=0}^{t-1} \left[ \left( \sum_{k=0}^{t-1} \eta \cdot \sigma_i^{2k+1} \right) \langle u_i, y \rangle \right] v_i$$

등비수열

$$\sum_{k=0}^{t-1} \eta \sigma_i^{2k+1} = \sigma_i \cdot \eta \sum_{k=0}^{t-1} (\sigma_i^2)^k = \sigma_i \cdot \eta \cdot \frac{1 - (\sigma_i^2)^t}{1 - \sigma_i^2} \approx 1 - (1 - \eta\sigma_i^2)^t$$

attenuation factor

최종) $\theta_t = V \cdot \sum_{i=1}^{d} \underbrace{(1-(1-\eta\sigma_i^2)^t)}_{\text{반복 횟수에 따라 줄어드는 계수}} \cdot \underbrace{\langle u_i, y\rangle}_{\text{label vector } y\text{가 data 방향 } u_i\text{에 갖는 projection}} \cdot \frac{1}{\sigma_i} \cdot v_i$

반복 횟수에 따라 줄어드는 계수 (early stopping 핵심역할)

label vector y가 data 방향 $u_i$에 갖는 projection

데이터 크기 (singular value)에 대한 보정

## 10.1.3 Risk of gradient descent iterates.

우리가 진짜 관심 있는것 = expected risk. $R(\theta) = E[(y-\langle x,\theta\rangle)^2]$

그리고 평균적으로 parameter $\theta^* \sim N(0, I)$라고 가정할때 GD iterates $\theta_t$에 대한 risk = bias + Variance 가능.

$$E[R(\theta_t)] = \underbrace{\sum_{i=1}^{d}(1-\eta\sigma_i^2)^{2t}}_{bias^2} + \underbrace{\sigma^2\sum_{i=1}^{d}\frac{(1-(1-\eta\sigma_i^2)^t)^2}{\sigma_i^2}}_{variance}$$

$\sigma_i$ = singular values of $X$
$\eta$ = learning rate (step size)
$t$ = number of iterations

Interpretations :

1) 반복이 적으면 ($t \Downarrow$) : $(1-\eta\sigma_i^2)^t \approx 1 \Rightarrow$ bias 가 크고, variance 작다 (underfitting)
2) 반복이 많으면 ($t \Uparrow$) : $(1-\eta\sigma_i^2)^t \to 0 \Rightarrow$ bias 가 작고, variance 커진다 (overfitting)

∴ 전체 risk는 t에 대해 (반복횟수에 대해) U-shaped curve를 가지며,
적절한 t에서 멈추는 것이 중요 !!! = $\boxed{\text{early stopping}}$

## 10.1.4 Comparison to ridge regression estimator

· Ridge Regression $\hat{\theta}_{ridge} = (X^TX+\lambda I)^{-1}X^Ty$

SVD

$$\hat{\theta}_{ridge} = V\cdot diag\left(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\right)V^T\theta^* + V\cdot diag\left(\frac{\sigma_i}{\sigma_i^2+\lambda}\right)U^Tz$$

$\Rightarrow$ Risk :

$$E[R(\hat{\theta}_{ridge})] = \underbrace{\sum_{i=1}^{d}\left(1-\frac{\sigma_i^2}{\sigma_i^2+\lambda}\right)^2}_{bias^2} + \underbrace{\sigma^2\sum_{i=1}^{d}\left(\frac{\sigma_i}{\sigma_i^2+\lambda}\right)^2}_{variance}$$

생긴게 비슷하네!

| | < Ridge Regression > | < GD + Early Stopping > |
|---|---|---|
| · Bias.Variance Trade-off | 둘다 $bias^2$ + variance 형태 | |
| · SVD 표현 | 모두 $V_i$-basis 에서 각 방향마다 shrinkage factor 적용 | |
| · Shrinkage 효과 | 각 방향마다 적절히 weight를 줄여서 overfitting 방지 | |
| · High-frequency 방향 억제 | 작은 $\sigma_i$ 에서는 (noise가 심한 방향) weight 작아짐 | |
| · Control Parameter | $\lambda$ (explicit) | t (iteration 횟수, implicit) |
| · Shrinkage 방식 | $\frac{\sigma_i^2}{\sigma_i^2+\lambda}$, $\lambda \to 0$ 일때 OLS | $1-(1-\eta\sigma_i^2)^t$, $t\to\infty$ 일때 OLS |
| · Noise Sensitivity (variance) | $\left(\frac{\sigma_i}{\sigma_i^2+\lambda}\right)^2$ | $\frac{(1-(1-\eta\sigma_i^2)^t)^2}{\sigma_i^2}$ |
| · 학습 방식 | 단일 closed-form | 반복 최적화 알고리즘 필요 |