Experimental Design

find k points out of n that $\min_{|S| \le k} \overbrace{R(\hat{f}) - R(f^*)}$ eg. $f^* = \theta^{*T} x$

$\{x_{(1)} \ldots x_{(k)}\} = S$ $\qquad Y = \theta^{*T} x + \varepsilon$

label budget k

$$X_{n \times d} = \begin{bmatrix} \longrightarrow \\ \longrightarrow \\ \longrightarrow \end{bmatrix}$$

(n: pool, d: features; rows: features; the matrix is the pool of samples)

Regression $R(\hat{f}) - R(f^*) = E_{X,Y,D}[(Y - \hat{f}(x))^2] - \underbrace{E_{XY}[(Y - f^*(x))^2]}_{\sigma^2 \text{ Irreducible error}}$

(D ← data (training))

$$E[(Y - \hat{f}(x))^2] = E_{X,Y}[(Y - f^*(x))^2] + E_{X,D}[(f^*(x) - \hat{f}(x))^2] \qquad Y = f^*(x) + \varepsilon, \ \varepsilon \overset{iid}{\sim} N(0, \sigma^2)$$

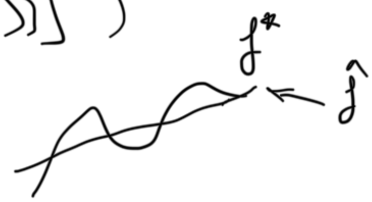
$$= E[(Y - f^*(x))^2] + E_x[(f^*(x) - E_D[\hat{f}(x)])^2]$$
$$+ E_{X,D}[(\hat{f}(x) - E_D[\hat{f}(x)])^2]$$

cross-terms vanished 1. $E[Y] = f^*(x)$ 2. $E[\hat{f}(x)] = E[\hat{f}(x)]$

$$= R(f^*) + \text{bias}^2 + \text{variance}$$

σ^2 irreducible

$$\left(2 E_X[(f^*(x) - E_D(\hat{f}(x))) E_D[\hat{f}(x) - E_D[\hat{f}(x)]]] \right) \quad (\to 0)$$



Bias can't be improved by sampling

$\Rightarrow$ ~~add~~ choice of sampling can only improve variance. ← measure of uncertainty

$\therefore$ Exp design (1-shot) & active learning (sequential) minimize variance.

Goal: How to choose $x_1 \ldots x_k$ to minimize $E_{X,D}[(\hat{f}(x) - E[\hat{f}(x)])^2]$? ($\hat{f}(x) = \hat{\theta}^T x$, $E[\hat{f}(x)] = \theta^{*T} x$)

$$= E[\langle \hat{\theta} - \theta^*, x \rangle^2]$$

$\hat{f}(x) = \hat{\theta}^T x \qquad \hat{\theta} = (X^T X)^{-1} X^T Y$ ← LS

$Y = \theta^{*T} X + \varepsilon$ $\quad X \in \begin{bmatrix} & & \\ & & \\ & d & \end{bmatrix}$ $\quad Y = \begin{bmatrix} \\ \\ \\ 1 \end{bmatrix}$

$$E[\hat{\theta}] = \theta^*$$

Uncertainty / variance of prediction

$$\rightarrow \quad E_D[\langle \hat{\theta} - \theta^*, x \rangle^2] = x^T \underbrace{E_D[(\hat{\theta} - \theta^*)(\hat{\theta} - \theta^*)^T]}_{cov(\hat{\theta})} x \quad \Leftarrow$$

$$cov(\hat{\theta}) = E[\, (X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1} \,]$$

$$= (X^T X)^{-1} X^T \sigma^2 I \, X (X^T X)^{-1}$$

$$= \sigma^2 (X^T X)^{-1} \quad \Leftarrow$$

$$\theta^* = (X^T X)^{-1} X^T X \theta^*$$

$$E[\varepsilon \varepsilon^T] = \sigma^2 I$$

A-optimality $\quad \min_{|S| \le k} var(\hat{\theta}_S) \quad S \subseteq D$

$$= \sigma^2 \, tr[(X_S^T X_S)^{-1}]$$

E/V-optimality $\quad \min_{|S| \le k} var(\langle \hat{\theta}_S, x \rangle)$ $\quad$ Transfer

$$= \sigma^2 \, x^T (X_S^T X_S)^{-1} x \quad \Leftarrow$$

Combinatorial opt / Intractable

$$\min_{|S| \le k} f(X_S^T X_S) \equiv f(\Sigma)$$

A-opt	$tr(\Sigma^{-1})$	Average	
E-opt	$\|\Sigma^{-1}\|_2$	Eigenvalue	$\max_{\|x\| \le 1} x^T M x = \lambda_{max}$
D-opt	$\det(\Sigma^{-1})$	Determinant	$\leftarrow$ Intractable $-\log\|\Sigma\|$ Convex
T-opt	$(tr(\Sigma))^{-1}$	Trace	
V-opt	$tr(X \Sigma^{-1} X^T)$	Variance (In-sample)	
G-opt	$\max(diag(X \Sigma^{-1} X^T))$	Global	

D-opt: $\max$ volume of region covered by chosen points $\leftarrow$

$=$

$X_S^T X_S = X^T W X$ $\quad W = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \end{bmatrix} = \text{diag}(w)$ $\quad w_i = \{0,1\}$

$n \times d$ $\quad \sum_i w_i \le k$

Continuous Relaxation $\quad w_i \in [0,1]$

$X^T W X$

$\rightarrow \min_{\substack{w_i \in [0,1] \\ \sum_i w_i \le k}} \text{tr}(X^T W X) = \sum_i w_i x_i x_i^T$

$\frac{w_i}{k} \equiv p_i \quad \sum_i \frac{w_i}{k} = \sum_i p_i \le 1$

Thm: $f\left(\sum_{i=1}^n w_i x_i x_i^T\right) \le \min_{|S| \le k} f(X_S^T X_S)$

$w_i \in [0,1]$ $\quad =$

How to choose subset $X_{\hat{S}}$ given $w_i \in [0,1]$?

Can we still get $f(X_{\hat{S}}^T X_{\hat{S}}) \le (1+\varepsilon) \min_{|S| \le k} f(X_S^T X_S)$?

→ Treat $\{\frac{w_i}{k}\}$ as p_i & sample with or without replacement k points

Thm: Approximation guarantee

	rough	A,V opt
$1+O(1)$	$k = \Omega(d)$	
$1+\varepsilon$	$k = \Omega\left(\frac{d \log d}{\varepsilon^2}\right)$ or $k = \Omega\left(\frac{d}{\varepsilon^2}\right)$	without replacement

More sophisticated rounding methods can get $1+\varepsilon$ approximation all for $k = \Omega\left(\frac{d}{\varepsilon}\right)$ for all optimality criteria.

Show for G-optimality $O(d^2)$ points suffice without replacement to get approximation ~~ratio~~ ratio of 1.

Fact: For every w s.t. $\sum_i w_i = 1$, $w_i \ge 0$

$\exists\, w'$ s.t. $\sum_i w_i' = 1$, $w_i \geq 0$ with $\|w'\|_0 = O(d^2)$.

s.t. $X^T W X = X^T w' X$

Caratheodory's Thm: If a point z lies in convex hull of set $\Omega \subseteq \mathbb{R}^m$ then z can be written as convex comb^n of atmost $m+1$ extremal points.

Let $\Omega = \{ \text{vec}(x_i x_i^T) \}_{i=1}^n \subseteq \mathbb{R}^{d^2}$

$z = \text{vec}\left(\sum_i w_i x_i x_i^T\right)$ lies in convex hull of Ω.

G-optimal procedure (with replacement)

Input: $X_{n \times d}$, budget k

1. Solve relaxed problem to get w.
2. Has $O(d^2)$ sparse solution w' with same objective as w.
3. Observe data pt. i $n_i = \lceil w_i k \rceil$ times (collect Y_i)
4. Compute $\hat{\theta} = (X^T X)^\dagger X^T Y$

Thm: W.p. $\geq 1-\delta$ $\forall\, x \in \mathcal{X}$

$$|\langle x, \hat{\theta} - \theta^* \rangle| \leq \sqrt{\|x^T (X^T X)^\dagger x\|} \sqrt{2\sigma^2 \log(2|\mathcal{X}|/\delta)}$$

$$= \sqrt{\frac{2d\sigma^2 \log(2|\mathcal{X}|/\delta)}{k}}.$$

Proof: $x^T(\hat{\theta} - \theta^*) = x^T (X^T X)^\dagger X^T \varepsilon =: z^T \varepsilon$ $\quad \varepsilon \sim N(0, \sigma^2 I)$

$$\sim N(0,\ \sigma^2 x^T (X^T X)^\dagger x)$$

$$\Pr(|x^T(\hat{\theta} - \theta^*)| > \varepsilon) \leq 2 e^{-\varepsilon^2 / 2\sigma^2 x^T (X^T X)^\dagger x}$$

$\leftarrow$ $\Pr(|\xi| > 0) \leq 2e^{-\varepsilon^2/2\sigma^2}$

Union bound $\forall\, x \in \mathcal{X}$ $\quad |\mathcal{X}|$

$P(A \cup B) \leq P(A) + P(B)$

wp $\geq 1-\delta$

$$|x^T(\hat{\theta} - \theta^*)| \leq \sqrt{2\sigma^2 \log \frac{2|\mathcal{X}|}{\delta}} \cdot \underbrace{\sqrt{x^T (X^T X)^{-1} x}}_{\sqrt{\frac{d}{k}}} \Rightarrow$$