Experimental Design

Not sequential, batch

How to choose a batch $x_1 \ldots x_k$ (k budget)

$$\min_{x_1 \ldots x_k} E[\|\hat{\mu} - \mu\|^2]$$

$\mu = E[x]$

$\hat{\mu}$ - reward estimate based on k pts.

Linear setting $\mu(x) = \theta^T x$

$\hat{\mu}(x) = \hat{\theta}^T x$, $\hat{\theta} = \underbrace{(X^T X)^{-1}}\underbrace{X^T y}$

$X_{k \times d}$ training point $y_{k \times 1}$

$$\min_{x_1 \ldots x_k} E[\|\hat{\theta}^T x - \theta^T x\|^2]$$

$$= \underbrace{\langle \hat{\theta} - \theta, x \rangle^2}$$

$var(\hat{\theta}^T x)$

"uncertainty"

$y = X\theta + \varepsilon$

if $y = \theta^T x + \varepsilon$

$E[\hat{\theta}]$

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

$= E[\cancel{(X^T X)^{-1}} \cancel{X^T X} \theta]$

$= \theta$ unbiased.

Note: exp design or selection of data points cannot reduce bias.

$$var(\hat{\theta}^T x) = E[\langle (X^T X)^{-1} X^T y - \theta, x \rangle^2]$$

$$= E[\langle (X^T X)^{-1} X^T (X\theta + \varepsilon) - \theta, x \rangle^2]$$

$$= E[\langle \cancel{\theta} + (X^T X)^{-1} X^T \varepsilon - \cancel{\theta}, x \rangle^2]$$

$\langle a, b \rangle^2 = b^T a a^T b$

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

$$= x^T (X^T X)^{-1} X^T \underbrace{E[\varepsilon \varepsilon^T]}_{\sigma^2 I} X (X^T X)^{-1} x$$

$\varepsilon \overset{iid}{\sim} N(0, \sigma^2 I)$

$\frac{d}{k} \rightarrow$

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

x - test pt.

X - training data matrix

variance of prediction $\hat{\theta}^T x$

- V-optimality $tr(X (X^T X)^{-1} X^T)$
- E-optimality $\lambda_{max}((X^T X)^{-1})$

$\underset{d \times 1}{x} \equiv X_{n \times d} =$

$$\max_{x : \|x\|_2 \leq 1} x^T (X^T X)^{-1} x =$$

$x^T M x$

G-optimality $\max(\mathrm{diag}(X(X^TX)^{-1}X^T))$ $= \lambda_{max}((X^TX)^{-1})$

T-optimality $tr((X^TX))^{-1}$

→ D-optimality $\det((X^TX)^{-1})$ ← Same as product of evals $\equiv$ max area of region spanned by chosen points

variance of parameters $\hat{\theta}$ { A-optimality $tr((X^TX)^{-1})$

$$\min_{X:|X|=k} f((X^TX)^{-1}) \equiv f(\Sigma)$$

Suppose given a pool of n points $\binom{n}{k}$ choices ⇒ Combinatorial NP-hard except D-opt.

$-\log\det$ convex

$\mathrm{var}(\hat{\theta}^T x) = \sigma^2 x^T (X^TX)^{-1} x$

$\mathrm{var}(\hat{\theta}) = \sigma^2 (X^TX)^{-1}$

if $X \overset{iid}{\sim} N(0, I)$ classical assumption

$X^TX \equiv I$

How to solve the A/V/D/T/E/G-optimality objectives?

$X - n \times d$

Original formulation $\min_{|S|\le k} f(X_S^TX_S) = \min f(X^TWX)$

$W = \mathrm{diag}(w)$

→ $w_i \in \{0,1\}$ NP-hard $W = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$

$\sum_i w_i = k$

Continuous relaxation
$$\min_{\substack{W=\mathrm{diag}(w) \\ w_i \in [0,1] \\ \sum_i w_i = k}} f(X^TWX)$$

need to do rounding on w_is

<u>Thm</u>: $\underset{w_i \in [0,1]}{f(X^TWX)} \le \min_{|S|\le k} f(X_S^TX_S)$

Different rounding techniques give different approximation

$$\min_{|S|\le k} f(X_S^TX_S) \le (1+\varepsilon) \underset{w_i \in [0,1]}{f(X^TWX)}$$

Consider no noise linear setting $y = x^T\theta \quad \theta \in \mathbb{R}^d$

$k \geq \underline{d}$

$O(d^2)$ or $O\left(\frac{d(d+1)}{2}\right)$ suffices to get approximation 1.

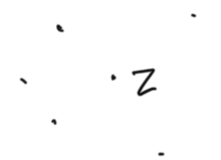
$$X^T W X = \sum_i w_i x_i x_i^T \leftarrow$$

For every w s.t. $\sum_i w_i = 1, w_i \geq 0$ $\quad \left(\sum_i w_i = k,\ w_i \in [0,1]\right)$

[w_i can be interpreted as prob of sampling a data point, when normalized by k]

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

Caratheodory's Theorem : if a point z lies in convex hull of a set $\Omega \in \mathbb{R}^m$ then z can be written as convex combination of atmost $m+1$ extremal pts.



$$\underbrace{X^T W X}_{} = \sum_i w_i \underbrace{x_i x_i^T}_{vec} = z_i \in \mathbb{R}^{d^2}$$

$d \times d$

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

G- optimal procedure (with replacement from pool) $\quad k \geq d^2$

Input : X $n \times d$, budget k on labels $k < n$

1. Solve relaxed continuous version of criterion to get w.
2. Has $O(d^2)$ sparse solution with same objective w
3. Observe data point i, $n_i = \lceil w_i k \rceil$ times to X_S
4. $\hat{\theta} = (X_S^T X_S)^{-1} X_S^T Y_S$ collect Y_S.

Improved results (in terms of budget) using more sophisticated rounding algos.

$1+\varepsilon$ approximation with $k = \Omega\left(\frac{d^2}{\varepsilon}\right)$ or $\Omega\left(\frac{d \log d}{\varepsilon}\right)$ for A, D, T, V, E, G

$x \in \mathcal{X}$ ... ε^2

using regret minimization.

$1+\varepsilon$ approximation with $k = \Omega\left(\frac{d}{\varepsilon}\right)$ using
{ Federov's algo. or Greedy algo. — for A/D opt.

Greedy – add 1 pt. at a time to maximize obj.

Federov – starts with k random points
swap one at time
two

Prediction variance $\sim d/k$ without assuming how X was obtained

Thm: wp $\geq 1-\delta$, $\forall x \in \mathcal{X}$ $\quad |\langle x, \hat{\theta}-\theta\rangle| \leq \sqrt{\|x^T(X_S^T X_S)^{-1} x\|} \cdot \sqrt{2\sigma^2 \log |\mathcal{X}|/\delta}$

if $x \in X$ $\quad |\langle x, \hat{\theta}-\theta\rangle| \leq \sqrt{\frac{d}{k}\sigma^2 \log \frac{|X|}{\delta}}$ $\leftarrow$

$\equiv$ var $\approx \frac{d}{k}$ upto log factors

~~Proof~~: $|\langle x, \hat{\theta}-\theta\rangle| \leq C$ $\qquad$ C – confidence / uncertainty

$\langle x, \hat{\theta}\rangle = \langle x, \theta\rangle \pm C$

Proof: $x^T(\hat{\theta}-\theta) = x^T\left((X_S^T X_S)^{-1} X_S^T Y_S - \theta\right) = x^T\left((X_S^T X_S)^{-1} X_S^T (X_S\theta + \varepsilon) - \theta\right)$

$= x^T (X_S^T X_S)^{-1} X_S^T \varepsilon \equiv Z^T \varepsilon \qquad \varepsilon \sim N(0, \sigma^2 I)$

$\sim N(0, \sigma^2 x^T (X_S^T X_S)^{-1} x)$

$P(|x^T(\hat{\theta}-\theta)| > \varepsilon) \leq 2e^{-\varepsilon^2 / 2\sigma^2 x^T (X_S^T X_S)^{-1} x} = \frac{\delta}{|\mathcal{X}|}$

Union bound $\forall x \in \mathcal{X}$

wp $\geq 1-\delta$ $\qquad |x^T(\hat{\theta}-\theta^*)| \leq \sqrt{2\sigma^2 \log \frac{2|\mathcal{X}|}{\delta}} \cdot \sqrt{\|x^T (X_S^T X_S)^{-1} x\|}$

$\sqrt{\frac{d}{k}}$

if $x \in X_{pool}$, $\|x^T (X_S^T X_S)^{-1} x\| \approx \frac{d}{k}$