Kiefer-Wolfowitz Thm &
G-optimal / D-optimal Designs.

Source: Chapter 21, Bandit Algorithms, L & S. (textbook)

Previously: Choose $\{A_s, s=1,2,\dots,t\}$ sequentially in $\mathbb{R}^d$ according to LinUCB rule. Develop $\{C_s, s=1,2,\dots,t\}$ confidence sets for $\{\hat{\theta}_s, s=1,\dots,t\}$ associated with the regularized least-square estimator.

A minmax problem — We are given a compact set $\mathcal{A} \subseteq \mathbb{R}^d$ (set could be finite/infinite). Let us define π to be a probability distribution over any finite subset of $\mathcal{A}$, denoted as $\mathrm{supp}(\pi)$. Let

$$g(\pi) = \max_{a \in \mathcal{A}} \|a\|^2_{V(\pi)^{-1}},$$

where $$V(\pi) = \sum_{a \in \mathcal{A}} \pi(a)\, a a^T.$$

Then, objective is to find a π^* s.t.

$$\boxed{\pi^* \in \operatorname*{argmin}_{\pi} g(\pi).}$$

(π^* is called as the G-optimal Design)

$\boxed{\text{A maximization problem}}$ — Same setting as above, with π a dist. over $\mathcal{A} \subseteq \mathbb{R}^d$.

Let $f(\pi) = \log \det(V(\pi))$.

Then, objective is to find $\hat{\pi}$ s.t.

$$\boxed{\hat{\pi} \in \underset{\pi}{\operatorname{argmax}}\ f(\pi)}$$

($\hat{\pi}$ is called as the <u>D-optimal Design</u>)

<u>Least Squares Summary</u>: Recall that:

$$Y_i = \langle A_i, \theta^* \rangle + \eta_i \quad , \quad i = 1, 2, \ldots, n$$

with $\{A_1, \ldots, A_n\} \in \mathbb{R}^d$, θ^* in $\mathbb{R}^d$, the following holds from Linear Least Squares Regression:

$$\underset{\text{design matrix}}{A} = \begin{pmatrix} A_1^T \\ A_2^T \\ \vdots \\ A_n^T \end{pmatrix} \begin{matrix} \uparrow \\ n \\ \downarrow \end{matrix} \quad \text{(width } d\text{)}$$

$$\underbrace{Y}_{\mathbb{R}^n} = A\theta^* + \underbrace{\eta}_{\mathbb{R}^n}, \qquad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}$$

$\{A_i\}_{i=1}^n$ spans $\mathbb{R}^d$.

Then, $\hat{\theta} = V^{-1} \sum_{i=1}^{n} A_i X_i$, unbiased estimate, where

$$V = \sum_{i=1}^{n} A_i A_i^T = (A^T A)$$

Further V^{-1} is the covariance matrix, i.e.,

$$E\left[(\hat{\theta}-\theta^*)(\hat{\theta}-\theta^*)^T\right] = V^{-1}$$

Further along any vector $x \in \mathbb{R}^d$, the variance is

$$E\left[\langle \hat{\theta}-\theta^*, x\rangle^2\right] = x^T V^{-1} x = \|x\|^2_{V^{-1}}$$

G-optimal Design

$$\pi^* \in \arg\min_{\pi} \max_{a \in A} \|a\|^2_{V(\pi)^{-1}}$$

Intuition: Equalize the variance along all "core" vectors in A.

D-optimal Design

$$\hat{\pi} \in \arg\max_{\pi} \log \det(V(\pi))$$

p.d., symmetric matrix

Intuition: $\det(V(\pi))$ is the product of eigenvalues. This scales inversely with the volume of the confidence ellipsoid around $(\hat{\theta}-\theta^*)$.

Ref: Optimal Design, Notes by Steven Buyske & Richard Trout, Rutgers University (Class notes for Stat. 591)

Geometry: Let π^* solve the problem above, and the corresponding $V(\pi^*) = \sum_{a \in \mathcal{A}} \pi^*(a)\, a a^T$.

Then, $\left(g(\pi^*), V(\pi^*)\right)$ defines the minimum volume centered ellipsoid containing $\mathcal{A}$. More concretely:

$$\text{Let} \quad \mathcal{E} = \left\{ x \in \mathbb{R}^d : \|x\|^2_{V^{-1}(\pi^*)} \le \underbrace{g(\pi^*)}_{=d} \right\}.$$

(we will show later).

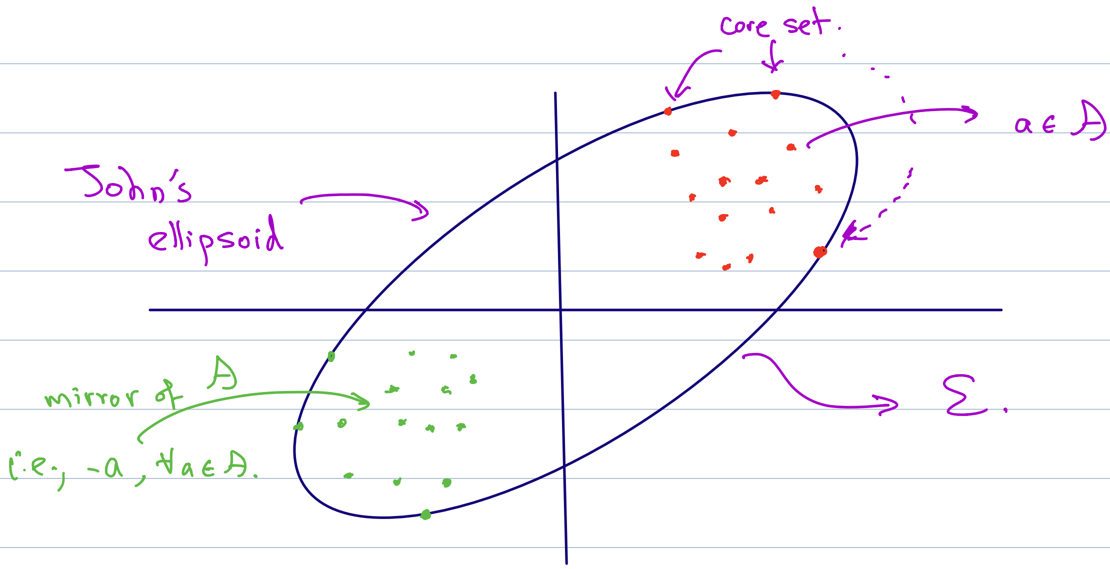

Figure 21.1 in Bandit Algorithms (textbook)

Then $\mathcal{E}$ is the minimum volume centered (at zero) ellipsoid containing the set $\mathcal{A}$.

Why is this useful? π^* a G-optimal Design for $\mathcal{A}$, and has finite support $\mathrm{supp}(\mathcal{A})$.

Now, consider the following design problem: We are allowed to pre-decide an integer $n \geq 1$, and a sequence of vectors $\{a_1, a_2, \cdots, a_n\}$, $a_i \in \mathcal{A}$. There is some unknown $\theta^* \in \mathbb{R}^d$, and we get the usual feedback, i.e., we observe:

$$X_i = \langle a_i, \theta^* \rangle + \eta_i \quad , \quad \eta_i \sim 1\text{-subG}.$$

We construct the least-square estimator $\hat{\theta}$, where

$$\hat{\theta} = V^{-1} \sum_{i=1}^{n} A_i X_i \quad , \quad \text{with} \quad V = \sum_{i=1}^{n} a_i a_i^T$$

Objective: Given (ε, δ), $\varepsilon > 0$, $\delta \in (0,1)$ and any $a \in \mathcal{A}$, we want

$$P\left(\langle \hat{\theta} - \theta^*, a \rangle \geq \varepsilon\right) \leq \delta$$

function of

$$\boxed{\text{Given } (\varepsilon, \delta) \text{, determine the smallest } n \text{, and associated } \{a_1, a_2, \cdots, a_n\} \text{ so that above bound holds.}}$$

The G-optimal Design provides an upper bound for n and a construction for $\{a_1, \cdots, a_n\}$. Textbook uses the term "accurate approximation".

<u>Construction</u>: π^* a G-optimal Design for $\mathcal{A}$, and $supp(\pi^*)$ finite (notation is slightly bad, as it should be $Supp_{\mathcal{A}}(\pi^*)$, as the support is a function of $\mathcal{A}$).

Let $$n_a = \left\lceil \frac{2\pi^*(a)\, g(\pi^*)}{\varepsilon^2} \ln\left(1/\delta\right) \right\rceil,$$

$$\forall\, a \in supp(\mathcal{A})$$

(Aside: Theorem below will show that $|supp(\mathcal{A})| \leq \frac{d(d+1)}{2}$)

Let $n = \sum_{a \in supp(\pi^*)} n_a$. We deterministically construct $\{a_1, a_2, \dots, a_n\}$, by replicating $a \in Supp(\pi^*)$ n_a times, $\forall\, a \in Supp(\pi^*)$.

Thus, $$V = \sum_{i=1}^{n} a_i a_i^T = \sum_{a \in Supp(\pi^*)} n_a\, a a^T$$

$$\geq \sum_{a \in Supp(\pi^*)} 2 \cdot \frac{\pi^*(a)\, g(\pi^*)}{\varepsilon^2} \ln(1/\delta) \cdot a a^T$$

$$= \frac{2 g(\pi^*)}{\varepsilon^2} \ln(1/\delta)\, V(\pi^*) \longrightarrow (**)$$

Recall that from Set 15, for any $a \in \mathbb{R}^d$, and $\delta \in (0,1)$,

$$\mathbb{P}\left(\langle \hat{\theta} - \theta^*, a \rangle \geq \sqrt{2 \|a\|^2_{V^{-1}} \ln(1/\delta)} \right) \leq \delta .$$

From $(**)$, we have for any $a \in \mathcal{A}$,

$$\|a\|^2_{V^{-1}} = a^T V^{-1} a \leq \underbrace{\left(a^T V(\pi^*)^{-1} a \right)}_{\leq g(\pi^*)} \frac{\varepsilon^2}{2 g(\pi^*) \ln(1/\delta)} .$$

$$\leq \frac{\varepsilon^2}{2 \ln(1/\delta)} .$$

$$\Rightarrow \sqrt{2 \ln(1/\delta) \|a\|^2_{V^{-1}}} \geq \varepsilon$$

$$\therefore \mathbb{P}\left(\langle \hat{\theta} - \theta^*, a \rangle \geq \varepsilon \right) \leq \delta$$

Further: $$n = \sum_{a \in \mathrm{Supp}(\pi^*)} n_a = \sum_{a \in \mathrm{Supp}(\pi^*)} \left\lceil \frac{2 \pi^*(a) g(\pi^*) \ln(1/\delta)}{\varepsilon^2} \right\rceil$$

$$\leq \underbrace{|\mathrm{Supp}(\pi^*)|}_{} + \frac{2 g(\pi^*)}{\varepsilon^2} \ln(1/\delta) .$$

core set $= \{ a \in \mathcal{A} : a \in \mathrm{Supp}(\pi^*) \}$

<u>Kiefer-Wolfowitz Thm (21.1 in text)</u>: Suppose $\mathcal{A} \subseteq \mathbb{R}^d$, is a compact set with $\mathrm{Span}(\mathcal{A}) = \mathbb{R}^d$. Then,

(i) $\pi^* \in \underset{\pi}{\mathrm{argmin}}\, g(\pi)$ $\iff$ (iii) $g(\pi^*) = d$.

$\Updownarrow$

(ii) $\pi^* \in \underset{\pi}{\mathrm{argmax}}\, \log \det(V(\pi))$

Further $\exists\, \pi^*$ as above with $\mathrm{Supp}(\pi^*) \leq \frac{d(d+1)}{2}$

<u>Remark:</u> ① $\pi^* \in \underset{\pi}{\mathrm{argmin}}\, g(\pi) \longrightarrow$ G-Optimal Design

② $\pi^* \in \underset{\pi}{\mathrm{argmax}}\, \log \det(V(\pi)) \longrightarrow$ D-Optimal Design.

Above Thm shows these are equivalent (essentially by establishing that the problems are duals of each other).

<u>Proof:</u> (Spcl case, where $|\mathcal{A}|$ is finite). Thus, $\pi \in \mathbb{R}^{|\mathcal{A}|}$, a finite dim. vector over $\mathcal{A}$. Recall:

$$g(\pi) = \max_{a \in \mathcal{A}} \|a\|^2_{V(\pi)^{-1}} \qquad f(\pi) = \log \det V(\pi).$$

Facts: f is concave, and $\underbrace{\nabla f(\pi)}_{\in \mathbb{R}^{|A|}} = \begin{pmatrix} \vdots \\ \|a\|^2_{V(\pi)^{-1}} \\ \vdots \end{pmatrix} \updownarrow |A|$

Also, for any π, we have

$$\sum_{a \in A} \pi(a) \|a\|^2_{V(\pi)^{-1}} = \sum_{a \in A} \pi(a) \cdot a^T V(\pi)^{-1} a$$

$$= \text{trace}\Big(\underbrace{\sum_a \pi(a)\, a a^T}_{V(\pi)} V(\pi)^{-1} \Big)$$

$$= \text{trace}(I) = d.$$

$$\therefore \quad g(\pi) = \max_{a \in A} \|a\|^2_{V(\pi)^{-1}} \underset{\max \geq \text{avg}}{\geq} \sum_{a \in A} \pi(a) \|a\|^2_{V(\pi)^{-1}} = d.$$

$\longrightarrow$ I

ⓐ. $\pi^* \in \arg\max_{\pi} \underbrace{\log \det V(\pi)}_{f(\pi)} \Longrightarrow \pi^* \in \arg\min_{\pi} \underbrace{\max_{a \in A} \|a\|^2_{V(\pi)^{-1}}}_{g(\pi).}$

f concave. From first-order optimality conditions, for any π, we have

$$0 \geq \langle \nabla f(\pi^*), \pi - \pi^* \rangle$$

$$= \sum_{a \in \mathcal{A}} \pi(a) \cdot \|a\|^2_{V(\pi^*)^{-1}} - \underbrace{\sum_{a \in \mathcal{A}} \pi^*(a) \|a\|^2_{V(\pi^*)^{-1}}}_{= d \quad (\text{from } \textcircled{I})}.$$

Now choose any $\tilde{a} \in \mathcal{A}$, and let
$\pi(\tilde{a}) = 1$, $\pi(a) = 0 \quad \forall a \neq \tilde{a}$.

Then, $0 \geq \|\tilde{a}\|^2_{V(\pi^*)^{-1}} - d$

$$\implies \|\tilde{a}\|^2_{V(\pi^*)^{-1}} \leq d \quad \text{for any } \tilde{a} \in \mathcal{A}$$

$$\implies \max_{a \in \mathcal{A}} \|a\|^2_{V(\pi^*)^{-1}} \leq d$$

$$\implies g(\pi^*) \leq d$$

Also, $g(\pi) \geq d \quad \forall \pi$ (from $\textcircled{I}$).

$\implies \pi^*$ is a minimizer for $g(\pi)$, and
$g(\pi^*) = d$.

(b) $g(\pi^*) = d \implies \pi^* \in \underset{\pi}{\arg\max}\, f(\pi)$.

We have $g(\pi^*) = \max_{a \in \mathcal{A}} \|a\|^2_{V(\pi^*)^{-1}} = d$

Now observe that:

$$< \nabla f(\pi^*), \pi - \pi^* >$$

$$= \sum_{a \in A} \pi(a) \cdot \underbrace{\|a\|^2_{V(\pi^*)^{-1}}}_{\leq d \ \forall a \in A} - \underbrace{\sum_{a \in A} \pi^*(a) \|a\|^2_{V(\pi^*)^{-1}}}_{= d.}$$

$$\because \max_{a \in A} \|a\|^2_{V(\pi^*)^{-1}} = d.$$

$$\leq 0$$

f concave, $< \nabla f(\pi^*), \pi - \pi^* > \leq 0$

$\Longrightarrow$ π^* is the maximizer of f.

(c) $\pi^* \in \arg\min_{\pi} \max_{a \in A} \|a\|^2_{V(\pi)^{-1}} \Longrightarrow g(\pi^*) = d.$

Recall $g(\pi) \geq d \quad \forall \pi \quad$ (from (I)).

We now demonstrate a $\hat{\pi}$ s.t. $g(\hat{\pi}) = d$. Thus, it immediately follows that $g(\pi^*) = d$.

Let $\hat{\pi}$ be the maximizer of $f(\cdot)$, i.e.,

$$\hat{\pi} \in \arg\max_{\pi} \log \det V(\pi).$$

From (a), we note $g(\hat{\pi}) = d$. Thus, we are done.

For showing that $\exists\ \pi^*$ s.t. $\text{Supp}(\pi^*) \leq \frac{d(d+1)}{2}$,

read text, Pf of Thm 21.1 $\blacksquare$