21 Optimal Design for Least Squares Estimators

In the preceeding chapters we introduced the linear bandit and showed how to construct confidence intervals for least squares estimators. We now study the problem of choosing actions for which these confidence intervals are small, which plays an important role in the analysis of stochastic linear bandits with finitely many arms (Chapter 22) or adversarial linear bandits (Part VI).

Let $\eta_1, \dots, \eta_n$ be a sequence of independent 1-subgaussian random variables and $A_1, \dots, A_n \in \mathbb{R}^d$ be a fixed sequence with $\text{span}(A_1, \dots, A_n) = \mathbb{R}^d$ and $Y_1, \dots, Y_n$ be given by $Y_t = \langle A_t, \theta_* \rangle + \eta_t$ for some $\theta_* \in \mathbb{R}^d$. Recall from the previous chapter that the least square estimator is $\hat{\theta} = V^{-1} \sum_{t=1}^n A_t Y_t$ with $V = \sum_{t=1}^n A_t A_t^\top$ the design matrix.

Unlike in previous chapters the least squares estimators used here are not regularized. This eases the calculations and the lack of regularization will not harm us in future applications.

For this choice we showed that for any $a \in \mathbb{R}^d$ it holds that

$$\mathbb{P}\left(\langle \hat{\theta} - \theta_*, a \rangle \geq \sqrt{2\|a\|_{V^{-1}}^2 \log\left(\frac{1}{\delta}\right)}\right) \leq \delta\,. \tag{21.1}$$

For our purposes, both A_t and x will usually be actions from some (possibly infinite) set $\mathcal{A} \subset \mathbb{R}^d$ and the question of interest is finding the shortest sequence of exploratory actions $A_1, \dots, A_n$ such that the confidence bound in the previous display is smaller than some threshold for all $a \in \mathcal{A}$. To solve this exactly is likely an intractable exercise in integer programming. But a highly accurate approximation turns out to be efficient for a broad class of action sets. Let $\pi : \mathcal{A} \to [0, 1]$ be a distribution on $\mathcal{A}$ so that $\sum_{a \in \mathcal{A}} \pi(a) = 1$ and $V(\pi) \in \mathbb{R}^{d \times d}$ and $g(\pi) \in \mathbb{R}$ be given by

$$V(\pi) = \sum_{a \in \mathcal{A}} \pi(a) a a^\top\,, \qquad g(\pi) = \max_{a \in \mathcal{A}} \|a\|_{V(\pi)^{-1}}^2\,. \tag{21.2}$$

In the subfield of statistics called optimal experimental design, the distribution π is called a **design** and the problem of finding the π that minimizes g is called the ***G*-optimal design problem**. So how to use this? Suppose that π is a design

and $a \in \mathrm{Supp}(\pi)$ and

$$n_a = \left\lceil \frac{\pi(a) g(\pi)}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) \right\rceil . \tag{21.3}$$

Then choosing each action $a \in \mathrm{Supp}(\pi)$ exactly n_a times is enough to ensure that

$$V = \sum_{a \in \mathrm{Supp}(\pi)} n_a a a^\top \geq \frac{g(\pi)}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) V(\pi) ,$$

which by Eq. (21.1) means that for any $a \in \mathcal{A}$, with probability $1 - \delta$,

$$\langle \hat{\theta} - \theta_*, a \rangle \leq \sqrt{\|a\|_{V^{-1}}^2 \log\left(\frac{1}{\delta}\right)} \leq \varepsilon .$$

By Eq. (21.3) the total number of actions required to ensure a confidence width of no more than ε is bounded by

$$n = \sum_{a \in \mathrm{Supp}(\pi)} n_a = \sum_{a \in \mathrm{Supp}(\pi)} \left\lceil \frac{\pi(a) g(\pi)}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) \right\rceil \leq |\,\mathrm{Supp}(\pi)| + \frac{g(\pi)}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) .$$

So how big are $|\,\mathrm{Supp}(\pi)|$ and $g(\pi)$? As we will now show, there exists a π^* that minimizes $g(\pi)$ such that $g(\pi) = d$ and $|\,\mathrm{Supp}(\pi)| \leq d(d+3)/2$. The first of these facts follows from the following theorem, while the latter is explained in Section 21.2.

THEOREM 21.1 (Kiefer–Wolfowitz) *The following are equivalent:*

1 π^* *is a minimizer of* g.
2 π^* *is a minimizer of* $f(\pi) = -\log\det V(\pi)$.
3 $g(\pi^*) = d$.

The theorem shows that G-optimal design is equivalent to the ***D*-optimal design problem** (D for 'determinant'), which is the objective in item (2) and (as we shall soon see) has a useful geometric interpretation.

21.1 Proof of Kiefer–Wolfowitz (†)

We follow the original proof by Kiefer and Wolfowitz [1960], which is direct and relies only on elementary linear algebra, convexity and calculus. Nevertheless, this section is not core to the rest of the book and could be skipped on a first pass. To begin we note that since no matter how the exploration distribution π is chosen it holds that $\sum_a \pi(a)\|a\|_{V(\pi)^{-1}}^2 = d$. Hence for all π there exists an $a \in \mathcal{A}$ such that $\|a\|_{V(\pi)^{-1}}^2 \geq d$. The proof will follow by showing that if π maximizes $\log\det(V(\pi))$, then $\|a\|_{V(\pi)^{-1}}^2 \leq d$ for all $a \in \mathcal{A}$. Suppose that π is a distribution

such that $\det V(\pi) > 0$ and

$$\frac{\partial}{\partial \alpha} \log\det V((1-\alpha)\pi + \alpha\pi')\Big|_{\alpha=0} \le 0 \qquad \text{for all distributions } \pi'. \tag{21.4}$$

Let us momentarily fix an alternative distribution π' and let A be a matrix such that $AV(\pi)A^\top = I$ and $AV(\pi')A^\top = B$ where B is diagonal with elements $b_1, \dots, b_d$ (such a matrix exists by simultaneous diagonalization. Then

$$\det V((1-\alpha)\pi + \alpha\pi') = \frac{\prod_{i=1}^d (1-\alpha+\alpha b_i)}{(\det A)^2}.$$

By noting that the sum of concave functions is concave and checking that $\log(1-\alpha+\alpha b_i)$ is concave it follows that $\log\det V((1-\alpha)\pi + \alpha\pi')$ is concave in $\alpha \in [0,1]$. It follows that for π with $\det V(\pi) > 0$ and satisfying Eq. (21.4) that $\pi = \operatorname{argmax}_\pi \log\det(V(\pi))$. The next step is a direct calculation of the derivative in Eq. (21.4) (details see Exercise 21.1):

$$\frac{\partial}{\partial \alpha} \log\det V((1-\alpha)\pi + \alpha\pi')\Big|_{\alpha=0} = \sum_{ij} V(\pi)^{-1}_{ij} V(\pi')_{ij} - d. \tag{21.5}$$

By letting $\pi'(a) = 1$ for some $a \in \mathcal{A}$ we have $\|a\|^2_{V(\pi)^{-1}} = \sum_{ij} V(\pi)^{-1}_{ij} V(\pi')_{ij} \le d$.

21.2 Minimum volume ellipsoids and John's theorem (†)

This section depends on a little background on convex optimization and especially the notion of duality. The classic reference is by Boyd and Vandenberghe [2004, Chap 5]. Let S^d_{++} be the space of (symmetric) positive definite matrices and recall that a d-dimensional ellipsoid is determined by its center $x_\circ \in \mathbb{R}^d$ and a positive definite matrix $H \in S^d_{++}$ and defined by $E(x_\circ, H) = \{x \in \mathbb{R}^d : \|x - x_\circ\|_{H^{-1}} \le 1\}$. Given a closed convex set $\mathcal{K} \subset \mathbb{R}^d$ it is a problem in convex geometry to find the ellipsoid E of smallest volume such that $\mathcal{K} \subseteq E$. Such an ellipsoid is called the **minimum-volume enclosing ellipsoid** (MVEE). The volume of an ellipsoid is easily evaluated by noting that if $L = \sqrt{H}$, then $\mathrm{vol}(E(x_\circ, H)) = \mathrm{vol}(E(0,H))$ and $E(0,H) = LB^d_2$ where $B^d_2 = \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$ is the unit ball and $LB^d_2 = \{Lx : x \in B^d_2\} \subset \mathbb{R}^d$. Therefore

$$\mathrm{vol}(E(x_\circ, H)) = \mathrm{vol}(B^d_2)\det(L) = \mathrm{vol}(B^d_2)\sqrt{\det(H)}.$$

To make the connection to optimal design we consider a modification of the problem of finding the MVEE by adding the restriction that the ellipsoid must be centered ($x_\circ = 0$), which is written as the following convex optimization problem:

$$\min_{H \in S^d_{++}} \log\det(H)$$
$$\text{subject to } \mathcal{K} \subseteq E(0,H).$$

If $\mathcal{K} = \mathrm{co}(\mathcal{A})$ is the convex hull of $\mathcal{A}$, then the dual of this problem is equivalent to the D-optimal design problem where the Lagrange multipliers play the role of the design π. The dual is

$$\max \log\det\left(\sum_{a\in\mathcal{A}} \lambda(a) a a^\top\right) - \sum_{a\in\mathcal{A}} \lambda(a) + d$$
$$\text{subject to } \lambda(a) \geq 0 \text{ for all } a \in \mathcal{A}\,. \tag{21.6}$$

As it happens this is one situation where strong duality holds, so the optimization problems are essentially equivalent. By introducing $\pi(a) = \lambda(a)/\sum_{a'\in\mathcal{A}} \lambda(a')$ it is easy to check (again by duality) that the above is equivalent to

$$\max \log\det\left(\sum_{a\in\mathcal{A}} \pi(a) a a^\top\right) + d\log d$$
$$\text{subject to } \pi \text{ being a distribution on } \mathcal{A}\,.$$

Of course, the $d\log d$ term does not depend on π, so this optimization problem is now equivalent to the D-optimal design problem that appeared in Theorem 21.1. Fritz John's celebrated result concerns the properties of the MVEE with no restriction on the center.

THEOREM 21.2 (John's theorem) *Let $\mathcal{K} \subset \mathbb{R}^d$ be convex, closed and assume that* $\mathrm{span}(\mathcal{K}) = \mathbb{R}^d$. *Then there exists a unique MVEE of $\mathcal{K}$. Furthermore, this MVEE is the unit ball B_2^d if and only if there exists $m \leq d(d+3)/2$ contact points ("the core set") $u_1, \ldots, u_m$ that belong to both $\mathcal{K}$ and the surface of B_2^d and there also exist positive reals $c_1, \ldots, c_m$ such that*

$$\sum_i c_i u_i = 0 \quad \text{and} \quad \sum_i c_i u_i u_i^\top = I, \tag{21.7}$$

To apply John's theorem we first massage the action set so that the MVEE provided by the theorem is centered, but without affecting the optimal design. Let $\mathcal{A}' = \{a : a \in \mathcal{A} \text{ or } -a \in \mathcal{A}\}$ and $\mathcal{K} = \mathrm{co}(\mathcal{A}')$ be the convex hull of $\mathcal{A}'$. Now take $E = E(x_\circ, H)$ to be the MVEE of $\mathcal{K}$, which by construction is centered so that $x_\circ = 0$. If $L = \sqrt{H}$, then the image of E under L^{-1} is B_2^d, which is the MVEE of convex set $L^{-1}\mathcal{K}$. Therefore by John's theorem there exists $u_1, \ldots, u_m \in L^{-1}\mathcal{K} \cap \partial B_2^d$ and positive reals $c_1, \ldots, c_m$ such that Eq. (21.7) holds. In fact, by the curvature of the ellipse we have $u_i \in L^{-1}\mathcal{A}' \cap \partial B_2^d$. Since the trace of a matrix is invariant under rotation,

$$d = \mathrm{trace}\left(\sum_i c_i u_i u_i^\top\right) = \sum_i c_i \,\mathrm{trace}(u_i u_i^\top) = \sum_i c_i\,.$$

This allows us to take

$$\pi(a) = \frac{1}{d}\sum_i c_i \mathbb{I}\left\{Lu_i = a \vee (Lu_i = -a \wedge -a \notin \mathcal{A})\right\}\,,$$

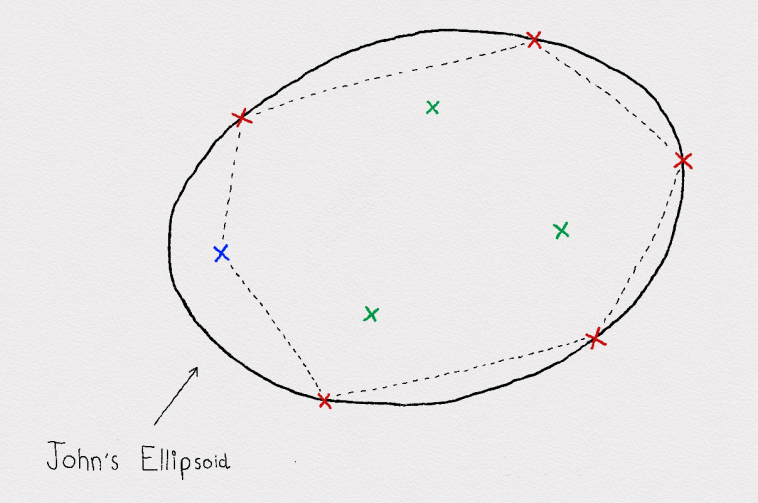

Figure 21.1 John's ellipsoid for the convex polytope (dashed line) over a small action set. The core set is marked in red. Actions on the boundary of the polytope (and not the core set) are blue, while the green actions are called interior points.

where the complicated expression is due to the fact that a and $-a$ might sometimes both be in $\mathcal{A}$. Therefore $V(\pi) = LL^\top/d$ and so

$$\sup_{a\in\mathcal{A}} \|a\|^2_{V(\pi)^{-1}} \le \sup_{u:\|u\|_2=1} \|L^{-1}(a)\|^2_{V(\pi)^{-1}} = d \sup_{u:\|u\|_2=1} u^\top L^{-1} LL^\top (L^\top)^{-1} u = d\,.$$

None of this is terribly surprising in light of Kiefer–Wolfowitz theorem, but John's theorem also provides a guarantee on the size of the core set, which means that the support of the G-optimal design π can be assumed to have cardinality at most $d(d+3)/2$.

21.3 Notes

1 In no applications will we require an exact solution to the design problem. In fact, finding a distribution π such that $g(\pi) \le (1+\varepsilon)g(\pi^*)$ will increase the regret of our algorithms by a factor of just $(1+\varepsilon)^{1/2}$.

2 When the action set is finite the computation of the optimal design is a convex problem for which there are numerous efficient approximation algorithms. The Franke-Wolfe algorithm is one such algorithm, which is known as Wynn's method in optimal experimental design and can be used to find a near-optimal solution for modestly sized problems. More sophisticated methods have also been investigated. A good place to start is: Vandenberghe et al. [1998]. If the action set is infinite, then the optimal design can often still be approximated efficiently. The most notable case is when there exists an efficient algorithm (a 'membership oracle') for the function $\mathbb{I}\{x \in \mathcal{K}\}$ for any $x \in \mathbb{R}^d$. For details on

this (and many other interesting algorithms involving convexity) see the book by Grötschel et al. [2012].

3 While the proof of Theorem 21.1 is sufficiently elementary to be included here, we do not know of a simple proof of John's theorem. Perhaps a reason for the additional difficulty is that John's proof implicitly shows that the cardinality of the core set is at most $d(d+3)/2$, which is not revealed at all by Kiefer–Wolfowitz. A proof of John's theorem may be found in the short book by Ball [1997].

21.4 Bibliographic remarks

According to our best knowledge, the connection to optimal experimental design through the Kiefer-Wolfowitz theorem and the proof that solely relied on this result has not been pointed out in the literature beforehand, though the connection between the Kiefer-Wolfowitz theorem and MVEEs is well known. Besides the previously mentioned book by Boyd and Vandenberghe [2004] there is also a recent book by Todd [2016] that discusses algorithmic issues as well as the duality. The theorem of Kiefer and Wolfowitz is due to them: Kiefer and Wolfowitz [1960]. John's theorem is due to John [1948]. The duality mentioned in the text was proved by Silvey and Sibson [1972].

21.5 Exercises

21.1 Prove the correctness of the derivative in Eq. (21.5).

Use the fact that the inverse of matrix A is $A^{-1} = M/\det(A)$ where M is the matrix of cofactors of A.

21.2 Find John's ellipsoid for each of the following sets and use it to derive the G-optimal design.

(a) The simplex: $\mathcal{K} = \mathrm{co}\{e_1, \dots, e_d\}$.
(b) The hypercube: $\mathcal{K} = \{x : \|x\|_\infty \le 1\}$.

21.3 Write a program that accepts as parameters a finite set $\mathcal{A} \subset \mathbb{R}^d$ and returns the G-optimal design $\pi : \mathcal{A} \to [0,1]$ that minimizes $g(\pi)$ given in Eq. (21.2).

The easiest 'pure' way to do this is to implement the Franke-Wolfe algorithm (see the notes). For more robust results we suggest a convex optimization library be used.