Course: Topics in Geometric Algorithms	Aug-Dec 2024
Instructors: Arindam Khan and Siddharth Barman	Indian Institute of Science
Presenter: Nirjhar Das	
Topic: John's Theorem and Optimal Design	

1 Preliminaries

We start with some basic definitions and two useful theorems in the study of convexity. In what follows, we will restrict ourselves to $\mathbb{R}^n$, $n \in \mathbb{N}$, the n-dimensional Euclidean space. The space is equipped with the Euclidean norm: $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$, where $x = [x_1, \ldots, x_n]^\intercal \in \mathbb{R}^n$ is any vector. A few fundamental objects of the space $\mathbb{R}^n$ are:

1. (Unit) Ball: $B_n \coloneqq \{x \in \mathbb{R}^n : \|x\| \leq 1\}$
2. (Unit) Sphere: $S^{n-1} \coloneqq \{x \in \mathbb{R}^n : \|x\| = 1\}$
3. Hyperplane: Given a vector $v \in \mathbb{R}^n$ and a scalar $t \in \mathbb{R}$, hyperplane is given by $H \coloneqq \{x \in \mathbb{R}^n : \langle x, v\rangle = t\}$
4. Half-space: For a hyperplane defined by $v \in \mathbb{R}^n$ and $t \in \mathbb{R}$, there are two half-spaces associated with it: $H_+ \coloneqq \{x \in \mathbb{R}^n : \langle x, v\rangle \geq t\}$ and $H_- \coloneqq \{x \in \mathbb{R}^n : \langle x, v\rangle \leq t\}$.

Remark 1.1. *We note that the sphere and the hyperplane are $(n-1)$-dimensional objects embedded in n-dimension. This is because the equality allows for only $n-1$ degrees of freedom.*

Now, let us recall that a set $S \subseteq \mathbb{R}^n$ is called convex if for any $x, y \in S$, we have $\lambda x + (1-\lambda)x \in S$, where $\lambda \in [0,1]$. An example of a convex set is the ball B_n itself. Similarly, a hyperplane as well as half-spaces are convex. However, the sphere S^{n-1} is not convex. An extremely useful concept in the study of convex geometry is the convex hull. The convex hull of a set $S \subseteq \mathbb{R}^n$, denoted by $conv(S)$ is the "smallest" convex set that contains S. Formally, $conv(S) = \cap\{A : \ A \text{ convex and } A \supseteq S\}$. Sometimes, it is more useful to write the alternate algebraic definition of convex hull: $conv(S) = \{\sum_{i=1}^k \lambda_i a_i : k \in \mathbb{N}, \ \forall\ i \in [k], a_i \in S, \lambda_i \in [0,1] \text{ and } \sum_{i=1}^k \lambda_i = 1\}$. The primary object of our study will be the convex body which is defined as follows:

Definition 1.1 (Convex Body). *A convex body is compact convex set with non-empty interior.*

Let us unpack the definition. A compact set in $\mathbb{R}^n$ is a set that is both closed and bounded[1]. Closed means that the limit point of a convergent sequence in the set is also contained in the set. Intuitively, it means that the set also includes the boundary, e.g. $[0,1]$ is closed but $(0,1)$ is not. Bounded means that the set is contained in a ball of radius $R > 0$, for some finite R, i.e., the set is contained in $RB_n \coloneqq \{x \in \mathbb{R}^n : \|x\| \leq R\}$. Next, we give some more definitions that will be useful in developing the later results:

1. Symmetric Convex Body: A convex body $K \subset \mathbb{R}^n$ such that if $x \in K$ then $-x \in K$.
2. Dilation of a convex body K: $\alpha K \coloneqq \{\alpha x : x \in K\}$, $\alpha \in \mathbb{R}$.
3. Translation of a convex body K: $b + K \coloneqq \{b + x : x \in K\}$, $b \in \mathbb{R}^n$.
4. Linear Transformation of a convex body K: $TK \coloneqq \{Tx : x \in K\}$, $K \in \mathbb{R}^{n \times n}$, T is invertible.
5. Ellipsoid: $\mathcal{E} = b + TB_n = \{b + x \in \mathbb{R}^n : \langle x, T^{-2}x\rangle \leq 1\}$, T is positive definite and $b \in \mathbb{R}^n$. We call b the center of the ellipsoid.

From the definition of a symmetric convex body, it is clear that $0 \in K$ if K is a symmetric convex body. Lastly, an ellipsoid is defined as a linear transformation of the unit ball followed by a translation.

Now, we state two fundamental results of convex geometry. The proofs of these theorems can be found in any standard references on convex analysis [Bertsekas et al., 2003, Rockafellar, 1970].

[1]Heine-Borel theorem shows this is a iff criterion in $\mathbb{R}^n$

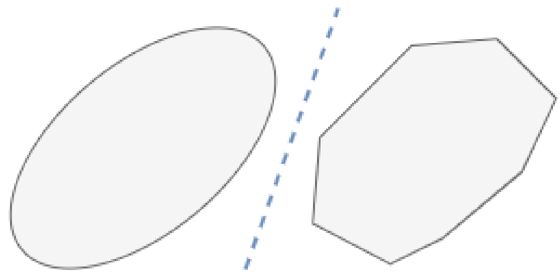

Figure 1: The blue hyperplane separates the two sets shown by shaded regions. Such a hyperplane need not be unique.

Theorem 1.1 (Carathéodory's Theorem)**.** *Let $A \subset \mathbb{R}^n$ and $x \in conv(A) \subset \mathbb{R}^n$ be any point in the convex hull of A. Then, x can be written as a convex combination of at most $n+1$ points in A. Specifically, there exists $\lambda_i \geq 0$, $a_i \in A$, $1 \leq i \leq n+1$, $\sum_{i=1}^{n+1} \lambda_i = 1$, such that*

$$x = \sum_{i=1}^{n+1} \lambda_i a_i$$

A particularly useful corollary of Carathéodory's theorem is when the set A is a finite collection of points, i.e., $A = \{a_1, \ldots, a_m\}$. Then, the theorem states that any point in the convex hull of A can be represented as a convex sum of at most $n+1$ points in A. The theorem is often used to prove the existence of small sized solutions to problems.

Next, we present another fundamental theorem of convex analysis.

Theorem 1.2 (Separating Hyperplane Theorem)**.** *Let A and B be two disjoint non-empty convex subsets of $\mathbb{R}^n$. Then there exists a hyperplane given by vector $v \neq 0$ and real $c \in \mathbb{R}$, such that,*

$$\langle x, v \rangle \leq c \leq \langle y, v \rangle, \quad \forall\ x \in A,\ y \in B$$

In the next section, we present a fundamental result of convex geometry called John's theorem. The theorem characterizes the geometry of convex bodies using ellipsoids that are contained in them. In section 3, we will see how studying this problem gives us useful results in designing small sized coresets.

2 Maximum Volume Ellipsoid

In Section 1, we have seen the definition of a convex body and an ellipsoid. The volume of a convex body K is the n-dimensional Lebesgue measure associated with the set K. The Lebesgue measure in define via intervals: the measure of a set $S = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n]$ is given by $Vol(S) = \prod_{i=1}^{n}(b_i - a_i)$. Further, by writing any convex body as a (countable) union of sets of like S (called box sets), we can define the volume of the convex body. Particularly, for an ellipsoid $\mathcal{E} = b + TB_n$, where T is a positive definite (p.d.) matrix and $b \in \mathbb{R}^n$, $Vol(\mathcal{E}) = \det(T)Vol(B_n)$.

Lemma 2.1. *Given a convex body K, the maximum volume ellipsoid contained in K exists and is unique.*

Proof. **Existence**: First we argue the existence of such an ellipsoid. We define $\mathcal{F} := \{(b, T) : b \in \mathbb{R}^n,\ T \text{ is p.d. } \in \mathbb{R}^{n \times n} \text{ s.t. } b + TB_n \subseteq K\}$. The set $\mathcal{F}$ is just the set of parameters that define the ellipsoids. The volume of the ellipsoids defined by elements of $\mathcal{F}$ can be thought of as the map $(b, T) \mapsto det(T)$, which is a continuous map. Therefore, if we can show that $\mathcal{F}$ is compact, then we have by Extreme Value Theorem [Rudin et al., 1964, Theorem 4.16], that there exists a $(b^*, T^*) \in \mathcal{F}$ such that the corresponding ellipsoid is of maximum volume.

Since the convex body K is bounded (by definition), there exists $R > 0$ such that $K \subseteq RB_n$. Hence, for all (b,T), $b + TB_n \subseteq RB_n$. Particularly, $b \in RB_n$, hence $\|b\| \le R$. Also, for any $x \in B_n$, $\|Tx\| = \|Tx + b - b\| \le \|Tx + b\| + \|b\| \le 2R$. Thus, $\|T\|_{op} \le 2R$. Hence, both the set of feasible b and feasible T are bounded. Next, we show the closedness of these sets.

Let us take any point k in the interior of K. We know that, by definition of interior, there exists an $\varepsilon > 0$ such that $\{y : \|k - y\| \le \varepsilon\} = k + \varepsilon B_n \subset K$. Specifically, $(k, \varepsilon \mathbf{I}_n) \in \mathcal{F}$. The volume of this ellipsoid is $\varepsilon^n Vol(B_n)$. We can now define $\mathcal{F}_\varepsilon = \{(b,T) \in \mathcal{F} : Vol(b + TB_n) \ge \varepsilon^n Vol(B_n)\}$. This set contains all ellipsoids of volume at least $\varepsilon^n Vol(B_n)$. Moreover, this set is non-empty. We will show closedness of this set. Let $\{(b_m, T_m)\}_m \subset \mathcal{F}_\varepsilon$ be a sequence that converges to (b,T), that is, $\|b_m - b\| \to 0$ and $\|T_m - T\|_{op} \to 0$.

We have for any $x \in B_n$, $\|(b_m + T_m x) - (b + Tx)\| \le \|b_m - b\| + \|T_m - T\|_{op}\|x\| \to 0$. Thus, the sets $b_m + T_m B_n$ converges pointwise to $b + TB_n$. Hence, $b + TB_n \subset K$ because otherwise, there exists $x' \in B_n$ such that $b + Tx' \notin K$. In other words, there exists an $\varepsilon' > 0$ such that the ε'-neighbourhood of $b + Tx'$ does not contain any point of K. However, $b_m + T_m x' \in K$ for all $m \in \mathbb{N}$, and we know by pointwise convergence, there exists an $M \in \mathbb{N}$ such that for all $m > M$, $\|(b_m + T_m x') - (b + Tx')\| < \varepsilon'$. This is a contradiction. Therefore, we must have $b + TB_n \subset K$. Hence, $(b,T) \in \mathcal{F}$.

Finally, we have by convergence of T_m to T[2] and by continuity of $\det(\cdot)$, $\det(T_m) \to \det(T)$. Thus, $\det(T) \ge \varepsilon^n$. Hence, $(b,T) \in \mathcal{F}_\varepsilon$. Now, we can conclude that $\mathcal{F}_\varepsilon$ is closed.

By closedness and boundedness of $\mathcal{F}_\varepsilon$, we have that $\mathcal{F}_\varepsilon$ is compact. Thus, the continuous function $(b,T) \mapsto \det(T)$ achieves the maximum value over this set. Moreover, since $\mathcal{F}_\varepsilon$ contains all ellipsoids of volume at least $\varepsilon^n Vol(B_n)$, we have the all other ellipsoids contained in K are of volume smaller than this quantity. Thus, the maximum over these two sets must be achieved on $\mathcal{F}_\varepsilon$. This proves the existence of the maximum volume ellipsoid.

Uniqueness: Now, we show the uniqueness of the maximum volume ellipsoid. Without loss of generality, assume that $\mathcal{E}_1 = B_n$ and $\mathcal{E}_2 = b + TB_n$ are the two different ellipsoids of maximum volume contained in K. From $Vol(\mathcal{E}_1) = Vol(\mathcal{E}_2)$, we have $\det(T) = 1$. Now consider the ellipsoid $\mathcal{E}_3 = \frac{b}{2} + \frac{\mathbf{I}_n + T}{2} B_n$. For any $x \in B_n$, $\frac{b}{2} + \frac{\mathbf{I}_n + T}{2} x = \frac{1}{2}x + \frac{1}{2}(b + Tx) \in K$ because $\mathcal{E}_1, \mathcal{E}_2 \subseteq K$ and K is convex. Thus, $\mathcal{E}_3 \subseteq K$. Moreover, $Vol(\mathcal{E}_3) = \det\left(\frac{\mathbf{I}_n + T}{2}\right) Vol(B_n) = \det\left(\frac{\mathbf{I}_n + T}{2}\right) Vol(\mathcal{E}_1)$. Thus, by maximality of $\mathcal{E}_1$, $\det\left(\frac{\mathbf{I}_n + T}{2}\right) \le 1$. Now, we have, by Minkowski's determinant inequality [Horn and Johnson, 2012, Theorem 7.8.21],

$$\det\left(\frac{\mathbf{I}_n + T}{2}\right)^{\frac{1}{n}} \ge \det\left(\frac{\mathbf{I}_n}{2}\right)^{\frac{1}{n}} + \det\left(\frac{T}{2}\right)^{\frac{1}{n}} = \frac{1}{2} + \frac{1}{2} = 1$$

Hence, $\det\left(\frac{\mathbf{I}_n + T}{2}\right) = 1$ which implies that $T = \mathbf{I}_n$ (because of equality condition of the Minkowski inequality). This in turn gives that $\mathcal{E}_3 = \frac{b}{2} + B_n$. Moreover, $b \neq 0$ because $\mathcal{E}_1 \neq \mathcal{E}_2$ by assumption. Consider the ellipsoid $\mathcal{E}'_3 = \frac{b}{2} + (\mathbf{I}_n + \delta bb^\intercal)B_n$, where $\delta > 0$. We have $Vol(\mathcal{E}'_3) = (1 + \delta)Vol(B_n)$. Hence, $\mathcal{E}'_3$ has a bigger volume than the maximum volume ellipsoids. Lastly, let us take any point in the ellipsoid $\mathcal{E}'_3$: $\frac{b}{2} + (\mathbf{I}_n + \delta bb^\intercal)x = x + (\frac{1}{2} + \delta b^\intercal x)b$, where $x \in B_n$. Note that $x \in K$ and $x + b \in K$, thus by convexity of K, $x + \lambda b \in K$ for all $\lambda \in [0,1]$. Further, we have $-\|b\| \le b^\intercal x \le \|b\|$. Consequently, by choosing $\delta = \frac{1}{4\|b\|}$, we have $\frac{1}{2} + \delta b^\intercal x \in [1/4, 3/4]$. This implies that, $x + (\frac{1}{2} + \delta b^\intercal x)b \in K$. Hence, $\mathcal{E}'_3 \subseteq K$. This a contradiction because this contained ellipsoid has volume larger than the maximum. Thus we conclude that the maximum volume ellipsoid must be unique. □

Now, we present John's theorem on the maximum volume ellipsoid contained in a convex body K. This theorem characterizes the points of contact between the maximum volume ellipsoid and the convex body. The theorem is formally stated below. John proved the implication of (i) to (ii) while Ball proved that the converse is also true.

Theorem 2.1 (John's Theorem with Ball's strengthening)**.** *Given a convex body K, the following are equivalent:*

(i) B_n is the maximum volume ellipsoid contained in K

[2] convergence in operator norm implies element-wise convergence, which in turn implies convergence with respect to any continuous function. See here for example.

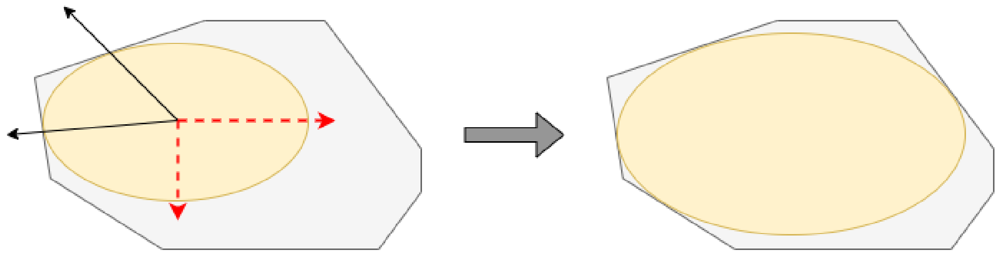

Figure 2: If the contact points are not well spread in all the directions, then it is possible to increase the ellipsoid in those directions in which contact points are not present. The red arrows indicate that we can increase the ellipsoid to find an ellipsoid of even bigger volume.

(ii) $B_n \subseteq K$ *and* $\exists\{u_1, u_2, \ldots, u_m\} \subset \partial K \cap \partial B_n$ *and* $\{c_1, \ldots, c_m\} \subset \mathbb{R}_+$ *such that (a)* $\sum_{i=1}^m c_i u_i = 0$ *and (b)* $\sum_{i=1}^m c_i u_i u_i^\intercal = \mathbf{I}_n$*. Moreover,* $m = O(n^2)$*.*

The theorem says that the contact points of the maximum volume ellipsoid with the convex body, when the ellipsoid is B_n, is spread in all directions well. Further, we can have a small subset of the contact points such that this condition holds. Note that if the contact points were not spread in all directions, then it would be possible to extend the ellipsoid in the directions where the contact points are not spread. Figure 2 shows this phenomenon.

Before proving John's theorem, let us look at an immediate corollary.

Corollary 2.1. *Given a convex body* K*, if* $\mathcal{E}$ *is the maximum volume enclosed ellipsoid, then* $\mathcal{E} \subseteq K \subseteq n\mathcal{E}$*. Further, if* K *is symmetric,* $\mathcal{E} \subseteq K \subseteq \sqrt{n}\mathcal{E}$*.*

To prove the corollary, we make an observation that will be useful.

Observation 2.1. *If* $u \in S^{n-1}$*, then the hyperplane at* u *that has all of* B_n *on one side is given by* $H = \{x \in \mathbb{R}^n : \langle x, v\rangle = 1\}$*. Moreover,* H *is the only such hyperplane.*

Proof. First we show that all of B_n lies in the same halfspace of H. For any $x \in B_n$, $\langle x, u\rangle \leq \|x\| \cdot \|u\| \leq 1$, where the first inequality is by Cauchy-Schwarz. Thus, all of B_n lies in the halfspace $H_- = \{x \in \mathbb{R}^n : \langle x, u\rangle \leq 1\}$. Now suppose there exists $v \in S^{n-1}$, $v \neq \pm u$, and $t \in \mathbb{R}$, such that the hyperplane $H' = \{x \in \mathbb{R}^n : \langle x, v\rangle = t\}$ passes through u and has B_n on one side. Therefore, $\langle u, v\rangle = t$. Thus, $H' = \{x \in \mathbb{R}^n : \langle x, v\rangle = \langle u, v\rangle\}$. Since $v \in S^{n-1} \subset B_n$, we must have v lies in one of the halfspaces of H'. Now, $\langle v, v\rangle = 1 > \langle u, v\rangle$, the inequality being (strict) again due to Cauchy-Schwarz. Thus, v lies in the halfspace $H'_+ = \{x \in \mathbb{R}^n : \langle x, v\rangle \geq \langle u, v\rangle\}$. However, for $-v \in B_n$, we have $\langle -v, v\rangle = -1 < \langle u, v\rangle$. Hence, $-v \notin H'_+$. Therefore, B_n does not lie on one side H'. This a contradiction and hence, no such v exists. Hence, the hyperplane at u that has B_n on one side is unique. □

The hyperplane mentioned in the above observation is called the supporting hyperplane at that point. Now, we are ready to give the proof of the corollary.

Proof. Without loss of generality, let B_n be the maximum volume enclosed ellipsoid of K (otherwise apply on K the inverse of the transformation that maps B_n to $\mathcal{E}$). Further, let the contact points be $\{u_1, \ldots, u_m\}$ and their corresponding coefficients from John's theorem be $\{c_1, \ldots, c_m\}$. Then, since u_i's lie on S^{n-1}, we have $\|u_i\| = 1$. Thus,

$$n = \operatorname{Tr}(\mathbf{I}_n) = \operatorname{Tr}\left(\sum_{i=1}^m c_i u_i u_i^\intercal\right) = \sum_{i=1}^m c_i \operatorname{Tr}(u_i u_i^\intercal) = \sum_{i=1}^m c_i \operatorname{Tr}(u_i^\intercal u_i) = \sum_{i=1}^m c_i \cdot 1 = \sum_{i=1}^m c_i$$

Note that the supporting hyperplane of K at u_i is also a supporting hyperplane of B_n at u_i, because all of K lies on one side of this hyperplane, and since $B_n \subseteq K$, all of B_n lies on one side of this hyperplane. However, by the above observation, we have that the unique supporting hyperplane of B_n at u_i is $H = \{x \in$

$\mathbb{R}^n : \langle x, u_i \rangle \leq 1\}$. Thus, H is also the unique supporting hyperplane of K at u_i. Hence, we have, for all $x \in K$, $\langle x, u_i \rangle \leq 1$. Now we can write:

$$
\begin{aligned}
\|x\|^2 = x^\intercal \mathbf{I}_n x &= x^\intercal \left(\sum_{i=1}^m c_i u_i u_i^\intercal \right) x && \text{(John's theorem)} \\
&= \sum_{i=1}^m c_i (x^\intercal u_i)(u_i^\intercal x) \\
&= \sum_{i=1}^m c_i \langle x, u_i \rangle^2 \\
&= \sum_{i=1}^m c_i \langle x, u_i \rangle^2 - 2\sum_{i=1}^m c_i \langle u_i, x \rangle + \sum_{i=1}^m c_i - \sum_{i=1}^m c_i && (\textstyle\sum_{i=1}^m c_i u_i = 0 \text{ by John's theorem}) \\
&= \sum_{i=1}^m c_i \left(\langle x, u_i \rangle - 1 \right)^2 - \sum_{i=1}^m c_i \\
&\leq \sum_{i=1}^m c_i (1 - \langle x, u_i \rangle) \cdot \max_{i \in [m]} (1 - \langle x, u_i \rangle) - \sum_{i=1}^m c_i && (\langle x, u_i \rangle \leq 1) \\
&= \left(\max_{i \in [m]} 1 - \langle x, u_i \rangle \right) \cdot \left(\sum_{i=1}^m c_i - \sum_{i=1}^m c_i \langle x, u_i \rangle \right) - \sum_{i=1}^m c_i \\
&= \left(\max_{i \in [m]} 1 - \langle x, u_i \rangle \right) \cdot \sum_{i=1}^m c_i - \sum_{i=1}^m c_i && (\textstyle\sum_{i=1}^m c_i u_i = 0) \\
&= \left(\max_{i \in [m]} -\langle x, u_i \rangle \right) \cdot \sum_{i=1}^m c_i \\
&\leq \|x\| n && (\text{Cauchy-Schwarz; } \textstyle\sum_{i=1}^m c_i = n)
\end{aligned}
$$

Therefore, $\|x\|^2 \leq \|x\| \cdot n$ which gives us $\|x\| \leq n$, for all $x \in K$. Thus, $K \subseteq nB_n$.

For the symmetric case, we have that $-x \in K$ whenever $x \in K$. Thus, $\langle -x, u_i \rangle \leq 1$ and $\langle x, u_i \rangle \leq 1$. Combining, we have $|\langle x, u_i \rangle| \leq 1$. Thus, $\|x\|^2 = \sum_{i=1}^m c_i \langle x, u_i \rangle^2 \leq \sum_{i=1}^m c_i \cdot 1 = n$. Hence, $\|x\| \leq \sqrt{n}$. Thus, $K \subseteq \sqrt{n} B_n$. □

The corollary implies that if the maximum volume ellipsoid is scaled up by at most n, the resulting ellipsoid will contain the convex body entirely. Indeed, the factor n is tight and an example of such a convex body is the $n-1$-regular simplex (the convex hull of n points that are pairwise equidistant)[3]. A symmetric body for which the $\sqrt{n}$ factor is tight is the unit cube. Now, give the proof of John's theorem.

Proof. We will give the proof for the general case. The symmetric case requires less algebraic manipulations because for a symmetric convex body, if u is a contact point between convex body K and its max volume ellipsoid, then so is $-u$. Therefore, the coefficient of u, c can be reduced to $c/2$ and the coefficient of $-u$ can be made $c/2$. This preserves the property (b) of the contact points while in the LHS of property (a) has one less summand. Repeating this argument, we see that property (a) is trivial for symmetric convex bodies. We therefore need to consider only property (b) for the symmetric case. In what follows, we describe the proof of John's theorem in full generality.

(ii) $\implies$ (i): This is the easier direction is to prove. We will start with this. The idea is show that volume of any ellipsoid contained in K is at most the volume of B_n. For this, let $\mathcal{E} = b + TB_n$ be any ellipsoid contained in K. For any $1 \leq i \leq m$, $b + Tu_i \in \mathcal{E}$. Moreover, since $H = \{x \in \mathbb{R}^n : \langle u_i, x \rangle \leq 1\}$ is the supporting hyperplane of K (see proof of corollarly 2.1) at u_i, and $\mathcal{E} \subseteq K$, we have $\langle u_i, b + Tu_i \rangle \leq 1$.

[3]This can be embedded in n dimension as follows: $v_i = [0, \ldots, 1, \ldots, 0] \in \mathbb{R}^n$, where the i-th entry is 1 and rest are 0's, $1 \leq i \leq n$. Then $(n-1)$-regular simplex is the convex hull of v_i's. See here for a calculation of ratio of radii of the inscribed and circumscribed hyperspheres.

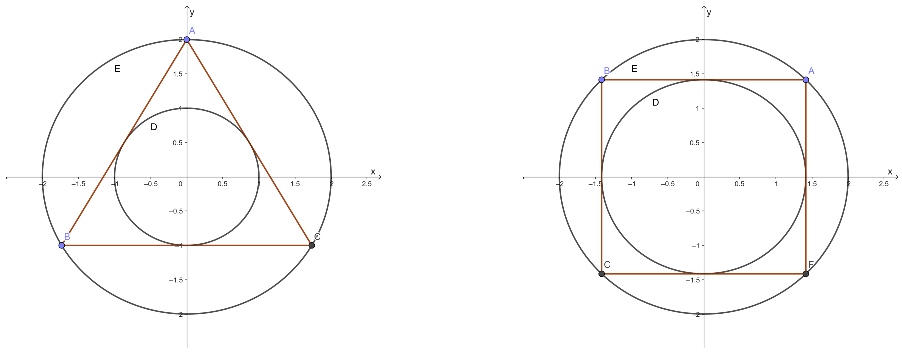

Figure 3: **Left**: The maximum volume contained ellipsoid for a triangle is its inscribed circle (in $\mathbb{R}^2$). A scaling of the incircle by $n = 2$ gives the circumcircle in the equilateral triangle case. Therefore the scaling is tight in this case. Moreover, in high dimensions, the generalization of the equilateral triangle is the regular simplex, for which the scaling is indeed tight. **Right**: The maximum volume contained ellipsoid of the square, when scaled by $\sqrt{n} = \sqrt{2}$ gives the circumcircle of the square, which contains it. Note that the square is a symmetric convex body. Moreover, the generalization of the square in high dimensions, called the hypercube, also exhibits the tight scaling of the factor $\sqrt{n}$ (i.e., a factor less than $\sqrt{n}$ does not suffice).

Therefore,

$$\begin{aligned} n = \sum_{i=1}^{m} c_i \geq \sum_{i=1}^{m} c_i \langle u_i, b + Tu_i \rangle &= \sum_{i=1}^{m} c_i \langle u_i, b \rangle + \sum_{i=1}^{m} c_i \langle u_i, Tu_i \rangle \\ &= 0 + \sum_{i=1}^{m} c_i \operatorname{Tr}\left(\langle u_i, Tu_i \rangle\right) && (\textstyle\sum_{i=1}^{m} c_i u_i = 0) \\ &= \sum_{i=1}^{m} c_i \operatorname{Tr}\left(u_i u_i^\intercal T\right) \\ &= \operatorname{Tr}\left((\sum_{i=1}^{m} c_i u_i u_i^\intercal) T \right) \\ &= \operatorname{Tr}(\mathbf{I}_n T) = \operatorname{Tr}(T) \end{aligned}$$

The volume of the ellipsoid $\mathcal{E}$ is $\det(T) Vol(B_n)$. Now, by AM-GM inequality $\det(T) = \prod_{i=1}^{n} \lambda_i \leq \left(\frac{\sum_{i=1}^{n} \lambda_i}{n} \right)^n = \left(\frac{\operatorname{Tr}(T)}{n} \right)^n \leq 1$, where λ_i's are eigenvalues of T. Hence, $Vol(\mathcal{E}) \leq Vol(B_n)$. Thus, B_n must be maximum volume ellipsoid contained in K.

(i) $\Longrightarrow$ (ii): We consider the following $\mathbb{R}^{n^2+n}$-dimensional space: $\{[vec(xx^\intercal), x] : x \in \mathbb{R}^n\}$, where $vec(M)$ is the vector created by stacking the rows of the matrix one after the other to get a long vector. We also use the notation $\langle X, Y \rangle$ to denote $\sum_{i=1}^{n} \sum_{j=1}^{n} X_{ij} Y_{ij}$ for two matrices $X, Y \in \mathbb{R}^{n \times n}$.

To prove this implication, we consider the convex hull $C = conv(\{[vec(uu^\intercal), u] : u \in \partial B_n \cap \partial K\})$. Our goal will be to show that $[vec(\frac{1}{n}\mathbf{I}_n), 0_n] \in C$, where 0_n is n-dimensional zero-vector. We do this via contradiction.

Suppose $[vec(\frac{1}{n}\mathbf{I}_n), 0_n] \notin C$. Thus, by the Separating Hyperplane Theorem (Theorem 1.2), there exists a vector $[vec(H), h]$ in $\mathbb{R}^{n^2+n}$ (where $H \in \mathbb{R}^{n \times n}$ and $h \in \mathbb{R}^n$) such that for all $p \in C$, $\langle [vec(H), h], p \rangle > \alpha > \langle [vec(H), h], [vec(\frac{1}{n}\mathbf{I}_n), 0_n] \rangle = \langle H, \frac{1}{n}\mathbf{I}_n \rangle + \langle h, 0 \rangle = \langle H, \frac{1}{n}\mathbf{I}_n \rangle$. Particularly,

$$\langle H, uu^\intercal \rangle + \langle h, u \rangle > \alpha > \langle H, \frac{1}{n}\mathbf{I}_n \rangle \quad \forall\ u \in \partial B_n \cap \partial K$$

Since H is operated on symmetric matrices, we can replace H with its symmetric version $(H + H^\intercal)/2$ which still preserves the separation property. Thus without loss of generality, let H be symmetric.

Further, we see that $\langle H+\sigma\mathbf{I}_n, uu^\intercal\rangle = \langle H, uu^\intercal\rangle + \sigma\|u\|^2 = \langle H, uu^\intercal\rangle + \sigma$, since $u \in \partial B_n = S^{n-1}$. Similarly, $\langle H+\sigma\mathbf{I}_n, \mathbf{I}_n\rangle = \langle H, \mathbf{I}_n\rangle + \sigma$. Thus, we can replace H with $H+\sigma\mathbf{I}_n$ for any σ. Particularly, we can choose σ such that $H+\sigma\mathbf{I}_n$ has trace equal to 0. Hereafter, without loss of generality let $\mathrm{Tr}(H) = 0$.

Now, we consider the ellipsoid $\mathcal{E}_\delta = -\frac{\delta}{2}(\mathbf{I}_n+\delta H)^{-1}h + (\mathbf{I}_n+\delta H)^{-\frac{1}{2}}B_n$. For small enough $\delta > 0$, the matrix $\mathbf{I}_n + \delta H$ will be positive definite, and therefore a true ellipsoid. The volume of this ellipsoid is

$$Vol(\mathcal{E}) = \det\left((\mathbf{I}_n+\delta H)^{-\frac{1}{2}}\right) Vol(B_n) = \frac{Vol(B_n)}{\sqrt{\det(\mathbf{I}_n+\delta H)}} > \frac{Vol(B_n)}{\sqrt{(\mathrm{Tr}(\mathbf{I}_n+\delta H)/n)^n}} = Vol(B_n)$$

where the inequality is via AM-GM, with strict inequality because H has at least one positive and one negative eigenvalue because H is non-zero whereas its trace is 0, thus $\mathbf{I}_n + \delta H$ cannot have same eigenvalues. Moreover, in the last equality, we have used the fact that $\mathrm{Tr}(\mathbf{I}_n+\delta H) = \mathrm{Tr}(\mathbf{I}_n) + \delta\,\mathrm{Tr}(H) = n$.

Note that the ellipsoid $\mathcal{E}_\delta$ can be alternately expressed as

$$\mathcal{E}_\delta = \{x \in \mathbb{R}^n : \left(x + \frac{\delta}{2}(\mathbf{I}_n+\delta H)^{-1}h\right)^\intercal (\mathbf{I}_n+\delta H)\left(x + \frac{\delta}{2}(\mathbf{I}_n+\delta H)^{-1}h\right) \leq 1\}$$

Setting $x = u$, where $u \in \partial B_n \cap \partial K$,

$$\begin{aligned}
&\left(u + \frac{\delta}{2}(\mathbf{I}_n+\delta H)^{-1}h\right)^\intercal (\mathbf{I}_n+\delta H)\left(u + \frac{\delta}{2}(\mathbf{I}_n+\delta H)^{-1}h\right) && (1)\\
&= \langle u, \mathbf{I}_n u\rangle + \delta\left(\langle u, Hu\rangle + \langle h, u\rangle\right) + \frac{\delta^2}{4}\langle h, (\mathbf{I}_n+\delta H)^{-1}h\rangle \\
&= \|u\|^2 + \delta\left(\langle H, uu^\intercal\rangle + \langle h, u\rangle\right) + \frac{\delta^2}{4}\langle h, (\mathbf{I}_n+\delta H)^{-1}h\rangle \\
&\geq \|u\|^2 + \delta\left(\langle H, uu^\intercal\rangle + \langle h, u\rangle\right) && (\mathbf{I}_n+\delta H \text{ is p.d.})\\
&= 1 + \delta\left(\langle H, uu^\intercal\rangle + \langle h, u\rangle\right) && (u \in \partial B_n)\\
&> 1 + \alpha && (\langle H, uu^\intercal\rangle + \langle h, u\rangle > \alpha)
\end{aligned}$$

Therefore, u lies outside $\mathcal{E}_\delta$. Note that as $\delta \downarrow 0$, $\mathcal{E}_\delta \to B_n$. Since ∂K is compact and the set $\partial K \cap \mathcal{E}$ gradually decreases to $\partial B_n \cap \partial K$, there exists a small enough δ such that the whole of $\mathcal{E}_\delta$ is inside K. Formally, we make the following arguments.

Claim 2.1. *There exists an $\varepsilon > 0$ such that $-\frac{1}{2}(H + \frac{1}{\varepsilon}\mathbf{I}_n)^{-1}h$ lies on S^{n-1} and $H + \frac{1}{\varepsilon}\mathbf{I}_n$ is positive definite.*

The claim can be easily proved by observing that the norm of the vector is given as a function of ε as $f(\varepsilon) = \frac{\varepsilon^2}{4}h^\intercal(\mathbf{I}_n+\varepsilon H)^{-2}h$. As $\varepsilon \downarrow 0$, we have $f(\varepsilon) \downarrow 0$, and when $\varepsilon \uparrow 1/|\lambda_{min}(H)|$ (note that $\lambda_{min}(H)$ is negative), $f(\varepsilon) \to \infty$. Therefore, by continuity of $f(\varepsilon)$, there exists a ε' such that $f(\varepsilon') = 1$. Moreover, since $\varepsilon' < 1/|\lambda_{min}(H)|$, $H + \frac{1}{\varepsilon'}\mathbf{I}_n$ is positive definite.

Using the claim above and Lemma 2.1 of [Hager, 2001], we have that $\min_{x\in S^{n-1}}\langle H, xx^\intercal\rangle + \langle h, x\rangle = -\frac{1}{4}\langle h, (\frac{1}{\varepsilon}\mathbf{I}_n + H)^{-1}(\frac{2}{\varepsilon}\mathbf{I}_n + H)(\frac{1}{\varepsilon}\mathbf{I}_n + H)^{-1}h\rangle < 0$ since $\mathbf{I}_n + \varepsilon H$ is positive definite.

Now, let w be a vector such that $\langle H, ww^\intercal\rangle + \langle h, w\rangle < 0$. Thus, with $u \in \partial B_n \cap \partial K$,

$$\begin{aligned}
0 > \langle H, ww^\intercal\rangle + \langle h, w\rangle &= \langle H, uu^\intercal\rangle + \langle h, u\rangle + \langle H, u(w-u)^\intercal\rangle + \langle H, (w-u)u^\intercal\rangle + \langle h, w-u\rangle \\
&> \alpha - \|w-u\| \cdot (2\|H\|_{op} + \|h\|)
\end{aligned}$$

$$\text{which implies that,}\quad \|w-u\| > \frac{\alpha}{2\|H\|_{op} + \|h\|}$$

Therefore, $w \in V \coloneqq \{x \in S^{n-1} : \forall\, u \in \partial B_n \cap \partial K, \|x-u\| \geq \frac{\alpha}{2\|H\|_{op}+\|h\|}\}$. Moreover, this also shows that $m \coloneqq \min_{x\in V}\langle H, xx^\intercal\rangle + \langle h, x\rangle < 0$. The set V is the set of points on S^{n-1} that are "far" from the contact points. We will show that all the points in ∂K which are obtained by extending the vectors in V to intersect ∂K are actually outside $\mathcal{E}_\delta$.

For a vector $v \in V$, let $t > 0$ be such that $tv \in \partial K$. Note that $t > 1$ because $B_n \subseteq K$ and $t \notin \partial B_n \cap \partial K$. Therefore,

$$\begin{aligned}
&\left(tv + \frac{\delta}{2}(\mathbf{I}_n + \delta H)^{-1}h\right)^\intercal (\mathbf{I}_n + \delta H) \left(tv + \frac{\delta}{2}(\mathbf{I}_n + \delta H)^{-1}h\right) && (2)\\
&= t^2\langle \mathbf{I}_n + \delta H, vv^\intercal\rangle + \delta\langle h, tv\rangle + \frac{\delta^2}{4}\langle h, (\mathbf{I}_n + \delta H)^{-1}h\rangle \\
&> t\langle \mathbf{I}_n + \delta H, vv^\intercal\rangle + t\delta\langle h, v\rangle && (\mathbf{I}_n + \delta H \text{ is p.d. and } t > 1)\\
&\geq t\|v\|^2 + \delta t\left(\langle H, vv^\intercal\rangle + \langle h, v\rangle\right) \\
&\geq t + \delta tm = t(1 + \delta m)
\end{aligned}$$

Let $t_m = \min_{v \in V} t$ s.t. $tv \in \partial K$ be the minimum scaling required for a vector in V to extend it to ∂K. For $\delta < \frac{1-(1/t_m)}{|m|}$, we have $t(1 + \delta m) > 1$. Thus, choosing this δ, we have tv lies outside $\mathcal{E}_\delta$ for all $v \in V$ and their corresponding values of t.

The final step is to show that all points in $S^{n-1} \setminus V$ also lie outside $\mathcal{E}_\delta$. This is rather easy and can be done as follows. For any $v \in S^{n-1} \setminus V$, there exists a $u \in \partial B_n \cap \partial K$ such that $\|v - u\| < \frac{\alpha}{2\|H\|_{op} + \|h\|}$. Therefore, we have, for a v and its corresponding u,

$$\begin{aligned}
&|\langle H, vv^\intercal\rangle + \langle h, v\rangle - \langle H, uu^\intercal\rangle + \langle h, u\rangle| \\
&= |\langle H, v(v-u)^\intercal\rangle + \langle H, u(v-u)^\intercal\rangle + \langle h, v - u\rangle| \\
&\leq (2\|H\|_{op} \cdot \|v - u\| + \|h\| \cdot \|v - u\|) && \text{(Cauchy-Schwarz)}\\
&= (2\|H\|_{op} + \|h\|) \cdot \|v - u\| \\
&< \alpha && \left(\|v - u\| < \tfrac{\alpha}{2\|H\|_{op} + \|h\|}\right)
\end{aligned}$$

Once again, let $t > 1$ be the scaling factor such that $tv \in \partial K$ for a $v \in S^{n-1} \setminus V$. Therefore, we have:

$$\begin{aligned}
&\left(tv + \frac{\delta}{2}(\mathbf{I}_n + \delta H)^{-1}h\right)^\intercal (\mathbf{I}_n + \delta H) \left(tv + \frac{\delta}{2}(\mathbf{I}_n + \delta H)^{-1}h\right) \\
&> t + \delta t\left(\langle H, vv^\intercal\rangle + \langle h, v\rangle\right) && \text{(see (2))}\\
&\geq t + \delta t\left(\langle H, uu^\intercal\rangle + \langle h, u\rangle - \alpha\right) \\
&\geq t + \delta t(1 + \alpha - \alpha) && \text{(see (1))}\\
&= t(1 + \delta) > 1
\end{aligned}$$

This shows that tv lies outside $\mathcal{E}_\delta$.

Thus, we conclude that $\mathcal{E}_\delta$ is completely inside K for a suitable value of $\delta > 0$. Hence, we have found an ellipsoid of volume bigger than that of B_n contained inside K. This contradicts the assumption that B_n is the maximum volume ellipsoid of K. Therefore, we must have $[vec(\mathbf{I}_n)/n, 0_n] \in C = conv\{[vec(uu^\intercal), u] : u \in \partial B_n \cap \partial K\}$. In other words, there exists $c_1, \ldots, c_m \geq 0$, $\sum_{i=1}^m c_i = 1$ and $u_1, \ldots, u_m \in \partial B_n \cap \partial K$ such that,

$$\sum_{i=1}^m c_i u_i = 0 \quad \text{and} \quad \sum_{i=1}^m c_i u_i u_i^\intercal = \frac{1}{n}\mathbf{I}_n$$

Moreover, by Carathéodory's Theorem, $m \leq \frac{n(n+1)}{2} + n + 1 = \frac{n(n+3)}{2} + 1$ (note that $n(n+1)/2$ suffices instead of n^2 because the convex hull contains only symmetric matrices). This completes the proof of John's theorem. □

Remark 2.1. *A convex body whose maximum volume enclosed ellipsoid is the unit ball B_n is said to be in John's position. John's theorem shows that there is an invertible map that can be used to map a given convex body to its John's position. Although finding such a transformation is computationally hard, good approximations (upto $(1+\varepsilon)$ factor) to the John ellipsoid can be computed in polynomial time. The algorithm is based on a Franke-Wolfe type coordinate ascent that maximizes the $\log\det(\cdot)$ of the p.d. matrix that encodes the transformation of the unit ball to John's ellipsoid. For a detailed discussion please refer to [Todd, 2016].*

2.1 Löwner Ellipsoid

Similar to the maximum volume ellipsoid contained in a convex body K, we can also show that every convex body has an ellipsoid of minimum volume that contains K. Moreover, a characterization of the contact points similar to John's theorem is also possible. For the symmetric case, it is easy to show this by applying John's theorem on the polar of K defined as $K^\circ \coloneqq \{y \in \mathbb{R}^n : \langle y, x\rangle \le 1 \;\forall\; x \in K\}$. Towards this, we state the following lemmas.

Claim 2.2. *For a convex body K, its polar K° is convex.*

Proof. Suppose $y_1, y_2 \in K^\circ$. Fix any $x \in K$. Then, $\langle y_1, x\rangle \le 1$ and $\langle y_2, x\rangle \le 1$. Thus, $\langle \lambda y_1 + (1-\lambda)y_2\rangle = \lambda\langle y_1, x\rangle + (1-\lambda)\langle y_2, x\rangle \le \lambda + 1 - \lambda = 1$, for any $\lambda \in [0,1]$. Therefore, $\lambda y_1 + (1-\lambda)y_2 \in K^\circ$, for all $\lambda \in [0,1]$. Hence K° is convex. □

Next, we state the following lemma about invertible transformation of convex bodies.

Lemma 2.2. *If K is a convex body in $\mathbb{R}^n$ with 0 in the interior of K, then for any invertible $T \in \mathbb{R}^{n\times n}$, it holds that $(TK)^\circ = T^{-\intercal}K^\circ$.*

Proof. We have by definition of polar

$$\begin{aligned}
(TK)^\circ &= \{y \in \mathbb{R}^n : \langle y, x\rangle \le 1 \;\forall\; x \in TK\} \\
&= \{y \in \mathbb{R}^n : \langle y, x\rangle \le 1 \;\forall\; x \in K\} \\
&= \{y \in \mathbb{R}^n : \langle T^\intercal y, x\rangle \le 1 \;\forall\; x \in K\} \\
&= \{T^{-\intercal} y : y \in \mathbb{R}^n, \langle y, x\rangle \le 1 \;\forall\; x \in K\} \\
&= T^{-\intercal}\{y \in \mathbb{R}^n : \langle y, x\rangle \le 1 \;\forall\; x \in K\} \\
&= T^{-\intercal} K^\circ
\end{aligned}$$

□

Lastly, we have the following property about polars of convex bodies.

Lemma 2.3. *If K and L are two convex bodies such that $K \subseteq L$, then we have $L^\circ \subseteq K^\circ$.*

Proof. If $x \in L^\circ$, then $\langle x, y\rangle \le 1$ for all $y \in L$. But $L \supseteq K$, thus, $\langle x, y\rangle \le 1$ for all $y \in K$, which implies that $x \in K^\circ$. Therefore, we have $L^\circ \subseteq K^\circ$. □

Now we are ready to prove the analog of John's theorem for Löwner ellipsoids, which is the ellipsoid of minimum volume containing a given convex body. Hereafter, assume K to be a symmetric convex body. By symmetry, the Löwner ellipsoid of K must have its center at 0.

Applying John's theorem on K°, let B_n be the maximum volume contained ellipsoid of K°. Then, since $B_n \subseteq K^\circ$, for any $y \in B_n$, we have that $\langle y, x\rangle \le 1$ for all $x \in K$ (by definition of polar). Letting $\widehat{x}$ be the unit vector in the direction of $x \in K$ (note that $\widehat{x} \in B_n$), we have $\langle x, \widehat{x}\rangle = \|x\| \le 1$. Thus, $K \subseteq B_n$. Now suppose TB_n, where $T \ne \mathbf{I}_n$ is a p.d. matrix, is the minimum volume ellipsoid containing K. Then we must have $\det(T) < 1$. However, since $K \subseteq TB_n$, applying lemma 2.3 $(TB_n)^\circ \subseteq K^\circ$. Again, by lemma 2.2, $T^{-1}B_n \subseteq K^\circ$ (note that $B_n^\circ = B_n$ and T is symmetric). Thus, $T^{-1}B_n$ must be an ellipsoid of volume smaller than B_n, i.e., $\det(T^{-1}) < 1$. Therefore, $1 > \det(T)\det(T^{-1}) = \det(TT^{-1}) = 1$, which is a contradiction. Hence, $\det(T) = 1$ and thus by uniqueness (which can be shown by similar arguments), B_n must be the smallest volume ellipsoid containing K.

In the next section, we will see an application of the idea of maximum volume ellipsoid in designing good estimators for the well known problem of linear regression.

3 Optimal Design

We first give a motivating example for the problem of experimental design. This is a well-studied problem in statistics and has recently seen plethora of applications in machine learning and reinforcement learning. This continues to be a rich area of study and what we describe here is just a starting point.

3.1 Motivation

Consider the following problem. We have been given $P = \{x_1, \ldots, x_N\} \subset \mathbb{R}^n$, a set of N vectors. Our goal is to learn an unknown parameter vector $\theta^* \in \mathbb{R}^n$. However, we are given a query model in which we can pick a vector z_t from the set P, for $t = 1, \ldots, T$ times, and for every z_t, we can observe

$$y_t = \langle z_t, \theta^* \rangle + \eta_t, \quad \text{where } \eta_t \sim \mathcal{N}(0, 1)$$

A standard way to solve this problem is to solve the least squares regression: $\widehat{\theta} \coloneqq \text{argmin}_{\theta \in \mathbb{R}^n} \sum_{t=1}^{T} \left(\langle z_t, \theta \rangle - y_t\right)^2$. The solution can be obtained in closed form by applying first order optimality condition (setting derivative w.r.t θ to 0) as $\widehat{\theta} = \left(\sum_{t=1}^{T} z_t z_t^{\intercal}\right)^{-1} \left(\sum_{t=1}^{T} y_t z_t\right)$. The matrix $V = \sum_{t=1}^{T} z_t z_t^{\intercal}$ is called the *design matrix*. We have the following result:

Proposition 3.1. *If η_t's are zero-mean, then $\widehat{\theta}$ is an unbiased estimator of θ^*. In other words, $\mathbb{E}[\widehat{\theta}] = \theta^*$.*

Proof.

$$\begin{aligned}
\widehat{\theta} &= V^{-1} \sum_{t=1}^{T} y_t z_t = V^{-1} \sum_{t=1}^{T} \left(\langle z_t, \theta^* \rangle + \eta_t\right) z_t = V^{-1} \sum_{t=1}^{T} z_t \langle z_t, \theta^* \rangle + V^{-1} \sum_{t=1}^{T} \eta_t z_t \\
&= V^{-1} \sum_{t=1}^{T} \left(z_t z_t^{\intercal}\right) \theta^* + V^{-1} \sum_{t=1}^{T} \eta_t z_t = V^{-1} V \theta^* + V^{-1} \sum_{t=1}^{T} \eta_t z_t = \theta^* + V^{-1} \sum_{t=1}^{T} \eta_t z_t \qquad (3)
\end{aligned}$$

$$\begin{aligned}
\text{therefore,} \quad \mathbb{E}[\widehat{\theta}] &= \theta^* + \mathbb{E}\left[V^{-1} \sum_{t=1}^{T} \eta_t z_t\right] \\
&= \theta^* + V^{-1} \sum_{t=1}^{T} \mathbb{E}[\eta_t] z_t = \theta^* \qquad (\mathbb{E}[\eta_t] = 0)
\end{aligned}$$

□

Although the estimator $\widehat{\theta}$ is unbiased, we desire stronger guarantees from our estimator. We ask the following question: does there exists a set of query vectors $\{z_1, \ldots, z_T\}, z_t \in P, 1 \leq t \leq T$, such that for any point $x \in P$, the deviation $|\langle x, \widehat{\theta} - \theta^* \rangle|$ is small. Put another way, we want our estimator $\widehat{\theta}$ to predict the value $\langle x, \theta^* \rangle$ as well as possible for all $x \in P$. The key challenge here is that the size of the set P, that is N, can be much larger than T. Thus, we cannot query every point in P sufficiently many times such that we have good estimate for every point. One therefore needs a cleverer way of querying. This is problem in statistics literature is called the *experimental design*.

Suppose z_t, $t \in [T]$, are chosen non-adaptively, meaning, one decides beforehand which points to query and this decision is unchanged by the observations y_t, $t \in [T]$. Restricting the queries to be non-adaptive might seem like a handicap but one can show that the gain from adaptive policies is not any better in the worst case. Let us try to apply some standard probability bounds on the deviation $\langle x, \widehat{\theta} - \theta^* \rangle$.

$$\begin{aligned}
\mathbb{P}[\langle x, \widehat{\theta} - \theta^* \rangle \geq \varepsilon] &= \mathbb{P}[\exp(\lambda \langle x, \widehat{\theta} - \theta^* \rangle) \geq \exp(\lambda \varepsilon)] && (\text{for } \lambda > 0) \\
&\leq \frac{\mathbb{E}[\exp(\lambda \langle x, \widehat{\theta} - \theta^* \rangle)]}{\exp(\lambda \varepsilon)} && (\text{by Markov's inequality}) \\
&= e^{-\lambda \varepsilon} \cdot \mathbb{E}[\exp(\lambda \langle x, V^{-1} \sum_{t=1}^{T} \eta_t z_t \rangle)] && (\text{see } (3)) \\
&= e^{-\lambda \varepsilon} \cdot \prod_{t=1}^{T} \mathbb{E}[\exp(\lambda \eta_t \langle x, V^{-1} z_t \rangle)] && (\eta_t z_t \text{ are independent for } t \in [T])
\end{aligned}$$

For a Gaussian random variable $\eta \sim \mathcal{N}(0,1)$, the moment generating function $\mathbb{E}[\exp(\lambda\eta)] = \exp(\frac{1}{2}\lambda^2)$ for any $\lambda \in \mathbb{R}$. Therefore,

$$\begin{aligned}
\mathbb{P}[\langle x, \widehat{\theta} - \theta^* \rangle \geq \varepsilon] &\leq e^{-\lambda\varepsilon} \cdot \prod_{t=1}^{T} \exp\left(\frac{1}{2}\lambda^2 \langle x, V^{-1} z_t\rangle^2\right) \\
&= e^{-\lambda\varepsilon} \cdot \exp\left(\frac{1}{2}\lambda^2 \sum_{t=1}^{T} \langle x, V^{-1} z_t\rangle^2\right) \\
&= e^{-\lambda\varepsilon} \cdot \exp\left(\frac{1}{2}\lambda^2 \sum_{t=1}^{T} (x^\intercal V^{-1} z_t)(z_t^\intercal V^{-1} x)\right) \\
&= e^{-\lambda\varepsilon} \cdot \exp\left(\frac{1}{2}\lambda^2 (x^\intercal V^{-1} \left(\sum_{t=1}^{T} z_t z_t^\intercal\right) V^{-1} x)\right) \\
&= e^{-\lambda\varepsilon} \cdot \exp\left(\frac{1}{2}\lambda^2 (x^\intercal V^{-1} x)\right)
\end{aligned}$$

Since the above inequality holds for all $\lambda > 0$, it particularly holds that $\mathbb{P}[\langle x, \widehat{\theta} - \theta^* \rangle \geq \varepsilon] \leq \inf_{\lambda > 0} \exp\left(\frac{1}{2}\lambda^2 (x^\intercal V^{-1} x) - \lambda\varepsilon\right)$. Minimizing the quadratic in the exponent, we obtain

$$\mathbb{P}[\langle x, \widehat{\theta} - \theta^* \rangle \geq \varepsilon] \leq \exp\left(-\frac{\varepsilon^2}{2\langle x, V^{-1} x\rangle^2}\right)$$

Therefore, to minimize the probability of deviation, we need to minimize $\langle x, V^{-1} x\rangle$. Moreover, we need to do so for all $x \in P$. Thus, the following natural question arises:

$$\min_{z_1,\ldots,z_T} \max_{x \in P} \langle x, V^{-1} x\rangle \quad \text{where } V = \sum_{t=1}^{T} z_t z_t^\intercal$$

The above problem is computationally hard as it involves an integer program (think of variables $w_i \in \{0,1\}$, $i \in [n]$, indicating whether $x_i \in P$ is to be included in $\{z_t\}_{t \in [T]}$ or not). We can relax the requirement of the problem to ask for a distribution over P such that when points to be queried are samples from this distribution, the expected design matrix gives a good deviation bound. In other words, we pose the following optimization problem:

$$\min_{\lambda \in \triangle(P)} \max_{x \in P} \langle x, V^{-1} x\rangle \quad \text{where } V = \sum_{i=1}^{N} \lambda_i x_i x_i^\intercal$$

where $\triangle(P)$ is the probability simplex over P, i.e, $\triangle(P) = \{\lambda \in [0,1]^N : \sum_{i=1}^{N} \lambda_i = 1\}$. The above problem is called the *G-optimal design* problem. This is a non-convex optimization problem and it is not clear how to solve this problem.

3.2 Minimum Volume Enclosing Ellipsoid

Let us now return to the problem of finding the minimum volume ellipsoid that contains our point set P. From John's theorem, we know that such an ellipsoid exists and has nice characterization of the contact points. We will see how this connects to the G-optimal design problem.

Consider a centered ellipsoid $\mathcal{E} = T^{1/2} B_n$ that contains P inside it (here T is a p.d. matrix). Alternatively, the ellipsoid is given by $\mathcal{E} = \{x \in \mathbb{R}^n : \langle x, T^{-1} x\rangle \leq 1\}$. Therefore, for all $x \in P$, we must have $\langle x, T^{-1} x\rangle \leq 1$. Moreover, the volume of $\mathcal{E}$ is $\det(T) Vol(B_n)$. Thus, the minimum volume ellipsoid containing P is given by:

$$\min_{T \text{ p.d.}} \det(T) \quad \text{s.t. } \langle x, T^{-1} x\rangle \leq 1 \ \forall\ x \in P$$

From Minkowski determinant inequality, it can be seen that $\log\det(\cdot)$ is a strictly concave function over the set of p.d. matrices. Moreover, the constraints above are non-convex in T. Therefore, we first replace T with T^{-1} and then change the objective from $\det(\cdot)$ to $\log\det(\cdot)$ (log being monotonically increasing preserves the optimization problem). Finally, we obtain the following optimization problem:

$$\min_{T \text{ p.d.}} -\log\det(T) \quad \text{s.t. } \langle x, Tx\rangle \leq 1 \ \forall\ x \in P \tag{P}$$

This is a convex program and can be solved via standard convex optimization algorithms. However, it is interesting to look at the dual of this problem. Let us write the Lagrangian of the optimization problem and then find the dual program. The Lagrangian is given by:

$$\mathcal{L}(T, \lambda) = -\log\det(T) + \sum_{i=1}^{N} \lambda_i(\langle x_i, Tx_i\rangle - 1)$$

From the KKT conditions, we have at optimality, $\nabla_T \mathcal{L} = 0$. This gives,

$$\nabla_T \mathcal{L} = -T^{-1} + \sum_{i=1}^{N} \lambda_i x_i x_i^\intercal = 0$$

$$\text{which gives } T = \left(\sum_{i=1}^{N} \lambda_i x_i x_i^\intercal\right)^{-1} \quad \text{iff RHS is p.d.}$$

If $V := \sum_{i=1}^{N} \lambda_i x_i x_i^\intercal$ is positive semi-definite (with at least one 0 eigenvalue), then let $k \in \mathbb{R}^n$ be chosen such that $Vk = 0$ and $\|k\| = 1$ and set $T = \mathbf{I}_n + \delta kk^\intercal$. Then, the Lagrangian becomes

$$\begin{aligned}
\mathcal{L} &= -\log\det(\mathbf{I}_n + \delta kk^\intercal) + \sum_{i=1}^{N} \mathrm{Tr}((\mathbf{I}_n + \delta kk^\intercal)\lambda_i x_i x_i^\intercal) - \sum_{i=1}^{N} \lambda_i \\
&= -\log\det(\mathbf{I}_n + \delta kk^\intercal) + \mathrm{Tr}((\mathbf{I}_n + \delta kk^\intercal)V) - \sum_{i=1}^{N} \lambda_i \\
&= -\log(1+\delta) + \mathrm{Tr}(V) - \sum_{i=1}^{N} \lambda_i \qquad (\det(\mathbf{I}_n + vv^\intercal) = 1 + \|v\|^2 \text{ for any } v\text{; and } Vk = 0)
\end{aligned}$$

Thus, by choosing δ arbitrarily large, the Lagrangian becomes unbounded and hence no solutions exists. Thus, the solution $T = V^{-1}$ is valid only when V is p.d. Hereafter, we shall assume that V is p.d. Putting $T = V^{-1}$ in the Lagrangian, we obtain,

$$\begin{aligned}
\min_{T \text{ p.d}} \mathcal{L} &= \log\det(V) + \sum_{i=1}^{n} \mathrm{Tr}(V^{-1}(\lambda_i x_i x_i^\intercal)) - \sum_{i=1}^{n} \lambda_i \\
&= \log\det(V) + \mathrm{Tr}(V^{-1} \sum_{i=1}^{n} \lambda_i x_i x_i^\intercal) - \sum_{i=1}^{N} \lambda_i \\
&= \log\det(V) + \mathrm{Tr}(\mathbf{I}_n) - \sum_{i=1}^{N} \lambda_i \qquad (V = \textstyle\sum_{i=1}^{N} \lambda_i x_i x_i^\intercal) \\
&= \log\det(V) + n - \sum_{i=1}^{N} \lambda_i
\end{aligned}$$

Let $\sum_{i=1}^{N} \lambda_i = n\alpha$ and let $\lambda_i' = \frac{\lambda_i}{n\alpha}$ and $V' =:= \sum_{i=1}^{N} \lambda_i' x_i x_i^\intercal$. Then we have $\sum_{i=1}^{N} \lambda_i' = 1$ and $V = n\alpha V'$. Rewriting in terms of λ_i', we get,

$$\begin{aligned}
\min_{T \text{ p.d.}} \mathcal{L} &= \log\det(n\alpha V') + n - n\alpha \\
&= \log\det(V') + n + n\log n + n(\log\alpha - \alpha)
\end{aligned}$$

The above expression is maximized for $\alpha = 1$. Thus,

$$\max_{\lambda \geq 0} \min_{T \text{ p.d}} \mathcal{L} = \max_{\lambda' \geq 0, \sum_{i=1}^{N} \lambda'_i = 1} \log\det(V')$$

Renaming V' as V and λ' as λ, we finally have,

$$\max_{\lambda \in \triangle(P)} \log\det(V) \tag{D}$$

This problem is known as the *D-Optimal design* problem. This is the dual of the original problem of finding the minimum volume ellipsoid containing P. We state the following theorem without giving its proof here (see [Todd, 2016, Section 2.1] for the proof).

Lemma 3.1. *The duality gap between* (P) *and* (D) *is* 0.

Therefore, it suffices to solve the dual problem to obtain a solution for the primal problem, that is, a solution to the D-optimal design problem also generates the solution to the (centered) minimum volume enclosing ellipsoid of P.

Remark 3.1. *We observe that if $V^* = \sum_{i=1}^{N} \lambda_i^* x_i x_i^\intercal$ is the solution of the D-optimal design problem, then the corresponding minimum volume centered ellipsoid is given by $(nV^*)^{-1/2} B_n$. Thus, although on the face the D-optimal design looks like a volume maximization problem, it is actually inverse of the required ellipsoid whose volume we want to minimize.*

Remark 3.2. *It is clear from the duality (by Slater's condition) that whenever $\lambda_i^* \neq 0$, we must have $\langle x_i, V^{*-1} x_i \rangle = n$ (recall that V^* here is actually $\frac{1}{n}$ times the V defined in the original minimum volume ellipsoid problem). Thus, the points $x \in P$ whose corresponding λ^*'s are non-zero, lie on the boundary of the ellipsoid. Hence, these are the contact points between the polytope $\mathcal{P}$ given by the convex hull of P and the (centered) minimum volume enclosing ellipsoid of $\mathcal{P}$. An example is shown in Figure 4 (left). The points with non-zero λ^* would correspond to the points E, F, H and J.*

Remark 3.3. *The centered assumption can be relaxed by recasting the problem into $n+1$-dimensions and rewriting x_i's as $[x_i, 1] \in \mathbb{R}^n$, that is, appending a 1 to the vectors $x_i \in P$. Thereafter, the D-optimal design can be solved and the resulting ellipsoid can be projected on to the hyperplane given by $\{x \in \mathbb{R}^n : \langle e_{n+1}, x \rangle = 1\}$, where e_{n+1} is the $(n+1)$-th standard basis vector. One such visualization is shown in Figure 4 (right). See [Todd, 2016, Section 2.3] for the details.*

In the next subsection we will finally see the connection between the D-optimal design problem and the G-optimal design problem.

3.3 Kiefer-Wolfowitz Theorem

Kiefer and Wolfowitz in 1960 showed a remarkable connection between the two design problems stated above. The result can be succinctly stated as follows:

Theorem 3.1 (Kiefer-Wolfowitz)**.** *Given a collection of points $P = \{x_1, \ldots, x_N\}$ such that $span(P) = \mathbb{R}^n$, the following statements are equivalent:*

(i) *λ^* is the solution of the G-optimal design problem.*

(ii) *λ^* is the solution of the D-optimal design problem.*

(iii) *The objective value of G-optimal design at λ^* is n.*

Proof. Firstly, we state the following identity for any matrix $A \in \mathbb{R}^{n \times n}$ with positive determinant whose elements are a function of $x \in \mathbb{R}$: $\frac{\partial}{\partial x} \log\det(A) = \text{Tr}\left(A^{-1} \frac{\partial A}{\partial x}\right)$.

Now recall that in the D-optimal design problem, we have $V = \sum_{i=1}^{N} \lambda_i x_i x_i^\intercal$, where $\lambda \in \triangle(P)$. Using the stated identity, we have,

$$\frac{\partial}{\partial \lambda_i} \log\det(V) = \text{Tr}\left(V^{-1} \frac{\partial V}{\partial \lambda_i}\right) = \text{Tr}\left(V^{-1}(x_i x_i^\intercal)\right) = \langle x_i, V^{-1} x_i \rangle \tag{4}$$

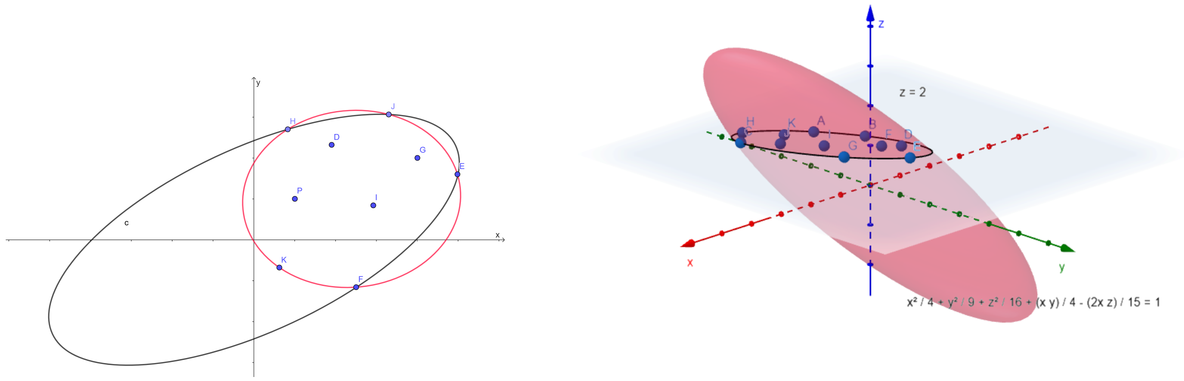

Figure 4: **Left**: The black ellipse is the centered minimum volume enclosing ellipsoid for the point set, whereas the red ellipse is the uncentered minimum volume enclosing ellipsoid. The D-optimal design problem will return the black ellipse, with the support of λ being points E, F, H and J. **Right**: While the points lie in 2D (represented as the blue plane $z = 2$ in 3D), by embedding them in 1 higher dimension, we are able to obtain a minimum volume enclosing ellipsoid (in this case, a 2D ellipse, shown by black outline) that is not centered. The red ellipsoid in 3D is the ellipsoid resulting from solving the D-optimal design in one higher dimension.

We also have,

$$\sum_{i=1}^{N} \lambda_i \langle x_i, V^{-1} x_i \rangle = \sum_{i=1}^{N} \operatorname{Tr}\left(V^{-1}(\lambda_i x_i x_i^\intercal)\right) = \operatorname{Tr}\left(V^{-1} \sum_{i=1}^{N} \lambda_i x_i x_i^\intercal\right) = \operatorname{Tr}(\mathbf{I}_n) = n \tag{5}$$

This further shows that for all $\lambda \in \triangle(P)$ and its corresponding V, the objective value of G-optimal design $\max_{x \in P} \langle x, V^{-1} x\rangle \geq n$ (**observation**).

Now, we start the proof by showing the implications cyclically.

(ii) $\implies$ (i) and (iii): Let V^* be the value of V at the maximizer λ^*. By concavity of $\log\det(\cdot)$, we have the necessity of the first order optimality condition, for any $\lambda \in \triangle(P)$,

$$\begin{aligned}
0 &\leq \langle \nabla_\lambda \log\det(V) \mid_{\lambda^*}, \lambda^* - \lambda \rangle \\
&= \sum_{i=1}^{N} \lambda_i^* \langle x_i, V^{*-1} x_i \rangle - \sum_{i=1}^{N} \lambda_i \langle x_i, V^{*-1} x_i \rangle \leq 0 && \text{(by identity 4)} \\
&= n - \sum_{i=1}^{N} \lambda_i \langle x_i, V^{*-1} x_i \rangle && \text{(by identity 5)} \\
\text{therefore, } & \sum_{i=1}^{N} \lambda_i \langle x_i, V^{*-1} x_i \rangle \leq n
\end{aligned}$$

Since the above inequality holds for any $\lambda \in \triangle(P)$, we can set $\lambda = e_i$, varying $i \in [N]$, where e_i is the i-th standard basis vector in $\mathbb{R}^N$. This gives us that for each $i \in [N]$, $\langle x_i, V^{*-1} x_i \rangle \leq n$. Hence, the objective value of G-optimal design for λ^* given by $\max_{x \in P} \langle x, V^{*-1} x \rangle \leq n$.

On the other hand, by observation above, $\max_{x \in P} \langle x, V^{*-1} x \rangle \geq n$.

Therefore, $\max_{x \in P} \langle x, V^{*-1} x \rangle = n$. Moreover, λ^* is a solution to the G-optimal design problem because it attains the minimum value possible for the G-optimal design problem.

(iii) $\implies$ (ii): Given that the objective value of G-optimal design at λ^* is n, we have, for any $\lambda \in \triangle(P)$,

$$\sum_{i=1}^{N} \lambda_i^* \langle x_i, V^{*-1} x_i \rangle \leq \max_{x \in P} \langle x, V^{*-1} x \rangle = n$$

Moreover, we have $\sum_{i=1}^{N} \lambda_i^* \langle x_i, V^{*-1} x_i \rangle = n$ by the identity 5. Therefore, we obtain,

$$\begin{aligned} 0 &\leq \sum_{i=1}^{N} \lambda_i^* \langle x_i, V^{*-1} x_i \rangle - \sum_{i=1}^{N} \lambda_i^* \langle x_i, V^{*-1} x_i \rangle \\ &= \langle \nabla_\lambda \log\det(V) \mid_{\lambda^*}, \lambda^* - \lambda \rangle \end{aligned}$$

By sufficiency of the first order optimality condition of the concave $\log\det(\cdot)$ function, we have that λ^* is a maximizer of the D-optimal design problem.

(i) $\Longrightarrow$ (iii): This is now easy to show. By the fact that (ii) $\Longrightarrow$ (i) and that the solution to the D-optimal design exists by John's theorem, we have for the solution λ^* of D-optimal design, which is also a solution of the G-optimal design, $\max_{x \in P} \langle x, V^{*-1} x \rangle = n$. Thus, any minimizer of the G-optimal design must attain value at most n. However, $\lambda \in \triangle(P)$ and its corresponding V, $\max_{x \in P} \langle x, V^{-1} x \rangle \geq n$ (by observation). Therefore, any minimizer of the G-optimal design attains value exactly n. □

Lemma 3.2. *There exists λ^*, a solution to the D-optimal design problem, such that size of support of λ^* is at most $\frac{n(n+1)}{2} + 1$.*

Proof. Letting $V^* = \sum_{i=1}^{N} \lambda_i^* x_i x_i^\intercal$, and using the *vec* notation, we have $vec(V^*) \in conv(\{vec(x_i x_i^\intercal) : i \in [N]\}) \subset \mathbb{R}^{n^2}$. However, since the convex hull contains only symmetric matrices, the actual dimension can be reduced to $n(n+1)/2$ by dropping the last $n(n-1)/2$ entries from $vec(x_i x_i)^\intercal$, $i \in [N]$. Therefore, by Carathéodory's theorem, there exists $\{z_1, \ldots, z_{\frac{n(n+1)}{2}+1}\} \subset P$ such that $V^* = \sum_{i=1}^{\frac{n(n+1)}{2}+1} \alpha_i z_i z_i^\intercal$, where $\sum_{i=1}^{\frac{n(n+1)}{2}+1} \alpha_i = 1$ and $\alpha_i \geq 0$ for all $i \in [\frac{n(n+1)}{2} + 1]$ □

This concludes the proof of the Kiefer-Wolfowitz equivalence theorem and the lemma that there is a solution to the D-optimal design problem with $O(n^2)$ support size. Going back to our motivating problem described in section 3.1, we see that if a point from P is queried $\lceil \lambda^* T \rceil$ times, where λ^* is the solution to the D-optimal design with support size of at most $\frac{n(n+1)}{2} + 1$, then

$$\sum_{i=1}^{N} \lceil \lambda_i^* T \rceil \leq T + \frac{n(n+1)}{2} + 1 \tag{6}$$

$$\widehat{V} \coloneqq \sum_{i=1}^{N} \lceil \lambda_i^* T \rceil x_i x_i^\intercal \succcurlyeq \sum_{i:\lambda_i > 0} \lambda_i^* T x_i x_i^\intercal = T V^* \tag{7}$$

where V^* is defined as usual with respect to λ^* and for two p.d. matrices A, B, $A \succcurlyeq B$ means $A - B$ is p.d. The first inequality shows that the total number of queries made is only at most $\frac{n(n+1)}{2} + 1$ more than T. The second inequality shows that for all $x \in P$, $\langle x, \widehat{V}^{-1} x \rangle \leq \langle x, \frac{1}{T} V^{*-1} x \rangle \leq \frac{n}{T}$. Hence, the probability of large deviation for the least squares regressor for any $x \in P$ is bounded as:

$$\mathbb{P}[\langle x, \widehat{\theta} - \theta^* \rangle] \leq \exp\left(-\frac{\varepsilon^2}{(n/T)}\right) = \exp\left(-\frac{T\varepsilon^2}{n}\right)$$

This shows us that by using the D-optimal design policy, we can get an exponential concentration in the estimates of $\langle x, \theta^* \rangle$ for all $x \in P$ even when there are $N \gg T$ points. Note that if we were to query every point in P equally often, the probability bound on the right would be $\exp\left(-\frac{T\varepsilon^2}{N}\right)$. This bound is significantly worse when $N \gg n$ (in fact, N can be exponential in n).

Remark 3.4 (Computation of D-optimal Design)**.** *There are several algorithms to approximately compute the solution to the D-optimal design. A solution which is within $(1-\varepsilon)$ of the maximum D-optimal objective value such that every constraint in the primal problem* (P) *is violated by a factor of at most $(1+\varepsilon)$, can be computed (i) with a certain algorithm that requires $O(n \log\log n + \frac{n}{\varepsilon})$ iterations with a preprocessing computation of $O(nN)$ (ii) with another algorithm that requires $O(n \log\log N + N + \frac{n}{\varepsilon})$ iterations. Both these algorithms require $O(n^2 + nN)$ computations per iterations. These algorithms mainly rely on Franke-Wolfe type coordinate ascent schemes. See [Todd, 2016, Chapter 3] for a description of these algorithms.*

Remark 3.5 (Approximation of G-optimal Design)**.** *If only a good approximation to the G-optimal design problem is desired, that is, if we just require that $\langle x, V^{-1}x\rangle \leq (1+\varepsilon)n$ for all $x \in P$, then there is an algorithm that requires (i) $O(n \log\log N + \frac{n}{\varepsilon})$ iterations when λ is initialized as uniform distribution, and (ii) $O(n \log\log n + \frac{n}{\varepsilon})$ iterations when a special initialization (requiring $O(n^2 N)$ preprocessing) is used. The computational complexity of each iteration is $O(n^2 + nN)$. However, the interesting fact is that since this algorithm is a coordinate ascent type algorithm, i.e., it only modifies λ in a few coordinates in every iteration, the support size of the final output is also $O(n \log\log n + \frac{n}{\varepsilon})$ with the latter initialization and it is $O(n \log\log N + \frac{n}{\varepsilon})$ with uniform initialization. This is much smaller than Carathéodory's guarantee. See [Todd, 2016, Chapter 3] for a description of these algorithms.*

Further Readings

Much of the presentation here has been referred from [Ball et al., 1997] and [Tkocz, 2018]. These resources provide a good introduction to the rich area of convex geometry. John's theorem was presented in John's original paper [John, 1948] studying extremum value problems in more generality. A more analytical proof of the theorem on ellipsoid can be found in [Har-Peled, 2011]. Another good resource is [Matousek, 2013] which also contains several other application of convex geometric tools in discrete problems. Approximating convex bodies with ellipsoids is a standard subroutine in several other important algorithms, for example, the Ellipsoid method and sampling from convex bodies.

The section on Optimal Design has been referred from [Todd, 2016] and [Lattimore and Szepesvári, 2020] with occasional glances at the original papers [Kiefer and Wolfowitz, 1960, Kiefer, 1974]. A detailed study of the vast area of optimal design in statistics and the equivalences therein can also be found in [Pukelsheim, 2006]. The computational complexity of approximating John's ellipsoid is also an active area of study with recent papers [Cohen et al., 2019].

References

Keith Ball et al. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31(1-58):26, 1997.

Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex analysis and optimization*, volume 1. Athena Scientific, 2003.

Michael B Cohen, Ben Cousins, Yin Tat Lee, and Xin Yang. A near-optimal algorithm for approximating the john ellipsoid. In *Conference on Learning Theory*, pages 849–873. PMLR, 2019.

William W. Hager. Minimizing a quadratic over a sphere. *SIAM Journal on Optimization*, 12(1):188–208, 2001. doi: 10.1137/S1052623499356071. URL https://doi.org/10.1137/S1052623499356071.

Sariel Har-Peled. *Geometric approximation algorithms*. Number 173. American Mathematical Soc., 2011.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Fritz John. Extremum problems with inequalities as subsidiary conditions, studies and essays presented to r. courant on his 60th birthday, january 8, 1948, 1948.

Jack Kiefer. General equivalence theory for optimum designs (approximate theory). *The annals of Statistics*, pages 849–879, 1974.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Jiri Matousek. *Lectures on discrete geometry*, volume 212. Springer Science & Business Media, 2013.

Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970. ISBN 9781400873173. doi: doi:10.1515/9781400873173. URL https://doi.org/10.1515/9781400873173.

Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

Tomasz Tkocz. Asymptotic convex geometry lecture notes. 2018.

Michael J Todd. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.