國立交通大學

應用數學系

碩 士 論 文

設計的 VC 維度

The VC Dimension of Block Design

研 究 生：林詒琪

指導教授：翁志文　教授

中 華 民 國 一百零一 年 一 月

設計的 VC 維度

The VC Dimension of Block Design

研 究 生：林詒琪　　Student：Yi-Chi Lin

指導教授：翁志文　　Advisor：Chih-Wen Weng

國 立 交 通 大 學

應 用 數 學 系

碩 士 論 文

A Thesis

Submitted to Department of Applied Mathematics

College of Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Applied Mathematics

January 2012

Hsinchu, Taiwan, Republic of China

中 華 民 國 一百零一 年 一 月

設計的 VC 維度

學生:林詒琪 指導教授:翁志文教授

國立交通大學應用數學系碩士班

摘 要

對一 t-(v, k, λ) 區族設計 $\mathcal{D} = (X, \mathcal{B})$ 而言, 其 VC 維度定義為 X 中滿足下列條件的子集A其元素個數的最大可能值: 每一個 A 的子集合 C 都存在一個在區族集 $\mathcal{B}$ 中的區族 B, 使得 $C = A \cap B$。在這篇論文我們探討 t-(v, k, λ) 區族設計的 VC 維度的基本性質, 並且運用他們完整的決定當 $\lambda = 1$ 且 $t = 2$ 和 $t = 3$ 時, 這兩類區族設計的 VC 維度。

The VC Dimension of a Block Design

student：Yi-Chi Lin Advisors：Dr. Chih-Wen Weng

Department of Applied Mathematics

National Chiao Tung University

ABSTRACT

The *Vapnik-Chervonenkis dimension* of a $t-(v,k,\lambda)$ design $D=(X, \mathcal{B})$ is the largest cardinality of a subset A of X such that for each subset $C \subseteq A$ there exists a block $B \in \mathcal{B}$ such that $C = A \cap B$. In this thesis we give some general properties of the Vapnik-Chervonenkis dimension of a $t-(v,k,\lambda)$ design, and use them to completely determine the Vapnik-Chervonenkis dimension of a $t-(v,k,1)$ design for $t = 2$ and $t = 3$.

誌　　　謝

首先感謝我的指導教授翁志文老師，謝謝老師不論在課業或論文方面給予我許多的指導，老師謝謝您。接下來感謝黃大原老師，指導我論文初步的方向。還有在交大學習的期間，我遇到了許多的師長、學長姐與同學們，感謝你們，沒有你們就不會有現在的我，你們的存在是不可或缺的，謝謝你們。

我還要感謝國立恆春工商職業學校校長李恆霖校長，余佳晉主任，林靜主任，郭姿伶組長及導師室的所有同仁，感謝你們在我進修的時候協助我處理課務及班級事務，讓我可以專心的進修。

最後我要感謝我的家人，爸爸、媽媽、妹妹與弟弟。無論是什麼情況，你們總是支持著我、鼓勵著我，讓我有繼續往前的動力，萬分的感謝你們。

目 錄

1 Introduction

Suppose that for opinion poll we want to select a small number of individuals representing all major sections of the society. First, we have to choose certain categories of people and then decide which of these groups are considered "important". According to our democratic principles, we shall measure the "importance" of a group by its size (in the percentage of the population) [3, page: 247]. Then the important groups will define a hypergraph H. The smallest number of people representing all important group is shattered. Vladimir Vapnik and Alexey Chervonenkis defined VC dimension (for Vapnik-Chervonenkis dimension) [1]. The VC dimension is a measure of the capacity of a statistical classification algorithm, defined as the cardinality of the largest set of points that the algorithm can shatter. In this thesis we completely determine the VC dimension of a t-$(v, k, 1)$ design for $t = 2$ and $t = 3$.

2 VC dimension of hypergraph

In this section, we shall give the definition of the VC dimension of a hypergraph, provide an example and a basic property. A hypergraph is a generalization of a graph, where an edge can connect any number of vertices. The formal definition is given below.

Definition 2.1. A *hypergraph* H is a pair $H = (V(H), E(H))$ where $V(H)$ is a set of elements, called *points* or *vertices*, and $E(H)$ is a set of non-empty subsets of $V(H)$ called *hyperedges* or *blocks*. Therefore, $E(H)$ is a subset of $P(X) \setminus \{\emptyset\}$, where $P(X)$ is the power set of X.

Example 2.2. Let $V(H) = \{1, 2, 3, 4, 5, 6, 7\}$ and

$$E(H) = \{123, 145, 167, 246, 257, 347, 356\}.$$

Then H is a hypergraph. See Figure 1 below for the diagram illustration.

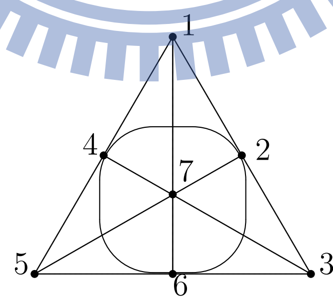

Figure 1.

Throughout this thesis, let $H = (V(H), E(H))$ denote a hypergraph.

Definition 2.3. A subset $A \subseteq V(H)$ is called *shattered* if for every $C \subseteq A$ there exists an $B \in E(H)$ such that $B \cap A = C$. The *Vapnik-Chervonenkis dimension* (or *VC dimension*) of H is the cardinality of the largest shattered subset of $V(H)$. It will be denoted by VC-dim(H). [3, page 247].

We give a necessary condition for a hypergraph to have VC dimension d.

Lemma 2.4. *If a hypergraph H has VC dimension d, then $|E(H)| \geq 2^d$.*

Proof. Let A be a shattered subset with $|A| = d$, so the number of subsets of A are 2^d. By the construction, we have $|E(H)| \geq 2^d$. $\square$

From the above lemma, a hypergraph without hyperedges has no VC dimension. A hypergraph with at least a hyperedge has VC dimension, since the empty set is shattered. The following lemma characterizes the hypergraph with VC dimension 0.

Lemma 2.5. *H has VC dimension 0 if and only if H has only one hyperedge.*

Proof. The sufficient condition follows immediately from Lemma 2.4. To prove the necessity, on the contrary assume H has two hyperedges $B \not\subseteq B'$. Pick $b \in B' - B$. Since the set $\{b\}$ is shattered, H has VC dimension at least 1. $\square$

3 Block designs

We shall consider a special class of hypergraphs hereafter.

Definition 3.1. Let t, v, k, and λ be positive integers such that $v \geq k \geq t$. A t-(v, k, λ) *design* is a hypergraph $(X, \mathcal{B})$ such that the following properties are satisfied:

(i) $|X| = v$;

(ii) each block contains exactly k points;

(iii) any t distinct points are contained in exactly λ blocks.

A t-(v, k, λ) design is *trivial* if $k = v$, i.e. there exists a unique block, which contains all the points. A *t-design* is a t-(v, k, λ) design for for some v, k, λ. Sometimes a t-design is referred as a *block design*.

The example in Example 2.2 is a 2-$(7, 3, 1)$ design. We give one more example of t-designs below.

Example 3.2. Let $X = \{1, 2, 3, 4, 5\}$ and

$$\mathcal{B} = \{123, 124, 125, 134, 135, 145, 234, 235, 245, 345\}.$$

It is easy to check that $(X, \mathcal{B})$ is a 3-$(5, 3, 1)$ design.

We are concerned in the VC dimension of a t-design. Note that from Lemma 2.5 a t-design has VC dimension 0 if and only if it is trivial. Hence we shall exclude the trivial t-design in our discussion. The following simple lemma provides an observation of how in general in this thesis we determine the VC dimension of a t-$(v, k, 1)$ design for t is 2 or 3.

Lemma 3.3. *Let $\mathcal{D} = (X, \mathcal{B})$ be the 2-$(7,3,1)$ design in Example 2.2, where $X = \{1,2,3,4,5,6,7\}$ and*

$$\mathcal{B} = \{123, 145, 167, 246, 257, 347, 356\}.$$

Then VC-dim$(\mathcal{D}) = 2$.

Proof. It is easy to check that $\{1,2\}$ is a shattered subset of X. Hence the VC dimension of $\mathcal{D}$ is at least 2. If VC-dim$(\mathcal{D}) \geq 3$, then by Lemma 2.4, $|\mathcal{B}| \geq 8$, a contradiction. Thus VC-dim$(\mathcal{D}) = 2$. □

The following basic properties of t-designs can be found from any text-books, e.g. [2, page 191].

Lemma 3.4. *Let $(X, \mathcal{B})$ be a t-(v,k,λ) design. Suppose that $Y \subseteq X$, where $|Y| = s \leq t$. Then there are exactly*

$$b_s = \frac{\lambda\binom{v-s}{t-s}}{\binom{k-s}{t-s}} \tag{1}$$

blocks in $\mathcal{B}$ that contain all the points in Y. □

Lemma 3.5. *Let $(X, \mathcal{B})$ be a t-(v,k,λ) design and $A \subseteq X$ with $|A| = d \leq t$. Then for $C \subseteq A$ with $|C| = i$, there are*

$$\sum_{j=0}^{d-i} (-1)^j b_{i+j} \binom{d-i}{j} \tag{2}$$

blocks B such that $B \cap A = C$.

Proof. Fix $C \subseteq A$ with $|C| = i$, and suppose $A - C = \{a_1, a_2, ..., a_{d-i}\} \subseteq X$. Set $S := \{B \in \mathcal{B} \mid C \subseteq B\}$, and $S_j = \{B \in \mathcal{B} \mid C \bigcup \{a_j\} \subseteq B\}$ for

$1 \le j \le d-i$. Note that $|\{B \in \mathcal{B} \mid B \cap A = C\}| = \left|\overline{S}_1 \bigcap \overline{S}_2 \bigcap \ldots \bigcap \overline{S}_{d-i}\right|$. By inclusion-exclusion principle,

$$\begin{aligned} & \left|\overline{S}_1 \bigcap \overline{S}_2 \bigcap \ldots \bigcap \overline{S}_{d-i}\right| \\ = & \sum_{k=0}^{d-i} (-1)^k \sum_{\substack{\alpha \subseteq \{1,2,3,\ldots,d-i\} \\ |\alpha| = k}} \left| \bigcap_{j \in \alpha} S_j \right| \\ = & \binom{d-i}{0} b_i - \binom{d-i}{1} b_{i+1} + \cdots + (-1)^{d-i} \binom{d-i}{d-i} b_d \\ = & \sum_{j=0}^{d-i} (-1)^j b_{i+j} \binom{d-i}{j}. \end{aligned}$$

□

Proposition 3.6. *Suppose a t-(v, k, λ) design $(X, \mathcal{B})$ has VC dimension at most t. Then the VC dimension of $(X, \mathcal{B})$ is the largest d such that the numbers in (2) are positive for all $0 \le i \le d$.*

Proof. This is immediate from the definition of VC dimension and Lemma 3.5. □

Proposition 3.7. *If the VC dimension of a nontrivial t-(v, k, λ) design $(X, \mathcal{B})$ is d, then*

$$d \le \min_{i \le t} \; i + \lfloor \log_2 b_i \rfloor.$$

Proof. Let A be a shattered subset of X, and $|A| = d$. For any $B \subseteq A$ with $|B| = i$, the number of subsets of A containing B are $2^{(d-i)}$. Since B is contained in exactly b_i blocks, we have $2^{(d-i)} \le b_i$, which is equivalent to $d - i \le \log_2 b_i$. □

We shall give an application of Proposition 3.7.

Corollary 3.8. *If* $(X, \mathcal{B})$ *is a* t-(v, k, λ) *design, then the VC dimension of* $(X, \mathcal{B})$ *is at most* $t + \log_2 \lambda$. *In particular, a* t-$(v, k, 1)$ *design has VC dimension at most* t.

Proof. This is immediate from Proposition 3.7. $\square$

Corollary 3.8 and Proposition 3.6 shed light on the determination of the VC dimension of a t-$(v, k, 1)$ design. We shall consider the cases $t = 2$ and $t = 3$ in the remaining of the thesis.

4 2-$(v, k, 1)$ design

We will determine the VC-dimension of a nontrivial 2-$(v, k, 1)$ design in this section. From Lemma 3.4, we have

$$b_0 = v(v-1)/(k(k-1)) > b_1 = (v-1)/(k-1) > \quad b_2 = 1.$$

We need the following lemma.

Lemma 4.1. *In a nontrivial 2-(v, k, λ) design,*

$$b_0 - 2b_1 + b_2 = \frac{\lambda(v-k)(v-k-1)}{k(k-1)}.$$

Proof. From (1),

$$\begin{aligned} b_0 - 2b_1 + b_2 &= \frac{\lambda v(v-1)}{k(k-1)} - \frac{2\lambda(v-1)}{k-1} + \lambda \\ &= \frac{\lambda v(v-1) - 2\lambda k(v-1) + \lambda k(k-1)}{k(k-1)} \\ &= \frac{\lambda(v-k)(v-k-1)}{k(k-1)}. \end{aligned}$$

□

Proposition 4.2. *The VC dimension d of a nontrivial 2-$(v, k, 1)$ design $(X, \mathcal{B})$ satisfies*

$$d = \begin{cases} 1, & \text{if } v = 3 \text{ and } k = 2; \\ 2, & \text{else.} \end{cases}$$

Proof. Note that $v > k$ since the design is nontrivial. Suppose $v = k + 1$. Pick distinct blocks $B, B' \in \mathcal{B}$. Note that $k \geq 2$ and $|B \cap B'| \leq 1$. Then

$$k + 1 = v \geq |B \cup B'| \geq |B| + |B'| - |B \cap B'| \geq 2k - 1.$$

This implies $k = 2$ and $v = 3$. Since a 2-$(3, 2, 1)$ design has three blocks, its VC dimension is at most 1 by Lemma 2.4, and indeed is 1 by Lemma 2.5, since it is a nontrivial 2-design. Suppose $v > k + 1$. Then

$$\begin{aligned} \sum_{j=0}^{2-i} (-1)^j b_{i+j} \binom{2-i}{j} &= \begin{cases} b_0 - 2b_1 + b_2, & i = 0; \\ b_1 - b_2, & i = 1; \\ b_2, & i = 2 \end{cases} \\ &> 0 \end{aligned}$$

by Lemma 4.1. This implies $d \geq 2$. By corollary 3.8, d is at most 2. Hence $d = 2$. □

5 3-$(v, k, 1)$ design

We determine the VC-dimension of a nontrivial 3-$(v, k, 1)$ design in this section. From Lemma 3.4, we have

$$b_0 = \frac{v(v-1)(v-2)}{k(k-1)(k-2)} > b_1 = \frac{(v-1)(k-2)}{(k-1)(k-2)} > b_2 = \frac{v-2}{k-2} > b_3 = 1.$$

To prove our main result, we need the following three lemmas.

Lemma 5.1. *In a nontrivial* 3-$(v, k, 1)$ *design,*

$$b_0 - 3b_1 + 3b_2 - b_3 = \frac{(v-k)(v-k-1)(v-k-2)}{k(k-1)(k-2)}.$$

Proof. From (1),

$$\begin{aligned}
& b_0 - 3b_1 + 3b_2 - b_3 \\
= & \frac{v(v-1)(v-2) - 3k(v-1)(v-2) + 3k(k-1)(v-2) - k(k-1)(k-2)}{k(k-1)(k-2)} \\
= & \frac{(v-1)(v-2)(v-k) - 2k(v-2)(v-k) + k(k-1)(v-k)}{k(k-1)(k-2)} \\
= & \frac{(v-k)[(v-1)(v-2) - 2k(v-2) + k(k-1)]}{k(k-1)(k-2)} \\
= & \frac{(v-k)(v-k-1)(v-k-2)}{k(k-1)(k-2)}.
\end{aligned}$$

□

Lemma 5.2. *In a nontrivial* 3-$(v, k, 1)$ *design,*

$$b_1 - 2b_2 + b_3 = \frac{(v-k)(v-k-1)}{(k-1)(k-2)}.$$

Proof. From (1),

$$\begin{aligned} b_1 - 2b_2 + b_3 &= \frac{(v-1)(v-2)}{(k-1)(k-2)} - \frac{2(k-1)(v-2)}{(k-1)(k-2)} + \frac{(k-1)(k-2)}{(k-1)(k-2)} \\ &= \frac{(v-2)(v-k)-(k-1)(v-k)}{(k-1)(k-2)} \\ &= \frac{(v-k)(v-k-1)}{(k-1)(k-2)}. \end{aligned}$$

□

Lemma 5.3. *In a nontrivial* 3-$(v,k,1)$ *design,*

$$b_0 - 2b_1 + b_2 = \frac{(v-2)(v-k)(v-k-1)}{k(k-1)(k-2)}.$$

Proof. From (1),

$$\begin{aligned} b_0 - 2b_1 + b_2 &= \frac{v(v-1)(v-2)}{k(k-1)(k-2)} - \frac{2(v-1)(v-2)}{(k-1)(k-2)} + \frac{v-2}{k-2} \\ &= \frac{(v-2)(v-k)(v-k-1)}{k(k-1)(k-2)}. \end{aligned}$$

□

Now we are ready to give the main result of this section.

Proposition 5.4. *The VC dimension d of a nontrivial* 3-$(v,k,1)$ *design satisfies*

$$d = \begin{cases} 1, & \text{if } (v,k) = (4,3)\ ; \\ 2, & \text{if } (v,k) = (5,3),\ \text{or } (6,4); \\ 3, & \text{else.} \end{cases}$$

Proof. By corollary 3.8, d is at most 3. We will use Proposition 3.6 to check the possibility for $d = 3$. Note that

$$\sum_{j=0}^{3-i}(-1)^j b_{i+j}\binom{3-i}{j} = \begin{cases} b_0 - 3b_1 + 3b_2 - b_3, & i = 0; \\ b_1 - 2b_2 + b_3, & i = 1; \\ b_2 - b_3, & i = 2; \\ b_3, & i = 3. \end{cases}$$
$$> 0$$

if $v > k + 2$ by Lemma 5.1-5.3. Hence if $v > k + 2$ then $d = 3$.

Suppose $v \leq k + 2$. Note that

$$1 = b_3 < b_2 = \frac{v-2}{k-2} \leq \left\lfloor \frac{k}{k-2} \right\rfloor = 1 + \left\lfloor \frac{2}{k-2} \right\rfloor.$$

Hence $k = 3$ and $b_2 = 2$ or $b_2 = 3$; or $k = 4$ and $b_2 = 2$.

(i) Suppose $k = 3$ and $b_2 = 2$. Then $v = 2k-2 = 4 = k+1$, $b_0 - 2b_1 + b_2 = 0$ by Lemma 5.3, and then $d = 1$.

(ii) Suppose $k = 3$ and $b_2 = 3$. Then $v = 3k - 4 = 5 = k + 2$, $b_0 - 3b_1 + 3b_2 - 1 = 0$ by Lemma 5.1, $b_0 - 2b_1 + b_2 > 0$ by Lemma 5.3, and then $d = 2$.

(iii) Suppose $k = 4$ and $b_2 = 2$. Then $v = 2k - 2 = 6 = k + 2$, $b_0 - 3b_1 + 3b_2 - 1 = 0$ by Lemma 5.1, $b_0 - 2b_1 + b_2 > 0$ by Lemma 5.3, and then $d = 2$.

□

References

[1] V. Vapnik and A. Chervonenkis, On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities, *Theory of Probability and Its Applications*, 1971, 16, 264-280.

[2] J. H. van Lint and R. M. Wilson, *A Course in Combinatorics*, Cambridge, Victoria, 1992.

[3] János Pach and Pankaj K. Agarwal, *Combinatorial Geometry*, John Wiley and Sons, New York, 1995.

[4] Douglas R. Stinson, *Combinatorial Designs*, Springer, 2003.