

Notes for High-Dimensional Probability Second Edition by
Roman Vershynin

Gallant Tsao

July 23, 2025

Contents

0	Appetizer: Using Probability to Cover a Set	4
0.1	Covering Geometric Sets	5
1	A Quick Refresher on Analysis and Probability	7
1.1	Convex Sets and Functions	7
1.2	Norms and Inner Products	7
1.3	Random Variables and Random Vectors	7
1.4	Union Bound	8
1.5	Conditioning	9
1.6	Probabilistic Inequalities	9
1.7	Limit Theorems	11
2	Concentration of Sums of Independent Random Variables	13
2.1	Why Concentration Inequalities?	13
2.2	Hoeffding Inequality	14
2.3	Chernoff Inequality	16
2.4	Application: Median-of-means Estimator	18
2.5	Application: Degrees of Random Graphs	20
2.6	Subgaussian Distributions	21
2.6.1	The Subgaussian Norm	23
2.7	Subgaussian Hoeffding and Khintchine Inequalities	23
2.7.1	Subgaussian Hoeffding Inequality	24
2.7.2	Subgaussian Khintchine Inequality	24
2.7.3	Maximum of Subgaussians	25
2.7.4	Centering	26
2.8	Subexponential Distributions	26
2.8.1	Subexponential Properties	26
2.8.2	The Subexponential Norm	28
2.9	Bernstein Inequality	29
3	Random Vectors in High Dimensions	32
3.1	Concentration of the Norm	32
3.2	Covariance Matrices and PCA	33
3.2.1	Learning from the Covariance Matrix	33
3.2.2	Principle Component Analysis	34
3.2.3	Isotropic Distributions	35
3.3	Examples of High-dimensional Distributions	35
3.3.1	Standard Normal	35
3.3.2	General Normal	36
3.3.3	Uniform on the Sphere	37
3.3.4	Uniform on a Convex Set	38
3.3.5	Frames	38
3.4	Subgaussian Distributions in High Dimensions	40
3.4.1	Gaussian, Rademacher, and More	40
3.4.2	Uniform on the Sphere	40
3.4.3	Non-examples	41
3.5	Application: Grothendieck Inequality and Semidefinite Programming	42
3.5.1	Semidefinite Programming	44
3.6	Application: Maximum Cut for Graphs	46
3.6.1	A Simple 0.5-approximation Algorithm	46
3.6.2	Semidefinite Relaxation	47
3.7	Kernel Trick and Tightening of Grothendieck Inequality	48
3.7.1	Tensors	49
3.7.2	Proof of Theorem 3.5.1	51
3.7.3	Kernels and Feature Maps	51

4	Random Matrices	52
4.1	A Quick Refresher on Linear Algebra	52
4.1.1	Singular Value Decomposition	52
4.1.2	Min-max Theorem	53
4.1.3	Frobenius and Operator Norms	54
4.1.4	The Matrix Norms and the Spectrum	54
4.1.5	Low-rank Approximation	55
4.1.6	Perturbation Theory	55
4.1.7	Isometries	57
4.2	Nets, Covering, and Packing	57
4.2.1	Covering Numbers and Volume	59
4.3	Application: Error Correcting Codes	60
4.4	Upper Bounds on Subgaussian Random Matrices	60
4.4.1	Computing the Norm on an ε net	61
4.4.2	The Norms of Subgaussian Random Matrices	61
4.4.3	Symmetric Matrices	63
4.5	Application: Community Detection in Networks	63
4.6	Two-sided Bounds on Subgaussian Matrices	63
4.7	Application: Covariance Estimation and Clustering	65
5	Concentration Without Independence	66
5.1	Concentration of Lipschitz Functions on the Sphere	66
5.1.1	Lipschitz Functions	66
5.1.2	Concentration via Isoperimetric Inequalities	66
5.1.3	Blow-up of Sets on the Sphere	67
5.1.4	Proof of Theorem 5.1.3	68
5.2	Concentration on Other Metric Measure Spaces	69
5.2.1	Gaussian Concentration	69
5.2.2	Hamming Cube	69
5.2.3	Symmetric Group	70
5.2.4	Riemannian Manifolds with Strictly Positive Curvature	70
5.2.5	Special Orthogonal Group	70
5.2.6	Grassmannian	71
5.2.7	Continuous Cube and Euclidean Ball	71
5.2.8	Densities of the Form $e^{-U(x)}$	71
5.2.9	Random Vectors with Independent Bounded Coordinates	72
5.3	Application: Johnson-Lindenstrauss Lemma	72
5.4	Matrix Bernstein Inequality	72
5.4.1	Matrix Calculus	72
5.4.2	Trace Inequalities	74
5.4.3	Proof of Matrix Bernstein Inequality	74
5.4.4	Matrix Hoeffding and Khintchine Inequalities	76
5.5	Application: Community Detection in Sparse Networks	77
5.6	Application: Covariance Estimation for General Distributions	77
5.7	Extra notes	77
6	Quadratic Forms, Symmetrization, and Contraction	78
6.1	Decoupling	78
6.2	Hanson-Wright Inequality	80
6.3	Symmetrization	80
6.4	Random Matrices with non-i.i.d. Entries	81
6.5	Application: Matrix Completion	82
6.6	Contraction Principle	82

7	Random Processes	85
7.1	Basic Concepts and Examples	85
7.1.1	Covariance and Increments	86
7.1.2	Gaussian Processes	86
7.2	Slepian, Sudakov-Fernique, and Gordon Inequalities	87
7.2.1	Gaussian Interpolation	88
7.2.2	Proof of Slepian Inequality	90
7.2.3	Sudakov-Fernique and Gordon Inequalities	91
7.3	Application: Sharp Bounds for Gaussian Matrices	92
7.4	Sudakov Inequality	92
7.4.1	Application for covering numbers in \mathbb{R}^n	93
7.5	Gaussian Width	94
7.5.1	Geometric Meaning of Width	95
7.5.2	Examples	96
7.5.3	Gaussian Complexity and Effective Dimension	97
7.6	Application: Random Projection of Sets	98
8	Chaining	99
8.1	Dudley Inequality	99
8.1.1	Variations and Examples	102
8.2	Application: Empirical Processes	103
8.3	VC Dimension	103
8.4	Application: Statistical Learning Theory	104
8.5	Generic Chaining	104
8.6	Chevet Inequality	104
9	Deviations of Random Matrices on Sets	101
9.1	Matrix Deviation Inequality	101
9.2	Random Matrices, Covariance Estimation, and Johnson-Lindenstrauss	101
9.3	Random Sections: The M^* Bound and Escape Theorem	101
9.4	Application: High-dimensional Linear Models	101
9.5	Application: Exact Sparse Recovery	101
9.6	Deviations of Random Matrices for General Norms	101
9.7	Two-sided Chevet Inequality and Dvoretzky-Milman Theorem	101

8 Chaining

This chapter concerns some of the central methods for bounding random processes (X_t) . We'll go over concepts such as chaining, VC theory, generic chaining methods, and bounds such as Talagrand's inequality and Chevet's inequality. We'll apply these to concepts such as Monte Carlo integration, empirical processes, and statistical learning theory.

8.1 Dudley Inequality

For a general Gaussian process $(X_t)_{t \in T}$, Sudakov inequality (Theorem 7.4.1) gives a *lower* bound on

$$\mathbb{E} \left[\sup_{t \in T} X_t \right]$$

in terms of the metric entropy of T . Now we'll go for an upper bound. Moreover, we generalize from Gaussian processes to subgaussian processes as well.

Definition 8.1.1. A random process $(X_t)_{t \in T}$ on a metric space (T, d) has subgaussian increments if there exists $K > 0$ such that

$$\|X_t - X_s\|_{\psi_2} \leq K d(t, s) \text{ for all } t, s \in T.$$

Example 8.1.2 (Gaussian processes). Let $(X_t)_{t \in T}$ be a Gaussian process on some set T . It naturally defines a *canonical metric* on T :

$$d(t, s) := \|X_t - X_s\|_{L^2}, \quad t, s \in T,$$

as we explained earlier. With respect to this metric, $(X_t)_{t \in T}$ clearly has subgaussian increments, with some absolute constant K .

Here is another (trivial) example: Any random process can be made to have subgaussian increments by defining the metric as $d(t, s) := \|X_t - X_s\|_{\psi_2}$.

Now we give a bound on a general subgaussian random process in terms of the metric entropy:

Theorem 8.1.3 (Dudley's integral inequality). Let $(X_t)_{t \in T}$ be a mean-zero random process on a metric space (T, d) with subgaussian increments as in Definition 8.1.1. Then

$$\mathbb{E} \left[\sup_{t \in T} X_t \right] \leq CK \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \varepsilon)} \, d\varepsilon.$$

Before going to the proof's let's compare Dudley's inequality with Sudakov's inequality (Theorem 7.4.1), which for Gaussian processes, says:

$$\mathbb{E} \left[\sup_{t \in T} X_t \right] \geq c \sup_{\varepsilon > 0} \varepsilon \sqrt{\log \mathcal{N}(T, d, \varepsilon)}.$$

Figure 8.1 below shows both bounds:

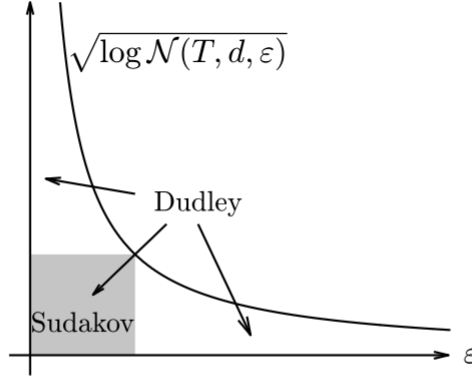


Figure 8.1 Dudley inequality bounds $\mathbb{E} \sup_{t \in T} X_t$ by the area under the curve. Sudakov inequality bounds it below by the largest area of a rectangle under the curve, up to constants.

There is a clear gap between the two bounds, and it turns out that metric entropy alone cannot close it - we will explore this later.

Dudley's inequality hints that $\mathbb{E} [\sup_{t \in T} X_t]$ is a *multiscale* quantity - to bound it, we need to look at T across all scales ε . That's exactly how the proof works! But let's prove a discrete version using dyadic scaled $\varepsilon = 2^{-k}$ like a Riemann sum, then move to the continuous version later.

Theorem 8.1.4 (Discrete Dudley's inequality). Let $(X_t)_{t \in T}$ be a mean-zero random process on a metric space (T, d) with subgaussian increments as from earlier. Then

$$\mathbb{E} \left[\sup_{t \in T} X_t \right] \leq CK \sum_{k \in \mathbb{Z}} 2^{-k} \sqrt{\log \mathcal{N}(T, d, \varepsilon)}.$$

The proof uses a technique called *chaining*. It is a multi-scaled version of the ε -net argument that we did in Theorem 4.4.3 and Theorem 7.6.1. In the ε -net argument, we approximate T by an ε -net \mathcal{N} so every point $t \in T$ is close to some $\pi(t) \in \mathcal{N}$, with $d(t, \pi(t)) \leq \varepsilon$. Then the increment condition gives

$$\|X_t - X_{\pi(t)}\|_{\psi_2} \leq K\varepsilon.$$

This leads to

$$\mathbb{E} \left[\sup_{t \in T} X_t \right] \leq \mathbb{E} \left[\sup_{t \in T} X_{\pi(t)} \right] + \mathbb{E} \left[\sup_{t \in T} (X_t - X_{\pi(t)}) \right].$$

We can handle the first term via union bound over $|\mathcal{N}| = \mathcal{N}(T, d, \varepsilon)$ points $\pi(t)$. However, the second term is unclear (if we were to use union bound) since there is both t and $\pi(t)$ in the supremum. To fix this, we don't stop at one net, but choose smaller and smaller ε to get better approximations $\pi_1(t), \pi_2(t), \dots$ to t with finer nets. This is the idea behind *chaining*.

Proof of Theorem 8.1.4. Step 1: Chaining setup. Without loss of generality, we may assume that $K = 1$ (because of C) and T is finite (Remark 7.2.1). Define the dyadic scale

$$\varepsilon_k = 2^{-k}, \quad k \in \mathbb{Z}$$

and choose ε_k -nets T_k of T so that

$$|T_k| = \mathcal{N}(T, d, \varepsilon_k).$$

Only a part of the dyadic scale will be needed. Since T is finite, there exists a small enough number $\kappa \in \mathbb{Z}$ (defining the coarsest net) and a large enough number $K \in \mathbb{Z}$ (defining the finest net), such that

$$T_\kappa = \{t_0\} \text{ for some } t_0 \in T, \quad T_K = T.$$

For a point $t \in T$, let $\pi_l(t)$ denote a closest point in T_k , so we have

$$d(t, \pi_k(t)) \leq \varepsilon_k.$$

Since $\mathbb{E}[X_{t_0}] = 0$ by assumption,

$$\mathbb{E} \left[\sup_{x \in T} X_t \right] = \mathbb{E} \left[\sup_{x \in T} (X_t - X_{t_0}) \right].$$

Let's write $X_t - X_{t_0}$ as a telescoping sum, walking from t_0 to t along a chain (aha!) of points $\pi_k(t)$ that mark progressively finer approximations of t :

$$X_t - X_{t_0} = (X_{\pi_\kappa(t)} - X_{t_0}) + (X_{\pi_{\kappa+1}(t)} - X_{\pi_\kappa(t)}) + \cdots + (X_t - X_{\pi_K(t)}),$$

see Figure 8.2 below for an illustration.

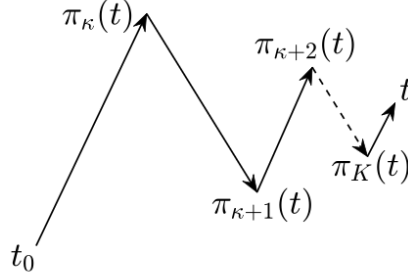


Figure 8.2 Chaining: a walk from a fixed point t_0 to an arbitrary point t in T along elements $\pi_k(T)$ of progressively finer nets of T

The first and last terms of this sum are zero by our definition earlier, so we have

$$X_t - X_{t_0} = \sum_{k=\kappa+1}^K (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}).$$

Since the supremum of the sum is bounded by the sum of the suprema, we get

$$\mathbb{E} \left[\sup_{t \in T} (X_t - X_{t_0}) \right] \leq \sum_{k=\kappa+1}^K \mathbb{E} \left[\sup_{t \in T} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) \right].$$

Step 2: Controlling the increments. In the equation above, it looks like we are taking the supremum over all of T in each summand, but really it is over the smaller set of pairs $(\pi_k(t), \pi_{k-1}(t))$. The number of such pairs is

$$|T_k| \cdot |T_{k-1}| = |T_k|^2,$$

A number that we can control via covering numbers from above. Moreover, for a fixed t , we can bound the increments in step 1 like this:

$$\begin{aligned} \|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\|_{\psi_2} &\leq d(\pi_k(t), \pi_{k-1}(t)) \quad (\text{By Definition 8.1.1}) \\ &\leq d(\pi_k(t), t) + d(t, \pi_{k-1}(t)) \quad (\text{By triangle inequality}) \\ &\leq \varepsilon_k + \varepsilon_{k-1} \quad (\text{By definition of } \pi_k(t)) \\ &\leq 2\varepsilon_{k-1}. \end{aligned}$$

Recall from Proposition 2.7.6 that the expected maximum of N subgaussian random variables is at most $CL\sqrt{\log N}$, where L is the largest ψ_2 norm. We can use this to bound each term:

$$\mathbb{E} \left[\sup_{t \in T} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) \right] \leq C\varepsilon_{k-1} \sqrt{\log |T_k|}.$$

Step 3: Summing up the increments. We have shown that

$$\mathbb{E} \left[\sup_{t \in T} (X_t - X_{t_0}) \right] \leq C \sum_{k=\kappa+1}^K \varepsilon_{k-1} \sqrt{\log |T_k|}.$$

Now plug in the values $\varepsilon_k = 2^{-k}$ and the bounds on $|T_k|$, we get

$$\mathbb{E} \left[\sup_{t \in T} (X_t - X_{t_0}) \right] \leq C_1 \sum_{k=\kappa+1}^K 2^{-k} \sqrt{\log \mathcal{N}(T, d, 2^{-k})}.$$

Hence the theorem is proved. \square

Let's now go for the proof for the integral form of Dudley's inequality.

Proof of Dudley's integral inequality (Theorem 8.1.3). To convert the sum from the discrete form into an integral, we express 2^{-k} as $2 \int_{2^{-k-1}}^{2^{-k}} d\varepsilon$. Then

$$\sum_{k \in \mathbb{Z}} 2^{-k} \sqrt{\log \mathcal{N}(T, d, 2^{-k})} = 2 \sum_{k \in \mathbb{Z}} \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log \mathcal{N}(T, d, \varepsilon)} d\varepsilon.$$

Within the limits of the integral, $2^{-k} \geq \varepsilon$, hence $\log \mathcal{N}(T, d, 2^{-k}) < \log \mathcal{N}(T, d, \varepsilon)$ and the sum is bounded by

$$2 \sum_{k \in \mathbb{Z}} \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log \mathcal{N}(T, d, \varepsilon)} d\varepsilon = 2 \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \varepsilon)} d\varepsilon,$$

and the proof is complete. \square

Actually, the discrete and integral Dudley inequalities are equivalent (Exercise 8.3).

8.1.1 Variations and Examples

Remark 8.1.5 (Dudley's inequality: supremum of increments). A quick look at the proof shows that chaining actually gives

$$\mathbb{E} \left[\sup_{t \in T} |X_t - X_{t_0}| \right] \leq CK \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \varepsilon)} d\varepsilon$$

for any fixed $t \in T$. We can combine with the same bound for $X_s - X_{t_0}$, then use the triangle inequality to get

$$\mathbb{E} \left[\sup_{t, s \in T} |X_t - X_s| \right] \leq CK \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \varepsilon)} d\varepsilon.$$

Remark 8.1.6 (Dudley's inequality: a high-probability bound). Dudley's inequality gives only an expectation bound, but chaining actually gives a high-probability bound. Assuming T is finite (avoid measurability issues), for every $u \geq 0$, the bound

$$\sup_{t, s \in T} |X_t - X_s| \leq CK \left[\int_0^\infty \sqrt{\log \mathcal{N}(T, d, \varepsilon)} d\varepsilon + u \cdot \text{diam}(T) \right]$$

holds with probability at least $1 - 2 \exp(-u^2)$ (Exercise 8.1). For Gaussian processes, this also follows directly from Gaussian concentration (Exercise 8.2).

Remark 8.1.7 (Limits of Dudley integral). Even though the Dudley integral goes over $[0, \infty]$, we can cap it at the diameter of T , since for $\varepsilon > \text{diam}(T)$, a single ε -ball covers T and so

$$\mathcal{N}(T, d, \varepsilon) = 1 \implies \log \mathcal{N}(T, d, \varepsilon) = 0.$$

Thus

$$\mathbb{E} \left[\sup_{t \in T} X_t \right] \leq CK \int_0^{\text{diam}(T)} \sqrt{\log \mathcal{N}(T, d, \varepsilon)} d\varepsilon.$$

If we apply Dudley's inequality for the canonical Gaussian process $\langle g, t \rangle$, just like we did with Sudakov's inequality in Corollary 7.4.2, we get the following:

Theorem 8.1.8 (Dudley's inequality in \mathbb{R}^n). The Gaussian width of any bounded set $Y \subset \mathbb{R}^n$ satisfies

$$w(T) \leq C \int_0^\infty \sqrt{\log \mathcal{N}(T, \varepsilon)} d\varepsilon,$$

where $\mathcal{N}(T, \varepsilon)$ is the smallest number of Euclidean balls with radius ε and centers in T that cover T .

Example 8.1.9 (Dudley's inequality is sharp for the Euclidean ball). Let's test Dudley's inequality for the unit Euclidean ball $T = B_2^n$. From Corollary 4.2.11,

$$\mathcal{N}(B_2^n, \varepsilon) \begin{cases} \leq (3/\varepsilon)^n & \text{for } \varepsilon \in (0, 1], \\ = 1 & \text{for } \varepsilon > 1 \end{cases}.$$

Then

$$w(B_2^n) \lesssim \int_0^1 \sqrt{n \log(3/\varepsilon)} d\varepsilon \lesssim \sqrt{n}.$$

This is in fact optimal: as we know from Example 7.5.6, $w(B_2^n) \asymp \sqrt{n}$.

Remark 8.1.10 (Dudley's inequality can be loose - but not too loose). In general, Dudley integral can overestimate the Gaussian width. Here is a bad example:

$$T = \left\{ \frac{e_k}{\sqrt{1 + \log k}}, k = 1, \dots, n \right\}$$

with e_k being the standard basis in \mathbb{R}^n . From exercise 8.4, we can see that

$$w(T) = O(1) \text{ while } \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \varepsilon)} d\varepsilon \rightarrow \infty$$

as $n \rightarrow \infty$. However, the good news:

- (a) Dudley equality is tight up to a logarithmic factor (Exercise 8.5);
- (b) We will use chaining to remove that logarithmic factor in Section 8.5.

8.2 Application: Empirical Processes

8.3 VC Dimension

VC dimensions measures how complex a class of Boolean functions is, where a Boolean function is a map $f : \omega \rightarrow \{0, 1\}$ on some set ω , and we are looking at some collection \mathcal{F} of these.

Definition 8.3.1. A subset $\Lambda \subseteq \Omega$ is shattered by a class of boolean functions \mathcal{F} if, for any possible binary labeling $g : \Lambda \rightarrow \{0, 1\}$, there is some function $f \in \mathcal{F}$ that matches it on Λ . Formally, this means the restriction of f onto Λ is g , i.e. $f(x) = g(x)$ for all $x \in \Lambda$.

The Vapnik-Chervonenkis dimension (VC dimension) of \mathcal{F} , denoted $\text{vc}(\mathcal{F})$, is the largest cardinality of a subset $\Lambda \subseteq \Omega$ that is shattered. If there is no largest one, then $\text{vc}(\mathcal{F}) = \infty$.

Let's go through a few examples to make the definition clearer:

Example 8.3.2 (Intervals). Let \mathcal{F} consist of the indicators of all closed intervals in \mathbb{R} :

$$\mathcal{F} = \{ \mathbf{1}_{[a, b]} : a, b \in \mathbb{R}, a \leq b \}.$$

We claim that

$$\text{vc}(\mathcal{F}) = 2.$$

We first show that $\text{vc}(\mathcal{F}) \geq 2$ by finding a two-point set $\Lambda \subset \mathbb{R}$ that is shattered by \mathcal{F} . Take, for example, $\Lambda = \{3, 5\}$. There are four possible binary labelings $g : \Lambda \rightarrow \{0, 1\}$ on this set, and each one can be obtained by restricting some interval indicator $f = \mathbf{1}_{[a,b]}$ onto Λ . For example, $g(3) = 1, g(5) = 0$ comes from $f = \mathbf{1}_{[2,4]}$. The other three cases are shown in Figure 8.5, so Λ is indeed shattered by \mathcal{F} .

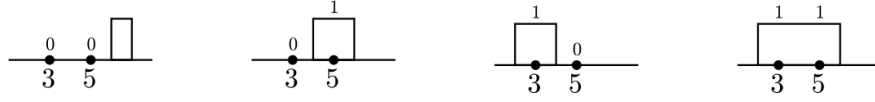


Figure 8.5 The binary function $g(3) = g(5) = 0$ is the restriction of $\mathbf{1}_{[6,7]}$ onto $\Lambda = \{3, 5\}$ (left). The function $g(3) = 0, g(5) = 1$ is the restriction of $\mathbf{1}_{[4,6]}$ onto Λ (middle left). The function $g(3) = 1, g(5) = 0$ is the restriction of $\mathbf{1}_{[2,4]}$ onto Λ (middle right). The function $g(3) = g(5) = 1$ is the restriction of $\mathbf{1}_{[2,6]}$ onto Λ (right).

To prove $\text{vc}(\mathcal{F}) < 3$, we need to show that no three-point set $\Lambda = \{p, q, r\}$ can be shattered by \mathcal{F} . To see this, assume $p < q < r$ and consider the labeling $g(p) = 1, g(q) = 0, g(r) = 1$. Then g cannot be a restriction of any indicator interval onto Λ (it is not linearly separable).

8.4 Application: Statistical Learning Theory

8.5 Generic Chaining

8.6 Chevet Inequality