Chapter 6

Yet even more on sampling

By Sariel Har-Peled, March 17, 2010[①]

> "I don't know why it should be, I am sure; but the sight of another man asleep in bed when I am up, maddens me. It seems to me so shocking to see the precious hours of a man's life - the priceless moments that will never come back to him again - being wasted in mere brutish sleep."
>
> – – Three men in a boat, Jerome K. Jerome

In this chapter, we will extend the sampling results shown in the previous chapter. In the process, we will prove stronger bounds on sampling and provide more general tools for using them. This chapter contains more advanced material and the casual reader will benefit from skipping it.

6.1 Introduction

6.1.1 A not quite new notion of dimension

We are interested in how much information can be extracted by random sampling of a certain size for a range space of VC dimension δ. To make things more interesting, we will extend the notion of VC dimension to real functions.

Definition 6.1.1 Let X be a ground set, and consider a set of functions $\mathcal{F}$ from X to the interval $\left[0,1\right]$. For a set $N = \{\mathsf{p}_1, \ldots, \mathsf{p}_d\} \subseteq \mathsf{X}$ and a set of real values $V = \left\{ v_1, \ldots, v_d \;\middle|\; v_i \in \left[0,1\right] \right\}$, and a function $f \in \mathcal{F}$, consider the subset

$$N_f = \left\{ \mathsf{p}_i \in N \;\middle|\; f(\mathsf{p}_i) \geq v_i \right\}$$

of N induced by f. We will refer to N_f as the ***induced subset*** of N formed by f and V.

We remind the reader that any subset $Y \subseteq S$, can be interpreted as its characteristic function (aka indicator function) $\mathbf{1}_Y : S \to \{0,1\}$, where $x \in Y$ if and only if $\mathbf{1}_Y(x) = 1$. A ***set system***

$(\mathsf{X}, \mathcal{R})$ is a pair, where $\mathcal{R}$ is a set of subsets of X. As such, given a set system $(\mathsf{X}, \mathcal{R})$, we can always interpret $\mathcal{R}$ as a set of characteristic functions of these sets. All the induced subsets would just be $\mathcal{R}$, if we set the threshold values $v_1, \ldots, v_d$ to all be 1. Thus, replacing the sets of $\mathcal{R}$ by real functions is a natural extension of the notion of a set system.

We next extend the notion of shattering.

Definition 6.1.2 Let X be a ground set, and consider a set of functions $\mathcal{F}$ from X to the interval $\Big[0, 1\Big]$. A subset $N = \{\mathsf{p}_1, \ldots, \mathsf{p}_d\} \subseteq \mathsf{X}$, with associated values $V = \{v_1, \ldots, v_d\}$ is ***shattered*** by $(\mathsf{X}, \mathcal{F})$ (or by $\mathcal{F}$) if

$$P_N = \left\{ N_f \;\middle|\; f \in \mathcal{F} \right\}$$

contains all possible (i.e., 2^d) subsets of S.

In particular, we now have a natural extension of the notion of dimension.

Definition 6.1.3 Given $\mathsf{S} = (\mathsf{X}, \mathcal{F})$ as above, its ***pseudo-dimension***, denoted by $\operatorname{pDim}(\mathcal{F})$, is the size of the largest subset of X that is shattered by $\mathcal{F}$.

Definition 6.1.4 Given a function $f \in \mathcal{F}$, and a probability distribution $\mathcal{D}$ on X, consider the average value of f on X. The ***measure*** of f is

$$\overline{m}(f) = \sum_{\mathsf{p} \in \mathsf{X}} \mathbf{Pr}\Big[\mathsf{p}\Big] f(\mathsf{p}) \qquad \text{and} \qquad \overline{m}(f) = \int_{\mathsf{p} \in \mathsf{X}} f(\mathsf{p})\, d\mathcal{D},$$

if the ground set is finite or infinite, respectively.

Intuitively, if $f = \mathbf{1}_Y$ is an indicator function of a set $Y \subseteq \mathsf{X}$, then this is no more than the measure of y in X; namely, if the underlying distribution is uniform, then $\overline{m}(\mathbf{1}_Y) = |Y| \,/\, |\mathsf{X}|$. Again, we are assuming here that X is finite, but the discussion can easily be extended to infinite domains, by replacing summation by integration. As usual, we are going to ignore this tedious (and somewhat insignificant) technicality.

Definition 6.1.5 Similarly, for an ordered sample $N = \langle \mathsf{p}_1, \ldots, \mathsf{p}_m \rangle$ of m points from X, and a function $f \in \mathcal{F}$, let

$$\overline{s} = \overline{s}_N(f) = \frac{1}{m} \sum_{i=1}^{m} f(\mathsf{p}_i)\,,$$

denote the ***estimate*** of f by N. The quantity $\overline{s}_N(f)$ can be interpreted as approximation to $\overline{m}(f)$ by the sample N.

Remark 6.1.6 Before the reader gets confused and impressed by the claptrap of pseudo-dimension, observe that we can easily construct a range space with the same VC dimension, so that we can work directly on this "alternative" range space. Indeed, let $\mathsf{X}' = \mathsf{X} \times [0, 1]$ and $\mathcal{G} = \left\{ \widehat{\mathbf{r}}(f) \;\middle|\; f \in \mathcal{F} \right\}$, where

$$\widehat{\mathbf{r}}(f) = \left\{ (x, y) \;\middle|\; x \in \mathsf{X}, y \in [0, 1] \text{ and } y \leq f(x) \right\}. \tag{6.1}$$

Now, consider the range space $\mathsf{T} = (\mathsf{X}', \mathcal{G})$.

A point $\mathsf{p} \in N$ with associated value v is in $\overline{s}_N(f)$ if and only if $f(\mathsf{p}) \geq v$. Namely, the point (p, v) lies below the graph of f. This is equivalent to $(\mathsf{p}, v) \in \widehat{\mathbf{r}}(f)$. As such, $\mathsf{S} = (\mathsf{X}, \mathcal{F})$ has pseudo-dimension δ if and only if T has VC dimension δ. Indeed, if S shatters the set $N = \{\mathsf{p}_1, \ldots, \mathsf{p}_d\} \subseteq \mathsf{X}$ and its associated real values $V = \left\{ v_1, \ldots, v_d \;\middle|\; v_i \in \left[0, 1\right] \right\}$, then T shatters the set of points $\left\{ (\mathsf{p}_1, v_1), \ldots, (\mathsf{p}_1, v_d) \right\}$.

Observe, also, that a distribution $\mathcal{D}$ over X induces a natural distribution on X'. Indeed, to pick a point randomly in the set $\mathsf{X} \times [0, 1]$, pick randomly a point $x \in \mathsf{X}$ according to the given distribution, and then pick uniformly $y \in [0, 1]$. The resulting pair (x, y) is in X', and this process defines a natural distribution on the elements of X'. In fact, given $f \in \mathcal{F}$, we have that $\overline{m}(f)$ is equal to the measure of $\widehat{\mathbf{r}}(f)$ in the ground set X'. Indeed, being slightly informal, we have that

$$\begin{aligned}\overline{m}(\widehat{\mathbf{r}}(f)) &= \int_{(x,y)\in\mathsf{X}'} \mathbf{Pr}[(x,y)]\, \mathbf{1}_{\widehat{\mathbf{r}}(f)}\Big((x,y)\Big)\, dx\, dy = \int_{x\in\mathsf{X}} \int_{y\in[0,1]} \mathbf{1}_{\widehat{\mathbf{r}}(f)}\Big((x,y)\Big)\, dy\, d\mathcal{D} \\ &= \int_{x\in\mathsf{X}} f(x)\; d\mathcal{D} = \overline{m}(f)\,.\end{aligned}$$

Thus, we immediately get the ε-net and ε-approximation theorems for the pseudo-dimension settings (since we proved these theorems for the VC dimension case).

6.1.2 The result

For a parameter $\nu > 0$, consider the distance function between two real numbers $r \geq 0$ and $s \geq 0$ defined as

$$d_\nu(r, s) = \frac{|r - s|}{r + s + \nu}.$$

It is easy to verify that (i) $0 \leq d_\nu(r, s) < 1$, (ii) $d_\nu(r, s) \leq d_\nu(u, v)$, for $u \leq r \leq s \leq v$, and (iii) $d_\nu(x, y) > 0$ if $x \neq y$.

We will elaborate on the properties of this distance function in Section 6.1.3 below.

Our purpose in this chapter, is to prove the following theorem.

Theorem 6.1.7 *Let $\alpha, \nu, \varphi > 0$ be parameters, and let $\mathsf{S} = (\mathsf{X}, \mathcal{F})$ be a range space, and $\mathcal{F}$ be a set of functions from X to $\left[0, 1\right]$, such that the pseudo-dimension of S is δ. We have, that for a random sample N (with repetition) from X of size*

$$O\left(\frac{1}{\alpha^2 \nu}\left(\delta \log \frac{1}{\nu} + \log \frac{1}{\varphi}\right)\right),$$

it holds

$$\forall f \in \mathcal{F} \quad d_\nu\Big(\overline{m}(f)\,,\, \overline{s}_N(f)\Big) < \alpha.$$

And this holds with probability $\geq 1 - \varphi$.

Before providing a proof of this theorem, we first demonstrate how this theorem implies the ε-net/sampling theorems, and in fact provide also some other results. But first, we must get a better handle on the distance function $d_\nu(\cdot, \cdot)$.

6.1.3 The curious distance function between real numbers

In the following, we will develop a unified result that encompass both the ε-net and ε-sample theorem. For technical reasons, it is convenient to work with the distance function described here between numbers in describing and deriving this result.

The proof of the following lemma is tedious, and is left to Exercise 6.4.2.

Lemma 6.1.8 *The triangle inequality holds for $d_\nu(\cdot,\cdot)$. Namely, for any $x, y, z \geq 0$ and $\nu > 0$, we have $d_\nu(x,y) + d_\nu(y,z) \geq d_\nu(x,z)$.*

Lemma 6.1.9 *The function $d_\nu(\cdot,\cdot)$ is a metric.*

Proof: Indeed, for any $x > 0$, we have $d_\nu(x,x) = 0$, and for any $x, y > 0$, it holds $d_\nu(x,y) = d_\nu(y,x)$. Finally, the triangle inequality is implied by Lemma 6.1.8. ■

Lemma 6.1.10 *Let $\alpha, \nu, \overline{m}, \overline{s}$ be non-negative numbers. Then $d_\nu(\overline{m}, \overline{s}) < \alpha$, if and only if $\overline{s} \in \mathcal{I} = (u_l, u_r)$, where*

$$u_l = \left(1 - \frac{2\alpha}{1+\alpha}\right)\overline{m} - \frac{\alpha\nu}{(1+\alpha)} \qquad \text{and} \qquad u_r = \left(1 + \frac{2\alpha}{1-\alpha}\right)\overline{m} + \frac{\alpha\nu}{(1+\alpha)}.$$

Proof: Indeed, we have $d_\nu\left(\overline{m}, \overline{s}\right) = \dfrac{|\overline{m} - \overline{s}|}{\overline{m} + \overline{s} + \nu} < \alpha$. If $\overline{m} > \overline{s}$ then this implies that

$$\overline{m} - \overline{s} < \alpha(\overline{m} + \overline{s} + \nu) \iff (1-\alpha)\,\overline{m} - \alpha\nu < (1+\alpha)\,\overline{s} \iff \frac{(1-\alpha)}{(1+\alpha)}\overline{m} - \frac{\alpha\nu}{(1+\alpha)} < \overline{s}$$
$$\iff u_l = \left(1 - \frac{2\alpha}{1+\alpha}\right)\overline{m} - \frac{\alpha\nu}{(1+\alpha)} < \overline{s}.$$

Since $\overline{m} \in \mathcal{I}$, this implies that $u_l < \overline{s} < \overline{m} \leq u_r$; namely, $\overline{s} \in \mathcal{I}$.

If $\overline{m} < \overline{s}$ then this implies that

$$\overline{s} - \overline{m} < \alpha(\overline{m} + \overline{s} + \nu) \iff (1-\alpha)\,\overline{s} < \alpha\nu + (1+\alpha)\,\overline{m} \iff \overline{s} < \frac{(1+\alpha)}{(1-\alpha)}\overline{m} + \frac{\alpha\nu}{(1+\alpha)}$$
$$\iff \overline{s} < u_r = \left(1 + \frac{2\alpha}{1-\alpha}\right)\overline{m} + \frac{\alpha\nu}{(1+\alpha)}.$$

Again, this implies that $u_l \leq \overline{m} \leq \overline{s} < u_r$, implying that $\overline{s} \in \mathcal{I}$. ■

Corollary 6.1.11 *For any real numbers, $\nu, \alpha, \overline{m}, \overline{s} > 0$, we have:*

(i) If $|\overline{s} - \overline{m}| \leq \Delta = (2\overline{m} + \nu)\dfrac{\alpha}{1+\alpha}$ then $d_\nu(\overline{m}, \overline{s}) < \alpha$.

(ii) If $d_\nu(\overline{m}, \overline{s}) < \alpha$ then $|\overline{s} - \overline{m}| \leq \Delta' = \dfrac{2\alpha}{1-\alpha}\overline{m} + \dfrac{\alpha\nu}{1+\alpha}$.

Name	Property $\forall \mathbf{r} \in \mathcal{F}$	Sample size
ε-net [HW87] Theorem 6.2.1	$\overline{m} = \overline{m}(r)$, $\overline{s} = \overline{s}(r)$ $\overline{m} \geq \varepsilon \quad \Rightarrow \quad \overline{s} > 0$	$O\left(\frac{\delta}{\varepsilon} \log \frac{1}{\varepsilon}\right)$
ε-approximation [VC71] Theorem 6.2.2	$\lvert\overline{m} - \overline{s}\rvert \leq \varepsilon$	$O\left(\frac{\delta}{\varepsilon^2}\right)$
Sensitive ε-approximation [Brö95, BCM99] Theorem 6.2.4	$\lvert\overline{m} - \overline{s}\rvert \leq \frac{\varepsilon}{2}\left(\sqrt{\overline{m}} + \varepsilon\right)$	$O\left(\frac{\delta}{\varepsilon^2} \log \frac{1}{\varepsilon}\right)$
Relative (ε, p)-approximation [CKMS06] Theorem 6.2.6	$\overline{m} \leq p \quad \Rightarrow \quad \overline{s} \leq (1+\varepsilon)p$ $\overline{m} \geq p \quad \Rightarrow$ $(1-\varepsilon)\overline{m} \leq \overline{s} \leq (1+\varepsilon)\overline{m}$	$O\left(\frac{\delta}{\varepsilon^2 p} \log \frac{1}{p}\right)$

Figure 6.1: The results hold for any range $\mathbf{r}$ in the given range space that have VC dimension δ, where $\overline{m} = \overline{m}(\mathbf{r})$ is the weight of $\mathbf{r}$, and $\overline{s} = \overline{s}(\mathbf{r})$ is its weight in the random sample. The samples have the required property (for all the ranges in the range space) with constant probability.

6.2 Applications

The following assumes that the reader is familiar and comfortable with ε-nets and ε-samples (see the previous chapter). The results implied by Theorem 6.1.7 are summarized in Figure 6.1.

We are given a range space $\mathsf{S} = (\mathsf{X}, \mathcal{F})$ of VC dimension δ, where X is a point set, and $\mathcal{F}$ is a set of ranges of X. In our settings, we will usually consider a finite subset $\mathsf{x} \subseteq \mathsf{X}$ and we will be interested in the range space induced by S on x. In particular, let N be a sample of X. For a range $\mathbf{r} \in \mathcal{F}$, let

$$\overline{m} = \overline{m}_{\mathsf{x}}(\mathbf{r}) = \frac{|\mathbf{r} \cap \mathsf{x}|}{|\mathsf{x}|} \quad \text{and} \quad \overline{s} = \overline{s}_N(\mathbf{r}) = \frac{|\mathbf{r} \cap N|}{|N|}.$$

Intuitively, $\overline{m}$ is the total weight of $\mathbf{r}$ in x, while $\overline{s}$ is the sample estimate for $\mathbf{r}$.

6.2.1 Getting the ε-net and ε-approximation theorems

We remind the reader that $N \subseteq \mathsf{x}$ is an ***ε-net*** for a finite range space$(\mathsf{x}, \mathcal{F})$, if for every $\mathbf{r} \in \mathcal{F}$, such that $\overline{m}_{\mathsf{x}}(\mathbf{r}) \geq \varepsilon$ implies that $N \cap \mathbf{r} \neq \emptyset$.

Theorem 6.2.1 ([HW87], ε-Net Theorem) *Let $\varphi, \varepsilon > 0$ be parameters and $\mathsf{S} = (\mathsf{X}, \mathcal{F})$ be a range space with VC dimension δ. Let $\mathsf{x} \subset \mathsf{X}$ be a finite subset. Then, a sample of size*

$$O\left(\frac{1}{\varepsilon}\left(\delta \log \frac{1}{\varepsilon} + \log \frac{1}{\varphi}\right)\right)$$

from x, is an ε-net for x with probability $\geq 1 - \varphi$.

Proof: Let $\alpha = 1/4$, $\nu = \varepsilon$, and apply Theorem 6.1.7. The sample size is

$$O\left(\frac{1}{\alpha^2 \nu}\left(\delta \log \frac{1}{\nu} + \log \frac{1}{\varphi}\right)\right) = O\left(\frac{1}{\varepsilon}\left(\delta \log \frac{1}{\varepsilon} + \log \frac{1}{\varphi}\right)\right).$$

Now, let $\mathbf{r} \in \mathcal{F}$ be a range such that $\overline{m} = \overline{m}(\mathbf{r}) \geq \varepsilon$. We have that $d_\nu(\overline{m}, \overline{s}) < \alpha = 1/4$, where $\overline{s} = \overline{s}(\mathbf{r})$ (and this holds with probably $\geq 1 - \varphi$ for all ranges). Then, by Corollary 6.1.11 (ii), we have that $\overline{s} \in (\overline{m} - \Delta', \overline{m} + \Delta')$, where

$$\Delta' = \frac{2\alpha}{1-\alpha}\overline{m} + \frac{\alpha\nu}{1+\alpha} = \frac{2}{3}\overline{m} + \frac{\varepsilon}{5}.$$

For $\overline{m} \geq \varepsilon$, we have $\overline{s} \geq \overline{m} - \Delta' = \overline{m}/3 - \varepsilon/5 \geq \varepsilon/15 > 0$, implying the claim. ∎

We remind the reader that $N \subseteq \mathsf{x}$ is an ***ε-sample*** for a finite range space $(\mathsf{x}, \mathcal{F})$, if for every $\mathbf{r} \in \mathcal{F}$, we have that $|\overline{m}_{\mathsf{x}}(\mathbf{r}) - \overline{s}_N(\mathbf{r})| \leq \varepsilon$.

Theorem 6.2.2 ([VC71], ε-sample theorem.) *Let $\varphi, \varepsilon > 0$ be parameters and $\mathsf{S} = (X, \mathcal{F})$ be a range space with VC dimension δ. Let $\mathsf{x} \subset X$ be a finite subset. A sample of size*

$$O\left(\frac{1}{\varepsilon^2}\left(\delta + \log\frac{1}{\varphi}\right)\right)$$

from x, is an ε-sample for $\mathsf{S} = (\mathsf{x}, \mathcal{F})$ with probability $\geq 1 - \varphi$.

Proof: Set $\alpha = \varepsilon/4$ and $\nu = 1/4$. We have, by Theorem 6.1.7, that for any $\mathbf{r} \in \mathcal{F}$, it holds

$$|r - s| \leq \frac{\varepsilon}{4}(r + s + \nu) \leq \varepsilon,$$

implying the claim. ∎

6.2.2 Sensitive ε-approximation

Another similar concept was introduced by [BCM99].

Definition 6.2.3 A sample $N \subseteq \mathsf{x}$ is a ***sensitive ε-approximation*** if

$$\forall \mathbf{r} \in \mathcal{F} \quad |\overline{m}(\mathbf{r}) - \overline{s}(\mathbf{r})| \leq \frac{\varepsilon}{2}\left(\sqrt{\overline{m}(\mathbf{r})} + \varepsilon\right).$$

Observe that a set N which is a sensitive ε-approximation is, simultaneously, both an ε^2-net and an ε-approximation.

The following theorem shows the existence of a asensitive ε-approximation. Note that the bound on its size is (slightly) better than the bound shown by [Brö95, BCM99].

Theorem 6.2.4 *A sample N from x of size*

$$O\left(\frac{1}{\varepsilon^2}\left(\delta \log\frac{1}{\varepsilon} + \log\frac{1}{\varphi}\right)\right).$$

is a sensitive ε-approximation, with probability $\geq 1 - \varphi$.

Proof: Let $\nu_i = i\varepsilon^2/c$, $\alpha_i = \sqrt{1/4i}$, for $i = 1, \ldots, M = \left\lceil c/\varepsilon^2 \right\rceil$, where c is some sufficiently large constant. As such, for $i = 1, \ldots, M$, we have that $\alpha_i^2 \nu_i = \varepsilon^2/(4c)$. Consider a single random sample N of size

$$U = O\left(\frac{1}{\varepsilon^2}\left(\delta \log \frac{1}{\varepsilon} + \log \frac{M}{\varphi}\right)\right) = O\left(\frac{1}{\varepsilon^2}\left(\delta \log \frac{1}{\varepsilon} + \log \frac{1}{\varphi}\right)\right).$$

For a fixed i, this sample comply with Theorem 6.1.7, with parameters ν_i and α_i, with probability at least $1 - \varphi/M$, since $O\left(\frac{1}{\alpha_i^2 \nu_i}\left(\delta \log \frac{1}{\nu_i} + \log \frac{M}{\varphi}\right)\right) = O(U)$.

In particular, Theorem 6.1.7 holds for N, with probability at least φ, for parameters α_i and ν_i, for all $i = 1, \ldots, M$. Indeed, the probability that it fails from any of value of i is bounded by $M(\varphi/M) = \varphi$.

Next, consider a range $\mathbf{r} \in \mathcal{F}$, such that $\overline{m} = \overline{m}(\mathbf{r}) \in [(i-1)\varepsilon^2/800, i\varepsilon^2/800]$ and $\overline{s} = \overline{s}(\mathbf{r})$.

If $i > 1$, we have that $\nu_i/2 \leq \overline{m} \leq \nu_i$, and as such

$$\alpha_i \overline{m} \leq \alpha_i \nu_i = \sqrt{\alpha_i^2 \nu_i \nu_i} \leq \sqrt{\alpha_i^2 \nu_i}\sqrt{2\overline{m}} = \sqrt{\frac{\varepsilon^2}{4c}}\sqrt{2\overline{m}} \leq \frac{\varepsilon\sqrt{\overline{m}}}{20}, \tag{6.2}$$

by making c sufficiently large. Now, we have that $d_{\nu_i}(\overline{s}, \overline{m}) \leq \alpha_i$, which implies, by Corollary 6.1.11, that

$$|\overline{s} - \overline{m}| \leq \Delta' = \frac{2\alpha_i}{1-\alpha_i}\overline{m} + \frac{\alpha_i \nu_i}{1+\alpha_i} \leq 4\alpha_i \overline{m} + \alpha_i \nu_i \leq 6\alpha_i \overline{m} \leq \varepsilon\sqrt{\overline{m}} \leq \frac{\varepsilon}{2}\Big(\sqrt{\overline{m}(\mathbf{r})} + \varepsilon\Big),$$

since $\alpha_i \leq 1/2$. For $i = 1$, we have $\overline{m} \leq \nu_1$, and

$$|\overline{s} - \overline{m}| \leq \Delta' = \frac{2\alpha_1}{1-\alpha_1}\overline{m} + \frac{\alpha_1 \nu_1}{1+\alpha_1} \leq 6\alpha_1 \nu_1 = 6\frac{\varepsilon^2}{4c} \leq \frac{\varepsilon^2}{2} \leq \frac{\varepsilon}{2}\Big(\sqrt{\overline{m}(\mathbf{r})} + \varepsilon\Big),$$

by picking $c \geq 3$. ∎

Looking on the bounds of sensitive ε-approximation as compared to ε-approximation, it's natural to ask whether its size can be improved, but observe that since such a sample is also an ε^2-net, and it is known that $\Omega\Big((\delta/\varepsilon^2)\log(1/\varepsilon)\Big)$ is a lower bound on the size of such a net [KPW92], this implies that such an improvement is impossible.

6.2.3 Relative ε-approximation

Definition 6.2.5 A subset $N \subset \mathsf{x}$ is a *relative* (p, ε)*-approximation* if for each $\mathbf{r} \in \mathcal{F}$, we have:

(i) If $\overline{m}(\mathbf{r}) \geq p$ then $(1-\varepsilon)\overline{m}(\mathbf{r}) \leq \overline{s}(\mathbf{r}) \leq (1+\varepsilon)\overline{m}(\mathbf{r})$.

(ii) If $\overline{m}(\mathbf{r}) \leq p$ then $\overline{s}(\mathbf{r}) \leq (1+\varepsilon)p$.

The concept was introduced by [CKMS06], except that property (ii) was not required. However, property (ii) is just an easy (but useful) "monotonicity" property that holds for all the constructions I am aware of.

There are relative approximations of size (roughly) $1/(\varepsilon^2 p)$. As such, relative approximation is interesting in the case where $p \ll \varepsilon$. Then, we can approximate ranges of weight larger than p with a sample that has only linear dependency on $1/p$. Otherwise, we would have to use the regular p-sample, and there the required sample is of size (roughly) $1/p^2$.

Theorem 6.2.6 *A sample N of size $O\left(\frac{1}{\varepsilon^2 p}\left(\delta \log \frac{1}{p} + \log \frac{1}{\varphi}\right)\right)$ is a relative (p,ε)-approximation with probability $\geq 1-\varphi$.*

Proof: Set $\nu = p/2$, and $\alpha = \varepsilon/9$, and apply Theorem 6.1.7. We get that, for any range $\mathbf{r} \in \mathcal{F}$, such that $\overline{m} = \overline{m}(\mathbf{r})$ and $s = \overline{s}(\mathbf{r})$, by Corollary 6.1.11 (ii), it holds that

$$|\overline{s} - \overline{m}| \leq \Delta' = \frac{2\alpha}{1-\alpha}\overline{m} + \frac{\alpha\nu}{1+\alpha} = \frac{2\varepsilon/9}{1-\varepsilon/9}\overline{m} + \frac{p\varepsilon}{18(1+\varepsilon/9)} \leq \frac{\varepsilon}{4}\overline{m} + \frac{p\varepsilon}{4},$$

where $\overline{s} = \overline{s}(\mathbf{r})$. For $\overline{m} \geq p$ this implies that $|\overline{s} - \overline{m}| \leq \varepsilon\overline{m}$. Similarly, if $\overline{m} \leq p$ then $\overline{s} \leq \overline{m} + p\varepsilon/2 \leq (1+\varepsilon)p$. ∎

In fact, one can slightly strengthen the concept by making it "sensitive".

Theorem 6.2.7 *A sample N of size $O\left(\frac{1}{\varepsilon^2 p}\left(\delta \log \frac{1}{p} + \log \frac{1}{\varphi}\right)\right)$ is a relative $\left(ip, \varepsilon/\sqrt{i}\right)$-approximation with probability $\geq 1-\varphi$, for all $i \geq 0$.*

Namely, for any range $\mathbf{r} \in \mathcal{F}$, such that $\overline{m}(\mathbf{r}) \geq ip$, we have

$$\left(1 - \frac{\varepsilon}{\sqrt{i}}\right)\overline{m}(\mathbf{r}) \leq \overline{s}(\mathbf{r}) \leq \left(1 + \frac{\varepsilon}{\sqrt{i}}\right)\overline{m}(\mathbf{r}). \tag{6.3}$$

Proof: Set $p_i = ip$ and $\varepsilon_i = \varepsilon/\sqrt{i}$, for $i = 1, \ldots, 1/p$. Now, apply Theorem 6.2.6, and observe that all the samples need are asymptotically of the same size; that is

$$O\left(\frac{1}{\varepsilon_i^2 p_i}\left(\delta \log \frac{1}{p_i} + \log \frac{1}{p\varphi}\right)\right) = O\left(\frac{1}{\varepsilon^2 p}\left(\delta \log \frac{1}{ip} + \log \frac{1}{p\varphi}\right)\right) = O\left(\frac{1}{\varepsilon^2 p}\left(\delta \log \frac{1}{p} + \log \frac{1}{\varphi}\right)\right).$$

As such, one can use the same sample to get this guarantee for all i, and the probability of this sample to fail for any i is at most $(1/p)p\varphi = \varphi$, as desired. ∎

Interestingly, sensitive approximation imply relative approximations.

Lemma 6.2.8 *Let $\varepsilon, p > 0$ be parameters, and let $\varepsilon' = \varepsilon\sqrt{p}$. Then, if N is sensitive ε'-approximation to the set system $(\mathsf{X}, \mathcal{F})$ then its also a relative (ε, p)-approximation.*

Proof: We know that $\forall \mathbf{r} \in \mathcal{F}$ it holds $|\overline{m}(\mathbf{r}) - \overline{s}(\mathbf{r})| \leq \frac{\varepsilon'}{2}\left(\sqrt{\overline{m}(\mathbf{r})} + \varepsilon'\right)$. As such, for $\mathbf{r} \in \mathcal{F}$, if $\overline{m}(\mathbf{r}) = \alpha p$ and $\alpha \geq 1$, then we have

$$|\overline{m}(\mathbf{r}) - \overline{s}(\mathbf{r})| \leq \frac{\varepsilon\sqrt{p}}{2}\left(\sqrt{\alpha p} + \varepsilon\sqrt{p}\right) = \frac{\varepsilon^2 p}{2} + \frac{\varepsilon}{2}\sqrt{\alpha}p \leq \left(\frac{\varepsilon^2}{2} + \frac{\varepsilon}{2}\right)\alpha p \leq \varepsilon\overline{m}(\mathbf{r}),$$

since $\varepsilon < 1$. This implies that N is a relative (ε, p)-approximation. ∎

6.3 Proof of Theorem 6.1.7

6.3.1 Why the sample works for a single function

Let us start by proving the claim for a single range. That is, we want to show that a random ample estimates a single range correctly. In this case, even this is not trivial.

We remind the reader that the ***variance*** of a random variable X is defined to be the quantity $\mathbf{V}[X] = \mathbf{E}\Big[(X - \mathbf{E}[X])^2\Big] = \mathbf{E}\Big[X^2\Big] - \mathbf{E}[X]^2$.

Lemma 6.3.1 *Let X be a random variable in the range $I = [0, M]$, with expectation μ. Then $\mathbf{V}[X] \leq \mu(M - \mu)$.*

Proof: For simplicity of exposition assume X is discrete. We then have that

$$\mathbf{V}[X] = \mathbf{E}\Big[X^2\Big] - \mu^2 = \sum_x x^2 \mathbf{Pr}[X = x] - \mu^2 \leq M \sum_x x \mathbf{Pr}[X = x] - \mu^2 = M\mu - \mu^2 = \mu(M - \mu).$$

■

Lemma 6.3.2 *Let $X_1, \ldots, X_\xi \in [0, M]$ be ξ random independent variables each with expectation μ. Let $Y = \sum_{i=1}^{\xi} X_i/\xi$. Then $\mathbf{E}[Y] = \mu$ and $\mathbf{V}[Y] \leq \mu(M - \mu)/\xi$.*

Proof: For any constant c, we have $\mathbf{V}[cX] = c^2 \mathbf{V}[X]$, and for two independent variables X and Y, we have $\mathbf{V}[X + Y] = \mathbf{V}[X] + \mathbf{V}[Y]$. As such, we have

$$\mathbf{V}[Y] = \frac{\mathbf{V}[X_1] + \ldots + \mathbf{V}\Big[X_\xi\Big]}{\xi^2} \leq \frac{\xi\,\mu\,(M - \mu)}{\xi^2} \leq \frac{\mu\,(M - \mu)}{\xi}.$$

■

Lemma 6.3.3 *Let $f : X \to [0, 1]$ be a function, and let N be a random sample (with repetition) from X of size ξ. Let $\overline{s} = \overline{s}_N(f)$ and $\overline{m} = \overline{m}(f)$. Then $\mathbf{E}[\overline{s}] = \overline{m}$ and $\mathbf{V}[\overline{s}] \leq \overline{m}(1 - \overline{m})/\xi$.*

Proof: Let X_i be the value of f on the ith sample of N, for $i = 1, \ldots, \xi$. Clearly, $\mathbf{E}[X_i] = \overline{m}(f)$. Observing that $\overline{s} = (\sum_{i=1}^{\xi} X_i)/\xi$ and using Lemma 6.3.2 implies the result. ■

Lemma 6.3.4 *Let α, ν be parameters, X a ground set, and f a function from X to $\Big[0, 1\Big]$. Then, for a random sample N (with repetition) from X of size $\xi = O\left(\frac{1}{\alpha^2 \nu}\right)$, it holds, with probability $\geq 3/4$, that $d_\nu(\overline{m}, \overline{s}) < \alpha/2$, where $\overline{m} = \overline{m}(f)$ and $\overline{s} = \overline{s}_N(f)$.*

Proof: Corollary 6.1.11 implies that if

$$|\overline{s} - \overline{m}| \leq \Delta = (2\overline{m} + \nu)\,\frac{\alpha/2}{1 + \alpha/2}$$

then $d_\nu(\overline{m}, \overline{s}) < \alpha/2$. As such, we need to bound the probability that $|\overline{s} - \overline{m}| \leq \Delta$. To this end, we will use Chebychev's inequality. Formally, as in the proof of Lemma 6.3.3, let $\overline{s} = (\sum_{i=1}^{\xi} X_i)/\xi$,

where X_i is the value of f on the ith sample point of N. Assume, for the time being, that for $t = 2$, it holds $\Delta \geq t\sqrt{\mathbf{V}[\overline{s}]}$. Then, by Chebychev's inequality (Theorem 6.6.5) applied to $\overline{s}$, we have that

$$\mathbf{Pr}[d_\nu(\overline{m}, \overline{s}) < \alpha/2] \geq \mathbf{Pr}[|\overline{s} - \overline{m}| \leq \Delta] \geq \mathbf{Pr}\left[|\overline{s} - \overline{m}| \leq t\sqrt{\mathbf{V}[\overline{s}]}\right] \geq 1 - \frac{1}{t^2} = \frac{3}{4},$$

since $\mathbf{E}[\overline{s}] = \overline{m}$.

Now, consider the quantity $t\sqrt{\mathbf{V}[\overline{s}]}$ (for $t = 2$) and observe that by Lemma 6.3.3, the variance of $\mathbf{V}[\overline{s}]$ (which is bounded by $\overline{m}(1-\overline{m})/\xi$) decreases as ξ increases, since $0 \leq \overline{m} \leq 1$. As such, we would like to pick ξ as small as possible such that $t\sqrt{\mathbf{V}[\overline{s}]} \leq 2\sqrt{\overline{m}(1-\overline{m})/\xi} \leq \Delta$. The last inequality can be rewritten as

$$2\sqrt{\frac{\overline{m}(1-\overline{m})}{\xi}} \leq \Delta = \frac{\alpha/2}{1+\alpha/2}(2\overline{m}+\nu) \quad \Longleftrightarrow \quad 2 \leq (2\overline{m}+\nu)\frac{\alpha/2}{1+\alpha/2}\sqrt{\frac{\xi}{\overline{m}(1-\overline{m})}}.$$

However, the last inequality holds for $\xi \geq 16/(\alpha^2\nu)$, since

$$2 \leq \frac{\sqrt{\alpha^2\nu\xi}}{2} \leq \frac{\alpha/2}{2}\sqrt{\frac{4\,\xi\,\overline{m}\nu}{\overline{m}(1-\overline{m})}} \leq \frac{\alpha/2}{1+\alpha/2}\sqrt{\frac{\xi\,(2\overline{m}+\nu)^2}{\overline{m}(1-\overline{m})}} \leq (2\overline{m}+\nu)\frac{\alpha/2}{1+\alpha/2}\sqrt{\frac{\xi}{\overline{m}(1-\overline{m})}}.$$

We conclude that for $\xi \geq 16/(\alpha^2\nu)$ it holds $\mathbf{Pr}[d_\nu(\overline{m}, \overline{s}) < \alpha/2] \geq 3/4$, as desired. ∎

6.3.2 Reduction to double sampling

In the above we proved that for a single function of $\mathcal{F}$ is estimated correctly with reasonable probability. To prove Theorem 6.1.7, we need to extend this to all the functions in $\mathcal{F}$. Specifically, we need to bound the probability that there exists a function $f \in \mathcal{F}$ such that the random sample S (of size m) fails to estimate it correctly, see Theorem 6.1.7.

The problem is, of course, that $\mathcal{F}$ is an infinite family. What we do in the following, is to reduce bounding the probability of a bad event over an infinite family, into a finite event that we can bound the probability of using a direct combinatorial argument.

To this end, for $f \in \mathcal{F}$, let $\mathcal{E}_{1,f}$ denotes the event the sample failed for the function f; specifically, we have

$$\mathcal{E}_{1,f} \quad \equiv \quad d_\nu\Big(\overline{m}(f)\,, \overline{s}_S(f)\Big) > \alpha.$$

The basic events in the probability space here are all the possible values of S, as such $\mathcal{E}_{1,f}$ is the set of values of S for which f is not being estimated correctly. Giving in to the urge to be formal, we have that $\mathcal{E}_{1,f} = \left\{ S \;\middle|\; S \in \mathsf{X}^m \text{ and } d_\nu(\overline{m}(f)\,, \overline{s}_S(f)) > \alpha \right\}$. For the sake of simplicity of exposition, we will try to be slightly less formal, but it is a good idea to keep track of the underlying probability space that we are arguing about.

As such, the bad event that S fails for any function of $\mathcal{F}$ is the event

$$\mathcal{E}_1 = \bigcup_{f \in \mathcal{F}} \mathcal{E}_{1,f},$$

and our purpose here is to bound $\mathbf{Pr}[\mathcal{E}_1]$.

Let T be a second sample of size m, and let

$$\mathcal{E}_{T \text{ good est. of } f} \quad \equiv \quad d_v\Big(\overline{m}, \overline{s}_T(f)\Big) < \frac{\alpha}{2}$$

be the event that f is being estimated correctly by T. Next, let

$$\mathcal{E}_{2,f} \quad \equiv \quad \mathcal{E}_{1,f} \cap \mathcal{E}_{T \text{ good est. of } f} = d_v\Big(\overline{m}, \overline{s}_S(f)\Big) > \alpha \quad \text{and} \quad d_v\Big(\overline{m}, \overline{s}_T(f)\Big) < \frac{\alpha}{2}.$$

By Lemma 6.3.4,for a fixed function f, we know that $\mathbf{Pr}\Big[\mathcal{E}_{T \text{ good est. of } f}\Big] \geq 1/2$, if m is sufficiently large.

Consider the event $\mathcal{E}_2 = \bigcup_{f \in \mathcal{F}} \mathcal{E}_{2,f}$. Intuitively, we expect that $\mathcal{E}_{2.f} \approx \mathcal{E}_{1,f}$ and thus $\mathcal{E}_2 \approx \mathcal{E}_1$, and as such, we can bound $\mathbf{Pr}[\mathcal{E}_2]$ instead of bounding $\mathbf{Pr}[\mathcal{E}_1]$. Clearly, $\mathcal{E}_2 \subseteq \mathcal{E}_1$, and

$$\mathbf{Pr}\Big[\mathcal{E}_2 \;\Big|\; \mathcal{E}_1\Big] = \frac{\mathbf{Pr}[\mathcal{E}_2 \cap \mathcal{E}_1]}{\mathbf{Pr}[\mathcal{E}_1]} = \frac{\mathbf{Pr}[\mathcal{E}_2]}{\mathbf{Pr}[\mathcal{E}_1]}.$$

However, if $\mathcal{E}_1$ happens (i.e., $S \in \mathcal{E}_1$), then there is at least one function $g^S \in \mathcal{F}$ that S fails for, and fix this function (conceptually, think about fixing both S and g^S). Now, S and T are independent, and the event $\mathcal{E}_{1,f}$ depends only on S and the event $\mathcal{E}_{T \text{ good est. of } f}$ depends only on T; namely, the events $S \in \mathcal{E}_{1,f}$ and $T \in \mathcal{E}_{T \text{ good est. of } f}$ are independent, for any f. Now, by Lemma 6.3.4, we have that

$$\begin{aligned}
\mathbf{Pr}\Big[\mathcal{E}_2 \;\Big|\; \mathcal{E}_1\Big] &= \mathbf{Pr}\Big[(S,T) \in \mathcal{E}_2 \;\Big|\; S \in \mathcal{E}_1\Big] \geq \mathbf{Pr}\Big[(S,T) \in \mathcal{E}_{2,g^S} \;\Big|\; S \in \mathcal{E}_1\Big] \\
&= \mathbf{Pr}\Big[\Big(S \in \mathcal{E}_{1,g^S}\Big) \cap \Big(T \in \mathcal{E}_{T \text{ good est. of } g^S}\Big) \;\Big|\; S \in \mathcal{E}_1\Big] = \mathbf{Pr}\Big[T \in \mathcal{E}_{T \text{ good est. of } g^S} \;\Big|\; S \in \mathcal{E}_1\Big] \\
&= \mathbf{Pr}\Big[\mathcal{E}_{T \text{ good est. of } g^S}\Big] \geq \frac{3}{4} \geq \frac{1}{2}.
\end{aligned}$$

This implies that $\mathbf{Pr}[\mathcal{E}_1] \leq 2\,\mathbf{Pr}[\mathcal{E}_2]$. Now, by the triangle inequality applied to $d_v(\cdot,\cdot)$ (see Lemma 6.1.8), we have that if $\mathcal{E}_{2,f}$ happens, then

$$\frac{\alpha}{2} + d_v\Big(\overline{s}_T(f), \overline{s}_S(f)\Big) > d_v\Big(\overline{m}, \overline{s}_T(f)\Big) + d_v\Big(\overline{s}_T(f), \overline{s}_S(f)\Big) \geq d_v\Big(\overline{m}, \overline{s}_S(f)\Big) > \alpha.$$

This implies that the event

$$\mathcal{E}'_{2,f} \quad \equiv \quad d_v\Big(\overline{s}_T(f), \overline{s}_S(f)\Big) > \frac{\alpha}{2}$$

happens, and let $\mathcal{E}'_2 = \bigcup_f \mathcal{E}'_{2,f}$. We have that if $\mathcal{E}_2$ happens then $\mathcal{E}'_2$ happens; that is $\mathcal{E}_2 \subseteq \mathcal{E}'_2$. We conclude that

$$\mathbf{Pr}[\mathcal{E}_1] \leq 2\,\mathbf{Pr}[\mathcal{E}_2] \leq 2\,\mathbf{Pr}[\mathcal{E}'_2].$$

Namely, we reduced the task of bounding the probability that there exists a bad function for the given sample, into a new event. The new event, is to decide if there is a function such that the two estimates for the two samples disagree considerably.

As such, from this point on, we are interested in bounding the probability of the event $\mathcal{E}'_2$. Again, giving in to our formal urge, the event $\mathcal{E}'_2$ can be stated as

$$\mathcal{E}'_2 = \left\{(S,T) \;\middle|\; \exists f \in \mathcal{F} \text{ such that } d_v\Big(\overline{s}_T(f), \overline{s}_S(f)\Big) > \frac{\alpha}{2}\right\}. \tag{6.4}$$

This does not look like a finite event and it is not. But we can bound the probability to be in this set by bounding a finite event. The reason being is that we can fix the content of $S \cup T$ and limit the probability argument into whats falls into S and what falls into T.

6.3.3 Double sampling

As in the proof of the ε-approximation theorem, we will bound the probability of failure of the sample, by using a double sampling argument. So, consider a sample $N = \langle \mathsf{p}_1, \ldots, \mathsf{p}_{2m} \rangle$ of size $2m$. Let Γ_{2m} be the set of permutations over $\{1, \ldots, 2m\}$, such that for $\sigma \in \Gamma_{2m}$, we have for $1 \leq i \leq m$ that either $\sigma(i) = i$ and $\sigma(m+i) = m+i$ or $\sigma(i) = m+i$ and $\sigma(m+i) = i$. Namely, a permutation $\sigma \in \Gamma_{2m}$ either flips the pair $(i, m+i)$, or preserve it, for every i.

Now, if we take any random sequence of independent samples, and remix it in any way, without looking on the values of the elements of the sequence, what we get is a random sequence.[②] In fact, this random sequence as the same distribution as the original random sequence.

As such, take a random permutation $\sigma \in \Gamma_{2m}$ and mix the random sample N to get a new sequence

$$N_\sigma = \langle \mathsf{p}_{\sigma(1)}, \mathsf{p}_{\sigma(2)}, \ldots, \mathsf{p}_{\sigma(2m)} \rangle .$$

Clearly, N and N_σ have *exactly* the same distribution. As such, bounding the probability for high disagreement between the first half of the sample of N and second half, can be done on N_σ. In particular, for $f \in \mathcal{F}$, let

$$\mu_1(f) = \frac{1}{m} \sum_{i=1}^{m} f(\mathsf{p}_{\sigma(i)}) \quad \text{and} \quad \mu_2(f) = \frac{1}{m} \sum_{i=m+1}^{2m} f(\mathsf{p}_{\sigma(i)}) . \tag{6.5}$$

We will be interested in bounding the quantity $d_\nu(\mu_1(f), \mu_2(f))$.

6.3.4 Some more definitions

Naturally, we can think about$(f(\mathsf{p}_1), \ldots, f(\mathsf{p}_{2m}))$ as a point in $[0,1]^{2m}$. As such, the ordered sample $N = \langle \mathsf{p}_1, \ldots, \mathsf{p}_{2m} \rangle$ induces a manifold in $\mathbb{R}^{2m}$, by considering the value of any function of $\mathcal{F}$ on the points of N. Formally, consider the set

$$\mathcal{M} = \left\{ f(N) = (f(\mathsf{p}_1), f(\mathsf{p}_2), \ldots, f(\mathsf{p}_{2m})) \;\middle|\; f \in \mathcal{F} \right\} .$$

The set $\mathcal{M}$ has pseudo-dimension δ. The reader can think of $\mathcal{M}$ as being a low dimensional manifold. From this point on, we will use the notation $f = f(N)$ and $f_i = f(\mathsf{p}_i)$, for $i = 1, \ldots, 2m$.

We remind the reader that the ℓ_1-norm of a point $p \in \mathbb{R}^{2m}$ is $\|p\|_1 = \sum_{i=1}^{2m} |p_i|$. The ***measure*** of $f \in \mathbb{R}^{2m}$ is the quantity

$$\overline{m}(f) = \frac{1}{2m} \sum_{i=1}^{2m} f_i .$$

The ***measure distance*** between $f, g \in \mathcal{M}$ is

$$d_{\overline{m}}(f,g) = \overline{m}(|f-g|) = \frac{1}{2m} \|f-g\|_1 . \tag{6.6}$$

Note, that since the triangle inequality holds for the $\|\cdot\|_1$, it also holds for $d_{\overline{m}}(\cdot)$ (which is thus a metric).

[②] Naturally, we can "lose" randomness by looking on the values. For example, think about sorting the sequence.

In the following discussion, fix $\alpha > 0$, and for any $f \in \mathcal{M}$, we would like to prove that $d_\nu(\mu_1(f), \mu_2(f)) \leq \alpha$. In particular, let $\mathcal{E}(f)$ denote the maximum value of ν, such that this inequality holds. That is, we have

$$\begin{aligned} d_{\mathcal{E}(f)}\Big(\mu_1(f), \mu_2(f)\Big) &= \frac{|\mu_1(f) - \mu_2(f)|}{\mu_1(f) + \mu_2(f) + \mathcal{E}(f)} = \alpha \\ &\iff \Big|\mu_1(f) - \mu_2(f)\Big| = \alpha\left(\mu_1(f) + \mu_2(f) + \mathcal{E}(f)\right) \\ &\iff \Big|\mu_1(f) - \mu_2(f)\Big| = \alpha\left(\frac{f_\Sigma}{m} + \mathcal{E}(f)\right), \end{aligned}$$

where $f_\Sigma = \sum_{i=1}^{2m} f_i$. As such, we define the ***error*** of f to be

$$\mathcal{E}(f) = \frac{\Big|\mu_1(f) - \mu_2(f)\Big|}{\alpha} - \frac{f_\Sigma}{m}. \tag{6.7}$$

This quantity is monotonically increasing as $\Big|\mu_1(f) - \mu_2(f)\Big|$ increasing, and as such it measures very roughly how similar are the two estimates of f to each other.[3] Namely, the usage of the term error to describe it, is used only in the strict

We would like to argue that, with good probability, this quantity is small for all $f \in \mathcal{M}$.

Observation 6.3.5 *For $g \in \mathcal{M}$, we have that $d_\nu\Big(\mu_1(g), \mu_2(g)\Big) \leq \alpha$ if and only if $\mathcal{E}(g) \leq \nu$.*

6.3.5 Bounding the error

The following lemma bounds the probability that a single function g in our family (now, a function is just a vector with $2m$ coordinates, since we care about its values only a specific $2m$ points) is being incorrectly estimated. The estimation here is a bit weird – we are randomly scrambling the coordinates g, see Eq. (6.5), and then compute how much the two parts of g diverge, see Eq. (6.7).

Lemma 6.3.6 *Let $g \in [0,1]^{2m}$, and let $\sigma \in \Gamma_{2m}$ be a random permutation (chosen uniformly). Then, for any $\nu > 0$, it holds that $\mathbf{Pr}[\mathcal{E}(g) \geq \nu] \leq 2\exp\Big(-\alpha^2 \nu m\Big)$.*

Proof: By Observation 6.3.5, the bad event (i.e., $\mathcal{E}(g) \geq \nu$) is equivalent to $d_\nu\Big(\mu_1(g), \mu_2(g)\Big) > \alpha$, that is

$$d_\nu\Big(\mu_1(g), \mu_2(g)\Big) = \frac{|\mu_1(g) - \mu_2(g)|}{\mu_1(g) + \mu_2(g) + \nu} > \alpha. \quad \Rightarrow \quad |\mu_1(g) - \mu_2(g)| > \alpha(\mu_1(g) + \mu_2(g) + \nu)\,.$$

Since $\mu_1(g) = (1/m)\sum_{i=1}^{m} g_{\sigma(i)}$ and $\mu_2(g) = (1/m)\sum_{i=m+1}^{2m} g_{\sigma(i)}$, by multiplying both sides by m, we get

$$\left|\sum_{i=1}^{m}\Big(g_{\sigma(i)} - g_{\sigma(m+i)}\Big)\right| > \alpha\left(\sum_{i=1}^{2m} g_i + \nu m\right) = \alpha(g_\Sigma + \nu m)\,.$$

[3] Ultimately, the form of $\mathcal{E}(f)$ rises from its usage in the following proofs. The reader might consider calling $\mathcal{E}(f)$ — the error of f — a mistake. Calling it the distortion of f, or the perversion of f might have been better. Maybe so, but words are just labels, weapons of fortune in conveying ideas. The reader is hopefully willing to suffer the arrows of mislabeling to get the underlying ideas. It is the nature of writing that authors rarely discuss such decisions, mainly because they are ad hoc, and lead to a pointless tangential discussions, as demonstrated by this footnote.

To bound the probability of this bad event, let us set $X_i = g_{\sigma(i)} - g_{\sigma(m+i)}$. Clearly, since $\sigma \in \Gamma_{2m}$ either keeps this pair or flips it (with equal probability), it follows that X_i is equal to either $g_i - g_{m+i} \in [-1, 1]$ or to $-(g_i - g_{m+i})$. In particular, $\mathbf{E}[X_i] = 0$, and $\mathbf{E}[\sum_i X_i] = 0$. By Hoeffding's inequality (Theorem 6.6.1), we have that the probability of the bad event is bounded by

$$\tau = \mathbf{Pr}\left[\left|\sum_{i=1}^{m} X_i\right| > \alpha\,(g_\Sigma + m\,\nu)\right] \leq 2\exp\left(-\frac{2\Big(\alpha(g_\Sigma + m\,\nu)\Big)^2}{\sum_{i=1}^{m} 4(g_i - g_{m+i})^2}\right).$$

Observe that $g_i - g_{m+i} \in [-1, 1]$, and as such $(g_i - g_{m+i})^2 \leq |g_i - g_{m+i}| \leq g_i + g_{m+i}$. In particular, $\sum_{i=1}^{m}(g_i - g_{m+i})^2 \leq \sum_{i=1}^{2m} g_i = g_\Sigma$. As such,

$$\begin{aligned}\tau &\leq 2\exp\left(-\frac{(\alpha(g_\Sigma + \nu m))^2}{2g_\Sigma}\right) \leq 2\exp\left(-\alpha^2\,\frac{(g_\Sigma)^2 + 2(g_\Sigma)\,\nu m + (\nu m)^2}{2g_\Sigma}\right)\\ &\leq 2\exp\Big(-\alpha^2\,\nu m\Big),\end{aligned}$$

as claimed. ∎

The following lemma testifies that a triangle type inequality holds for the error.

Lemma 6.3.7 *Consider $f, g \in [-1, 1]^{2m}$, then we have $\mathcal{E}(f + g) \leq \mathcal{E}(f) + \mathcal{E}(g)$.*

Proof: Since $\mu_1(\cdot)$ and $\mu_2(\cdot)$ are linear functions, we have

$$\begin{aligned}\mathcal{E}(f + g) &= \frac{|\mu_1(f + g) - \mu_2(f + g)|}{\alpha} - \frac{(f + g)_\Sigma}{m}\\ &= \frac{|\mu_1(f) - \mu_2(f) \;+\; \mu_1(g) - \mu_2(g)|}{\alpha} - \frac{f_\Sigma}{m} - \frac{g_\Sigma}{m}\\ &\leq \frac{|\mu_1(f) - \mu_2(f)|}{\alpha} - \frac{f_\Sigma}{m} \;+\; \frac{|\mu_1(g) - \mu_2(g)|}{\alpha} - \frac{g_\Sigma}{m} = \mathcal{E}(f) + \mathcal{E}(g).\end{aligned}$$

∎

One of the key new ingredients in the proof of Theorem 6.1.7, is the idea that if two functions $f, h \in [0, 1]^{2m}$ are close to each other (in the ℓ_1-norm) then they have similar error. To this end, consider the function $g = f - h$, and observe that it must have a small ℓ_1-norm. As such, we can strengthen Lemma 6.3.6 for g, if the vector g is short (i.e., its ℓ_1-norm is small). Intuitively, what we are proving is that if h has a small error, and $f - h$ is "short", then, with high probability, $f = h + g = h + (f - h)$ would also have small error, since the error complies with the triangle inequality.

Lemma 6.3.8 *Let $\alpha, \nu, c > 0$ be some constants, where $c \leq 2/3$. Also, let $g \in [-1, 1]^{2m}$, such that $\|g\|_1 \leq cvm$, where $\|g\|_1 = \sum_{i=1}^{2m} |g_i|$ is the ℓ_1-norm of g. Then, for a random permutation $\sigma \in \Gamma_{2m}$ (chosen uniformly), we have that $\mathbf{Pr}[\mathcal{E}(g) > \nu] \leq 2\exp\left(-\frac{\alpha^2 \nu m}{36c}\right)$.*

Proof: Arguing as in Lemma 6.3.6, the desired probability τ is bounded by

$$\tau \leq 2\exp\left(-\frac{2\Big(\alpha(g_\Sigma + \nu m)\Big)^2}{4\sum_{i=1}^{m}(g_i - g_{m+i})^2}\right).$$

Now, observe that

$$\sum_{i=1}^{m}(g_i - g_{m+i})^2 \leq 2\sum_{i=1}^{m}|g_i - g_{m+i}| \leq 2\sum_{i=1}^{2m}|g_i| \leq 2c\nu m,$$

since $g_i \in [-1, 1]$. Furthermore, the quantity $(g_\Sigma + \nu m)^2$ is minimized, under the condition $\sum_{i=1}^{2m}|g_i| \leq c\nu m$, when $g_\Sigma = -c\nu m$, which implies that

$$\Big(g_\Sigma + \nu m\Big)^2 \geq ((1-c)\nu m)^2 \geq \frac{\nu^2 m^2}{9},$$

since $c \leq 2/3$. This implies that

$$\tau \leq 2\exp\left(-2\alpha^2\frac{(g_\Sigma + \nu m)^2}{4\sum_{i=1}^{m}(g_i - g_{m+i})^2}\right) =\leq 2\exp\left(-\alpha^2\frac{\nu^2 m^2/9}{4c\nu m}\right) \leq 2\exp\left(-\frac{\alpha^2\nu m}{36c}\right).$$

■

6.3.6 On the number of distinct functions

Let $\mathsf{S} = (\mathsf{X}, \mathcal{F})$ be a range space with ***pseudo-dimension*** δ. We are interested in how many "truly" distinct concepts S really has.

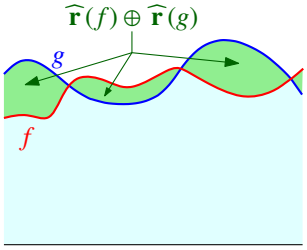
Definition 6.3.9 A subset $P \subseteq \mathcal{F}$ is an ***ε-packing*** of $\mathcal{F}$ if for any pair of functions $f, g \in P$, we have that

$$d_{\overline{m}}(f, g) = \overline{m}(|f - g|) \geq \varepsilon.$$

Furthermore, for any function $h \in \mathcal{F}$, there exists a function $h' \in P$, such that $d_{\overline{m}}(h, h') \leq \varepsilon$.

In the following we bound the size of such an ε-packing P.

For two sets X and Y, we denote by $X \oplus Y = (X \cup Y) \setminus (X \cap Y)$. For two functions f and g, the symmetric difference between $\widehat{\mathbf{r}}(f) \oplus \widehat{\mathbf{r}}(g)$ is the region lying between the two functions, see figure on the right.



Lemma 6.3.10 *Let $\mathsf{S} = (\mathsf{X}, \mathcal{F})$ be a range space with* **pseudo-dimension** *δ. Let $\varepsilon > 0$ be a parameter, and $P \subseteq \mathcal{F}$ be an ε-packing of S. Then $|P| = 1/\varepsilon^{O(\delta)}$.*

Proof: As suggested in Remark 6.1.6, it would be easier to work in the induced range space

$$\mathsf{T} = (\mathsf{X}', \mathcal{G}),$$

where $\mathcal{G} = \left\{\widehat{\mathbf{r}}(f) \;\middle|\; f \in \mathcal{F}\right\}$, see Eq. (6.1). The key observation here is that the VC dimension of T is δ.

Now, for $f \in \mathcal{F}$, we have that $\overline{m}(f) = \overline{m}(\widehat{\mathbf{r}}(f))$, where here $\overline{m}(\widehat{\mathbf{r}}(f))$ is just the measure of the set $\widehat{\mathbf{r}}(f)$. For any $f, g \in P$, we have

$$\varepsilon \leq \overline{m}(|f - g|) = \overline{m}\Big(\widehat{\mathbf{r}}(f) \oplus \widehat{\mathbf{r}}(g)\Big).$$

So, let us consider the range set

$$\mathcal{H} = \left\{ \mathbf{r} \oplus \mathbf{r}' \;\middle|\; \mathbf{r}, \mathbf{r}' \in \mathcal{G} \right\},$$

and the resulting range space $\mathsf{U} = (\mathsf{X}', \mathcal{H})$. Each range in it is formed by combining two ranges of T, and by Theorem 6.6.2, the VC dimension of U is $O(\delta)$.

By Theorem 6.6.3 (which we proved independently of the discussion in this chapter), the range space U has an ε-net $N \subseteq \mathsf{X}'$ of size $k = O((\delta/\varepsilon)\log(1/\varepsilon))$; specifically, a random sample of this size is the desired net with constant probability, and as such a net exists.

Now, consider two distinct functions $f, g \in P$. We have that $\overline{m}(|f - g|) \geq \varepsilon$, and the above discussion implies that $\overline{m}(\widehat{\mathbf{r}}(f) \oplus \widehat{\mathbf{r}}(g)) \geq \varepsilon$. As such, since $\mathbf{r} = \widehat{\mathbf{r}}(f) \oplus \widehat{\mathbf{r}}(g)$ is a range of U, and N is an ε-net of U, it follows that $\mathbf{r} \cap N \neq \emptyset$. Namely, $\widehat{\mathbf{r}}(f) \cap N \neq \widehat{\mathbf{r}}(g) \cap N$. Namely, $\widehat{\mathbf{r}}(f)$ and $\widehat{\mathbf{r}}(g)$ are distinct ranges, when we project U to N. We conclude that for

$$F' = P_{|N} = \left\{ \widehat{\mathbf{r}}(f) \cap N \;\middle|\; f \in P \right\},$$

it holds

$$|P| = |F'| \leq \mathcal{G}_{O(\delta)}(|N|) = |N|^{O(\delta)} = k^{O(\delta)} = 1/\varepsilon^{O(\delta)},$$

by Lemma 6.6.4. ∎

6.3.7 Chaining

We need to show that the error $\mathcal{E}(f) \leq \nu$ for all $f \in \mathcal{M}$. The problem is that $\mathcal{M}$ is potentially an infinite set, and our tools can handle only a finite set of points (i.e., Lemma 6.3.6). To overcome this problem, we define a canonical set of packings of $\mathcal{M}$. So, let

$$\varepsilon_j = \frac{\nu}{2^{2j+4}},$$

for $j \geq 1$. Let $P_{-1} = \{h\}$, for any arbitrary point on $h \in \mathcal{M}$.

Next, for $j \geq 0$, let P_j be a ε_j-packing of $\mathcal{M}$. Here we are using the measure distance between points, which is $d_{\overline{m}}(f, g) = (1/2m)\sum_{i=1}^{2m} |f_i - g_i|$, see Eq. (6.6). As such, for any $f, g \in P_j$, it holds that $d_{\overline{m}}(f, g) \geq \varepsilon_j$.

One can build the packing P_j, for $j \geq 1$, by starting from the set P_{j-1}. If there is point $f \in \mathcal{M}$, such that

$$\operatorname{dist}(f, P_j) = \min_{g \in P_j} \overline{m}(f - g) \geq \varepsilon_j,$$

then we add f to P_j. We continue in this fashion till no such point can be found. Note, that the ordering in which we consider the points of $\mathcal{M}$ does not matter. Clearly, the resulting set is a ε_j-packing. More importantly, we have the property that

$$P_1 \subseteq P_2 \subseteq \cdots \subseteq P_j \subseteq \cdots$$

In particular, let $\mathcal{M}^* = \cup_{j=0}^{\infty} P_j$. The set $\mathcal{M}^*$ is dense in $\mathcal{M}$, and as such it is sufficient to prove our claim for points on $\mathcal{M}^*$ (because the error function $\mathcal{E}(\cdot)$ is continuous). To see the density claim, observe that for any $\mathsf{p} \in \mathcal{M}$, there exists a sequence $\mathsf{p}_j \in P_j$, such that $d_{\overline{m}}(\mathsf{p})\,\mathsf{p}_j \leq \varepsilon_j$. As such, $\lim_{j\to\infty} \mathsf{p}_j = \mathsf{p}$.

For a point $f \in \mathcal{M}^*$, let $\widehat{f_j}$ be its nearest neighbor in P_j. Formally,

$$\widehat{f_j} = \arg\min_{g\in P_j} \overline{m}(|f - g|)\,.$$

Since P_j is an ε_j-packing, we have that for any $f \in \mathcal{M}^*$, it holds that $d_{\overline{m}}\Big(f, \widehat{f_j}\Big) \leq \varepsilon_j$. As such, observe that

$$d_{\overline{m}}\Big(\widehat{f_j}, \widehat{f_{j+1}}\Big) \leq d_{\overline{m}}\Big(\widehat{f_j}, f\Big) + d_{\overline{m}}\Big(f, \widehat{f_{j+1}}\Big) \leq \varepsilon_j + \varepsilon_{j+1} \leq 2\varepsilon_j. \tag{6.8}$$

Also $f = \widehat{f_1} + \sum_{j=1}^{\infty}\Big(\widehat{f_{j+1}} - \widehat{f_j}\Big)$. As such, since the triangle inequality holds for the error function (see Lemma 6.3.7), we have

$$\mathcal{E}(f) \leq \mathcal{E}\Big(\widehat{f_1}\Big) + \sum_{j=1}^{\infty} \mathcal{E}\Big(\widehat{f_{j+1}} - \widehat{f_j}\Big)\,.$$

For $j \geq 1$, let

$$\nu_j = \nu\frac{\sqrt{j+1}}{3\cdot 2^j}. \tag{6.9}$$

And assume that, for all f and j, it holds that

$$\mathcal{E}\Big(\widehat{f_1}\Big) \leq \nu_1 \text{ and } \mathcal{E}\Big(\widehat{f_{j+1}} - \widehat{f_j}\Big) \leq \nu_j. \tag{6.10}$$

Then, we have that

$$\mathcal{E}(f) \leq \nu_1 + \sum_{j=1}^{\infty} \nu_j = \nu\left(\frac{\sqrt{2}}{3\cdot 2} + \sum_{j=1}^{\infty} \frac{\sqrt{j+1}}{3\cdot 2^j}\right) \leq \nu,$$

as can be easily verified.

Now, if the above holds for all $f \in \mathcal{M}^*$ then this implies the desired result (that is Theorem 6.1.7). Indeed, Observation 6.3.5 implies that this would imply that for all $f \in \mathcal{M}$, we have that $d_\nu\Big(\mu_1(g)\,, \mu_2(g)\Big) \leq \alpha/2$ (for the sake of simplicity of exposition, we are fidgeting with the constants a bit here). This imply that $\mathcal{E}'_2$ does not happen (see Eq. (6.4)), which implies that sample estimates correctly all functions, which is what we needed prove.

Thus, to complete the proof, we need to bound the probabilities that the assumptions of Eq. (6.10) do not hold.

Lemma 6.3.10 implies the following result.

Corollary 6.3.11 *For $j \geq 1$, we have $\left|P_j\right| \leq 1/\varepsilon_j^{c\delta}$, where c is some constant.*

Lemma 6.3.12 *If m is of size as specified by Theorem 6.1.7, then it holds that* $\mathbf{Pr}\Big[\exists f \in P_1 \;\; \mathcal{E}(f) \geq \nu_1\Big] \leq \varphi/2$.

Proof: By Corollary 6.3.11, the size of P_1 is bounded by $1/\varepsilon_1^{c\delta} = \left(2^6/\nu\right)^{c\delta} = (64/\nu)^{c\delta}$. We remind the reader that $\nu_1 = \nu\frac{\sqrt{2}}{6}$ (see Eq. (6.9)), and $m = O\left(\frac{1}{\alpha^2\nu}\left(\delta \log\frac{1}{\nu} + \log\frac{1}{\varphi}\right)\right)$, see Theorem 6.1.7. As such, by Lemma 6.3.6, for some constant c_1, we have

$$\begin{aligned}\mathbf{Pr}\left[\exists f \in P_1 \;\; \mathcal{E}(f) \geq \nu_1\right] &\leq 2\left|P_1\right| \exp\left(-\alpha^2\nu_1 m\right) \leq 2\left(\frac{64}{\nu}\right)^{c\delta} \exp\left(-c_1\alpha^2\nu\frac{1}{\alpha^2\nu}\left(\delta\log\frac{1}{\nu} + \log\frac{1}{\varphi}\right)\right) \\ &\leq 2\left(\frac{64}{\nu}\right)^{c\delta} \exp\left(-c_1\delta\log\frac{1}{\nu} - c_1\log\frac{1}{\varphi}\right) = 2\left(\frac{64}{\nu}\right)^{c\delta} \nu^{c_1\delta}\varphi^{c_1} \leq \frac{\varphi}{2},\end{aligned}$$

by making m (and thus c_1) sufficiently large. ∎

Lemma 6.3.13 *If m is of size as specified by Theorem 6.1.7, then for any $j \geq 1$ and $f \in \mathcal{M}$, it holds that*

$$\mathbf{Pr}\left[\exists f \in \mathcal{M} \;\; \mathcal{E}\left(\widehat{f_{j+1}} - \widehat{f_j}\right) \geq \nu_j\right] \leq \frac{\varphi}{2^{j+1}}.$$

Proof: Fix j, and consider the set

$$X = \left\{\widehat{f_{j+1}} - \widehat{f_j} \;\middle|\; f \in \mathcal{M}\right\}.$$

For any $x \in X$, we have that $\overline{m}(|x|) \leq 2\varepsilon_j$, by Eq. (6.8). As such, by definition (see Eq. (6.6)), we have

$$\forall x \in X \quad \|x\|_1 \leq 4\varepsilon_j m = \frac{4}{2^{2j+4}}\nu m,$$

which implies that we can apply Lemma 6.3.8 to its vectors, with $c = c_j = 2^{-2j-3}/4c$‘.

Furthermore, an element of X is formed by the difference of a point of P_j and P_{j+1}, and as such

$$|X| \leq \left|P_j\right| \cdot \left|P_{j+1}\right| \leq \frac{1}{\varepsilon_j^{c\delta} \cdot \varepsilon_{j+1}^{c\delta}} \leq \varepsilon_{j+1}^{-2c\delta} \leq \left(\frac{2^{2j+4}}{\nu}\right)^{2c\delta},$$

by Corollary 6.3.11. We remind the reader that $\nu_j = \nu\frac{\sqrt{j+1}}{3\cdot 2^j}$ (see Eq. (6.9)), and

$$m = O\left(\frac{1}{\alpha^2\nu}\left(\delta\log\frac{1}{\nu} + \log\frac{1}{\varphi}\right)\right),$$

see Theorem 6.1.7. As such, by Lemma 6.3.8, we have

$$\begin{aligned}\mathbf{Pr}\left[\exists x \in X \;\; \mathcal{E}(x) \geq \nu_j\right] &\leq 2\left|X\right| \exp\left(-\frac{\alpha^2\nu_j m}{9c_j}\right) \leq 2\left(\frac{2^{2j+4}}{\nu}\right)^{2c\delta} \exp\left(-\frac{4\alpha^2\left(\nu\frac{\sqrt{j+1}}{3\cdot 2^j}\right)m}{9\cdot 2^{-2j-3}}\right) \\ &= 2\left(\frac{2^{2j+4}}{\nu}\right)^{2c\delta} \exp\left(-\frac{32\cdot 2^j\sqrt{j+1}}{27}\alpha^2\nu m\right).\end{aligned}$$

Now, $m = O\left(\frac{1}{\alpha^2\nu}\left(\delta \log \frac{1}{\nu} + \log \frac{1}{\varphi}\right)\right)$ and there exists a constant c_2, such that

$$\begin{aligned}
\mathbf{Pr}\left[\exists x \in X \;\; \mathcal{E}(x) \geq \nu_j\right] &\leq 2\left(\frac{2^{2j+4}}{\nu}\right)^{2c\delta} \exp\left(-2^j\sqrt{j+1}\alpha^2\nu\left(\frac{c_2}{\alpha^2\nu}\left(\delta \log \frac{1}{\nu} + \log \frac{1}{\varphi}\right)\right)\right) \\
&\leq \exp\left(1 + 2c\delta(2j+4) + 2c\delta \ln \frac{1}{\nu} - 2^j\sqrt{j+1}\left(c_2\left(\delta \log \frac{1}{\nu} + \log \frac{1}{\varphi}\right)\right)\right) \\
&\leq \exp(-(j+1) - \ln \varphi) \leq \frac{\varphi}{2^{j+1}},
\end{aligned}$$

by making m (and thus c_2) sufficiently large. ∎

Lemma 6.3.12 and Lemma 6.3.13 imply that the probability that our assumptions of Eq. (6.10) fail is at most

$$\frac{\varphi}{2} + \sum_{j=1}^{\infty} \frac{\varphi}{2^{j+1}} \leq \varphi,$$

implying that Theorem 6.1.7 holds. ■

6.4 Exercises

Exercise 6.4.1 (Playing with $d_\nu(\cdot,\cdot)$) [2 Points]
Prove the following properties of $d_\nu(\cdot,\cdot)$:

(i) For all $r, s \geq 0$, we have $0 \leq d_{r,s}(<) 1$.

(ii) For all non-negative $r \leq s \leq t$, we have $d_\nu(r,s) \leq d_\nu(r,t)$ and $d_\nu(s,t) \leq d_\nu(r,t)$.

(iii) For $0 \leq r, s \leq M$, we have $\frac{|r-s|}{2M+\nu} \leq d_\nu(r,s) \leq d_\nu(r,t) \leq \frac{|r-s|}{\nu}$.

Exercise 6.4.2 (Distance between numbers and the triangle inequality.) [10 Points]
Prove Lemma 6.1.8. Namely, for any $r, s \geq 0$ and $\nu > 0$, consider the function

$$d_\nu(r,s) = \frac{|r-s|}{r+s+\nu}.$$

Prove that the triangle inequality holds for $d_\nu(\cdot,\cdot)$. Namely, we have $d_\nu(x,y) + d_\nu(y,z) > d_\nu(x,z)$.
(Warning: This exercise is tedious.)

6.5 Bibliographical notes

Theorem 6.1.7 is from Lin *et al.* [LLS01], and we *very* loosely followed their presentation. Their work is the pinnacle of a long sequence of papers by Pollard [Pol86] and Haussler [Hau92].

The author came up with the proof of Lemma 6.3.10, although better bounds are known, see Haussler [Hau95].

Another take on the proof of Theorem 6.1.7. The key step in the proof Theorem 6.1.7, was showing that the manifold $\mathcal{M}$ embeds into a short interval, when projected into a specifically chosen random vector. Lemma 6.3.10 tells us that the manifold $\mathcal{M}$ is low dimensional. From this point on, the proof now uses the chaining technique due to Kolmogorov. A somewhat similar argument was used recently to show that manifolds with low doubling dimension embed with low distortion [AHY07] (see also [IN07]). This interpretation of spaces of low VC dimension (so sorry, pseudo-dimension) as being low dimensional in the sense of low dimension manifold is quite interesting.

6.6 From other lectures

Theorem 6.6.1 (Hoeffding's inequality.) *Let $X_1, \ldots, X_n$ be independent random variables, where $X_i \in \left[a_i, b_i\right]$, for $i = 1, \ldots, n$. Then, for the random variable $S = X_1 + \cdots + X_n$, and any $\eta > 0$, we have*

$$\mathbf{Pr}\left[\left|S - \mathbf{E}[S]\right| \geq \eta\right] \leq 2\exp\left(-\frac{2\eta^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Theorem 6.6.2 *Let $\mathsf{S}_1 = \left(X, \mathcal{R}^1\right), \ldots, \mathsf{S}_k = \left(X, \mathcal{R}^k\right)$ be range spaces with VC dimension $\delta_1, \ldots, \delta_k$, respectively. Next, let $f(\mathbf{r}_1, \ldots, \mathbf{r}_k)$ be a function that maps any k tuple of sets $\mathbf{r}_1 \in \mathcal{R}^1, \ldots, \mathbf{r}_k \in \mathcal{R}^k$ into a subset of X. Consider the range set*

$$\mathcal{R}' = \left\{ f(\mathbf{r}_1, \ldots, \mathbf{r}_k) \;\middle|\; \mathbf{r}_1 \in \mathcal{R}_1, \ldots, \mathbf{r}_k \in \mathcal{R}_k \right\}$$

and the associated range space $\mathsf{T} = (X, \mathcal{R}')$. Then, the VC dimension of T is bounded by $O(k\delta \lg k)$, where $\delta = \max_i \delta_i$.

Theorem 6.6.3 (ε-net theorem, [HW87]) *Let $(X, \mathcal{R})$ be a range space of VC-dimension δ, x be a finite subset of X and suppose that $0 < \varepsilon \leq 1$ and $\varphi < 1$. Let N be a set obtained by m random independent draws from x, where*

$$m \geq \max\left(\frac{4}{\varepsilon} \lg \frac{4}{\varphi}, \frac{8\delta}{\varepsilon} \lg \frac{16}{\varepsilon} \right). \tag{6.11}$$

Then N is an ε-net for x with probability at least $1 - \varphi$.

Lemma 6.6.4 (Sauer's Lemma) *If $(X, \mathcal{R})$ is a range space of VC-dimension δ with $|X| = n$ then $|\mathcal{R}| \leq \mathcal{G}_\delta(n)$.*

Theorem 6.6.5 (Chebychev inequality) *Let X be a random variable with $\mu_x = \mathbf{E}[X]$ and σ_x be the standard deviation of X. That is $\sigma_X^2 = \mathbf{E}\left[(X - \mu_x)^2\right]$. Then, $\mathbf{Pr}[|X - \mu_X| \geq t\sigma_X] \leq \dfrac{1}{t^2}$.*

Bibliography

[AHY07] P. Agarwal, S. Har-Peled, and H. Yu. Embeddings of surfaces, curves, and moving points in Euclidean space. In *Proc. 23rd Annu. ACM Sympos. Comput. Geom.*, pages 381–389, 2007.

[BCM99] H. Brönnimann, B. Chazelle, and J. Matoušek. Product range spaces, sensitive sampling, and derandomization. *SIAM J. Comput.*, 28:1552–1575, 1999.

[Brö95] H. Brönnimann. *Derandomization of Geometric Algorithms*. Ph.D. thesis, Dept. Comput. Sci., Princeton University, Princeton, NJ, May 1995.

[CKMS06] E. Cohen, H. Kaplan, Y. Mansour, and M. Sharir. Approximations with relative errors in range spaces of finite VC dimension. manuscript, 2006.

[Hau92] D. Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992.

[Hau95] D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *J. Comb. Theory Ser. A*, 69(2):217–232, 1995.

[HW87] D. Haussler and E. Welzl. ε-nets and simplex range queries. *Discrete Comput. Geom.*, 2:127–151, 1987.

[IN07] P. Indyk and A. Naor. Nearest neighbor preserving embeddings. *ACM Trans. Algo.*, 2007. To appear.

[KPW92] J. Komlós, J. Pach, and G. Woeginger. Almost tight bounds for ε-nets. *Discrete Comput. Geom.*, 7:163–173, 1992.

[LLS01] Y. Li, P. M. Long, and A. Srinivasan. Improved bounds on the sample complexity of learning. *J. Comput. Syst. Sci.*, 62(3):516–527, 2001.

[Pol86] D. Pollard. Rates of uniform almost-sure convergence for empirical processes indexed by unbounded classes of functions. Manuscript, 1986.

[VC71] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971.