Some variants on a theorem of Vapnik

by Arne Hole

Department of Mathematics, University of Oslo
Box 1053, Blindern, N-0316 Oslo, Norway
e-mail: arneh@math.uio.no

Abstract

We prove some reformulated and slightly sharpened versions of a theorem of Vapnik on uniform convergence of estimated probabilities. This theorem has applications, amongst other places, in the theory of generalization ability of neural nets.

1. Introduction

Vapnik's theorems on uniform convergence of estimated probabilities [7] form a corner stone for many developments in learning theory (see eg. [1] [2] [3] [5]). We will deal here with the theorem concerning *relative* deviations of estimated probabilities, or more specifically, with the version appearing in [1]. The main difference in form between that result and theorems 1 and 2 of this paper, is that the theorems given here are expressed in terms of an arbitrary map ϕ instead of a supremum operation or existence quantifier. Moreover, via the parameter α, theorems 1 and 2 incorporate a certain type of possible a priori knowledge about the map ϕ in their bounds. Interpreting in terms of learning (cf. section 4 below), the parameter α can be viewed as measuring the amount a priori knowledge about the generalizing ability of ϕ.

Taking $\alpha = 1$ in theorem 1 and translating into the usual form, we obtain (corollary 1) an improvement on the bound given in [1] by a factor of two, and on the bound given in [7] by a factor of four. Under a certain additional assumption, corollary 3 further improves the bound by a factor of two.

The proofs we give follow the path of the original one in [7] closely. Concerning lemma 2, it may be remarked that [1] contains a much shorter and more elegant argument for bounding the quantity Γ_z^r. However, the bound we obtain below (by sharpening the original argument) turns out to be a factor of two better than the one in [1].

2. Notation

The letter $\wp$ denotes power set, and card denotes cardinality. The composition of a mapping ϕ with a mapping ψ is denoted $\psi \circ \phi$, ie. $\psi \circ \phi(\xi) = \psi(\phi(\xi))$. The conditional probability of the event A given the event B is written $P(A \mid B)$, where

P is the probability measure. Define the step function $\chi : \mathbf{R} \to \mathbf{R}$ by $\chi(t) = 1$ for $t > 0$ and $\chi(t) = 0$ for $t \leq 0$.

3. Basic results

From now on, let X and S be fixed sets, where X is equipped with a probability measure P. The product measure on X^m defined by P will be denoted P^m for each integer $m \geq 1$. Let ν be a fixed probability measure on $X^m \times S$ such that the marginal of ν on X^m is P^m. The elements $x = (x_1, \ldots, x_m) \in X^m$ will be referred to as *samples* of length m. For each sample $x \in X^m$ and each $r \subseteq X$, let

$$x \cap r = \{i \mid 1 \leq i \leq m \text{ and } x_i \in r\},$$

and put

$$\hat{P}_x(r) = \frac{1}{m} \cdot \text{card}(x \cap r).$$

If $x \in X^m$, $H \subseteq \wp(X)$ and $\beta \in [0, 1]$, we then define

$$(x \cap H)^{[\beta]} = \{x \cap r \mid r \in H \text{ and } \hat{P}_x(r) \geq \beta\}.$$

Put $\Delta_H^{[\beta]}(x) = \text{card}(x \cap H)^{[\beta]}$. Finally, define

$$\Delta_H^{[\beta]}(m) = \max_{x \in X^m} \Delta_H^{[\beta]}(x)$$

We write $\Delta_H^{[\beta]}(x)$ and $\Delta_H^{[\beta]}(m)$ for $\beta \leq 0$ simply as $\Delta_H(x)$ and $\Delta_H(m)$, respectively.

If $z \in X^{2m}$ is a sample of length $2m$, we use the notation $z = (z_a, z_b)$, where $z_a, z_b \in X^m$. For each $r \subseteq X$ and $z \in X^{2m}$, let

$$K_r(z) = \frac{\hat{P}_{z_a}(r) - \hat{P}_{z_b}(r)}{\sqrt{1 - \hat{P}_z(r) + 2/m}}.$$

A class $R \subseteq \wp(X)$ of Borel sets is called *well behaved* if the two quantities

$$P^m \left\{ \sup_{r \in R} \frac{\hat{P}_x(r) - P(r)}{\sqrt{1 - P(r)}} > \delta \right\}$$

$$P^{2m} \left\{ \sup_{r \in R} K_r(z) > \delta \right\}$$

are both well defined for each $\delta > 0$. A map $\phi : X^m \times S \to \wp(X)$ is called *good* if its image $\phi(X^m \times S)$ is well behaved, and the two quantities

$$A(\phi) = \nu \left\{ \frac{\hat{P}_x(\phi(x, \sigma)) - P(\phi(x, \sigma))}{\sqrt{1 - P(\phi(x, \sigma))}} > \delta \right\}$$

$$B(\phi) = P^{2m} \left\{ \sup_{\sigma \in S} K_{\phi^+(z, \sigma)}(z) > \delta \right\}$$

are both well defined for all $\delta > 0$, where $\phi^+ : X^{2m} \times S \to \wp(X)$ is the map given by $\phi^+(z, \sigma) = \phi(z_a, \sigma)$. Let S_{2m} denote the group of permutations on $\{1, \ldots, 2m\}$. If $z \in X^{2m}$, we define $\pi(z) = (z_{\pi(1)}, \ldots, z_{\pi(2m)})$. If $z, z' \in X^{2m}$ are such that there is a $\pi \in S_{2m}$ with $\pi(z) = z'$, then z' is called a *permutation* of z. Write $S_{2n} = \{\pi_i\}_{i=1}^{(2m)!}$. For given $r \subseteq X$, $z \in X^{2m}$ and $\delta > 0$, let

$$\Gamma_z^r = \frac{1}{(2m)!} \sum_{i=1}^{(2m)!} \chi[K_r(\pi_i(z)) - \delta]$$

Lemma 1. *For all $\delta > 0$, $m \geq 4/\delta^2$ and good ϕ, we have $A(\phi) \leq 4B(\phi)$.*

Lemma 2. *For all $z \in X^{2m}$, $r \subseteq X$, $\delta > 0$ and $m \geq 4/\delta^2$, we have*

$$\Gamma_z^r < \frac{1}{2} e^{-m\delta^2/4}$$

These lemmas are proved in appendices 1 and 2, respectively. From them, the basic version of Vapnik's theorem follows quite easily.

Theorem 1. *Let $\alpha, \beta \in [0, 1]$, $\delta \in (0, 1)$, and let $m \geq 4/\delta^2$ be an integer. Assume that $\phi : X^m \times S \to \wp(X)$ is good, and that $P(\phi(x, \sigma)) \geq 1 - \alpha$ for all (x, σ). Then*

$$A(\phi) < 2\Delta e^{-m\delta^2/4}$$

where

$$\Delta = \tau \Delta_{\phi(X^m \times S)}(2m) + (1 - \tau) \Delta_{\phi(X^m \times S)}^{[1-\alpha-\beta]}(2m)$$

and $\tau < \min\{1, 7\Delta_{\phi(X^m \times S)}(4m) e^{-m\beta^2/2}\}$.

Proof. Observe that $\phi^+(X^{2m} \times S) = \phi(X^m \times S)$. We have

$$B(\phi) \leq \int_{X^{2m}} \sup_{r \in \phi(X^m \times S)} \chi\left[K_r(z) - \delta\right] \, dP^{2m}(z).$$

Call the expression on the right hand side of the equation above $C(\phi)$. Let $\pi \in S_{2m}$ be a permutation. Since the distribution P^{2m} on X^{2m} is invariant under permutations, we get

$$C(\phi) = \int_{X^{2m}} \frac{1}{(2m)!} \sum_{i=1}^{(2m)!} \sup_{r \in \phi(X^m \times S)} \chi\left[K_r(\pi_i(z)) - \delta\right] \, dP^{2m}(z).$$

Now let $z \in X^{2m}$ be fixed. We split up $\phi(X^m \times S)$ into equivalence classes defined by the relation

$$r_1 \sim r_2 \iff z \cap r_1 = z \cap r_2.$$

It is clear that the map $r \mapsto K_r(\pi_i(z))$ will be constant on each equivalence class in $\phi(X^m \times S)$, for all permutations $\pi_i \in S_{2m}$. The number of equivalence classes is $\Delta_{\phi(X^m \times S)}(z)$. Thus we may form a finite set $repr(z)$ consisting of one element from each equivalence class. Then

$$\sup_{r \in \phi(X^m \times S)} K_r(\pi_i(z)) = \sup_{r \in repr(z)} K_r(\pi_i(z)).$$

So

$$C(\phi) = \int_{X^{2m}} \frac{1}{(2m)!} \sum_{i=1}^{(2m)!} \sup_{r \in repr(z)} \chi\left[K_r(\pi_i(z)) - \delta\right] dP^{2m}(z),$$

and thus (repacing the sup operation with a sum)

$$C(\phi) \leq \int_{X^{2m}} \sum_{r \in repr(z)} \left\{ \frac{1}{(2m)!} \sum_{i=1}^{(2m)!} \chi[K_r(\pi_i(z)) - \delta] \right\} dP^{2m}(z). \tag{1}$$

Let $U_{bad} \subseteq X^{2m}$ be the set of all $z \in X^{2m}$ such that there is an $r \in \phi(X^m \times S)$ with $P(r) \geq 1 - \alpha$ and $\hat{P}_z(r) < 1 - \alpha - \beta$. Let U_{good} be the complement of U_{bad} in X^{2m}. Using the standard version of Vapnik's theorem (theorem A.2, ch. 6 in [7]), remembering that $\phi(X^m \times S)$ is assumed to be well behaved, we get

$$P^{2m}(U_{bad}) < 7\Delta_{\phi(X^m \times S)}(4m)e^{-m\beta^2/2}$$

Assume $z \in U_{good}$. Then for all $r \in repr(z)$ we have $\hat{P}_r(z) \geq 1 - \alpha - \beta$, so

$$\text{card}(repr(z)) \leq \Delta^{[1-\alpha-\beta]}_{\phi(X^m \times S)}(2m)$$

for all $z \in U_{good}$. Using lemma 2 and substituting this in (1), we obtain

$$\begin{aligned} C(\phi) &< \frac{1}{2} \int_{U_{bad}} \Delta_{\phi(X^m \times S)}(2m)e^{-m\delta^2/4} \; dP^{2m}(z) \\ &+ \frac{1}{2} \int_{U_{good}} \Delta^{[1-\alpha-\beta]}_{\phi(X^m \times S)}(2m)e^{-m\delta^2/4} \; dP^{2m}(z) \\ &< \frac{1}{2}\tau\Delta_{\phi(X^m \times S)}(2m)e^{-m\delta^2/4} + \frac{1}{2}(1-\tau)\Delta^{[1-\alpha-\beta]}_{\phi(X^m \times S)}(2m)e^{-m\delta^2/4} \end{aligned}$$

Combining this with lemma 1 and the definition of $C(\phi)$, we get the theorem. ■

Remark. Instead of using the original theorem A.2 of [7], we could have used the sharpened version given in [5]. However, the improvement this would lead to is not essential in the present context.

Corollary 1 (Standard version). *Let $\delta > 0$, and assume that $R \subseteq \wp(X)$ is well behaved. Then for all integers $m \geq 4/\delta^2$ we have*

$$P^m \left\{ \sup_{r \in R} \frac{\hat{P}_x(r) - P(r)}{\sqrt{1 - P(r)}} > \delta \right\} < 2\Delta_R(2m) e^{-m\delta^2/4}$$

Proof. Let $S = \{0\}$, and let ν be the unique measure on $X^m \times S$ such that the marginal of ν on X^m is P^m. Let the map $\phi : X^m \times S \to R$ be such that the following condition is satisfied: Given $x \in X^m$, if there is an $r \in R$ such that

$$\frac{\hat{P}_x(r) - P(r)}{\sqrt{1 - P(r)}} > \delta,$$

then

$$\frac{\hat{P}_x(\phi(x,0)) - P(\phi(x,0))}{\sqrt{1 - P(\phi(x,0))}} > \delta$$

Since R is well behaved, ϕ is good. Clearly

$$A(\phi) = P^m \left\{ \sup_{r \in R} \frac{\hat{P}_x(r) - P(r)}{\sqrt{1 - P(r)}} > \delta \right\}$$

Now apply theorem 1, taking $\alpha = 1$ and noting that $\phi(X^m \times S) \subseteq R$. ■

4. Applications to learning

In this section, I will indicate how the preceeding formalism fits in with a special type of learning situation. First I will make some additional definitions and prove the relevant corollary of theorem 1. Then I will comment briefly on interpretation. A more detailed discussion is given in [4].

Let F be some function class, and let $\theta : F \to \wp(X)$ and $\lambda : X^m \times S \to F$ be maps. We write $\lambda(x, \sigma)$ as λ_x^s. For all $f \in F$, integers $m \geq 1$, $x \in X^m$, and $t \in [0,1]$, let

$$E(f, \theta, x) = (1/m)\text{card}\{i \mid 1 \leq i \leq m \text{ and } x_i \notin \theta(f)\}$$
$$E(f, \theta) = 1 - P(\theta(f))$$
$$\Omega(t, \theta) = \nu\{E(\lambda_x^\sigma, \theta, x) \leq t\}$$

Corollary 2. *Assume that $\theta \circ \lambda$ is good. Let $\alpha, \beta \in [0,1]$, $\gamma \in [0,1)$, $\epsilon \in (0,1)$, and let $m \geq 4/(\epsilon(1-\gamma)^2)$ be an integer. If $E(\lambda_x^\sigma, \theta) \leq \alpha$ for all $(x, \sigma) \in X^m \times S$, then*

$$\nu\Big\{ E(\lambda_x^\sigma, \theta) > \epsilon \;\Big|\; E(\lambda_x^\sigma, \theta, x) \leq \gamma\epsilon \Big\} < \frac{2\Delta}{\Omega_\Lambda(\gamma\epsilon, \theta) e^{\epsilon m(1-\gamma)^2/4}}$$

where Δ is as in theorem 1, with $\theta(F)$ substituted for $\phi(X^m \times S)$.

Proof. Let Q_1 and Q_2 be the events (relative to ν) given by

$$Q_1 = \{E(\lambda_x^\sigma, \theta) > \epsilon\} \qquad Q_2 = \{E(\lambda_x^\sigma, \theta, x) \leq \gamma\epsilon\}$$

Assume that $(x, \sigma) \in X^m \times S$ is such that Q_1 and Q_2 are both true. Let $r = \theta(\lambda_x^\sigma)$. Then $\hat{P}_x(r) \geq 1 - \gamma\epsilon$ and $1 - P(r) > \epsilon$, so

$$\hat{P}_x(r) \geq (1-\gamma)(1-P(r)) + P(r)$$

In other words,

$$\begin{aligned} \frac{\hat{P}_x(r) - P(r)}{\sqrt{1-P(r)}} &\geq (1-\gamma)\sqrt{1-P(r)} \\ &> (1-\gamma)\sqrt{\epsilon} \end{aligned}$$

Applying theorem 1 with $\phi = \theta \circ \lambda$ and $\delta = (1-\gamma)\sqrt{\epsilon}$, it follows that

$$\nu(Q_1 \cap Q_2) \leq 2\Delta e^{-\epsilon m(1-\gamma)^2/4}$$

where Δ is as stated in the corollary. But then by the law of conditional probability,

$$\nu(Q_1 \mid Q_2) = \frac{\nu(Q_1 \cap Q_2)}{\nu(Q_2)} = \frac{\nu(Q_1 \cap Q_2)}{\Omega(\gamma\epsilon, \theta)}$$

which gives the result. ■

Now some words on interpretation. For convenience, I discuss this in terms of learning with neural networks. Of course, the formalism can be interpreted in terms of many other types of learning models as well.

There are two different classes of situations I want to distinguish between, namely learning with and without noise.

Situation 1: Noiseless learning. In this case, we may call the set X the *input space*. We assume that in addition we are given another set Y, called the *output space*. The class F is the function class defined by the neural network architecture we work with, and F consists of functions $f : X \to Y$. We imagine that we are given a "target" $f_0 : X \to Y$ which we want the network to learn. The map λ is called the *learning algorithm*, and the map θ may be called a *learning criterion*. The idea is that the set $\theta(f) \subseteq X$ should be interpreted as the region of the input space where f behaves "acceptably" relative to the target f_0. Then $E(f, \theta, x)$ represents the error of f on x, and $E(f)$ is the global error of f. Each $x \in X^m$ can be viewed as a "training sequence" of length m, and the set S represents possible stochastic elements in the training process. The quantity $\Omega(t, \theta)$ is the probability that the "learned function" λ_x^σ has error $\leq t$ on the training sequence x. Corollary 2 says

something about the probability that the learned function has small global error given that is has small error on the training sequence.

Situation 2: Learning with noise. Here we take $X = Y \times Z$, where Y and Z represents the input and output space, respectively. Now F consists of functions $f : Y \to Z$, and the probability measure P itself represents target. The set $\theta(f)$ can be viewed as the region of the input-output space which is acceptable relative to f. Otherwise, everything works roughly as in the noiseless situation.

5. Another variant

If one makes the additional assumption that ϵm is an integer, the bounds given in corollaries 1 and 2 can be improved by a factor of two. The proofs parallel the ones given above, and I will only briefly indicate the differences.

If $r \subseteq X$ is measurable, let $P^\sharp(r)$ be the smallest real number t such that $t \geq P(r)$ and tm is an integer. We call a class $R \subseteq \wp(X)$ of Borel sets *well $\sharp$-behaved* if the two quantities

$$P^m \left\{ \sup_{r \in R} \frac{\hat{P}_x(r) - P^\sharp(r)}{\sqrt{1 - P^\sharp(r)}} > \delta \right\}$$

$$P^{2m} \left\{ \sup_{r \in R} K_r(z) > \delta \right\}$$

are both well defined for all fixed $\delta > 0$. Also, a map $\phi : X^m \times S \to \wp(X)$ is called *$\sharp$-good* if its image $\phi(X^m \times S)$ is well $\sharp$-behaved, and the two quantities

$$A^\sharp(\phi) = \nu \left\{ \frac{\hat{P}_x(\phi(x,\sigma)) - P^\sharp(\phi(x,\sigma))}{\sqrt{1 - P^\sharp(\phi(x,\sigma))}} > \delta \right\}$$

$$B(\phi) = P^{2m} \left\{ \sup_{\sigma \in S} K_{\phi^+(z,\sigma)}(z) > \delta \right\}$$

are both well defined for all $\delta > 0$, where $\phi^+ : X^{2m} \times S \to \wp(X)$ is the associated map given by $\phi^+(z,\sigma) = \phi(z_a, \sigma)$. Then we have the following result.

Theorem 2. *Let $\alpha, \beta \in [0,1]$, $\delta \in (0,1)$, and let $m \geq 4/\delta^2$ be an integer. Assume that $\phi : X^m \times S \to \wp(X)$ is $\sharp$-good, and that $P(\phi(x,\sigma)) \geq 1 - \alpha$ for all (x,σ). Then*

$$A^\sharp(\phi) < \Delta e^{-m\delta^2/4},$$

where Δ is as in theorem 1.

Proof. I claim that in this case, for all $\delta > 0$, $m \geq 4/\delta^2$ and $\sharp$-good ϕ, we have $A(\phi) \leq 2B(\phi)$. The proof of this is analogous to the proof of lemma 1, except for the following: The probability of the event $\hat{P}_{z_b}(r) \leq P^\sharp(r)$ is equal to $Pr(W \leq \lceil mq \rceil)$,

where W is binomially distributed (m, q) with $q = P(r)$, and $\leq \lceil mq \rceil$ means the smallest integer $\geq mq$. It follows [6] that this probability is $\geq 1/2$, which gives the claim. The rest of the proof is identical to the corresponding part of the proof of theorem 1. ■

Corollary 3. *Let $\delta > 0$, and assume that $R \subseteq \wp(X)$ is well $\sharp$-behaved. Then for all integers $m \geq 4/\delta^2$ we have*

$$P^m \left\{ \sup_{r \in R} \frac{\hat{P}_x(r) - P^\sharp(r)}{\sqrt{1 - P^\sharp(r)}} > \delta \right\} < \Delta_R(2m) e^{-m\delta^2/4}$$

Proof. Analogous to the proof of corollary 1, using theorem 2 instead of theorem 1. ■

Corollary 4. *Assume that $\theta \circ \lambda$ is $\sharp$-good. Let $\alpha, \beta \in [0,1]$, $\gamma \in [0,1)$, $\epsilon \in (0,1)$, and let $m \in \{1,2,3,\ldots\}$ be such that $m \geq 4/(\epsilon(1-\gamma)^2)$ and ϵm is an integer. If $E(\lambda_x^\sigma, \theta) \leq \alpha$ for all $(x, \sigma) \in X^m \times S$, then*

$$\nu\Big\{ E(\lambda_x^\sigma, \theta) > \epsilon \;\Big|\; E(\lambda_x^\sigma, \theta, x) \leq \gamma\epsilon \Big\} < \frac{\Delta}{\Omega_\Lambda(\gamma\epsilon, \theta) e^{\epsilon m (1-\gamma)^2/4}}$$

where Δ is as in theorem 1, with $\theta(F)$ substituted for $\phi(X^m \times S)$.

Proof. As before, let Q_1 and Q_2 be the events given by $Q_1 = \{E(\lambda_x^\sigma, \theta) > \epsilon\}$ and $Q_2 = \{E(\lambda_x^\sigma, \theta, x) \leq \gamma\epsilon\}$. Assume that $(x, \sigma) \in X^m \times S$ is such that Q_1 and Q_2 are both true. Then $\hat{P}_x(\theta(\lambda_x^\sigma)) \geq 1 - \gamma\epsilon$ and $1 - P(\theta(\lambda_x^\sigma)) > \epsilon$. But since ϵm is an integer, it follows that $1 - P^\sharp(\theta(\lambda_x^\sigma)) > \epsilon$. The rest is analogous to the proof of corollary 2. ■

Appendix 1: Proof of lemma 1

Since the marginal of ν on X^m is P^m, clearly

$$A(\phi) \leq P^{2m} \left\{ \sup_{\sigma \in S} \frac{\hat{P}_{z_a}(\phi^+(z,\sigma)) - P(\phi^+(z,\sigma))}{\sqrt{1 - P(\phi^+(z,\sigma))}} > \delta \right\}$$

Assume that $z \in X^{2m}$ is such that there is an $\sigma \in S$ with $r := \phi^+(z, \sigma)$ satisfying

$$\frac{\hat{P}_{z_a}(r) - P(r)}{\sqrt{1 - P(r)}} > \delta,$$

that is,

$$\hat{P}_{z_a}(r) > P(r) + \delta \cdot \sqrt{1 - P(r)}. \tag{2}$$

Since $\hat{P}_{z_a}(r) \leq 1$, this implies

$$1 - P(r) > \delta^2. \tag{3}$$

Assume that we also have

$$\hat{P}_{z_b}(r) \leq P(r). \tag{4}$$

Proceeding as in [7], it can be shown by straightforward algebra that under the conditions (2), (3) and (4), we have $K_r(z) > \delta$.

Now let us examine the condition (4) more closely. When ϕ^+ is given, the event (4) is independent of the event (2) under the probability measure P^{2m} on X^{2m}. The reason is that (2) depends only on the first m coordinates of the sample z, while (4) depends on the last m only. The probability of (4) in X^{2m} is equal to $Pr(W \geq \lceil mq \rceil)$, where W is binomially distributed (m, q) with $q = 1 - P(r)$, and $\lceil mq \rceil$ means the smallest integer $\geq mq$. Since $m \geq 4/\delta^2$ and $q = 1 - P(r) > \delta^2$ by (3), we have $m > 4/q$. It follows from a result in [6] that the probability of (4) is $\geq 1/4$, and the lemma follows. ■

Appendix 2: Proof of lemma 2

We have

$$\Gamma_z^r = \sum_k \frac{\binom{n}{k}\binom{2m-n}{m-k}}{\binom{2m}{m}},$$

where the sum runs through all k such that $\max(0, n-m) \leq k \leq \min(n, m)$ and

$$\frac{k/m - (n-k)/m}{\sqrt{1 - n/2m + 2/m}} > \delta. \tag{5}$$

The condition (5) is equivalent to

$$k > \frac{n}{2} + \frac{\zeta m}{2}, \qquad \text{where} \qquad \zeta = \delta\sqrt{\frac{2m-n+4}{2m}}.$$

Put $s = \min(n, m)$ and $T = \max(0, n-m)$. Following [7], for $T \leq k \leq s$ we let

$$p(k) = \frac{\binom{n}{k}\binom{2m-n}{m-k}}{\binom{2m}{m}}$$

$$q(k) = \frac{p(k+1)}{p(k)} = \frac{(n-k)(m-k)}{(k+1)(m+k+1-n)}$$

$$d(k) = \sum_{i=k}^{s} p(i)$$

Using that $q(t)$ is monotone decreasing, we then get [7]

$$d(k+1) \leq q(k) \sum_{i=k}^{s-1} p(i) < q(k) \sum_{i=k}^{s} p(i) = q(k)d(k).$$

Now let

$$j = \begin{cases} n/2+1 & \text{if } n \text{ is even} \\ n/2+1/2 & \text{if } n \text{ is odd.} \end{cases}$$

Since $m > 2/\delta^2$, we have $\delta > \sqrt{2/m}$. Also $\zeta \geq \delta\sqrt{2/m}$, so $\zeta > 2/m$. Thus $\zeta m/2 > 1$, so $k > n/2+1$ in the case we are considering. Thus we always have $k > j$. By repetition of the argument above

$$d(k) < d(j) \prod_{i=j}^{k-1} q(i)$$

for $k > j$. Since the expression for $p(k)$ is symmetric around $n/2$, it follows that $d(j) \leq 1/2$. So

$$d(k) < \frac{1}{2} \prod_{i=j}^{k-1} q(i)$$

for $k > j$. Let $t = k - (n-1)/2$. Then

$$q(k) = \frac{(n+1)/2 - t}{(n+1)/2 + t} \cdot \frac{(m-(n-1)/2) - t}{(m-(n-1)/2) + t}.$$

By taking logarithms on both sides and using linear approximation at the origin, we obtain [7] the bound

$$\ln q(k) \leq -2\left[\frac{2}{n+1} + \frac{2}{2m-n+1}\right](k - \frac{n-1}{2}) = -K(k - \frac{n-1}{2}),$$

where $K = 8(m+1)/[(n+1)(2m-n+1)]$. This estimate holds for $(n-1)/2 \leq k \leq s$, ie. it is sufficient for all the values of k we are interested in. Assume first that n is even, so that $j = n/2+1$. Then

$$\begin{aligned} \ln(2d(k)) < \ln(\prod_{i=j}^{k-1} q(i)) &= \sum_{i=j}^{k-1} \ln q(i) \leq -K \cdot \sum_{i=j}^{k-1} \left(i - \frac{n-1}{2}\right) \\ &= -K \cdot \sum_{i=j}^{k-1} \left(i - \frac{n}{2} + \frac{1}{2}\right) = -K \cdot \sum_{i=1}^{k-n/2-1} \left(i + \frac{1}{2}\right) \\ &= -\frac{K}{2}\left[(k - \frac{n}{2})^2 - 1\right] \end{aligned}$$

Then assume n is odd, so that $j = n/2+1/2$. Then

$$\begin{aligned} \ln(2d(k)) < \ln(\prod_{i=j}^{k-1} q(i)) &= \sum_{i=j}^{k-1} \ln q(i) \leq -K \cdot \sum_{i=j}^{k-1} \left(i - \frac{n-1}{2}\right) \\ &= -K \cdot \sum_{i=1}^{k-(n-1)/2-1} i = -\frac{K}{2}\left[(k - \frac{n}{2})^2 - \frac{1}{4}\right] \end{aligned}$$

So $\ln(2d(k)) < -(K/2)[(k-n/2)^2 - 1]$ for all k such that $j < k \leq s$. Let k be the smallest integer such that $k - n/2 > \zeta m/2$. Then $\Gamma = d(k)$, and therefore

$$\ln(2\Gamma) < -\frac{K}{2}\left[\left(\frac{\zeta m}{2}\right)^2 - 1\right] = -\frac{K}{8}(\zeta^2 m^2 - 4).$$

Substituting for K and ζ, we now get

$$\begin{aligned}\ln(2\Gamma) &< -\frac{m+1}{(n+1)(2m-n+1)} \cdot [(2m-n+4)m\delta^2/2 - 4] \\ &= -\frac{m+1}{(n+1)(2m-n+1)} \cdot [(2m-n+1)m\delta^2/2 + 3m\delta^2/2 - 4]\end{aligned}$$

Since $m > 4/\delta^2$, we have $3m\delta^2/2 > 6 > 4$. So

$$\ln(2\Gamma) < -\frac{m+1}{(n+1)(2m-n+1)} \cdot [(2m-n+1)m\delta^2/2] = -\frac{(m+1)m\delta^2}{2(n+1)}.$$

The expression on the right hand side reaches its maximum for $n = 2m$, so

$$\Gamma_r^z < \frac{1}{2}\exp\left(-\frac{(m+1)m\delta^2}{2(2m+1)}\right) < \frac{1}{2}e^{-m\delta^2/4}. \quad \blacksquare$$

REFERENCES

[1] Anthony, M., and Shawe-Taylor, J. 1993. A result of Vapnik with applications. *Discrete Appl. Math.* **47**, 207-217.

[2] Baum, E., and Haussler, D. 1989. What Size Net Gives Valid Generalization? *Neural Comp.* **1**, 151-160.

[3] Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. 1989. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.* **36**, 929-965.

[4] Hole, A. 1995. Vapnik-Chervonenkis generalization bounds for real valued neural networks. *Preprint series, Institute of mathematics, University of Oslo.*

[5] Parrando, J., and Van den Broeck, C. 1993. Vapnik-Chervonenkis bounds for generalization *J. Phys. A.* **26**, 2211-2223.

[6] Slud, E. 1977. Distribution inequalities for the binomial law. *Annals of Prob.* **5**(3), 404-412.

[7] Vapnik, V. 1982. *Estimation of Dependencies Based on Empirical Data.* Springer-Verlag, New York.