8 Supplementary Material

8.1 Proof of Theorem 1

Our proof of Theorem 1 uses techniques from Gurvits and Koiran (1997).

Theorem. *Let $\mathcal{F}$ be a class of real-valued, continuous functions over a set $\mathcal{X}$, with a finite co-VC-dimension D. Let $g(\mathbf{x})$ be a function in the convex hull of $\mathcal{F}$: $g(\mathbf{x}) = \sum_{i=1}^{N} w_i f_i(\mathbf{x})$, with $\sum_{i=1}^{N} w_i = 1$ and $f_i \in \mathcal{F}$. Assume that functions $f_i(\mathbf{x})$ are upper-bounded by M and that the quantity $\int f_i(\mathbf{x})\,d\mathbf{x}$ is lower-bounded by B for all f_i. Let P be the probability measure over functions $\{f_1, \ldots, f_N\}$ such that $P(f_i) = w_i$. A sampling operation is taken to draw K functions $\{h_1, \ldots, h_K\}$ independently from P. Then, for any $\mathbf{x} \in \mathcal{X}$,*

$$P\left\{\frac{1}{K}\sum_{i=1}^{K} h_i(\mathbf{x}) \notin [(1-\zeta)g(\mathbf{x}), (1+\zeta)g(\mathbf{x})]\right\} < 8(2K)^D \exp\left(-\frac{\zeta^2}{4}\frac{B}{M}K\right) \tag{8}$$

Proof. Given a function $f \in \mathcal{F}$, denote the expected value of f over $\mathcal{X}$ as $P(f) = \int f(\mathbf{x})d\mathbf{x}$. As the distribution of $\mathbf{x}$ is usually unknown, in practice, an i.i.d. sample of K inputs, $\mathbf{x}_i \in \mathcal{X}$, $1 \le i \le K$, is usually used to approximate $P(f)$. Denote the approximation as $v(f) = \frac{1}{K}\sum_{i=1}^{K} f(\mathbf{x}_i)$. In Section 8.4, we provide an inequality that bounds the relative difference between the two quantities $P(f)$ and $v(f)$, using the pseudo-dimension of the function class $\mathcal{F}$:

$$\begin{aligned}&P\{v(f) \notin [(1-\zeta)P(f), (1+\zeta)P(f)]\}\\ &= P\left\{\frac{1}{K}\sum_{i=1}^{K} f(\mathbf{x}_i) \notin \left[(1-\zeta)P(f), (1+\zeta)P(f)\right]\right\}\\ &< 8\,\Pi(2K)\exp\left(-\frac{\zeta^2}{4}\frac{B}{M}K\right)\end{aligned} \tag{9}$$

where B is a lower bound of $P(f)$ and M is an upper bound of $f(\mathbf{x})$; $\Pi(2K)$ is a quantity called the "growth function" that satisfies $\Pi(2K) \le (2K)^H$ where H is the pseudo-dimension of $\mathcal{F}$.

As we aim to bound $\sum_{i=1}^{K} h_i(\mathbf{x})$ instead of $\sum_{i=1}^{K} h(\mathbf{x}_i)$, we make use of co-VC-dimension in the dual, instead of pseudo-dimension. By the sampling operation in our assumption, we have that for every $\mathbf{x} \in \mathcal{X}$ and each h_i, $E[h_i(\mathbf{x})] = g(\mathbf{x})$. Following techniques in Gurvits and Koiran (1997), we make the substitutions: $f(\mathbf{x}_i) \leftarrow h_i(\mathbf{x})$, $P(f) \leftarrow g(\mathbf{x})$, and $\Pi(2K) \le (2K)^D$ where D is the co-VC-dimension, into the inequality (9). Then, for any $\mathbf{x} \in \mathcal{X}$, we have

$$P\left\{\frac{1}{K}\sum_{i=1}^{K} h_i(\mathbf{x}) \notin [(1-\zeta)g(\mathbf{x}), (1+\zeta)g(\mathbf{x})]\right\} < 8(2K)^D \exp\left(-\frac{\zeta^2}{4}\frac{B}{M}K\right)$$

□

8.2 Max-Norm Reweighting Scheme

While multiplying two weighted sums of Gaussians, we make use of the max-norm reweighting scheme in Wrigley et al. (2017) to make multiplications more effective. Specifically, the term M/B in the convergence rate expression (6) suggests that its minimization will lead to a faster convergence. For a weighted sum of functions $\phi(\mathbf{x}) = \sum_{i=1}^{K} w_i\psi_i(\mathbf{x})$ and a reweighted representation $\phi'(\mathbf{x}) = \sum_{i=1}^{K} w_i'\psi_i'(\mathbf{x})$, the max-norm scheme to minimize the M/B ratio is to set $w_i' \propto w_i \max_{\mathbf{x}} \psi_i(\mathbf{x})$. For a weighted sum of Gaussians, $\phi(\mathbf{x}) = \sum_{i=1}^{K} w_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, to minimize the ratio, we set

$$\begin{aligned} w_i' &= w_i \max_{\mathbf{x}} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{w_i}{(2\pi)^{d/2}\sqrt{\det \boldsymbol{\Sigma}}}\\ \psi_i'(\mathbf{x}) &= \frac{w_i}{w_i'}\psi_i(\mathbf{x}) = \exp\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\end{aligned} \tag{10}$$

where we note that the maximum of a Gaussian is achieved at $\mathbf{x} = \boldsymbol{\mu}$ and equal to $\left((2\pi)^{d/2}\sqrt{\det \boldsymbol{\Sigma}}\right)^{-1}$. The resulting sum of functions is in effect a weighted sum of Gaussian exponential components. Multiplying two Gaussian exponentials yields another exponential, with a constant factor s different from c in (7).

$$s = \exp\left(-\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right) \tag{11}$$

8.3 Closed-Form Transition Update

In this section, we derive the closed-form transition formula previously presented in Section 4.2. To facilitate integration over product of Gaussians, we re-express Gaussians in a different form. We use results stated in Koller and Friedman (2009).

Definition 7 (Canonical Form)**.** *A canonical form $\mathcal{C}(\mathbf{x}; \mathbf{K}, \mathbf{h}, g)$ where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{K} \in \mathbb{R}^{d\times d}$, $\mathbf{h} \in \mathbb{R}^d$ and g is a scalar, is defined as*

$$\mathcal{C}(\mathbf{x}; \mathbf{K}, \mathbf{h}, g) = \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{K}\mathbf{x} + \mathbf{h}^T\mathbf{x} + g\right) \tag{12}$$

A Gaussian function, $\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ over $\mathbf{x}\in\mathbb{R}^d$, can be equivalently expressed in canonical form $\mathcal{C}(\mathbf{x};\mathbf{K},\mathbf{h},g)$ with $\mathbf{K}=\boldsymbol{\Sigma}^{-1}$, $\mathbf{h}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, and

$$g=-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\frac{d}{2}\log 2\pi-\frac{1}{2}\log\det\boldsymbol{\Sigma}$$

The product of two canonical forms over the same $\mathbf{x}$:

$$\begin{aligned}&\mathcal{C}(\mathbf{x};\mathbf{K}_1,\mathbf{h}_1,g_1)\cdot\mathcal{C}(\mathbf{x};\mathbf{K}_2,\mathbf{h}_2,g_2)\\ =&\,\mathcal{C}(\mathbf{x};\mathbf{K}_1+\mathbf{K}_2,\mathbf{h}_1+\mathbf{h}_2,g_1+g_2)\end{aligned}$$

When we have two canonical forms over different scopes $\mathbf{x}$ and $\mathbf{y}$, we extend the scopes of both to make them match and then perform the above multiplication. The extension of scope is by adding zero entries to both the $\mathbf{K}$ matrices and the $\mathbf{h}$ vectors.

Next, consider the marginalization operation. Let $\mathcal{C}(\mathbf{x},\mathbf{y};\mathbf{K},\mathbf{h},g)$ be a canonical form over $\{\mathbf{x},\mathbf{y}\}$ where

$$\mathbf{K}=\begin{pmatrix}\mathbf{K}^{xx} & \mathbf{K}^{xy}\\ \mathbf{K}^{yx} & \mathbf{K}^{yy}\end{pmatrix},\quad \mathbf{h}=\begin{pmatrix}\mathbf{h}^{x}\\ \mathbf{h}^{y}\end{pmatrix}$$

The marginalization of this canonical form onto the variables $\mathbf{x}$ is the integral over the variables $\mathbf{y}\in\mathbb{R}^d$, $\int\mathcal{C}(\mathbf{x},\mathbf{y};\mathbf{K},\mathbf{h},g)\,d\mathbf{y}$. The result of the integration is a canonical form $\mathcal{C}(\mathbf{x};\mathbf{K}',\mathbf{h}',g')$ given by:

$$\begin{aligned}\mathbf{K}'&=\mathbf{K}^{xx}-\mathbf{K}^{xy}(\mathbf{K}^{yy})^{-1}\mathbf{K}^{yx}\\ \mathbf{h}'&=\mathbf{h}^{x}-\mathbf{K}^{xy}(\mathbf{K}^{yy})^{-1}\mathbf{h}^{y}\\ g'&=g+\frac{1}{2}\left(d\log 2\pi-\log\det\mathbf{K}^{yy}+(\mathbf{h}^y)^T\mathbf{K}^{yy}\mathbf{h}^y\right)\end{aligned}$$

Moreover, according to Petersen et al., the inverse of a matrix in block representation can be expressed as,

$$\begin{aligned}&\begin{pmatrix}\mathbf{K}^{xx} & \mathbf{K}^{xx'}\\ \mathbf{K}^{x'x} & \mathbf{K}^{x'x'}\end{pmatrix}^{-1}\\ =&\begin{pmatrix}\mathbf{M}_1^{-1} & -(\mathbf{K}^{xx})^{-1}\mathbf{K}^{xx'}\mathbf{M}_2^{-1}\\ -\mathbf{M}_2^{-1}\mathbf{K}^{x'x}(\mathbf{K}^{xx})^{-1} & \mathbf{M}_2^{-1}\end{pmatrix}\end{aligned}$$

where

$$\begin{aligned}\mathbf{M}_1&=\mathbf{K}^{xx}-\mathbf{K}^{xx'}(\mathbf{K}^{x'x'})^{-1}\mathbf{K}^{x'x}\\ \mathbf{M}_2&=\mathbf{K}^{x'x'}-\mathbf{K}^{x'x}(\mathbf{K}^{xx})^{-1}\mathbf{K}^{xx'}\end{aligned}$$

We are now ready to derive the closed-form transition formula. Consider a Gaussian function over $\mathbf{x}$ and $\mathbf{x}'$: $\mathcal{N}(\mathbf{x},\mathbf{x}';\boldsymbol{\mu},\boldsymbol{\Sigma})$, expressed in its corresponding canonical form $\mathcal{C}(\mathbf{x},\mathbf{x}';\mathbf{K},\mathbf{h},g)$. Assume that parameters $\mathbf{K}\in\mathbb{R}^{(2d)^2}$ and $\mathbf{h}\in\mathbb{R}^{2d}$ are given by:

$$\mathbf{K}=\begin{pmatrix}\mathbf{K}^{xx} & \mathbf{K}^{xx'}\\ \mathbf{K}^{x'x} & \mathbf{K}^{x'x'}\end{pmatrix},\quad \mathbf{h}=\begin{pmatrix}\mathbf{h}^{x}\\ \mathbf{h}^{x'}\end{pmatrix}$$

Denote belief and transition as weighted sums of Gaussians.

$$\begin{aligned}bel_{t-1}(\mathbf{x})&=\sum_{i=1}^{K_1}w_i\,\mathcal{N}(\mathbf{x};\mathbf{a}_i,\mathbf{A}_i)\\ &=\sum_{i=1}^{K_1}w_i\,\mathcal{C}(\mathbf{x};\mathbf{J}_i,\mathbf{m}_i,n_i)\\ &=\sum_{i=1}^{K_1}w_i\,\mathcal{C}(\mathbf{x},\mathbf{x}';\begin{pmatrix}\mathbf{J}_i & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix},\begin{pmatrix}\mathbf{m}_i\\ \mathbf{0}\end{pmatrix},n_i)\\ p(\mathbf{x}\mid\mathbf{u},\mathbf{x}')&=\sum_{j=1}^{K_2}v_j\,\mathcal{N}(\mathbf{x},\mathbf{x}';\boldsymbol{b}_j,\mathbf{B}_j)\\ &=\sum_{j=1}^{K_2}v_j\,\mathcal{C}(\mathbf{x},\mathbf{x}';\mathbf{K}_j,\mathbf{h}_j,g_j)\\ &=\sum_{j=1}^{K_2}v_j\,\mathcal{C}(\mathbf{x},\mathbf{x}';\begin{pmatrix}\mathbf{K}_j^{xx} & \mathbf{K}_j^{xx'}\\ \mathbf{K}_j^{xx'} & \mathbf{K}_j^{x'x'}\end{pmatrix},\begin{pmatrix}\mathbf{h}_j^{x}\\ \mathbf{h}_j^{x'}\end{pmatrix},g_j)\end{aligned}$$

Following these notations, by the multiplication and marginalization formulae, the transition update

$$\begin{aligned}&\bar{bel}_t(\mathbf{x})\\ =&\int p(\mathbf{x}\mid\mathbf{u},\mathbf{x}')\;bel_{t-1}(\mathbf{x}')d\mathbf{x}'\\ =&\int\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}w_iv_j\\ &\mathcal{C}(\mathbf{x},\mathbf{x}';\begin{pmatrix}\mathbf{J}_i+\mathbf{K}_j^{xx} & \mathbf{K}_j^{xx'}\\ \mathbf{K}_j^{xx'} & \mathbf{K}_j^{x'x'}\end{pmatrix},\begin{pmatrix}\mathbf{m}_i+\mathbf{h}_j^{x}\\ \mathbf{h}_j^{x'}\end{pmatrix},n_i+g_j)\;d\mathbf{x}'\\ =&\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}w_iv_j\,\mathcal{C}(\mathbf{x};\mathbf{L}_{ij},\mathbf{q}_{ij},m_{ij})\end{aligned}$$

with parameters

$$\begin{aligned}\mathbf{L}_{ij}&=\mathbf{J}_i+\mathbf{K}_j^{xx}-\mathbf{K}_j^{xx'}(\mathbf{K}_j^{x'x'})^{-1}\mathbf{K}_j^{xx'}\\ \mathbf{q}_{ij}&=\mathbf{m}_i+\mathbf{h}_j^{x}-\mathbf{K}_j^{xx'}(\mathbf{K}_j^{x'x'})^{-1}\mathbf{h}_j^{x'}\\ m_{ij}&=n_i+g_j+\\ &\quad\frac{1}{2}\left(d\log 2\pi-\log\det\mathbf{K}_j^{x'x'}+(\mathbf{h}_j^{x'})^T\mathbf{K}_j^{x'x'}\mathbf{h}_j^{x'}\right)\end{aligned}$$

Next, we seek to express above parameters back in the moments parameterization. By our representation of $bel_{t-1}(\mathbf{x})$ and $p(\mathbf{x}\mid\mathbf{u},\mathbf{x}')$:

$$\begin{aligned}\mathbf{J}_i=\mathbf{A}_i^{-1},&\quad\mathbf{m}_i=\mathbf{A}_i^{-1}\mathbf{a}_i,\\ \mathbf{K}_j=\mathbf{B}_j^{-1},&\quad\mathbf{h}_j=\mathbf{B}_j^{-1}\boldsymbol{b}_j\end{aligned}$$

Matching the above parameters with the block inversion formula, it is easy to see

$$\mathbf{C}_{ij}=\mathbf{L}_{ij}^{-1}=\left(\mathbf{A}_i^{-1}+(\mathbf{B}_j^{xx})^{-1}\right)^{-1}$$

$$\mathbf{c}_{ij} = \mathbf{C}_{ij}\left(\mathbf{A}_i^{-1}\mathbf{a}_i + (\mathbf{B}_j^{xx})^{-1}\boldsymbol{b}_j^x\right)$$

A Gaussian with mean $\mathbf{c}_{ij}$ and covariance $\mathbf{C}_{ij}$ is associated with scalar c_{ij} in its canonical form,

$$c_{ij} = -\frac{1}{2}\mathbf{c}_{ij}^T(\mathbf{C}_{ij})^{-1}\mathbf{c}_{ij} - \frac{d}{2}\log 2\pi - \frac{1}{2}\log\det\mathbf{C}_{ij}$$

Calculation of the leading constant z_{ij} requires matching the following condition

$$\log z_{ij} = m_{ij} - c_{ij}$$

With algebraic manipulations, it is easy to see z_{ij} is given by the following expression

$$z_{ij} = \left(\det\left(\mathbf{A}_i + \mathbf{B}_j^{xx}\right)\right)^{-1/2} \cdot \exp\left(-\frac{1}{2}\mathbf{a}_i^T\mathbf{A}_i^{-1}\mathbf{a}_i - \frac{1}{2}\boldsymbol{b}_j^T\mathbf{B}_j^{-1}\boldsymbol{b}_j + \frac{1}{2}(\mathbf{h}_j^{x'})^T\mathbf{K}_j^{x'x'}\mathbf{h}_j^{x'} + \frac{1}{2}\mathbf{c}_{ij}^T(\mathbf{C}_{ij})^{-1}\mathbf{c}_{ij}\right)$$

8.4 Multiplicative Bounds for Expected and Empirical Values of Real-Valued Functions

In this section, we provide our derivation of a bound between the expected value and sampled, empirical values of real-valued functions. A one-sided multiplicative inequality is provided in Vapnik (1998), while we provide our derivation of the other sided inequality. We then combine these two relative bounds to produce the inequality (9) which we previously used. Our derivation closely follows theorems from Vapnik (1998), especially Theorems 4.2, 4.2*, 5.2, 5.3 and 5.3*.

Let $\mathcal{F}$ denote a function class of indicator or real-valued functions. Given a set of l data points $z_1, \ldots, z_l$ from the distribution $\mathcal{Z}$, the averaged empirical value v over a function $f \in \mathcal{F}$ is defined as

$$v(f) = \frac{1}{l}\sum_{i=1}^{l} f(z_i)$$

while the expected value $P(f)$ is

$$P(f) = \int f(z)\,dz$$

In the following, for Theorem 5, $\mathcal{F}$ consists of indicator functions. For Theorems 6 and 7, $\mathcal{F}$ is a set of real-valued functions. Let $\Pi(l)$ denote the growth function that satisfies the inequality

$$\Pi(l) \leq l^h$$

where h is the VC-dimension or pseudo-dimension of $\mathcal{F}$.

Theorem 5. *(cf. Theorem 4.2 and 4.2* in Vapnik (1998)) The inequality*

$$P\left\{\sup_{f\in\mathcal{F}} \frac{v(f) - P(f)}{\sqrt{P(f)}} > \epsilon\right\} < 4\,\Pi(2l)\exp\left(-\frac{\epsilon^2 l}{4}\right) \tag{13}$$

holds true.

Proof. Consider two events constructed from a random and independent sample of size $2l$:

$$\mathcal{Q}_1 = \left\{z : \sup_{f\in\mathcal{F}} \frac{v_1(A_f) - P(A_f)}{\sqrt{P(A_f)}} > \epsilon\right\},$$

$$\mathcal{Q}_2 = \left\{z : \sup_{f\in\mathcal{F}} \frac{v_1(A_f) - v_2(A_f)}{\sqrt{v(A_f) + \frac{1}{2\log l}}} > \epsilon\right\},$$

where A_f is the event

$$A_f = \{z : f(z) = 1\}$$

$P(A_f)$ is the probability of event A_f:

$$P(A_f) = \int f(z)\,dz$$

$v_1(A_f)$ is the frequency of event A_f computed from the first half-sample $z_1, \ldots, z_l$ of the sample $z_1, \ldots, z_{2l}$:

$$v_1(A_f) = \frac{1}{l}\sum_{i=1}^{l} f(z_i)$$

and $v_2(A_f)$ is the frequency of event A_f computed from the second half-sample $z_{l+1}, \ldots, z_{2l}$:

$$v_2(A_f) = \frac{1}{l}\sum_{i=l+1}^{2l} f(z_i)$$

Denote $v(A_f) = \frac{1}{2}(v_1(A_f) + v_2(A_f))$. Note that in the case $l \leq \epsilon^{-2}$, the assertion of the theorem is trivial as the right-hand side of the inequality exceeds one. Accordingly we shall prove the theorem as follows: First we show that for $l > \epsilon^{-2}$ the inequality $P(\mathcal{Q}_1) < 4P(\mathcal{Q}_2)$ is valid (Lemma 5.1), and then we bound $P(\mathcal{Q}_2)$ (Lemma 5.2). □

Lemma 5.1. *(cf. Lemma 4.1 in Vapnik (1998)) For* $l > \max\{\exp(-\frac{\sqrt{\epsilon^2+4}+\epsilon}{2\sqrt{2}\epsilon}), [1 - \frac{1}{4}(-\epsilon + \sqrt{\epsilon^2+4})^2]^{-1}\}$, *the probability*

$$P(\mathcal{Q}_1) < \frac{1}{4}P(\mathcal{Q}_2)$$

is valid.

Proof. Assume that $\mathcal{Q}_1$ has occurred. This means that there exists event A^* such that for the first half-sample the equality

$$v_1(A^*) - P(A^*) > \epsilon\sqrt{P(A^*)}$$

is fulfilled. Since $v_1(A^*) < 1$, this implies that $\epsilon\sqrt{P(A^*)} + P(A^*) < 1$.

Let $f(x) = x^2 + \epsilon x - 1$ with $x \in [0,1]$. Then $f'(x) = 2x+\epsilon > 0$ for all x. Hence $f(x)$ is strictly increasing on $[0,1]$. Notice $f(0) = -1 < 0$, $f(1) = \epsilon > 0$. Thus there exists a root for $f(x) = 0$ on the interval $[0,1]$. There are two solutions to $f(x) = 0$: $x_1 = \frac{1}{2}(-\epsilon - \sqrt{\epsilon^2+4}) < 0$ (rejected), $x_2 = \frac{1}{2}(-\epsilon + \sqrt{\epsilon^2+4}) \in [0,1]$. Hence $\epsilon\sqrt{P(A^*)} + P(A^*) < 1$ implies $\sqrt{P(A^*)} < \frac{1}{2}(-\epsilon + \sqrt{\epsilon^2+4})$.

Assume that for the second half-sample the frequency of the event A^* is less than the probability $P(A^*)$:

$$v_2(A^*) < P(A^*)$$

Under these conditions, we prove that event $\mathcal{Q}_2$ will definitely occur. To do this, we bound the quantity

$$\mu = \frac{v_1(A^*) - v_2(A^*)}{\sqrt{v(A^*) + \frac{1}{2\log l}}}$$

under the conditions

$$\begin{aligned} v_1(A^*) &> P(A^*) + \epsilon\sqrt{P(A^*)} \\ v_2(A^*) &< P(A^*) \\ \sqrt{P(A^*)} &< \frac{1}{2}(-\epsilon + \sqrt{\epsilon^2+4}) \end{aligned}$$

For this purpose, we find the minimum of the function

$$T = \frac{x-y}{\sqrt{x+y+c}}$$

in the domain $0 < a \le x \le 1$, $0 < y \le b$, $c > 0$. We have

$$\begin{aligned} \frac{\partial T}{\partial x} &= \frac{1}{2}\frac{x+3y+2c}{(x+y+c)^{3/2}} > 0 \\ \frac{\partial T}{\partial y} &= -\frac{1}{2}\frac{3x+y+2c}{(x+y+c)^{3/2}} < 0 \end{aligned}$$

Consequently, T attains its minimum in the admissible domain at the boundary points $x = a$ and $y = b$.

Specific to the quantity μ, the above boundary points are equivalent to the conditions when $x = v_1(A^*) = P(A^*) + \epsilon\sqrt{P(A^*)}$ and $y = v_2(A^*) = P(A^*)$. Thus, the quantity μ is bounded from below,

$$\mu \ge \frac{\epsilon\sqrt{2P(A^*)}}{\sqrt{2P(A^*) + \epsilon\sqrt{P(A^*)} + \frac{\sqrt{2}}{2\log l}}}$$

From the given conditions, observe that

$$\begin{aligned} l &> \exp\left(-\frac{\sqrt{\epsilon^2+4}+\epsilon}{2\sqrt{2}\epsilon}\right) && \Leftrightarrow \\ 2\log l &> -\frac{\sqrt{\epsilon^2+4}+\epsilon}{\sqrt{2}\epsilon} && \Leftrightarrow \\ \frac{\sqrt{2}}{2\log l} &< -\frac{1}{2}\epsilon(\sqrt{\epsilon^2+4}-\epsilon) \end{aligned}$$

Since $\sqrt{P(A^*)} < \frac{1}{2}(-\epsilon + \sqrt{\epsilon^2+4})$ and $\frac{\sqrt{2}}{2\log l} < -\frac{1}{2}\epsilon(-\epsilon + \sqrt{\epsilon^2+4})$, we have:

$$\mu \ge \frac{\epsilon\sqrt{2P(A^*)}}{\sqrt{2P(A^*) + \frac{1}{2}\epsilon(-\epsilon+\sqrt{\epsilon^2+4}) - \frac{1}{2}\epsilon(-\epsilon+\sqrt{\epsilon^2+4})}} = \epsilon$$

Thus, if $\mathcal{Q}_1$ occurs and the condition $v_2(A^*) < P(A^*)$ is satisfied, then $\mathcal{Q}_2$ occurs as well.

The second half-sample is chosen independently of the first one. By Corollary 3 in Greenberg and Mohri (2014), the event

$$v_2(A^*) < P(A^*)$$

occurs with probability exceeding 1/4 if

$$\begin{aligned} P(A^*) &< \frac{1}{4}(-\epsilon + \sqrt{\epsilon^2+4})^2 < 1 - \frac{1}{l} \Rightarrow \\ l &> [1 - \frac{1}{4}(-\epsilon + \sqrt{\epsilon^2+4})^2]^{-1} \end{aligned}$$

This is fulfilled by the condition of the lemma. Thus, we have

$$P(\mathcal{Q}_2) > \frac{1}{4}P(\mathcal{Q}_1)$$

□

Lemma 5.2. *(cf. Lemma 4.2 in Vapnik (1998)) For $l > \exp\left(-\frac{\sqrt{\epsilon^2+4}+\epsilon}{\sqrt{2}\epsilon}\right)$, the bound*

$$P(\mathcal{Q}_2) < \Pi(2l)\exp\left(-\frac{\epsilon^2 l}{4}\right)$$

is valid.

Proof. Denote by $R_A(Z^{2l})$ the quantity

$$R_A(Z^{2l}) = \frac{v_1(A) - v_2(A)}{\sqrt{v(A) + 1/(2\log l)}}$$

then the estimated probability equals

$$P(\mathcal{Q}_2) = \int_{Z(2l)} \theta\left[\sup_{A\in S} R_A(Z^{2l}) - \epsilon\right] dF(Z^{2l})$$

where θ is the sign function. Here the integration is carried out over the space of all possible samples of size $2l$.

Consider now all possible permutations T_i, $i = 1, 2, \ldots, (2l)!$ of the sequence $z_1, \ldots, z_{2l}$. For each such permutation the equality

$$\int_{Z(2l)} \theta \left[\sup_{A \in S} R_A(Z^{2l}) - \epsilon \right] dF(Z^{2l}) = \int_{Z(2l)} \theta \left[\sup_{A \in S} R_A(T_i Z^{2l}) - \epsilon \right] dF(Z^{2l})$$

is valid. Therefore the equality

$$\begin{aligned} P(\mathcal{Q}_2) &= \int_{Z(2l)} \theta \left[\sup_{A \in S} R_A(Z^{2l}) - \epsilon \right] dF(Z^{2l}) \\ &= \int_{Z(2l)} \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta \left[\sup_{A \in S} R_A(T_i Z^{2l}) - \epsilon \right] dF(Z^{2l}) \end{aligned} \tag{14}$$

is valid.

Now consider the integrand. Since the sample $z_1, \ldots, z_{2l}$ is fixed, instead of the system of events S one can consider a finite system of events S^* which contains one representative for each one of the equivalence classes. Thus the equality

$$\frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta \left[\sup_{A \in S} R_A(T_i Z^{2l}) - \epsilon \right] = \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta \left[\sup_{A \in S^*} R_A(T_i Z^{2l}) - \epsilon \right]$$

is valid. Furthermore,

$$\begin{aligned} &\frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta \left[\sup_{A \in S} R_A(T_i Z^{2l}) - \epsilon \right] \\ &< \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \sum_{A \in S^*} \theta \left[R_A(T_i Z^{2l}) - \epsilon \right] \\ &= \sum_{A \in S^*} \left\{ \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta \left[R_A(T_i Z^{2l}) - \epsilon \right] \right\} \end{aligned}$$

The expression in braces is the probability of greater than ϵ deviation of the frequencies in two half-samples for a fixed event A and a given composition of a complete sample. This probability equals

$$\Gamma = \sum_k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l}$$

where m is the number of occurrences of event A in a complete sample; k is the number of occurrences of the event in the first half-sample and runs over these values:

$$\begin{gathered} \max(0, m-l) \leq k \leq \min(m, l) \\ \frac{k}{l} - \frac{m-k}{l} > \epsilon \sqrt{\frac{m}{2l} + \frac{1}{2 \log l}} \end{gathered}$$

Denote by ϵ^* the quantity

$$\sqrt{\frac{m}{2l} + \frac{1}{2 \log l}} \epsilon = \epsilon^*$$

Using this notation the constraints become

$$\begin{gathered} \max(0, m-l) \leq k \leq \min(m, l) \\ \frac{k}{l} - \frac{m-k}{l} > \epsilon^* \end{gathered} \tag{15}$$

It can be shown[2] that the following bound on the quantity Γ under constraints (15) is valid:

$$\Gamma < \exp\left(- \frac{(l+1)(\epsilon^*)^2 l^2}{(m+1)(2l-m+1)} \right) \tag{16}$$

Substituting in ϵ^*,

$$\begin{aligned} \Gamma &< \exp\left(- \frac{(l+1)\epsilon^2 l^2}{(m+1)(2l-m+1)} \left(\frac{m}{2l} + \frac{1}{2 \log l}\right) \right) \\ &< \exp\left(- \frac{(l+1)\epsilon^2 l^2}{(m+1)(2l-m+1)} \frac{m+1}{2l} \right) \end{aligned}$$

The second inequality is derived by noting $\frac{1}{\log l} > \frac{1}{l}$. For the inequality, $\Gamma < \exp\left(- \frac{(l+1)\epsilon^2 l^2}{(m+1)(2l-m+1)} \frac{m+1}{2l} \right)$, the right-hand side reaches its maximum at $m = 0$. Thus,

$$\Gamma < \exp\left(- \frac{\epsilon^2 l}{4} \right) \tag{17}$$

Substituting (17) into the right-hand side of (14) and integrating, we have

$$\begin{aligned} P(\mathcal{Q}_2) &= \int_{Z(2l)} N^S(Z^{2l}) \exp\left(- \frac{\epsilon^2 l}{4} dF(Z^{2l}) \right) \\ &< \Pi(2l) \exp\left(- \frac{\epsilon^2 l}{4} \right) \end{aligned}$$

□

The above theorems are for indicator functions. We next consider the case of real-valued functions, whose probability bounds are directly dependent on the above binary bounds.

[2] See Section 4.13 of Vapnik (1998).

Theorem 6. *(cf. Theorem 5.2 in Vapnik (1998)) Let $\mathcal{F}$ be a set of real-valued, non-negative functions. Let $\Pi(l)$ be the growth function of indicators for this set of functions. Let auxiliary function $D(f) = \int_0^\infty \sqrt{P\{f(z) > c\}}dc$. Then, the inequality*

$$P\left\{\sup_{f\in\mathcal{F}} \frac{v(f) - P(f)}{D(f)} > \epsilon\right\} < 4\,\Pi(2l)\exp\left(-\frac{\epsilon^2 l}{4}\right) \tag{18}$$

is valid.

Proof. Consider the expression

$$\sup_{f\in\mathcal{F}} \frac{v(f) - P(f)}{D(f)}$$
$$= \sup_{f\in\mathcal{F}} \frac{\lim_{n\to\infty}\left[\sum_{i=1}^\infty \frac{1}{n} v\left\{f(z) > \frac{i}{n}\right\} - \sum_{i=1}^\infty \frac{1}{n} P\left\{f(z) > \frac{i}{n}\right\}\right]}{D(f)} \tag{19}$$

We show that if inequality

$$\sup_{f\in\mathcal{F}} \frac{v\left\{f(z) > \frac{i}{n}\right\} - P\left\{f(z) > \frac{i}{n}\right\}}{\sqrt{P\left\{f(z) > \frac{i}{n}\right\}}} \le \epsilon \tag{20}$$

is fulfilled, then the inequality

$$\sup_{f\in\mathcal{F}} \frac{v(f) - P(f)}{D(f)} \le \epsilon \tag{21}$$

is fulfilled as well.

Indeed, equation (19) and inequality (20) imply that

$$\sup_{f\in\mathcal{F}} \frac{v(f) - P(f)}{D(f)}$$
$$\le \sup_{f\in\mathcal{F}} \frac{\epsilon \lim_{n\to\infty} \sum_{i=1}^\infty \frac{1}{n}\sqrt{P\{f(z) > \frac{i}{n}\}}}{D(f)} = \sup_{f\in\mathcal{F}} \frac{\epsilon D(f)}{D(f)} = \epsilon$$

Therefore probability of event (20) does not exceed probability of event (21). This means that the probability of the complementary events are connected by the inequality

$$P\left\{\sup_{f\in\mathcal{F}} \frac{v(f) - P(f)}{D(f)} > \epsilon\right\}$$
$$\le P\left\{\sup_{f\in\mathcal{F}} \frac{v\left\{f(z) > \beta\right\} - P\left\{f(z) > \beta\right\}}{\sqrt{P\left\{f(z) > \beta\right\}}} > \epsilon\right\}$$

In Theorem 5 we bounded the right-hand side of this inequality. Using this bound we prove the theorem. □

Theorem 7. *(cf. Theorem 5.3 and 5.3* in Vapnik (1998)) Assume that functions f are bounded above by M: $0 \le f(z) \le M, f \in \mathcal{F}$. Then, the inequality*

$$P\left\{\sup_{f\in\mathcal{F}} \frac{v(f) - P(f)}{\sqrt{P(f)}} > \epsilon\right\} < 4\,\Pi(2l)\exp\left(-\frac{\epsilon^2 l}{4M}\right) \tag{22}$$

is valid.

Proof. The proof is based on Holder's inequality for two functions. We say that function $f(z)$ belongs to space $L_p(a,b)$ if $\int_a^b |f(z)|^p\, dz \le \infty$. The values a, b are not necessarily finite. Holder's inequality states that for functions $f(z) \in L_p(a,b)$ and $g(z) \in L_q(a,b)$, where

$$\frac{1}{p} + \frac{1}{q} = 1, p > 0, q > 0$$

then

$$\int_a^b |f(z)g(z)|\, dz \le \left(\int_a^b |f(z)|^p\, dz\right)^{\frac{1}{p}} \left(\int_a^b |g(z)|^q\, dz\right)^{\frac{1}{q}}$$

Consider the function

$$D(f) = \int_0^\infty \sqrt{P\left\{f(z) > c\right\}}dc$$

For a bounded set of functions, we can rewrite this expression in the form

$$D(f) = \int_0^M \sqrt{P\left\{f(z) > c\right\}}dc$$

Now let us denote $f(z) = \sqrt{P\left\{f(z) > c\right\}}$ and denote $g(z) = 1$. Using these notations we utilize Holder's inequality. We obtain

$$D(f) = \int_0^M \sqrt{P\{f(z) > t\}}dt$$
$$< \left(\int_0^M P\{f(z) > t\}dt\right)^{1/2} M^{1/2}$$

Taking into account this inequality, we obtain

$$P\left\{\sup_{f\in\mathcal{F}} \frac{v(f) - P(f)}{\sqrt{P(f)}} > \epsilon M^{1/2}\right\}$$
$$\le P\left\{\sup_{f\in\mathcal{F}} \frac{v(f) - P(f)}{\int \sqrt{P\{f(z) > t\}dt}} > \epsilon\right\}$$

Using the bound on the right-hand side of this inequality given by Theorem 6, we obtain the desired inequality (22). □

8.4.1 Combining Inequalities

Under similar settings as Theorem 7, the following inequality is provided in the original book by Vapnik (1998):

$$P\left\{\sup_{f\in\mathcal{F}} \frac{P(f)-v(f)}{\sqrt{P(f)}} > \epsilon\right\} < 4\,\Pi(2l)\exp\left(-\frac{\epsilon^2 l}{4M}\right) \tag{23}$$

Combining the above inequality with Theorem 7, we obtain the following two inequalities for a bounded, real-valued function f: $0 \le f(z) \le M, f \in \mathcal{F}$:

$$P\left\{\sup_{f\in\mathcal{F}} \frac{P(f)-v(f)}{\sqrt{P(f)}} > \epsilon\right\} < 4\,\Pi(2l)\exp\left(-\frac{\epsilon^2 l}{4M}\right)$$
$$P\left\{\sup_{f\in\mathcal{F}} \frac{v(f)-P(f)}{\sqrt{P(f)}} > \epsilon\right\} < 4\,\Pi(2l)\exp\left(-\frac{\epsilon^2 l}{4M}\right)$$

Equivalently, for all $f \in \mathcal{F}$,

$$P\left\{\frac{P(f)-v(f)}{\sqrt{P(f)}} < \epsilon\right\} > 1 - 4\,\Pi(2l)\exp\left(-\frac{\epsilon^2 l}{4M}\right)$$
$$P\left\{\frac{v(f)-P(f)}{\sqrt{P(f)}} < \epsilon\right\} > 1 - 4\,\Pi(2l)\exp\left(-\frac{\epsilon^2 l}{4M}\right)$$

Setting $\epsilon = \zeta\sqrt{P(f)}$ for a given ζ, we get

$$P\left\{v(f) > (1-\zeta)P(f)\right\} > 1 - 4\,\Pi(2l)\exp\left(-\frac{\zeta^2 l P(f)}{4M}\right)$$
$$P\left\{v(f) < (1+\zeta)P(f)\right\} > 1 - 4\,\Pi(2l)\exp\left(-\frac{\zeta^2 l P(f)}{4M}\right)$$

Consequently

$$P\left\{v(f) < (1-\zeta)P(f)\right\} < 4\,\Pi(2l)\exp\left(-\frac{\zeta^2 l P(f)}{4M}\right)$$
$$P\left\{v(f) > (1+\zeta)P(f)\right\} < 4\,\Pi(2l)\exp\left(-\frac{\zeta^2 l P(f)}{4M}\right)$$

The two events $E_1 = \{f : v(f) \le (1-\zeta)P(f)\}$ and $E_2 = \{f : v(f) \ge (1+\zeta)P(f)\}$ are mutually exclusive. Hence the probability of the union,

$$\begin{aligned} P(E_1 \cup E_2) &= P\left(v(f) \notin [(1-\zeta)P(f), (1+\zeta)P(f)]\right) \\ &= P(E_1) + P(E_2) \\ &= 8\,\Pi(2l)\exp\left(-\frac{\zeta^2 l P(f)}{4M}\right) \\ &< 8\,\Pi(2l)\exp\left(-\frac{\zeta^2}{4}\frac{B}{M} l\right) \end{aligned}$$

This gives the desired inequality we used in Section 8.1, where B is a lower bound of the value $P(f)$, and M is an upper bound of functions $f \in \mathcal{F}$.