1 Covering and Packing Numbers

Let \( T \) be a subset of a metric space \((X, \rho)\). By an \( \epsilon \)-packing subset of \( T \), we mean a finite set \( t_1, \ldots, t_n \) of points in \( T \) for which \( \rho(t_i, t_j) > \epsilon \) whenever \( i \neq j \). The \( \epsilon \)-packing number, \( N(\epsilon, T) \), of \( T \) is defined as the largest \( n \) for which there exists an \( \epsilon \)-packing subset of size \( n \).

By an \( \epsilon \)-cover of \( T \), we mean a finite set of points \( x_1, \ldots, x_m \) in \( X \) for which
\[
\sup_{t \in T} \min_{1 \leq i \leq m} \rho(t, x_i) \leq \epsilon.
\]
The \( \epsilon \)-covering number, \( M(\epsilon, T) \), of \( T \) is defined as the smallest \( m \) for which there exists an \( \epsilon \)-cover of \( T \) of size \( m \).

**Lemma 1.1.** For every \( \epsilon > 0 \), we have
\[
M(\epsilon, T) \geq N(\epsilon, T) \leq M(\epsilon/2, T).
\]

**Proof.** If \( n = N(\epsilon, T) \), then there exists a maximal set of points \( t_1, \ldots, t_n \) with \( \rho(t_i, t_j) > \epsilon \). Because of maximality, every other point \( t \in T \) should be within \( \epsilon \) of some \( t_i \) and thus \( t_1, \ldots, t_n \) form an \( \epsilon \)-cover of \( T \) which implies that \( M(\epsilon, T) \leq n = N(\epsilon, T) \).

Conversely, for every \( \epsilon \)-packing subset \( t_1, \ldots, t_n \) of \( T \), the closed balls \( B(t_i, \epsilon/2), i = 1, \ldots, n \) are disjoing and hence every \( \epsilon/2 \)-cover of \( T \) must have one point in each of the balls \( B(t_i, \epsilon/2) \). As a result, an \( \epsilon/2 \)-cover of \( T \) must have at least \( n \) points. This implies that \( M(\epsilon/2, T) \geq N(\epsilon, T) \). \( \square \)

**Lemma 1.2** (Volumetric Argument). Let \( T = \mathcal{X} \) denote the ball in \( \mathbb{R}^d \) of radius \( \Gamma \) centered at the origin. Then, under the usual Euclidean metric, we have
\[
M(\epsilon, T) \geq \left( \frac{\Gamma}{\epsilon} \right)^d \quad \text{and} \quad N(\epsilon, T) \leq \left( 1 + \frac{2\Gamma}{\epsilon} \right)^d.
\]

**Proof.** For every \( \epsilon \)-cover of \( T \), the whole set \( T \) is clearly contained in the union of the balls of radius \( \epsilon \) with centers in the cover. Therefore, the volume of \( T \)
must be smaller than the sum of the volumes of these balls. Thus the number of points in the \( \epsilon \)-cover must be at least \( (\frac{\Gamma}{\epsilon})^d \). Therefore, \( M(\epsilon, T) \geq (\frac{\Gamma}{\epsilon})^d \).

For every \( \epsilon \)-packing subset of \( T \), the balls of radius \( \epsilon/2 \) with centers in the packing set are all disjoint and their union is contained in the ball of radius \( \Gamma + (\epsilon/2) \) centered at the origin. Consequently, the sum of the volumes of these balls is smaller than the volume of the ball of radius \( \Gamma + (\epsilon/2) \) centered at the origin. Therefore, the number of points in the \( \epsilon \)-packing subset is at most \( (1 + (2\Gamma/\epsilon))^d \). Thus \( N(\epsilon, T) \geq (1 + (2\Gamma/\epsilon))^d \). \( \square \)

2 Smooth Functions

2.1 One Dimension

Fix \( \alpha > 0 \). Let \( \beta \) denote the largest integer that is STRICTLY smaller than \( \alpha \). For example, if \( \alpha = 5 \), then \( \beta = 4 \) and if \( \alpha = 5.2 \), then \( \beta = 5 \).

The class \( \mathcal{S}_\alpha \) is defined to consist of functions \( f \) on \([0, 1]\) that satisfy all the following properties:

1. \( f \) is continuous on \([0, 1]\).
2. \( f \) is differentiable \( \beta \) times on \((0, 1)\).
3. \(|f^{(k)}(x)| \leq 1\) for all \( k = 0, \ldots, \beta \) and \( x \in [0, 1] \) where \( f^{(0)}(x) := f(x) \).
4. \(|f^{(\beta)}(x) - f^{(\beta)}(y)| \leq |x - y|^\alpha - \beta \) for all \( x, y \in (0, 1) \).

Let \( \rho \) denote the supremum metric on \( \mathcal{S}_\alpha \) defined by

\[
\rho(f, g) := \sup_{x \in [0, 1]} |f(x) - g(x)| = \sup_{x \in (0, 1)} |f(x) - g(x)|. \tag{1}
\]

**Theorem 2.1.** There exist positive constants \( \epsilon_0, C_1 \) and \( C_2 \) depending on \( \alpha \) alone such that for all \( 0 < \epsilon \leq \epsilon_0 \), we have

\[
M(\epsilon, \mathcal{S}_\alpha) \leq \exp \left( C_1 \epsilon^{-1/\alpha} \right) \quad \text{and} \quad N(\epsilon, \mathcal{S}_\alpha) \geq \exp \left( C_2 \epsilon^{-1/\alpha} \right).
\]

We will use the Taylor’s theorem in the proof. For every \( f \in \mathcal{S}_\alpha \) and \( x, x+h \in (0, 1) \), we can write

\[
f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!} f''(x) + \cdots + \frac{h^{\beta-1}}{(\beta-1)!} f^{(\beta-1)}(x) + \frac{h^\beta}{\beta!} f^{(\beta)}(y).
\]
Thus if 
\[ R_f(x, h) := f(x+h) - f(x) - hf'(x) - \frac{h^2}{2!} f''(x) - \cdots - \frac{h^{\beta-1}}{(\beta-1)!} f^{(\beta-1)}(x) - \frac{h^\beta}{\beta!} f^{(\beta)}(x), \]
then 
\[ |R_f(x, h)| = \frac{h^\beta}{\beta!} |f^{(\beta)}(x) - f^{(\beta)}(y)| \leq \frac{|h|^\alpha}{\beta!}. \]
Thus for every function \( f \) in \( S_\alpha \) and \( x, x + h \in (0, 1) \), we have
\[
f(x + h) = \sum_{k=0}^{\beta} \frac{h^k}{k!} f^{(k)}(x) + R_f(x, h) \quad \text{with } |R_f(x, h)| \leq |h|^\alpha / (\beta!). \quad (2)
\]
Moreover, for every \( f \in S_\alpha \) and \( 0 \leq i \leq \beta \), the derivative \( g = f^{(i)} \in S_{\alpha-i} \) and hence using the above, we obtain
\[
f^{(i)}(x + h) = \sum_{k=0}^{\beta-i} \frac{h^k}{k!} f^{(i+k)}(x) + R_{f^{(i)}}(x, h) \quad \text{with } |R_{f^{(i)}}(x, h)| \leq |h|^{\alpha-i} / (\beta!). \quad (3)
\]

**Proof of Theorem 2.1.** Throughout, \( C(\alpha) \) will be a generic positive constant that depends on \( \alpha \) alone and whose value might be different in different appearances.

Let us first prove the upper bound on \( M(\epsilon, S_\alpha) \). Fix \( \epsilon > 0 \).

Fix \( x \in (0, 1) \). Suppose that two functions \( f \) and \( g \) in \( S_\alpha \) satisfy
\[
|f^{(k)}(x) - g^{(k)}(x)| \leq \epsilon_k \quad \text{for all } k = 0, \ldots, \beta
\]
for some \( \epsilon_0, \ldots, \epsilon_\beta > 0 \). What then is a good upper bound on \( |f(x+h) - g(x+h)| \) for some \( h \) such that \( x + h \in (0, 1) \)? Write
\[
|f(x+h) - g(x+h)| = \left| \sum_{k=0}^\beta \frac{h^k}{k!} \left( f^{(k)}(x) - g^{(k)}(x) \right) + R_f(x, h) - R_g(x, h) \right|
\]
\[
\leq \sum_{k=0}^\beta \frac{|h|^k \epsilon_k}{k!} + 2|h|^{\alpha} / \beta!.
\]
Therefore if \( |h| \leq \epsilon^{1/\alpha} \) and \( \epsilon_k = \epsilon^{1-(k/\alpha)} \), then
\[
|f(x+h) - g(x+h)| \leq \epsilon \left( \sum_{k=0}^\beta \frac{1}{k!} + \frac{2}{\beta!} \right) = C(\alpha) \epsilon
\]
whenever (4) is satisfied and \( |h| \leq \epsilon^{1/\alpha} \).
The upshot is that if we consider a grid of points \( \epsilon^{1/\alpha} \)-apart in \((0, 1)\) and then, at each grid point, cover the \( k \)th derivative of functions in \( S_\alpha \) to within \( \epsilon_k = \epsilon^{1-(k/\alpha)} \), then we would end up with an \( C(\alpha)\epsilon \) cover in the supremum metric for \( S_\alpha \).

Let \( x_1, \ldots, x_s \) be an \( \epsilon^{1/\alpha} \)-grid of points in \((0, 1)\) so that \( s \leq C(\alpha)\epsilon^{-1/\alpha} \). Define, for each \( f_0 \in S_\alpha \),

\[
G(f_0) := \left\{ f \in S_\alpha : \left[ \frac{f^{(k)}(x_i)}{\epsilon_k} \right] \leq \left[ \frac{f_0^{(k)}(x_i)}{\epsilon_k} \right] \text{ for all } i = 1, \ldots, s \text{ and } k = 0, \ldots, \beta \right\},
\]

where \([x]\) denotes the largest integer that is less than or equal to \( x \). From the above argument, the number of distinct sets \( G(f_0) \) as \( f_0 \) ranges over \( S_\alpha \) is an upper bound on the \( C(\alpha)\)-covering number of \( S_\alpha \). Clearly, \( G(f_0) \) only depends on the numbers \([f_0^{(k)}(x_i)/\epsilon_k] \) for \( i = 1, \ldots, s \) and \( k = 0, \ldots, \beta \) and so the number of distinct sets \( G(f_0) \) is at most the cardinality of

\[
I := \left\{ \left[ \frac{f^{(k)}(x_i)}{\epsilon_k} \right], i = 1, \ldots, s \text{ and } k = 0, \ldots, \beta : f \in S_\alpha \right\}.
\]

Let us assume that \( x_1 < \cdots < x_s \). We can start counting from \( x_1 \). What is the number of possible values of the vector \([f^{(k)}(x_1)/\epsilon_k], k = 0, \ldots, \beta \) as \( f \) ranges over \( S_\alpha \)? Since \(|f^{(k)}(x_1)| \leq 1\), clearly this number is bounded from above by:

\[
\frac{1}{\epsilon_0} \frac{1}{\epsilon_1} \cdots \frac{1}{\epsilon_\beta} \leq \left( \frac{1}{\epsilon} \right)^{\beta+1}.
\]

The above number is actually smaller than \( \epsilon^{-\beta-1} \) but this bound is good enough for us. Now we come to \( x_2 \). Given the values of \([f^{(k)}(x_1)/\epsilon_k], k = 0, \ldots, \beta \), what is the number of possible values of the vector \([f^{(k)}(x_2)/\epsilon_k], k = 0, \ldots, \beta \)? Suppose, for each \( 0 \leq k \leq \beta \),

\[
A_k := \left[ \frac{f^{(k)}(x_1)}{\epsilon_k} \right] \quad \text{so that } A_k\epsilon_k \leq f^{(k)}(x_1) < (A_k + 1)\epsilon_k.
\]

We fix \( 0 \leq i \leq \beta \) and use (3) with \( x = x_1, h = x_2 - x_1 \) to get:

\[
\left| f^{(i)}(x_2) - \sum_{k=0}^{\beta-i} \frac{h^k}{k!} f^{(i+k)}(x_1) \right| \leq \frac{|h|^{\alpha-i}}{\beta!}.
\]

As a result

\[
\left| f^{(i)}(x_2) - \sum_{k=0}^{\beta-i} \frac{h^k}{k!} A_{i+k} \right| \leq \left| f^{(i)}(x_2) - \sum_{k=0}^{\beta-i} \frac{h^k}{k!} f^{(i+k)}(x_1) \right| + \sum_{k=0}^{\beta-i} \frac{h^k}{k!} (f^{(i+k)}(x_1) - A_{i+k})
\]

\[
\leq \frac{|h|^{\alpha-i}}{\beta!} + \sum_{k=0}^{\beta-i} \frac{|h|^{k}}{k!} \epsilon_{i+k}.
\]
Because \( x_1, \ldots, x_s \) form a \( \epsilon^{1/\alpha} \)-grid, we have \( |h| = |x_2 - x_1| \leq C(\alpha)\epsilon^{1-\alpha} \). It can then be checked that

\[
|f(i)(x_2) - \sum_{k=0}^{\beta-i} h^k A_{i+k}| \leq C(\alpha) \frac{\epsilon}{\epsilon^{1/\alpha}} = C(\alpha)\epsilon_i.
\]

Thus given the values of \( [f^{(k)}(x_1)/\epsilon_k] \) for \( k = 0, \ldots, \beta \), the derivative \( f(i)(x_2) \) takes values in an interval of length at most \( C(\alpha)\epsilon_i \). Therefore, \( [f(i)(x_2)/\epsilon_i] \) takes at most a constant, \( C(\alpha) \), number of values.

We have thus proved that given the values of \( [f^{(k)}(x_1)/\epsilon_k] \) for \( k = 0, \ldots, \beta \), the number of possible values of \((f^{(k)}(x_2)/\epsilon_k), k = 0, \ldots, \beta\) is at most \( C(\alpha)\epsilon_i \). Therefore, \( [f(i)(x_2)/\epsilon_i] \) takes at most a constant, \( C(\alpha) \), number of values.

The exact same conclusion holds if \( x_1 \) and \( x_2 \) are replaced by \( x_j \) and \( x_j+1 \) for any \( 1 \leq j \leq s-1 \). Combining this with (6), we obtain that the cardinality of \( I \), defined in (5), is at most

\[
|I| \leq \left(\frac{1}{\epsilon}\right)^{\beta+1} (C(\alpha))^s \leq \left(\frac{1}{\epsilon}\right)^{\beta+1} (C(\alpha))^{C(\alpha)\epsilon^{-1/\alpha}} \leq \exp\left((\beta+1)\log(1/\epsilon) + \epsilon^{-1/\alpha}C(\alpha)\log C(\alpha)\right) \leq \exp\left(C(\alpha)\epsilon^{-1/\alpha}\right).
\]

This completes the proof of \( M(\epsilon,\mathcal{S}_\alpha) \leq \exp(C(\alpha)\epsilon^{-1/\alpha}) \).

We now prove the lower bound on \( N(\epsilon,\mathcal{S}_\alpha) \). There exists a function \( f_0 \) defined on the whole real line \( \mathbb{R} \) such that

1. \( f_0(x) = 0 \) for \( x \notin (0,1) \) and \( f_0(x) > 0 \) for \( x \in (0,1) \).
2. \( f_0 \) restricted to the interval \([0,1]\) lies in \( \mathcal{S}_\alpha \).
3. \( f_0(1/2) = \max_{0 \leq x \leq 1} f_0(x) > 0 \).

For example, one can take \( f_0(x) = ce^{-x/\beta}e^{-1/(1-x)} \) for \( 0 < x < 1 \) and \( 0 \) outside \((0,1)\) and choose the constant \( c \) so that \( f_0 \in \mathcal{S}_\alpha \). Given such a function, consider points

\[
0 < a_1 < b_1 < a_2 < b_2 < \cdots < a_s < b_s < 1
\]

where \( b_i - a_i = \epsilon^{1/\alpha} \) and \( s \geq C(\alpha)\epsilon^{-1/\alpha} \). For each \( j = 1, \ldots, s \), define

\[
g_i(x) := (b-a)^\alpha f_0 \left( \frac{t-a}{b-a} \right).
\]
It is easy to check that \( g_i \in \mathcal{S}_\alpha \) and \( g_i \) is supported on \((a_i, b_i)\). Now for \( \tau \in \{0,1\}^s \), define the function

\[
u_\tau(x) := \sum_{i=1}^s \tau_i g_i(x).
\]

Once again, it is straightforward to see that \( \nu_\tau \in \mathcal{S}_\alpha \) for each \( \tau \in \{0,1\}^k \). Moreover, if \( \tau, \tau' \in \{0,1\}^s \) with \( \tau_j \neq \tau'_j \), then

\[
\rho(\nu_\tau, \nu_{\tau'}) \geq g_j \left( \frac{a_j + b_j}{2} \right) = (b_j - a_j)\alpha f_0(1/2) \geq f_0(1/2)\epsilon = C(\alpha)\epsilon.
\]

In other words, \( \nu_\tau, \tau \in \{0,1\}^s \) forms a \( C(\alpha)\epsilon \)-packing subset of \( \mathcal{S}_\alpha \) in the metric \( \rho \). This proves \( N(\epsilon, \mathcal{S}_\alpha) \geq \exp(C(\alpha)s) \geq \exp(C_2\epsilon^{-1/\alpha}) \). The proof of Theorem 2.1 is complete.

Distances between functions in \( \mathcal{S}_\alpha \) can also be measured by the \( L_p \) metric as opposed to the supremum metric (the supremum metric is also known as the \( L_\infty \) metric). The \( L_p \) metric is defined by:

\[
\|f - g\|_p := \left( \int_0^1 |f(x) - g(x)|^p \, dx \right)^{1/p} \quad \text{for } p \geq 1.
\]

Because the \( L_p \) metric is less than or equal to the supremum metric, \( \rho \), it follows from Theorem 2.1 that \( M(\epsilon, \mathcal{S}_\alpha, L_p) \leq \exp(C_1\epsilon^{-1/\alpha}) \). In fact, it is also true that \( N(\epsilon, \mathcal{S}_\alpha, L_p) \geq \exp(C_2\epsilon^{-1/\alpha}) \). This can be seen using the Varshamov-Gilbert lemma together with the lower bound construction in the proof of Theorem 2.1.

Recall that the Varshamov-Gilbert lemma states that there exists a subset \( W \) of \( \{0,1\}^s \) with cardinality, \( |W| \geq \exp(s/8) \) and such that the Hamming distance \( \Delta(\tau, \tau') := \sum_{i=1}^s \{ \tau_i \neq \tau'_i \} > s/4 \) for all \( \tau, \tau' \in W \) with \( \tau \neq \tau' \). Consider the finite class of functions \( \nu_\tau, \tau \in W \) where \( \nu_\tau \) are from the proof of Theorem 2.1. By construction, it is easy to see that for every \( \tau, \tau' \in W \) with \( \tau \neq \tau' \), we get

\[
\|\nu_\tau - \nu_{\tau'}\|_p^p = \Delta(\tau, \tau')\|g_1\|_p^p \geq \frac{s}{4}\|g_1\|_p^p
\]

and

\[
\|g_1\|_p^p = \int_{a_1}^{b_1} (b_1 - a_1)^{p \alpha} f_0\left(\frac{t-a_1}{b_1-a_1}\right) \, dt = (b_1 - a_1)^{p \alpha} \int_0^1 f_0^p(x)(b_1-a_1) \, dx = C(\alpha)(b_1-a_1)^{p \alpha+1}.
\]

Thus for \( \tau, \tau' \in W \) with \( \tau \neq \tau' \),

\[
\|\nu_\tau - \nu_{\tau'}\|_p \geq C(\alpha)s(b_1 - a_1)^{p \alpha+1} \geq C(\alpha)\epsilon^{-1/\alpha} \epsilon^{(p \alpha+1)/\alpha} = C(\alpha)\epsilon^p
\]

which implies that \( \|\nu_\tau - \nu_{\tau'}\|_p \geq C(\alpha)\epsilon \) whenever \( \tau, \tau' \in W \) with \( \tau_i \neq \tau'_i \). This proves that \( N(\epsilon, \mathcal{S}_\alpha, L_p) \geq \exp(C_2\epsilon^{-1/\alpha}) \).

6
2.2 Multidimensions

Once again, $\alpha > 0$ and $\beta$ is the largest integer that is strictly smaller than $\alpha$.

For a vector $p = (p_1, \ldots, p_d)$ consisting of nonnegative integers $p_1, \ldots, p_d$, let $< p > := p_1 + \cdots + p_d$. Let

$$D^p := \partial^{< p >} / \partial x_1^{p_1} \cdots \partial x_d^{p_d}$$

and $h^p = h_1^{p_1} h_2^{p_2} \cdots h_d^{p_d}$.

where $h = (h_1, \ldots, h_d)$. We also use $|h|$ for $\sqrt{h_1^2 + \cdots + h_d^2}$.

The class $S_{\alpha,d}$ is defined to consist of all functions $f$ on $[0, 1]^d$ that satisfy:

1. $f$ is continuous on $[0, 1]^d$.
2. All partial derivatives $D^p$ of $f$ exist on $(0, 1)^d$ for $< p > \leq \beta$.
3. $|D^p f(x)| \leq 1$ for all $p$ with $< p > \leq \beta$ and $x \in [0, 1]^d$.
4. $|D^p f(x) - D^p f(y)| \leq |x - y|^\alpha / \beta$ for all $p$ with $< p > = \beta$ and $x, y \in (0, 1)^d$.

Once again, we consider the supremum metric defined by $\rho(f, g) := \sup_{x \in [0, 1]^d} |f(x) - g(x)|$.

**Theorem 2.2.** There exist positive constants $\epsilon_0, C_1$ and $C_2$ depending only on $\alpha$ and the dimension $d$ such that for all $0 < \epsilon \leq \epsilon_0$, we have

$$M(\epsilon, S_{\alpha,d}) \leq \exp(C_1 \epsilon^{-d/\alpha})$$

and

$$N(\epsilon, S_{\alpha,d}) \geq \exp(C_2 \epsilon^{-d/\alpha})$$

The analogue of Taylor’s formula (2) is now

$$f(x+h) = \sum_{k=0}^\beta \sum_{p:<p>=k} \frac{h^p D^p f(x)}{k!} + R_f(x, h) \quad \text{with } |R_f(x, h)| \leq |h|^\alpha / \beta!$$

and the analogue of (3) would be

$$D^q f(x+h) = \sum_{k=0}^{\beta-<q>} \sum_{p:<p>=k} \frac{h^p D^{p+q} f(x)}{k!} + R(x, h) \quad \text{with } |R(x, h)| \leq |h|^{\alpha-<q>} / \beta!$$

With these, the proof of Theorem 2.2 would proceed just like the proof of Theorem 2.1.

3 Application to Regression

Considering the problem of estimating a function $f \in S_{\alpha}$ from observations:

$$(x_1, y_1), \ldots, (x_n, y_n)$$

where

$$y_i = f(x_i) + \xi_i \quad i = 1, \ldots, n.$$
We assume that \( x_1, \ldots, x_n \) are independent and uniformly distributed on \([0, 1]\). Also, we assume that \( \xi_1, \ldots, \xi_n \) are independent normal random variables with mean zero and variance one. Finally, we assume that \( \xi_1, \ldots, \xi_n \) and \( x_1, \ldots, x_n \) are independent.

For each \( f \in S_\alpha \), let \( P_f \) denote the distribution of the data \( z := ((x_1, y_1), \ldots, (x_n, y_n)) \) when the true function is \( f \). Clearly, \( P_f \) has the density \( p_f \) on \([0, 1]^n \times \mathbb{R}^n\) given by

\[
p_f(z) := \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(y_i - f(x_i))^2}{2} \right).
\]

Let \( L(f, g) \) denote the loss function (for the moment, assume that it is arbitrary). Let \( F \) be a finite subset of \( S_\alpha \) and consider the estimator:

\[
\hat{f} := \arg\min_{f \in F} \sum_{i=1}^n (y_i - f(x_i))^2.
\]

Here is simple technique for bounding the risk of this estimator. For every \( g \in F \), we have

\[
L(f, \hat{f}) = \log \left( e^{L(f, \hat{f})} \right)
\leq \log \left( e^{L(f, \hat{f})} \frac{p_f(z)}{p_g(z)} \right)
\leq \log \left( \sum_{u \in F} e^{L(f, u)} \sqrt{p_u(z) \frac{p_f(z)}{p_g(z)}} \right)
\leq \log \left( \sum_{u \in F} e^{L(f, u)} \frac{p_u(z)}{p_g(z)} \right) + \frac{1}{2} \log \left( \frac{p_f(z)}{p_g(z)} \right)
\]

Taking expectations on both sides with respect to \( z \sim P_f \), we have

\[
E_f L(f, \hat{f}) \leq E_f \log \left( \sum_{u \in F} e^{L(f, u)} \sqrt{p_u(z) \frac{p_f(z)}{p_g(z)}} \right) + \frac{1}{2} D(P_f || P_g).
\]

Because \( \log \) is concave, we can take the expectation on the right hand side inside the log to obtain:

\[
E_f L(f, \hat{f}) \leq \log \left( \sum_{u \in F} e^{L(f, u)} E_f \sqrt{p_u(z) \frac{p_f(z)}{p_g(z)}} \right) + \frac{1}{2} D(P_f || P_g).
\]

Suppose now that we choose the loss function to be

\[
L(f, u) = -\log E_f \sqrt{p_u(z) \frac{p_f(z)}{p_g(z)}}.
\]
We then get
\[ E_f L(f, \hat{f}) \leq \log |F| + \frac{1}{2} D(P_f || P_g). \]
Because \( g \in F \) was arbitrary, we get the risk bound:
\[ E_f L(f, \hat{f}) \leq \log |F| + \frac{1}{2} \min_{g \in F} D(P_f || P_g). \]
This further implies (by taking a supremum over \( f \in \mathcal{S}_\alpha \) on both sides)
\[
\sup_{f \in \mathcal{S}_\alpha} E_f L(f, \hat{f}) \leq \log |F| + \frac{1}{2} \sup_{f \in \mathcal{S}_\alpha} \min_{g \in F} D(P_f || P_g). \tag{10}
\]
In this problem, it is easy to see that
\[
\mathbb{E}_f \sqrt{p_f(z) p_u(z)} = \left( \int_0^1 \exp \left( -\frac{(u(x) - f(x))^2}{4} \right) dx \right)^n
\]
and
\[
D(P_f || P_g) = \frac{n}{2} \int_0^1 (f(x) - g(x))^2 dx.
\]
The loss function (9) therefore equals
\[
L(f, u) = -n \log \left( \int_0^1 \exp \left( -\frac{(u(x) - f(x))^2}{4} \right) dx \right).
\]
The bound (10) can now be written as
\[
\sup_{f \in \mathcal{S}_\alpha} -E_f \log \left( \int_0^1 \exp \left( -\frac{(f(x) - \hat{f}(x))^2}{4} \right) dx \right) \leq \frac{\log |F|}{n} + \frac{1}{4} \inf_{f \in \mathcal{S}_\alpha} \sup_{g \in F} \int_0^1 (f(x) - g(x))^2 dx.
\]
If we now fix \( \epsilon > 0 \) and take \( F \) to be an \( \epsilon \)-cover of \( \mathcal{S}_\alpha \) under the \( L_2 \) metric, we get, by Theorem 2.1, that the right hand side is bounded from above by a positive constant multiple of
\[
\frac{1}{n} \epsilon^{-1/\alpha} + \epsilon^2
\]
By taking \( \epsilon = n^{-\alpha/(2\alpha+1)} \), we get that
\[
\inf_{f} \sup_{f \in \mathcal{S}_\alpha} -E_f \log \left( \int_0^1 \exp \left( -\frac{(f(x) - \hat{f}(x))^2}{4} \right) dx \right) \leq C(\alpha)n^{-2\alpha/(2\alpha+1)}.
\]
Because \( -\log x \geq 1 - x \), we get
\[
\inf_{f} \sup_{f \in \mathcal{S}_\alpha} \mathbb{E}_f \left( 1 - \int_0^1 \exp \left( -\frac{(f(x) - \hat{f}(x))^2}{4} \right) dx \right) \leq C(\alpha)n^{-2\alpha/(2\alpha+1)}. \tag{11}
\]
Now because both the functions \( f \) and \( \hat{f} \) are in \( \mathcal{S}_\alpha \), they are bounded by 1 and thus for each \( x \in [0, 1] \), we have

\[
\frac{(f(x) - \hat{f}(x))^2}{4} \leq 1.
\]

Because on \([0, 1]\), the convex function \( e^{-t} \) lies below the chord joining the points \((0, 1)\) and \((1, 1/e)\), we get

\[
1 - e^{-t} \geq (1 - 1/e)t \quad \text{for } 0 \leq t \leq 1.
\]

Using (11) and the above for \( t = (f(x) - \hat{f}(x))^2/4 \) we get that

\[
R_{\text{minimax}} := \inf_{\hat{f}} \sup_{f \in \mathcal{S}_\alpha} \mathbb{E}_f \int (f(x) - \hat{f}(x))^2 \, dx \leq C(\alpha) n^{-2\alpha/(2\alpha + 1)}.
\]

(12)

Note that the property of the parameter space \( \mathcal{S}_\alpha \) that was used to derive the above risk bound is the covering number bound given by Theorem 2.1.

Using the lower bound:

\[
R_{\text{minimax}} \geq \sup_{\eta, \epsilon} \frac{\eta}{2} \left( 1 - \frac{\log 2 + \log M(\epsilon) + \epsilon^2}{\log N(\eta)} \right)
\]

that we derived previously, one can show that \( R_{\text{minimax}} \geq C(\alpha) n^{-2\alpha/(2\alpha + 1)} \) using only the covering and packing numbers derived in Theorem 2.1.