Book review:
High-Dimensional Statistics (Martin J. Wainwright)

Chapter 2–4: Basic Tail and Concentration Bounds & Uniform Bounds

Suehyun Kim

29 July 2025

Causal Inference Lab.
Seoul National University

Outline

Chapter 2: Basic tail and concentration bounds
- Classical bounds
- Martingale-based methods
- Lipschitz functions of Gaussian variables

Chapter 4: Uniform laws of large numbers
- Motivation
- A uniform law via Rademacher complexity
- Upper bounds on the Rademacher complexity

Outline

From Markov to Chernoff

In many settings, we are interested in whether *a random variable is close to its mean or median.*

- Deviation inequalities (tail bounds): $P(X - \mu \geq t)$
- Concentration inequalities: $P(|X - \mu| \geq t)$

Two classical inequalities form the foundation for bounding tail probabilities:

- **Markov's inequality**: Controlling the moments.
- **Chernoff bound**: Controlling the mgf.

From Markov to Chernoff

Proposition (Markov's inequality)

For a non-negative random variable X with finite mean,

$$P(X \geq t) \leq \frac{\mathbb{E}[X]}{t} \qquad \text{for all } t > 0.$$

- If X has a central moment of order k, the following is a direct corollary:

$$P(|X - \mu| \geq t) \leq \frac{\mathbb{E}[|X - \mu|^k]}{t^k} \quad \text{for all } t > 0. \tag{1}$$

- Setting $k = 2$ yields the well-known *Chebyshev's inequality*.

From Markov to Chernoff

Proposition (Chernoff bound)

For a random variable X with an mgf $\phi(\lambda) = \mathbb{E}[e^{\lambda(X-\mu)}]$ defined on $|\lambda| \leq b$,

$$P(X - \mu \geq t) \leq \inf_{\lambda \in [0,b]} e^{-\lambda t}\, \mathbb{E}[e^{\lambda(X-\mu)}]. \tag{2}$$

Proof. From Markov's inequality,

$$P(X - \mu \geq t) = P(e^{\lambda(X-\mu)} \geq e^{\lambda t}) \leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}}.$$

Optimize the choice of λ to obtain the tightest result. *Remark.* The moment bound (1) with an optimal choice of k is never worse than the Chernoff bound (2).

Tail behavior of random variables

Consequently, it is natural to classify random variables in terms of their mgfs, or more intuitively, *the lightness of their tails*.

(i) **Sub-Gaussian variables**

- Eventually dominated by a Gaussian variable.
- The exponent term in the tail probabilities scales *quadratically*.

(ii) **Sub-exponential variables**

- Tails are heavier than sub-Gaussians - eventually the exponent term in the tail probabilities scales *linearly*.
- A variable is sub-exponential if and only if its mgf exists in a neighborhood around zero.

Tail behavior of random variables

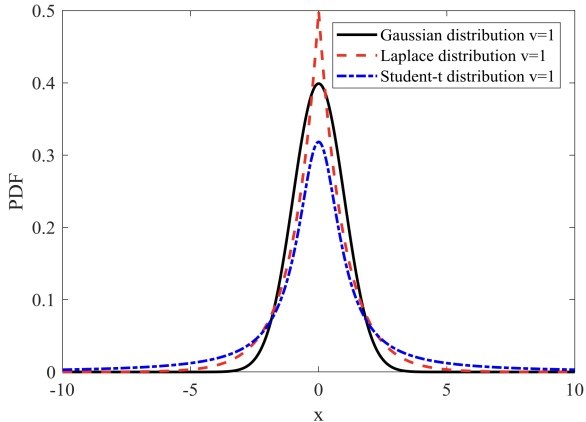

Figure reference: Chu et al. (2020), A High-Resolution and Low-Frequency Acoustic Beamforming Based on Bayesian Inference and Non-Synchronous Measurements.

Sub-Gaussian variables and Hoeffding bounds

Definition 2.2 (Sub-Gaussianity)

A random variable X with mean $\mu = \mathbb{E}[X]$ is *sub-Gaussian* if there is a positive number σ such that

$$\mathbb{E}[e^{\lambda(X-\mu}] \leq e^{\sigma^2\lambda^2/2} \quad \text{for all } \lambda \in \mathbb{R}.$$

- The constant σ is referred to as the *sub-Gaussian parameter*.
- Any Gaussian variable with variance σ^2 is sub-Gaussian with parameter σ.

Sub-Gaussian variables and Hoeffding bounds

The following proposition characterizes the tail behavior of sub-Gaussians:

Proposition (Sub-Gaussian tail bound)

If X is sub-Gaussian with parameter σ, it satisfies the *upper deviation inequality*

$$P(X - \mu \geq t) \leq e^{-\frac{t^2}{2\sigma^2}} \quad \text{for all } t \geq 0.$$

Proof. Applying the Chernoff bound, $P(X - \mu \geq t) \leq \inf_{\lambda \in \mathbb{R}} e^{-\lambda t}\, \mathbb{E}[e^{\lambda(X-\mu)}] = e^{-\frac{t^2}{2\sigma^2}}$.

Since X is sub-Gaussian if and only if $-X$ is sub-Gaussian, any sub-Gaussian variable satisfies the *concentration inequality*

$$P(|X - \mu| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}} \quad \text{for all } t \in \mathbb{R}.$$

Sub-Gaussian variables and Hoeffding bounds

Example 2.3 (Rademacher variables)

A Rademacher random variable ε takes the values $\{-1, +1\}$ equiprobably. We claim that it is sub-Gaussian with parameter $\sigma = 1$.

Using the power-series expansion for the exponential, we obtain

$$\begin{aligned}
\mathbb{E}[e^{\lambda\varepsilon}] = \frac{1}{2}(e^{\lambda} + e^{-\lambda}) &= \frac{1}{2}\left\{\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!} + \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\right\} \\
&= \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!} \\
&\leq 1 + \sum_{k=1}^{\infty}\frac{\lambda^{2k}}{2^k k!} \\
&= e^{\lambda^2/2}.
\end{aligned}$$

Sub-Gaussian variables and Hoeffding bounds

Example 2.4 (Bounded random variables)

Let X be zero-mean, and supported on some interval $[a, b]$. Letting X' be an independent copy, for any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}_X[e^{\lambda X}] = \mathbb{E}_X[e^{\lambda(X - \mathbb{E}_{X'}[X'])}] \leq \mathbb{E}_{X,X'}[e^{\lambda(X - X')}],$$

where the last inequality follows from the convexity of the exponential.

Letting ε be an independent Rademacher variable, note that the distribution of $X - X'$ is the same as that of $\varepsilon(X - X')$, so that we have

$$\mathbb{E}_{X,X'}[e^{\lambda(X - X')}] = \mathbb{E}_{X,X'}[\mathbb{E}_\varepsilon[e^{\lambda\varepsilon(X - X')}]] \leq \mathbb{E}_{X,X'}[e^{\frac{\lambda^2(X - X')^2}{2}}],$$

with the inequality following from the previous example.

Sub-Gaussian variables and Hoeffding bounds

Example 2.4 (Bounded random variables; continued)

Since $|X - X'| \leq b - a$, we are guaranteed that

$$\mathbb{E}_{X,X'}[e^{\frac{\lambda^2(X-X')^2}{2}}] \leq e^{\frac{\lambda^2(b-a)^2}{2}},$$

and hence X is sub-Gaussian with parameter at most $\sigma = b - a$.

Remarks.

- The proof technique is a simple example of a *symmetrization argument* - we introduce an independent copy X', and symmetrize the problem using an independent Rademacher variable.
- The sub-Gaussian parameter can be further sharpened to $\sigma = \frac{b-a}{2}$.

Sub-Gaussian variables and Hoeffding bounds

Just as the property of Gaussianity is preserved by linear operations, so is the property of sub-Gaussianity.

- If X_1 and X_2 are independent sub-Gaussians with parameters σ_1 and σ_2, then $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\sigma_1 + \sigma_2}$.
- Consequently, we obtain an important result, known as the *Hoeffding bound*.

Proposition 2.5 (Hoeffding bound)

Suppose that the variables X_i, $i = 1, \ldots, n$, are independent, and X_i has mean μ_i and sub-Gaussian parameter σ_i. Then, for all $t \geq 0$, we have

$$P\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right).$$

Sub-Gaussian variables and Hoeffding bounds

Remark.

The Hoeffding bound is often stated only for the special case of *bounded random variables*. If $X_i \in [a, b]$ for all $i = 1, \dots, n$, it is sub-Gaussian with parameter $\sigma = \frac{b-a}{n}$, so that we obtain the bound

$$P\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{2t^2}{n(b-a)^2}\right).$$

We conclude the discussion on sub-Gaussianity with equivalent characterizations of sub-Gaussian variables.

Sub-Gaussian variables and Hoeffding bounds

Theorem 2.6 (Equivalent characterizations of sub-Gaussian variables)

Given any zero-mean random variable X, the following properties are equivalent:

(i) There is a constant $\sigma \geq 0$ such that

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2 \sigma^2}{2}} \quad \text{for all } \lambda \in \mathbb{R}.$$

(ii) There is a constant $c \geq 0$ and Gaussian variable $Z \sim N(0, \tau^2)$ such that

$$P(|X| \geq s) \leq c\, P(|Z| \geq s) \quad \text{for all } s \geq 0.$$

(iii) There is a constant $\theta \geq 0$ such that

$$\mathbb{E}[X^{2k}] \leq \frac{(2k)!}{2^k k!}\, \theta^{2k} \quad \text{for all } k = 1, 2, \ldots .$$

(iv) There is a constant $\sigma \geq 0$ such that

$$\mathbb{E}\left[e^{\frac{\lambda X^2}{2\sigma^2}}\right] \leq \frac{1}{\sqrt{1-\lambda}} \quad \text{for all } \lambda \in [0, 1).$$

Sub-exponential variables and Bernstein bounds

Sometimes, the notion of sub-Gaussianity is too restrictive; sub-exponentiality offers a more relaxed condition.

Definition 2.7 (Sub-exponentiality)

A random variable X with mean $\mu - \mathbb{E}[X]$ is *sub-exponential* if there are non-negative parameters (ν, α) such that

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\nu^2\lambda^2}{2}} \quad \text{for all } |\lambda| < \frac{1}{\alpha}.$$

- Here, we have two separate parameters, ν (corresponding to the variability) and α (corresponding to the range).
- Any sub-Gaussian variable is sub-exponential, with parameters $(\nu, \alpha) = (\sigma, 0)$.

Sub-exponential variables and Bernstein bounds

Example 2.8 (Sub-exponential but not sub-Gaussian)

Let $Z \sim N(0,1)$, and consider the random variable $X = Z^2$.

- Since $\mathbb{E}[e^{\lambda(X-1)}] = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}$, the mgf of X is infinite for $\lambda > \frac{1}{2}$, and thus X is not sub-Gaussian.
- However, X is sub-exponential with parameters $(\nu, \alpha) = (2, 4)$.
  For $|\lambda| < 1/4$, we have

$$\begin{aligned}\mathbb{E}[e^{\lambda(X-1)}] &= \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \\ &\leq e^{2\lambda^2}.\end{aligned}$$

Sub-exponential variables and Bernstein bounds

Proposition 2.9 (Sub-exponential tail bound)

Suppose that X is sub-exponential with parameters (ν, α). Then,

$$P(X - \mu \geq t) \leq \begin{cases} e^{-\frac{t^2}{2\nu^2}} & \text{if } 0 \leq t \leq \frac{\nu^2}{\alpha}, \\ e^{-\frac{t}{2\alpha}} & \text{for } t > \frac{\nu^2}{\alpha}. \end{cases}$$

- The behavior of X is sub-Gaussian toward the center, however the tail eventually decays like a exponential variable.
- As with the Hoeffding inequality, a concentration inequality bounding $P(|X - \mu| \geq t)$ can be derived with an additional factor of 2.

Sub-exponential variables and Bernstein bounds

Proof of Prop 2.9.
WLOG assume $\mu = 0$, and apply the Chernoff bound:

$$P(X \geq t) \leq \inf_{\lambda} e^{-\lambda t}\, \mathbb{E}[e^{\lambda X}] \leq \exp\left(-\lambda t + \frac{\lambda^2 \nu^2}{2}\right), \quad \text{for all } \lambda \in [0, \alpha^{-1}).$$

Now, let $g(\lambda, t) = -\lambda t + \frac{\lambda^2 \nu^2}{2}$, and consider the quantity $g^*(t) = \inf_{\lambda \in [0, \alpha^{-1})} g(\lambda, t)$. Note that, for each fixed $t \geq 0$, $g(\lambda, t)$ is a quadratic function of λ that attains its minimum $-\frac{t^2}{2\nu^2}$ at $\lambda^* = \frac{t}{\nu^2}$.

If $\lambda^* \leq \alpha^{-1}$, we obtain the global minimum $g^*(t) = -\frac{t^2}{2\nu^2}$ on $0 \leq t \leq \frac{\nu^2}{\alpha}$. Otherwise, the constrained minimum is achieved at the boundary $\lambda^\dagger = \alpha^{-1}$, giving the bound

$$g^*(t) = g(\lambda^\dagger, t) = -\frac{t}{\alpha} + \frac{1}{2\alpha}\frac{\nu^2}{\alpha} \leq -\frac{t}{2\alpha}.$$

Sub-exponential variables and Bernstein bounds

Direct calculation of the mgf may be impractical in many settings. *Bernstein's condition* offers a sufficient criterion for sub-exponentiality by controlling the moments.

Definition (Bernstein's condition)

For a random variable X with mean μ and variance σ^2, *Bernstein's condition with parameter b* is satisfied if

$$|\mathbb{E}[(X-\mu)^k]| \leq \frac{1}{2} k!\, \sigma^2\, b^{k-2} \quad \text{for } k = 2, 3, \ldots . \tag{3}$$

Sub-exponential variables and Bernstein bounds

When X satisfies the Bernstein condition, then it is sub-exponential with parameters determined by σ^2 and b:

Proposition 2.10 (Bernstein-type bound)

For any variable satisfying the Bernstein condition (3), we have

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}} \quad \text{for all } |\lambda| < \frac{1}{b},$$

and, moreover, the concentration inequality

$$P(|X-\mu| \geq t) \leq 2e^{-\frac{t^2}{2(\sigma^2+bt)}} \quad \text{for all } t \geq 0.$$

Sub-exponential variables and Bernstein bounds

Proof of Prop 2.10.
By the power series expansion of the exponential, we have

$$\begin{aligned}\mathbb{E}[e^{\lambda(X-\mu)}] &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbb{E}[(X-\mu)^k]}{k!} \\ &\leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2} \sum_{k=3}^{\infty} (|\lambda| b)^{k-2}.\end{aligned}$$

Then, for any $|\lambda| < 1/b$, we obtain

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq 1 + \frac{\lambda^2\sigma^2/2}{1-b|\lambda|} \leq e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}},$$

using the summation of the geometric series and the bound $1 + t \leq e^t$.

Moreover, it follows that X is $(\sqrt{2}\sigma, 2b)$-sub-exponential, since $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\lambda^2(\sqrt{2}\sigma)^2}{2}}$.

For the concentration inequality, set $\lambda = \frac{t}{bt+\sigma^2} \in [0, 1/b)$ in the Chernoff bound.

Sub-exponential variables and Bernstein bounds

Remark 1.

- Like the sub-Gaussian property, the sub-exponential property is *preserved under summation for independent random variables*.
- If $\{X_k\}_{k=1}^n$ is an independent sequence of random variables with mean μ_k and sub-exponential parameters (ν_k, α_k), we can bound the of $\sum_{k=1}^n (X_k - \mu_k)$ as

$$\mathbb{E}[e^{\lambda \sum_{k=1}^n (X_k - \mu_k)}] = \prod_{k=1}^n \mathbb{E}[e^{\lambda (X_k - \mu_k}] \leq \prod_{k=1}^n e^{\frac{\lambda^2 \nu_k^2}{2}},$$

valid for all $|\lambda| < (\max_{k=1,\ldots,n} \alpha_k)^{-1}$.

Sub-exponential variables and Bernstein bounds

Remark 1 (continued).

- Consequently, $\sum_{k=1}^{n}(X_k - \mu_k)$ is (ν_*, α_*)-sub-exponential, with

$$\nu_* = \sqrt{\sum_{k=1}^{n} \nu_k^2}, \qquad \alpha_* = \max_{k=1,\dots,n} \alpha_k.$$

- This observation leads to the tail bound

$$P\left(\frac{1}{n}\sum_{k=1}^{n}(X_k - \mu_k) \geq t\right) \leq \begin{cases} e^{-\frac{nt^2}{2(\nu_*^2/n)}} & \text{if } 0 \leq t \leq \frac{\nu_*^2}{n\alpha_*}, \\ e^{-\frac{nt}{2\alpha_*}} & \text{for } t > \frac{\nu_*^2}{n\alpha_*}. \end{cases}$$

Sub-exponential variables and Bernstein bounds

Remark 2.

- Prop 2.10 has an important consequence even for *bounded random variables*.
- Suppose $|X - \mu| < b$. Then X satisfies the Bernstein condition with parameter $c = b/3$, since

$$\mathbb{E}[|X - \mu|^k] \leq \sigma^2 b^{k-2} = \sigma^2 3^{k-2} c^{k-2} \leq \frac{k!}{2} \sigma^2 c^{k-2},$$

  using the inequality $3^{k-2} \leq \frac{k!}{2}$ for all $k \geq 2$.
- Since the Bernstein bound involves both the variance σ^2 and the bound b, it is substantially better than the sub-Gaussian bound when $\sigma^2 \ll b^2$.

Sub-exponential variables and Bernstein bounds

Remark 2 (continued).

- Bernstein bound is also often stated for the bounded case where X_i are independent mean-zero variables with $|X_i| \leq b$ for all $i = 1, \ldots, n$. Letting $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2$, we have

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(\frac{nt^2}{2\sigma^2 + 2ct/3}\right) \quad \text{for all } t \geq 0.$$

- In general, for bounded random variables, *Bennett's inequality* can be used to provide sharper control on the tails.

Sub-exponential variables and Bernstein bounds

Remark 3. If a variable is known to be bounded only from above, it is still possible to derive *one-sided Bernstein-type bounds*:

Proposition 2.14 (One-sided Bernstein's inequality)

If $X \leq b$ almost surely, then

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq \exp\left(\frac{\frac{\lambda^2}{2}\mathbb{E}[X^2]}{1-\frac{b\lambda}{3}}\right) \quad \text{for all } \lambda \in [0, 3/b).$$

Consequently, given n independent random variables such that $X_i \leq b$ almost surely, we have

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq \delta\right) \leq \exp\left(-\frac{n\delta^2}{2\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] + \frac{b\delta}{3}\right)}\right).$$

Sub-exponential variables and Bernstein bounds

Example 2.11 (χ^2-variables)

Let $Y := \sum_{k=1}^{n} Z_k^2$ be a χ^2-distributed with n degrees of freedom, where $Z_k \overset{i.i.d.}{\sim} N(0,1)$.

- We have shown in Example 2.8 that Z_k^2 is $(2,4)$-sub-exponential.
- Consequently, Y is $(2\sqrt{n},4)$-sub-exponential, and thus we obtain the two-sided tail bound

$$P\left(\left|\frac{1}{n}\sum_{k=1}^{n} Z_k^2 - 1\right| \geq t\right) \leq 2e^{-nt^2/8} \quad \text{for all } t \in (0,1).$$

The concentration of χ^2-variables plays an important role in analyzing procedures based on random projections, as illustrated in the next example.

Sub-exponential variables and Bernstein bounds

Example 2.12 (Johnson-Lindenstrauss embedding)

Suppose that we are given $N \geq 2$ distinct vectors $\{u^1, \ldots, u^N\}$, with each vector lying in $\mathbb{R}^d$.

If the data dimension d is large, it might be expensive to store and manipulate the original dataset. Thus, one might be interested in projecting the vectors onto a space of lower dimension.

- We aim to achieve *dimensionality reduction* by constructing a mapping $F : \mathbb{R}^d \to \mathbb{R}^m$ with $m \ll d$, while preserving some key features.
- The *Johnson-Lindenstrauss embedding* preserves *pairwise distances* with a multiplicative tolerance $\delta \in (0, 1)$, so that

$$(1 - \delta) \leq \frac{\|F(u^i) - F(u^j)\|_2^2}{\|u^i - u^j\|_2^2} \leq (1 + \delta) \quad \text{for all pairs } u^i \neq u^j. \tag{4}$$

Sub-exponential variables and Bernstein bounds

Example 2.12 (Johnson-Lindenstrauss embedding; continued)

We can construct such a mapping F as follows:

- First, form a random matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$ filled with independent $N(0,1)$ entries.
- Then, define a linear mapping $F : \mathbb{R}^d \to \mathbb{R}^m$ via $u \mapsto \mathbf{X}u/\sqrt{m}$.

We now verify that F satisfies the bound (4) with high probability.

- Let x_i denote the i-th row of $\mathbf{X}$, and consider some fixed $u \neq 0$.
- Since x_i is a standard normal vector, the variable $\langle x_i, u/\|u\|_2 \rangle$ is also standard Gaussian.
- Hence, the quantity

$$Y := \frac{\|\mathbf{X}u\|_2^2}{\|u\|_2^2} = \sum_{i=1}^{m} \langle x_i, u/\|u\|_2 \rangle$$

  follows a χ^2-distribution with m degrees of freedom, due to the independence of rows.

Sub-exponential variables and Bernstein bounds

Example 2.12 (Johnson-Lindenstrauss embedding; continued)

- Therefore, applying the result from the previous example, we obtain

$$P\left(\left|\frac{\|\mathbf{X}u\|_2^2}{m\|u\|_2^2} - 1\right| \geq \delta\right) \leq 2e^{-m\delta^2/8} \quad \text{for all } \delta \in (0,1).$$

- Rearranging and recalling the definition of F yields the bound

$$P\left(\frac{\|F(u)\|_2^2}{\|u\|_2^2} \notin [1-\delta, 1+\delta]\right) \leq 2e^{-m\delta^2/8} \quad \text{for any fixed } 0 \neq u \in \mathbb{R}^d.$$

Sub-exponential variables and Bernstein bounds

Example 2.12 (Johnson-Lindenstrauss embedding; continued)

- Applying the union bound with $\binom{N}{2}$ distinct data points, we conclude that

$$P\left(\frac{\|F(u^i - u^j)\|_2^2}{\|u^i - u^j\|_2^2} \notin [1-\delta, 1+\delta] \quad \text{for some } u^i \neq u^j\right) \leq 2\binom{N}{2} e^{-m\delta^2/8}.$$

- For any $\epsilon \in (0,1)$, this probability can be driven below ϵ by choosing $m < \frac{16}{\delta^2}\log(N/\epsilon)$. Note that this quantity does not depend on the original dimension d, and scales only logarithmically with the number of data points N.

Sub-exponential variables and Bernstein bounds

Theorem 2.13 (Equivalent characterizations of sub-exponential variables)

For a zero-mean random variable X, the following statements are equivalent:

(i) There are non-negative numbers (ν, α) such that

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\nu^2\lambda^2}{2}} \quad \text{for all } |\lambda| < \frac{1}{\alpha}.$$

(ii) There is a positive number $c_0 > 0$ such that $\mathbb{E}[e^{\lambda X} < \infty$ for all $|\lambda| \leq c_0$.

(iii) There are constants $c_1, c_2 > 0$ such that

$$P(|X| \geq t) \leq c_1 e^{-c_2 t} \quad \text{for all } t > 0.$$

(iv) The quantity $\gamma := \sup_{k \geq 2} \left[\frac{\mathbb{E}[X^k]}{k!}\right]^{1/k}$ is finite.

Outline

Background on martingales

- So far, we have covered two elementary bounds—namely, the *Hoeffding and Bernstein bounds*—which provide useful results for sub-Gaussian, sub-exponential, or bounded variables.
- Often, we are also interested in the behavior of $f(X) - \mathbb{E}[f(X)]$. When the function f satisfies a certain condition, called the *bounded difference property*, we can derive a Hoeffding-like bound.
- Such a bound is obtained via a telescoping decomposition

$$f(X) - \mathbb{E}[f(X)] = Y_n - Y_0 = \sum_{k=1}^{n} (Y_k - Y_{k-1}),$$

where the sequence $Y_k = \mathbb{E}[f(X) \mid X_1, \ldots, X_k]$ forms a specific type of martingale known as the *Doob martingale*. Here, we aim to prove the bounded difference inequality using bounds on martingale difference sequences.

Background on martingales

Given a probability space $(\Omega, \mathcal{F}, P)$, a nested sequence of sub σ-fields of $\mathcal{F}$ is called a *filtration*, meaning $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ for all $k \geq 1$. A sequence of random variables $\{Y_k\}_{k=1}^{\infty}$ is *adapted* to the filtration $\{\mathcal{F}_k\}_{k=1}^{\infty}$ if each Y_k is $\mathcal{F}_k$-measurable.

Definition 2.15 (Martingale)

Given a sequence $\{Y_k\}_{k=1}^{\infty}$ of random variables adapted to a filtration $\{\mathcal{F}_k\}_{k=1}^{\infty}$, the pair $\{(Y_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ is a *martingale* if, for all $k \geq 1$,

$$\mathbb{E}[|Y_k|] < \infty \quad \text{and } \mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] = Y_k.$$

Roughly speaking, if a variable is a martingale, *the best prediction of tomorrow is given by today's value*.

Background on martingales

Remarks.

- If the filtration is defined by another sequence of r.v.s $\{X_k\}_{k=1}^{\infty}$ via the canonical σ-fields $\mathscr{F}_k := \sigma(X_1, \dots, X_k)$, we say that $\{Y_k\}_{k=1}^{\infty}$ is a martingale w.r.t. $\{X_k\}_{k=1}^{\infty}$.
- If a sequence is martingale with respect to itself (i.e. with $\mathscr{F}_k = \sigma(Y_1, \dots, Y_k)$, we simply say that $\{Y_k\}_{k=1}^{\infty}$ forms a martingale sequence.
- In general, the notion of martingale can be defined for stochastic processes as follows. For a filtered probability space $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t\geq 0}, P)$, an adapted process $\{(X_t, \mathscr{F}_t)\}_{t\geq 0}$ is a martingale if

$$\mathbb{E}[|X_t|] < \infty \quad \text{and } \mathbb{E}[X_t \mid \mathscr{F}_s] = X_s \quad \text{for all } s < t.$$

Background on martingales

Another useful notion is that of martingale difference sequences.

Definition (Martingale difference sequence)

An adapted sequence $\{(D_k, \mathscr{F}_k)\}_{k=1}^{\infty}$ is called a *martingale difference sequence* if, for all $k \geq 1$,

$$\mathbb{E}[|D_k|] < \infty \quad \text{and } \mathbb{E}[D_{k+1} \mid \mathscr{F}_k] = 0.$$

- As suggested by their name, such difference sequences arise in a natural way from martingales by defining $D_k = Y_k - Y_{k-1}$.
- For any martingale sequence $\{Y_k\}_{k=0}^{n}$, we have the telescoping decompositon $Y_n - Y_0 = \sum_{k=1}^{n} D_k$, where $\{D_k\}_{k=1}^{n}$ is a martingale difference sequence.

Background on martingales

Example 2.17 (Doob construction)

Given a sequence of independent random variables $\{X_k\}_{k=1}^n$, define the sequence $Y_k := \mathbb{E}[f(X) \mid X_1, \dots X_k]$, and suppose that $\mathbb{E}[|f(X)|] < \infty$. We claim that $\{Y_k\}_{k=0}^n$ is a martingale w.r.t. $\{X_k\}_{k=1}^n$.

- Writing $X_1^k = (X_1, \dots, X_k)$, we indeed have

$$\mathbb{E}[|Y_k|] = \mathbb{E}[|\mathbb{E}[f(X) \mid X_1^k]|] \leq \mathbb{E}[|f(X)|] < \infty,$$

  due to Jensen's inequality.

- Moreover, by the tower property,

$$\mathbb{E}[Y_{k+1} \mid X_1^k] = \mathbb{E}[\mathbb{E}[f(X) \mid X_1^{k+1}] \mid X_1^k] = \mathbb{E}[f(X) \mid X_1^k] = Y_k,$$

  and the second condition is also satisfied.

- Note that $D_k := Y_k - Y_{k-1}$ is a martingale difference sequence.

Concentration bounds for martingale difference sequences

We begin by stating and proving a general *Bernstein-type bound for a martingale difference sequence*, which can be used to bound the quantity $Y_n - Y_0$, or the sum $\sum_{k=1}^n D_k$ itself.

Theorem 2.19

Let $\{(D_k, \mathscr{F}_k)\}_{k=1}^\infty$ be a martingale difference sequence, and suppose that $\mathbb{E}[e^{\lambda D_k} \mid \mathscr{F}_{k-1}] \leq e^{\lambda^2 \nu_k^2/2}$ almost surely for any $|\lambda| < 1/\alpha_k$.Then the following hold:

(a) The sum $\sum_{k=1}^n D_k$ is sub-exponential with parameters $\left(\sqrt{\sum_{k=1}^n \nu_k^2}, \alpha_*\right)$, where $\alpha_* = \max_{k=1,\ldots,n} \alpha_k$.

(b) The sum satisfies the concentration inequality

$$P\left(\left|\sum_{k=1}^n D_k\right| \geq t\right) \leq \begin{cases} 2e^{-\frac{t^2}{2\sum_{k=1}^n \nu_k^2}} & \text{if } 0 \leq t \leq \frac{\sum_{k=1}^n \nu_k^2}{\alpha_*}, \\ 2e^{-\frac{t}{2\alpha_*}} & \text{if } t > \frac{\sum_{k=1}^n \nu_k^2}{\alpha_*}. \end{cases}$$

Concentration bounds for martingale difference sequences

Proof of Thm 2.19.

For any scalar λ such that $|\lambda| < \frac{1}{\alpha_*}$, conditioning on $\mathscr{F}_{n-1}$ and applying iterated expectation yields

$$\begin{aligned}\mathbb{E}[e^{\lambda \sum_{k=1}^{n} D_k}] &= \mathbb{E}[e^{\lambda \sum_{k=1}^{n-1} D_k} \mathbb{E}[e^{\lambda D_n} \mid \mathscr{F}_{n-1}]] \\ &\leq \mathbb{E}[e^{\lambda \sum_{k=1}^{n-1} D_k}] e^{\lambda^2 v_n^2/2},\end{aligned}$$

where the inequality follows from our assumption.

Iterating this procedure yields the bound $\mathbb{E}[e^{\lambda \sum_{k=1}^{n} D_k}] \leq e^{\lambda^2 \sum_{k=1}^{n} v_k^2/2}$, valid for all $|\lambda| < \alpha_*^{-1}$. The tail bound follows from applying the Bernstein bound from Prop 2.9. □

Concentration bounds for martingale difference sequences

Again, we need to isolate sufficient and easily checkable conditions for the differences D_k to be sub-exponential. Since bounded random variables are sub-Gaussian, we obtain the following corollary:

Corollary 2.20 (Azuma-Hoeffding)

Let $\{(D_k, \mathscr{F}_k)\}_{k=1}^{\infty}$ be a martingale difference sequence for which there are constants $\{(a_k, b_k)\}_{k=1}^{n}$ such that $D_k \in [a_k, b_k]$ almost surely for $k = 1, \dots, n$. Then, for all $t \geq 0$,

$$P\left(\left|\sum_{k=1}^{n} D_k\right| \geq t\right) \leq 2e^{-\frac{2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}}.$$

Proof Since $D_k \in [a_k, b_k]$ a.s., the $(D_k \mid \mathscr{F}_{k-1}) \in [a_k, b_k]$ a.s.; use a similar argument with the sub-Gaussian parameter $(b_k - a_k)/2$ and the Hoeffding bound.

Concentration bounds for martingale difference sequences

A key application of Cor 2.20 concerns functions with the following property:

Definition (Bounded difference property)

Given vectors $x, x' \in \mathbb{R}^n$ and an index $k \in \{1, \dots, n\}$, define a new vector $x^{\setminus k} \in \mathbb{R}^n$ via

$$x_j^{\setminus k} := \begin{cases} x_j & \text{if } j \neq k, \\ x'_k & \text{if } j = k. \end{cases}$$

We say that $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the *bounded difference property* with parameters $(L_1, \dots, L_n)$ if, for all $k = 1, \dots, n$,

$$|f(x) - f(x^{\setminus k})| \leq L_k \quad \text{for all } x, x' \in \mathbb{R}^n.$$

Concentration bounds for martingale difference sequences

Corollay 2.21 (Bounded difference inequality; *McDiarmid's inequality*)

Suppose that f satisfies the bounded difference property with parameters $(L_1, \dots, L_n)$ and the random vector $X = (X_1, \dots, X_n)$ has independent compoments. Then,

$$P(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq 2e^{-\frac{2t^2}{\sum_{k=1}^n L_k^2}} \quad \text{for all } t \geq 0.$$

Remark. In the special case when f is L-Lipschitz w.r.t. the Hamming norm defined via the metric $d_H(x, y) = \sum_{i=1}^n \mathbf{1}(x_i \neq y_i)$ for $x, y \in \mathbb{R}^n$, we obtain

$$P(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq 2e^{-\frac{2t^2}{nL^2}} \quad \text{for all } t \geq 0.$$

Concentration bounds for martingale difference sequences

Proof of Cor 2.21.
Recall the Doob martingale and its associated martingale difference sequence

$$D_k = \mathbb{E}[f(X) \mid X_1, \dots, X_k] - \mathbb{E}[f(X) \mid X_1, \dots, X_{k-1}].$$

We claim that D_k lies in an interval of length at most L_k almost surely. Define the random variables

$$A_k := \inf_x \mathbb{E}[f(X) \mid X_1, \dots, X_{k-1}, x] - \mathbb{E}[f(X) \mid X_1, \dots, X_{k-1}],$$
$$B_k := \sup_x \mathbb{E}[f(X)|X_1, \dots, X_{k-1}, x] - \mathbb{E}[f(X) \mid X_1, \dots, X_{k-1}].$$

By definition, $D_k - A_k = \mathbb{E}[f(X) \mid X_1, \dots, X_k] - \inf_x \mathbb{E}[f(X) \mid X_1, \dots, X_{k-1}, x]$, so $D_k \geq A_k$ a.s., and similarly, $D_k \leq B_k$ a.s..

Concentration bounds for martingale difference sequences

Proof of Cor 2.21. (continued)
Observe that by the independence of X_k, we have

$$\mathbb{E}[f(X) \mid x_1, \dots, x_k] = \mathbb{E}_{k+1}[f(x_1, \dots, x_k, X_{k+1}, \dots, X_n)] \quad \text{for all } (x_1, \dots, x_k),$$

where $\mathbb{E}_{k+1}$ denotes the expectation over $(X_{k+1}, \dots, X_n)$. Consequently, we have

$$\begin{aligned} B_k - A_k &= \sup_x \mathbb{E}_{k+1}[f(X_1, \dots, X_{k-1}, x, X_{k+1}, \dots, X_n)] - \inf_x \mathbb{E}_{k+1}[f(X_1, \dots, X_{k-1}, x, X_{k+1}, \dots, X_n)] \\ &\le sup_{x,y} |\mathbb{E}_{k+1}[f(X_1, \dots, X_{k-1}, x, X_{k+1}, \dots, X_n)] - \mathbb{E}_{k+1}[f(X_1, \dots, X_{k-1}, y, X_{k+1}, \dots, X_n)]| \\ &\le L_k, \end{aligned}$$

so we obtain the desired result as a corollary of the Azuma-Hoeffding inequality.

Concentration bounds for martingale difference sequences

Example 2.23 (U-statistics)

Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a symmetric function of its arguments. Given an i.i.d. sequence X_k, $k \geq 1$, of random variables, the quantity

$$U := \frac{1}{\binom{n}{2}} \sum_{j<k} g(X_j, X_k)$$

is known as a pairwise *U-statistic*.

- For instance, if $g(s,t) = |s-t|$, then U is an unbiased estimator of the mean absolute pairwise deviation $\mathbb{E}[|X_1 - X_2|]$.
- Other examples of the U-statistic include the Mann-Whitney U-statistic and Kendall's tau.
- Note that U is *not* a sum of independent r.v.s, although the dependence is relatively weak.

Concentration bounds for martingale difference sequences

Example 2.23 (U-statistics; continued)

- Suppose g is b-uniformly bounded, so that $\|b\|_\infty < \infty$.
- Viewing U as a function $f(x) = f(x_1, \dots, x_n)$, for any given coordinate k, we have

$$\begin{aligned} |f(x) - f(x^{\setminus k})| &\le \frac{1}{\binom{n}{2}} \sum_{j \neq k} |g(x_j, x_k) - g(x_j, x_k')| \\ &\le \frac{(n-1)(2b)}{\binom{n}{2}} = \frac{4b}{n}, \end{aligned}$$

  so that the bounded differences property holds with $L_k = \frac{4b}{n}$ in each coordinate.
- Therefore, by Cor 2.21, we obtain

$$P(|U - \mathbb{E}[U]| \ge t) \le 2e^{-\frac{nt^2}{8b^2}}.$$

Outline

Lipschitz functions of Gaussian variables

Another useful bound can be derived for *Lipschitz functions of Gaussian random variables*.

Definition (L-Lipschitz function)

We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to the Euclidean norm $\|\cdot\|_2$ if

$$|f(x) - f(y)| \leq L\|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^n.$$

Lipschitz functions of Gaussian variables

The following result guarantees that any L-Lipschitz function of i.i.d. standard Gaussian variables is sub-Gaussian with parameter at most L, regardless of the dimension n.

Theorem 2.26

Let $(X_1, \ldots, X_n)$ be a vector of i.i.d. standard Gaussian variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz w.r.t. the Euclidean norm. Then the variable $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most L, and hence

$$P(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq 2e^{-\frac{t^2}{2L^2}} \quad \text{for all } t \geq 0.$$

Lipschitz functions of Gaussian variables

Remarks.

- Roughly speaking, $f(X)$ behaves like a scalar Gaussian variable with variance L^2.
- While we omit the proof of Thm 2.26, it relies on properties specific to the standard Gaussian distribution.
- Hence, the dimension-free concentration of Lipschitz functions *need not hold for sub-Gaussian distributions in general* without additional assumptions such as convexity.
- However, similar concentration results do hold for other non-Gaussian distributions, including the uniform distribution on the sphere and strictly log-concave distributions.

Lipschitz functions of Gaussian variables

Thm 2.26 is useful for a broad range of problems. Although we do not know the explicit value of $\mathbb{E}[f(X)]$, we can still obtain some strong concentration bounds:

Example 2.29 (Order statistics)

Given a random vector $(X_1, \dots, X_n)$, consider the *order statistics* $X_{(1)} \le \cdots \le X_{(n)}$.

- It can be shown that $|X_{(k)} - Y_{(k)}| \le \|X - Y\|_\infty \le \|X - Y\|_2$ for all $k = 1, \dots, n$.
- Consequently, when $X_i \overset{\text{i.i.d.}}{\sim} N(0,1)$, we obtain

$$P(|X_{(k)} - \mathbb{E}[X_{(k)}]| \ge \delta) \le 2e^{-\frac{\delta^2}{2}} \quad \text{for all } \delta \ge 0,$$

since each order statistic is 1-Lipschitz.

Lipschitz functions of Gaussian variables

Example 2.29 (Singular values of Gaussian random matrices)

For integers $n > d$, let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a random matrix with i.i.d. $N(0,1)$ entries, and denote its singular values by

$$\sigma_1(\mathbf{X}) \geq \cdots \geq \sigma_d(\mathbf{X}) \geq 0.$$

- By Weyl's theorem, we have

$$\max_{k=1,\ldots,d} |\sigma_k(\mathbf{X}) - \sigma_k(\mathbf{Y})| \leq \|\mathbf{X} - \mathbf{Y}\|_2 \leq \|\mathbf{X} - \mathbf{Y}\|_2,$$

  where $\|\cdot\|_2$ and $\|\cdot\|_F$ are the operator 2-norm and the Frobenius norm, respectively.

- Note that the Frobenius norm plays the role of the Euclidean norm. Consequently, each singular value is a 1-Lipschitz function of the random matrix, and we obtain

$$P(|\sigma_k(\mathbf{X}) - \mathbb{E}[\sigma_k(\mathbf{X})]| \geq \delta) \leq 2e^{-\frac{\delta^2}{2}} \quad \text{for all } \delta \geq 0.$$

Outline

Outline

Motivation

- In the previous chapter, we discussed concentration inequalities that allow us to bound quantities of the form $|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]|$ for a fixed function f.
- In this chapter, we strengthen these results by deriving *uniform bounds*.
- That is, given a function class $\mathscr{F}$, we explore deviations of $\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]\right|$.
- *Why are uniform bounds so important?* To motivate this, we briefly review two key concepts in both classical and modern statistics: the **plug-in principle** and **empirical risk minimization (ERM)**.

The plug-in principle

- In statistical settings, a typical use of the *empirical cdf* if to construct estimators of various quantities associated with the *population cdf*.
- Many such estimation problems can be formulated in terms of a functional $\gamma : F \mapsto \gamma(F)$, which maps a cdf F to a real number $\gamma(F)$.
- The *plug-in principle* suggests replacing the unknown F with the empirical cdf $\widehat{F}_n$, thereby obtaining an estimate of $\gamma(F)$.

Example 4.1 (Expectation functionals)

Given some integrable function g, we define the *expectation functional* γ_g via $\gamma_g(F) := \int g(x) \, \mathrm{d}F(x)$.
Then, the plug-in estimate of $\mathbb{E}[g(X)]$ is given by $\gamma_g(\widehat{F}_n) = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$.

The plug-in principle

A natural question arises: *when does the plug-in estimator $\gamma(\widehat{F}_n)$ converge to $\gamma(F)$ in probability (or almost surely)?*

- This question can be addressed in a unified manner for many functionals by defining a notion of continuity.
- Given a pair of cdfs F and G, we can measure the distance between F and G using the sup-norm $\|G - F\|_\infty := sup_{t \in \mathbb{R}} |G(t) - F(t)|$.
- We say that the functional γ is *continuous at F in the sup-norm* if, for all $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\|G - F\|_\infty \leq \delta \implies |\gamma(G) - \gamma(F)| \leq \epsilon.$$

The plug-in principle

- For any continuous functional, the question of consistency is now reduced to the convergence of the random variable $\|\widehat{F_n} - F\|_\infty$.
- In light of this perspective, we revisit the well-known Glivenko–Cantelli theorem.

Theorem 4.4 (Glivenko-Cantelli)

For any distribution, the empirical cdf $\widehat{F_n}$ is a strongly consistent estimator of the population cdf in the uniform norm, meaning that

$$\|\widehat{F_n} - F\|_\infty \xrightarrow{a.s.} 0.$$

Uniform laws for more general function classes

We now consider uniform laws in broader contexts.

Let $\mathscr{F}$ be a class of integrable real-valued functions with domain $\mathscr{X}$, and let $\{X_i\}_{i=1}^n$ be a collection of i.i.d. samples from some distribution P over $\mathscr{X}$. Consider the random variable

$$\|P_n - P\|_{\mathscr{F}} := \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right|.$$

Definition 4.5

We say that $\mathscr{F}$ is a *Glivenko-Cantelli class* for P if $\|P_n - P\|_{\mathscr{F}}$ converges to zero in probability as $n \to \infty$.

Remark. When almost sure convergence holds, we say that $\mathscr{F}$ satisfies a strong Glivenko-Cantelli law.

Empirical risk minimization

Variables of the form $\|P_n - P\|_{\mathcal{F}}$ are ubiquitous in statistics, especially in methods based on *empirical risk minimization*.

- Consider an indexed family of probability distributions $\{P_\theta \mid \theta \in \Theta\}$, and suppose we are given samples $X := \{X_i\}_{i=1}^n$, drawn i.i.d. according to P_{θ^*} for some unknown θ^*.
- Let $\theta \mapsto \mathscr{L}_\theta(X)$ be a loss function that measures the fit between parameter θ and the sample X.
- The principle of empirical risk minimization is based on the *empirical risk*

$$\widehat{R_n}(\theta, \theta^*) := \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}_\theta(X_i).$$

- The empirical risk should be contrasted with the *population risk* $R(\theta, \theta^*) := \mathbb{E}_{\theta^*}[\mathscr{L}_\theta(X)]$.

Empirical risk minimization

In practice, one minimizes the empirical risk over some subset Θ_0 of the full space Θ. The statistical question is how to bound the *excess risk*,

$$E(\hat{\theta}, \theta^*) := R(\hat{\theta}, \theta^*) - \inf_{\theta \in \Theta_0} R(\theta, \theta^*).$$

- For simplicity, assume that there exists some $\theta_0 \in \Theta_0$ such that $R(\theta_0, \theta^*) = \inf_{\theta \in \Theta_0} R(\theta, \theta^*)$.
- Then, the excess risk can be decomposed as

$$E(\hat{\theta}, \theta^*) = \underbrace{\left\{R(\hat{\theta}, \theta^*) - \hat{R}_n(\hat{\theta}, \theta^*)\right\}}_{T_1} + \underbrace{\left\{\hat{R}_n(\hat{\theta}, \theta^*) - \hat{R}_n(\theta_0, \theta^*)\right\}}_{T_2 \leq 0} + \underbrace{\left\{\hat{R}_n(\theta_0, \theta^*) - R(\theta_0, \theta^*)\right\}}_{T_3}.$$

Empirical risk minimization

$$E(\hat{\theta}, \theta^*) = \underbrace{\left\{R(\hat{\theta}, \theta^*) - \hat{R}_n(\hat{\theta}, \theta^*)\right\}}_{T_1} + \underbrace{\left\{\hat{R}_n(\hat{\theta}, \theta^*) - \hat{R}_n(\theta_0, \theta^*)\right\}}_{T_2 \leq 0} + \underbrace{\left\{\hat{R}_n(\theta_0, \theta^*) - R(\theta_0, \theta^*)\right\}}_{T_3}.$$

- The term T_2 is non-positive by the definition of θ.
- The term $T_3 = \frac{1}{n}\sum_{i=1}^n \mathscr{L}_{\theta_0}(X_i) - \mathbb{E}_X[\mathscr{L}_{\theta_0}(X)]$ can be dealt with techniques from Chapter 2, since θ_0 is an unknown but deterministic quantity.
- However, we need stronger results to control $T_1 = \mathbb{E}_X[\mathscr{L}_{\hat{\theta}}(X)] - \frac{1}{n}\sum_{i=1}^n \mathscr{L}_{\hat{\theta}}(X_i)$, since $\hat{\theta}$ is a random quantity depending on the sample X.
- Suppose we can control $\|P_n - P\|_{\mathfrak{L}(\Theta_0)}$ via a *uniform law* over the loss function class $\mathfrak{L}(\Theta_0) := \{x \mapsto \mathscr{L}_\theta(x), \theta \in \Theta_0\}$. Since T_1 and T_3 are both dominated by $\|P_n - P\|_{\mathfrak{L}(\Theta_0)}$, we conclude that the excess risk is at most $2\|P_n - P\|_{\mathfrak{L}(\Theta_0)}$.

Empirical risk minimization

The analysis of ERM is fundamental in *statistical learning theory*. Suppose we aim to learn a target function f_* by computing the empirical risk minimizer $\tilde{f}$ over a predefined hypothesis space.

The error $f_* - \tilde{f}$ can be decomposed as follows:

$$f_* - \tilde{f} = \underbrace{f_* - f^0}_{\text{approximation error}} + \underbrace{f^0 - \hat{f}}_{\text{generalization error}} + \underbrace{\hat{f} - \tilde{f}}_{\text{optimization error}},$$

- $f^0 = \arg\min_{f \in \mathcal{F}} \mathbb{E}[\mathscr{L}(Y, f(X))]$: Expected risk minimizing function.
- $\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}(Y_i, f(X_i))$: Empirical risk minimizing function.
- $\tilde{f}$: Approximation of $\hat{f}$ by optimization.

Empirical risk minimization

In particular, uniform bounds play an essential role in controlling the *generalization error* in statistical learning.

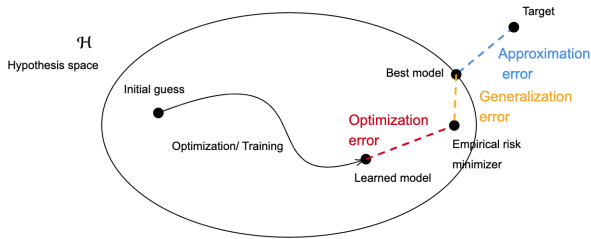

Figure reference: https://dcn.nat.fau.eu/breaking-the-curse-of-dimensionality-with-barron-spaces/

Outline

Rademacher complexity

- So far in our discussion of empirical risk minimization, it is natural to think that the generalization ability in a statistical learning problem depends on the "size" of the function class $\mathscr{F}$.
- If $\mathscr{F}$ is finite, its cardinality serves as a measure of its size. However, in many practical scenarios, $\mathscr{F}$ is infinite, and we need a more refined notion.
- One such notion is *Rademacher complexity*, which quantifies the ability of a function class to fit random binary noise. A function class with high Rademacher complexity is prone to overfitting, leading to poor generalization in ERM.

Rademacher complexity

Let $\mathscr{F}$ be the function class of our interest. For any fixed collection $x_1^n := (x_1, \dots, x_n)$ of points, consider the subset of $\mathbb{R}^n$ given by

$$\mathscr{F}(x_1^n) := \{(f(x_1), \dots, f(x_n)) \mid f \in \mathscr{F}\},$$

i.e. $\mathscr{F}(x_1^n)$ is the set of all possible realizations of applying f to $x_1^n \in \mathbb{R}^n$.

- If $\mathscr{F}$ is a class of binary classifiers, then $|\mathscr{F}(x_1^n)| < 2^n$.
- Intuitively, the cardinality of $\mathscr{F}(x_1^n)$ reflects the variability that can be captured through $\mathscr{F}$, given a fixed sample.

Rademacher complexity

Definition (Rademacher complexity)

Let ε_i be i.i.d. Rademacher variables. Regarding the sample x_1^n as fixed, define the *empirical Rademacher complexity* of a function class $\mathscr{F}$ as

$$\mathscr{R}(\mathscr{F}(x_1^n)/n) := \mathbb{E}_\varepsilon \left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) \right| \right].$$

Then, for random samples $X_1^n := \{X_i\}_{i=1}^n$, the *Rademacher complexity* of $\mathscr{F}$ is given by

$$\mathscr{R}_n(\mathscr{F}) := \mathbb{E}_X[\mathscr{R}(\mathscr{F}(X_1^n)/n)] = \mathbb{E}_{X,\varepsilon} \left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right].$$

i.e. Rademacher complexity is the average of the maximum correlation between $(f(X_1), \dots, f(X_n))$ and the "noise" vector $(\varepsilon_1, \dots, \varepsilon_n)$.

Rademacher complexity

The following theorem presents a fundamental result, providing a connection between Rademacher complexity and the Glivenko-Cantelli property for uniformly bounded function classes.

Theorem 4.10

Let $\mathscr{F}$ be b-uniformly bounded, meaning that $\|f\|_\infty \leq b$ for all $f \in \mathscr{F}$. Then, for any $n \geq 1$ and $\delta \geq 0$, we have

$$\|P_n - P\|_{\mathscr{F}} \leq 2\mathscr{R}_n(\mathscr{F}) + \delta$$

with P-probability at least $1 - \exp\left(\frac{n\delta^2}{2b^2}\right)$.

Consequently, as long as $\mathscr{R}_n(\mathscr{F}) = o(1)$, we have $\|P_n - P\|_{\mathscr{F}} \xrightarrow{\text{a.s.}} 0$.

Rademacher complexity

Proof of Thm 4.10.

Step 1. Concentration around mean

First, we claim that $\|P_n - P\|_{\mathscr{F}} - \mathbb{E}[\|P_n - P\|_{\mathscr{F}}] \leq t$ with P-probability at least $1 - \exp\left(-\frac{nt^2}{2b^2}\right)$.

For notational convenience, define the recentered functions $\bar{f}(x) := f(x) - \mathbb{E}[f(X)]$, and write $\|P_n - P\|_{\mathscr{F}} = \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(X_i) \right|$.

Now, regarding the samples as fixed, consider the following function

$$G(x_1, \dots, x_n) := \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) \right|.$$

Rademacher complexity

Proof of Thm 4.10. (continued)

We claim that G satisfies the bounded differences property with a uniform constant $2b/n$. Since G is invariant to permutations of x_i's, it suffices to bound $|G(x) - G(y)|$ when the first coordinate x_1 is perturbed.

For any function $\bar{f} = f - \mathbb{E}[f]$, we have

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}\bar{f}(x_i)\right| - \sup_{h\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\bar{h}(y_i)\right| &\le \left|\frac{1}{n}\sum_{i=1}^{n}\bar{f}(x_i) - \frac{1}{n}\sum_{i=1}^{n}\bar{f}(y_i)\right| \\ &\le \frac{1}{n}|\bar{f}(x_1) - \bar{f}(y_1)| \\ &\le \frac{2b}{n}.\end{aligned}$$

Taking the supremum on both sides, we obtain $G(x) - G(y) \le 2b/n$, and thus $|G(x) - G(y)| \le 2b/n$ by symmetry. Therefore, we obtain the desired result using the bounded differences inequality (Prop 2.21).

Rademacher complexity

Proof of Thm 4.10. (continued)

Step 2. Upper bound on mean

Now, it remains to prove $\mathbb{E}[\|P_n - P\|_{\mathscr{F}}] \leq 2\mathscr{R}_n(\mathscr{F})$.

Letting $(Y_1, \ldots, Y_n)$ be a second i.i.d. sequence independent of $(X_1, \ldots, X_n)$, we use a *symmetrization argument* as follows:

$$\begin{aligned}
\mathbb{E}\left[\|P_n - P\|_{\mathscr{F}}\right] &= \mathbb{E}_X\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\{f(X_i) - \mathbb{E}_Y[f(Y_i)]\}\right|\right] \\
&= \mathbb{E}_X\left[\sup_{f\in\mathscr{F}}\left|\mathbb{E}_Y\left[\frac{1}{n}\sum_{i=1}^{n}\{f(X_i) - f(Y_i)\}\right]\right|\right] \\
&\leq \mathbb{E}_{X,Y}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\{f(X_i) - f(Y_i)\}\right|\right].
\end{aligned}$$

Rademacher complexity

Proof of Thm 4.10. (continued)

Let $(\varepsilon_1, \ldots, \varepsilon_n)$ be an i.i.d. sequence of Rademacher variables, independent of X and Y.

Then, $\varepsilon_i(f(X_i) - f(Y_i)) \overset{d}{=} f(X_i) - f(Y_i)$, yielding

$$\begin{aligned}\mathbb{E}_{X,Y}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\{f(X_i) - f(Y_i)\}\right|\right] &= \mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i) - f(Y_i))\right|\right] \\ &\leq 2\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right] = 2\mathscr{R}_n(\mathscr{F}).\end{aligned}$$

Combining **Steps 1 & 2** gives the desired result. □

Necessary conditions with Rademacher complexity

- In the proof of Thm 4.10, we used the *symmetrization technique*, relating the random variable $\|P_n - P\|_{\mathscr{F}}$ to its symmetrized version

$$\|S_n\|_{\mathscr{F}} := \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=-1}^{n} \varepsilon_i f(X_i) \right|.$$

- Note that the expectation of $\|S_n\|_{\mathscr{F}}$ is the Rademacher complexity of $\mathscr{F}$.
- Nonetheless, one may wonder whether much was lost in bounding $\|P_n - P\|_{\mathscr{F}}$ with $\|S_n\|_{\mathscr{F}}$. Thus, it is worthwhile to explore the relationship between these two quantities.

Necessary conditions with Rademacher complexity

The following "sandwich" result provides valuable insight:

Proposition 4.11

For any convex non-decreasing function $\Phi : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}_{X,\varepsilon}\left[\Phi\left(\frac{1}{2}\|S_n\|_{\bar{\mathscr{F}}}\right)\right] \leq \mathbb{E}_X[\Phi(\|P_n - P\|_{\mathscr{F}})] \leq \mathbb{E}_{X,\varepsilon}\left[\Phi\left(2\|S_n\|_{\mathscr{F}}\right)\right],$$

where $\bar{\mathscr{F}} := \{f - \mathbb{E}[f], f \in \mathscr{F}\}$ is the recentered function class.

In particular, when applied with $\Phi(t) = t$, we obtain

$$\frac{1}{2}\mathbb{E}_{X,\varepsilon}\left[\|S_n\|_{\bar{\mathscr{F}}}\right] \leq \mathbb{E}_X[\|P_n - P\|_{\mathscr{F}}] \leq 2\mathbb{E}_{X,\varepsilon}\left[\|S_n\|_{\mathscr{F}}\right].$$

Necessary conditions with Rademacher complexity

A consequence of Prop 4.11 is that $\|P_n - P\|_{\mathscr{F}}$ can also be bounded from below using the Rademacher complexity.

Proposition 4.12

For any b-uniformly bounded function class $\mathscr{F}$, and any $n \geq 1, \delta \geq 0$, we have

$$\|P_n - P\|_{\mathscr{F}} \geq \frac{1}{2}\mathscr{R}_n(\mathscr{F}) - \frac{\sup_{f \in \mathscr{F}} |\mathbb{E}[f]|}{2\sqrt{n}} - \delta$$

with P-probability at least $1 - \exp\left(\frac{n\delta^2}{2b^2}\right)$.

- As a result, if the Rademacher complexity $\mathscr{R}_n(\mathscr{F})$ remains bounded away from zero, then $\|P_n - P\|_{\mathscr{F}}$ cannot converge to zero in probability.
- Thus, for a uniformly bounded function class, the Rademacher complexity provides a *necessary and sufficient condition for it to be Glivenko-Cantelli*.

Outline

Upper bounds on the Rademacher complexity

- We end this chapter with a few elementary techniques for bounding the Rademacher complexity, particularly those applicable to function classes with *polynomial discrimination*.
- We also explore the related notion of *Vapnik-Chervonenkis (VC) dimension* and their properties.
- More advanced techniques are addressed in Chapter 5, including those involving metric entropy and chaining arguments.

Classes with polynomial discrimination

Recall that, by definition, the cardinality of $\mathscr{F}(x_1^n)$ provides a sample-dependent measure of the complexity of $\mathscr{F}$.

- If $\mathscr{F}$ is a class of classifiers (or binary-valued functions), then $|\mathscr{F}(x_1^n)| \leq 2^n$.
- We are interested in classes for which $|\mathscr{F}(x_1^n)|$ grows *polynomially* with n.

Definition 4.13 (Polynomial discrimination)

A class $\mathscr{F}$ of functions with domain $\mathscr{X}$ has *polynomial discrimination of order* $\nu \geq 1$ if, each positive integer n and collection x_1^n of n points in $\mathscr{X}$, the set $\mathscr{F}(x_1^n)$has cardinality upper bounded as

$$|\mathscr{F}(x_1^n)| \leq (n+1)^\nu.$$

Classes with polynomial discrimination

The polynomial discrimination property provides a straightforward approch to controlling the VC dimension:

Lemma 4.14

Suppose that $\mathscr{F}$ has polynomial discrimination of order v. Then for all positive integers n and any collection of points $x_1^n = (x_1, \dots, x_n)$,

$$\mathscr{R}(\mathscr{F}(x_1^n)/n) = \mathbb{E}_\varepsilon \left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right] \leq 4D(x_1^n) \sqrt{\frac{v \log(n+1)}{n}},$$

where $D(x_1^n) := \sup_{f \in \mathscr{F}} \sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}}$ is the ℓ_2-radius of the set $\mathscr{F}(x_1^n)/\sqrt{n}$.

Classes with polynomial discrimination

Remarks.

- Note that Lemma 4.14 bounds the *empirical Rademacher complexity*.
- However, in the special case where $\mathscr{F}$ is b-uniformly bounded so that $D(x_1^n) \leq b$ for all samples, we obtain

$$\mathscr{R}_n(\mathscr{F}) \leq 4b\sqrt{\frac{v \log(n+1)}{n}} \quad \text{for all } n \geq 1. \tag{5}$$

  Combined with Thm 4.10, this implies that any bounded function class with polynomial discrimination is Glivenko–Cantelli.
- One such instance is the *class of indicator functions* with polynomial discrimination, which are uniformly bounded by $b = 1$.

Classes with polynomial discrimination

With new tools in hand, we revisit the classical Glivenko-Cantelli theorem from a more quantitative perspective.

Corollary 4.15 (Classical Glivenko-Cantelli)

Let $F(t) = P(X \leq t)$ be the cdf of a random variable X, and let $\widehat{F}_n$ be the empirical cdf based on n i.i.d. samples $X_i \sim P$. Then,

$$P\left(\|\widehat{F}_n - F\|_\infty \geq 8\sqrt{\frac{\log(n+1)}{n}} + \delta \right) \leq e^{-\frac{n\delta^2}{2}} \quad \text{for all } \delta \geq 0,$$

and hence $\|\widehat{F}_n - F\|_\infty \overset{a.s.}{\to} 0$.

Remark. The sharpest bound is given by $P(\|\widehat{F}_n - F\|_\infty \geq \delta) \leq 2e^{-2n\delta^2}$, due to Massart (1990).

Classes with polynomial discrimination

Proof of Cor 4.15.

For a given sample x_1^n, consider the set $\mathscr{F}(x_1^n)$, where $\mathscr{F}$ is the set of all indicator functions of the half-intervals $(-\infty, t]$ for $t \in \mathbb{R}$.

Reordering the samples as $x_{(1)} \leq \cdots \leq x_{(n)}$, the real line is split into at most $n+1$ pieces.

Thus, for any sample x_1^n, we have $|\mathscr{F}(x_1^n)| \leq n+1$. Applying the inequality from (5) with $b = 1$ and $v = 1$, we obtain $\mathscr{R}_n(\mathscr{F}) \leq 4\sqrt{\frac{\log(n+1)}{n}}$.

The claim then follows from Thm 4.10. □

Vapnik-Chervonenkis dimension

The theory of *Vapnik–Chervonenkis (VC) dimension* provides a more efficient means of verifying the polynomial discrimination property for a function class.

We focus on function classes $\mathcal{F}$ consisting of $\{0,1\}$ binary-valued functions, i.e., indicator functions. We begin by defining the concept of *shattering*.

Definition 4.16 (Shattering and VC dimension)

Given a class $\mathcal{F}$ of binary-valued functions, we say that the set $x_1^n = (x_1, \ldots, x_n)$ is *shattered* by $\mathcal{F}$ if $|\mathcal{F}(x_1^n)| = 2^n$.

The *VC dimension* $v(\mathcal{F})$ is the largest integer n for which there is *some* collection $x_1^n = (x_1, \ldots, x_n)$ of n points that is shattered by $\mathcal{F}$. When $v(\mathcal{F})$ is finite, the function class $\mathcal{F}$ is said to be a *VC class*.

Vapnik-Chervonenkis dimension

Remarks.

- Note that there is a 1-1 correspondence between a class of sets $\mathscr{S}$ and its corresponding class of indicator functions $\mathscr{F}$. Accordingly, we use the notations $\mathscr{S}(x_1^n)$ and $\nu(S)$.
- For a given set class $\mathscr{S}$, the *shatter coefficient* of order n is defined as

$$s(\mathscr{S}, n) := \max_{(x_1,\dots,x_n)} |\mathscr{S}(x_1^n)|.$$

  Thus, x_1^n is shattered by $\mathscr{S}$ if $s(\mathscr{S}, n) = 2^n$.

Vapnik-Chervonenkis dimension

Example 4.17 (Intervals in $\mathbb{R}$)

First, consider the class of all indicator functions for left-sided half-intervals on $\mathbb{R}$, namely $\mathscr{S}_{\text{left}} := \{(\infty, a] \mid a \in \mathbb{R}\}$.

- It is implicit in the proof of Cor 4.15 that $\nu(\mathscr{S}_{\text{left}}) = 1$:
  - Any singleton set x_1 can be picked out by $\mathscr{S}_{\text{left}}$.
  - However, for any $x_1 < x_2$, it is impossible to find a left-sided interval that contains x_2 but not x_1.
- Indeed, we have shown more specifically that $|\mathscr{S}_{\text{left}}| \leq n + 1$.

Vapnik-Chervonenkis dimension

Example 4.17 (Intervals in $\mathbb{R}$; continued)

Now, consider the class of all two-sided intervals on $\mathbb{R}$, namely $\mathscr{S}_{\text{two}} := \{(b, a] \mid a, b \in \mathbb{R} \text{ s.t. } b < a\}$.

- It is easy to verify that $\nu(\mathscr{S}_{\text{two}}) = 2$:
  - Any two-point set can be shattered by $\mathscr{S}_{\text{two}}$.
  - However, given three distinct points $x_1 < x_2 < x_3$, $\mathscr{S}_{\text{two}}$ cannot pick out the subset $\{x_1, x_3\}$.
- Moreover, by an argument similar to Cor 4.15, we can obtain a crude bound $|\mathscr{S}_{\text{two}}| \leq (n+1)^2$.

Vapnik-Chervonenkis dimension

The previous example showed two function classes with finite VC dimension, both of which also had polynomial discrimination. Is this merely a coincidence?

In fact, *any finite VC class has polynomial discrimination with degree at most the VC dimension*:

Proposition 4.18 (Vapnik-Chervonenkis, Sauer and Shelah)

Consider a set class $\mathscr{S}$ with $\nu(\mathscr{S}) < \infty$. Then, for any collection of points $x_1^n = (x_1, \dots, x_n)$ with $n > \nu(\mathscr{S})$, we have

$$|\mathscr{S}(x_1^n)| \leq \sum_{i=0}^{\nu(\mathscr{S})} \binom{n}{i} \leq \left(\frac{en}{\nu(\mathscr{S})}\right)^{\nu(\mathscr{S})}.$$

In particular, we have $|\mathscr{S}(x_1^n)| \leq (n+1)^{\nu(\mathscr{S})}$ for $n > \nu(\mathscr{S})$.

Controlling the VC dimension

Since classes with finite VC dimension have polynomial discrimination, it is of interest to develop techniques for controlling the VC dimension.

To begin with, the property of having a finite VC dimension is preserved under a number of basic operations:

Proposition 4.19

Let $\mathscr{S}$ and $\mathscr{T}$ be set classes, each with finite VC dimensions. Then the following set classes are also of finite VC dimension:

(a) The set class $\mathscr{S}^c := \{S^c \mid S \in \mathscr{S}\}$.

(b) The set class $\mathscr{S} \sqcup \mathscr{T} := \{S \cup T \mid S \in \mathscr{S}, T \in \mathscr{T}\}$.

(c) The set class $\mathscr{S} \sqcap \mathscr{T} := \{S \cap T \mid S \in \mathscr{S}, T \in \mathscr{T}\}$.

Controlling the VC dimension

For any real-valued function $g : \mathscr{X} \to \mathbb{R}$, define its *subgraph at level zero* by the subset $S_g := \{x \in \mathscr{X} \mid g(x) \leq 0\}$. Analogously, for a function class $\mathscr{G}$, we obtain the subgraph class $\mathscr{S}(\mathscr{G}) := \{S_g \mid g \in \mathscr{G}\}$.

In many cases, the function class $\mathscr{G}$ has a *vector space structure*. The following proposition allows us to upper bound the VC dimension of the associated set class $\mathscr{S}(\mathscr{G})$.

Proposition 4.20 (Finite-dimensional vector spaces)

Let $\mathscr{G}$ be a vector space of functions $g : \mathbb{R}^d \to \mathbb{R}$ with dimension $\dim(\mathscr{G}) < \infty$. Then the subgraph class $\mathscr{S}(\mathscr{G})$ has VC dimension at most $\dim(\mathscr{G})$.

Controlling the VC dimension

We demonstrate the use of Prop 4.20 through several examples:

Example 4.21 (Linear functions in $\mathbb{R}^d$ and half-spaces)

- For a pair $(a, b) \in \mathbb{R}^d \times \mathbb{R}$, define the linear function $f_{a,b}(x) := \langle a, x \rangle + b$, and consider the family $\mathscr{L}^d := \{f_{a,b} \mid (a, b) \in \mathbb{R}^d \times \mathbb{R}\}$.
- The associated subgraph class corresponds to the collection of all *half-spaces* of the form $H_{a,b} := \{x \in \mathbb{R}^d \mid \langle a, x \rangle + b \leq 0\}$.
- Since $\mathscr{L}^d$ is a vector space of dimension $d + 1$, it follows that $\mathscr{S}(\mathscr{L}^d)$ has VC dimension at most $d + 1$.

Remark. In general, it can be shown that the VC dimension of $\mathscr{S}(\mathscr{L}^d)$ is $d + 1$ for all dimensions.

Controlling the VC dimension

Example 4.22 (Spheres in $\mathbb{R}^d$)

- Consider the sphere $S_{a,b} := \{x \in \mathbb{R}^d \mid \|x - a\|_2 \leq b\}$, where $(a, b) \in \mathbb{R}^d \times \mathbb{R}_+$ specify its center and radius, respectively, and let $\mathscr{S}^d_{\text{sphere}}$ denote the collection of all such spheres.
- If we define the function

$$f_{a,b}(x) := \|x\|_2^2 - 2\sum_{j=1}^{d} a_j x_j + \|a\|_2^2 - b^2,$$

then we have $S_{a,b} = \{x \in \mathbb{R}^d \mid f_{a,b}(x) \leq 0\}$, so that the sphere $S_{a,b}$ is a subgraph of the function $f_{a,b}$.

Controlling the VC dimension

Example 4.22 (Spheres in $\mathbb{R}^d$; continued)

- Define a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^{d+2}$ via $\phi(x) := (1, x_1, \dots, x_d, \|x\|_2^2)$, and consider functions of the form $g_c(x) := \langle c, \phi(x) \rangle$, where $c \in \mathbb{R}^{d+2}$.
- The family of functions $\{g_c,\ c \in \mathbb{R}^{d+2}\}$ is a vector space of dimension $d+2$, and it contains the function class $\{f_{a,b},\ (a,b) \in \mathbb{R}^d \times \mathbb{R}_+\}$.
- Consequently, the VC dimension of $\mathscr{S}^d_{\text{sphere}}$ is at most $d+2$.

Remark. While this bound is adequate for many cases, it can be further sharpened to $d+1$.