Chapter 4

Some asymptotic results for the bootstrap distribution of the sample mean

4.1 Introduction

In Chapter 1, we argued that Edgeworth expansions and saddlepoint approximations have provided much of the theoretical underpinning for the bootstrap. They give a mathematical basis for assessing its properties and comparing its performance with other techniques.

Edgeworth expansions provide the order of magnitude (in probability) of the absolute error between a bootstrap distribution and the true distribution it estimates. Results are now known for many statistics of practical interest, such as smooth functions of a multidimensional sample mean (Hall, 1992) and M-estimators (Lahiri, 1992, 1994), and are often cited in support of the bootstrap.

Saddlepoint approximations offer a different form of evidence, namely the order of the *relative* error in a bootstrap approximation. This information is particularly relevant when the magnitude of the feature of the underlying population of interest, such as a tail probability, may be very small. Statistics studied from this perspective include sample means (Jing, Feuerverger and Robinson, 1994) and smooth functions of sample means (Robinson and Skovgaard, 1998).

However, these are not the only aspects of bootstrap performance which merit consideration. In this chapter, we take a different approach, as a first attempt to answer the basic question, 'What is the probability that the bootstrap performs badly?'. A mathematical formulation of this problem involves appropriate choices both of a statistic and of a distance between distributions. We work with the univariate sample mean, and the Mallows distance (Mallows, 1972), whose properties were exploited effectively in a bootstrap context by Bickel and Freedman (1981). The main objective is to study the rate of decay of the probability that the distance between the true distribution of the normalised sample mean and its bootstrap approximation exceeds a given threshold.

In Sections 4.2 and 4.3, we review the Mallows distance and show how to reduce our bootstrap problem to one of studying the Mallows distance between a distribution and the empirical distribution of a sample. The main essence of the results in Sections 4.4 and 4.5 is that rate of decay of the probability of poor bootstrap performance depends on the tail of the underlying population. In Section 4.4, we give an explicit bound on the probability of the Mallows distance exceeding a threshold, and show that, under certain tail and smoothness conditions, this bound may decay exponentially in the sample size; that is, the bound is no more than $e^{-n\delta}$, for some $\delta > 0$ and sufficiently large sample sizes n. This may be interpreted as a mathematical statement that in such cases the probability of poor bootstrap performance decays satisfactorily. However, by choosing a distribution with a sufficiently heavy tail, we can ensure the bound decays no faster than $\exp(-n^\beta)$, for any given $\beta \in (0, 1)$.

Section 4.5 provides further supporting evidence. For example, where the underlying population is of bounded support, it is shown that a large deviations upper bound exists on the probability of the Mallows distance exceeding a threshold. However, for distributions with heavy (polynomial) tails, the empirical distributions fail to satisfy a large deviations principle in the Mallows topology. This shows the delicacy of Sanov's theorem, which says that the empirical measures do satisfy a large deviations principle in the (coarser) weak topology. Results are not known for populations with infinite but light tails, such as the exponential and normal distributions, and these remain interesting topics for further research. The proofs omitted in the main text are given in Section 4.6.

4.2 The Mallows distance on the real line

Let $\mathcal{F}$ denote the set of all distribution functions on the real line and, for $r \geq 1$, let $\mathcal{F}_r = \{F \in \mathcal{F} : \int_{-\infty}^{\infty} |x|^r \, dF(x) < \infty\}$. For $F, G \in \mathcal{F}_r$, the Mallows metric $d_r(F, G)$ is defined by

$$d_r(F, G) = \inf_{\mathcal{T}_{X,Y}} \left\{\mathbb{E}|X - Y|^r\right\}^{1/r},$$

where $\mathcal{T}_{X,Y}$ is the set of all joint distributions of pairs of random variables X and Y whose marginal distributions are F and G respectively. In a slight abuse of notation, we also write $d_r(X, Y)$ for $d_r(F, G)$, where this will cause no confusion. The following results about d_r are proved in Bickel and Freedman (1981) and Major (1978):

(a) If $(F_n) \in \mathcal{F}$ and $F \in \mathcal{F}$, then $d_r(F_n, F) \to 0$ as $n \to \infty$ if and only if, for every bounded, continuous function $g : \mathbb{R} \to \mathbb{R}$, we have

$$\int_{-\infty}^{\infty} g(x)\, dF_n(x) \to \int_{-\infty}^{\infty} g(x)\, dF(x)$$

as $n \to \infty$, and also

$$\int_{-\infty}^{\infty} |x|^r\, dF_n(x) \to \int_{-\infty}^{\infty} |x|^r dF(x)$$

as $n \to \infty$. Thus, convergence in the Mallows metric d_r is equivalent to convergence in distribution together with convergence of the rth absolute moments.

(b) If $a \in \mathbb{R}$ and X, Y are random variables with finite rth absolute moments, then

$$d_r(aX, aY) = |a| d_r(X, Y).$$

(c) The infimum in the definition of the Mallows metric is attained by the following construction: let $U \sim U(0,1)$, and define $X = F^{-1}(U)$, $Y = G^{-1}(U)$. Here, F^{-1} and G^{-1} are the left-continuous versions of the respective inverse functions, so that, for example, $F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$. Thus

$$d_r(F, G) = \left(\int_0^1 |F^{-1}(p) - G^{-1}(p)|^r\, dp \right)^{1/r}.$$

It is therefore more convenient in much of what follows to work with the set

$$\mathcal{G}_r = \left\{ G : (0,1) \to \mathbb{R} : G \text{ is left-continuous, increasing and } \int_0^1 |G(p)|^r dp < \infty \right\}$$

equipped with the L_r-norm restricted to this set:

$$\|G\|_r = \left(\int_0^1 |G(p)|^r\, dp \right)^{1/r}.$$

The map from $(\mathcal{F}_r, d_r)$ to $(\mathcal{G}_r, \|\cdot\|_r)$ which sends a distribution function to its left-continuous inverse is a distance-preserving bijection.

(d) Suppose X and Y have distributions in $\mathcal{F}_2$. Then

$$d_2(X, Y)^2 = d_2\big(X - \mathbb{E}(X)\,,\, Y - \mathbb{E}(Y)\big)^2 + \big(\mathbb{E}(X) - \mathbb{E}(Y)\big)^2.$$

(e) Suppose $X_1, \ldots, X_n$ are independent, $Y_1, \ldots, Y_n$ are independent, that all the distributions are in $\mathcal{F}_2$, and that $\mathbb{E}(X_i) = \mathbb{E}(Y_i)$ for $i = 1, \ldots, n$. Then

$$d_2\Bigl(\sum_{i=1}^n X_i\,,\, \sum_{i=1}^n Y_i\Bigr)^2 \le \sum_{i=1}^n d_2(X_i, Y_i)^2.$$

Equality is attained if $X_1, \ldots, X_n$ are independent $N(\mu, \sigma_X^2)$ random variables and $Y_1, \ldots, Y_n$ are independent $N(\mu, \sigma_Y^2)$ random variables, for some $\mu \in \mathbb{R}$ and $\sigma_X^2, \sigma_Y^2 \ge 0$.

We add two further properties in Proposition 4.2.1 below. The completeness of the Mallows metric on a separable metric space is already known (Dobrushin, 1970), but we can give a much simpler argument for the case of distributions on the real line.

Proposition 4.2.1. *The metric space $(\mathcal{F}_r, d_r)$ is separable and complete.*

Proof. To show separability, consider the set

$$\mathcal{H}_r = \{H \in \mathcal{G}_r : H(p) \in \mathbb{Q} \text{ for all } p \in \mathbb{Q} \cap (0,1)\}.$$

Note that any function in $\mathcal{G}_r$ is determined by its values at the rational points in $(0,1)$, and that $\mathcal{H}_r$ is countable. Moreover, given $\epsilon > 0$ and any $G \in \mathcal{G}_r$, we can choose values $H(p)$ for $p \in \mathbb{Q} \cap (0,1)$ such that

$$|H(p) - G(p)| \le \epsilon$$

and $H(p_1) \le H(p_2)$ whenever $p_1 \le p_2$. Extending H to a left-continuous function on $(0,1)$ (which is necessarily increasing), we have $|H(p) - G(p)| \le \epsilon$ for all $p \in (0,1)$, so

$$\|H - G\|_r \le \epsilon,$$

and moreover,

$$\|H\|_r \le \|G\|_r + \|H - G\|_r \le \|G\|_r + \epsilon,$$

so $H \in \mathcal{H}_r$. Hence $\mathcal{H}_r$ is dense in $\mathcal{G}_r$, for each r. Consequently, the distribution functions corresponding to the functions in $\mathcal{H}_r$ are dense in $\mathcal{F}_r$.

Now suppose (F_n) is a Cauchy sequence in $(\mathcal{F}_r, d_r)$. Let $U \sim U(0,1)$, and for each $n \in \mathbb{N}$, let $X_n = F_n^{-1}(U)$. Then X_n has distribution function F_n, and for each $m, n \in \mathbb{N}$, we have

$$d_r(F_m, F_n) = \big(\mathbb{E}|X_m - X_n|^r\big)^{1/r}.$$

Thus (X_n) is a Cauchy sequence in L_r. But L_r is complete (Billingsley, 1995, p. 243), so there exists a random variable $X \in L_r$ such that

$$\mathbb{E}|X_n - X|^r \to 0$$

as $n \to \infty$. Hence if F is the distribution function of X, then

$$d_r(F_n, F) \leq \big(\mathbb{E}|X_n - X|^r\big)^{1/r} \to 0$$

as $n \to \infty$. □

4.3 The Mallows distance and the bootstrap

Suppose $X_1, \ldots, X_n$ are independent random variables, each having distribution function F with mean μ and finite variance. Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ denote the sample mean. If we are interested in making inference about μ, a natural root to consider is $n^{1/2}(\bar{X}_n - \mu)$, whose sampling distribution under F we denote by $H_n(F)$. Conditional on $X_1, \ldots, X_n$, let $X_1^*, \ldots, X_n^*$ be independent and identically distributed random variables drawn from the empirical distribution of the sample, whose distribution function, $\hat{F}_n$, is given by

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}$$

for $x \in \mathbb{R}$. The non-parametric bootstrap approximates the sampling distribution of $n^{1/2}(\bar{X}_n - \mu)$ by that of $n^{1/2}(\bar{X}_n^* - \bar{X}_n)$, where $\bar{X}_n^* = n^{-1}\sum_{i=1}^n X_i^*$. In other words, conditional on $X_1, \ldots, X_n$, we approximate $H_n(F)$ by $H_n(\hat{F}_n)$. The calculation below, which follows Shao and Tu (1995), shows how the properties of the Mallows distance d_2 outlined in Section 4.2 make it suitable for studying the performance of the bootstrap approximation in this context:

$$\begin{aligned}
d_2\big(H_n(\hat{F}_n), H_n(F)\big) &= d_2\left(\frac{1}{n^{1/2}}\sum_{i=1}^n (X_i^* - \bar{X}_n)\,,\, \frac{1}{n^{1/2}}\sum_{i=1}^n (X_i - \mu)\right) \\
&\leq \frac{1}{n^{1/2}}\left(\sum_{i=1}^n d_2(X_i^* - \bar{X}_n\,,\, X_i - \mu)^2\right)^{1/2} \\
&= d_2(X_1^* - \bar{X}_n\,,\, X_1 - \mu) \\
&\leq d_2(X_1^*, X_1) \\
&= d_2(\hat{F}_n, F).
\end{aligned}$$

Thus, in particular, the distance between the distribution of the root of interest, $H_n(F)$, and its bootstrap approximation, $H_n(\hat{F}_n)$, is stochastically dominated by the distance between the true and empirical distributions.

It follows immediately by property (a) in Section 4.2 and the strong law of large numbers that $d_2(\hat{F}_n, F) \to 0$ almost surely as $n \to \infty$. In straightforward cases, we can give a limiting distribution for $n^{1/2} d_2(\hat{F}_n, F)$, as in Theorem 4.3.2 below. Let $D = D[0,1]$ denote the space of left-continuous, real-valued functions on $[0,1]$ possessing right limits at each point. We may equip D with the uniform norm, so that for $x, y \in D$, we define

$$\|x - y\|_\infty = \sup_{p \in [0,1]} |x(p) - y(p)|.$$

A technical complication in Theorem 4.3.2 arises from the fact that the normed space $(D, \|\cdot\|_\infty)$ is non-separable, and the σ-algebra, $\mathcal{D}$, generated by the open balls is strictly smaller than the Borel σ-algebra, $\mathcal{D}_{\text{Borel}}$, generated by the open sets. This creates measurability problems, as explained in Chibisov (1965), which lead us to work with the space $(D, \mathcal{D}, \|\cdot\|_\infty)$. A consequence of using the ball σ-algebra is that we must make a slight modification to the notion of weak convergence, in line with Billingsley (1999), p. 67:

Definition 4.3.1. *If $(Y_n)_{n\geq 0}$ is a sequence of random elements of $(D, \mathcal{D}, \|\cdot\|_\infty)$, we write $Y_n \xrightarrow{d^\circ} Y_0$ as $n \to \infty$ if*

$$\mathbb{E}\big(f(Y_n)\big) \to \mathbb{E}\big(f(Y_0)\big)$$

as $n \to \infty$, for all bounded, continuous functions $f : D \to \mathbb{R}$ which are $\mathcal{D}$-measurable.

Recall that a Brownian bridge $B = \big(B(p)\big)_{0\leq p\leq 1}$ is a zero mean Gaussian process with

$$\operatorname{Cov}\big(B(p), B(q)\big) = p(1-q)$$

for $p \leq q$. For $p \in (0,1)$, let $\xi_p = \inf\{x \in \mathbb{R} : F(x) \geq p\}$.

Theorem 4.3.2. *Suppose that the limits $\xi_0 = \lim_{p\searrow 0} \xi_p$ and $\xi_1 = \lim_{p\nearrow 1} \xi_p$ exist in $\mathbb{R}$, and that F has a density f such that $f(\xi_p)$ is positive and continuous for $p \in [0,1]$. Let $B = \big(B(p)\big)_{0\leq p\leq 1}$ denote a Brownian bridge. Then*

$$n^{1/2} d_2(\hat{F}_n, F) \xrightarrow{d} \left(\int_0^1 \frac{B^2(p)}{f^2(\xi_p)}\, dp\right)^{1/2}$$

as $n \to \infty$.

Proof. Theorem 1 on pp. 640–641 of Shorack and Wellner (1986), together with Corollary 1 on p. 48 of the same book, give that

$$f(\xi_p)\, n^{1/2}\big(\hat{F}_n^{-1}(p) - \xi_p\big) \xrightarrow{d^\circ} B(p)$$

on $(D, \mathcal{D}, \|\cdot\|_\infty)$, as $n \to \infty$. Now, with probability one, B belongs to the space $(C[0,1], \|\cdot\|_\infty)$ of continuous real-valued functions on $[0,1]$ equipped with the uniform norm, and moreover this space is separable. We can therefore apply the version of the continuous mapping theorem for $\overset{d^\circ}{\to}$ convergence (Billingsley, 1999, pp. 67–68) to a composition map $h(p) = h_2\big(h_1(p)\big)$ from $(D, \mathcal{D}, \|\cdot\|_\infty)$ to $\mathbb{R}$. The individual maps $h_1 : (D, \mathcal{D}, \|\cdot\|_\infty) \to (D, \mathcal{D}, \|\cdot\|_\infty)$ and $h_2 : (D, \mathcal{D}, \|\cdot\|_\infty) \to \mathbb{R}$ are defined by

$$h_1(G)(p) = \frac{G^2(p)}{f^2(\xi_p)} \quad \text{and} \quad h_2(G) = \left(\int_0^1 G(p)\,dp\right)^{1/2}.$$

Observe that the continuity of h_1 follows from the fact that $f(\xi_p)$ attains its (positive) infimum for some $p \in [0,1]$. We conclude that

$$n^{1/2} d_2(\hat{F}_n, F) = \left(\int_0^1 n\big(\hat{F}_n^{-1}(p) - \xi_p\big)^2\,dp\right)^{1/2} \overset{d^\circ}{\to} \left(\int_0^1 \frac{B^2(p)}{f^2(\xi_p)}\,dp\right)^{1/2}$$

as $n \to \infty$. The result follows on noting that any bounded, continuous function from $\mathbb{R}$ to $\mathbb{R}$ is (Borel) measurable. □

4.4 An exponential bound?

The following inequality, which is derived in Serfling (1980) from a lemma of Hoeffding (1963), is crucial for obtaining the main bound of this section in Theorem 4.4.3.

Lemma 4.4.1. *Let $F \in \mathcal{F}$, and suppose $p \in (0,1)$ is such that there exists a unique $x \in \mathbb{R}$ such that $F(x_-) \le p \le F(x)$, where $F(x_-) = \lim_{y \nearrow x} F(y)$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}(|\hat{F}_n^{-1}(p) - \xi_p| \ge \epsilon) \le 2e^{-2n\delta_\epsilon^2},$$

where $\delta_\epsilon = \min\{F(\xi_p + \epsilon) - p\,,\, p - F(\xi_p - \epsilon)\}$.

Let $B = \{p \in (0,1) :$ there exist $x_0 < x_1$ satisfying $F(x_0) = F(x_1) = p\}$. If $p \in B$, then F is constant in a right-neighbourhood of ξ_p, so B is countable.

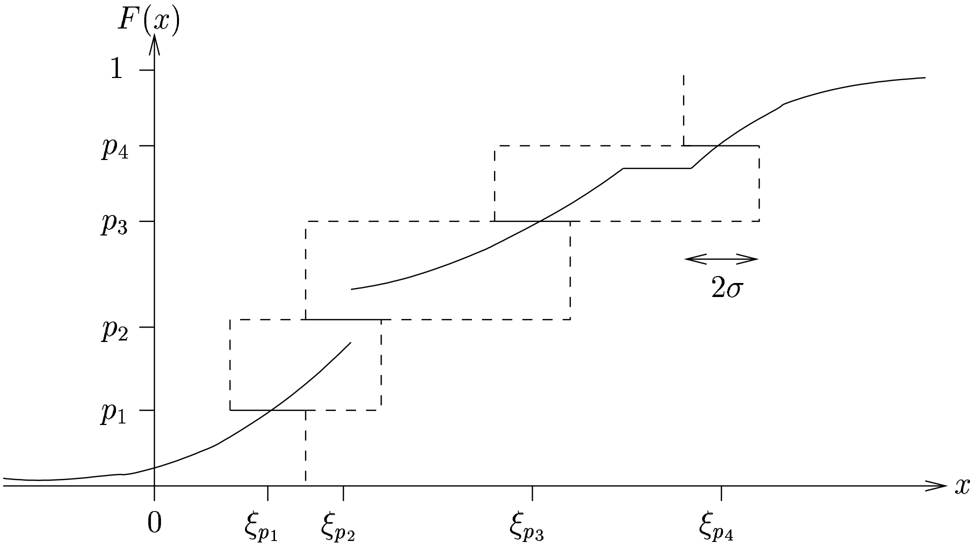

Figure 4.1: A plot of a distribution function F and the construction used in the proof of Theorem 4.4.2.

Theorem 4.4.2. *Let $F \in \mathcal{F}_2$, $\sigma > 0$ and $n \in \mathbb{N}$. Suppose $p_1, \ldots, p_n$ are such that $p_i \in \big((i-1)/n, i/n\big]$ and $p_i \notin B$ for $i = 1, \ldots, n$. Let*

$$\epsilon = \int_0^{p_1} (\xi_{p_1} + \sigma - \xi_p)^2 \, dp + \sum_{i=1}^{n-1} (\xi_{p_{i+1}} - \xi_{p_i} + 2\sigma)^2 (p_{i+1} - p_i) + \int_{p_n}^1 \big(\xi_p - (\xi_{p_n} - \sigma)\big)^2 \, dp.$$

Then

$$\mathbb{P}\big(d_2(\hat{F}_n, F)^2 > \epsilon\big) \leq 2 \sum_{i=1}^n e^{-2n\delta_\sigma(p_i)^2},$$

where $\delta_\sigma(p) = \min\{F(\xi_p + \sigma) - p \, , \, p - F(\xi_p - \sigma)\}$.

Proof. See Figure 4.1. We write

$$\begin{aligned} d_2(\hat{F}_n, F)^2 &= \int_0^1 \big(\hat{F}_n^{-1}(p) - \xi_p\big)^2 \, dp \\ &= \int_0^{p_1} \big(\hat{F}_n^{-1}(p) - \xi_p\big)^2 \, dp + \sum_{i=1}^{n-1} \int_{p_i}^{p_{i+1}} \big(\hat{F}_n^{-1}(p) - \xi_p\big)^2 \, dp + \int_{p_n}^1 \big(\hat{F}_n^{-1}(p) - \xi_p\big)^2 \, dp. \end{aligned}$$

Observe that $\hat{F}_n^{-1}(p)$ is constant for $p \in (0, p_1)$, so that if $\xi_{p_1} - \sigma \leq \hat{F}_n^{-1}(p_1) \leq \xi_{p_1} + \sigma$, then

$$\int_0^{p_1} \big(\hat{F}_n^{-1}(p) - \xi_p\big)^2 \, dp \leq \int_0^{p_1} (\xi_{p_1} + \sigma - \xi_p)^2 \, dp.$$

A similar argument applies for the interval $(p_n, 1)$, and since $\hat{F}_n$ is a increasing function, it follows that for $i = 1, \ldots, n-1$,

$$\int_{p_i}^{p_{i+1}} \big(\hat{F}_n^{-1}(p) - \xi_p\big)^2 \, dp \leq (\xi_{p_{i+1}} - \xi_{p_i} + 2\sigma)^2 (p_{i+1} - p_i),$$

whenever

$$\xi_{p_i} - \sigma \leq \hat{F}_n^{-1}(p_i) \leq \xi_{p_i} + \sigma \quad \text{and} \quad \xi_{p_{i+1}} - \sigma \leq \hat{F}_n^{-1}(p_{i+1}) \leq \xi_{p_{i+1}} + \sigma.$$

For $i = 1, \ldots, n$, let

$$B_i = \{\xi_{p_i} - \sigma \leq \hat{F}_n^{-1}(p_i) \leq \xi_{p_i} + \sigma\}.$$

Then by Lemma 4.4.1,

$$\mathbb{P}\big(d_2(\hat{F}_n, F)^2 > \epsilon\big) \leq \mathbb{P}\Big(\bigcup_{i=1}^n B_i^c\Big) \leq \sum_{i=1}^n \mathbb{P}(B_i^c) \leq 2 \sum_{i=1}^n e^{-2n\delta_\sigma(p_i)^2},$$

where $\delta_\sigma(p)$ is as stated in the theorem. □

We are particularly interested in values of $p_1, \ldots, p_n$ satisfying

(a) $p_1 \in \big(1/(2n), 1/n\big]$, $p_n \in \big(1 - 1/n, 1 - 1/(2n)\big]$ and $p_i \in \big(i/n, (i+1)/n\big]$ for $i = 2, \ldots, n-1$

(b) $p_i \notin B$ for $i = 1, \ldots, n$.

Theorem 4.4.3. *Given any $\epsilon > 0$, there exists $\sigma > 0$ such that for sufficiently large n, and all $p_1, \ldots, p_n$ satisfying conditions (a) and (b) above, we have*

$$\mathbb{P}\big(d_2(\hat{F}_n, F)^2 > \epsilon\big) \leq 2 \sum_{i=1}^n e^{-2n\delta_\sigma(p_i)^2}. \tag{4.1}$$

Proof. The proof is a matter of showing that the positive ϵ in Theorem 4.4.2 can be made arbitrarily small by choosing $\sigma > 0$ suitably small and n sufficiently large. Observe that for $x > 0$,

$$x^2\bigl(1 - F(x)\bigr) = x^2 \int_x^\infty dF(y) \le \int_x^\infty y^2\, dF(y),$$

and since $F \in \mathcal{F}_2$, we may apply the dominated convergence theorem to conclude that $1 - F(x) = o(x^{-2})$ as $x \to \infty$, and similarly $F(x) = o(x^{-2})$ as $x \to -\infty$. Thus $(1-p)\xi_p^2 \to 0$ as $p \to 1$ and $p\xi_p^2 \to 0$ as $p \to 0$. Hence, given $\epsilon > 0$, we may choose n_0 large enough such that

$$\xi_{1/(2n)}^2 \le \frac{\epsilon n}{16} \quad \text{and} \quad \xi_{1-1/(2n)}^2 \le \frac{\epsilon n}{16}$$

as well as

$$\int_0^{1/n} (\xi_{1/n} - \xi_p)^2\, dp \le \frac{\epsilon}{8} \quad \text{and} \quad \int_{1-1/n}^1 (\xi_p - \xi_{1-1/n})^2\, dp \le \frac{\epsilon}{8}$$

for $n \ge n_0$. For such n, and for $p_1, \ldots, p_n$ satisfying conditions (a) and (b),

$$\begin{aligned}\sum_{i=1}^{n-1} (\xi_{p_{i+1}} - \xi_{p_i})^2 (p_{i+1} - p_i) &\le \max_{1 \le i \le n-1} (p_{i+1} - p_i)(\xi_{p_n} - \xi_{p_1})^2 \\ &\le \frac{2}{n}\bigl(\xi_{1-1/(2n)} - \xi_{1/(2n)}\bigr)^2 \\ &\le \frac{4}{n}\bigl(\xi_{1-1/(2n)}^2 + \xi_{1/(2n)}^2\bigr) \\ &\le \frac{\epsilon}{2}.\end{aligned}$$

Finally, choose $\sigma > 0$ small enough such that, for all $n \ge n_0$,

$$\int_0^{1/n} (\xi_{1/n} + \sigma - \xi_p)^2\, dp \le \frac{\epsilon}{6}, \quad \int_{1-1/n}^1 \bigl(\xi_p - (\xi_{1-1/n} - \sigma)\bigr)^2\, dp \le \frac{\epsilon}{6},$$

and

$$\sum_{i=1}^{n-1} (\xi_{p_{i+1}} - \xi_{p_i} + 2\sigma)^2 (p_{i+1} - p_i) \le \frac{2\epsilon}{3}.$$

Theorem 4.4.2 now completes the proof. □

Remark: Since $d_2(\hat{F}_n, F)$ is stochastically dominated by $d_2\big(H_n(\hat{F}_n), H_n(F)\big)$, the same bound (4.1) holds for $\mathbb{P}\big\{d_2\big(H_n(\hat{F}_n), H_n(F)\big)^2 > \epsilon\big\}$ under the conditions of the theorem.

Although the bound (4.1) appears at first sight to give a very satisfactory mathematical answer to the original question posed in the introduction concerning the probability of poor bootstrap performance for the sample mean, it is in fact not always the case that (4.1) is a genuine exponential bound in n. For instance, if for $x > 1$ and some $m > 3$,

$$F(x) = 1 - \frac{1}{x^{m-1}},$$

so that F has density $f(x) = (m-1)/x^m$ for $x > 1$, and $\xi_p = (1-p)^{-1/(m-1)}$, then

$$\sum_{i=1}^{n} e^{-2n\delta_\sigma(p_i)^2} \geq e^{-2n\sigma^2 f^2(\xi_{1-1/n})} = \exp\big(-2(m-1)^2\sigma^2 n^{(m-3)/(m-1)}\big).$$

Note that the power of n may be made arbitrarily close to zero by choosing m sufficiently close to 3. The problems here are caused by the heavy tails in the underlying distribution. The following result, however, gives simple conditions under which the bound (4.1) decays exponentially in n.

Corollary 4.4.4. *Suppose that the limits $\xi_0 = \lim_{p\searrow 0}\xi_p$ and $\xi_1 = \lim_{p\nearrow 1}\xi_p$ exist in $\mathbb{R}$, and that F has a density f such that $f(\xi_p)$ is positive and continuous for $p \in [0,1]$. Then given any $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$ such that, for all sufficiently large $n \in \mathbb{N}$,*

$$\mathbb{P}\big(d_2(\hat{F}_n, F)^2 > \epsilon\big) \leq e^{-n\delta}.$$

Proof. Let

$$\alpha = \inf_{p\in[0,1]} f(\xi_p),$$

so that $\alpha > 0$. Then by the mean value theorem, $\delta_\sigma(p) \geq \alpha\sigma$ for each $p \in [0,1]$. By Theorem 4.4.3 therefore, given $\epsilon > 0$, there exists $\sigma = \sigma(\epsilon) > 0$ such that, for sufficiently large n,

$$\mathbb{P}\big(d_2(\hat{F}_n, F)^2 > \epsilon\big) \leq 2ne^{-2n\alpha^2\sigma^2}.$$

Hence the result holds for any $\delta \in (0, 2\alpha^2\sigma^2)$. □

4.5 A Large Deviations Principle?

In the light of Section 4.4, it is natural to ask whether the sequence of empirical distribution functions $(\hat{F}_n)$ satisfies a large deviations principle (LDP) in the topology generated by the Mallows metric. The answer is in general negative, though it is true under certain conditions which are described in this section. First, we recall some standard definitions and results on large deviations, which may be found in Dembo and Zeitouni (1995).

Definition 4.5.1. *Let $\mathcal{X}$ be a topological space. A function $I : \mathcal{X} \to [0, \infty]$ is called a rate function if it is lower semi-continuous; that is, if for each $\alpha \in [0, \infty)$, the level set $\{x \in \mathcal{X} : I(x) \leq \alpha\}$ is closed. A good rate function is a rate function for which the level sets are compact subsets of $\mathcal{X}$.*

Let $\bar{A}$ and A° denote the closure and interior, respectively, of a set A, and let $\mathcal{B}$ denote the Borel σ-algebra of $\mathcal{X}$. Write $M(\mathcal{X})$ for the space of probability measures on $\mathcal{X}$.

Definition 4.5.2. *A sequence of probability measures (μ_n) on $(\mathcal{X}, \mathcal{B})$ satisfies an LDP with rate function I if, for all $A \subseteq \mathcal{B}$,*

$$-\inf_{x \in A^\circ} I(x) \leq \liminf_{n \to \infty} \frac{1}{n} \log \mu_n(A) \leq \limsup_{n \to \infty} \frac{1}{n} \log \mu_n(A) \leq -\inf_{x \in \bar{A}} I(x).$$

Remark: If $\mathcal{X} = \mathbb{R}$, and (μ_n) satisfies an LDP, we may also say that (F_n) satisfies an LDP, where F_n is the distribution function corresponding to μ_n. Similarly, if (X_n)

is a sequence of random elements of $\mathcal{X}$ such that X_n is distributed according to μ_n, we may also say (X_n) satisfies an LDP in $\mathcal{X}$.

If μ and ν are probability measures, we write $\nu \ll \mu$ if ν is absolutely continuous with respect to μ. In this case, we also write $d\nu/d\mu$ for the Radon-Nikodym derivative of ν with respect to μ. If $\phi : \mathcal{X} \to \mathbb{R}$ is a bounded, continuous function, $x \in \mathbb{R}$ and $\delta > 0$, define an open set in $M(\mathcal{X})$ by

$$U_{\phi,x,\delta} = \Big\{\nu \in M(\mathcal{X}) : \Big|\int_{\mathcal{X}} \phi(y)\, d\nu(y) - x\Big| < \delta\Big\}.$$

The collection $\{U_{\phi,x,\delta}\}$ generates the weak topology, and the Borel σ-algebra in $M(\mathcal{X})$, equipped with the weak topology, is the σ-algebra generated by the open sets in the weak topology.

Theorem 4.5.3 (Sanov's Theorem). *Let $\mathcal{X}$ be a complete, separable metric space, and let μ be a probability measure on $\mathcal{X}$. If $X_1, \ldots, X_n$ are independent and identically distributed according to μ, and $\hat{\mu}_n$ denotes their empirical measure, then the sequence $(\hat{\mu}_n)$ satisfies an LDP in $M(\mathcal{X})$, equipped with the weak topology, with good rate function*

$$I(\nu) = \begin{cases} \int_{-\infty}^{\infty} \frac{d\nu}{d\mu}(x) \log\big(\frac{d\nu}{d\mu}(x)\big)\, d\mu(x) & \text{if } \nu \ll \mu \\ \infty & \text{otherwise.} \end{cases}$$

Proposition 4.5.4 (Contraction Principle). *Let $\mathcal{X}$ and $\mathcal{Y}$ be Hausdorff spaces, and let $f : \mathcal{X} \to \mathcal{Y}$ be a continuous function. If (X_n) satisfies an LDP in $\mathcal{X}$ with good rate function I, then $\big(f(X_n)\big)$ satisfies an LDP in $\mathcal{Y}$ with good rate function*

$$J(y) = \inf\{I(x) : f(x) = y\}.$$

Proposition 4.5.5 (Inverse Contraction Principle). *Let $\mathcal{X}$ and $\mathcal{Y}$ be Hausdorff spaces, and let $f : \mathcal{X} \to \mathcal{Y}$ be a continuous bijection. Suppose (X_n) is a sequence of random elements of $\mathcal{X}$ such that the following two conditions hold:*

(a) $\big(f(X_n)\big)$ *satisfies an LDP in* $\mathcal{Y}$ *with rate function* $J(y)$

(b) for all $\alpha \in [0,\infty)$ *there exists a compact set* K_α *in* $\mathcal{X}$ *such that*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(X_n \notin K_\alpha) \leq -\alpha.$$

Then (X_n) *satisfies an LDP in* $\mathcal{X}$ *with good rate function* $I(x) = J\big(f(x)\big)$.

Remark: Condition (b) is usually known as exponential tightness. Since $(\mathcal{F}_2, d_2)$ is complete and separable, it follows from Lemma 2.6 of Lynch and Sethuraman (1987), that exponential tightness is a necessary condition for the sequence of empirical distribution functions $(\hat{F}_n)$ to satisfy an LDP with a good rate function.

In trying to strengthen the topology in which we hope an LDP will hold, we therefore first need to characterise the compact sets in $(\mathcal{F}_2, d_2)$. This task is complicated by the following lemma, whose proof is given in Section 4.6.

Lemma 4.5.6. *For each* $r \geq 1$, *each* $F \in \mathcal{F}_r$ *and every* $\epsilon > 0$, *the closed ball* $\bar{B}(F,\epsilon) = \{G \in \mathcal{F}_r : d_r(F,G) \leq \epsilon\}$ *is not compact.*

Nevertheless, the next two lemmas, whose proofs are also deferred to Section 4.6, provide enough compact sets to study exponential tightness in $(\mathcal{F}_2, d_2)$. In fact, we work with compact sets in $(\mathcal{G}_2, \|\cdot\|_2)$ for convenience. We let $\mathcal{H}$ denote the set of pairs (H_1, H_2) of functions $H_1 : (0,1) \to [0,\infty)$ and $H_2 : (0,1) \to [0,\infty)$ such that $H_1(\epsilon) \to 0$ as $\epsilon \to 0$, and $H_2(1-\epsilon) \to 0$ as $\epsilon \to 0$.

Lemma 4.5.7. *For* $(H_1, H_2) \in \mathcal{H}$, *let*

$$K_{H_1,H_2} = \Big\{G \in \mathcal{G}_2 : \int_0^\epsilon G^2(p)\,dp \leq H_1^2(\epsilon) \text{ and } \int_{1-\epsilon}^1 G^2(p)\,dp \leq H_2^2(1-\epsilon)\ \forall \epsilon \in (0,1)\Big\}.$$

Then K_{H_1,H_2} *is compact in* $(\mathcal{G}_2, \|\cdot\|_2)$.

Lemma 4.5.8. *If K is a compact subset of $(\mathcal{G}_2, \|\cdot\|_2)$, then there exists a pair $(H_1, H_2) \in \mathcal{H}$ such that $K \subseteq K_{H_1,H_2}$.*

Theorem 4.5.9. *If F has bounded support, then the sequence $(\hat{F}_n)$ of empirical distribution functions is exponentially tight in $(\mathcal{F}_2, d_2)$.*

Proof. It suffices to find a pair $(H_1, H_2) \in \mathcal{H}$ such that $\mathbb{P}(\hat{F}_n^{-1} \notin K_{H_1,H_2}) = 0$ for all $n \in \mathbb{N}$. If F has bounded support, then $\xi_0 = \lim_{p\searrow 0} \xi_p$ and $\xi_1 = \lim_{p\nearrow 1} \xi_p$ exist in $\mathbb{R}$. For $\epsilon \in (0,1)$, let

$$H_1(\epsilon) = \epsilon^{1/2} \max(|\xi_0|, |\xi_1|) \quad \text{and} \quad H_2(\epsilon) = H_1(1-\epsilon).$$

Then $(H_1, H_2) \in \mathcal{H}$, and

$$\int_0^{\epsilon} \big(\hat{F}_n^{-1}(p)\big)^2 \, dp \le \epsilon \max(\xi_0^2, \xi_1^2) = H_1^2(\epsilon),$$

and similarly $\int_{1-\epsilon}^1 \big(\hat{F}_n^{-1}(p)\big)^2 \, dp \le H_2^2(1-\epsilon)$. Hence $\mathbb{P}(\hat{F}_n^{-1} \notin K_{H_1,H_2}) = 0$ for all $n \in \mathbb{N}$. □

Corollary 4.5.10. *If F has bounded support, then $d_2\big(H_n(\hat{F}_n), H_n(F)\big)$ satisfies the large deviations upper bound for semi-infinite intervals with a good rate function. In other words, there exists a good rate function I such that, for every $\epsilon > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big\{d_2\big(H_n(\hat{F}_n), H_n(F)\big) \ge \epsilon\big\} \le -\inf_{x \ge \epsilon} I(x).$$

Proof. By Sanov's theorem, Theorem 4.5.9 and the Inverse Contraction Principle, the sequence $(\hat{F}_n)$ satisfies an LDP in $(\mathcal{F}_2, d_2)$ with good rate function

$$I_1(F') = \begin{cases} \int_{-\infty}^{\infty} \frac{d\mu_{F'}}{d\mu_F}(x) \log\Big(\frac{d\mu_{F'}}{d\mu_F}(x)\Big) \, dF(x) & \text{if } \mu_{F'} \ll \mu_F \\ \infty & \text{otherwise,} \end{cases}$$

where μ_F and $\mu_{F'}$ are the probability measures corresponding to the distribution functions F and F' respectively. Since the function $\psi : (\mathcal{F}_2, d_2) \to \mathbb{R}$ defined by

$\psi(F') = d_2(F', F)$ is continuous, the Contraction Principle implies that $d_2(\hat{F}_n, F)$ satisfies an LDP in $\mathbb{R}$ with good rate function

$$I(x) = \inf\{I_1(F') : F' \in \mathcal{F}_2,\ d_2(F', F) = x\}.$$

Since $d_2\big(H_n(\hat{F}_n), H_n(F)\big)$ is stochastically dominated by $d_2(\hat{F}_n, F)$, the result for $d_2\big(H_n(\hat{F}_n), H_n(F)\big)$ follows. □

We say that $F \in \mathcal{F}_2$ has a polynomial tail if there exists an $m > 2$ such that $x^m\big(1 - F(x)\big) \to \infty$ as $x \to \infty$, or $|x|^m F(x) \to \infty$ as $x \to -\infty$.

Theorem 4.5.11. *If $F \in \mathcal{F}_2$ has a polynomial tail, then the sequence $(\hat{F}_n)$ of empirical distribution functions is not exponentially tight in $(\mathcal{F}_2, d_2)$.*

Proof. It suffices to show that for any $(H_1, H_2) \in \mathcal{H}$, we have

$$\mathbb{P}(\hat{F}_n^{-1} \notin K_{H_1,H_2}) \ge e^{-n}$$

for sufficiently large $n \in \mathbb{N}$. Now, we may assume without loss of generality that $x^m\big(1 - F(x)\big) \to \infty$ as $x \to \infty$, for some $m > 2$. Let $X_{(n)} = \max_{1 \le i \le n} X_i$. Then

$$\mathbb{P}(\hat{F}_n^{-1} \notin K_{H_1,H_2}) \ge \mathbb{P}\big(X_{(n)} > n^{1/2} H_2(1 - 1/n)\big) = 1 - F\big(n^{1/2} H_2(1 - 1/n)\big)^n.$$

Choose $n_0 \in \mathbb{N}$ large enough such that $H_2(1 - 1/n) \le 1$ and $1 - F(n^{1/2}) \ge n^{-m/2}$ for all $n \ge n_0$. Then, for $n \ge n_0$,

$$\mathbb{P}(\hat{F}_n^{-1} \notin K_{H_1,H_2}) \ge 1 - F(n^{1/2})^n \ge 1 - \Big(1 - \frac{1}{n^{m/2}}\Big)^n \ge 1 - \exp(-n^{-(m/2-1)}).$$

But, for sufficiently large n,

$$1 - \exp(-n^{-(m/2-1)}) \ge \frac{1}{2n^{m/2-1}} \ge e^{-n}.$$

□

Remark: In view of the remark following the statement of the Inverse Contraction principle (Proposition 4.5.5), Theorem 4.5.11 shows that the sequence $(\hat{F}_n)$ does not satisfy an LDP in $(\mathcal{F}_2, d_2)$.

4.6 Appendix

Proof of **Lemma 4.5.6**.

We prove that the closed ball is not sequentially compact. Fix $\epsilon > 0$, $F \in \mathcal{F}_r$, and, for $p \in (0,1)$, let $\xi_p = \inf\{x \in \mathbb{R} : F(x) \geq p\}$. Consider the sequence of distribution functions (F_n) given by

$$F_n(x) = \begin{cases} F(x) & \text{if } x < \xi_{1-1/n} \\ 1 - 1/n & \text{if } \xi_{1-1/n} \leq x < \xi_{1-1/n} + \epsilon n^{1/r} \\ F(x - \epsilon n^{1/r}) & \text{if } x \geq \xi_{1-1/n} + \epsilon n^{1/r}. \end{cases}$$

Thus

$$F_n^{-1}(p) = \begin{cases} \xi_p & \text{if } p \leq 1 - 1/n \\ \xi_p + \epsilon n^{1/r} & \text{if } p > 1 - 1/n. \end{cases}$$

It follows that

$$d_r(F_n, F) = \left(\int_0^1 |F_n^{-1}(p) - \xi_p|^r \, dp \right)^{1/r} = \left(\int_{1-1/n}^1 \epsilon^r n \, dp \right)^{1/r} = \epsilon.$$

On the other hand, we have $|F_n(x) - F(x)| \leq 1/n$ for all $x \in \mathbb{R}$ and $n \in \mathbb{N}$, so if a subsequence (F_{n_k}) satisfied $d_r(F_{n_k}, G) \to 0$ as $k \to \infty$, then we would have to have $G \equiv F$. Since $d_r(F_{n_k}, F) = \epsilon$ for all $k \in \mathbb{N}$, no convergent subsequence can exist. □

Proof of **Lemma 4.5.7**.

Take a sequence $(G_n) \in K_{H_1,H_2}$. For each $m \in \mathbb{N}$, choose $\epsilon_m \in (0, 1/2)$ such that $H_1(\epsilon) \leq 1/m$ and $H_2(1-\epsilon) \leq 1/m$ for each $\epsilon \in (0, \epsilon_m]$. We claim that there exists an infinite subset N_1 of $\mathbb{N}$ such that

$$\int_{\epsilon_1}^{1-\epsilon_1} \bigl(G_{n_1}(p) - G_{n_2}(p)\bigr)^2 \, dp \leq 1$$

for all $n_1, n_2 \in N_1$. To see why this is the case, observe first that $G(\epsilon_1) \geq -H_1(\epsilon_1)/\epsilon_1^{1/2}$ and $G(1-\epsilon_1) \leq H_2(1-\epsilon_1)/\epsilon_1^{1/2}$ for each $G \in K_{H_1,H_2}$. Now we may partition the interval $[\epsilon_1, 1-\epsilon_1]$ into d equally spaced divisions and note that given any $\delta > 0$, there exist $-H_1(\epsilon_1)/\epsilon_1^{1/2} \leq x_0 \leq x_1 \leq \ldots \leq x_d \leq H_2(1-\epsilon_1)/\epsilon_1^{1/2}$ and an infinite subset N_1 of $\mathbb{N}$ such that

$$\Bigl|G_n\Bigl(\epsilon_1 + \frac{i}{d}(1-2\epsilon_1)\Bigr) - x_i\Bigr| \leq \delta$$

for all $i = 0, 1, \ldots, d$ and $n \in N_1$. For $i = 0, 1, \ldots, d$, let $p_i = \epsilon_1 + i(1-2\epsilon_1)/d$. Then, for $n_1, n_2 \in N_1$,

$$\begin{aligned}
\int_{\epsilon_1}^{1-\epsilon_1} & \bigl(G_{n_1}(p) - G_{n_2}(p)\bigr)^2 \, dp \\
&= \sum_{i=1}^{d} \int_{p_{i-1}}^{p_i} \bigl(G_{n_1}(p) - G_{n_2}(p)\bigr)^2 \, dp \\
&\leq \frac{1}{d} \sum_{i=1}^{d} \bigl(x_i - x_{i-1} + 2\delta\bigr)^2 \\
&\leq \frac{1}{d} \Bigl\{ \Bigl(\frac{H_2(1-\epsilon_1)}{\epsilon_1^{1/2}} + \frac{H_1(\epsilon_1)}{\epsilon_1^{1/2}}\Bigr)^2 + 4\delta \Bigl(\frac{H_2(1-\epsilon_1)}{\epsilon_1^{1/2}} + \frac{H_1(\epsilon_1)}{\epsilon_1^{1/2}}\Bigr) + 4\delta^2 \Bigr\} \\
&\leq 1,
\end{aligned}$$

for sufficiently small $\delta > 0$, and sufficiently large $d \in \mathbb{N}$.

In a similar manner, we may find a sequence of infinite subsets (N_m) of $\mathbb{N}$ with $N_1 \supseteq N_2 \supseteq \ldots$, such that for each $m \in \mathbb{N}$,

$$\int_{\epsilon_m}^{1-\epsilon_m} \bigl(G_{n_1}(p) - G_{n_2}(p)\bigr)^2 \, dp \leq \frac{1}{m}$$

for all $n_1, n_2 \in N_m$. Now construct the diagonal subsequence (n_k), by taking n_k to be the kth smallest element of N_k, for each $k \in \mathbb{N}$. The sequence (G_{n_k}) is a subsequence of the original sequence (G_n), and is a Cauchy sequence in $(\mathcal{G}_2, \|\cdot\|_2)$ because, for $k \le l$,

$$\begin{aligned}\int_0^1 \bigl(G_{n_k}(p) - G_{n_l}(p)\bigr)^2\,dp &\\ \le 2\int_0^{\epsilon_k} \bigl(G_{n_k}^2(p) + G_{n_l}^2(p)\bigr)\,dp &+ \int_{\epsilon_k}^{1-\epsilon_k} \bigl(G_{n_k}(p) - G_{n_l}(p)\bigr)^2\,dp \\ &\qquad + 2\int_{1-\epsilon_k}^{1} \bigl(G_{n_k}^2(p) + G_{n_l}^2(p)\bigr)\,dp \\ \le 4\bigl(H_1^2(\epsilon_k) + H_2^2(1-\epsilon_k)\bigr) + \frac{1}{k} & \\ \to 0 &\end{aligned}$$

as $k \to \infty$. But $(\mathcal{G}_2, \|\cdot\|_2)$ is complete, so (G_{n_k}) converges in $(\mathcal{G}_2, \|\cdot\|_2)$, so K_{H_1,H_2} is sequentially compact. □

Proof of **Lemma 4.5.8**.

Suppose the lemma is false. Then without loss of generality, we may assume there exist $\epsilon > 0$ and a sequence $(G_n) \in K$ such that

$$\int_0^{1/n} G_n^2(p)\,dp \ge \epsilon.$$

Since K is compact, there exist a strictly increasing sequence $(n_k) \in \mathbb{N}$, $G \in K$ and $k_0 \in \mathbb{N}$ such that

$$\int_0^1 \bigl(G_{n_k}(p) - G(p)\bigr)^2\,dp \le \frac{\epsilon}{4}$$

for all $k \ge k_0$. By restricting attention to a further subsequence if necessary, we may assume

$$\int_{1/n_{k+1}}^{1/n_k} G_{n_k}^2(p)\,dp \ge \frac{\epsilon}{2}$$

for each $k \in \mathbb{N}$. But then, by Minkowski's inequality,

$$\begin{aligned}
\left(\int_0^1 G^2(p)\,dp\right) &\geq \sum_{k=k_0}^{\infty} \int_{1/n_{k+1}}^{1/n_k} G^2(p)\,dp \\
&\geq \sum_{k=k_0}^{\infty} \left(\int_{1/n_{k+1}}^{1/n_k} G_{n_k}^2(p)\,dp - \int_{1/n_{k+1}}^{1/n_k} \bigl(G_{n_k}(p) - G(p)\bigr)^2\,dp\right)^2 \\
&\geq \sum_{k=k_0}^{\infty} \left(\frac{\epsilon}{2} - \frac{\epsilon}{4}\right)^2 \\
&= \infty,
\end{aligned}$$

which contradicts the fact that $G \in (\mathcal{G}_2, \|\cdot\|_2)$. □