Introduction to Asymptotic Theory - Part I

Stéphane Guerrier, Mucyo Karemera & Samuel Orso

Data Analytics Lab
University of Geneva

This document was prepared with the help of
Gaetan Bakalli (U. Geneva), Dr. Roberto Molinari (Penn State), Lionel Voirol (U. Geneva)
Haotian Xu (U. Geneva) & Yuming Zhang (U. Geneva)

November 26, 2019

Statistical Estimators

In this course, we present an introduction to the study of the properties of statistical estimators. We will mainly focus on asymptotic properties and start by considering a general class of estimators.

Definition 1.1 (Extremum Estimator).

Many estimators have a **common structure**, which is often useful to study their asymptotic properties. One structure or framework is the class of estimators that maximize some objective function, referred to as extremum estimators, which can be defined as follows:

$$\hat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \ \widehat{Q}_n(\boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\Theta}$ denote, respectively, the parameter vector of interest and its set of possible values.

Remark:

The vast majority of statistical estimators can be represented as extremum estimator. This is for example for the case for least squares, maximum likelihood or (generalized) method of moment estimators.

Example: Least Squares Estimator

Consider the linear model $\boldsymbol{y} = \mathbf{X}\beta_0 + \boldsymbol{\varepsilon}$ where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a full-ranked constant matrix, $[\boldsymbol{\varepsilon}]_i \overset{iid}{\sim} (0, \sigma_\varepsilon^2)$ and $\beta \in \mathcal{B} \subseteq \mathbb{R}^p$. Let $\hat{\beta}$ denote the Least Squares Estimator (LSE) of β_0, i.e.

$$\hat{\beta} := \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{y}.$$

Moreover, this estimator is an extremum estimator since it can expressed as:

$$\hat{\beta} = \underset{\beta \in \mathcal{B}}{\operatorname{argmax}} \; -||\boldsymbol{y} - \mathbf{X}\beta||_2^2,$$

similarly to our definition given in (1). ●

Example: Maximum Likelihood Estimator

Let $Z_1, \ldots, Z_n$ be an iid sample with pdf $f(Z|\boldsymbol{\theta}_0)$. The Maximum Likelihood Estimator (MLE) is given by

$$\hat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \; \frac{1}{n} \sum_{i=1}^{n} \log \left[f\left(z_i | \boldsymbol{\theta} \right) \right]. \tag{2}$$

Therefore, the MLE can be seen as an example of extremum estimator with

$$\widehat{Q}_n(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \log \left[f\left(z_i | \boldsymbol{\theta} \right) \right].$$

●

Remark: In (2) we are actually using a *normalized* log-likelihood instead of the actual log-likelihood. This has (in the vast majority of cases) no impact on the estimator but the normalized form is more convenient to use when we let $n \to \infty$.

Example: Generalized Method of Moment

Consider the same iid sample as in the previous example and suppose that there is a “moment function” vector $\mathbf{g}(Z|\boldsymbol{\theta})$ such that $\mathbb{E}[\mathbf{g}(Z|\boldsymbol{\theta}_0)] = 0$. Then, a possible estimator of $\boldsymbol{\theta}_0$ is

$$\hat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argmax}} \; - \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{g}(z_i|\boldsymbol{\theta}) \right]^T \widehat{\mathbf{W}} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{g}(z_i|\boldsymbol{\theta}) \right], \tag{3}$$

where $\widehat{\mathbf{W}}$ is an positive definite matrix of appropriate dimension. Such estimators are called Generalized Method of Moments (GMM) estimators. They belong to the class of extremum estimators. ●

Remark: Instead of (3) we will often consider an alternative (but equivalent) definition of such estimator, i.e.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argmax}} \; -||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\theta})||^2_{\widehat{\mathbf{W}}},$$

where $||\mathbf{x}||^2_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{x}$, and where $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}(\boldsymbol{\theta})$ denote, respectively, the empirical and model based moments.

Example: A simple GMM Estimator

Let $Z_i \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma_0^2)$ and $\boldsymbol{\theta}_0 = (\mu_0, \sigma_0^2)^T$. Suppose we wish to estimate $\boldsymbol{\theta}_0$ by matching the first three empirical moments with their theoretical counterparts. In this case, a reasonable moment function or condition defining a GMM estimator is given by:

$$\mathbf{g}(Z|\boldsymbol{\theta}) := \begin{bmatrix} Z - \mu \\ Z^2 - \left(\mu^2 + \sigma^2\right) \\ Z^3 - \left(\mu^3 + 3\mu\sigma^2\right) \end{bmatrix}.$$

However, it can be noticed that $\frac{1}{n}\sum_{i=1}^n \mathbf{g}(z_i|\boldsymbol{\theta}) = \hat{\gamma} - \gamma(\boldsymbol{\theta})$, where $\hat{\gamma}$ and $\gamma(\boldsymbol{\theta})$ denote, respectively, the empirical and model-based moments, i.e.

$$\hat{\gamma} := \frac{1}{n}\sum_{i=1}^n \begin{bmatrix} z_i \\ z_i^2 \\ z_i^3 \end{bmatrix}, \quad \gamma(\boldsymbol{\theta}) := \begin{bmatrix} \mu \\ \mu^2 + \sigma^2 \\ \mu^3 + 3\mu\sigma^2 \end{bmatrix}.$$

Then, we can write our GMM estimator of $\boldsymbol{\theta}_0$ as:

$$\hat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}}\ ||\hat{\gamma} - \gamma(\boldsymbol{\theta})||_{\widehat{\mathbf{W}}}^2 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\ -||\hat{\gamma} - \gamma(\boldsymbol{\theta})||_{\widehat{\mathbf{W}}}^2. \tag{4}$$

Consistency of Statistical Estimators

In the next slides we will discuss the conditions for the consistency of extremum estimator, which is often denoted as $\hat{\boldsymbol{\theta}} \xrightarrow{\mathcal{P}} \boldsymbol{\theta}_0$. We start by defining consistency.

Definition 1.2 (Consistency).

The estimator $\hat{\boldsymbol{\theta}}$ is said to be consistent if it converges in probability to $\boldsymbol{\theta}_0$, i.e. for all $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr\left(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_2 \geq \varepsilon\right) = 0.$$

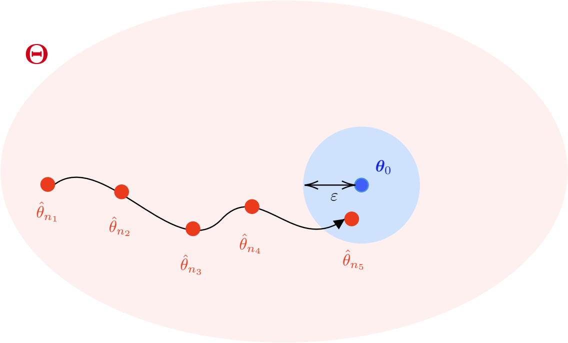
More precisely, by the definition of the limit, this means that the estimator $\hat{\boldsymbol{\theta}}$ is consistent if $\forall \varepsilon > 0$, $\forall \delta > 0$, $\exists n^* \geq 0$ such that $\forall n \geq n^*$ we have

$$\Pr\left(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_2 \geq \varepsilon\right) \leq \delta.$$

Consistency - Interpretation

Interpretation:

In Layman's term consistency simply means that if n is "large enough" $\hat{\boldsymbol{\theta}}$ will be **arbitrarily close to θ_0** (i.e. inside of an hypersphere of radius ε centered at $\boldsymbol{\theta}_0$). This also means the procedure (i.e. our estimator) based on unlimited data will be able to identify the underlying truth (i.e. $\boldsymbol{\theta}_0$).



Remark: Norm Equivalency

In the previous definition, we used $|| \cdot ||_2$ norm. Actually, we could have used any other norm $|| \cdot ||_k$ of $\mathbb{R}^p$, where $k \in \mathbb{N}^* \cup \{\infty\}$. We recall that $|| \cdot ||_k$ is defined as

$$||x||_k = \begin{cases} \left(\sum_{i=1}^p |x_i|^k\right)^{1/k} & \text{if } k \in \mathbb{N}^*, \\ \sup\limits_{i=1,\ldots,p} |x_i| & \text{if } k = \infty. \end{cases}$$

This comes from the following theorem:

Theorem 1.3 (Norm equivalency in finite dimension).

*If $p < \infty$ then **all norms of $\mathbb{R}^p$ are equivalent**. In particular, for all $k \in \mathbb{N}^* \cup \{\infty\}$, there exist $C_1, C_2 > 0$ such that for all $x \in \mathbb{R}^p$ we have*

$$C_1||x||_k \leq ||x||_2 \leq C_2||x||_k.$$

Consistency equivalence

Theorem 1.3 leads to the following Corollary:

Corollary 1.4.

If $p < \infty$, then for all $k \in \mathbb{N}^* \cup \{\infty\}$

$$\lim_{n\to\infty} \Pr\left(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_2 \geq \varepsilon\right) = 0,$$

if and only if

$$\lim_{n\to\infty} \Pr\left(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_k \geq \varepsilon\right) = 0.$$

Remark 1.

In $\mathbb{R}^\infty$ the norm equivalency result does not hold true.

Related Theorems

Consistency is often proven using the following two important results (see e.g. Chapter 1 of DasGupta 2008).

Theorem 1.5 (Weak Law of Large Number).

Suppose X_i are iid random variables with finite mean μ (i.e. $\mathbb{E}[X_i] = \mu$) and finite variance. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then $\bar{X}_n \xrightarrow{\mathcal{P}} \mu$.

Theorem 1.6 (Continuous Mapping Theorem).

Suppose $Y_n \xrightarrow{\mathcal{P}} \mu$, then $g(Y_n) \xrightarrow{\mathcal{P}} g(\mu)$ if $g(\cdot)$ is a continuous function.

A simple example on the consistency is presented in Appendix A ▸ Go to Appendix A.

Consistency of Extremum Estimators

When considering real-life problems the approach based on Theorem 1.5 presented in Appendix A is in general not flexible enough and we generally rely on the results such Theorem 2.1 of Newey and McFadden 1994, which is presented below.

Theorem 1.7 (Consistency of Extremum Estimators).

If there is a function $Q_0(\boldsymbol{\theta})$ such that:

(C.1) *$Q_0(\boldsymbol{\theta})$ is uniquely maximized in $\boldsymbol{\theta}_0$,*

(C.2) *$\boldsymbol{\Theta}$ is compact*[a]*,*

(C.3) *$Q_0(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$,*

(C.4) *$\widehat{Q}_n(\boldsymbol{\theta})$ converges uniformly in probability to $Q_0(\boldsymbol{\theta})$*[b]*,*

then we have $\hat{\boldsymbol{\theta}} \xrightarrow{\mathcal{P}} \boldsymbol{\theta}$.

[a] Compact, in finite dimension, means that $\boldsymbol{\Theta}$ is both closed (i.e. containing all its limit points) and bounded (i.e. all its points are contained in a ball).

[b] $\widehat{Q}_n(\boldsymbol{\theta})$ is said to converges uniformly in probability to $Q_0(\boldsymbol{\theta})$ if $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \xrightarrow{\mathcal{P}} 0$

Remarks: Consistency of Extremum Estimators

This theorem is an important result, which provides a general approach to prove the consistency of a large class of estimators. A few remarks on the conditions of this result:

- Condition (C.1) is substantive and there are well-known examples where it fails. We will discuss further on how this assumption can (in some cases) be verified in practice.
- Condition (C.2) is also substantive as it requires that there exist some known bounds on the parameters. In practice, this assumption is often neglected although it is in most cases unrealistic to assume it.
- Conditions (C.3) and (C.4) are often referred to as "standard regularity conditions". They are typically satisfied. The verification of these conditions will be discussed further in this document.

Theorem 1.7: Sketch of the proof

The basic idea of the proof is the following. Under Condition (C.1) we have

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\; Q_0(\boldsymbol{\theta}).$$

Condition (C.4) implies that

$$\widehat{Q}_n(\boldsymbol{\theta}) \xrightarrow{\mathcal{P}} Q_0(\boldsymbol{\theta}),$$

therefore, it seems logical that

$$\hat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\; \widehat{Q}_n(\boldsymbol{\theta}) \xrightarrow{\mathcal{P}} \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\; Q_0(\boldsymbol{\theta}) = \boldsymbol{\theta}_0.$$

Unfortunately, without additional conditions this simple proof is not correct. The main reason is that additional conditions (i.e. uniform convergence as well as Conditions (C.2) and (C.3)) are needed to ensure the validity of the above convergence given in orange. A formal proof of this results is given in Appendix B and is of course good to know! ▸ Go to Theorem Appendix B

Consistency of Z-estimators

Estimators are not always expressed as extremum. We see such an example with Z-estimators.

Theorem 1.8 (Consistency of Z-estimators).

Let $\hat{\boldsymbol{\theta}} := \operatorname{argzero}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \hat{\mathbf{g}}_n(\boldsymbol{\theta})$*. If there exists a function* $\mathbf{g}_0(\boldsymbol{\theta})$ *such that:*

(C.1) $\mathbf{g}_0(\boldsymbol{\theta})$ *has a unique root in* $\boldsymbol{\Theta}$ *at* $\boldsymbol{\theta}_0$[a]

(C.2) $\boldsymbol{\Theta}$ *is compact*

(C.3) $\mathbf{g}_0(\boldsymbol{\theta})$ *is continuous in* $\boldsymbol{\theta}$,

(C.4) $\hat{\mathbf{g}}_n(\boldsymbol{\theta})$ *converges uniformly in probability to* $\mathbf{g}_0(\boldsymbol{\theta})$,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\hat{\mathbf{g}}_n(\boldsymbol{\theta}) - \mathbf{g}_0(\boldsymbol{\theta})\| \xrightarrow{\mathcal{P}} 0,$$

then we have $\hat{\boldsymbol{\theta}} \xrightarrow{\mathcal{P}} \boldsymbol{\theta}$.

[a] this means that $\mathbf{g}_0(\boldsymbol{\theta}) = 0$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Exercises:

1) Prove Theorem 1.5.
2) Prove Corollary 1.4.
3) In $\mathbb{R}^\infty$, are the $|| \; ||_\infty$-consistency and the $|| \; ||_2$-consistency equivalent?
4) Prove Theorem 1.6 using Definition 1.2.
5) Prove Theorem 1.8.

Exercises

6) Consider the following estimator:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argmax}} \; -||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\theta})||_{\mathbf{W}}^2,$$

where $||\mathbf{x}||_{\mathbf{A}}^2 = \mathbf{x}^T\mathbf{A}\mathbf{x}$, and where $\hat{\boldsymbol{\mu}} \in \mathbb{R}^p$, $\boldsymbol{\mu}(\boldsymbol{\theta}) \in \mathbb{R}^p$, $\mathbf{W} \in \mathbb{R}^{p \times p}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$. Note that the $\mathbf{W}$ may be random and/or depend on $\boldsymbol{\theta}$ but we assume that $\mathbf{W} > 0$.

1. Is it possible to show that

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argzero}} \; \hat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\theta})?$$

   If so, do we need any additional conditions?

2. Simplify the conditions of Theorem 1.7 (or 1.8) for this estimator.

References

DasGupta, Anirban (2008). *Asymptotic theory of statistics and probability*. Springer Science & Business Media.

Newey, W. K. and D. McFadden (1994). “Large Sample Estimation and Hypothesis Testing, V in Handbook of Econometrics”. In: vol. 4. Elsevier, Amsterdam.

Appendix A: Consistency - a simple example

Let $X_i \overset{iid}{\sim} \mathcal{E}(\lambda_0),\ \lambda_0 \in \mathbb{R}^+,\ i = 1, ..., n.$ We wish to show that the MLE for λ_0 is consistent. Then, we have that the density of X is given by (assuming $X \geq 0$):

$$f(X|\lambda) := \lambda \exp\left(-\lambda X\right).$$

Therefore, the normalized log-likelihood function is given by

$$\mathcal{L}(\lambda|X_1, ..., X_n) = \log(\lambda) - \lambda \bar{X}_n,$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. By taking the first derivative we obtain:

$$\frac{\partial}{\partial\lambda}\,\mathcal{L}(\lambda|X_1, ..., X_n) = \frac{1}{\lambda} - \bar{X}_n,$$

which implies that MLE is such that $\frac{\partial}{\partial\lambda}\,\mathcal{L}(\hat{\lambda}|X_1, ..., X_n) = 0$. We obtain

$$\frac{1}{\hat{\lambda}} - \bar{X}_n = 0 \implies \hat{\lambda} = \bar{X}_n^{-1}.$$

Appendix A: Consistency - a simple example

Finally, we verify that

$$\frac{\partial^2}{\partial \lambda^2} \, \mathcal{L}(\lambda | X_1, ..., X_n) = -\frac{1}{\lambda^2} < 0,$$

implying that $\hat{\lambda}$ is the maxima of $\mathcal{L}(\lambda|X_1, ..., X_n)$. Therefore, the MLE is a function of the sample mean $\bar{X}_n$. In this case, the consistency of $\hat{\lambda}$ is implied by Theorems 1.5 and 1.6. Indeed, it follows from the Weak Law of Large Number (i.e. Theorem 1.5) that $\bar{X}_n \xrightarrow{\mathcal{P}} \mu$, where μ is given by

$$\mu := \mathbb{E}[X_i] = \int_0^\infty x\lambda_0 \exp\left(-\lambda_0 x\right) dx = \frac{1}{\lambda_0}.$$

Since the function $f(x) = 1/x$ is continuous in $\mathbb{R}^+$ we obtain by the Continuous Mapping Theorem (i.e. Theorem 1.6) that $\hat{\lambda} \xrightarrow{\mathcal{P}} \lambda_0$, which concludes our example. ▸ Return to Theorem 1.5 ●

Appendix B: Proof of Theorem 1.7

Let $\mathcal{G}$ be the ε-ball centered at $\boldsymbol{\theta}_0$ i.e. $\mathcal{G} = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||_2 < \varepsilon\}$ for some $\varepsilon > 0$. Then $\hat{\boldsymbol{\theta}} \xrightarrow{\mathcal{P}} \boldsymbol{\theta}_0$ is equivalent to

$$\lim_{n\to\infty} \Pr(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_2 \geq \varepsilon) = \lim_{n\to\infty} \Pr(\hat{\boldsymbol{\theta}} \notin \mathcal{G}) = 0.$$

We define $\gamma = Q_0(\boldsymbol{\theta}_0) - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\mathcal{G}} \; Q_0(\boldsymbol{\theta})$ which is strictly positive by 1.7. Then we have that $\hat{\boldsymbol{\theta}} \notin \mathcal{G}$ implies

$$Q_0(\hat{\boldsymbol{\theta}}) \leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\mathcal{G}} \; Q_0(\boldsymbol{\theta}) = Q_0(\boldsymbol{\theta}_0) - \gamma$$

and therefore

$$\lim_{n\to\infty} \Pr(\hat{\boldsymbol{\theta}} \notin \mathcal{G}) \leq \lim_{n\to\infty} \Pr(\mathcal{A})$$

where $\mathcal{A} = \left\{ Q_0(\hat{\boldsymbol{\theta}}) \leq Q_0(\boldsymbol{\theta}_0) - \gamma \right\}$.

Appendix B: Proof of Theorem 1.7

Next, we define the following events:

$$\begin{aligned} \mathcal{B} &= \left\{ \left| \widehat{Q}_n(\hat{\boldsymbol{\theta}}) - Q_0(\hat{\boldsymbol{\theta}}) \right| > \gamma/3 \right\} \\ \mathcal{C} &= \left\{ \left| \widehat{Q}_n(\boldsymbol{\theta}_0) - Q_0(\boldsymbol{\theta}_0) \right| > \gamma/3 \right\} \\ \mathcal{D} &= \left\{ \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \widehat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta}) \right| > \gamma/3 \right\} \end{aligned}$$

and we have that

$$\begin{aligned} \Pr(\mathcal{A}) \leq \Pr(\mathcal{A} \cup (\mathcal{B} \cup \mathcal{C})) &= \Pr\left(\mathcal{A} \cap (\mathcal{B} \cup \mathcal{C})^c\right) + \Pr(\mathcal{B} \cup \mathcal{C}) \\ &= \Pr(\mathcal{A} \cap \mathcal{B}^c \cap \mathcal{C}^c) + \Pr(\mathcal{B} \cup \mathcal{C}). \end{aligned}$$

Appendix B: Proof of Theorem 1.7

It is easy to verify that $\Pr\left(\mathcal{A} \cap \mathcal{B}^c \cap \mathcal{C}^c\right) = \Pr(\varnothing)$ because if $\mathcal{A}$, $\mathcal{B}^c$ and $\mathcal{C}^c$ occur simultaneously we have that

$$\begin{aligned}\widehat{Q}_n(\hat{\boldsymbol{\theta}}) &= Q_0(\hat{\boldsymbol{\theta}}) + \widehat{Q}_n(\hat{\boldsymbol{\theta}}) - Q_0(\hat{\boldsymbol{\theta}}) \leq Q_0(\hat{\boldsymbol{\theta}}) + |\widehat{Q}_n(\hat{\boldsymbol{\theta}}) - Q_0(\hat{\boldsymbol{\theta}})| \\ &\leq Q_0(\hat{\boldsymbol{\theta}}) + \gamma/3 \leq Q_0(\boldsymbol{\theta}_0) - 2\gamma/3 = Q_0(\boldsymbol{\theta}_0) - \widehat{Q}_n(\boldsymbol{\theta}_0) + \widehat{Q}_n(\boldsymbol{\theta}_0) - 2\gamma/3 \\ &\leq |Q_0(\boldsymbol{\theta}_0) - \widehat{Q}_n(\boldsymbol{\theta}_0)| + \widehat{Q}_n(\boldsymbol{\theta}_0) - 2\gamma/3 \leq \widehat{Q}_n(\boldsymbol{\theta}_0) - \gamma/3 < \widehat{Q}_n(\boldsymbol{\theta}_0)\end{aligned}$$

which contradicts (1). Moreover, the probability $\Pr\left(\mathcal{B} \cup \mathcal{C}\right)$ can be bounded as follow

$$\begin{aligned}\Pr\left(\mathcal{B} \cup \mathcal{C}\right) &= \Pr\left(\sup_{\boldsymbol{\theta} \in \left\{\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}\right\}} \left|\widehat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})\right| > \gamma/3\right) \\ &\leq \Pr\left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left|\widehat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})\right| > \gamma/3\right) = \Pr\left(\mathcal{D}\right).\end{aligned}$$

Appendix B: Proof of Theorem 1.7

Finally, we have that

$$\begin{aligned}\lim_{n\to\infty} \Pr(\hat{\boldsymbol{\theta}} \notin \mathcal{G}) &\leq \lim_{n\to\infty} \Pr(\mathcal{A}) \leq \lim_{n\to\infty} \Pr\left(\mathcal{A} \cap \mathcal{B}^c \cap \mathcal{C}^c\right) + \Pr\left(\mathcal{B} \cup \mathcal{C}\right) \\ &= \Pr(\varnothing) + \lim_{n\to\infty} \Pr\left(\mathcal{B} \cup \mathcal{C}\right) \leq \lim_{n\to\infty} \Pr(\mathcal{D}) = 0,\end{aligned}$$

which conclusively concludes the proof. ■

‣ Return to Sketch of the proof