The Empirical Process in SLT

Rohan Hitchcock

Introduction

The Empirical Process in SLT

- The empirical process $\psi_n(w)$ controls the difference between the true and empirical KL-divergence of a model.
- Recall $K(w) = \mathbf{E}[f(X, w)]$ and $K_n(w) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} f(X_i, w)$ where $f(x, w) = -\log \dfrac{p(x\,|\,w)}{q(x)}$.
- We define $\psi_n(w) = \dfrac{1}{\sqrt{n}} \sum_{i=1}^{n} \dfrac{K(w) - f(X_i, w)}{\sqrt{K(w)}}$ so that:

$$K_n(w) = K(w) - \sqrt{K(w)/n}\,\psi_n(w)$$

- Also consider an empirical process $\xi_n(u)$ on the resolution:

$$K_n(g(u)) = u^{2k} - \frac{1}{\sqrt{n}} u^k \xi_n(u)$$

- Proofs of (e.g.) the Free Energy Formula require establishing some kind of convergence $\psi_n(w) \to \psi(w)$ and $\xi_n(u) \to \xi(u)$.

Goal for this talk

- Understand convergence of sequences of random functions like $\psi_n(w)$ and $\xi_n(u)$.
- This is the main focus of the Grey Book Chapter 5.
- I am to clarify some confusing aspects of that chapter.
- We are focused on real-valued functions on compact spaces with the supremum norm topology.

Set-up

- Consider continuous $f : \mathbb{R}^N \times W \to \mathbb{R}$, where $W \subseteq \mathbb{R}^d$ is compact.
- Fix a probability distribution $q(x)$ on $\mathbb{R}^N$ and iid $X_1, X_2, \ldots \sim q(x)$.
- Define a sequence of (random) functions

$$\psi_n(w) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(f(X_i, w) - \mathbf{E} f(X_i, w) \right)$$

Goal: Prove, in some suitable sense, "$\psi_n(w) \to \psi(w)$ as $n \to \infty$", where $\psi(w)$ is continuous.

Background

Function-valued random variables

- We are concerned with **function-valued random variables**.
- Two measurable spaces:
  - All functions $W \to \mathbb{R}$, denoted $\mathbb{R}^W$, with the *cylindrical σ-algebra* $\mathscr{B}_{\text{cyl}}$
    - ($\mathscr{B}_{\text{cyl}}$ is the smallest σ-algebra s.t. all projections $\pi : \mathbb{R}^W \to \mathbb{R}^F$, $F \subseteq W$ finite are measurable).
    - $\mathscr{B}_{\text{cyl}}$ is "coarse": it is "easy" to find distributions on $(\mathbb{R}^W, \mathscr{B}_{\text{cyl}})$, but $\mathscr{B}_{\text{cyl}}$ doesn't specify many interesting properties.
  - Continuous functions $W \to \mathbb{R}$, denoted $C = Cts(W, \mathbb{R})$, with the *Borel σ-algebra* $\mathscr{B}$
    - ($\mathscr{B}$ is the smallest σ-algebra s.t. all open sets of C are measurable).
    - $\mathscr{B}$ is "compatible with the sup-norm topology on C"

Two important theorems

Kolmogorov's Extension Theorem: Given probability measures Ψ_F on each $\mathbb{R}^F$ where $F \subseteq W$ is finite, can be used to prove that there exists a measure Ψ on $(\mathbb{R}^W, \mathscr{B}_{\text{cyl}})$ such that the pushforward to any $\mathbb{R}^F$ is Ψ_F.

- To construct a random function ψ:
  - Define the intended joint distribution of $(\psi(w_1), \ldots, \psi(w_k))$ for all finite subsets $\{w_1, \ldots, w_k\} \subseteq W$.
  - Kolmogorov gives you a random variable ψ on $(\mathbb{R}^W, \mathscr{B}_{\text{cyl}})$ with the desired finite projections.
  - See Grey Book, Remark 5.6.

Two important theorems

Prohorov's Theorem: Given a sequence Ψ_n of probability measures on $(C, \mathscr{B})$, can be used to show convergence in distribution $\Psi_n \to \Psi$ to another probability measure Ψ on $(C, \mathscr{B})$.

- To construct a random function ψ:
  - Define continuous random functions ψ_n.
  - Show the sequence converges.

Aside: Constructing Brownian motion

- Want a continuous random function $B : [0,\infty) \to \mathbb{R}$ with independent normal increments $B_{t+\Delta t} - B_t \sim \mathcal{N}(0,\Delta t)$.
- **Kolmogorov's Extension Theorem:**

$Cts([0,\infty), \mathbb{R}) \notin \mathcal{B}_{\text{cyl}}$!

Define distributions on $\mathbb{R}^F$ —Kolmogorov→ Obtain distribution β on $\mathbb{R}^{[0,\infty)}$ → Show $\beta(Cts([0,\infty), \mathbb{R}) = 1$ ✗

→ Show $B \sim \beta$ possesses an 'a.s. continuous modification'

- **Prohorov's Theorem:**

Define cts, finite time, approximations $B_n \in Cts([0,1], \mathbb{R})$ —Prohorov→ Obtain finite time cts Brownian motion on $Cts([0,1], \mathbb{R})$ → Glue copies into a process with sample paths in $Cts([0,\infty), \mathbb{R})$

Sketching convergence $\psi_n \to \psi$

Reminder

- $W \subseteq \mathbb{R}^d$ compact.
- Continuous $f : \mathbb{R}^N \times W \to \mathbb{R}$.
- $C = Cts(W, \mathbb{R})$ and $\mathscr{B}$ the Borel σ-algebra.
- $q(x)$ a pdf on $\mathbb{R}^N$
- $X_1, X_2, \ldots \sim q(x)$ iid.
- C-valued random variables:

$$\psi_n(w) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(f(X_i, w) - \mathbf{E} f(X_i, w) \right)$$

Goal: Prove $\psi_n(w) \to \psi(w)$ in distribution on $(C, \mathscr{B})$, for some C-valued $\psi(w)$.

Tightness and Prohorov's Theorem

Let Ψ_n be a sequence of probability measures on $(C, \mathscr{B})$.

Definition: We say Ψ_n is *tight* if for every $\epsilon > 0$ there exists a compact $K \subseteq C$ such that $\Psi_n(K) > 1 - \epsilon$ for all n.

A sequence of C-valued random variables ψ_n is *tight* if the corresponding distributions $\Psi_n(A) = \mathbf{P}(\psi_n \in A)$ are tight.

Prohorov's Theorem: Suppose W is compact and ψ_n is tight. Then ψ_n converges in distribution to some C-valued random variable ψ if and only if every finite projection $(\psi_n(w_1), \ldots, \psi_n(w_k))$ converges in distribution as $n \to \infty$.

(Really, this is a corollary of Prohorov's Theorem, which is about relating tightness and *relative compactness*.)

Compactness in C

- So it suffices to show ψ_n is tight.
- How do we show subsets of C are compact? (closed balls are not compact!)

Arzelà-Ascoli Theorem: Let $W \subseteq \mathbb{R}^d$ be compact. A subset $F \subseteq C$ has compact closure if and only if:

1. F is *uniformly bounded*: there exists $M > 0$ such that $\sup_{w \in W} f(w) < M$ for all $f \in F$
2. F is *equicontinuous:* for every $\epsilon > 0$ there exists $\delta > 0$ such that for all $f \in F$

$$\sup_{\|w - w'\| \leq \delta} |f(w) - f(w')| < \epsilon$$

- Only uniform boundedness is discussed in the Grey Book's convergence proof.

$L^s(q)$-valued real analytic functions

- For $s \geq 1$, consider the Banach space $L^s(q)$ of functions $h : \mathbb{R}^N \to \mathbb{R}$ which satisfy

$$\int |h(x)| q(x) dx < \infty$$

Definition: We say $f : \mathbb{R}^N \times W \to \mathbb{R}$ is an $L^s(q)$*-valued real analytic function* if for any $w^* \in W$ there exist coefficient functions $a_\alpha(x) \in L^s(q)$ indexed by $\alpha \in \mathbb{N}^d$ such that

$$f(x, w) = \sum_{\alpha \in \mathbb{N}^d} a_\alpha(x)(w - w^*)^\alpha$$

for all $w \in W'$ where W' is an open nbhd of w^*. Here, the RHS is required to converge absolutely in $L^s(q)$ for all $w \in W'$. This means that

$$\sum_{\alpha \in \mathbb{N}^d} \|a_\alpha\|_s |w - w^*| < \infty$$

for all $w \in W'$.

Lemma for uniform boundedness

Lemma (Grey Book Theorem 5.8)**:** Suppose W is compact. Let $s \geq 2$ be an even integer, and suppose $f : \mathbb{R}^N \times W \to \mathbb{R}$ is $L^s(q)$-valued real analytic. Then

$$\mathbf{E}[\sup_{w \in W} [\,|\psi_n(w)|^s]\,] < \infty$$

- Recall Markov's inequality: $\mathbf{P}(Y \geq M) \leq M^{-1}\mathbf{E}[Y]$, where $Y \geq 0$.
- Hence $\mathbf{P}(\sup_{w \in W} |\psi_n(w)| \geq M) \leq AM^{-s}$ for any $M \geq 0$.
- Hence for any $\epsilon > 0$ there exists $M > 0$ such that

$$\mathbf{P}(\sup_{w \in W} |\psi_n(w)| < M) \geq 1 - \epsilon$$

Lemma for equicontinuity

Lemma: Suppose W is compact. Let $s \geq 2$ be an even integer, and suppose $f : \mathbb{R}^N \times W \to \mathbb{R}$ is $L^s(q)$-valued real analytic. Then there exists $\delta_{\text{max}} > 0$ and $B > 0$ such that for any $\delta \in (0, \delta_{\text{max}})$

$$\mathbf{E}\left[\sup_{\|w-w'\| \leq \delta} |\psi_n(w) - \psi_n(w')|^s\right] \leq B\delta^s$$

Lemma: Suppose $f : \mathbb{R}^N \times W \to \mathbb{R}$ is $L^s(q)$-valued real analytic. Then the partial derivative $\partial_{w_j} f(w, x)$ is $L^s(q)$-valued real analytic.

Lemma for equicontinuity

Lemma: Suppose W is compact. Let $s \geq 2$ be an even integer, and suppose $f : \mathbb{R}^N \times W \to \mathbb{R}$ is $L^s(q)$-valued real analytic. Then there exists $\delta_{\text{max}} > 0$ and $B > 0$ such that for any $\delta \in (0, \delta_{\text{max}})$

$$\mathbf{E}\left[\sup_{\|w-w'\|\leq\delta} |\psi_n(w) - \psi_n(w')|^s\right] \leq B\delta^s$$

- Via Markov's inequality as before, this implies for any $\epsilon > 0$ and $\eta > 0$ that there exists $\delta > 0$ such that

$$\mathbf{P}\left(\sup_{\|w-w'\|\leq\delta} |\psi_n(w) - \psi_n(w')| < \eta\right) < 1 - \epsilon$$

- Tightness can now be proved via a standard argument.
- Let $\epsilon > 0$ be given.
- Let M be such that, for $F' = \{f \in C \mid \sup_{w \in W} |f(w)| < M\}$, we have $\mathbf{P}(\psi_n \in F') \geq 1 - \epsilon/2$
- For $k = 1,2,\ldots$ let $\delta_k > 0$ be such that, for $F_k = \{f \in C \mid \sup_{\|w-w'\| \leq \delta_k} |f(w) - f(w')| < 1/k\}$, we have $\mathbf{P}(\psi_n \in F_k) \geq 1 - 2^{-k}\epsilon/2$
- Let $F = F' \cap \bigcap_k F_k$ and $K = \overline{F}$.
- K is compact by Arzelà-Ascoli and $\mathbf{P}(\psi_n \in K) \geq 1 - \epsilon$.
- Therefore ψ_n is tight.

Theorem (Grey Book Theorem 5.9)**:** Suppose W is compact. Let $s \geq 2$ be an even integer, and suppose $f : \mathbb{R}^N \times W \to \mathbb{R}$ is $L^s(q)$-valued real analytic. Then:

1. ψ_n is tight.
2. For any finite subset $\{w_1, \ldots, w_k\} \subseteq W$ the vector $(\psi_n(w_1), \ldots, \psi_n(w_k))$ converges as $n \to \infty$ to a normal distribution with mean zero and covariance $\Sigma_{ij} = \text{cov}(f(X, w_i), f(X, w_j))$. *This is the usual central limit theorem.*
3. $\psi_n \to \psi$ in distribution, where ψ is continuous and every $(\psi(w_1), \ldots, \psi(w_k))$ has the above normal distribution. *By Prohorov's Theorem.*

Comparison to the Grey Book

- The Grey Book establishes uniform boundedness in Theorem 5.8.
- The proof of tightness (Example 5.3) does not discuss Arzelà-Ascoli or establish equicontinuity.
- The proof of convergence (Theorem 5.9) discusses the cylindrical σ-algebra on $\mathbb{R}^W$, which doesn't seem relevant.
- Because equicontinuity is not discussed, the fact that the w-derivatives of f inherit the property of being $L^s(q)$-valued real analytic isn't mentioned.

Bibliography

- (Grey Book) S. Watanabe (2009) *Algebraic Geometry and Statistical Learning Theory.*
- P. Billingsley (1999) *Convergence of Probability Measures.*