Aspects of symmetric Gamma process mixtures

Zacharie Naulet (Paris-Dauphine University)

Joint work with:

Judith Rousseau (Paris-Dauphine University)
Éric Barat (Commissariat à l'Energie Atomique)

Colloque JPS | 20th April 2016

Outline

Frequentist vs Bayes

Frequentist approach

- Choose a model $\mathcal{P}_n = \{P_\theta^n : \theta \in \Theta\}$.
- Observations $Y := (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$ are random variables with joint distribution $P_{\theta_0}^n \in \mathcal{P}_n$, θ_0 is unknown but assumed to be deterministic.
- Build an estimator $\hat{\theta}_n(Y)$ of θ_0 (ideally converging to θ_0 under P_{θ_0}).

Frequentist vs Bayes

Frequentist approach

- Choose a model $\mathcal{P}_n = \{P_\theta^n : \theta \in \Theta\}$.
- Observations $Y := (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$ are random variables with joint distribution $P_{\theta_0}^n \in \mathcal{P}_n$, θ_0 is unknown but assumed to be deterministic.
- Build an estimator $\hat{\theta}_n(Y)$ of θ_0 (ideally converging to θ_0 under P_{θ_0}).

Bayesian approach

- Observations $Y = (Y_1, \ldots, Y_n)$ and parameter θ are random variables with joint distribution Π on $\mathcal{Y}^n \times \Theta$.
- $P_\theta^n(\cdot) = \Pi(\cdot|\theta)$.
- Marginal of Π on Θ, Π_θ, is called the prior distribution.
- The model is the probability space $(\Theta, \Sigma, \Pi_\theta)$.

Frequentist vs Bayes

Frequentist approach

- Choose a model $\mathcal{P}_n = \{P_\theta^n : \theta \in \Theta\}$.
- Observations $Y := (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$ are random variables with joint distribution $P_{\theta_0}^n \in \mathcal{P}_n$, θ_0 is unknown but assumed to be deterministic.
- Build an estimator $\hat{\theta}_n(Y)$ of θ_0 (ideally converging to θ_0 under P_{θ_0}).

Bayesian approach

- Observations $Y = (Y_1, \ldots, Y_n)$ and parameter θ are random variables with joint distribution Π on $\mathcal{Y}^n \times \Theta$.
- $P_\theta^n(\cdot) = \Pi(\cdot|\theta)$.
- Marginal of Π on Θ, Π_θ, is called the prior distribution.
- The model is the probability space $(\Theta, \Sigma, \Pi_\theta)$.

In both cases, the model is

1. Parametric if Θ is a finite-dimensional vector space.
2. Non parametric if Θ is an infinite-dimensional vector space.

Bayesian estimation

The conditional distribution $\Pi_{\theta|Y}$ is called the posterior distribution, and is given by the Bayes rule :

$$\Pi_{\theta|Y}(U|B) = \frac{\Pi_{Y|\theta}(B|U)\Pi_{\theta}(U)}{\Pi_{Y}(B)}.$$

Bayesian estimation

The conditional distribution $\Pi_{\theta|Y}$ is called the posterior distribution, and is given by the Bayes rule :

$$\Pi_{\theta|Y}(U|B) = \frac{\Pi_{Y|\theta}(B|U)\Pi_{\theta}(U)}{\Pi_{Y}(B)}.$$

Bayesian point estimator

- Posterior mean : $\hat{\theta}_n(Y) = \int_\Theta \theta \, d\Pi(\theta|Y_1, \ldots, Y_n)$.
- If the posterior is dominated on Θ,
  Maximum a Posterior (MAP) : $\hat{\theta}_n = \arg\max_{\theta\in\Theta} \pi(\theta|Y_1, \ldots, Y_n)$.

Credible intervals

- U is a credible interval with level α if,

$$\Pi(\theta \in U|Y_1, \ldots, Y_n) = 1 - \alpha.$$

Bayesian estimation

Two kinds of Bayesians :

1. Classical : The classical Bayesian believe in the existence of a true parameter to be estimated from the data (*e.g.* Laplace, Bayes)
2. Subjectivist : The subjectivist Bayesian rejects the idea of a true parameter, there are no objectives probability models (*e.g.* Diaconis, De Finetti)

More details in

Persi Diaconis and David Freedman (1986). "On the consistency of Bayes estimates". In: *The Annals of Statistics*, pp. 1–26.

Bayesian estimation

Two kinds of Bayesians :

1. Classical : The classical Bayesian believe in the existence of a true parameter to be estimated from the data (*e.g.* Laplace, Bayes)
2. Subjectivist : The subjectivist Bayesian rejects the idea of a true parameter, there are no objectives probability models (*e.g.* Diaconis, De Finetti)

More details in
Persi Diaconis and David Freedman (1986). "On the consistency of Bayes estimates". In: *The Annals of Statistics*, pp. 1–26.

If we are a classical Bayesian, we probably want that our posterior distribution converges (in some sense) to a degenerate distribution at θ_0, as the data increase.

Bayesian estimation

Two kinds of Bayesians :

1. Classical : The classical Bayesian believe in the existence of a true parameter to be estimated from the data (*e.g.* Laplace, Bayes)
2. Subjectivist : The subjectivist Bayesian rejects the idea of a true parameter, there are no objectives probability models (*e.g.* Diaconis, De Finetti)

More details in

Persi Diaconis and David Freedman (1986). "On the consistency of Bayes estimates". In: *The Annals of Statistics*, pp. 1–26.

If we are a classical Bayesian, we probably want that our posterior distribution converges (in some sense) to a degenerate distribution at θ_0, as the data increase.

- Frequentist consistency : An estimator $\hat{\theta}_n(Y)$ is consistent at θ_0 (in the distance d) if $d(\hat{\theta}_n, \theta_0) \to 0$ in $P_{\theta_0}^{\infty}$-probability.
- Bayesian consistency : The sequence of posterior distributions $\{\Pi_n(\cdot|Y)\}$ is consistent at θ_0 (in the distance d) if for all $\epsilon > 0$,

$$\Pi_n(\{\theta : d(\theta, \theta_0) \geq \epsilon\} | Y_1, \ldots, Y_n) \to 0, \quad P_{\theta_0}^{\infty}\text{-a.s.}$$

Outline

Problem Statement

We consider the nonparametric (direct or indirect) regression problem with $f : \mathbb{R}^d \to \mathbb{C}$ and data (X_i, Y_i), $i = 1, \ldots, n$, with

$$E[Y_i|X_i] = T(f)(X_i), \quad X_i \in \mathcal{X},$$

from a Bayesian perspective.

Problem Statement

We consider the nonparametric (direct or indirect) regression problem with $f : \mathbb{R}^d \to \mathbb{C}$ and data (X_i, Y_i), $i = 1, \dots, n$, with

$$E[Y_i|X_i] = T(f)(X_i), \quad X_i \in \mathcal{X},$$

from a Bayesian perspective.

Canonical example: Gaussian mean regression, with $f : \mathbb{R}^d \to \mathbb{R}$:

$$\begin{aligned} Y_i|X_i, \epsilon_i &= f(X_i) + \epsilon_i, \quad i = 1, \dots, n \\ \epsilon_1, \dots, \epsilon_n &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \\ f &\sim \Pi. \end{aligned}$$

The posterior distribution $\Pi(\cdot|Y_1, \dots, Y_n)$ is given by

$$\underbrace{\Pi(f \in U|Y_1, Y_2, \dots)}_{\textit{posterior}} \propto \int_U \underbrace{L(f|Y_1, Y_1, \dots)}_{\textit{likelihood}} \underbrace{\Pi(df)}_{\textit{prior}}.$$

Prior distributions on function spaces

Brief (non exhaustive) state of the art for prior distributions on function spaces.

Regression:

- Gaussian processes (Rasmussen, 2004)

Density estimation:

- Dirichlet processes mixtures (Escobar and West, 1995)

Idea: Use Kernel mixtures models in regression problems

- Abramovich, Sapatinas, and Silverman (2000),
- Wolpert, Ickstadt, and Hansen (2003),
- Pillai et al. (2007) and Pillai (2008),
- Wolpert, Clyde, and Tu (2011),
- Malou (2014),
- This talk, and Naulet and Barat (2015).

Kernel mixtures models

Let

- $\mathcal{G}$ be a measurable space
- $\mathcal{M}(\mathcal{G})$ be the set of signed (or complex-valued) measures on $\mathcal{G}$
- $\Pi_*(dQ)$ be a prior distribution on $\mathcal{M}(\mathcal{G})$
- $\Phi : \mathcal{G} \times \mathbb{R}^d \to \mathbb{R}$ be a kernel function

Then $\Pi_*(dQ)$ induces a prior distribution on an abstract space of functions $f : \mathbb{R}^d \to \mathbb{R}$ through the mapping

$$\mathcal{M}(\mathcal{G}) \ni Q \mapsto \int_{\mathcal{G}} \Phi(x; \cdot)\, dQ(x).$$

Let $\Pi(df)$ denote this prior distribution :

$$f \sim \Pi(df) \iff \begin{array}{l} f(\cdot) = \int_{\mathcal{G}} \Phi(x; \cdot)\ dQ(x) \\ Q \sim \Pi_*(dQ). \end{array}$$

Kernel mixtures models

Examples of prior distributions on $\mathcal{M}(\mathcal{G})$ (*ie.* random measures):

- Dirichlet processes,
- Completely Random Measures (Kingman, 1967; Kingman, 1992; Naulet and Barat, 2015).
- Lévy Random Measures (Wolpert, Clyde, and Tu, 2011; Pillai, 2008; Rajput and Rosinski, 1989; Barndorff-Nielsen and Schmiegel, 2004).

Examples of kernels:

- Location-scale kernels :

$$f(\cdot) := \int \sigma^{-1} g(\cdot/\sigma)\, dQ(\mu, \sigma),$$

- Location-modulation kernels :

$$f(\cdot) := \int g(\cdot - \mu) \cos(\langle \omega, \cdot \rangle + \theta)\, dQ(\mu, \omega, \theta),$$

- ...

Symmetric Gamma Random Measures

Symmetric Gamma Random Measures (SGRM) are distributions over space of signed measures (random signed measure)

Symmetric Gamma Random Measures

Symmetric Gamma Random Measures (SGRM) are distributions over space of signed measures (random signed measure)

We let,

1. $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and
2. $(\mathcal{G}, \Sigma_{\mathcal{G}})$ be a measurable space.

Symmetric Gamma Random Measures

Symmetric Gamma Random Measures (SGRM) are distributions over space of signed measures (random signed measure)

We let,

1. $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and
2. $(\mathcal{G}, \Sigma_{\mathcal{G}})$ be a measurable space.

Definition

Let α be a finite measure on $(\mathcal{G}, \Sigma_{\mathcal{G}})$ and $\eta > 0$. A random measure $Q : \Omega \times \Sigma_{\mathcal{G}} \to [-\infty, \infty]$ is a symmetric Gamma random measure with parameters (α, η) if for all disjoints set $A_1, \ldots, A_n \in \Sigma_{\mathcal{G}}$ the random variables $Q(\cdot, A_1), \ldots, Q(\cdot, A_n)$ are independent random variables with the distribution of the difference of two independent Gamma $(\alpha(A_i), \eta)$ random variables.

Symmetric Gamma Random Measures (continued)

Some properties of SGRM :

- Q has almost-surely the series representation $Q = \sum_{i=1}^{\infty} \beta_i \delta_{x_i}$, with
- Q is a signed measure
- $\sum_{i=1}^{\infty} |\beta_i|$ has Gamma $(2\alpha(\mathcal{G}), \eta)$ distribution (hence almost-surely finite)
- Both β_i's and x_i's are random.

... one last thing.

It is not clear *a priori* whether or not the integral with respect to Q in the definition of the mixture converges (and in what sense).

For those who are interested, see details in

- Robert L Wolpert, Merlise A Clyde, and Chong Tu (2011). "Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels". In: *The Annals of Statistics* 39.4, pp. 1916–1962
- Jan Rosiński (1987). *Bilinear random integrals*.

Outline

Outline

Preliminaries

Remember that a sequence of posterior distributions converges at rate $\epsilon_n \to 0$ toward f_0 (in the distance d) if there is a constant $M > 0$ such that

$$\Pi(\{f : d(f, f_0) \geq M\epsilon_n\} | Y_1, \ldots, Y_n) \to 0, \quad P_{f_0}^{\infty}\text{-a.s.}$$

Question : Are kernel mixtures model consistent ?
No general answer, because

- The posterior depends on the likelihood function.
- It certainly depends on the choice of the metric d.

But Bayesians have a very general theorem for consistency, namely :

Theorem (Doob's consistency theorem, 1948)

Suppose that the parameter space Θ and the sample space $\mathcal{Y}$ are complete, separable, metric, and endowed with their respective Borel σ-algebra. Assume that $\Theta \ni f \mapsto P_f$ is one-to-one. Then the sequence of posterior distributions is consistent Π-almost-surely.

The unfortunate point in Doob's theorem is Π-almost-surely.
In nonparametric situation, it is always possible that the true parameter belongs to a null set of the prior.

In fact, null sets of the prior distribution can be large in a topological sense.
Famous example can be found in
David A Freedman (1963). "On the asymptotic behavior of Bayes' estimates in the discrete case". In: *The Annals of Mathematical Statistics*, pp. 1386–1403.

Thus, not all priors are suitable in nonparametric estimation (unless you're a subjectivist Bayesian), and we should provide conditions to ensure consistency.

Schwartz's theory (non iid observations)

Let define,

- $K_i(f_0, f) := \int \log \frac{dP_{f_0,i}}{dP_{\theta,i}} \, dP_{f_0,i}$,
- $V_i(f_0, f) := \int (\log \frac{dP_{f_0,i}}{dP_{\theta,i}} - K_i(f_0, f))^2 \, dP_{f_0,i}$,
- $KL(f_0, \epsilon_n) := \{f \in \Theta : \sum_{i=1}^n K_i(f_0, f) \leq n\epsilon_n^2, \ \sum_{i=1}^n V_i(f_0, f) \leq n\epsilon_n^2\}$.

Theorem (Schwartz's theorem : elementary version)

Assume that for a sequence $\epsilon_n \to 0$ with $n\epsilon_n^2 \to \infty$ the following holds

1. $\Pi(KL(f_0, \epsilon_n)) \geq \exp(-n\epsilon_n^2)$
2. *there exist a sequence (ϕ_n) of test-functions such that:*

$$P_{f_0}^n \phi_n \to 0, \qquad \sup_{f, d(f,f_0) \geq \epsilon} P_f^n (1 - \phi_n) \leq e^{-3n\epsilon_n^2}.$$

Then,

$$\Pi(\{f \in \Theta : d(f, f_0) \geq \epsilon_n\} | Y_1, \ldots, Y_n) \to 0 \quad P_{f_0}^\infty\text{-}a.s.$$

Approach to consistency : existence of tests

But, in general when Θ is infinite-dimensional, we cannot find test-functions for testing $H_0 : f = f_0$ vs $H_1 : d(f, f_0) \geq \epsilon$.

But sometimes, we can find tests functions for testing "balls", $H_0 : f = f_0$ vs $H_1 : d(f, f_1) \leq \epsilon/2$ with $d(f_1, f_0) \geq \epsilon$.

- Imagine that we can cover the set $\{f : d(f, f_0) \geq \epsilon\}$ with finitely many (say N) balls

$$B_j := \{f \in \Theta : d(f, f_j) \leq \epsilon/2\}.$$

  (Likewise $d(f, f_0) \geq \epsilon/2$ for all $f \in B_j$ and all j).
- We can have test-functions $\phi_n^{(j)}$, $j = 1, \ldots, N$ for testing $f = f_0$ against $f \in B_j$, each satisfying $P_{f_0}^n \phi_n^{(j)} \leq \mathrm{e}^{-Kn}$ and $\sup_{f \in B_j} P_f^n (1 - \phi_n^{(j)}) \leq \mathrm{e}^{-Kn}$.
- Then we can build the test-functions $\phi_n = \max_j \phi_n^{(j)}$, and
  - $P_{f_0}^n \phi_n \leq \sum_{j=1}^N P_{f_0} \phi_n^{(j)} \leq N \mathrm{e}^{-Kn}$
  - $\sup_{f, d(f, f_0) \geq \epsilon/2} P_f^n (1 - \phi_n) \leq \min_{j=1,\ldots,N} \sup_{f \in B_j} (1 - \phi_n^{(j)}) \leq \mathrm{e}^{-Kn}$.

Approach to consistency : existence of tests

Unfortunately (especially for metric inducing strong topologies), it is often impossible to cover $\{f : d(f, f_0) \geq \epsilon\}$ with finitely many balls of finite radius.

But, if we can find sets $\Theta_n \subset \Theta$ (called a *sieve*), such that

- Θ_n can be covered with $N_n < +\infty$ balls of radius $\epsilon/2$,
- N_n is not growing too fast as n increase.
- $\Theta_n \nearrow \Theta$ in some sense.

Then the previous reasoning still holds.

Improved Schwartz's theorem

If the following conditions are met for n large enough ($K > 0$ universal constant)

- Existence of exponentially consistent test, $(\phi_i)_{i=1}^n$ such that for all $f_1 \in \Theta$ with $d(f_1, f_0) > \epsilon$,

$$P_{f_0}^{(n)} \phi_n \leq e^{-3Kn\epsilon_n^2}, \qquad \sup_{f \in \Theta, d(f,f_1) \leq \epsilon/2} P_f^{(n)}(1 - \phi_n) \leq e^{-3Kn\epsilon_n^2},$$

- Existence of sets $\Theta_n \subset \Theta$ such that,

$$\Pi(\Theta \backslash \Theta_n) \leq e^{-3Kn\epsilon_n^2}$$
$$\log N(\epsilon/2, \Theta_n, d) \leq K' n\epsilon_n^2, \quad K' < K$$

- Prior positivity of Kullback-Leibler neighborhoods.

$$\Pi\left(KL(f_0, \epsilon_n)\right) > \exp(-Kn\epsilon_n^2).$$

Then, for all $M < +\infty$

$$\Pi(\{f \in \Theta \ : \ d(f, f_0) > M\epsilon_n\} | Y_1, \ldots, Y_n) \to 0 \quad P_{f_0}^{\infty}\text{-a.s.}$$

Outline

Model and results

Let $x_1, \ldots, x_n \in [0,1]^d$. We consider the simple model,

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n$$
$$\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$
$$f(x) = \int_{\mathcal{S} \times \mathbb{R}^d} g\left(A^{-1}x - \mu\right) \, Q(dAd\mu), \quad \forall x \in \mathbb{R}^d$$
$$Q \sim \mathrm{SGRM}(\alpha, \eta).$$

Assumptions

- Too many to be listed.
- $d_n(f_1, f_2)^2 := n^{-1} \sum_{i=1}^n |f_1(x_i) - f_2(x_i)|^2$

Results

Suppose that $f_0 \in \mathcal{C}^\beta[0,1]^d$. Under the previous assumptions, there is $\zeta > 0$, $M > 0$ such (1) holds for the location-scale prior with $\epsilon_n = (n/\log n)^{-\beta/(2\beta+d+1/2)}(\log n)^\zeta$.

$$\lim_{n\to\infty} \Pi\left(\{f \in \Theta \ : \ d_n(f, f_0) > M\epsilon_n\} | Y_1, \ldots, Y_n\right) = 0 \quad P_{f_0}^\infty\text{-a.s.} \tag{1}$$

Conclusion

Thank you.

References I

Abramovich, F, T Sapatinas, and BW Silverman (2000). "Stochastic expansions in an overcomplete wavelet dictionary". In: *Probability Theory and Related Fields* 117.1, pp. 133–144 (cit. on p. 14).

Barndorff-Nielsen, Ole E and Jürgen Schmiegel (2004). "Lévy-based spatial-temporal modelling, with applications to turbulence". In: *Russian Mathematical Surveys* 59.1, p. 65 (cit. on p. 16).

Diaconis, Persi and David Freedman (1986). "On the consistency of Bayes estimates". In: *The Annals of Statistics*, pp. 1–26 (cit. on pp. 8–10).

Escobar, Michael D and Mike West (1995). "Bayesian density estimation and inference using mixtures". In: *Journal of the american statistical association* 90.430, pp. 577–588 (cit. on p. 14).

Freedman, David A (1963). "On the asymptotic behavior of Bayes' estimates in the discrete case". In: *The Annals of Mathematical Statistics*, pp. 1386–1403 (cit. on p. 24).

Kingman, JFC (1967). "Completely random measures". In: *Pacific Journal of Mathematics* 21.1 (cit. on p. 16).

References II

Kingman, John Frank Charles (1992). *Poisson processes*. Vol. 3. Oxford university press (cit. on p. 16).

Malou, Eddy (2014). "Congolexicomatisation des lois du marché." In: (cit. on p. 14).

Naulet, Zacharie and Eric Barat (2015). "Adaptive Bayesian nonparametric regression using mixtures of kernels". In: *arXiv preprint arXiv:1504.00476* (cit. on pp. 14, 16).

Pillai, Natesh S (2008). "Lévy random measures: Posterior consistency and applications". PhD thesis. Duke University (cit. on pp. 14, 16).

Pillai, Natesh S et al. (2007). "Characterizing the function space for Bayesian kernel models". In: *Journal of Machine Learning Research* 8, pp. 1769–1797 (cit. on p. 14).

Rajput, Balram S and Jan Rosinski (1989). "Spectral representations of infinitely divisible processes". In: *Probability Theory and Related Fields* 82.3, pp. 451–487 (cit. on p. 16).

Rasmussen, Carl Edward (2004). "Gaussian processes in machine learning". In: *Advanced Lectures on Machine Learning*. Springer, pp. 63–71 (cit. on p. 14).

References III

Rosiński, Jan (1987). *Bilinear random integrals*. (Cit. on p. 20).
Wolpert, Robert L, Merlise A Clyde, and Chong Tu (2011). "Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels". In: *The Annals of Statistics* 39.4, pp. 1916–1962 (cit. on pp. 14, 16, 20).
Wolpert, Robert L, Katja Ickstadt, and Martin B Hansen (2003). "A nonparametric Bayesian approach to inverse problems". In: *Bayesian statistics* 7, pp. 403–417 (cit. on p. 14).

Sketch of the proof of the Schwartz theorem I

We assume that (but we could have not) that the model is dominated by λ and we write $p_{f,i} = \frac{dP_{f,i}}{d\lambda}$ and $Y^n := (Y_1, \ldots, Y_n)$.

Then,

$$\begin{aligned} & P_{f_0}^n \Pi(f : d(f, f_0) \geq \epsilon_n | Y^n) \\ & = P_{f_0}^n [\Pi(f : d(f, f_0) \geq \epsilon_n | Y^n) \phi_n] + P_{f_0}^n [\Pi(f : d(f, f_0) \geq \epsilon_n | Y^n)(1 - \phi_n)] \\ & \leq P_{f_0}^n \phi_n + P_{f_0}^n [\Pi(f : d(f, f_0) \geq \epsilon_n | Y^n)(1 - \phi_n)] \end{aligned}$$

Wa can rewrite,

$$\Pi(f : d(f, f_0) \geq \epsilon_n | Y^n) = \frac{\int_{d(f,f_0) \geq \epsilon_n} \frac{p_f(Y_1, \ldots, Y_n)}{p_{f_0}(Y_1, \ldots, Y_n)} d\Pi(f)}{\int_\Theta \frac{p_f(Y_1, \ldots, Y_n)}{p_{f_0}(Y_1, \ldots, Y_n)} d\Pi(f)}$$

Sketch of the proof of the Schwartz theorem II

. Let $\Lambda_n(Y_1, \ldots, Y_n) = n^{-1} \sum_{i=1}^n \log \frac{p_f(Y_i)}{p_{f_0}(Y_i)}$.

- the integrand in the denominator can be rewritten as $e^{-n\Lambda_n(Y_1,\ldots,Y_n)}$.
- Λ_n behaves like the KL-divergence for large n
- Lower bound the integral by integrating on the smaller set $KL(f_0, \epsilon_n)$
- Use Chebychev and the assumption $\Pi(KL(f_0, \epsilon_n)) \gtrsim \exp(-n\epsilon_n^2)$ to show that the event that the denominator is smaller than $\exp(-2n\epsilon_n^2)$ has probability $\to 0$ as $n \to \infty$.

2. Bounding the numerator

- It suffices to consider the event A_n that the denominator is greater than $\exp(-2n\epsilon_n^2)$. Hence by an application of Fubini's theorem,

$$
\begin{aligned}
&P_{f_0}^n[\mathbb{1}_{A_n}\Pi(f : d(f, f_0) \geq \epsilon_n | Y^n)(1 - \phi_n)] \\
&\qquad \leq \exp(2n\epsilon_n^2) P_{f_0}^n(1 - \phi_n) \int \frac{dP_f(Y^n)}{dP_{f_0}} \, d\Pi(f) \\
&\qquad \leq \exp(2n\epsilon_n^2) P_f^n(1 - \phi_n).
\end{aligned}
$$