Notation of posterior consistency

- Model: \( X^{(n)} =: (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} f \)
- Prior: \( f \sim \Pi \)
- Posterior:
  \[
  \Pi_n(B|X^{(n)}) = \frac{\int_B \prod_{i=1}^n f(y_i) d\Pi(f)}{\int \prod_{i=1}^n f(y_i) d\Pi(f)}
  \]

**Consistency**

The posterior distribution \( \Pi_n(\cdot|X^{(n)}) \) is said to be consistent at \( f_0 \) if for every neighborhood \( U \) of \( f_0 \), \( \Pi_n(U|X^{(n)}) \to 1 \) almost surely under \( f_0 \).
The choice of topology determined the geometry of the neighborhoods.

Weak neighborhoods:

\[ U_\epsilon(f_0) = \{ f : d_W(f, f_0) < \epsilon \}, \]

where \( d_W \) is a metric that metrizes the topology of weak convergence.

Recall (Weak convergence): \( f_n \rightarrow f \) weakly if and only if

\[ \lim_{n \rightarrow \infty} \int \psi f_n = \int \psi f \] for any bounded continuous function \( \psi \).

\( d_W \) is a metric such that \( d_W(f_n, f) \rightarrow 0 \) if and only if

\[ \int \psi f_n \rightarrow \int \psi f \] for any bounded continuous function \( \psi \).

This leads to weak consistency.
Strong consistency

- Strong neighborhoods:

\[ U_\varepsilon(f_0) = \{ f : \int |f_n - f_0| < \varepsilon \} = \{ f : \|f_n - f_0\|_1 < \varepsilon \} \]

- This leads to strong or \( L_1 \) consistency.

- Other metrics on densities \((p, q)\) given dominating measure \(\mu\):
  - Total variation distance:
    \[ d_{TV}(p, q) = \frac{1}{2} \int |p - q|d\mu \]
  - Hellinger distance:
    \[ d_H(p, q) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2 d\mu} \]
  - Kullback-Leibler divergence:
    \[ KL(p; q) = \int \left( \log \frac{p}{q} \right) p d\mu \]
- $\text{KL}(p; q) \geq 0$ for any densities $(p, q)$ by applying Jensen’s inequality.

Inequalities:

$$\|p - q\|_1^2 \leq 4d_H^2(p, q) \leq 4\|p - q\|_1$$

$$\text{KL}(p, q) \geq \|p - q\|_1^2/2$$

### Kullback-Leibler property

A density $f_0$ is said possess the **Kullback-Leibler property** relative to a prior $\Pi$ if $\Pi(f : \text{KL}(f_0; f) < \varepsilon) > 0$ for every $\varepsilon > 0$. This is denoted $f_0 \in \text{KL}(\Pi)$, and we also say that $f_0$ belongs to the **Kullback-Leibler support** of $\Pi$. 

Schwartz’s theorem

Schwartz (1965)
If $f_0 \in \text{KL}(\Pi)$ and for every neighborhood $U$ of $f_0$ there exists test $\phi_n$ such that $P^n_0 \phi_n \to 0$ and $\sup_{f \in U^c} P^n(1 - \phi_n) \to 0$, then the posterior distribution $\Pi_n(U|X^{(n)}) \to 1 \text{ a.s. } P_0^\infty$.

- Interpretation: posterior consistency is guaranteed by large KL support and model identifiability condition.
- Existence of consistent tests may be not convenient to verify.
- However, consistent tests under the weak topology on the set of probability measures always exists.
- Therefore, we have

Weak consistency
If $f_0 \in \text{KL}(\Pi)$, the posterior distribution is weakly consistent at $f_0$. 
The Kullback-Leibler (KL) property plays a central role to verify posterior consistency of numerous processes. Processes (under certain conditions) that possess the KL property include:

- Póly trees
- Kernel mixtures (Wu and Ghosal, 2008):
  \[ f_{F,\psi}(x) = \int \psi(x; \theta, \phi) dF(\theta) \] if priors on \( F \) and \( \phi \) have full support.
- Normal mixtures (Tokdar, 2006):
  \[ p_{F,\sigma} = \phi_\sigma * F \] with \( F \sim \Pi \) and \( \sigma \sim p_\sigma \).
  - Sufficient conditions: \( \sigma^{-2} \sim \text{Inverse-Gamma}(a, b) \) with \( a > 1 \), and \( F \sim \text{DP}(\alpha P_0) \) where \( P_0 \) has compact support or sub-Gaussian tails (i.e., \( \Pr(|X| \geq t) \lesssim \exp(-t^2/(2\sigma^2)) \)) for any \( t \in \mathbb{R} \).
- DP mixture (Ghosal et al., 1999)
- Bernstein polynomial priors (Petrone and Wasserman, 2002)
- Bayesian histograms (Gasparini, 1996)
A natural follow-up question after consistency is how fast the posterior concentrates.

A sequence $\epsilon_n$ is a posterior convergence rate at $f_0$ with respect to the pseudo-metric $d$ if

$$\Pi_n(f : d(f,f_0) \geq M_n \epsilon_n) \to 0$$

in $P_0^{(n)}$-probability, for every $M_n \to \infty$.

Equivalently, we can verify the condition

$$E_{f_0}^{n} \Pi_n(\theta : d(f,f_0) \geq M_n \epsilon_n) \to 0$$
We define “a” rate rather than “the” rate of convergence, because any rate slower than a convergence rate is also a convergence rate.

We are interested in a fastest decreasing sequence $\varepsilon_n$.

Generally we are happy if our rate is equal to or close to an “optimal” rate.

Nevertheless, we often make statements like “$\varepsilon_n$ is the rate of convergence”; this should not cause confusion.

**Lemma**

If $\Theta \in \mathbb{R}$, and $E(\theta|X^{(n)}) = \theta_0 + O_p(\varepsilon_n)$ and $\text{Var}(\theta|X^{(n)}) = O_p(\varepsilon_n^2)$ with respect to the distribution generated by the true parameter $\theta_0$, then $\varepsilon_n$ is a posterior convergence rate at $\theta_0$ with respect to the Euclidean distance.
Example: Dirichlet process

- Suppose $X_1, \ldots, X_n | P \sim P$ and $P \sim \text{DP}(\alpha P)$.
- Then the posterior distribution of $P$ is $\text{DP}(\alpha P + P_n)$ where $P_n$ is the empirical distribution of the observation.
- It follows that for a set $A$, the posterior mean of $P(A) = P_0(A) + O_p(n^{-1/2})$, and the posterior variance of $P(A) = O_p(n^{-1/2})$.
- Therefore, the posterior convergence rate using $d(P_1, P_2) = |P_1(A) - P_2(A)|$ is $n^{-1/2}$. 
In a general theorem given by Ghosal, Ghosh & van der Vaart (2000), posterior rates are calculated by satisfying three conditions:

- Entropy calculation $\log N(\varepsilon_n, \mathcal{F}_n, d) \lesssim n\varepsilon_n^2$
- Sieve selection: $-\log \Pi(\mathcal{F}_n^c) \gtrsim n\varepsilon_n^2$
- Prior concentration:

$$-\log \Pi(f : \text{KL}(f_0, f) \leq \varepsilon_n^2, V(f_0, f) \leq \varepsilon_n^2) \lesssim n\varepsilon_n^2,$$

where $\mathcal{F}_n \subset \mathcal{F}$ (the space of all densities) and $V(f_0, f) = \int f_0 \{\log(f_0/f)\}^2$.

Here $\varepsilon_n$ is a sequence such that $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$.

While there is a rich literature on contraction rates of many nonparametric Bayesian methods, it is still an ongoing research area that is very active.
Consistency is a frequentist concept.

Why do Bayesian care about consistency?

For an objective Bayesian who believes in an unknown true model, consistency validates the Bayesian method, i.e., it ensures data override prior opinions as they grow indefinitely.

For a subjective Bayesian who think no true parameter exists:

- A degenerate prior models perfect knowledge. Thus consistency entails convergence towards perfect knowledge.
- Merging of opinions. Calculations of different Bayesians with different priors will tend to agree iff their posteriors are consistent, where “agree” is with respect to the predictive distributions of future observations.
Characterizing a contraction rate is a significant refinement of posterior consistency.

It links nonparametric Bayes to the minimax theory of general statistical estimation.

It also provides guidance to select priors or hyperparameters in practice.

This leads to “adaptive” methods which do not require oracle knowledge about the unknown parameter.
Consider the nonparametric regression setting that was studied: \( y_i = f_0(x_i) + \varepsilon_i \), where \( \varepsilon_i \sim N(0, \sigma^2) \) and \( x_i \in \mathbb{R}^d \).

The empirical Bayes method in Homework 1 Problem 3(ii) is popular, but its theoretical properties have become known only recently (Szabó, 2013).

How about the fully Bayes approach to put a prior on the length scale parameter? And what priors should we use?

If \( f_0 \) belongs to the Hölder space of \( \alpha \)-smooth functions, then the minimax optimal rate to estimate \( f_0 \) is known as \( n^{-\alpha/(2\alpha+d)} \).

When applying Gaussian process priors with properly selected length scale parameters that depend on \( \alpha \) (say \( \ell = n^{-1/(2\alpha+1)} \) when \( d = 1 \)), the posterior distributions achieve the minimax optimal rate up to a logarithmic factor.
However, $\alpha$ of course is unknown to us.

Practical selection of tuning parameters in a nonparametric approach, Bayesian or non-Bayesian, often relies on empirical rules such as empirical Bayes and cross-validation.

If the optimal length scale is not known, why not put a prior on it thus randomly rescales the Gaussian process in use?

Van der Vaart and van Zanten (2009, AOS) showed that if the prior of the length scale satisfies some tail constraint, then the resulting posterior is nearly minimax optimal for any $\alpha$.

Therefore, we obtain a fully Bayes procedure that is adaptive to any unknown $\alpha$!

Specifically, the commonly used Gamma distribution on $\ell^{-d}$ satisfies the tail constraint.

Another example is squared periodic kernel GPs for closed curves in Li and Ghosal (2017, AOS).