2

Priors on Function Spaces

A functional parameter in nonparametric statistics may be restricted only by qualitative assumptions, such as smoothness (as in standard nonparametric regression), but may also satisfy additional constraints, such as positivity (density, Poisson regression), integration to unity (probability density), monotonicity (distribution, or link function), etc. A prior distribution for such a parameter must take the restrictions into account. Given only qualitative restrictions, a probability measure on a general function space is a appropriate, whereas a special construction or a transformation is necessary to adhere to further restrictions. In this chapter we start with examples of priors on general function spaces, and next discuss priors satisfying constraints. Examples of the first type are random series and stochastic processes.

2.1 Random Basis Expansion

Given a set of "basis functions" $\psi_j : T \to \mathbb{R}$ on some domain $T \subset \mathfrak{X}$, indexed by j in some set J, a prior on functions $f : T \to \mathbb{R}$ can be constructed by writing

$$f = \sum_{j \in J} \beta_j \psi_j, \tag{2.1}$$

and putting priors on the coefficients β_j in this representation. There are many choices of bases: polynomials, trigonometric functions, wavelets, splines, spherical harmonics, etc. Each basis is suited to represent a certain type of function, where the quality is typically measured by the number of basis functions needed to attain a good approximation. The approximation properties of the basis together with the prior on the coefficients determine the suitability of the prior for a given application. If the coefficients are unrestricted by each other (the usual situation), then independent priors on the coefficients are reasonable. Normal priors are commonly used if the coefficients can take on every real value, typically with zero mean and variances chosen to control the "energy" at the different levels of the expansion.

Given infinitely many basis functions, the convergence of the *random series* (2.1) is not guaranteed. When the target functions form a (subset of a) Hilbert space, it is natural and convenient to use an orthonormal basis, and convergence is easy to establish.

Lemma 2.1 (Hilbert space) *If the functions* $(\psi_j : j \in \mathbb{N})$ *form an orthonormal basis of a Hilbert space* $\mathbb{H}$, *then the series* (2.1) *converges in* $\mathbb{H}$ *a.s. if and only if* $\sum_j \beta_j^2 < \infty$ *a.s. If the* β_j *are independent and for every* k *the vector* $(\beta_1, \ldots, \beta_k)$ *has full support* $\mathbb{R}^k$, *then the corresponding prior has support* $\mathbb{H}$. *In particular, for* $\beta_j \sim \mathrm{Nor}(\mu_j, \tau_j^2)$ *these conditions* *are verified if* $\sum_{j=1}^{\infty}\mu_j^2 < \infty$ *and* $\sum_{j=1}^{\infty}\tau_j^2 < \infty$. *If* $\beta_1, \beta_2, \ldots$ *are also independent, then these conditions are necessary.*

Proof The first assertion is immediate from the fact that a deterministic series $\sum_{j=1}^{\infty}\beta_j\psi_j$ converges in $\mathbb{H}$ if and only if $\sum_j \beta_j^2 < \infty$. That $\{\psi_j: j \in \mathbb{N}\}$ forms an orthonormal basis means that any element $f_0 \in \mathbb{H}$ can be written $f_0 = \sum_j \beta_{0j}\psi_j$, where $\sum_{j>k}\beta_{j0}^2 \to 0$ as $k \to \infty$. That f_0 is in the support of the prior therefore follows from the fact that the event $\{(\beta_1, \ldots, \beta_k): \sum_{j\le k}|\beta_j - \beta_{j0}|^2 < \epsilon, \sum_{j>k}\beta_j^2 < \epsilon\}$ has positive probability under the prior, for every $\epsilon > 0$ and sufficiently large k.

For a Gaussian sequence $(\beta_1, \beta_2, \ldots)$, we have $\mathrm{E}\sum_j \beta_j^2 = \sum_j(\mu_j^2 + \tau_j^2)$, and hence $\sum_j \beta_j^2$ converges in mean if (and only if) the given series converges. The series $\sum_j \beta_j^2$ then also converges almost surely. For the necessity of the condition if the variables β_j are independent, we first note that $\beta_j \to 0$ as $j \to \infty$ and normality imply that $\mu_j \to 0$ and $\tau_j \to 0$. Furthermore, by the three-series theorem, convergence of the series $\sum_j \mathrm{E}(\beta_j^2 \mathbb{1}\{|\beta_j| < 1\})$ is necessary for almost sure convergence, and this can be translated into convergence of $\sum_j(\tau_j^2\kappa_{j,2} + 2\tau_j\mu_j\kappa_{j,1} + \mu_j^2\kappa_{j,0})$, where $\kappa_{j,k} = \mathrm{E}(Z^k\mathbb{1}\{A_j\})$ for a standard normal variable Z and $A_j = \{|\tau_j Z + \mu_j| < 1\}$. Since the events A_j increase to the full space as $\mu_j \to 0$ and $\tau_j \to 0$, it follows that $\kappa_{j,k} \to 1$, for $k = 0, 2$, and $\kappa_{j,1} \to 0$. Therefore $\sum_j(\mu_j^2 + \tau_j^2) < \infty$. □

For convergence of an infinite series (2.1) relative to other norms, there exist many precise results in the literature, but they depend on the basis functions, on the prior on the coefficients and on the desired type of convergence. The following lemma gives a simple condition for convergence relative to the *supremum norm* $\|f\|_\infty = \sup_{t\in T}|f(t)|$.

Lemma 2.2 (Uniform norm) *If* $\sum_{j=1}^{\infty}\|\psi_j\|_\infty \mathrm{E}|\beta_j| < \infty$, *then the series* (2.1) *converges uniformly, in mean. If* $\beta_1, \beta_2, \ldots$ *are independent, then the uniform convergence is also in the almost sure sense. If any finite set of coefficients* $(\beta_1, \ldots, \beta_k)$ *has full support* $\mathbb{R}^k$, *then the corresponding prior has support equal to the closure of the linear span of the functions* ψ_j *in the set of bounded functions relative to the uniform norm.*

Proof By the monotone convergence theorem, $\mathrm{E}\sum_{j=1}^{\infty}|\beta_j|\|\psi_j\|_\infty = \sum_{j=1}^{\infty}\mathrm{E}|\beta_j|\|\psi_j\|_\infty < \infty$. This implies the almost sure pointwise absolute convergence of the series $\sum_{j=1}^{\infty}|\beta_j|\psi_j$ and hence its pointwise convergence, almost surely. The uniform convergence in mean of the series is next immediate from the inequality $\mathrm{E}\|\sum_{j>n}\beta_j\psi_j\|_\infty \le \sum_{j>n}|\beta_j|\mathrm{E}\|\psi_j\|_\infty$. That the series converges almost surely if the variables are independent follows by the Ito-Nisio theorem, which says that a series of independent random elements in a separable Banach space converges almost surely if and only if it converges in mean. By Markov's inequality, $\mathrm{P}(\|\sum_{j>k}\beta_j\psi_j\|_\infty < \epsilon) \to 1$, as $k \to \infty$, and hence it is certainly positive for sufficiently large k. If $(\beta_1, \ldots, \beta_k)$ has full support, then $\mathrm{P}(\|\sum_{j\le k}\beta_j\psi_j - f\|_\infty < \epsilon) > 0$, for every $\epsilon > 0$ and every f in the linear span of $\psi_1, \ldots, \psi_k$. Combined this gives the last assertion. □

For a finite set of basis functions, the sum (2.1) is always well defined, but the prior has a finite-dimensional support and hence is not truly nonparametric. This may be remedied by using several dimensions simultaneously, combined with a prior on the dimension.

Basis functions are often constructed so that for certain target functions f of interest, there exist coefficients $f_1, f_2, \ldots$ such that for every $J \in \mathbb{N}$ and k the dimension of the domain of the functions,

$$\Big\| f - \sum_{j=1}^{J} f_j \psi_j \Big\| \lesssim \Big(\frac{1}{J}\Big)^{\alpha/k} \|f\|_\alpha^*. \tag{2.2}$$

Then truncating the basis at J terms allows us to approximate the function, in a given norm $\|\cdot\|$, with error bounded by a power of $1/J$. The power is typically determined by a *regularity parameter* α (e.g. the number of times f is differentiable) divided by the dimension k of the domain. The approximation is moderated by a quantitative measure $\|f\|_\alpha^*$ of the regularity, for instance the norm of the αth derivative. Results of this nature can be found in approximation theory for every of the standard types of bases, including wavelets, polynomials, trigonometric functions and splines, where the norms used in the inequality depend on the type of basis (see Appendix E for a brief review).

A rough statistical message is as follows. The truncated series $\sum_{j=1}^J f_j \psi_j$ gives a J-dimensional parametric model, and hence ought to be estimable with mean square error of the order J/n, for n expressing the informativeness of the data (e.g. n independent replicates), i.e. a mean square error "per parameter" of the order $1/n$. Under (2.2) the bias incurred by estimating the truncation rather than f is of the order $(1/J)^{\alpha/k}$. Equating square bias and mean square error gives

$$\Big(\frac{1}{J}\Big)^{2\alpha/k} \asymp \frac{J}{n}.$$

The solution $J \asymp n^{k/(2\alpha+k)}$ gives an estimation error of the order $n^{-\alpha/(2\alpha+k)}$. This is the typical optimal nonparametric rate of estimation for functional parameters on a k-dimensional domain that are regular of order α. It is attained by standard procedures such as kernel estimation, truncated series, penalized splines, etc.

In the Bayesian setting we put a prior on the coefficients f_j and hope that the posterior distribution will come out right. The informal calculation suggests that if the unknown function is a priori thought to be regular of order α, then a series prior that mostly charges the first $J \asymp n^{k/(2\alpha+k)}$ coefficients might do a good job. For instance, we might put noninformative priors on the first $n^{k/(2\alpha+k)}$ coefficients and put the remaining coefficients equal to zero, or we might make the priors shrink to 0 at an appropriate speed as $j \to \infty$ (e.g. Gaussian priors $\mathrm{Nor}(0, \tau_j^2)$ with τ_j tending to zero). A problem is that the truncation point, or rate of decay, must depend on the regularity α, which we may only have guessed. A natural Bayesian solution is to put a prior on α, or perhaps more simply on the number J of terms in the truncated series, or on the scale of the variance.

We analyze constructions of this type in Chapters 9, 10 and 11.

2.2 Stochastic Processes

A *stochastic process* indexed by a set T is a collection $W = (W(t)\colon t \in T)$ of random variables defined on a common probability space. A *sample path* $t \mapsto W(t)$ of W is a "random function," and hence the law of W, the "law of the set of sample paths," is a prior on a space of functions $f\colon T \to \mathbb{R}$.

For a precise definition, W must be defined as a measurable map $W\colon (\Omega, \mathcal{U}, \mathrm{P}) \mapsto (\mathfrak{F}, \mathscr{F})$ from a probability space into a function space $\mathfrak{F}$, equipped with a σ-field $\mathscr{F}$. This is not always easy, but there are many ready examples of stochastic processes.

2.2.1 Gaussian Processes

A *Gaussian process* W is a stochastic process such that finite-dimensional distributions of $(W(t_1), \ldots, W(t_m))$, for all $t_1, \ldots, t_m \in T$, $m \in \mathbb{N}$, are multivariate normally distributed. The finite-dimensional distributions are completely specified by the *mean function* $\mu\colon T \to \mathbb{R}$ and *covariance kernel* $K\colon T \times T \to \mathbb{R}$, given by

$$\mu(t) = \mathrm{E}[W(t)], \qquad K(s, t) = \mathrm{cov}\,(W(s), W(t)).$$

The mean function is a deterministic shift of the sample paths, and in prior construction the Gaussian process is often taken with zero mean and a desired offset is made part of the statistical model, possibly with its own prior.

The finite-dimensional distributions determine the process as a measurable map in $\mathbb{R}^T$, but for uncountable domains T, they do not determine the sample paths. However, there may exist a version with special sample paths, often regular of some type, and the process may be seen as a map into the space of functions of this type. For instance, a Brownian motion is a Gaussian process with $T = [0, \infty)$ with mean zero and covariance function $\min(s, t)$, and can be realized to have continuous sample paths (which are then even regular of order 1/2), so that it can be viewed as a map in the space of continuous functions.

A function $K\colon T \times T \to \mathbb{R}$ is called *nonnegative-definite* if all matrices of the type $((K(t_i, t_j)))$, for $t_1, \ldots, t_m \in T$ and $m \in \mathbb{N}$, are nonnegative. Any such nonnegative-definite function is the covariance function of some Gaussian process. Thus there is a great variety of such processes. For some applications with $T \subset \mathbb{R}^k$ covariance functions of the type $K(s, t) = \psi(s - t)$ are of special interest as they generate *stationary processes*: the corresponding (mean zero) process is invariant in law under translations: the distribution of $(W(t + h)\colon t \in T)$ is the same for every $h \in \mathbb{R}^k$.

Many Gaussian processes have "full" support, for instance in a space of continuous functions, and therefore are suitable to model a function with unrestricted shape taking values in $\mathbb{R}$. However, their performance as a prior strongly depends on the fine properties of the covariance function and sample paths. This performance may also be influenced by changing the *length scale* of the process: using $t \mapsto W(\lambda t)$ as a prior process, for some positive number λ, possibly chosen itself from a prior distribution.

A random series (2.1) with independent normal variables β_j as coefficients, is an example of a Gaussian process prior. Conversely, every Gaussian process can be represented as a random series with independent normal coefficients (see Theorem I.25). The special choice with the basis functions ψ_j equal to the eigenfunctions of the covariance kernel is called the *Karhunen-Loève expansion*.

Gaussian priors are conjugate in the Gaussian nonparametric regression model. Outside this linear Gaussian setting, computation of the posterior distribution is considerably more complicated. We discuss computational schemes and many results on Gaussian priors in Chapter 11.

2.2.2 Increasing Processes

Because Gaussian variables can assume every real value, Gaussian processes are unsuitable as priors for monotone functions $f\colon \mathbb{R} \to \mathbb{R}$ or $f\colon [0,\infty) \to \mathbb{R}$, such as cumulative distribution and hazard functions, or for monotone link functions in generalized linear models. Independent increment processes with nonnegative increments are the most popular examples of priors for increasing functions. Gamma processes and, more generally, Lévy processes with nonnegative increments (*subordinators*, see Appendix J) are in this class. These processes have the remarkable property that they increase by jumps only. Their sample paths either converge to a finite positive random variable or diverge to infinity, a.s.

Given an increasing process $S = (S(t)\colon t \geq 0)$, we may define a random cumulative distribution function (c.d.f.) F on $[0,\tau]$ by

$$F(t) = S(t)/S(\tau),$$

for any $\tau < \infty$, and also for $\tau = \infty$ if the sample paths of S are bounded. For S the gamma process, this leads to the Dirichlet process, discussed in Chapter 4.

An increasing process with unbounded sample paths may be used as a random cumulative hazard function, which will subsequently induce a prior on distributions. This approach is fruitful in the context of Bayesian survival analysis, discussed in Chapter 13. A similar but technically different idea is to consider a random cumulative distribution function defined by $F(t) = 1 - e^{-S(t)}$. If S is not continuous, then it is not the cumulative hazard function of F, but it can be close to it.

Increasing processes, and hence priors on sets of increasing functions, can be constructed without the measure-theoretic problems associated with the construction of (general) random measures. We focus on increasing functions with domain $\mathbb{R}$. The idea is to specify the desired distribution of every finite-dimensional marginal $(F(t_1),\ldots,F(t_k))$, with $t_1,\ldots,t_k$ belonging to a given countable, dense subset of $\mathbb{R}$ – for instance, the set of rational numbers $\mathbb{Q}$. Provided these distributions are consistent, Kolmogorov's consistency theorem allows construction of a full process. The following proposition shows that this will automatically have right-continuous, nondecreasing sample paths if the bivariate and univariate marginal distributions satisfy certain conditions.

Proposition 2.3 *For every $k \in \mathbb{N}$ and every $t_1,\ldots,t_k$ in a countable dense set $Q \subset \mathbb{R}$ let $\mu_{t_1,\ldots,t_k}$ be a probability distribution on $\mathbb{R}^k$ such that*

(i) *if $T' \subset T$, then $\mu_{T'}$ is the marginal distribution obtained from μ_T by integrating out the coordinates $T \setminus T'$;*

(ii) $t_1 < t_2$ *implies* $\mu_{t_1,t_2}\{(x_1, x_2)\colon x_1 \leq x_2\} = 1$;
(iii) $t_n \downarrow t$ *implies* $\mu_{t_n} \rightsquigarrow \mu_t$.

Then there exists a stochastic process $(F(t)\colon t \in \mathbb{R})$ *with nondecreasing, right-continuous sample paths such that* $(F(t_1), \ldots, F(t_k))$ *has distribution* $\mu_{t_1,\ldots,t_k}$ *for every* $t_1, \ldots, t_k \in Q$ *and* $k \in \mathbb{N}$. *If, moreover,*

(iv) $t \downarrow -\infty$ *implies* $\mu_t \rightsquigarrow \delta_0$ *and* $t \uparrow \infty$ *implies* $\mu_t \rightsquigarrow \delta_1$, *for* δ_x *the degenerate measure at* x,

then the sample paths of the stochastic process F *are cumulative distribution functions on* $\mathbb{R}$.

Proof Because the marginal distributions are consistent by assumption (i), Kolmogorov's consistency theorem implies existence of a stochastic process $(F(t)\colon t \in Q)$ with the given marginal distributions. Because only countably many exceptional null sets arise in relation (ii), the sample paths of F are nondecreasing a.s. To show a.s. right continuity on Q, fix $t \in Q$ and let $t_k \downarrow t$, $t_k \in Q$. Monotonicity gives that $F^*(t) := \lim_{k\to\infty} F(t_k) \geq F(t)$ a.s., while property (iii) implies that $F(t_k) \rightsquigarrow F(t)$. This implies that $F^*(t) = F(t)$ a.s. (see Problem L.1). Again the exceptional null sets can be pulled together to a single null set, which implies a.s. right continuity simultaneously at every point.

Any right-continuous, nondecreasing function $F\colon Q \to \mathbb{R}$ can be uniquely extended to a function $F\colon \mathbb{R} \to \mathbb{R}$ with the same properties by simply taking a right limit at every $t \in \mathbb{R} \setminus Q$.

This function is a cumulative distribution function if and only if its limits at $-\infty$ and ∞ are 0 and 1, respectively. Under the additional assumption (iv), this can be established a.s. $[\mu]$ by a similar argument. □

Example 2.4 (Subordinators) As an example, consider the construction of *subordinators*, with given marginal distributions μ_t. Assume that $(\mu_t\colon t \geq 0)$ are a continuous semigroup (i.e. $\mu_{s+t} = \mu_s * \mu_t$ for all $s, t \geq 0$ and $\mu_t \rightsquigarrow \delta_0$ as $t \to 0$) of probability measures concentrated on $(0, \infty)$, and define $\mu_{t_1,\ldots,t_k}$ for rational numbers $0 < t_1 < t_2 < \cdots < t_k$ as the joint distribution of the cumulative sums of k independent variables with distributions $\mu_{t_1}, \mu_{t_2-t_1}, \ldots, \mu_{t_k-t_{k-1}}$, respectively. The semigroup properties make the marginal distributions consistent and ensure property (iii), whereas (ii) is immediate from the nonnegativity of the support of μ_t. The random distribution function F has independent increments with $F(t) - F(s) \sim \mu_{t-s}$.

2.3 Probability Densities

Probability densities are functions that are nonnegative and integrate to 1. In this section we discuss priors that observe these constraints. A prior induced by a prior on distributions that charges absolutely continuous distributions only (e.g. certain Pólya tree processes, Section 3.7) would be an alternative.

Natural metrics on the set of all densities relative to a given dominating measure μ are the *total variation distance* and the *Hellinger distance*

$$d_{TV}(p,q) = \tfrac{1}{2}\int |p-q|\,d\mu,$$

$$d_H(p,q) = \sqrt{\int (\sqrt{p}-\sqrt{q})^2\,d\mu}.$$

For a standard Borel sample space and a σ-finite dominating measure, the set of μ-densities is Polish under these metrics, and hence the support is well defined. Because the two distances induce the same topology, the total variation and Hellinger supports are the same.

It will be seen in Chapter 6 that support relative to the Kullback-Leibler divergence is crucial for posterior consistency. The *Kullback Leibler divergence* between two densities p and q is defined as

$$K(p;q) = \int \Big(\log\frac{p}{q}\Big)\,p\,d\mu.$$

Because it is not a metric, the notion of a "support" does not follow from general principles. We say that a density p_0 is in the *Kullback-Leibler support* of a prior Π if, for all $\epsilon > 0$,

$$\Pi\Big(p\colon K(p_0;p) < \epsilon\Big) > 0.$$

Because $d_H^2(p,q) \le K(p;q)$, for any pair of probability densities p, q, the Kullback-Leibler support is always smaller than the Hellinger support. See Appendix B for further details on these distances and discrepancies.

2.3.1 Exponential Link Function

Given a measurable function $f\colon \mathfrak{X} \to \mathbb{R}$ with a μ-integrable exponential e^f, we can define a probability density p_f by

$$p_f(x) = e^{f(x)-c(f)}, \qquad (2.3)$$

where $c(f) = \log\int e^f\,d\mu$ is the norming constant. We can construct a prior on densities by putting a prior on the function f.

If the prior on f is a random series, then the induced prior (2.3) is an *exponential family*

$$\bar{p}_\beta(x) = \exp\Big(\sum_j \beta_j\psi_j(x) - c(\beta)\Big).$$

The basis functions ψ_j are the sufficient statistics of the family. If the series is infinite, then the exponential family is also of infinite dimension.

The prior (2.3) will have full support relative to the Kullback-Leibler divergence as soon as the prior on f has full support relative to the uniform norm. This follows from the following lemma, which even shows that the Hellinger distance and Kullback-Leibler divergence between two densities of the form (2.3) are essentially bounded by the uniform norm between the corresponding exponents. The discrepancy $V_2(p,q)$ is the square Kullback-Leibler variation, defined in (B.2).

Lemma 2.5 *For any measurable functions $f, g\colon \mathfrak{X} \to \mathbb{R}$,*

(i) $d_H(p_f, p_g) \le \|f - g\|_\infty \, e^{\|f-g\|_\infty/2}$,
(ii) $K(p_f; p_g) \lesssim \|f - g\|_\infty^2 \, e^{\|f-g\|_\infty}(1 + \|f - g\|_\infty)$,
(iii) $V_2(p_f; p_g) \lesssim \|f - g\|_\infty^2 \, e^{\|f-g\|_\infty}(1 + \|f - g\|_\infty)^2$.

Furthermore, if $f(x_0) = g(x_0)$ for some $x_0 \in \mathfrak{X}$, then

(iv) $\|f - g\|_\infty \le 2D(f, g)\|p_f - p_g\|_\infty$, *for* $D(f, g) = (\min_x(p_f(x) \wedge p_g(x))^{-1}$.

Proof The triangle inequality and simple algebra give

$$d_H(p_f, p_g) = \left\| \frac{e^{f/2}}{\|e^{f/2}\|_2} - \frac{e^{g/2}}{\|e^{g/2}\|_2} \right\|_2 \le 2\frac{\|e^{f/2} - e^{g/2}\|_2}{\|e^{g/2}\|_2}.$$

Because $|e^{f/2} - e^{g/2}| \le e^{g/2}e^{|f-g|/2}|f - g|/2$ for any $f, g \in \mathbb{R}$, the square of the right side is bounded by

$$\frac{\int e^g e^{|f-g|}|f - g|^2 \, d\nu}{\int e^g \, d\nu} \le e^{\|f-g\|_\infty} \|f - g\|_\infty^2.$$

This proves assertion (i) of the lemma.

Next assertions (ii) and (iii) follow from (i) and the equivalence of K, V_2 and the squared Hellinger distance if the quotient of the densities is uniformly bounded (see Lemma B.2). From $g - \|f - g\|_\infty \le f \le g + \|f - g\|_\infty$ it follows that $c(g) - \|f - g\|_\infty \le c(f) \le c(g) + \|f - g\|_\infty$. Therefore $\|\log(p_f/p_g)\|_\infty = \|f - g - c(f) + c(g)\|_\infty \le 2\|f - g\|_\infty$. Assertions (ii) and (iii) now follow by Lemma B.2.

For the final assertion we note first that $|f - g - c(f) + c(g)| = |\log p_f - \log p_g|$ is bounded above by $D(f, g)|p_f - p_g|$. Evaluating this as $x = x_0$, we see that $|c(f) - c(g)| \le D(f, g)|p_f(x_0) - p_g(x_0)|$. The result follows by reinserting this in the initial inequality. □

2.3.2 Construction through Binning

Consider a partition of $\mathbb{R}$ into intervals of length $h > 0$, and randomly distribute probabilities to the intervals, for instance by the techniques to be discussed in Section 3.3, and distribute the mass uniformly inside the intervals. This yields a random histogram of bandwidth h. Finally, give h a nonsingular prior that allows arbitrary small values of h with positive probability.

Since any integrable function can be approximated in $\mathbb{L}_1$-distance by a histogram of sufficiently small bandwidth, the resulting prior has full $\mathbb{L}_1$-support if the weights themselves are chosen to have full support equal to the countable unit simplex.

The idea is easily extendable to $\mathbb{R}^k$ or regularly shaped subsets of $\mathbb{R}^k$.

2.3.3 Mixtures

A powerful technique to construct a prior on densities via one on probabilities is through mixtures. For given probability density functions $x \mapsto \psi(x, \theta)$ indexed by a parameter θ and measurable in its two arguments, and a probability distribution F, let

$$p_F(x) = \int \psi(x, \theta)\, dF(\theta).$$

Then a prior on F induces a prior on densities. The function ψ is referred to as the *kernel function* and the measure F as the *mixing distribution*.

The choice of a kernel will depend on the sample space. For the entire Euclidean space, a location-scale kernel, such as the normal kernel, is often appropriate. For the unit interval, beta distributions form a flexible two-parameter family. On the positive half line, mixtures of gamma, Weibull or lognormal distributions may be used. The use of a uniform kernel leads to random histograms as discussed in the method of binning in Section 2.3.2.

Mixtures can form a rich family. For instance, location-scale mixtures, with a kernel of the form $x \mapsto \sigma^{-1}\psi((x-\mu)/\sigma)$ for some fixed density ψ and $\theta = (\mu, \sigma)$, may approximate any density in the $\mathbb{L}_1$-sense if σ is allowed to approach 0. This is because

$$\int \frac{1}{\sigma}\psi\left(\frac{\cdot - \mu}{\sigma}\right) f(\mu)\, d\mu \to f(\cdot), \qquad \text{as } \sigma \to 0, \tag{2.4}$$

in $\mathbb{L}_1$ by *Fejér's theorem* (cf. Helson (1983)). Thus, a prior on densities may be induced by putting a prior on the mixing distribution F and a prior on σ, which supports zero. Similarly, one can consider location-scale mixture $\int \sigma^{-1}\psi((x-\mu)/\sigma)\, dF(\mu, \sigma)$, where F is a bivariate distribution on $\mathbb{R} \times (0, \infty)$.

The mixing distribution F may be given a prior following any of the methods discussed in Chapter 3. A Dirichlet process prior is particularly attractive, as it leads to efficient computational algorithms (cf. Chapter 5) and a rich theory of convergence properties (cf. Chapter 7 and 9).

2.3.4 Feller Approximation

In the previous section several different choices of kernels were mentioned, mostly motivated by the underlying sample space (full line, half line, interval, etc.). The choice of kernel is also intimately connected to the approximation scheme it generates. For instance, motivation for a location-scale kernel comes from (2.4), whereas beta mixtures can be based on the theory for Bernstein polynomials, as reviewed in Section E.1. Feller (1968) described a general constructive approximation scheme for continuous functions on an interval, which may be used to propose a kernel in a canonical way.

A *Feller random sampling scheme* on an interval I is a collection $\{Z_{k,x}: k \in \mathbb{N}, x \in I\}$ of random variables such that $\mathrm{E}(Z_{k,x}) = x$ and $V_k(x) := \mathrm{var}(Z_{k,x}) \to 0$ as $k \to \infty$. Then $Z_{k,x} \rightsquigarrow \delta_x$, and hence, for any bounded, continuous real-valued function F,

$$A(x; k, F) := \mathrm{E}(F(Z_{k,x})) \to F(x), \qquad k \to \infty.$$

Thus $A(\cdot; k, F)$ is an approximation of the function F. If $Z_{k,x}$ possesses a density $z \mapsto g_k(z; x)$ relative to some dominating measure ν_k, then we can write the approximation in the form

$$A(x; k, F) = \int F(t) g_k(t; x)\, d\nu_k(t) = \int\int_{[z,\infty)} g_k(t; x)\, d\nu_k(t)\, dF(z).$$

Thus a (continuous) distribution function F can be approximated using mixtures of the kernel $\bar{G}_k(z;x) := \int_{[z,\infty)} g_k(t;x)\,d\nu_k(t)$. Under suitable conditions mixtures of the derivative $(\partial/\partial x)\bar{G}_k(z;x)$ of this kernel will approximate a corresponding density function.

An example of a Feller scheme is $Z_{k,x} = k^{-1}\sum_{i=1}^k Y_{i,x}$, for i.i.d. random variables $Y_{i,x}$ with mean x and finite variance. Consider specializing the distribution of $Y_{i,x}$ to a natural exponential family, given by a density of the form $p_\theta(y) = \exp(\theta y - \psi(\theta))$ relative to some σ-finite dominating measure independent of θ. The density $z \mapsto g_k(z;x)$ of $Z_{k,x}$ relative to some (other) dominating measure ν_k can then be written in the form $g_k(z;x) = \exp(k\theta z - k\psi(\theta))$, and the usual identities for exponential families show that $\mathrm{E}(Z_{k,x}) = \mathrm{E}Y_{i,x} = \psi'(\theta)$ and $\mathrm{var}(Y_{i,x}) = \psi''(\theta)$. The first implies that the parameter $\theta(x)$ of the family must be chosen to solve $\psi'(\theta(x)) = x$. Differentiating across this identity with respect to x and using the second identity we see that $\theta'(x) = 1/\psi''(\theta(x)) = 1/V(x)$ for $V(x) = \mathrm{var}(Y_{i,x})$. This permits us to express $(\partial/\partial x)g_k(z;x)$ as $kg_k(z;x)(z-x)/V(x)$, leading to the approximation scheme for the density of the cumulative distribution function F given by

$$a(x;k,F) = \int h_k(x;z)\,dF(z), \tag{2.5}$$

for

$$h_k(x;z) = \frac{k}{V(x)}\int_{[z,\infty)} (t-x)g_k(t;x)\,d\nu_k(t).$$

The following proposition shows that this mixture indeed approaches $f = F'$ as $k \to \infty$. It is implicitly assumed that the equation $\psi'(\theta(x)) = x$ is solvable: the approximation scheme only works on the interval of possible mean values of the given exponential family (the image of its natural parameter set under the increasing map ψ').

Proposition 2.6 *If F has bounded and continuous density f, then $a(x;k,F) \to f(x)$ as $k \to \infty$.*

Proof Without loss of generality set $x = 0$; otherwise redefine $Z_{k,x}$ to $Z_{k,x} - x$. Abbreviate $Z_{k,0}$ to Z_k, $V(0)$ to V, and $h_k(0;z)$ to $h_k(z)$. The latter function can be written in the form $h_k(z) = k\mathrm{E}\mathbb{1}\{Z_k \geq z\}Z_k/V$. It is nonnegative for every z, as is immediate for $z \geq 0$, and follows for $z \leq 0$ by rewriting it as $-k\mathrm{E}\mathbb{1}\{Z_k < z\}Z_k/V$, using the fact that $\mathrm{E}Z_k = 0$. Furthermore, by Fubini's theorem, for every z,

$$\bar{H}_k(z) := \int_z^\infty h_k(t)\,dt = \mathrm{E}kZ_k^2\mathbb{1}\{Z_k \geq z\}/V - z\mathrm{E}kZ_k\mathbb{1}\{Z_k \geq z\}/V.$$

For fixed k the first term tends to $\mathrm{E}kZ_k^2/V = 1$ and to 0 as z tends to $-\infty$ or ∞, respectively. Because $\mathrm{E}Z_k = 0$, the second term can also be written $-z\mathrm{E}kZ_k\mathbb{1}\{Z_k < z\}/V$. Therefore its absolute value is bounded above by $\mathrm{E}kZ_k^2\mathbb{1}\{Z_k < z\}/V$ as $z < 0$ and by $\mathrm{E}kZ_k^2\mathbb{1}\{Z_k \geq z\}/V$ as $z > 0$ and hence it tends to zero both as $z \to -\infty$ and as $z \to \infty$. We conclude that $\bar{H}_k$ is a survival function, for every fixed k.

By the law of large numbers $\mathrm{P}(Z_k > z) \to 0$ as $k \to \infty$ for $z > 0$. Because $\sqrt{k}Z_k$ is uniformly square integrable in k (it has variance V and tends to a Nor$(0,V)$-variable), it follows for $z > 0$ by the same bounds as previously that $\bar{H}_k(z) > 2\mathrm{E}kZ_k^2\mathbb{1}\{Z_k \geq z\}/V \to 0$, as $k \to \infty$. Similarly we see that $\bar{H}_k(z) \to 1$ for any $z < 0$. We conclude that $H_k \rightsquigarrow \delta_0$, whence $a(0; k, F) = \int f(z)\, dH_k(z) \to f(0)$, by the assumed continuity and boundedness of f. □

2.4 Nonparametric Normal Regression

Consider a continuous response variable Y which depends on the value of a covariate X through the regression equation $Y = f(X) + \varepsilon$, for a function $f\colon \mathfrak{X} \to \mathbb{R}$ on a measurable space $\mathfrak{X}$, and a stochastic error variable ε. If $\varepsilon |\, X \sim \mathrm{Nor}(0, \sigma^2)$, then the conditional density of Y given $X = x$ is

$$p_{f,\sigma}(y|\, x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-f(x))^2/(2\sigma^2)}.$$

In the *random design* model the covariate X is a random variable. We denote its marginal distribution by G and write $p_{f,\sigma}$ for the joint density of a single observation (X, Y), relative to the product of G and Lebesgue measure. In the *fixed design* model the covariates are deterministic values $x_1, \dots, x_n$. We denote their empirical distribution by G, and use the same notation $p_{f,\sigma}$ for the density of a vector of $(Y_1, \dots, Y_n)$ of independent responses at these covariate values.

A prior on (f, σ) induces a prior on $p_{f,\sigma}$. The following lemma shows that a prior on f with large support relative to the $\mathbb{L}_2(G)$-norm combined with a suitable independent prior on σ induces a prior with a large Hellinger or Kullback-Leibler support.

Lemma 2.7 *For any measurable functions $f, g\colon \mathfrak{X} \to \mathbb{R}$ and any $\sigma, \tau > 0$,*

(i) $d_H(p_{f,\sigma}, p_{g,\tau}) \le \|f - g\|_{2,G}/(\sigma + \tau) + 2|\sigma - \tau|/(\sigma + \tau)$.
(ii) $K(p_{f,\sigma}; p_{g,\tau}) = \frac{1}{2}\|f - g\|_{2,G}^2/\tau^2 + \frac{1}{2}\big[\log(\tau^2/\sigma^2) + (\sigma^2/\tau^2) - 1\big]$.
(iii) $V_{2,0}(p_{f,\sigma}; p_{g,\tau}) = (\sigma/\tau)^2\|f - g\|_{2,G}^2/\tau^2 + \frac{1}{2}(\sigma^2 - \tau^2)^2/\tau^4$.

Proof For (i) we first find by direct calculation using the Gaussian integral, that the affinity between two univariate normal distributions is given by

$$\rho_{1/2}\big(\mathrm{Nor}(f, \sigma^2), \mathrm{Nor}(g, \tau^2)\big) = \sqrt{1 - (\sigma - \tau)^2/(\sigma^2 + \tau^2)}\, e^{-|f-g|^2/(4\sigma^2+4\tau^2)}.$$

Next we integrate with respect to G to find the affinity between $p_{f,\sigma}$ and $p_{g,\tau}$, use the relationship between the square Hellinger distance and affinity (see Lemma B.5(i)), and finally use the inequalities $|1 - \sqrt{1-s}| \le s$, for $s \in [0, 1]$, and $1 - e^{-t} \le t$, for $t \ge 0$. Equalities (ii) and (iii) similarly can first be established when G is a Dirac measure by calculations on the normal distribution, and next be integrated for a general G. □

2.5 Nonparametric Binary Regression

In the binary regression model we try to predict a binary outcome (zero-one response, membership in one of two classes, success or failure, etc.) by the value of a covariate x. If we model the outcome as a Bernoulli variable with success probability given as a function $x \mapsto p(x)$ of the covariate, then it is natural to predict (or classify) a new example x according to whether $p(x)$ exceeds a certain threshold $c \in [0, 1]$ or not (e.g. $c = 1/2$).

To estimate the function p from “training data” we put a prior on the set of functions $p: \mathfrak{X} \to [0, 1]$ on the covariate space. Because the prior should observe that the range is the unit interval, a general prior on functions is unsuitable. This may be remedied by using a (inverse) *link function* $H: \mathbb{R} \to [0, 1]$ that maps the real line to the unit interval with derivative $h = H'$. We may then use a general prior on functions f and model the probability response function as

$$p(x) = H(f(x)).$$

Monotone link functions (in particular cumulative distribution functions) are customary. In particular, the *logistic link function* $H(t) = (1 + e^{-t})^{-1}$ and *Gaussian link function* $H(t) = \Phi(t)$, the cumulative distribution function of the standard normal distribution, are popular.

The conditional density of a Bernoulli variable $Y|\, x \sim \mathrm{Bin}(1, H(f(x))$ is

$$p_f(y|\, x) = H(f(x))^y (1 - H(f(x)))^{1-y}, \qquad y \in \{0, 1\}. \tag{2.6}$$

For sampling instances of (X, Y), this is the relevant density (relative to the marginal distribution of the covariate X and counting measure on $\{0, 1\}$), if we assume that the covariate distribution is not informative about the response.

The following lemma allows us to derive the support of the induced prior on p_f from the support of the prior on f. For the logistic link function, the induced prior has full Kullback-Leibler support as soon as the prior on f has full support relative to the $\mathbb{L}_2(G)$-metric, for G the marginal distribution of the covariate. The Gaussian link is slightly less well behaved, but a similar result holds with the uniform norm if f_0 is bounded.

Lemma 2.8 *For any measurable functions $f, g: \mathfrak{X} \to \mathbb{R}$ and any $r \geq 1$,*

(i) $\|p_f - p_g\|_r = 2^{1/r}\|H(f) - H(g)\|_{r,G} \leq 2^{1/r}\|h\|_\infty \|f - g\|_{2,G}$.
(ii) $d_H(p_f, p_g) \leq \sqrt{I(h)}\|f - g\|_{2,G}$, *for* $I(h) = \int (h'/h)^2\, dH$.
(iii) $K(p_f; p_g) \lesssim \|(f - g)S(|f| \vee |g|)^{1/2}\|_{2,G}^2$.
(iv) $V_2(p_f; p_g) \lesssim \|(f - g)S(|f| \vee |g|)\|_{2,G}^2$.

Here $S = 1$ for H the logistic link and $S(u) = |u| \vee 1$ for the Gaussian link. Furthermore, the inequalities are true with $S = 1$ for any strictly monotone, continuously differentiable link function if the functions f, g take values in a fixed compact interval.

Proof Assertion (i) follows, because, for any x,

$$|p_f(0|\, x) - p_g(0|\, x)| = |p_f(1|\, x) - p_g(1|\, x)| = |H \circ f(x) - H \circ g(x)|.$$

The second inequality follows by bounding the right side by the mean value theorem. For (ii) we argue similarly, but must bound the differences $|\sqrt{H \circ f} - \sqrt{H \circ g}|$ and $|\sqrt{1 - H \circ f} - \sqrt{1 - H \circ g}|$. The appropriate derivatives are now $\frac{1}{2}h/\sqrt{H}$ and $\frac{1}{2}h/\sqrt{1 - H}$, which are both bounded by $\frac{1}{2}\sqrt{I(h)}$, since $h(x)^2 = |\int_{-\infty}^x h'(s)\, ds|^2 \leq I(h)H(x)$, in view of the Cauchy-Schwarz inequality, and the analogous inequality in the right tail.

To derive (iii) we express the Kullback-Leibler divergence as $K(p_f; p_g) = \int \psi_f(g)\, dG$, where

$$\psi_u(v) = H(u) \log \frac{H(u)}{H(v)} + (1 - H(u)) \log \frac{1 - H(u)}{1 - H(v)}.$$

The function ψ_u possesses derivative $\psi_u'(v) = (h/(H(1-H)))(v)(H(v)-H(u))$. Here the function $h/(H(1-H))$ is identically equal to 1 for the logistic link and is asymptotic to the identity function, as its argument tends to ±∞ for the Gaussian link; hence it is bounded in absolute value by the function S as given. Because the function ψ_u vanishes at u, the mean value theorem gives that $|\psi_f(g)| \leq S(|f| \vee |g|)|f-g|$, and assertion (iii) follows.

For the proof of (iv) we write $V_2(p_f, p_g) = \int \phi_f(g)\, dG$, where the function ϕ_f is as ψ_f, but with the logarithmic factors squared. The result follows because both $\log(H(g)/H(f))$ and $\log[(1-H(g))/(1-H(f)]$ are bounded in absolute value by $S(|f| \vee |g|)|f-g|$. □

2.6 Nonparametric Poisson Regression

In the Poisson regression model, a response variable is Poisson distributed with mean parameter $\lambda(x)$ depending on a covariate x. In the nonparametric situation, $x \mapsto \lambda(x)$ is an unknown function $\lambda\colon \mathfrak{X} \to (0, \infty)$. A link function $H\colon \mathbb{R} \to (0, \infty)$ maps a general function $f\colon \mathbb{R} \to \mathbb{R}$ to the positive half line through

$$\lambda(x) = H(f(x)).$$

A prior on f induces a prior on λ. The exponential link function $H(f) = e^f$ is a canonical choice, but functions that increase less rapidly may have advantages, as seen in Example 2.11, below.

The conditional density of a Poisson variable $Y|\,x \sim \mathrm{Poi}(H(f(x))$ is given by

$$p_f(y|\,x) = \frac{1}{y!} e^{-H(f(x))} H(f(x))^y, \qquad y = 0, 1, 2, \ldots.$$

The following lemma relates the relevant distances on p_f to the $\mathbb{L}_2(G)$-distance on f, for G, the marginal distribution of the variable x.

Lemma 2.9 *For any measurable functions* $f, g\colon \mathfrak{X} \to (a, b) \subset \mathbb{R}$, *and* $\|K\|_\infty = \sup\{|K(s)|\colon a < s < b\}$, *the uniform norm of a function with domain* (a, b),

(i) $d_H(p_f, p_g) \leq \|H'/\sqrt{H}\|_\infty \, \|f-g\|_{2,G}$;
(ii) $K(p_f; p_g) \leq (C_f \|H'/H\|_\infty^2 + C_f \|H''/H\|_\infty + \|H''\|_\infty)\, \|f-g\|_{2,G}^2$, *for* $C_f = \sup_x |H(f(x))|$;
(iii) $V_{2,0}(p_f; p_g) \leq C_f \|H'/H\|_\infty^2 \, \|f-g\|_{2,G}^2$.

Proof The three distances between Poisson distributions Q_λ with mean λ are given by

$$d_H^2(Q_\lambda, Q_\mu) = 2(1 - e^{-(\sqrt{\lambda}-\sqrt{\mu})^2/2}) \leq (\sqrt{\lambda} - \sqrt{\mu})^2,$$
$$K(Q_\lambda; Q_\mu) = \mu - \lambda + \lambda \log \frac{\lambda}{\mu},$$
$$V_{2,0}(Q_\lambda; Q_\mu) = \lambda\Big(\log \frac{\lambda}{\mu}\Big)^2.$$

The quantities in the lemma are obtained by replacing λ and μ in the right sides by $H(f(x))$ and $H(g(x))$, and integrating over x with respect to G. For (i) and (iii) we use that the functions $g \mapsto \sqrt{H(g)} - \sqrt{H(f)}$ and $g \mapsto \sqrt{H(f)} \log H(f)/H(g)$ have first derivatives $\frac{1}{2}H'/\sqrt{H}$ and $-\sqrt{H(f)}(H'/H)$, respectively. For (ii) we use that the first derivative of the function $g \mapsto H(g) - H(f) + H(f) \log H(f)/H(g)$ vanishes at $g = f$ and that the second derivative of the function takes the form $H'' - H(f)(H''/H) + H(f)(H'/H)^2$. □

Example 2.10 (Exponential link) For the exponential link function $H(x) = e^x$ we may take the interval (a, b) equal to $(-\infty, M)$, for some constant M. Then the constants $\|H'/\sqrt{H}\|_\infty$ and $\|H''\|_\infty$ in the lemma are bounded by e^M, while the constants $\|H'/H\|_\infty$ and $\|H''/H\|_\infty$ are equal to 1. If f takes its values in $(-\infty, M]$, then $C_f \leq e^M$, and all multiplicative constants in the lemma are bounded by a multiple of e^M.

Example 2.11 (Quadratically increasing link) The function $H: \mathbb{R} \to (0, \infty)$ defined by $H(x) = e^x$, for $x \leq 1$, and $H(x) = e + e(x-1) + e(x-1)^2/2$, for $x \geq 1$, is twice continuously differentiable, and such that the constants $\|H'/\sqrt{H}\|_\infty$, $\|H''\|_\infty$, $\|H'/H\|_\infty$ and $\|H''/H\|_\infty$ are all finite when computed for the full real line $(a, b) = \mathbb{R}$. Thus, for this link function, the Hellinger distance on the densities p_f is bounded by the $\mathbb{L}_2(G)$-distance on the functions f. If f is bounded above, then $C_f < \infty$, and the Kullback-Leibler discrepancies between p_f and any other density p_g will be bounded by the square $\mathbb{L}_2(G)$-distance between f and g.

2.7 Historical Notes

Random basis expansions have been in use for a long time, beginning with the work of Poincaré. The term "infinite-dimensional exponential family" seems to be first mentioned by Verdinelli and Wasserman (1998) and Barron et al. (1999). Gaussian processes as priors for density estimation were first used by Leonard (1978) and later by Lenk (1988). Nonparametric mixtures of kernels with a Dirichlet process as mixing distribution appeared in Ghorai and Rubin (1982), Ferguson (1983) and Lo (1984). Feller approximation schemes to construct priors were considered by Petrone and Veronese (2010). Lemmas 2.5 and 2.8 come from van der Vaart and van Zanten (2008a).

Problems

2.1 (Hjort 1996) For density estimation on $\mathbb{R}$, use finite Hermite polynomial expansion

$$\sum_{j=1}^{m} \phi_\sigma(x-\mu)\Big[1 + \sum_{j=3}^{m} \frac{\kappa_j}{j!} H_j((x-\mu)/\sigma)\Big]$$

for large m, where κ_js are cumulants and the Hermite polynomials H_js are given by $(d^j/dx^j)\phi(x) = (-1)^j H_j(x)\phi(x)$, to construct a prior on a density.

The above method has the drawback that cumulants need not be finite for many densities. Show that the alternative expansion

$$\sum_{j=1}^{m} \phi_\sigma(x-\mu)\Big[\sum_{j=0}^{m} \frac{\delta_j}{\sqrt{j!}} H_j(\sqrt{2}(x-\mu)/\sigma)\Big]$$

leads to finite values of δ_js for any density and hence can be used for constructing a prior on a probability density.

2.2 (Petrone and Veronese 2010) Show that $a(x; k, F)$ in (2.5) is indeed the density of $A(x; k, F)$.

2.3 (Petrone and Veronese 2010) Show that in a Feller sampling scheme:

(a) if the sampling scheme is based on $Y \sim \mathrm{Nor}(x, \sigma^2)$, then the kernel $h_k(x; z)$ is normal;
(b) if the sampling scheme is based on $Y \sim \mathrm{Bin}(1, x)$, then the kernel $h_k(x; z)$ is a Bernstein polynomial;
(c) if the sampling scheme is based on $Y \sim \mathrm{Poi}(x)$, then the kernel $h_k(x; z)$ is gamma;
(d) if the sampling scheme is based on $Y \sim \mathrm{Ga}(x)$, then the kernel $h_k(x; z)$ is inverse-gamma.

2.4 (Log Lipschitz link function) Let $\Psi: \mathbb{R} \to (0, \infty)$ be a monotone function whose logarithm $\log \Psi$ is uniformly Lipschitz with constant L. For a function $f: \mathfrak{X} \to \mathbb{R}$ on a measurable space $\mathfrak{X}$ such that $c(f) := \int \Psi(f)\, d\mu < \infty$, define the probability density $p_f(x) = \Psi(f(x))/c(f)$. Show that $d_H(p_f, p_g) \lesssim L\|f - g\|_\infty e^{L\|f-g\|_\infty}$ and $(K + V_{2,0})(p_f; p_g) \lesssim L^2\|f - g\|_\infty e^{4L\|f-g\|_\infty}$. This generalizes Lemma 2.5 (i) to (iii), which consider the special case that Ψ is the exponential function. [Hint: Derive first that $|\Psi(f)^a - \Psi(g)^a| \le \Psi(f)^a e^{aL\epsilon}$ if $\|f - g\|_\infty = \epsilon$, for $a = 1, 2$. This yields the bound on the Hellinger distance as in the proof of Lemma 2.5. Furthermore, it shows that $c(g) \le c(f)(1 + L\epsilon e^{L\epsilon})$, which together with the Lipschitz property gives a uniform bound on $\log(p_f/p_g)$.]

2.5 (Lipschitz link function) For given measurable functions $\Psi: \mathbb{R} \to (0, \infty)$ and $f: \mathfrak{X} \to \mathbb{R}$ such that $c(f) := \int \Psi(f)\, d\mu < \infty$, define the probability density $p_f(x) = \Psi(f(x))/c(f)$. For $r \ge 1$, show that $\|p_f^{1/r} - p_g^{1/r}\|_r \le 2\|\Psi(f)^{1/r} - \Psi(g)^{1/r}\|_r/c(f)^{1/r}$ and $|c(f)^{1/r} - c(g)^{1/r}| \le \|\Psi(f)^{1/r} - \Psi(g)^{1/r}\|_r$, and also that $\|p_f - p_g\|_\infty \le (1 + \mu(\mathfrak{X})\|p_g\|_\infty)\|\Psi(f) - \Psi(g)\|_\infty/c(f)$. Conclude that if Ψ is uniformly Lipschitz on an interval that contains the ranges of f and g, then $\|p_f - p_g\|_1 \lesssim \mu(\mathfrak{X})\|f - g\|_\infty/c(f)$ and $|c(f) - c(g)| \lesssim \mu(\mathfrak{X})\|\|f - g\|_\infty$, and also that $\|p_f - p_g\|_\infty \le (1 + \mu(\mathfrak{X})\|p_g\|_\infty)\|f - g\|_\infty/c(f)$. Furthermore, conclude that if $\sqrt{\Psi}$ is Lipschitz on an interval that contains the ranges of f and g, then $d_H(p_f, p_g) \lesssim \mu(\mathfrak{X})\|f - g\|_\infty/c(f)$ and $|c(f)^{1/2} - c(g)^{1/2}| \lesssim \mu(\mathfrak{X})\, \|f - g\|_\infty$.

2.6 (van der Vaart and van Zanten 2008a) In the setting of Lemma 2.8, let H be the cumulative distribution function Φ of the standard normal distribution. Show that $\max\{K(p_w, p_{w_0}), V_2(p_w, p_{w_0})\} \lesssim \|w - w_0\|_{2,G_0}^2 + \|w - w_0\|_{4,G}^4$, where $dG_0 = (w_0^2 \vee 1)\, dG$.