# ON BAYESIAN INFERENCE FOR SOME STATISTICAL INVERSE PROBLEMS WITH PARTIAL DIFFERENTIAL EQUATIONS

Richard Nickl
University of Cambridge
r.nickl@statslab.cam.ac.uk

Communicated by Sara van de Geer

This letter (*Bernoulli News*, Vol. 24 (2), 2017) summarizes key ideas from the Ethel Newbold prize lecture. We discuss recent results that provide theoretical support for nonparametric Bayes solutions of statistical inverse problems arising in some partial differential equation model problems of parabolic, elliptic and transport type.

## §0. Introduction

Inverse problems form a vast and well-studied area within applied mathematics, statistics and numerical analysis. Just as in other areas of data science, it has become increasingly intractable to understand the real world performance of algorithms designed to solve these problems without considering the effects of *statistical noise*. Often a method that works for noiseless data needs to be substantially modified to deal with the presence of random measurement error. But even when algorithms such as Tikhonov regularisers are shown to be robust to perturbations of the signal via some convergence rate analysis, much is to be gained from interpreting an inverse problem in a genuinely statistical way: First, when applied appropriately, the theory of statistical inference can be used to provide *recovery guarantees* for commonly used algorithms, which in turn permit *uncertainty quantification*, a contemporary word for the statistical practise of reporting confidence regions and the associated process of algorithm-based decision making, such as rejection of scientific hypotheses at a certain significance level. Second, as we shall explain below, a notion of microscopic statistical fluctuations of inverse problem solvers can be introduced, allowing for refined comparisons of the infinitesimal behaviour of competing algorithms. From this a notion of statistical optimality emerges which blends classical 'efficiency ideas' due to C.F. Gauß and R.A. Fisher with analytic questions about the 'information operators' underpinning every inverse problem. In particular Bayesian methods as suggested by Stuart (2010) can be shown to provide optimal solutions for inverse problems in this sense.

## §1. Statistical Inverse Problems and PDEs

A large family of important inverse problems arise in the area of partial differential equations (PDEs). Typically some partial differential operator $\mathcal{L}_f$ acting on functions $u : \mathcal{O} \to \mathbb{R}$ defined on some regular bounded domain $\mathcal{O} \subset \mathbb{R}^d$ is given, and the *coefficient* $f$ is the unknown functional parameter of interest. Data is given in the form of some solution $u$ of an operator equation $F(\mathcal{L}_f, u) = 0$ subject to some boundary conditions guaranteeing a solution. Prototypical examples are solutions $u = u_f$ of divergence form elliptic PDEs

$$\mathcal{L}_f u \equiv \nabla \cdot (f_1 \nabla u) - f_0 u = 0 \text{ on } \mathcal{O} \qquad (1)$$
$$\text{s.t. } u = g \text{ on } \partial\mathcal{O}$$

where $f_1$ models the coefficient of the partial differential operator and $f_0$ is a potential term. We can treat either $f_0$ or $f_1$ as the unknown function $f$ here. Under suitable conditions the map $f \mapsto u_f$ is then injective and we can ask the question of how to infer the value of $f$ given $u_f$ corrupted by additive Gaussian white noise. Applications of such models in engineering and physics are abundant.

We may further introduce some time evolution dynamics on a time interval $[0, T]$, for instance by considering solutions $u(x, t)$ to the parabolic PDE

$$\frac{\partial u(x,t)}{\partial t} - \mathcal{L}_{f,x} u(x,t) = 0 \;\; \forall (x,t) \in \mathcal{O} \times [0,T], \quad (2)$$

subject to an initial condition $u(\cdot, 0) = g$ and some boundary conditions. Typically here we will discard the potential term in $\mathcal{L}_f$ (i.e., set $f_0 = 0$) and, instead of considering a divergence form operator, explicitly model the 'drift' $f_1$ and 'diffusion' coefficient $f_2$ separately; in the scalar case $d = 1$ for instance

$$\mathcal{L}_f(\cdot) = f_1 \frac{d}{dx} + \frac{f_2^2}{2} \frac{d^2}{(dx)^2}.$$

Then $u_f$ is the solution to the heat equation described by the semigroup dynamics with infinitesimal generator $\mathcal{L}_f$. Identifying the functional parameters $f_1, f_2$ from some observations in such a diffusion model is of fundamental importance in many applications in modern science, e.g., in biology, physics and economics.

Our third example is a first order PDE with boundary data. Consider the transport equation

$$v \cdot \nabla_x u(x,v) + a(x) u(x,v) = f(x), \; x \in \mathcal{O}, \; v \in S^{d-1}, \quad (3)$$

subject to the boundary condition $u(x, v) = 0$ for $x \in \partial\mathcal{O}, v \cdot \nu(x) \geq 0$, where $\nu(x)$ is the outer normal at $x$. Here $a$ is a known attenuation coefficient and $f$ an unknown source function. Along each straight line the last PDE becomes an ordinary differential equation that is easily solved. The influx trace of this solution $u = u_{f,a}(x, v), x \in \partial\mathcal{O}, v \cdot \nu(x) \leq 0$, is precisely the (attenuated) $X$-ray transform of the source function $f$, and the inverse problem is to reconstruct $f$ based on this boundary data. When $\mathcal{O}$ is the unit disk and $a = 0$ this equals the standard problem of reconstructing $f$ from its *Radon transform* $\mathcal{R}(f)$, the workhorse of modern 'non-invasive' computerised tomography methods. More generally such $X$-ray transforms are the basis of many modern scientific imaging methods such as PET and SPECT.

## §2. Statistical noise and measurement models

It is natural to assume that physical measurements in inverse problems arise in a statistical fashion. Observations are always discrete, and if we sample the solution $u_f$ of our PDE at a number of 'design points' $x_i$ (such as different geodesics along which a Radon transform is shot), we can model the measurement errors as independent random variables $g_i$. Each $g_i$ being itself a superposition of many independent random effects, a Gaussian model for the $g_i$'s is approximately correct in view of the central limit theorem. Formally, for our data is then

$$Y_i = u_f(x_i) + g_i, \quad i = 1, \dots, n; \; g_i \text{ i.i.d. } N(0, 1). \quad (4)$$

By standard arguments from asymptotic statistics (see Reiß (2008) or Chapter 1 in Giné and Nickl (2016)) this discrete measurement model is asymptotically (as $n \to \infty$) equivalent to observing the continuous functional equation

$$Y = u_f + \varepsilon\mathbb{W} \;\text{ in } \mathbb{H}, \quad \varepsilon = \frac{1}{\sqrt{n}} \quad (5)$$

where $\mathbb{W}$ is a Gaussian white noise process on the Hilbert space $\mathbb{H}$ that is the natural range of $u_f$. While $\mathbb{W}$ can be defined by its action on this Hilbert space, it does not define a proper random element in it. For instance in the elliptic case (1), $\mathbb{H} = L^2(\mathcal{O})$ but $\mathbb{W}$ defines a random variable only in a negative Sobolev space $H^{-\beta}, \beta > d/2$. Thus even if $u_f$ is a smooth function, our data $Y$ will be 'rough', and solving for $f$ in the presence of noise with such 'large support' is a non-obvious task.

When considering a time evolution PDE, the additive noise model just described may be relevant too. However, stochastic noise may propagate through the entire system with time, and in this case the theory of stochastic differential equations (SDEs) can come to our aid to provide a consistent measurement model. More precisely, a Markov process $(Y_t : t \geq 0)$ with transition semigroup operator $P_t = e^{t\mathcal{L}_f}, t \geq 0$, provides solutions to the SDE

$$dY_t = f_1(Y_t)dt + f_2(Y_t)dW_t, t \geq 0, \quad (6)$$

with $(W_t)$ a Brownian motion, effectively describing the diffusion of a particle that, when positioned at $x$ at a certain time, has 'infinitesimal' drift $f_1(x)$ with Gaussian noise of variance $f_2^2(x)$. A realistic measurement model for the parabolic PDE (2) then consists of observing the entire trajectory of the Markov process $(Y_t : 0 \leq t \leq T)$ until time $T$, or of discrete samples $Y_0, Y_\Delta, \dots, Y_{n\Delta}$ thereof, paralleling the situations described in (5), (4) for i.i.d. noise.

## §3. The Bayesian approach

All the above problems share the common structure that we observe

$$\textit{data } Y \textit{ drawn from some distribution } P_{u_f}$$

where $u_f$ is some forward operator and $f$ the unknown function. As suggested in Stuart (2010) (see also Dashti and Stuart (2016)), it is tempting to take the Bayesian approach and model $f$ by some prior probability distribution $\Pi$ in function space. Even though the models for $f$ from the previous section are typically infinite-dimensional, their near 'Gaussian' character permits the use of basic tools from probability theory to deduce 'Bayes' formula'

$$f \sim \Pi, \; Y|f \sim P_{u_f} \;\Rightarrow f|Y \sim \frac{dP_{u_f}(Y)d\Pi(f)}{\int dP_{u_f}(Y)d\Pi(f)} \quad (7)$$

where $dP_{u_f}$ is a density with respect to a suitable dominating measure. We can then extract information on $f$ from the posterior distribution $\Pi(\cdot|Y)$ of $f|Y$. For the infinite-dimensional, or 'nonparametric', models relevant here, a large class of priors have been developed in the area of 'Bayesian Nonparametrics', we refer to the recent monograph Ghosal and van der Vaart (2017). While the focus of this letter is not Bayesian *computation*, we note here that modern MCMC methodology can be used successfully to sample from posterior distributions, and to numerically evaluate point estimates of $f$, such as the posterior mean or mode (see Dashti and Stuart (2016)). The Bayesian approach thus gives concrete algorithms that can be used in real world inverse problems including all the PDE examples introduced above. Moreover, this methodology is attractive for statistical scientists because the spread of the posterior distribution automatically delivers an estimate of the uncertainty in the reconstruction, and hence suggests 'confidence' intervals.

The performance of Bayesian algorithms of course crucially depends on the choice of the prior $\Pi$, which in our 'nonparametric' setting serves solely as a regularisation tool and does not represent any subjective beliefs. This can be nicely illustrated by the fact that for linear inverse problems and Gaussian priors $\Pi$ with associated reproducing kernel Hilbert space $\mathcal{H}$, the posterior mean can be shown to coincide with the usual Tikhonov regulariser which solves

$$\min_f \left[ \sum_{i=1}^n (Y_i - u_f(x_i))^2 + \|f\|_{\mathcal{H}}^2 \right], \quad (8)$$

so that the prior choice is somehow dual to the choice of the penalty function in a standard optimisation based estimator for $f$. For example Matérn or integrated Brownian motion priors will generate regularisers with commonly used penalties arising from standard Sobolev norms.

In light of the previous observation it becomes of crucial importance to study the performance of Bayesian inversion in some 'objective' way that is independent of the prior choice, as otherwise posterior inferences would only be reproducing prior guesses that do not represent anything in particular about the real world. This is not just a 'philosophical' debate about Bayesian or non-Bayesian statistics, but a question of plain common sense, just as the choice of the penalty function $\|\cdot\|_{\mathcal{H}}$ in (8) is not a philosophical question. Trying to understand the 'frequentist' validity of Bayesian inference is a classical topic in mathematical statistics that goes back to Laplace (1812), and which has undergone vigorous development in the last two decades. It can help to provide objective foundations for prior based inference methods also in contemporary inverse problems.

## §4. Posterior contraction rates to the true parameter.

The posterior distribution $\Pi(\cdot|Y)$ arising from the formalism (7) is a (through $Y$) *random* probability measure in function space. Henceforth we assume that the data $Y$ are generated from a fixed unknown probability distribution $P_{f_0} \equiv P_{u_{f_0}}$, where $f_0$ represents an arbitrary, hypothetically 'true', value. The first question we can ask is about 'consistency' of the posterior random measure in the sense that we want it to concentrate most of its mass near $f_0$, at least in the 'large sample' or 'small noise' limit where $n \to \infty$ or $\varepsilon \to 0$, respectively. Formally we want to find an as fast as possible rate $\delta_n$ (or $\delta_\varepsilon$) such that in some metric $d$ on function space,

$$\Pi(f : d(f, f_0) \geq \delta_n|Y) \to 0 \tag{9}$$

as $n \to \infty$ and in $P_{u_{f_0}}$-probability. Tools for this have been developed in remarkable depth and breadth in Bayesian Non-parametrics for direct problems, a key idea being 'robust testing in Hellinger distance' – see Ghosal and van der Vaart (2017) or Sections 7.1 and 7.3 in Giné and Nickl (2016). These methods however do not obviously adapt to the inverse problems setting, and new ideas are required. For linear inverse problems some tools exist, see Knapik et al. (2011); Agapiou et al. (2013); Ray (2013); Kekkonen et al. (2016); van Waaij and van Zanten (2016), covering in particular the problem involving the transport PDE (3) appearing with Radon transforms and the SDE problem (6) with $\sigma = 1$ and continuous data $(Y_t : 0 \leq t \leq T)$. Bute none of these proofs give a strategy to prove contraction rates for general, non-linear, inverse problems. In Ray (2013) an idea of Giné and Nickl (2011) is picked up to construct tests for linear problems replacing 'robust testing' by techniques from concentration of measure theory and nonparametric statistics. This allows to obtain contraction rates outside of the conjugate setting, an approach that generalises to the non-linear setting, as demonstrated in the recent papers Nickl and Söhl (2017); Nickl (2017) where the parabolic and elliptic problems from above were considered, respectively. In both it was found that the posterior contracts at optimal rates (in a minimax sense). For instance in the elliptic case (1) with $f_1 = 1$ known but unknown potential $f_0 \in C^s(\mathcal{O})$ a positive $s$-times continuously differentiable function on $\mathcal{O}$, if the observations are given in model (5), the contraction rates in $L^2(\mathcal{O})$-distance for a uniform wavelet prior are (up to log-factors)

$$\delta_\varepsilon \approx \varepsilon^{\frac{2s}{(2s+4+d)}} \quad \text{as noise level } \varepsilon \to 0.$$

In the parabolic case (2), when considering discrete (low frequency) data $Y_\Delta, \ldots, Y_{n\Delta}$ in the scalar diffusion model (6) with a suitable hierarchical prior construction, then if $f_1 \in C^{s-1}, f_2 \in C^s$, one obtains in (9) the rates

$$\delta_n = n^{-(s-1)/(2s+3)} \text{ for the drift coefficient } f_1$$

$$\delta_n = n^{-s/(2s+3)} \text{ for the diffusion coefficient } f_2,$$

as sample size $n$ increases, again up to log-factors, and in $L^2$-distance. The proof techniques employed in (Nickl and Söhl (2017); Nickl (2017)) depend on a few standard properties of the elliptic and parabolic problems that feature also in general inverse problems: Main analytic ingredients are a *stability estimate* for the forward problem that allows to control $\|f - g\|$ in terms of $\|u_f - u_g\|'$, in suitable norms $\|\cdot\|, \|\cdot\|'$, and a dual form of the usual *regularity estimates* for solutions of PDEs such as $\|u_f - u_g\|_{L^2} \lesssim \|f - g\|_{H^{-\alpha}}$ where $H^{-\alpha}$ is a negative Sobolev space with exponent $\alpha$ corresponding to the ill-posedness of the problem. If such estimates are available then tools from nonparametric statistics can be applied to deduce contraction rates for priors that generally do not require identification of a SVD-type basis underlying the forward operator.

## §5. Microscopic fluctuations of solutions of inverse problems and the Fisher information operator

Once it is known that the posterior concentrates near the true value $f_0$ in a certain distance, it is natural to consider the fluctuations of $f|Y$ near $f_0$ when scaled by some inverse 'contraction rate'. The previous results were obtained for the distance function $d$ induced by the $L^2$-norm, and one may thus initially consider the statistical fluctuations of the random variable

$$Z_\varepsilon = \delta_\varepsilon^{-1}(\bar f|Y - f_0) \text{ in } L^2.$$

Here some subtle geometric obstructions occur, some of which were already noted in the simplest infinite sequence space model in Freedman (1999)'s Wald lecture. Perhaps the easiest way to understand the issue is to anticipate the results that will follow: after centring $Z_\varepsilon$ at its expectation, the marginal distributions of the process $(\varepsilon^{-1}\langle Z_\varepsilon - EZ_\varepsilon, \psi\rangle_{L^2} : \psi \in C^\infty)$ along smooth projection directions $\psi$ will be seen to converge weakly in probability to a fixed non-degenerate Gaussian process $(\mathbb{G}(\psi) : \psi \in C^\infty)$. Note that here we have re-scaled by the larger $\varepsilon^{-1}$ instead of $\delta_\varepsilon^{-1}$. Since all sub-sequential distributional limits of random variables in function space are determined by the limits of such marginal distributions, a functional limit theorem in $L^2$ would also have to hold at rate $\delta_\varepsilon = \varepsilon$. But this would imply a contraction theorem as in (9) at that rate, which is impossible in light of lower bounds provided by statistical minimax theory for such estimation problems in Gaussian white noise (Chapter 6 in Giné and Nickl (2016)).

The way to overcome, or rather side-step, these obstructions, was set out in the papers Castillo and Nickl (2013, 2014). The idea is to determine *maximal* families $\Psi$ of functions $\psi$ for which the Gaussian asymptotics

$$(\varepsilon^{-1}\langle Z_\varepsilon - EZ_\varepsilon, \psi\rangle_{L^2} : \psi \in \Psi) \to (\mathbb{G}(\psi) : \psi \in \Psi) \tag{10}$$

can be obtained – by analogy to the classical result from parametric statistics these are often called 'nonparametric Bernstein - von Mises theorems'. Due to the maximality requirement such results are somehow 'dual' to obtaining contraction rates in $L^2$ (by using arguments from interpolation theory), but they allow to go beyond a mere convergence rate analysis: the Gaussian process $\mathbb{G}$ identifies the precise microscopic fluctuations of the posterior near $f_0$. While the results in Castillo and Nickl (2013, 2014) are confined to 'direct' problems in nonparametric regression and probability density estimation, in the recent articles Nickl (2017); Monard et al. (2017) the first such nonparametric Bernstein-von Mises theorems have been proved for PDE type inverse problems in the white noise model (5) (see also Nickl and Söhl (2017) for a non-linear inverse problem with jump processes).

To understand the nature of the microscopic fluctuations, let us first consider the transport PDE problem where the observations consist of the $X$-ray transform $u_{a,f} \equiv I_a(f)$ of the unknown source function $f$. When $\mathcal{O}$ is the unit disk the forward operator equals the standard Radon transform, but even in the general setting the linear operator $I_a$, known as the attenuated $X$-ray transform, is well studied in integral geometry. In Monard et al. (2017), using techniques from micro-local analysis, it is proved that the 'information' operator $I_a^* I_a$, where $I_a^*$ is a natural adjoint operator, has an inverse $(I_a^* I_a)^{-1}$ that maps $C^\infty(\mathcal{O})$ isomorphically into $\{g/\sqrt{d_\mathcal{O}} : g \in C^\infty(\mathcal{O})\}$, where $d_\mathcal{O} = d(\cdot, \partial\mathcal{O})$ is the distance function to the boundary $\partial\mathcal{O}$ of $\mathcal{O}$. For natural Gaussian priors for $f$ and posterior draws $f|Y \sim \Pi(\cdot|Y)$, it is then proved that, whenever $\psi \in C^\infty(\mathcal{O})$, as $\varepsilon \to 0$,

$$\varepsilon^{-1}\langle f|Y - E^\Pi[f|Y], \psi\rangle_{L^2} \to^d N(0, \|I_a(I_a^* I_a)^{-1}\psi\|^2) \tag{11}$$

in $P_{f_0}$-probability, where the norm on the right hand side is a natural $L^2$-norm on 'geodesic space'. The limiting covariance can be shown to be minimal in the sense that it attains the semi-parametric Cramér-Rao lower bound (or 'inverse Fisher information') for estimating $\langle f, \psi\rangle_{L^2}$ near $f_0$. The Gaussian nature of the posterior distribution combined with the Paley-Zygmund inequality then also shows that the Tikhonov regulariser $\hat f$ minimising (8) in this problem with any Sobolev-norm penalty satisfies, for any $\psi \in C^\infty(\mathcal{O})$ and as $\varepsilon \to 0$,

$$\varepsilon^{-1}\langle \hat f - f_0, \psi\rangle_{L^2} \to^d N(0, \|I_a(I_a^* I_a)^{-1}\psi\|^2),$$

a result that is of interest also outside of the Bayesian context (although its proof is 'Bayesian').

The findings in the transport PDE case foreshadow the general principle: the microscopic fluctuations of optimal inverse problem solvers will depend on the inverse Fisher information operator $(I_a^* I_a)^{-1}$, and its existence combined with mapping properties play a crucial role in proving Bernstein-von Mises theorems. For non-linear inverse problems, the information operator that has to be inverted is found after linearisation. This is demonstrated in Nickl (2017) for a prototypical elliptic PDE case (1) with $f_1 = 1$ and unknown potential $f = f_0$: basic perturbation arguments for the Schrödinger equation imply that in this case the role of $I_a$ is replaced by $V_f[\cdot/u_f]$, where $V_f$ is the inverse of the Schrödinger operator $S_f(u) = \Delta u - fu$, derived via PDE techniques or using semi-group theory for killed Brownian motion (see Chung and Zhao (1995)). In this case by self-adjointness of $V_f$ the Cramer-Rao lower bound simplifies and the Bernstein-von Mises theorem becomes

$$\varepsilon^{-1}\langle f|Y - E^\Pi[f|Y], \psi\rangle_{L^2} \to^d N(0, \|S_{f_0}[\psi/u_{f_0}]\|^2_{L^2}) \tag{12}$$

in $P_{f_0}$-probability as $\varepsilon \to 0$, for every compactly supported $\psi \in C^\infty(\mathcal{O})$, and for $f|Y$ drawn from a posterior distribution corresponding to a natural uniform wavelet prior.

The limit theorems (11), (12) single out the Gaussian limit process $(\mathbb{G}(\psi) : \psi \in \Psi_a)$ towards which the centred posterior distribution will converge. We can then return to the program laid out in (10) and look for maximal classes $\Psi$ of functionals $\psi$ for which this convergence occurs simultaneously. As shown in Nickl (2017) maximal such classes can be characterised in terms of the sample continuity properties of the limiting Gaussian process, and in the elliptic PDE case equals a ball in the space $C_c^\alpha(\mathcal{O})$ of compactly supported $\alpha$-Hölder functions with critical threshold $\alpha > 2 + d/2$. It is then further shown in the main theorem in Nickl (2017) that indeed the posterior distribution converges weakly to the law of $\mathbb{G}$ for the topology of uniform convergence on $\Psi$, in $P_{f_0}$-probability, giving the first optimality result of its kind for the Bayesian solution of a PDE-type non-linear inverse problem.

The Bernstein-von Mises theorems introduced here have important applications to the frequentist justification of Bayesian inference methods. Particularly they imply that Bayesian 'credible regions' and 'error bars' amount to proper confidence sets according to the usual 'frequency' interpretation of statistical significance. In particular, Bayesian inferences that have 95% posterior credibility will have approximately 0.95 chance of returning the correct decision in repeated trials. Once a Bernstein-von Mises theorem is at hand these facts are not specific to inverse problems and follow the general ideas developed in Castillo and Nickl (2013, 2014).

To conclude, the ideas presented here allow to derive precise microscopic fluctuations of inverse problem solvers from a careful study of the information operator underlying a given inverse problem. They provide a general template to prove similar results in various other settings. The main attraction of Bernstein-von Mises type results is perhaps that they reveal finer properties of an inverse problem than a mere convergence rate analysis does, via the covariance structure of the limiting Gaussian process. They also raise the interesting open question whether standard numerical inverse solvers attain the statistical information bounds that emerge from our theory, or whether they are potentially outperformed by Bayesian methods when interpreted as statistical algorithms.

**Acknowledgement**. I would like to thank Gabriel P. Paternain for helpful discussions. The author acknowledges support by European Research Council (ERC) grant No. 647812.

# References

Agapiou, S., Larsson, S. and Stuart, A. M. (2013), Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Process. Appl.* **123** 3828–3860.

Castillo, I., and Nickl, R. (2013) Nonparametric Bernstein - von Mises theorems in Gaussian white noise. *Annals of Statistics* **41** 1999-2028.

Castillo, I., and Nickl, R. (2014) On the Bernstein - von Mises phenomenon for nonparametric Bayes procedures. *Annals of Statistics* **42** 1941-1969.

Chung, K.L. and Zhao, Z.X. (1995). *From Brownian motion to Schrödinger's equation.* Springer, 1995.

Dashti, M., and Stuart, A.M. (2010), "The Bayesian approach to inverse problems," *In: Handbook of Uncertainty Quantification*, Springer, 2016.

Freedman, M. (1999) On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Annals of Statistics* **27** 1119–1140.

Ghosal, S., and van der Vaart, A. W. (2017), *Fundamentals of Nonparametric Bayesian Inference.* Cambridge University Press, New York

Giné, E., and Nickl, R. (2011) Rates of contraction for posterior distributions in $L^r$-metrics, $1 \le r \le \infty$. *Annals of Statistics* **39** 2883?2911.

Giné, E., and Nickl, R. (2016) *Mathematical foundations of infinite-dimensional statistical models.* Cambridge University Press, New York

Kekkonen, H., Lassas, M. and Samuli Siltanen. (2016). Posterior consistency and convergence rates for Bayesian inversion with hypoelliptic operators. *Inverse Problems* **32**

Knapik, B., van der Vaart, A.W. and van Zanten, H. (2011) Bayesian inverse problems with Gaussian priors. *Annals of Statistics*, **39** 2626–2657.

Laplace, P.-S. M. de (1812) *Théorie analytiques des probabilités.* Courcier, Paris.

Monard, F., Nickl, R., and Paternain, G.P. (2017) Efficient nonparametric Bayesian inference for $X$-ray transforms. *arXiv preprint.*

Nickl, R. (2017) Bernstein - von Mises theorems for statistical inverse problems I: Schödinger equation. *arXiv preprint.*

Nickl, R. and Söhl, J. (2017) Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *Annals of Statistics* **45** 1664–1693

Nickl, R. and Söhl, J. (2017) Bernstein - von Mises theorems for statistical inverse problems II: compound Poisson processes. *arXiv.*

Ray, K. (2013) Bayesian inverse problems with non-conjugate priors. *Electronic Journal of Statistics* **7** 2516–2549.

Reiß, M. (2008), Asymptotic equivalence for nonparametric regression with multivariate and random design. *Annals of Statistics*, **36**, 1957-1982.

Stuart, A.M. (2010), Inverse problems: a Bayesian perspective. *Acta Numerica*, **19**, 451-559.

van Waaij, J. and van Zanten, J.H. (2016). Gaussian process methods for one-dimensional diffusions: optimal rates and adaptation. *Electronic Journal of Statistics* **10** 628?645.