The Bayesian Approach to Inverse Problems

Andrew M. Stuart,

Mathematics Institute
Warwick University
CV4 7AL, UK
e-mail: a.m.stuart@warwick.ac.uk

Abstract: These notes give various details, along with references to the literature, that I will not, for reasons of brevity, give during the lectures themselves.

1. Goal of Lectures

There is an intimate relationship between optimization and probability and the goal of the lectures is to explore this connection in the context of inverse problems. We start by explaining aspects of this relationship in the finite dimensional context, as stated in Theorems 1.1 and 1.2. We then proceed to demonstrate results analogous to these in the infinite dimensional setting arising from the regularization of inverse problems. We will view the inverse problem through Bayesian regularization, and demonstrate how Tikhonov regularized minimization and gradient flow arise in this context.

We will be interested in minimizing a least squares functional Φ subject to regularization. In particular we aim to show the following, both in a Hilbert space setting: (i) that probability maximization is related to minimization of Tikhonov regularized version of Φ, denoted by I; (ii) that random walks combined with judicious accept/reject lead to an approximation of a noisy gradient flow for I, and hence to a stochastic annealing method.

2. Notation

Throughout $\left(H, \|\cdot\|, \langle\cdot,\cdot\rangle\right)$ denotes a Hilbert space and X denotes a Banach space continuously embedded in H. We also denote the finite dimensional Euclidean Hilbert space by $\left(\mathbb{R}^N, |\cdot|, \langle\cdot,\cdot\rangle\right)$. Because Gaussian measure will play a central role we also use the concept of weighted inner-products and norms in H: for any self-adjoint positive operator A on H we define

$$\langle\cdot,\cdot\rangle_A = \langle A^{-\frac{1}{2}}\cdot, A^{-\frac{1}{2}}\cdot\rangle, \quad \|\cdot\|_A = \|A^{-\frac{1}{2}}\cdot\|.$$

Similarly in $\mathbb{R}^N$ we write

$$|\cdot|_A = |A^{-\frac{1}{2}}\cdot|.$$

3. Overview in $\mathbb{R}^N$

Discussion of Theorem 1.1 Let $Z = \int_{\mathbb{R}^N} \exp\left(-I(u)\right)du$. Theorem 1.1 can be proved directly from the identity

$$
\begin{aligned}
\mu\left(B_\delta(z)\right) &= \frac{1}{Z}\int_{|u-z|<\delta}\exp\left(-I(u)\right)du \\
&= \frac{1}{Z}\exp\left(-I(z)\right)\int_{|v|<\delta}\exp\left(I(z)-I(z+v)\right)dv \\
&= \frac{1}{Z}\exp\left(-I(z)\right)\int_{|v|<\delta}dv + e(\delta)
\end{aligned}
$$

1

where

$$\frac{e(\delta)}{\int_{|v|<\delta} dv} \to 0$$

as $\delta \to 0$, using continuity of I.

Discussion of Theorem 1.2 The SDE has the form

$$\frac{du}{dt} = -A\nabla I(u) + \sqrt{2A}\frac{dW}{dt}.$$

The Fokker-Planck equation (see [Gar85]) for this SDE is thus

$$\partial_t \rho = \nabla \cdot J(\rho)$$
$$J(\rho) = A\nabla I(u)\rho + \nabla \cdot (A\rho).$$

The invariance of the measure can be checked by showing that the density $\rho_\infty(u) \propto \exp(-I(u))$ is a steady solution of the Fokker-Planck equation because $J(\rho_\infty) \equiv 0$. The properties of $\omega-$limit points for gradient flows are discussed in [Hal88]. The ergodicity of the noisy gradient flow is proved in [MSH02], although the almost sure pathwise result is not stated explicitly, but rather is a consequence of the exponential convergence of expectatations, using the ideas in Chapter 17 of [MT93]; see also [PV01, PV03, PV05] for results on convergence of path averages of SDEs.

Simulated Annealing This is a generic term based on the (physically motivated) idea that gradually cooling a system (of atoms) can be a way of finding (possibly global) optima of a functional (energy). In the context of SDEs this corresponds to choosing a cooling schedule in which $\tau \to 0$ as $t \to \infty$; see [GH86]. However the idea is particularly useful in the development of optimization strategies which do not require evaulation of gradients and, instead, use random walks combined with judicious accept/reject mechanisms; see [KJV83, Čer85, BT93].

4. Motivating Examples

Groundwater Flow Variants on this inverse problem are studied in [BK89] and [Ric81], for example. The results from Lemma 2.1 are established in [DS10].

Oceanography The problem we describe is known as *Lagrangian data assimilation* and is a simplified model for the collection of data concerning ocean flows through the use of Lagrangian instruments, advected by the ocean, and transmitting positional data via GPS; see [KIJ03, SKJI06, VIJ10]. The results from Lemma 2.2 are eastablished in [CDRS09].

Comments Note that both problems a *under-determined* in the sense that the aim is to recover a function (infinite dimensions) from a finite set of measurements. This observation motivates not only the need for regularization, but also for the probabilistic approach. Prior modelling assumptions are needed to determine a solution to the problem, and a probabilistic viewpoint provides one natural and clear way to state these modelling assumptions.

5. Bayesian Approach to Inverse Problems

Setup The material in this section is concerned primarily with the Bayesian formulation of inverse problems; see [KS05] for an introduction to this subject. Our perspective is to "regularize then discretize" and not the other way around. Hence we need to formulate the Bayesian inverse problem on infinite dimensional (function) spaces. We primarily adopt the use of Gaussian random field priors. This approach is overviewed in [Stu10]. There is some current interest in extending prior measures beyond the Gaussian case and a particularly promising avenue is the develpment of so-called *Besov priors*, tuned to wavelet representation of functions, and introduced recently in [LSS09]; see [DHS] for further analaysis of these priors, and application to the groundwater flow inverse problem. There is also some interest in the use of *uniform priors* which give rise to random fields attaining almost sure upper and lower bounds [SS10]; such probability measures are widely studied in the numerical analysis community in the context of uncertainty quantification [GS11].

Tikhonov Regularization This is a widely studied subject; see [BK89, EHN96] and the references therein. Theorem 3.2 for *weakly* convergent subsequence is a classical result in the theory of the calculus of variations, see [Dac89]; the fact that minimizing sequences contain *strongly* convergent subsequences may be found in [KS80].

Gaussian Measures A comprehensive reference for Gaussian measures is [Bog98]. This book works in the generality of locally convex topological spaces, and has material on Gaussian measure on Banach and Hilbert spaces as particular cases. A similar book, focussed on the setting in Banach spaces, is [Lif95]. For a simpler introduction to the Hilbert space setting see the relevant material in [DZ92]. Regularity of functions drawn from Gaussian measures is discussed in some detail in [DZ92]. Theorem 3.3 appears as Theorem 3.5.7 and Corollary 3.5.8 in [Bog98]. We will sometimes use X to denote a Banach space continuously embedded in H. If $\mu_0 = N(0,C)$ is a Gaussian measure on H, and if $\mu_0(X) = 1$ then the Hilbert space E with inner-product $\langle C^{-\frac{1}{2}}\cdot, C^{-\frac{1}{2}}\cdot \rangle$ is compactly embedded in X (and hence H); see Corollary 3.2.4 in [Bog98]. Using this we may adapt Theorem 3.2 as follows (see Theorem 5.4 in [Stu10]):

Modified Theorem 3.2 *Let $\mu_0(X) = N(0,C)$ denote a Gaussian measure on $H \supseteq X$ and assume that $\mu_0(X) = 1$. Then there exists $\bar{u} \in E$ such that $I(\bar{u}) = \inf\{I(u) : u \in E\}$. Furthermore, any minimizing sequence contains a subsequence $\{u_n\}$ converging strongly to $\bar{u}$ in E.*

Bayesian Regularization The conditioning argument used to pass from the expression for $d\nu/d\nu_0$ to that for $d\mu/d\mu_0$ may be found in [Dud02]. The particular form of the statement, as given below in the **Conditioning Lemma**, may be found in [HSV07]. The Lipschitz continuity of expectations with respect to changes in data may be expressed in terms of the Hellinger metric and this approach is adopted in [Stu10], Theorem 4.2.

Conditioning Lemma *Let ν, ν_0 be probability measures on $S \times T$ where $(S, \mathbb{A})$ and $(T, \mathbb{B})$ are measurable spaces. Denote by (x, y) with $x \in S$ and $y \in T$ an element of $S \times T$. Assume that $\nu \ll \nu_0$ and that ν has Radon-Nikodym derivative ϕ with respect to ν_0. Assume further that the conditional distribution of $x|y$ under ν_0, denoted by $\nu_0^y(dx)$, exists. Then the conditional distribution of $x|y$ under ν, denoted $\nu^y(dx)$, exists and $\nu^y \ll \nu_0^y$. The Radon-Nikodym derivative is given by*

$$\frac{d\nu^y}{d\nu_0^y}(x) = \begin{cases} \frac{1}{c(y)}\phi(x,y), & \text{if } c(y) > 0, \text{ and} \\ 1 & \text{else} \end{cases} \tag{5.1}$$

with $c(y) = \int_S \phi(x,y)\, d\nu_0^y(x)$ for all $y \in T$.

6. MAP Estimators and Tikhonov Regularization

The work in this section is based on the forthcoming preprint [DS11]. The proof is very similar to that presented above (Theorem 1.1) in the finite dimensional case. However care has to be taken because $\|u\|_C$ is infinite μ_0 almost everywhere in $B_\delta(z)$. The property of Gaussian measures which is the first display in the proof of Theorem 4.3 may be found as Proposition 3 in section 18 of [Lif95] and is the key estimate which overcomes this technical difficulty.

7. Simulated Annealing and Gradient Flows

The work in this section is based on the forthcoming preprint [PST11a]. Related SPDE limit theorems may be found in [MPS11], and [PST11b]; however these latter two papers concern algorithms which are not defined on Hilbert space, and only approach an SPDE through finite dimensional approximation, and with a "Courant-like" restriction on the time-step to achieve convergence. In this context note that the proposal we study is, given current state u,

$$v = (1 - 2\delta)^{\frac{1}{2}}u + \sqrt{2\tau\delta}\xi, \quad \xi \sim N(0, C). \tag{7.1}$$

The proposal studied in [MPS11] appears very similar:

$$v = u + \sqrt{2\tau\delta}\xi, \quad \xi \sim N(0, C). \tag{7.2}$$

3

Yet the resulting algorithm is only defined after truncation to finite dimensions, and then requires $\delta = \mathcal{O}(N^{-1})$ in order to obtain the SPDE as a limit. In contrast the propsal in (7.1) gives rise to an algorithm defined on Hilbert space; if it is applied after truncation to finite dimensions then convergence to the SPDE limit is found without any restriction on δ in terms of N.

MCMC methods were introduced by solid state physicists in the seminal paper [MRTT53] and subsequently generalized to Metropolis-Hastings methods on $\mathbb{R}^n$ in [Has70]. The books [RC99, Liu01] contain overviews from the perspective of applied statistics. The formulation of Metropolis-Hastings methods on a general state space is given in [Tie98]; this is the natural setting for our random walk-based methods on Hilbert space. The particular random walk that we analyze is introduced in [BS09, BRS09].

The basic theory of SPDEs (or SDEs in Hilbert space) may be found in [DZ92]. The particular SPDE which arises as a diffusion limit of our Markov chain is studied in [HSV07], and is called the *preconditioned* SPDE. In particular, Theorem 5.3 is proved, under much weaker conditions than we state here, in [HSV07]. The idea of deriving diffusion limts for MCMC methods was systematically developed in the papers [RGG97, RR98, RR01]. Those papers concerned finite dimensional target measures with a product structure (independence amongst coordinates) and as a consequence it is possible to prove a diffusion limit for a single component of the vector in the target space as the dimension of the system grows. However, for our problem there are correlations among different components (say coefficients in an expansion in the basis determined by C) and so we obtain diffusion limit to an SDE in infinite dimensions: Theorem 5.5. The idea of using a continous time interpolant of the discrete time Markov chain is developed in the simple context of Euler approximation of SDEs in [Mao97, HMS03]. The invariance principle for the noise process $m^{(k)}$ uses a result from the paper [Ber86].

References

[Ber86]   E. Berger. Asymptotic behaviour of a class of stochastic approximation procedures. *Probability theory and related fields*, 71(4):517–552, 1986.

[BK89]   H.T. Banks and K. Kunisch. *Estimation techniqiues for distributed parameter systems.* Birkhäuser, 1989.

[Bog98]   V.I. Bogachev. *Gaussian Meausures.* American Mathematical Society, 1998.

[BRS09]   A. Beskos, G. O. Roberts, and A. M. Stuart. Optimal scalings for local Metropolis-Hastings chains on non-product targets in high dimensions. *Ann. Appl. Prob.*, 19:863–898, 2009.

[BS09]   A. Beskos and A.M. Stuart. MCMC methods for sampling function space. In *Invited Lectures, Sixth International Congress on Industrial and Applied Mathematics, ICIAM07, Editors Rolf Jeltsch and Gerhard Wanner*, pages 337–364. European Mathematical Society, 2009.

[BT93]   D. Bertsimas and J. Tsitsiklis. Simulated annealing. *Statistical Science*, 8(1):10–15, 1993.

[CDRS09]   S.L. Cotter, M. Dashti, J.C. Robinson, and A.M. Stuart. Bayesian inverse problems for functions and applications to fluid mechanics. *Inverse Problems*, 25:doi:10.1088/0266–5611/25/11/115008, 2009.

[Čer85]   V. Černỳ. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of optimization theory and applications*, 45(1):41–51, 1985.

[Dac89]   B. Dacarogna. *Direct Methods in the Calculus of Variations.* Springer, New York, 1989.

[DHS]   M. Dashti, S. Harris, and A.M. Stuart. Besov priors for bayesian inverse problems.

[DS10]   M. Dashti and A.M. Stuart. Uncertainty quantification and weak approximation of an elliptic inverse problem. *Submitted*, http://arxiv.org/abs/1102.0143, 2010.

[DS11]   M. Dashti and A.M. Stuart. Map estimators and the bayesian approach to inverse problems. *in preparation*, http://arxiv.org/abs/, 2011.

[Dud02]   R.M. Dudley. *Real Analysis and Probability.* Cambridge University Press, Cambridge, 2002.

[DZ92]   G. DaPrato and J. Zabczyk. *Stochastic equations in infinite dimensions*, volume 44 of *Encyclopedia of Mathematics and its Applications.* Cambridge University Press, Cambridge, 1992.

[EHN96]   H.K. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, 1996.

[Gar85]   C. W. Gardiner. *Handbook of stochastic methods*. Springer-Verlag, Berlin, second edition, 1985. For physics, chemistry and the natural sciences.

[GH86]   S. Geman and C.R. Hwang. Diffusions for global optimization. *SIAM Journal on Control and Optimization*, 24:1031, 1986.

[GS11]   C.J. Gittelson and Ch. Schwab. Sparse tensor discretizations of high-dimensional PDEs. *Acta Numerica*, 20, 2011.

[Hal88]   J.K Hale. *Asymptotic behavior of dissipative systems*, volume 25 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1988.

[Has70]   W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[HMS03]   D.J. Higham, X. Mao, and A.M. Stuart. Strong convergence of Euler-type methods for nonlinear stochastic differential equations. *SIAM Journal on Numerical Analysis*, 40(3):1041–1063, 2003.

[HSV07]   M. Hairer, A. M. Stuart, and J. Voss. Analysis of SPDEs arising in path sampling, part II: The nonlinear case. *Annals of Applied Probability*, 17:1657–1706, 2007.

[KIJ03]   L. Kuznetsov, K. Ide, and C.K.R.T Jones. A method for assimilation of Lagrangian data. *Mon. Wea. Rev.*, 131(10):2247–2260, 2003.

[KJV83]   S. Kirkpatrick, D.G. Jr., and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[KS80]   D. Kinderlehrer and G. Stampacchia. *An Introduction to Variational Inequalities and Their Applications*. SIAM, 1980.

[KS05]   J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*, volume 160 of *Applied Mathematical Sciences*. Springer, 2005.

[Lif95]   M.A. Lifshits. *Gaussian Random Functions*, volume 322 of *Mathematics and its Applications*. Kluwer, Dordrecht, 1995.

[Liu01]   Jun S. Liu. *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. Springer, 2001.

[LSS09]   M. Lassas, E. Saksman, and S. Siltanen. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Problems and Imaging*, 3:87–122, 2009.

[Mao97]   Xuerong Mao. *Stochastic differential equations and their applications*. Horwood Publishing Series in Mathematics & Applications. Horwood Publishing Limited, Chichester, 1997.

[MPS11]   J.C. Mattingly, N.S. Pillai, and A.M. Stuart. Spde limits of the random walk metropolis algorithm in high dimensions. *Ann. Appl. Prob*, 2011.

[MRTT53] N. Metropolis, R.W. Rosenbluth, M.N. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[MSH02]   J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Process. Appl.*, 101(2):185–232, 2002.

[MT93]   S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Communications and Control Engineering Series. Springer-Verlag London Ltd., London, 1993.

[PST11a] N.S. Pillai, A.M. Stuart, and A.H. Thiery. Gradient flow from a random walk in hilbert space. *in preparation*, http://arxiv.org/abs/, 2011.

[PST11b] N.S. Pillai, A.M. Stuart, and A.H. Thiery. Optimal scaling and diffusion limits for the langevin algorithm in high dimensions. *Arxiv preprint arXiv:1103.0542*, 2011.

[PV01]   E. Pardoux and A. Yu. Veretennikov. On the Poisson equation and diffusion approximation. I. *Ann. Probab.*, 29(3):1061–1085, 2001.

[PV03]   È. Pardoux and A. Yu. Veretennikov. On Poisson equation and diffusion approximation. II. *Ann. Probab.*, 31(3):1166–1192, 2003.

[PV05]   E. Pardoux and A. Yu. Veretennikov. On the Poisson equation and diffusion approximation. III. *Ann. Probab.*, 33(3):1111–1133, 2005.

[RC99]   C.P. Robert and G.C. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, 1999.

[RGG97]   G.O. Roberts, A. Gelman, and W.R. Gilks. Weak convergence and optimal scaling of random

walk Metropolis algorithms. *Ann. Appl. Prob.*, 7:110–120, 1997.

[Ric81]   G.R. Richter. An inverse problem for the steady state diffusion equation. *SIAM Journal on Applied Mathematics*, 41(2):210–221, 1981.

[RR98]   G.O. Roberts and J. Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *J. Roy. Stat. Soc. B*, 60:255–268, 1998.

[RR01]   G.O. Roberts and J. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16:351–367, 2001.

[SKJI06] H. Salman, L. Kuznetsov, C.K.R.T Jones, and K. Ide. A method for assimilating Lagrangian data into a shallow-water equation ocean model. *Mon. Wea. Rev.*, 134(4):1081–1101, 2006.

[SS10]   Ch. Schwab and A.M. Stuart. Sparse deterministic approximation of Bayesian inverse problems. *Submitted*, http://arxiv.org/abs/1103.4522, 2010.

[Stu10]   A.M. Stuart. Inverse problems: a Bayesian approach. *Acta Numerica*, 19, 2010.

[Tie98]   L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab.*, 8(1):1–9, 1998.

[VIJ10]   G. Vernieres, K. Ide, and C.K.R.T. Jones. Lagrangian data assimilation, an application to the gulf of mexico. *Physica D*, 2010.