Well-posedness of Bayesian inverse problems in function spaces: analysis and algorithms

T. J. Sullivan[1,2]

Summer School / Workshop on Computational Mathematics and Data Science
University of Oulu, FI, 22–24 August 2018

[1]Free University of Berlin, DE
[1]Zuse Institute Berlin, DE

Introduction and overview

Contents

Introduction and overview

What are inverse problems?

Inverse problems

- An **inverse problem** consists of recovering an unknown $u \in \mathcal{U}$ from an observation $y \in \mathcal{Y}$. Usually we have in hand an idealised **forward operator** $G \colon \mathcal{U} \to \mathcal{Y}$ and we think that $y \approx G(u)$.
- Given $y \in \mathcal{Y}$, the problem of finding $u \in \mathcal{U}$ such that $y = G(u)$ is often **ill-posed**:
  - there may be **no** such u (because $y \notin G(\mathcal{U})$); or
  - there may be **multiple** such u (because G is not injective); or
  - there may be a unique solution u, but it may depend **sensitively** on y (because G has a discontinuous inverse).
- These issues become even more prominent when we accept that there are inevitable observational errors, so we are more likely to see e.g.

$$y = G(u) + \varepsilon,$$

with ε random, perhaps with known distribution.

Examples of inverse problems

- In **numerical weather prediction** we are given noisy observations y of today's weather state $u =$ (air temperature, pressure, velocity, ...) at various sites around the world, and must reconstruct the complete current state u and use it to predict tomorrow's state... and do all this within six hours!
- In **X-ray tomography** we aim to recover a compactly supported function $u \colon \mathbb{R}^d \to \mathbb{R}$ from a noisy and partial observation of its Radon transform

$$Gu \colon \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R} \qquad (Gu)(\widehat{n}, s) := \int_{\{x \in \mathbb{R}^d \mid x \cdot \widehat{n} = s\}} u(x) \, \mathrm{d}x.$$

- In an **elliptic inverse problem** we are given noisy observations y of $(p(x_j))_{j=1}^J$ at $x_1, \dots, x_J \in D \subset \mathbb{R}^d$, and $f \colon D \to \mathbb{R}$, and seek $a = \exp(u) \colon D \to \mathbb{R}^{d \times d}$ such that

$$-\nabla \cdot (a \nabla p) = f, \qquad p|_{\partial D} = 0.$$

A variation on this problem — with applied currents and voltage observations at the boundary — underlies electrical impedance tomography (Somersalo et al., 1992; Dunlop and Stuart, 2016).

Variational approaches to inverse problems

- The **variational approach** to the inverse problem is to seek $u^\star \in \mathcal{U}$ that approximately solves the "impossible" equation $y = G(u)$ by minimising a weighted misfit:

$$u^\star \in \operatorname*{arg\,min}_{u \in \mathcal{U}} \Phi(u; y).$$

- In the case $\varepsilon \sim \mathcal{N}(0, \Gamma)$ on $\mathcal{Y} = \mathbb{R}^J$, we take

$$\Phi(u; y) = \tfrac{1}{2} \| y - G(u) \|_\Gamma^2 = \tfrac{1}{2} \| \Gamma^{-1/2} (y - G(u)) \|^2;$$

  when G is a linear operator, the Gauss–Markov theorem from basic statistics tells us that this choice yields an unbiased estimate $u^\star$ with minimal error and variance.

- Furthermore, we often have a prior belief that u "should" be close to some known value, or have desired properties such as sparsity, or smoothness, or having some jumps but not too many, so we seek a **regularised solution**

$$u^\star \in \operatorname*{arg\,min}_{u \in \mathcal{U}} \Phi(u; y) + R(u),$$

  where e.g. $R(u) = \frac{1}{2} \|u\|_2^2$, $\|u\|_1$, $\|u\|_{\mathrm{TV}}$, …

- A Bayesian probabilist/statistician gets around all of these problems by regarding $\boldsymbol{u}$, $\boldsymbol{y}$, and $\boldsymbol{\varepsilon}$ as random variables.
- The equation "$\boldsymbol{y} = G(\boldsymbol{u}) + \boldsymbol{\varepsilon}$" is treated as defining the conditional distribution of $\boldsymbol{y}|\boldsymbol{u}$.
- The Bayesian also has a prior belief about $\boldsymbol{u}$ before seeing the data, $\boldsymbol{u} \sim \mu$.
- The Bayesian **solution** to the inverse problem is the posterior distribution of $\boldsymbol{u}|\boldsymbol{y} = y$.

Bayes' rule for finite sets $\mathcal{U}$, $\mathcal{Y}$

The prior probability mass function $p_{\boldsymbol{u}} \colon \mathcal{U} \to [0, 1]$ of $\boldsymbol{u}$, the mass function $p_{\boldsymbol{y}|\boldsymbol{u}}(\cdot|u)$ of $\boldsymbol{y}|\boldsymbol{u} = u$, and the posterior mass function $p_{\boldsymbol{u}|\boldsymbol{y}}(\cdot|y)$ of $\boldsymbol{u}|\boldsymbol{y} = y$ satisfy

$$p_{\boldsymbol{u}|\boldsymbol{y}}(u|y) = \frac{p_{\boldsymbol{y}|\boldsymbol{u}}(y|u) p_{\boldsymbol{u}}(u)}{Z(y)}, \qquad Z(y) = \sum_{u' \in \mathcal{U}} p_{\boldsymbol{y}|\boldsymbol{u}}(y|u') p_{\boldsymbol{u}}(u').$$

Formally, the variational approach amounts to setting $R(u) = -\log p_{\boldsymbol{u}}(u)$ and finding a point of maximum Bayesian posterior probability... see the lectures of **Tapio Helin**!

- A Bayesian probabilist/statistician gets around all of these problems by regarding $\boldsymbol{u}$, $\boldsymbol{y}$, and $\boldsymbol{\varepsilon}$ as random variables.
- The equation "$\boldsymbol{y} = G(\boldsymbol{u}) + \boldsymbol{\varepsilon}$" is treated as defining the conditional distribution of $\boldsymbol{y}|\boldsymbol{u}$.
- The Bayesian also has a prior belief about $\boldsymbol{u}$ before seeing the data, $\boldsymbol{u} \sim \mu$.
- The Bayesian **solution** to the inverse problem is the posterior distribution of $\boldsymbol{u}|\boldsymbol{y} = y$.

Bayes' rule for finite-dimensional spaces $\mathcal{U}, \mathcal{Y}$

The prior density $\rho_{\boldsymbol{u}} \colon \mathcal{U} \to \mathbb{R}$ of $\boldsymbol{u}$, the density $\rho_{\boldsymbol{y}|\boldsymbol{u}}(\,\cdot\,|u)$ of $\boldsymbol{y}|\boldsymbol{u} = u$, and the posterior density $\rho_{\boldsymbol{u}|\boldsymbol{y}}(\,\cdot\,|y)$ of $\boldsymbol{u}|\boldsymbol{y} = y$ satisfy

$$\rho_{\boldsymbol{u}|\boldsymbol{y}}(u|y) = \frac{\rho_{\boldsymbol{y}|\boldsymbol{u}}(y|u)\rho_{\boldsymbol{u}}(u)}{Z(y)}, \qquad Z(y) = \int_{\mathcal{U}} \rho_{\boldsymbol{y}|\boldsymbol{u}}(y|u')\rho_{\boldsymbol{u}}(u')\,\mathrm{d}u'.$$

Formally, the variational approach amounts to setting $R(u) = -\log \rho_{\boldsymbol{u}}(u)$ and finding a point of maximum Bayesian posterior probability… see the lectures of **Tapio Helin**!

Bayesian conditioning

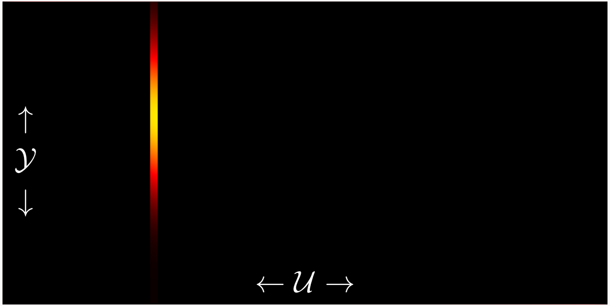
Prior density $\rho_{\boldsymbol{u}}$ on $\mathcal{U}$:

Bayesian conditioning

Prior density $\rho_{\boldsymbol{u}}$ on $\mathcal{U}$:

Observation density $\rho_{\boldsymbol{y}|\boldsymbol{u}}(y|u)$ on $\{u\} \times \mathcal{Y}$

$\implies$ joint density $\rho_{(\boldsymbol{u},\boldsymbol{y})}(u,y) = \rho_{\boldsymbol{u}}(u)\rho_{\boldsymbol{y}|\boldsymbol{u}}(y|u)$ on $\mathcal{U} \times \mathcal{Y}$:



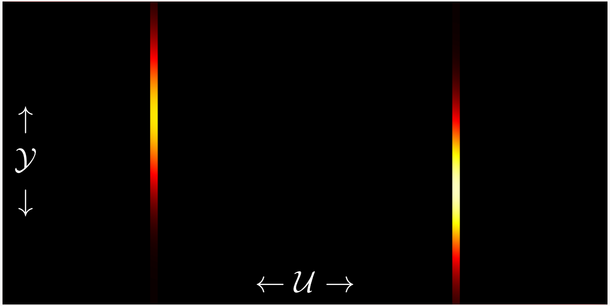
u

Bayesian conditioning

Prior density $\rho_{\boldsymbol{u}}$ on $\mathcal{U}$:

Observation density $\rho_{\boldsymbol{y}|\boldsymbol{u}}(y|u)$ on $\{u\} \times \mathcal{Y}$

$\Longrightarrow$ joint density $\rho_{(\boldsymbol{u},\boldsymbol{y})}(u,y) = \rho_{\boldsymbol{u}}(u)\rho_{\boldsymbol{y}|\boldsymbol{u}}(y|u)$ on $\mathcal{U} \times \mathcal{Y}$:



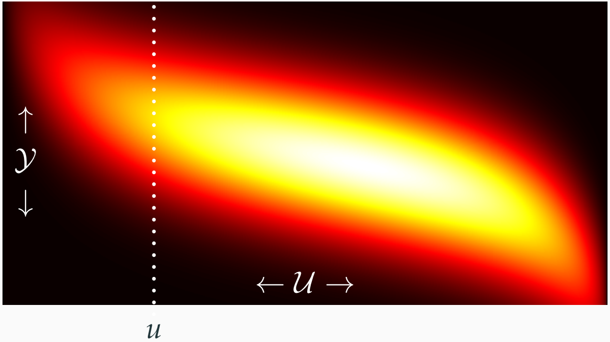
u

Bayesian conditioning

Prior density $\rho_{\boldsymbol{u}}$ on $\mathcal{U}$:

Observation density $\rho_{\boldsymbol{y}|\boldsymbol{u}}(y|u)$ on $\{u\} \times \mathcal{Y}$

$\implies$ joint density $\rho_{(\boldsymbol{u},\boldsymbol{y})}(u,y) = \rho_{\boldsymbol{u}}(u)\rho_{\boldsymbol{y}|\boldsymbol{u}}(y|u)$ on $\mathcal{U} \times \mathcal{Y}$:

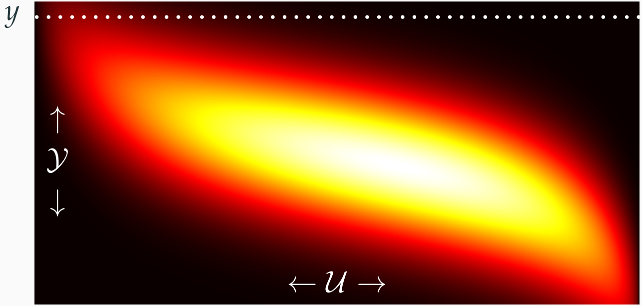


Bayesian conditioning

Prior density $\rho_{\boldsymbol{u}}$ on $\mathcal{U}$:

Observation density $\rho_{\boldsymbol{y}|\boldsymbol{u}}(y|u)$ on $\{u\} \times \mathcal{Y}$

$\implies$ joint density $\rho_{(\boldsymbol{u},\boldsymbol{y})}(u,y) = \rho_{\boldsymbol{u}}(u)\rho_{\boldsymbol{y}|\boldsymbol{u}}(y|u)$ on $\mathcal{U} \times \mathcal{Y}$:

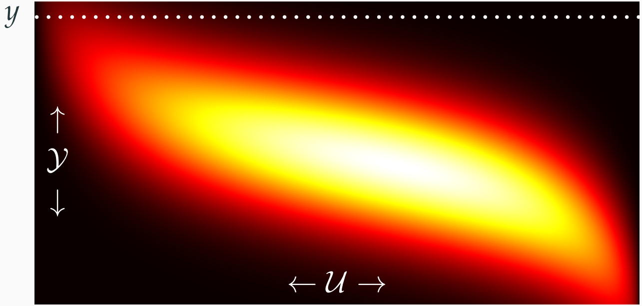


Bayesian conditioning

Prior density $\rho_{\boldsymbol{u}}$ on $\mathcal{U}$:

Observation density $\rho_{\boldsymbol{y}|\boldsymbol{u}}(y|u)$ on $\{u\} \times \mathcal{Y}$

$\implies$ joint density $\rho_{(\boldsymbol{u},\boldsymbol{y})}(u,y) = \rho_{\boldsymbol{u}}(u)\rho_{\boldsymbol{y}|\boldsymbol{u}}(y|u)$ on $\mathcal{U} \times \mathcal{Y}$:



Posterior density $\rho_{\boldsymbol{u}|\boldsymbol{y}}(u|y) \propto \rho_{\boldsymbol{y}|\boldsymbol{u}}(y|u)\rho_{\boldsymbol{u}}(u)$ on $\mathcal{U}$:

Problem: on spaces of functions, fields etc., there is nothing like a Lebesgue measure with respect to which we could form probability density functions!

Theorem 1 (e.g. Yamasaki, 1985, Part B, Section 5)

Let $\mathcal{U}$ be a normed vector space over $\mathbb{R}$ and $\mathcal{B}(\mathcal{U})$ its Borel σ-algebra. If $\dim \mathcal{U} < \infty$, then there is a ***Lebesgue measure*** *μ on $(\mathcal{U}, \mathcal{B}(\mathcal{U}))$ that is simultaneously*

- *strictly positive, i.e. for all $x \in \mathcal{U}$ and $r > 0$, $\mu(B(x,r)) > 0$;*
- *locally finite, i.e. for all $x \in \mathcal{U}$ and small enough $r > 0$, $\mu(B(x,r)) < \infty$;*
- *translation invariant, i.e. for all $x \in \mathcal{U}$ and $A \in \mathcal{B}(\mathcal{U})$, $\mu(A) = \mu(x + A)$;*

and it is unique up to multiplication by a constant. If $\dim \mathcal{U} = \infty$, then no such μ exists.

Indeed, on an infinite-dimensional normed space $\mathcal{U}$, the zero measure $\mu \equiv 0$ is the only σ-finite measure μ on $(\mathcal{U}, \mathcal{B}(\mathcal{U}))$ for which being μ-measure zero is preserved by all translations.

- Suppose that we wish to reconstruct a "blocky" image $u \colon [0,1] \to \mathbb{R}$ from a noisy, J-dimensional linear observation: $\boldsymbol{y} = Au + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Gamma)$, $\Gamma \in \mathbb{R}^{J \times J}_{++}$.
- We discretise using $(N+1) \in \mathbb{N}$ pixels, represent u by the piecewise linear interpolation of $u^{(N)} := (u_n)_{n=0}^{N}$, $u_n := u(\frac{n}{N})$, and examine the naïve "posterior distribution" on $\mathbb{R}^{N+1}$

$$\rho_{\boldsymbol{u}^{(N)}|\boldsymbol{y}}(u^{(N)}|y) \propto \exp\left(-\tfrac{1}{2}\|\Gamma^{-1/2}(y - Au^{(N)})\|^2 + \alpha_N \|u^{(N)}\|_{\mathrm{TV}}\right)$$
$$= \exp\left(-\frac{1}{2}\sum_{j=1}^{J} |(\Gamma^{-1/2}(y - Au^{(N)}))_j|^2 + \alpha_N \sum_{n=1}^{N} |u_n - u_{n-1}|\right)$$

Theorem 2 (Lassas and Siltanen, 2004)

There is no choice of $(\alpha_N)_{N \in \mathbb{N}}$ such that the prior $\rho_{\boldsymbol{u}} \propto \exp(-\alpha_N \| \cdot \|_{\mathrm{TV}})$ and the posterior $\rho_{\boldsymbol{u}^{(N)}|\boldsymbol{y}}(\cdot | y)$ both have "sensible" limits as $N \to \infty$.

The mission of these lectures

- A mathematical problem is called **well-posed** if, given the problem data, there exists a solution, the solution is unique, and the solution depends continuously on the data.
- Interpreted naïvely, inverse problems are typically ill-posed.
- These lectures show how Bayesian inverse problems are well-posed: the posterior distribution is uniquely determined and depends continuously on the problem setup with respect to a suitable statistical distance, namely the Hellinger metric.
- We focus on the mathematical structure from the function space viewpoint[1], and hence on theory and algorithms that are general enough to consider big data, and big unknowns — assuming that infinite-dimensional is big enough for you?

[1] The function-space-down viewpoint complements the "Finnish school" of finite-dimensions-up *grid independence* (Lehtinen, 1999) or *discretisation invariance* (Lassas and Siltanen, 2004; Lassas et al., 2009).

Acknowledgments

- These notes follow a mathematical theme initiated by the seminal article of Stuart (2010) and further developed in lecture notes and texts such as Dashti and Stuart (2017), Schillings and Sprungk (2017), and Sullivan (2015).
- My research has been primarily supported by the Free University of Berlin, the Zuse Institute Berlin, and the German Research Foundation (DFG); additional support has been received from the US National Science Foundation via SAMSI, and the Alan Turing Institute.

Introduction and overview

Notation

Notation I

- $\mathcal{U}, \mathcal{V}, \ldots$ will be separable **Banach or Hilbert spaces**,[2] almost always over $\mathbb{R}$.
- Here, $\mathcal{U}'$ denotes the **dual space** of $\mathcal{U}$, the space of continuous linear maps from $\mathcal{U}$ to $\mathbb{R}$; we write $\langle \ell | u \rangle$ for the scalar value of $\ell \in \mathcal{U}'$ acting on $u \in \mathcal{U}$.
- If $\mathcal{H}$ is a Hilbert space and $\Gamma \colon \mathcal{H} \to \mathcal{H}$ is a symmetric, positive-definite, linear operator, then we define the **Γ-weighted inner product and norm** by

$$\langle u, v \rangle_{\Gamma} := \langle \Gamma^{-1/2} u, \Gamma^{-1/2} v \rangle_{\mathcal{H}} \qquad \| u \|_{\Gamma} := \langle u, u \rangle_{\Gamma}^{1/2} = \| \Gamma^{-1/2} u \|_{\mathcal{H}}.$$

- $\mathcal{M}(\mathcal{U})$ denotes the set of all (non-negative, σ-additive, σ-finite, Borel) measures on $\mathcal{U}$.
- $\mathcal{P}(\mathcal{U}) \subset \mathcal{M}(\mathcal{U})$ denotes the set of probability measures on $\mathcal{U}$.

---

[2] Because all our spaces $\mathcal{U}$ will be separable and completely metrisable, we gain two technical advantages: (1) the **Borel σ-algebra** $\mathcal{B}(\mathcal{U})$ generated by the open balls coincides with the **cylindrical σ-algebra** $\mathcal{E}(\mathcal{U}) := \sigma(\{\ell^{-1}(E) \mid \ell \in \mathcal{U}', E \in \mathcal{B}(\mathbb{R})\})$ generated by $\mathcal{U}'$, and (2) all probability measures on $\mathcal{U}$ are inner regular a.k.a. Radon.

Notation II

- Notation for the **Lebesgue integral** / **expected value** of $f\colon \mathcal{U} \to \mathbb{R}$ with respect to $\mu \in \mathcal{M}(\mathcal{U})$:
$$\int_{\mathcal{U}} f \, \mathrm{d}\mu \equiv \int_{\mathcal{U}} f(u) \, \mathrm{d}\mu(u) \equiv \int_{\mathcal{U}} f(u) \, \mu(\mathrm{d}u) \equiv \mathbb{E}_{\mu}[f].$$
- $L^p(\mathcal{U}, \mu; \mathcal{V}) := \{ f\colon \mathcal{U} \to \mathcal{V} \mid \|f\|_{L^p(\mu)} := (\int_{\mathcal{U}} \|f(u)\|_{\mathcal{V}}^p \, \mathrm{d}\mu(u))^{1/p} < \infty \}$, modulo equality μ-a.e., are the **Lebesgue** or **Lebesgue–Bochner spaces**.
- When $\mu, \nu \in \mathcal{M}(\mathcal{U})$ and $\nu(E) = 0 \implies \mu(E) = 0$, we say that μ is **absolutely continuous** with respect to ν and write $\mu \ll \nu$. In this case, the Radon–Nikodym theorem ensures the existence of a ν-a.e. unique **density function** or **Radon–Nikodym derivative** $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} \colon \mathcal{U} \to \mathbb{R}$ such that
$$\int_{\mathcal{U}} f \, \mathrm{d}\mu = \int_{\mathcal{U}} f \cdot \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \, \mathrm{d}\nu \qquad \text{for all } f \in L^1(\mu).$$

Notation III

- For $\mu \in \mathcal{M}(\mathcal{U})$ and $\nu \in \mathcal{M}(\mathcal{V})$, $\mu \otimes \nu \in \mathcal{M}(\mathcal{U} \times \mathcal{V})$ denotes their **product measure**, completely determined by its values on rectangles:
$$(\mu \otimes \nu)(A \times B) := \mu(A)\nu(B) \qquad \text{for } A \in \mathcal{B}(\mathcal{U}), B \in \mathcal{B}(\mathcal{V}).$$
- The **push-forward** or **image measure** of $\mu \in \mathcal{M}(\mathcal{U})$ under $f \colon \mathcal{U} \to \mathcal{V}$ is $f_\sharp \mu \in \mathcal{M}(\mathcal{V})$,
$$(f_\sharp \mu)(B) := \mu(f^{-1}(B)) = \mu(\{u \in \mathcal{U} \mid f(u) \in B\}) \qquad \text{for } B \in \mathcal{B}(\mathcal{V}),$$
with the resulting **change-of-variables formula** for integration:
$$\int_{\mathcal{V}} g(v)\, (f_\sharp \mu)(\mathrm{d}v) = \int_{\mathcal{U}} g(f(u))\, \mu(\mathrm{d}u) \qquad \text{for } g \in L^1(\mathcal{V}, f_\sharp \mu; \mathbb{R}).$$
- The **law** or **distribution** $\mu_{\boldsymbol{u}} \in \mathcal{P}(\mathcal{U})$ of a random variable $\boldsymbol{u} \colon \Omega \to \mathcal{U}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is the push-forward measure $\mu_{\boldsymbol{u}} := \boldsymbol{u}_\sharp \mathbb{P}$, i.e.
$$\mu_{\boldsymbol{u}}(A) = \mathbb{P}(\{\omega \in \Omega \mid \boldsymbol{u}(\omega) \in A\}) = \mathbb{P}[\boldsymbol{u} \in A] \qquad \text{for } A \in \mathcal{B}(\mathcal{U}),$$
and random variables $\boldsymbol{u}, \boldsymbol{y}$ are **independent** exactly when $\mu_{(\boldsymbol{u},\boldsymbol{y})} = \mu_{\boldsymbol{u}} \otimes \mu_{\boldsymbol{v}}$.

Probability on function spaces

Three main ways to reason about random elements of infinite-dimensional spaces, such as spaces of functions:

- series expansions with respect to a given basis, with random coefficients;
- work in a coordinate-free way — easiest if there is some nice structure, e.g. Gaussian;
- (only briefly discussed here) for Gaussian random functions $\boldsymbol{u} \colon D \to \mathbb{R}$, express everything in terms of a mean function $m \colon D \to \mathbb{R}$ and a covariance kernel $c \colon D \times D \to \mathbb{R}$, $c(x, x') = \operatorname{Cov}(\boldsymbol{u}(x), \boldsymbol{u}(x'))$.

In this section we also discuss how to condition random variables that take values in such spaces.

Probability on function spaces

Random series

- The first, easiest, way to build probability measures on an infinite-dimensional space $\mathcal{U}$ is to fix a basis $(\psi_n)_{n\in\mathbb{N}}$ of $\mathcal{U}$ and to form random series expansions in this basis:

$$\boldsymbol{u} := \sum_{n\in\mathbb{N}} \xi_n \psi_n$$

  where the real-valued random variables $(\xi_n)_{n\in\mathbb{N}}$ are chosen appropriately.
- For example, take $(\psi_n)_{n\in\mathbb{N}}$ to be a Fourier or wavelet basis of $\mathcal{U} = L^2([0,1], \mathrm{d}x; \mathbb{R})$.
- Once the finite-dimensional distributions of $\boldsymbol{\xi} := (\xi_n)_{n\in\mathbb{N}}$ are determined, the Daniell–Kolmogorov extension theorem implies that $\boldsymbol{\xi}$ is a $\mathbb{R}^{\mathbb{N}}$-valued random variable[3] and its law $\mu_{\boldsymbol{\xi}} \in \mathcal{P}(\mathbb{R}^{\mathbb{N}})$.
- Question: what conditions do we need to impose on $\mathcal{U}$, $(\psi_n)_{n\in\mathbb{N}}$, and $\boldsymbol{\xi}$ to ensure that $\boldsymbol{u}$ is a well-defined $\mathcal{U}$-valued random variable, i.e. that $\mu_{\boldsymbol{u}} \in \mathcal{P}(\mathcal{U})$?

[3] Again, countability/separability is helpful: Daniell–Kolmogorov would only say that $\boldsymbol{\xi}$ is measurable w.r.t. the product σ-algebra on $\mathbb{R}^{\mathbb{N}}$, but fortunately the product σ-algebra of a countable product of separable metric spaces is the Borel σ-algebra of the product.

Example 3

Consider the Hilbert space

$$\ell^2 = \left\{ v = (v_n)_{n\in\mathbb{N}} \;\middle|\; \|v\|_{\ell^2}^2 := \sum_{n\in\mathbb{N}} |v_n|^2 < \infty \right\}$$

and a random sequence $\xi := (\xi_n)_{n\in\mathbb{N}}$ with independent components $\xi_n \sim \mathrm{Unif}(-\frac{1}{n}, \frac{1}{n})$. Then the random sequence $\xi \in \ell^2$ a.s., since we have the upper bound

$$\|\xi\|_{\ell^2}^2 = \sum_{n\in\mathbb{N}} |\xi_n|^2 \leq \sum_{n\in\mathbb{N}} \frac{1}{n^2} = \frac{\pi^2}{6} < \infty.$$

If now $\mathcal{H}$ is a separable Hilbert space with complete orthonormal basis $(\psi_n)_{n\in\mathbb{N}}$, and $u := \sum_{n\in\mathbb{N}} \xi_n \psi_n$, then Parseval's identity says that $\|u\|_{\mathcal{H}} = \|\xi\|_{\ell^2}$, so $u \in \mathcal{H}$ a.s.

Example 4

One way to characterise the Sobolev space $H^s(-\pi, \pi)$, $s \in \mathbb{N} \cup \{0\}$, is via the decay of Fourier modes (indexed by $n \in \mathbb{Z}$ rather than $n \in \mathbb{N}$):

$$u = \sum_{n \in \mathbb{Z}} u_n \psi_n \in H^s \iff \sum_{n \in \mathbb{Z}} (1 + |n|^2)^s |u_n|^2 < \infty,$$

$$\psi_n(x) := (2\pi)^{-1/2} e^{inx}.$$

Thus, if we take $\boldsymbol{u} := \sum_{n \in \mathbb{Z}} \xi_n \psi_n$ with independent coefficients $\xi_n \sim \mathrm{Unif}\big(-\frac{1}{1+|n|}, \frac{1}{1+|n|}\big)$, then $\boldsymbol{u} \in H^0 = L^2$ a.s.

If we take $\boldsymbol{u} := \sum_{n \in \mathbb{Z}} \xi_n \psi_n$ with independent coefficients $\xi_n \sim \mathrm{Unif}\big(-\frac{1}{(1+|n|)^2}, \frac{1}{(1+|n|)^2}\big)$, then $\boldsymbol{u} \in H^1$ a.s. Furthermore, by the Sobolev embedding theorem, H^1 embeds into C^0 in one space dimension, so $\boldsymbol{u}$ is a.s. a continuous function.

Theorem 5 (Kolmogorov's two-series theorem)

Let $(x_n)_{n\in\mathbb{N}}$ be a sequence of independent $\mathbb{R}$-valued random variables. If $\sum_{n\in\mathbb{N}} \mathbb{E}[x_n]$ converges absolutely and $\sum_{n\in\mathbb{N}} \mathbb{E}[x_n^2]$ converges, then $\sum_{n\in\mathbb{N}} x_n$ converges almost surely.

Exercise 6

Let $\mathcal{H}$ be a separable Hilbert space with complete orthonormal basis $(\psi_n)_{n\in\mathbb{N}}$. Let $u := \sum_{n\in\mathbb{N}} \sigma_n \eta_n \psi_n$, where the $\eta_n \sim \mathcal{N}(0,1)$ are i.i.d. Use Kolmogorov's two-series theorem to show that, if $(\sigma_n)_{n\in\mathbb{N}} \in \ell^2$, then $u \in \mathcal{H}$ a.s.

Hence construct Gaussian random functions that a.s. lie in the Sobolev spaces H^s.

Harder: show that $(\sigma_n \eta_n)_{n\in\mathbb{N}}$ is a.s. bounded if and only if $\sum_{n\in\mathbb{N}} \exp(-\varepsilon/2\sigma_n^2)$ is finite for some $\varepsilon > 0$, and a.s. tends to 0 if and only if this series is finite for all $\varepsilon > 0$.

Similar methods, with Γ-distributed coefficients and wavelets ψ_n, can be used to define random functions in the Besov spaces (Lassas et al., 2009; Dashti et al., 2012).

To handle coefficients whose tails are much heavier than Gaussians, we need

Theorem 7 (Kolmogorov's three-series theorem)

Let $(x_n)_{n\in\mathbb{N}}$ *be a sequence of independent* $\mathbb{R}$*-valued random variables. Then* $\sum_{n\in\mathbb{N}} x_n$ *converges almost surely if and only if there exists* $A \in [0, \infty]$ *such that*

- $\sum_{n\in\mathbb{N}} \mathbb{P}[|x_n| \geq A]$ *converges; and*
- $\sum_{n\in\mathbb{N}} \mathbb{E}[x_n \mathbb{1}_{|x_n|<A}]$ *converges absolutely; and*
- $\sum_{n\in\mathbb{N}} \mathbb{E}[x_n^2 \mathbb{1}_{|x_n|<A}]$ *converges.*

Examples of such heavy-tailed coefficients include the α-stable distributions, e.g. the $\alpha = 1$ **Cauchy distribution** $\mathcal{C}(\delta, \gamma)$ with location $\delta \in \mathbb{R}$ and scale $\gamma > 0$,

$$\frac{\mathrm{d}\mathcal{C}(\delta, \gamma)}{\mathrm{d}x}(x) = \frac{1}{\gamma\pi}\frac{1}{1 + ((x-\delta)/\gamma)^2}.$$

Suppose that the basis $(\psi_n)_{n\in\mathbb{N}}$ and $q > 0$ are such that $\underline{v} := (v_n)_{n\in\mathbb{N}} \mapsto \sum_{n\in\mathbb{N}} v_n\psi_n$ is a continuous embedding of the sequence space ℓ^q of coefficients into $\mathcal{U}$, i.e.

$$\left\| \sum_{n\in\mathbb{N}} v_n\psi_n \right\|_{\mathcal{U}} \leq C\|\underline{v}\|_{\ell^q}, \qquad \|\underline{v}\|_{\ell^q}^q := \sum_{n\in\mathbb{N}} |v_n|^q.$$

For an orthonormal basis or a Riesz basis of a Hilbert space, we can take $q = 2$; in general, we might be forced to take $q = 1$ — or smaller, in quasi-normed spaces.

Theorem 8 (Well-definedness of $\mathcal{U}$-valued Cauchy random variables; Sullivan, 2017)

The random series $\boldsymbol{u} := \sum_{n\in\mathbb{N}} \boldsymbol{\xi}_n\psi_n$ with $\boldsymbol{\xi}_n \sim \mathcal{C}(\delta_n, \gamma_n)$ converges a.s. if $\underline{\gamma} \in \ell^1$, $\underline{\delta} \in \ell^q$, and

$$[\underline{\gamma}]_{\ell\log\ell} := \sum_{n\in\mathbb{N}} \bigl|\gamma_n \log|\gamma_n|\bigr| < \infty, \qquad \text{if } q = 1.$$

Sketch proof of Theorem 8

Write $\boldsymbol{\xi}_n = \delta_n + \gamma_n \boldsymbol{\eta}_n$ with $\boldsymbol{\eta}_n \sim \mathcal{C}(0,1)$. For $q = 1$, the key estimates are, for any $A > 0$,

$$\mathbb{P}\big[|\gamma \boldsymbol{\eta}_n| \geq A\big] = 1 - \frac{2}{\pi} \arctan \frac{A}{\gamma_n},$$

$$\mathbb{E}\big[|\gamma_n \boldsymbol{\eta}_n| \mathbb{1}_{|\gamma_n \boldsymbol{\eta}_n| < A}\big] = \frac{\gamma_n}{\pi} \log\left(1 + \frac{A^2}{\gamma_n^2}\right),$$

$$\mathbb{E}\big[|\gamma_n \boldsymbol{\eta}_n|^2 \mathbb{1}_{|\gamma_n \boldsymbol{\eta}_n| < A}\big] = \frac{2A\gamma_n}{\pi} + \frac{2\gamma_n^2}{\pi} \arctan \frac{A}{\gamma_n}.$$

These quantities are summable over $n \in \mathbb{N}$ under the above assumptions on $\underline{\gamma}$. Thus, Kolmogorov's three-series theorem gives us $\mathbb{P}$-a.s. convergence of $\sum_{n \in \mathbb{N}} \boldsymbol{\xi}_n \psi_n$ with respect to the norm on $\mathcal{U}$, so $\boldsymbol{u} \in \mathcal{U}$ $\mathbb{P}$-a.s.

Note that the assumption that $[\underline{\gamma}]_{\ell \log \ell} < \infty$ is slightly stronger than $\|\underline{\gamma}\|_{\ell^1} < \infty$; without it, we lose summability of the first truncated moments. $\square$

Probability on function spaces

Gaussian measures

Gaussian measures on $\mathbb{R}^n$

Definition 9

The **Gaussian measure** or **normal distribution** $\mathcal{N}(m, \sigma^2)$ on $\mathbb{R}$ with mean $m \in \mathbb{R}$ and variance $\sigma^2 > 0$ is the measure with Lebesgue density

$$\frac{\mathrm{d}\mathcal{N}(m, \sigma^2)}{\mathrm{d}x}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2} \right).$$

We sometimes think of the Dirac measure / point mass centred at m as a degenerate Gaussian measure, one with zero variance.

Definition 10

The **Gaussian measure** or **normal distribution** $\mathcal{N}(m, C)$ on $\mathbb{R}^n$ with mean $m \in \mathbb{R}^n$ and symmetric, positive-definite covariance matrix $C \in \mathbb{R}^{n \times n}$ has Lebesgue density

$$\frac{\mathrm{d}\mathcal{N}(m, C)}{\mathrm{d}x}(x) = \frac{1}{\sqrt{\det(2\pi C)}} \exp\left(-\tfrac{1}{2} \| C^{-1/2}(x-m) \|^2 \right).$$

Definition 11

The **characteristic function** or **Fourier transform** $\widehat{\mu}\colon \mathcal{U}' \to \mathbb{C}$ of $\mu \in \mathcal{P}(\mathcal{U})$ is

$$\widehat{\mu}(\ell) := \int_{\mathcal{U}} \exp(i\langle \ell | x \rangle)\, \mu(\mathrm{d}x).$$

- The characteristic function of μ completely characterises it, i.e.

$$\mu = \nu \in \mathcal{P}(\mathcal{U}) \iff \widehat{\mu} = \widehat{\nu}\colon \mathcal{U}' \to \mathbb{C}.$$

- For any continuous linear map $T\colon \mathcal{U} \to \mathcal{V}$,

$$\widehat{T_{\sharp}\mu}(\ell) = \widehat{\mu}(T^{*}\ell) \qquad \text{for all } \ell \in \mathcal{V}',$$

  where $T^{*}: \mathcal{V}' \to \mathcal{U}'$ is the **adjoint operator** to T, i.e. $\langle T^{*}\ell | u \rangle := \langle \ell | Tu \rangle$.
- In the finite-dimensional Gaussian case, for $\mu \in \mathcal{P}(\mathbb{R}^{n})$,

$$\mu = \mathcal{N}(m, C) \iff \widehat{\mu}(\ell) = \exp\bigl(i\ell \cdot m - \tfrac{1}{2}\ell \cdot C\ell\bigr).$$

  (We can use this formula to define Gaussians with indefinite covariance C.)

Theorem 12 (e.g. Bogachev, 1998, Theorem 2.2.4)

Let $\mathcal{U}$ be a separable Banach space. The following are equivalent and, if one (and hence any) holds, then $\mu \in \mathcal{P}(\mathcal{U})$ is called a ***Gaussian measure****:*

- *for every $\ell \in \mathcal{U}'$, $\ell_\sharp \mu \in \mathcal{P}(\mathbb{R})$ is Gaussian;*
- *for every continuous linear map $T \colon \mathcal{U} \to \mathbb{R}^n$ with $n \in \mathbb{N}$, $T_\sharp \mu \in \mathcal{P}(\mathbb{R}^n)$ is Gaussian;*
- *the characteristic function $\widehat{\mu} \colon \mathcal{U}' \to \mathbb{C}$ of μ is*

$$\widehat{\mu}(\ell) = \exp\bigl(i\langle \ell | m \rangle - \tfrac{1}{2} Q(\ell)\bigr)$$

  for some $m \in \mathcal{U}$ and some non-negative quadratic form Q on $\mathcal{U}'$.

A random variable $\boldsymbol{u} \colon \Omega \to \mathcal{U}$ is ***Gaussian*** *if its law $\mu_{\boldsymbol{u}} \in \mathcal{P}(\mathcal{U})$ is a Gaussian measure.*

Example 13

The series with Gaussian random coefficients from earlier are Gaussian in the new sense.

Definition 14

The **mean** of $\mu \in \mathcal{P}(\mathcal{U})$ is $m_\mu \in \mathcal{U}$ if

$$\langle \ell | m_\mu \rangle = \int_{\mathcal{U}} \langle \ell | x \rangle \, \mu(\mathrm{d}x) \quad \text{for all } \ell \in \mathcal{U}'.$$

If $m_\mu = 0$, then μ is said to be **centred**.

The **covariance operator** of μ is the linear operator $C_\mu \colon \mathcal{U}' \to \mathcal{U}''$,

$$\langle C_\mu k | \ell \rangle := \int_{\mathcal{U}} \langle k | x - m_\mu \rangle \langle \ell | x - m_\mu \rangle \, \mu(\mathrm{d}x) \quad \text{for } k, \ell \in \mathcal{U}'.$$

We also often think of C_μ as a bilinear form on $\mathcal{U}'$, or, when $\mathcal{U}$ is Hilbert, $C_\mu \colon \mathcal{U} \to \mathcal{U}$, with

$$\langle C_\mu u, v \rangle = \int_{\mathcal{U}} \langle u, x - m \rangle \langle v, x - m \rangle \, \mu(\mathrm{d}x).$$

The inverse covariance operator $P_\mu := C_\mu^{-1}$ is called the **precision operator**.

A Gaussian measure always has a mean and a covariance operator, and in fact it is characterised by these two objects: its characteristic function is

$$\widehat{\mathcal{N}(m, C)}(\ell) = \exp\bigl(i\langle \ell | m \rangle - \tfrac{1}{2}\langle C\ell | \ell \rangle\bigr).$$

Exercise 15

Show that if $T \colon \mathcal{U} \to \mathcal{V}$ is a continuous linear map and $\boldsymbol{x} \sim \mathcal{N}(m, C) \in \mathcal{P}(\mathcal{U})$ and $\boldsymbol{y} \sim \mathcal{N}(n, D) \in \mathcal{P}(\mathcal{V})$ are independent, then $T\boldsymbol{x} + \boldsymbol{y} \sim \mathcal{N}(Tm + n, TCT^{*} + D)$.

Theorem 16 (Fernique, 1970; cf. Ledoux, 1996, Theorem 4.1)

Let $\mu = \mathcal{N}(m, C)$ on a separable Banach space $\mathcal{U}$. Then, for some $\alpha > 0$,

$$\int_{\mathcal{U}} \exp(\alpha \|x - m\|^2) \, \mu(\mathrm{d}x) < +\infty.$$

Hence, for all $k \geq 0$, $\int_{\mathcal{U}} \|x\|^k \, \mu(\mathrm{d}x) < \infty$, and C is a bounded linear operator with $\mathcal{R}(C) \subseteq \mathcal{U}$ in the sense of the usual embedding $\mathcal{U} \hookrightarrow \mathcal{U}''$. (Indeed, any $\alpha < (2\|C\|)^{-1}$ will work.)

Sazonov's theorem

Covariance operators are the infinite-dimensional versions of symmetric positive-definite matrices, and for finite-variance measures they always have finite trace:

$$\operatorname{tr}(K) := \sum_{n \in \mathbb{N}} \langle K\psi_n, \psi_n \rangle$$

for any orthonormal basis $(\psi_n)_{n \in \mathbb{N}}$ of $\mathcal{H}$, and (by Lidskiĭ's theorem) this equals the sum of the eigenvalues of K, counted with multiplicity.

Theorem 17 (Sazonov, 1958)

On a separable Hilbert space $\mathcal{H}$, let $\mu = \mathcal{N}(m, C) \in \mathcal{P}(\mathcal{H})$. Then $C \colon \mathcal{H} \to \mathcal{H}$ is symmetric, positive semi-definite, and of trace class (a.k.a. nuclear, and in particular compact) with

$$\operatorname{tr}(C) = \int_{\mathcal{H}} \|x - m\|^2 \, \mu(\mathrm{d}x) < \infty.$$

Conversely, if $K \colon \mathcal{H} \to \mathcal{H}$ is positive semi-definite, symmetric, and of trace class, then there is a Gaussian measure $\mu \in \mathcal{P}(\mathcal{H})$ such that $C_\mu = K$.

Proof of Theorem 17 i

For symmetry, let $u, v \in \mathcal{H}$. Then

$$\begin{aligned}
\langle Cu, v \rangle &= \int_{\mathcal{H}} \langle u, x-m \rangle \langle v, x-m \rangle \, \mu(\mathrm{d}x) \\
&= \int_{\mathcal{H}} \langle v, x-m \rangle \langle u, x-m \rangle \, \mu(\mathrm{d}x) \\
&= \langle Cv, u \rangle \\
&= \langle u, Cv \rangle.
\end{aligned}$$

For non-negativity, let $u \in \mathcal{H}$:

$$\langle Cu, u \rangle = \int_{\mathcal{H}} |\langle u, x-m \rangle|^2 \, \mu(\mathrm{d}x) \geq 0.$$

Incidentally, $\langle Cu, u \rangle = 0 \iff u \perp (x-m)$ for μ-a.a. $x \in \mathcal{H}$.

Proof of Theorem 17 II

Finally, let $(\psi_n)_{n\in\mathbb{N}}$ be a complete orthonormal basis of $\mathcal{H}$. Then

$$
\begin{aligned}
\operatorname{tr}(C) &= \sum_{n\in\mathbb{N}} \langle C\psi_n, \psi_n \rangle \\
&= \sum_{n\in\mathbb{N}} \int_{\mathcal{H}} |\langle \psi_n, x - m \rangle|^2 \, \mu(\mathrm{d}x) \\
&= \int_{\mathcal{H}} \sum_{n\in\mathbb{N}} |\langle \psi_n, x - m \rangle|^2 \, \mu(\mathrm{d}x) && \text{(Fubini)} \\
&= \int_{\mathcal{H}} \|x - m\|^2 \, \mu(\mathrm{d}x) && \text{(Parseval)} \\
&\leq 2\|m\|^2 + 2 \int_{\mathcal{H}} \|x\|^2 \, \mathrm{d}\mu(x) < \infty. && \text{(Fernique)}
\end{aligned}
$$

The converse is an exercise. Hint: choose a complete orthonormal system of eigenvectors of K and define a Gaussian measure on $\mathcal{H}$ by a random series in this basis. $\square$

Translating Gaussian measures I

- Recall that the Lebesgue measure $\mu \in \mathcal{M}(\mathbb{R}^n)$ is **translation invariant**:

$$\mu(A) = \mu(v + A) \qquad \text{for all } v \in \mathbb{R}^n, A \in \mathcal{B}(\mathbb{R}^n).$$

- In terms of the translation map $T_v \colon \mathbb{R}^n \to \mathbb{R}^n$, $T_v(x) := v + x$, $(T_v)_\sharp \mu = \mu$.
- We do not expect $(T_v)_\sharp \mu = \mu$ for a Gaussian measure $\mu = \mathcal{N}(m, C)$, which obviously has more mass near m than far away from m.
- What is true is that $(T_v)_\sharp \mathcal{N}(m, C) = \mathcal{N}(m + v, C)$ is equivalent to $\mathcal{N}(m, C)$ (i.e. each has a density with respect to the other):

Exercise 18

Let $\mu = \mathcal{N}(m, C) \in \mathcal{P}(\mathbb{R}^n)$. Show that

$$\frac{\mathrm{d}(T_v)_\sharp \mu}{\mathrm{d}\mu}(x) = \exp\left(\langle v, x - m \rangle_C - \frac{\|v\|_C^2}{2} \right).$$

There are two important Hilbert spaces associated to a Gaussian measure μ:

Definition 19

The **reproducing kernel Hilbert space** of $\mu = \mathcal{N}(m, C)$ is the Hilbert space

$$\mathcal{U}'_\mu := \overline{\{\ell - \langle \ell | m \rangle \mid \ell \in \mathcal{U}'\}} \text{ w.r.t. } \langle \cdot, \cdot \rangle_{L^2(\mu)},$$

and we extend $C_\mu \colon \mathcal{U}' \to \mathcal{U}''$ to $C_\mu \colon \mathcal{U}'_\mu \to \mathcal{U}''$ by

$$\langle C_\mu \xi | \ell \rangle := \int_{\mathcal{U}} \xi(x) \langle \ell | x - m \rangle \, \mu(\mathrm{d}x).$$

The **Cameron–Martin space** of μ is the Hilbert (!) space

$$\mathcal{H}_\mu := \{h \in \mathcal{U} \mid \|h\|_{\mathcal{H}_\mu} < \infty\}, \qquad \|h\|_{\mathcal{H}_\mu} := \sup\{\langle \ell | h \rangle \mid \ell \in \mathcal{U}', \langle C_\mu \ell | \ell \rangle \leq 1\}.$$

N.B. Each $\xi \in \mathcal{U}'_\mu$ is a $\mathbb{R}$-valued Gaussian, $\xi \sim \mathcal{N}(0, \|\xi\|^2_{L^2(\mu)})$.

The translation formula for finite-dimensional Gaussian measures holds in general, but only for *some* translation vectors v, by the **Cameron–Martin theorem**:

Theorem 20 (e.g. Bogachev, 1998, Sections 2.3 and 2.4)

Let $\mu = \mathcal{N}(m, C)$ on a separable Banach space $\mathcal{U}$.

- $\mathcal{H}_\mu = \mathcal{R}(C \colon \mathcal{U}'_\mu \to \mathcal{U}) = \overline{\mathcal{R}(C \colon \mathcal{U}' \to \mathcal{U})}$ *with respect to* $\langle Ck, C\ell \rangle_\mu := \langle k, \ell \rangle_{L^2(\mu)}$.
- *If $\mathcal{U}$ is Hilbert, then* $\mathcal{H}_\mu = \overline{\mathcal{R}(C^{1/2} \colon \mathcal{U} \to \mathcal{U})}$ *w.r.t.* $\langle h, k \rangle_C := \langle C^{-1/2} h, C^{-1/2} k \rangle_{\mathcal{U}}$.
- $\mathcal{H}_\mu = \{ v \in \mathcal{U} \mid (T_v)_* \mu := \mathcal{N}(m + v, C)$ *is equivalent to* $\mu \}$, *with*

$$\frac{\mathrm{d}(T_v)_\sharp \mu}{\mathrm{d}\mu}(x) = \exp\left(\langle C^{-1} v | x - m \rangle - \frac{\langle C^{-1} v | v \rangle}{2} \right) = \exp\left(\langle v, x - m \rangle_C - \frac{\|v\|_C^2}{2} \right).$$

- *For $m = 0$, $\mathcal{H}_\mu$ is the intersection of all linear subspaces of $\mathcal{U}$ that have full μ-measure. However, if $\dim \mathcal{U}'_\mu = \infty$, then $\mu(\mathcal{H}_\mu) = 0$.*

Support of Gaussian measures

As with the series examples earlier, it is usual relatively easy to show that $\mu(E) = 1$ for the right set $E \subseteq \mathcal{U}$. It is often harder to find the "smallest" such E, the **support** of μ,

$$\operatorname{supp}(\mu) := \bigcap_{\substack{E \subseteq \mathcal{U} \text{ closed} \\ \mu(E)=1}} E = \{ x \in \mathcal{U} \mid \text{for all } r > 0,\, \mu(B(x,r)) > 0 \}.$$

Fortunately, for Gaussian measures, the support is easy to describe:

Theorem 21 (e.g. Bogachev, 1998, Theorems 3.2.3 and 3.6.1)

Let $\mu = \mathcal{N}(m_\mu, C_\mu)$ on a separable Banach space $\mathcal{U}$. Then $\mathcal{R}(C_\mu \colon \mathcal{U}'_\mu \to \mathcal{U}'') \subseteq \mathcal{U}$ in the sense of the usual embedding $\mathcal{U} \hookrightarrow \mathcal{U}''$ and

$$\operatorname{supp}(\mu) = m_\mu + \overline{\mathcal{H}_\mu} = m_\mu + \overline{\mathcal{R}(C_\mu)},$$

where the closure is taken in the norm of $\mathcal{U}$, not the norm of $\mathcal{H}_\mu$.

Equivalence and singularity of Gaussian measures

Recall that $\mu, \nu \in \mathcal{P}(\mathcal{U})$ are **equivalent** if each has a density with respect to the other, and are **mutually singular** if there is a set E with $\mu(E) = 0$ but $\nu(E) = 1$.

Theorem 22 (Feldman, 1958; Hájek, 1958)

Let $\mu, \nu \in \mathcal{P}(\mathcal{U})$ be Gaussian measures on a normed space $\mathcal{U}$. Then either μ and ν are equivalent, or they are mutually singular.

Example 23

- Translating a Gaussian measure μ by a vector not in its Cameron–Martin space $\mathcal{H}_\mu$ induces mutual singularity.
- Dilating a Gaussian measure μ by any non-unit constant induces mutual singularity.
- Non-uniform dilations, e.g. when $(C_\mu^{-1/2} C_\nu^{1/2})(C_\mu^{-1/2} C_\nu^{1/2})^* - I$ is a Hilbert–Schmidt operator, can preserve equivalence. (Similar properties hold for Cauchy measures; Lie and Sullivan, 2018.)

Probability on function spaces

Gaussian processes

Definition 24

A family $\xi = (\xi_t)_{t\in T}$ of random variables is a **Gaussian process** on an index set T if every finite linear combination of the components ξ_t is Gaussian on $\mathbb{R}$. Set

$$m(t) := \mathbb{E}[\xi_t], \qquad c(s,t) := \mathbb{E}\big[(\xi_s - \mathbb{E}[\xi_s])(\xi_t - \mathbb{E}[\xi_t])\big].$$

Theorem 25 (e.g. Bogachev, 1998, Prop. 2.3.9)

The law μ_ξ of a Gaussian process $\xi = (\xi_t)_{t\in T}$ is a Gaussian measure on $\mathcal{U} = \mathbb{R}^T$ with the topology of pointwise convergence, with some mean $m \in \mathcal{U}$ and $C_\mu(\mathcal{U}'_\mu) \subset \mathcal{U}$. Thought of as a bilinear form on $\mathcal{U}'$, its covariance operator is determined by

$$C_\mu(\delta_s, \delta_t) = c(s,t),$$

where $\langle \delta_s | u \rangle := u(s)$. A choice of $c\colon T^2 \to \mathbb{R}$ generates a centred Gaussian process / measure if and only if $(c(s_i,t_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$ is non-negative for all $s_1,\dots,s_n,t_1,\dots,t_n \in T$.

Example 26

Gaussian white noise on T corresponds to $c(s,t) = \mathbb{1}_{s=t}$

Example 27 (e.g. Bogachev, 1998, Example 2.3.11 and Remark 2.3.13)

The **Wiener process** or **Brownian motion** on $T = [0,1]$ corresponds to $c(s,t) = \min(s,t)$. This Gaussian measure is supported on $C^0([0,1];\mathbb{R})$ and has $H^1([0,1];\mathbb{R})$ as its Cameron–Martin space. This measure is also the law of

$$\boldsymbol{u}(t) := \sum_{n \in \mathbb{N}} \boldsymbol{\eta}_n \int_0^t \psi_n(s) \, \mathrm{d}s,$$

with $(\psi_n)_{n\in\mathbb{N}}$ any orthonormal basis of $L^2([0,T];\mathbb{R})$, and $\boldsymbol{\eta}_n \sim \mathcal{N}(0,1)$ i.i.d.

See the lectures of **Matt Dunlop** for many more examples and applications of Gaussian processes, especially compositions of such objects.

Probability on function spaces

Conditional probability

Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. How do we condition (update) one random variable $\boldsymbol{x} \colon \Omega \to \mathcal{X}$ given knowledge of another (hopefully related) one $\boldsymbol{y} \colon \Omega \to \mathcal{Y}$?

- Let $\mu_{\boldsymbol{x}} := \boldsymbol{x}_\sharp \mathbb{P} \in \mathcal{P}(\mathcal{X})$ (resp. $\mu_{\boldsymbol{y}} \in \mathcal{P}(\mathcal{Y})$) be the law of $\boldsymbol{x}$ (resp. $\boldsymbol{y}$).
- Let $\sigma(\boldsymbol{y}) \subseteq \mathcal{F}$ denote the σ-algebra generated by $\boldsymbol{y}$, i.e. the coarsest σ-algebra on Ω so that $\boldsymbol{y}$ is $(\sigma(\boldsymbol{y}), \mathcal{B}(\mathcal{Y}))$-measurable:

$$\sigma(\boldsymbol{y}) := \{\boldsymbol{y}^{-1}(E) \mid E \in \mathcal{B}(\mathcal{Y})\}.$$

Definition 28

For two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ with $\mathbb{E}[\|\boldsymbol{x}\|] < \infty$, the **conditional expectation** of $\boldsymbol{x}$ given $\boldsymbol{y}$ is a ($\mathbb{P}$-a.s. unique) random variable $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}] \colon \Omega \to \mathcal{X}$ such that

- $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}]$ is $\sigma(\boldsymbol{y})$-measurable; and
- for all $E \in \sigma(\boldsymbol{y})$,

$$\int_E \boldsymbol{x}(\omega)\, \mathbb{P}(\mathrm{d}\omega) = \int_E \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}](\omega)\, \mathbb{P}(\mathrm{d}\omega).$$

Exercise 29 (A discrete example)

A fair cubical die has sides 1, 2, and 3 coloured red, 4 and 5 coloured green, and 6 coloured blue. Let x be the numerical outcome of the die roll, and y the colour outcome.

Ω	$x(\omega)$	$y(\omega)$	$\mathbb{E}[x\|y](\omega)$
$\{\omega \in \Omega \mid x(\omega) = 1\}$	1	r	???
$\{\omega \in \Omega \mid x(\omega) = 2\}$	2	r	???
$\{\omega \in \Omega \mid x(\omega) = 2\}$	3	r	???
$\{\omega \in \Omega \mid x(\omega) = 3\}$	4	g	???
$\{\omega \in \Omega \mid x(\omega) = 4\}$	5	g	???
$\{\omega \in \Omega \mid x(\omega) = 5\}$	6	b	???

Complete the last column. Hint: you only need to consider $E = y^{-1}(r), y^{-1}(g), y^{-1}(b)$; $\sigma(y)$-measurability means that $\mathbb{E}[x|y]$ must be constant on each of these three sets, and the corresponding constant value must be...

Conditional probability

Definition 30

For two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ with $\mathbb{E}[\|\boldsymbol{x}\|] < \infty$, the **conditional probability** $\mathbb{P}[\boldsymbol{x} \in A|\boldsymbol{y}] : \Omega \to [0,1]$ that $\boldsymbol{x}$ lies in $A \in \mathcal{B}(\mathcal{X})$ given $\boldsymbol{y}$ is the random variable

$$\mathbb{P}[\boldsymbol{x} \in A|\boldsymbol{y}] := \mathbb{E}[\mathbb{1}_{\boldsymbol{x} \in A}|\boldsymbol{y}]$$

By the Doob–Dynkin lemma (e.g. Kallenberg (1997, Lemma 1.13)), there is a measurable function $\psi = \psi_{\boldsymbol{x},\boldsymbol{y}} : \mathcal{Y} \to \mathcal{X}$ such that $\psi(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}]$ $\mathbb{P}$-a.s. We consider

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y} = y] := \psi(y)$$

to be the conditional expectation of $\boldsymbol{x}$ given the (possibly probability zero) event $\boldsymbol{y} = y$. Similarly, we consider

$$\mathbb{P}[\boldsymbol{x} \in A|\boldsymbol{y} = y] := \mathbb{E}[\mathbb{1}_{\boldsymbol{x} \in A}|\boldsymbol{y} = y].$$

to be the conditional probability that $\boldsymbol{x} \in A$ given $\boldsymbol{y} = y$.

$$\begin{tikzcd}
\Omega \arrow[r, "\boldsymbol{x}"] \arrow[d, "\boldsymbol{y}"'] \arrow[dr, "{\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}]}"] & \mathcal{X} \\
\mathcal{Y} \arrow[r, "\psi"'] & \mathcal{X}
\end{tikzcd}$$

Definition 31

For two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$, a **regular conditional distribution** of $\boldsymbol{x}$ given $\boldsymbol{y}$ is a function $\mu_{\boldsymbol{x}|\boldsymbol{y}} \colon \mathcal{Y} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ such that

- for all $y \in \mathcal{Y}$, $\mu_{\boldsymbol{x}|\boldsymbol{y}}(y, \cdot\,) \in \mathcal{P}(\mathcal{X})$;
- for all $A \in \mathcal{B}(\mathcal{X})$, $\mu_{\boldsymbol{x}|\boldsymbol{y}}(\,\cdot\,, A) \colon \mathcal{Y} \to [0,1]$ is measurable;
- for all $A \in \mathcal{B}(\mathcal{X})$, as random variables $\Omega \to [0,1]$,

$$\mu_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{y}, A) = \mathbb{P}[\boldsymbol{x} \in A | \boldsymbol{y}] \quad \mathbb{P}\text{-a.s.}.$$

A closely related concept is the **disintegration** of $\nu \in \mathcal{P}(\mathcal{Z})$ with respect to $\pi \colon \mathcal{Z} \to \mathcal{Y}$, which amounts to conditioning ν onto the fibres $\pi^{-1}(y)$ — we happen to only care about horizontal slices of the form $\mathcal{X} \times \{y\}$ in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ (Chang and Pollard, 1997).

Example 32

Take $\mathcal{X} = \mathcal{Y} = [0, 1]$ and let $(\boldsymbol{x}, \boldsymbol{y})$ be uniformly distributed in the square $\mathcal{X} \times \mathcal{Y}$.

- One regular conditional distribution for $\boldsymbol{x}|\boldsymbol{y}$ is the obvious one:

$$\mu_{\boldsymbol{x}|\boldsymbol{y}}(y, \cdot) := \mathrm{Unif}(\mathcal{X}) \text{ for all } y \in \mathcal{Y}.$$

  This is common sense: regardless of the value of $\boldsymbol{y}$, $\boldsymbol{x}$ should be uniformly distributed.

- However, the following is *also* a regular conditional distribution for $\boldsymbol{x}|\boldsymbol{y}$:

$$\nu_{\boldsymbol{x}|\boldsymbol{y}}(y, \cdot) := \begin{cases} \mathrm{Unif}(\mathcal{X}) & \text{for all } y \in \mathcal{Y}, y \neq \frac{1}{2}, \\ \mathrm{Unif}([\frac{1}{4}, \frac{1}{2}]) & \text{for all } y = \frac{1}{2}. \end{cases}$$

  This looks rather strange: $\boldsymbol{x}$ is no longer uniform on $\mathcal{X}$ if $\boldsymbol{y}$ happens to take the value $y = \frac{1}{2}$. But this is fine: the exceptional set of y's has $\mu_{\boldsymbol{y}}$-measure zero.

Thus, a regular conditional distribution for $\boldsymbol{x}|\boldsymbol{y}$ is a measurable family of probability measures $\mu_{\boldsymbol{x}|\boldsymbol{y}}(y, \cdot) \in \mathcal{P}(\mathcal{X})$, parametrised by $y \in \mathcal{Y}$, satisfying

$$\begin{aligned} \mu_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{y}, A) &= \mathbb{P}[\boldsymbol{x} \in A | \boldsymbol{y}] \quad \mathbb{P}\text{-a.s.} \\ \text{i.e.} \quad \mu_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{y}, A) &= \mathbb{E}[\mathbb{1}_{\boldsymbol{x} \in A} | \boldsymbol{y}] \quad \mathbb{P}\text{-a.s.} \end{aligned}$$

In practice, the condition that we check is an integral formulation of the above: that, for all $E \in \sigma(\boldsymbol{y}) \subseteq \mathcal{F}$ of the form $E = \boldsymbol{y}^{-1}(B)$, $B \in \mathcal{B}(\mathcal{Y})$,

$$\begin{aligned} \int_E \mu_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{y}(\omega), A) \, \mathbb{P}(\mathrm{d}\omega) &= \int_E \mathbb{E}[\mathbb{1}_{\boldsymbol{x} \in A} | \boldsymbol{y}](\omega) \, \mathbb{P}(\mathrm{d}\omega) \\ \text{i.e.} \quad \int_B \mu_{\boldsymbol{x}|\boldsymbol{y}}(y, A) \, \mu_{\boldsymbol{y}}(\mathrm{d}y) &= \int_{\boldsymbol{y}^{-1}(B)} \mathbb{1}_{\boldsymbol{x} \in A}(\omega) \, \mathbb{P}(\mathrm{d}\omega). \end{aligned}$$

Here the LHS is rewritten using the change-of-variables formula for integration, and the RHS is rewritten using the defining relation for conditional expectation.

The existence and essential uniqueness of regular conditional probabilities / disintegrations is very technical to show but holds under general conditions that are certainly enough for our purposes:

Theorem 33 (e.g. Kallenberg, 1997, Theorem 5.3)

When $\mathcal{X}$ and $\mathcal{Y}$ are separable Banach spaces — or even just separable and completely metrisable spaces — a regular conditional distribution $\mu_{\boldsymbol{x}|\boldsymbol{y}}$ always exists.

Furthermore, $\mu_{\boldsymbol{x}|\boldsymbol{y}}$ is essentially unique in the sense that if $\nu_{\boldsymbol{x}|\boldsymbol{y}}$ also satisfies the requirements to be a regular conditional distribution for $\boldsymbol{x}|\boldsymbol{y}$, then

$$\begin{aligned} & \mu_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{y}, \cdot\,) = \nu_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{y}, \cdot\,) \quad \mathbb{P}\text{-a.s.,} \\ \text{i.e.} \quad & \mu_{\boldsymbol{x}|\boldsymbol{y}}(y, \cdot\,) = \nu_{\boldsymbol{x}|\boldsymbol{y}}(y, \cdot\,) \quad \text{for } \mu_{\boldsymbol{y}}\text{-a.e. } y \in \mathcal{Y}. \end{aligned}$$

The language of regular conditional probability can be used to give a rigorous meaning to Bayes' formula, *without* needing density functions. Here we consider the usual introductory case of a nonlinear forward operator corrupted by centred, additive, finite-dimensional Gaussian noise:

Theorem 34 (RCP Bayes' formula)

Let $\mathcal{X}$ and $\mathcal{Y}$ be separable Banach spaces with $\dim \mathcal{Y} < \infty$. *For* $\boldsymbol{x} \sim \mu_{\boldsymbol{x}} \in \mathcal{P}(\mathcal{X})$ *and* $\boldsymbol{y} = G(\boldsymbol{x}) + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Gamma)$ *independently of* $\boldsymbol{x}$, *the distribution*

$$\mu_{\boldsymbol{x}|\boldsymbol{y}}(y, \mathrm{d}x) := \frac{\exp(-\frac{1}{2}\|y - G(x)\|_{\Gamma}^{2})}{Z(y)} \mu_{\boldsymbol{x}}(\mathrm{d}x),$$

$$Z(y) := \int_{\mathcal{X}} \exp(-\tfrac{1}{2}\|y - G(x)\|_{\Gamma}^{2}) \, \mu_{\boldsymbol{x}}(\mathrm{d}x)$$

is a regular conditional distribution of $\boldsymbol{x}$ *given* $\boldsymbol{y}$.

Proof of the RCP Bayes' formula I

Since $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Gamma)$, the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x} = x$ is absolutely continuous with respect to Lebesgue measure $\mathrm{d}y$ on $\mathcal{Y}$ with the following form:

$$\mu_{\boldsymbol{y}|\boldsymbol{x}=x}(\mathrm{d}y) = \frac{\exp(-\frac{1}{2}\|y - G(x)\|_\Gamma^2)}{\sqrt{\det(2\pi\Gamma)}}\,\mathrm{d}y$$

Thus, the joint distribution $\mu_{(\boldsymbol{x},\boldsymbol{y})} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ of $(\boldsymbol{x}, \boldsymbol{y})$ is determined by its values on rectangles $A \times B$, $A \in \mathcal{B}(\mathcal{X})$, $B \in \mathcal{B}(\mathcal{Y})$, as

$$\mathbb{P}[\boldsymbol{x} \in A, \boldsymbol{y} \in B] = \mu_{(\boldsymbol{x},\boldsymbol{y})}(A \times B) = \int_A \int_B \frac{\exp(-\frac{1}{2}\|y - G(x)\|_\Gamma^2)}{\sqrt{\det(2\pi\Gamma)}}\,\mathrm{d}y\,\mu_{\boldsymbol{x}}(\mathrm{d}x).$$

In particular, taking $A = \mathcal{U}$, we obtain the marginal distribution $\mu_{\boldsymbol{y}}$ of $\boldsymbol{y}$:

$$\mathbb{P}[\boldsymbol{y} \in B] = \mu_{\boldsymbol{y}}(B) = \int_B \frac{Z(y)}{\sqrt{\det(2\pi\Gamma)}}\,\mathrm{d}y.$$

Proof of the RCP Bayes' formula II

As an exercise, check that $Z(y) > 0$ for all $y \in \mathcal{Y}$. (Hint: you can even prove a positive lower bound that depends only on $\|y\|$, Γ, G, and $\mu_{\boldsymbol{x}}$.)

Hence, for each $y \in \mathcal{Y}$, $\mu_{\boldsymbol{x}|\boldsymbol{y}}(y, \cdot\,)$ is a well-defined probability measure on $\mathcal{X}$, as it is a re-weighting of $\mu_{\boldsymbol{x}} \in \mathcal{P}(\mathcal{X})$ by a non-negative, integrable function with unit integral against $\mu_{\boldsymbol{x}}$.

The measurability of $y \mapsto \mu_{\boldsymbol{x}|\boldsymbol{y}}(y, A)$ for all $A \in \mathcal{B}(\mathcal{X})$ is rather techical and not the main point of the proof, so we omit it.

It remains to show that, for all $A \in \mathcal{B}(\mathcal{X})$, $\mu_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{y}, A) = \mathbb{P}[\boldsymbol{x} \in A | \boldsymbol{y}]$ $\mathbb{P}$-a.s.; as discussed earlier, it is enough to show that, for all $B \in \mathcal{B}(\mathcal{Y})$,

$$\int_B \mu_{\boldsymbol{x}|\boldsymbol{y}}(y, A)\, \mu_{\boldsymbol{y}}(\mathrm{d}y) = \int_{\boldsymbol{y}^{-1}(B)} \mathbb{1}_{\boldsymbol{x} \in A}(\omega)\, \mathbb{P}(\mathrm{d}\omega)$$

Proof of the RCP Bayes' formula III

Again writing $\mu_{(\boldsymbol{x},\boldsymbol{y})}$ for the joint distribution of $(\boldsymbol{x},\boldsymbol{y})$,

$$
\begin{aligned}
\int_{\boldsymbol{y}^{-1}(B)} \mathbb{1}_{\boldsymbol{x}\in A}(\omega)\,\mathbb{P}(\mathrm{d}\omega) &= \mathbb{P}[\boldsymbol{x}\in A, \boldsymbol{y}\in B] = \mu_{(\boldsymbol{x},\boldsymbol{y})}(A\times B) \\
&= \int_A \left(\int_B \frac{\exp(-\frac{1}{2}\|y-G(x)\|_\Gamma^2)}{\sqrt{\det(2\pi\Gamma)}}\,\mathrm{d}y \right) \mu_{\boldsymbol{x}}(\mathrm{d}x) \\
&= \int_B \left(\int_A \frac{\exp(-\frac{1}{2}\|y-G(x)\|_\Gamma^2)}{\sqrt{\det(2\pi\Gamma)}}\,\mu_{\boldsymbol{x}}(\mathrm{d}x) \right) \mathrm{d}y && \text{(Fubini)} \\
&= \int_B \left(\int_A \mu_{\boldsymbol{x}|\boldsymbol{y}}(y,\mathrm{d}x) \right) \frac{Z(y)}{\sqrt{\det(2\pi\Gamma)}}\,\mathrm{d}y && \text{(definition of } Z(y)\text{)} \\
&= \int_B \mu_{\boldsymbol{x}|\boldsymbol{y}}(y,A)\,\mu_{\boldsymbol{y}}(\mathrm{d}y), && \text{(formula for } \mu_{\boldsymbol{y}}\text{)}
\end{aligned}
$$

as required. □

Theorem 34 (RCP Bayes' formula)

Let $\mathcal{X}$ and $\mathcal{Y}$ be separable Banach spaces with $\dim \mathcal{Y} < \infty$. *For* $\boldsymbol{x} \sim \mu_{\boldsymbol{x}} \in \mathcal{P}(\mathcal{X})$ *and* $\boldsymbol{y} = G(\boldsymbol{x}) + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Gamma)$ *independently of* $\boldsymbol{x}$*, the distribution*

$$\mu_{\boldsymbol{x}|\boldsymbol{y}}(y, \mathrm{d}x) := \frac{\exp(-\frac{1}{2}\|y - G(x)\|_{\Gamma}^{2})}{Z(y)} \mu_{\boldsymbol{x}}(\mathrm{d}x), \quad Z(y) := \int_{\mathcal{X}} \exp(-\tfrac{1}{2}\|y - G(x)\|_{\Gamma}^{2})\, \mu_{\boldsymbol{x}}(\mathrm{d}x)$$

is a regular conditional distribution of $\boldsymbol{x}$ *given* $\boldsymbol{y}$.

A Bayesian's solution to an inverse problem is a "posterior distribution"; evaluating the RCP Bayes' rule at the observed data y gives rigorous meaning to this "posterior":

Definition 35

When $\mu = \mu_{\boldsymbol{x}}$ is the prior distribution of $\boldsymbol{x}$, we define $\mu^{y} := \mu_{\boldsymbol{x}|\boldsymbol{y}}(y, \cdot\,)$ to be the **Bayesian posterior distribution** for $\boldsymbol{x}$ given the observation $\boldsymbol{y} = y$.

Expressions for Bayes' formula

The above expression for the posterior distribution,

$$\mu^y(\mathrm{d}x) := \frac{\exp(-\frac{1}{2}\|y - G(x)\|_\Gamma^2)}{Z(y)} \mu(\mathrm{d}x),$$

$$Z(y) := \int_{\mathcal{X}} \exp(-\tfrac{1}{2}\|y - G(x)\|_\Gamma^2)\, \mu(\mathrm{d}x)$$

can be rearranged to give the density of μ^y with respect to the prior μ:

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu}(x) = \frac{\exp(-\frac{1}{2}\|y - G(x)\|_\Gamma^2)}{Z(y)}.$$

In the case that $\dim \mathcal{X} < \infty$ and $\mu(\mathrm{d}x) = \rho(x)\,\mathrm{d}x$, we get the familiar Bayesian formula for the posterior density ρ^y:

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu}(x) = \frac{\mathrm{d}\mu^y}{\mathrm{d}x}(x) \Big/ \frac{\mathrm{d}\mu}{\mathrm{d}x}(x) = \frac{\exp(-\frac{1}{2}\|y - G(x)\|_\Gamma^2)}{Z(y)}$$

$$\text{i.e.} \quad \rho^y(x) = \frac{\exp(-\frac{1}{2}\|y - G(x)\|_\Gamma^2)}{Z(y)} \rho(x).$$

Theorem 36 (Anderson and Trapp, 1975)

Let $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2) \sim \mathcal{N}(m, C)$ *on a direct sum* $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ *of separable Hilbert spaces, with mean* $m = (m_1, m_2)$ *and positive-definite covariance operator* C*. For* $i, j = 1, 2$*, let*

$$C_{ij} \colon \mathcal{H}_j \to \mathcal{H}_i, \qquad \langle C_{ij} k_j, \ell_i \rangle_{\mathcal{H}_i} := \mathbb{E}\Big[\langle k_j, \boldsymbol{x}_j - m_j \rangle_{\mathcal{H}_j} \langle \ell_i, \boldsymbol{x}_i - m_i \rangle_{\mathcal{H}_i} \Big]$$

for all $k_i, \ell_i \in \mathcal{H}_i$*,* $k_j, \ell_j \in \mathcal{H}_j$*, so that* $C \colon \mathcal{H} \to \mathcal{H}$ *is decomposed in block form as*

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix};$$

in particular, $\boldsymbol{x}_i \sim \mathcal{N}(m_i, C_{ii})$*, and* $C_{21} = C_{12}^*$*. Then* C_{22} *is invertible and, for each* $x_2 \in \mathcal{H}_2$*,*

$$(\boldsymbol{x}_1 | \boldsymbol{x}_2 = x_2) \sim \mathcal{N}\Big(m_1 + C_{12} C_{22}^{-1} (x_2 - m_2), \underbrace{C_{11} - C_{12} C_{22}^{-1} C_{21}}_{\substack{\text{shorted operator /} \\ \text{Schur complement}}} \Big).$$

Linear Gaussian inverse problems

Let G be a continuous linear operator, $\boldsymbol{u} \sim \mu = \mathcal{N}(m_0, C_0)$, and suppose that we observe $\boldsymbol{y} = G\boldsymbol{u} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Gamma)$ independent of $\boldsymbol{u}$. Thus,

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m_0 \\ Gm_0 \end{bmatrix}, \begin{bmatrix} C_0 & C_0G^* \\ GC_0 & \Gamma + GC_0G^* \end{bmatrix} \right).$$

The Gaussian conditioning theorem implies that $(\boldsymbol{u}|\boldsymbol{y} = y) \sim \mathcal{N}(m_1, C_1)$, where

$$\begin{aligned} m_1 &= m_0 + C_0G^*(\Gamma + GC_0G^*)^{-1}(y - Gm_0) = m_0 + C_1G^*\Gamma^{-1}(y - Gm_0), \\ C_1 &= C_0 - C_0G^*(\Gamma + GC_0G^*)^{-1}GC_0. \end{aligned}$$

The repeated sequential application of this formula (with the prior μ being the evolution of a previous state estimate under some dynamical system) underlies data assimilation techniques like the Kálmán filter (Law et al., 2015; Reich and Cotter, 2015)... see also the lectures of **Simo Särkkä**.

Well-posedness of BIPs

In this section we

- quantify the distance between probability measures using the Hellinger metric;
- study the stability of Bayesian inverse problems (BIPs) with forward equation

$$\boldsymbol{y} = G(\boldsymbol{u}) + \boldsymbol{\varepsilon}$$

  with $\boldsymbol{\varepsilon}$ finite-dimensional and Gaussian;
- discuss how to extend the analysis to more general additive problems.

Well-posedness of BIPs

Quantifying posterior stability

Definition 37

The **Hellinger distance** between $\mu, \nu \in \mathcal{P}(\mathcal{U})$, both absolutely continuous with respect to $\pi \in \mathcal{M}(\mathcal{U})$, is

$$d_{\mathrm{H}}(\mu,\nu)^2 := \frac{1}{2}\int_{\mathcal{U}} \left| \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}} - \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\pi}} \right|^2 \mathrm{d}\pi = \frac{1}{2}\left\| \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}} - \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\pi}} \right\|_{L^2(\pi)}^2 .$$

Exercise 38

Show that d_{H} defines a metric on $\mathcal{P}(\mathcal{U})$ that is independent of the choice of π, and that

$$d_{\mathrm{H}}(\mu,\nu)^2 = 1 - \int_{\mathcal{U}} \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\frac{\mathrm{d}\nu}{\mathrm{d}\pi}}\,\mathrm{d}\pi = 1 - \mathbb{E}_{\nu}\left[\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}}\right]$$

whenever the relevant densities exist.

Lemma 39 (Kraft, 1955)

The Hellinger and total variation distances are topologically equivalent:

$$d_{\mathrm{H}}(\mu, \nu)^2 \leq d_{\mathrm{TV}}(\mu, \nu) \leq \sqrt{2}\, d_{\mathrm{H}}(\mu, \nu),$$

where

$$d_{\mathrm{TV}}(\mu, \nu) := \sup_{E \in \mathcal{B}(\mathcal{U})} |\mu(E) - \nu(E)| = \frac{1}{2} \int_{\mathcal{U}} \left| \frac{\mathrm{d}\mu}{\mathrm{d}\pi} - \frac{\mathrm{d}\nu}{\mathrm{d}\pi} \right| \mathrm{d}\pi.$$

One can also show by Pinsker's inequality that convergence in d_{H} or d_{TV} is implied by, but is generally strictly weaker than, convergence with respect to the Kullback–Leibler divergence (relative entropy)

$$D_{\mathrm{KL}}(\mu \| \nu) := \int_{\mathcal{U}} \left(\log \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \right) \mathrm{d}\mu.$$

For $a, b \geq 0$,

$$|a-b|^2 = |a-b||a-b| \leq |a-b||a+b| = |a^2-b^2|.$$

Hence,

$$d_{\mathrm{H}}(\mu,\nu)^2 := \frac{1}{2}\int_{\mathcal{U}} \left| \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}} - \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\pi}} \right|^2 \mathrm{d}\pi \leq \frac{1}{2}\int_{\mathcal{U}} \left| \frac{\mathrm{d}\mu}{\mathrm{d}\pi} - \frac{\mathrm{d}\nu}{\mathrm{d}\pi} \right| \mathrm{d}\pi =: d_{\mathrm{TV}}(\mu,\nu).$$

In the other direction, since $a^2 - b^2 = (a+b)(a-b)$,

$$\begin{aligned} d_{\mathrm{TV}}(\mu,\nu) &= \frac{1}{2}\int_{\mathcal{U}} \left| \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}} + \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\pi}} \right| \left| \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}} - \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\pi}} \right| \mathrm{d}\pi \\ &\leq \frac{1}{2}\sqrt{\int_{\mathcal{U}} \left| \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}} + \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\pi}} \right|^2 \mathrm{d}\pi} \sqrt{\int_{\mathcal{U}} \left| \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}} - \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\pi}} \right|^2 \mathrm{d}\pi} \\ &\leq \frac{1}{2} \cdot \sqrt{4} \cdot \sqrt{2 d_{\mathrm{H}}(\mu,\nu)^2} \end{aligned}$$

Exercise 40

Show that, for $f \in L^2(\mathcal{U}, \mu) \cap L^2(\mathcal{U}, \nu)$,

$$\left|\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]\right| \leq 2\sqrt{\mathbb{E}_\mu[f^2] + \mathbb{E}_\nu[f^2]}\, d_{\mathrm{H}}(\mu, \nu).$$

Exercise 41

Show that the Hellinger distance between non-degenerate Gaussian measures $\mu_0 = \mathcal{N}(m_0, C_0)$ and $\mu_1 = \mathcal{N}(m_1, C_1)$ on $\mathbb{R}^n$ is

$$d_{\mathrm{H}}(\mu_0, \mu_1)^2 = 1 - \frac{\det C_0^{1/4} \det C_1^{1/4}}{\det C_{1/2}^{1/2}} \exp\left(-\frac{1}{8}\|m_0 - m_1\|^2_{C_{1/2}}\right),$$

where $C_{1/2} := \frac{1}{2}(C_0 + C_1)$.

Harder: try the case of two Gaussians on a separable Hilbert space $\mathcal{H}$.

The following crucial lemma tells us how close two re-weightings μ_1 and μ_2 of a common reference probability measure ν by two potentials Φ_1 and Φ_2 are in the Hellinger sense:

Lemma 42

Let $\nu \in \mathcal{P}(\mathcal{U})$ and $\Phi_i \in L^2(\mathcal{U}, \nu; [0, \infty))$ for $i = 1, 2$. Define $\mu_i \in \mathcal{P}(\mathcal{U})$ by

$$\mu_i(\mathrm{d}u) := \frac{1}{Z_i} \exp(-\Phi_i(u))\, \nu(\mathrm{d}u), \qquad Z_i := \int_{\mathcal{U}} \exp(-\Phi_i(u))\, \nu(\mathrm{d}u).$$

Then

$$\begin{aligned} |Z_1 - Z_2| &\leq \|\Phi_1 - \Phi_2\|_{L^2(\nu)}, \\ d_{\mathrm{H}}(\mu_1, \mu_2) &\leq \frac{1}{\sqrt{2}\min(Z_1, Z_2)} \|\Phi_1 - \Phi_2\|_{L^2(\nu)}. \end{aligned}$$

From the definitions of μ_1, μ_2, and d_{H}, we have

$$
\begin{aligned}
d_{\mathrm{H}}(\mu_1,\mu_2)^2 &= \frac{1}{2}\int_{\mathcal{U}}\left|\frac{e^{-\Phi_1/2}}{Z_1^{1/2}} - \frac{e^{-\Phi_2/2}}{Z_2^{1/2}}\right|^2 \mathrm{d}\nu \\
&\leq \int_{\mathcal{U}}\left|\frac{e^{-\Phi_1/2}}{Z_1^{1/2}} - \frac{e^{-\Phi_2/2}}{Z_1^{1/2}}\right|^2 + \left|\frac{e^{-\Phi_2/2}}{Z_1^{1/2}} - \frac{e^{-\Phi_2/2}}{Z_2^{1/2}}\right|^2 \mathrm{d}\nu \\
&= I_1 + I_2,
\end{aligned}
$$

where the inequality follows from $(a-c)^2 \leq 2(a-b)^2 + 2(b-c)^2$, and

$$
\begin{aligned}
I_1 &:= \frac{1}{Z_1}\int_{\mathcal{U}}\left|e^{-\Phi_1/2} - e^{-\Phi_2/2}\right|^2 \mathrm{d}\nu, \\
I_2 &:= Z_2\left|Z_1^{-1/2} - Z_2^{-1/2}\right|^2 = \frac{1}{Z_1}\left|Z_2^{1/2} - Z_1^{1/2}\right|^2.
\end{aligned}
$$

We now bound I_1, I_2, and $|Z_1 - Z_2|$.

Since, for $a, b \geq 0$, $|e^{-a} - e^{-b}| \leq |a - b|$,

$$I_1 := \frac{1}{Z_1} \int_{\mathcal{U}} \left| e^{-\Phi_1/2} - e^{-\Phi_2/2} \right|^2 \mathrm{d}\nu \leq \frac{1}{Z_1} \int_{\mathcal{U}} \frac{|\Phi_1 - \Phi_2|^2}{4} \mathrm{d}\nu = \frac{1}{4Z_1} \|\Phi_1 - \Phi_2\|_{L^2(\nu)}^2.$$

Since, for $a, b \geq 0$, $|a^{1/2} - b^{1/2}| \leq \frac{1}{2} \min(a, b)^{-1/2} |a - b|$,

$$I_2 := \frac{1}{Z_1} \left| Z_2^{1/2} - Z_1^{1/2} \right|^2 \leq \frac{1}{Z_1 \cdot 2^2 \cdot \min(Z_1, Z_2)} |Z_1 - Z_2|^2.$$

We bound the difference of normalising constants using Jensen's inequality:

$$|Z_1 - Z_2|^2 \leq \int_{\mathcal{U}} |e^{-\Phi_1} - e^{-\Phi_2}|^2 \,\mathrm{d}\nu \leq \int_{\mathcal{U}} |\Phi_1 - \Phi_2|^2 \,\mathrm{d}\nu = \|\Phi_1 - \Phi_2\|_{L^2(\nu)}^2.$$

Proof of Lemma 42 III

Now observe that, for $i = 1, 2$, since the potentials are non-negative,

$$Z_i := \int_{\mathcal{U}} e^{-\Phi_i} \, \mathrm{d}\nu \leq \int_{\mathcal{U}} 1 \, \mathrm{d}\nu = 1,$$

and so

$$\min(Z_1, Z_2)^2 \leq Z_i^2 \leq Z_i.$$

Hence, we have $d_{\mathrm{H}}(\mu_1, \mu_2)^2 \leq C \|\Phi_1 - \Phi_2\|_{L^2(\nu)}^2$ with

$$C = \frac{1}{4Z_1} + \frac{1}{4Z_1 \min(Z_1, Z_2)} \leq \frac{1}{4 \min(Z_1, Z_2)^2} + \frac{1}{4 \min(Z_1, Z_2)^2} = \frac{1}{2 \min(Z_1, Z_2)^2}$$

and taking square roots completes the proof. $\square$

Well-posedness of BIPs

Finite-dim. additive Gaussian noise

- $y = G(u) + \varepsilon$ with $G\colon \mathcal{U} \to \mathcal{Y}$, with $\dim \mathcal{Y} < \infty$.
- The noise $\varepsilon \sim \mathcal{N}(0, \Gamma)$ is independent of the prior μ on u, and Γ is invertible.
- The potential Φ is the weighted misfit

$$\Phi(u; y) = \frac{1}{2}\|y - G(u)\|_\Gamma^2 = \frac{1}{2}\|\Gamma^{-1/2}(y - G(u))\|^2.$$

- G is Lipschitz and in $L^4(\mathcal{U}, \mu; \mathcal{Y})$.
- The posterior of interest is $\mu^y \in \mathcal{P}(\mathcal{U})$,

$$\mu^y(\mathrm{d}u) := \frac{\exp(-\frac{1}{2}\|y - G(u)\|_\Gamma^2)}{Z}\,\mu(\mathrm{d}u), \quad Z := \int_{\mathcal{U}} \exp\Big(-\tfrac{1}{2}\|y - G(u)\|_\Gamma^2\Big)\,\mu(\mathrm{d}u).$$

The Lipschitz assumption on G ensures that the *forward* problem is well posed; the L^4 assumption comes from the quadratic misfit $\times$ the L^2 nature of Lemma 42.

The first stability result shows that the normalising constant and the posterior are locally Lipschitz functions of the observed data y:

Theorem 43

Under the usual assumptions, for every $r > 0$, there is a constant $K = K_{r,G,\Gamma,\mu}$ such that, whenever $\|y\|, \|\widetilde{y}\| \leq r$,

$$|Z - \widetilde{Z}| \leq K\|y - \widetilde{y}\|,$$
$$d_{\mathrm{H}}(\mu^{y}, \mu^{\widetilde{y}}) \leq K\|y - \widetilde{y}\|,$$

where

$$\mu^{\widetilde{y}}(\mathrm{d}u) := \frac{\exp(-\frac{1}{2}\|\widetilde{y} - G(u)\|_{\Gamma}^{2})}{\widetilde{Z}}\,\mu(\mathrm{d}u), \qquad \widetilde{Z} := \int_{\mathcal{U}} \exp\left(-\tfrac{1}{2}\|\widetilde{y} - G(u)\|_{\Gamma}^{2}\right)\mu(\mathrm{d}u).$$

We apply the Hellinger stability lemma with $\nu = \mu$, $\Phi_1(u) := \frac{1}{2}\|y - G(u)\|_\Gamma^2$, and $\Phi_2(u) := \frac{1}{2}\|\widetilde{y} - G(u)\|_\Gamma^2$. The square-integrability assumption on Φ_1 is satisfied, since

$$\begin{aligned}
\|\Phi_1\|_{L^2(\mu)}^2 &= \int_{\mathcal{U}} \left(\frac{1}{2}\|y - G(u)\|_\Gamma^2 \right)^2 \mu(\mathrm{d}u) \\
&\leq \int_{\mathcal{U}} \left(\|y\|_\Gamma^2 + \|G(u)\|_\Gamma^2 \right)^2 \mu(\mathrm{d}u) \\
&\leq 2 \int_{\mathcal{U}} \left(\|y\|_\Gamma^4 + \|G(u)\|_\Gamma^4 \right) \mu(\mathrm{d}u) \\
&\leq 2\|\Gamma^{-1}\|^2 \left(r^4 + \|G\|_{L^4(\mu)}^4 \right) < \infty,
\end{aligned}$$

and similarly for Φ_2. Hence, by the Hellinger stability lemma,

$$d_{\mathrm{H}}(\mu^y, \mu^{\widetilde{y}}) \leq \frac{1}{\sqrt{2}\min(Z, \widetilde{Z})} \|\Phi_1 - \Phi_2\|_{L^2(\mu)}.$$

For the quadratic form $Q(v) := \frac{1}{2}\|v\|_\Gamma^2 = \frac{1}{2}\langle v, \Gamma^{-1}v\rangle$, a quick calculation using its gradient $\nabla Q(v) = \Gamma^{-1}v$ and Hessian $\nabla^2 Q(v) = \Gamma^{-1}$ yields

$$|Q(v) - Q(w)| \leq 2\|\Gamma^{-1}\| \max(\|v\|, \|w\|)\|v - w\|.$$

Hence, the potential is locally Lipschitz with respect to y:

$$|\Phi_1(u) - \Phi_2(u)| = |Q(y - G(u)) - Q(\tilde{y} - G(u))| \leq 2\|\Gamma^{-1}\|(r + \|G(u)\|)\|y - \tilde{y}\|;$$

hence,

$$\|\Phi_1 - \Phi_2\|_{L^2(\mu)} \leq 2\|\Gamma^{-1}\|(2r^2 + 2\|G\|_{L^2(\mu)}^2)^{1/2}\|y - \tilde{y}\|.$$

So, by the Hellinger stability lemma, $|Z - \widetilde{Z}| \leq \|\Phi_1 - \Phi_2\|_{L^2(\mu)} \leq K_{r,G,\Gamma,\mu}\|y - \tilde{y}\|$.

Proof of Theorem 43 III

Let $R > 0$ be large enough that the ball $B(0, R) \subset \mathcal{U}$ has strictly positive μ-mass. Since $\frac{1}{2}\|y - G(u)\|_\Gamma^2 \le \|\Gamma^{-1}\|(r^2 + \|G(u)\|^2)$, and since $\|G(\cdot)\| \le \|G(0)\| + R\,\mathrm{Lip}(G)$ on $B(0, R)$,

$$
\begin{aligned}
Z &= \int_{\mathcal{U}} \exp\left(-\tfrac{1}{2}\|y - G(u)\|_\Gamma^2\right) \mu(\mathrm{d}u) \\
&\ge \int_{B(0,R)} \exp\left(-\tfrac{1}{2}\|y - G(u)\|_\Gamma^2\right) \mu(\mathrm{d}u) \\
&\ge \exp\big(-\|\Gamma^{-1}\|^2(r^2 + (\|G(0)\| + R\,\mathrm{Lip}(G))^2\big)\mu(B(0, R)) \\
&=: Z_{\min}(r, G, \Gamma, \mu) > 0,
\end{aligned}
$$

and similarly for $\widetilde{Z}$. In summary, we obtain

$$
d_{\mathrm{H}}(\mu^y, \mu^{\widetilde{y}}) \le \frac{1}{\sqrt{2} Z_{\min}} 2\|\Gamma^{-1}\|(2r^2 + 2\|G\|_{L^2(\mu)}^2)^{1/2}\|y - \widetilde{y}\| = K_{r,G,\Gamma,\mu}\|y - \widetilde{y}\|,
$$

as claimed. $\square$

Theorem 44

Under the usual assumptions, for all $r > 0$, there exists $K = K_{r,y,\mu}$ such that, whenever $\|G\|_{L^4(\mu)}, \|\widetilde{G}\|_{L^4(\mu)} \leq r$, and with

$$\widetilde{\mu}^y(\mathrm{d}u) := \frac{\exp(-\frac{1}{2}\|y - \widetilde{G}(u)\|_\Gamma^2)}{\widetilde{Z}}\,\mu(\mathrm{d}u), \qquad \widetilde{Z} := \int_{\mathcal{U}} \exp\Big(-\tfrac{1}{2}\|y - \widetilde{G}(u)\|_\Gamma^2\Big)\,\mu(\mathrm{d}u),$$

we have

$$|Z - \widetilde{Z}| \leq K_{r,y,\mu}\|G - \widetilde{G}\|_{L^4(\mu)},$$

$$d_{\mathrm{H}}(\mu^y, \widetilde{\mu}^y) \leq \frac{K_{r,y,\mu}}{\min(Z, \widetilde{Z})}\|G - \widetilde{G}\|_{L^4(\mu)}.$$

Corollary 45

Under the usual assumptions, suppose that $G_N \colon \mathcal{U} \to \mathcal{Y}$, for $N \in \mathbb{N}$, are such that

$$\|G(u) - G_N(u)\| \leq M_1(u)\psi(N),$$
$$\max(\|G(u)\|, \|G_N(u)\|) \leq M_2(u),$$

for all $N \in \mathbb{N}$ and μ-a.a. $u \in \mathcal{U}$, where $\psi \colon \mathbb{N} \to \mathbb{R}$ is measurable, and $M_1, M_2 \in L^2(\mathcal{U}, \mu; \mathbb{R})$ have $M_1 M_2 \in L^2(\mathcal{U}, \mu; \mathbb{R})$. Then, for

$$\mu^{y,N}(\mathrm{d}u) := \frac{\exp(-\frac{1}{2}\|y - G_N(u)\|_\Gamma^2)}{Z^h}\mu(\mathrm{d}u) \quad Z^N := \int_{\mathcal{U}} \exp\left(-\tfrac{1}{2}\|y - G_N(u)\|_\Gamma^2\right)\mu(\mathrm{d}u),$$

we have the same rate of convergence for the BIP as the forward problem:

$$d_{\mathrm{H}}(\mu^y, \mu^{y,N}) \leq K_{y,M_1,M_2,\mu}\psi(N).$$

Finite-dimensional approximation — a subtlety

- In practice, we usually compute on a finite-dimensional subspace $\mathcal{U}_N \subset \mathcal{U}$: the approximate forward model is $G_N \colon \mathcal{U}_N \to \mathcal{Y}$ and the approximate prior and posterior are in $\mathcal{P}(\mathcal{U}_N) \subset \mathcal{P}(\mathcal{U})$.
- We do *not* assert that these approximate prior/posterior measures are close to the idealised ones! In fact they are generally Hellinger distance 1 apart, for all N!
- What *is* true, under suitable assumptions, is that product of the "posterior" on $\mathcal{U}_N$ with the prior on $\mathcal{U}_N^\perp$, is close to the ideal posterior.

Corollary 46

Suppose that $\mathcal{U}_N$ is complemented, i.e. there is a closed subspace $\mathcal{U}_N^\perp \subset \mathcal{U}$ such that $\mathcal{U} = \mathcal{U}_N \oplus \mathcal{U}_N^\perp$, that $G_N \colon \mathcal{U} \to \mathcal{Y}$ is independent of the $\mathcal{U}_N^\perp$-component of its input, and that the prior μ factors as $\mu = \mu_N \otimes \mu_N^\perp$ with $\mu_N \in \mathcal{P}(\mathcal{U}_N)$, $\mu_N^\perp \in \mathcal{P}(\mathcal{U}_N^\perp)$. Then the posterior $\mu^{y,N}$ factors as $\mu^{y,N} = \widetilde{\mu}_N^y \otimes \mu_N^\perp$ and converges to μ^y under the assumptions of Corollary 45.

$$
\begin{aligned}
Z^N &= \int_{\mathcal{U}} \exp\left(-\tfrac{1}{2}\|y - G_N(u)\|_\Gamma^2\right) \mu(\mathrm{d}u) \\
&= \int_{\mathcal{U}_N} \int_{\mathcal{U}_N^\perp} \exp\left(-\tfrac{1}{2}\|y - G_N(u_N + u_N^\perp)\|_\Gamma^2\right) \mu_N^\perp(\mathrm{d}u_N^\perp)\mu_N(\mathrm{d}u_N) \\
&= \int_{\mathcal{U}_N} \int_{\mathcal{U}_N^\perp} \exp\left(-\tfrac{1}{2}\|y - G_N(u_N)\|_\Gamma^2\right) \mu_N^\perp(\mathrm{d}u_N^\perp)\mu_N(\mathrm{d}u_N) \\
&= \int_{\mathcal{U}_N} \exp\left(-\tfrac{1}{2}\|y - G_N(u_N)\|_\Gamma^2\right) \mu_N(\mathrm{d}u_N).
\end{aligned}
$$

and

$$
\begin{aligned}
\mu^{y,N}(\mathrm{d}u) &= (Z^N)^{-1} \exp\left(-\tfrac{1}{2}\|y - G_N(u)\|_\Gamma^2\right) \mu(\mathrm{d}u) \\
&= (Z_N^y)^{-1} \exp\left(-\tfrac{1}{2}\|y - G_N(u_N)\|_\Gamma^2\right) \mu_N(\mathrm{d}u_N)\mu_N^\perp(\mathrm{d}u_N^\perp).
\end{aligned}
$$

The rest follows from Corollary 45. $\square$

Theorem 47

Under the usual assumptions, let $\widetilde{\mu} \in \mathcal{P}(\mathcal{U})$ be another prior measure and set

$$\widetilde{\mu}^y(\mathrm{d}u) := \frac{\exp(-\frac{1}{2}\|y - G(u)\|_\Gamma^2)}{\widetilde{Z}}\,\widetilde{\mu}(\mathrm{d}u), \qquad \widetilde{Z} := \int_{\mathcal{U}} \exp\Big(-\tfrac{1}{2}\|y - G(u)\|_\Gamma^2\Big)\,\widetilde{\mu}(\mathrm{d}u).$$

Then

$$|Z - \widetilde{Z}| \leq d_{\mathrm{H}}(\mu, \widetilde{\mu}),$$
$$d_{\mathrm{H}}(\mu^y, \widetilde{\mu}^y) \leq \frac{2}{\min(Z, \widetilde{Z})} d_{\mathrm{H}}(\mu, \widetilde{\mu}).$$

Exercise 48

Prove Theorem 44, Corollary 45, and Theorem 47, in each case using a similar strategy to Theorem 43.

Stability — summary and warning

The above results indicate that BIPs are well-posed in the sense that the posterior μ^y is a locally Lipschitz function of the prior μ, the forward model G, and the data y.

Warning!

The Lipschitz constants in the above stability results scale like Z^{-1}, which blows up as the data becomes very precise, e.g. $\varepsilon \sim \mathcal{N}(0, \Gamma / \beta)$, $\beta > 0$:

$$\begin{aligned} Z &:= \int_{\mathcal{U}} \exp\left(-\tfrac{1}{2} \| y - G(u) \|_{\Gamma/\beta}^2 \right) \mu(\mathrm{d}u) \\ &= \int_{\mathcal{U}} \left(\exp\left(-\tfrac{1}{2} \| y - G(u) \|_{\Gamma}^2 \right) \right)^{\beta} \mu(\mathrm{d}u) \xrightarrow[\beta \to \infty]{} \mu[G = y] = 0, \text{ typically.} \end{aligned}$$

Similar blow-up occurs for fixed Γ, but a growing number of i.i.d. observations. Thus, the above Hellinger well-posedness results are not uniform and become ineffective in the limit of small observational noise or large data.

Using moment-based arguments, Owhadi et al. (2015) construct examples of priors $\mu, \widetilde{\mu} \in \mathcal{P}(\mathcal{U})$ with $d_{\mathrm{H}}(\mu, \widetilde{\mu}) \approx 0$ but $d_{\mathrm{H}}(\mu^y, \widetilde{\mu}^y) \approx 1$ for precise enough y.

Well-posedness of BIPs

General additive noise

- We now sketch out how to extend the above theory to models of the form

$$\boldsymbol{y} = G(\boldsymbol{u}) + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ can be non-Gaussian and $\mathcal{Y}$ can be infinite-dimensional.

Example 49

On a "nice" domain $D \subseteq \mathbb{R}^d$, recover the initial condition $u_0 \in L^2(D; \mathbb{R})$ of the heat equation $\partial_t u_t = \Delta u_t$ from a noisy observation of u_1. Here, $G = e^{-\Delta}$, which is so strongly smoothing that the (naïve) inverse problem is highly ill-conditioned.

- Obviously we can no longer rely on Lebesgue densities for $\mu_{\boldsymbol{\varepsilon}}$, $\mu_{\boldsymbol{y}|\boldsymbol{u}=u}$, or $\mu_{\boldsymbol{y}}$.
- More importantly, the potential $\Phi \colon \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$ must be allowed to take negative values and even to be unbounded below. Thus, in general, "potential" $\neq$ "misfit".
- See Stuart (2010, Remark 3.8) and Kasanický and Mandel (2017).

- As usual, $(\boldsymbol{u}, \boldsymbol{y})$ is assumed to be a well-defined $\mathcal{U} \times \mathcal{Y}$-valued random variable.
- The marginal distribution of $\boldsymbol{u}$ is the Bayesian prior $\mu \in \mathcal{P}(\mathcal{U})$.
- The noise $\boldsymbol{\varepsilon}$ is distributed according to $\nu^0 \in \mathcal{P}(\mathcal{Y})$, independently of $\boldsymbol{u}$.
- So $\boldsymbol{y} | \boldsymbol{u} = u$ is distributed according to ν^u, the translate of ν^0 by $G(u)$, which is assumed to be absolutely continuous with respect to ν^0, with

$$\frac{\mathrm{d}\nu^u}{\mathrm{d}\nu^0}(y) = \exp(-\Phi(u; y)).$$

- $\Phi \colon \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$ is the **negative log-likelihood** or simply **potential**.
- Thus $(\boldsymbol{u}, \boldsymbol{y}) \sim \mu(\mathrm{d}u)\nu^u(\mathrm{d}y)$, and a similar argument to before gives us the regular conditional distribution / posterior distribution

$$\mu^y(\mathrm{d}u) := \mu_{\boldsymbol{u}|\boldsymbol{y}}(y, \mathrm{d}u) = \frac{\exp(-\Phi(u; y))}{Z(y)} \mu(\mathrm{d}u),$$

$$Z(y) := \int_{\mathcal{U}} \exp(-\Phi(u; y)) \, \mu(\mathrm{d}u).$$

Cameron–Martin revisited

- For the purposes of integrating over u, it is permitted to change Φ by adding any ν^0-a.s. *finite* $g(y)$ independent of u and renormalising, absorbing the change into the normalising constant $Z(y)$.
- If $\varepsilon \sim \nu^0 = \mathcal{N}(0, \Gamma)$ on a Banach space $\mathcal{Y}$, then the Cameron–Martin theorem gives

$$\begin{aligned} \frac{\mathrm{d}\nu^u}{\mathrm{d}\nu^0}(y) &= \exp\bigl(\langle G(u), y\rangle_\Gamma - \tfrac{1}{2}\|G(u)\|_\Gamma^2\bigr), \\ \text{and so} \quad \Phi(u; y) &= \tfrac{1}{2}\|G(u)\|_\Gamma^2 - \langle G(u), y\rangle_\Gamma. \end{aligned}$$

- When $\dim \mathcal{Y} < \infty$ we tend to work with the familiar misfit

$$\Psi(u; y) := \Phi(u; y) + \tfrac{1}{2}\|y\|_\Gamma^2 = \tfrac{1}{2}\|y - G(u)\|_\Gamma^2.$$

- This Ψ is *not* an admissible potential when $\dim \mathcal{Y} = \infty$, because $\frac{1}{2}\|y\|_\Gamma^2 = \infty$ ν^0-a.s. (Recall that the Cameron–Martin space of an infinite-dimensional Gaussian measure has measure zero.)

Further modifications to the theory

- Since Φ can be unbounded below, one has to introduce a lower bound on Φ such as $\Phi(u;y) \geq L(r)$ when $\|u\|, \|y\| \leq r$. This lower bound will need to be exponentially integrable with respect to the prior μ, and its integral will appear as a constant in bounds on $|Z - \widetilde{Z}|$, $d_{\mathrm{H}}(\mu^y, \widetilde{\mu}^y)$ etc.
- Remember while quadratic functions are exponentially integrable with respect ot Gaussian priors, this is definitely not the case for all functions and all priors. In particular, for heavy-tailed priors, the class of potentials for which one can get well-posedness is more restricted than in the Gaussian case.

Markov chain Monte Carlo

The challenge: sampling the posterior

- Except in very special cases (e.g. Gaussian prior, linear G, additive Gaussian ε) we do not have a closed-form expression for the posterior distribution μ^y. Nevertheless, we need to access μ^y, and to integrate quantities of interest $f\colon \mathcal{U} \to \mathbb{R}$ against it.
- Monte Carlo methods aim to draw a sequence $(\boldsymbol{x}_j)_{j\in\mathbb{N}}$ of i.i.d. samples from μ^y and to approximate integrals by sample averages:

$$\int_{\mathcal{U}} f(x)\,\mu^y(\mathrm{d}x) \approx \frac{1}{J}\sum_{j=1}^{J} f(\boldsymbol{x}_j).$$

- Markov chain Monte Carlo (MCMC) methods generate samples that asymptotically satisfy these requirements via a Markov chain with limiting invariant distribution μ^y.
- Since there is nothing especially Bayesian about this procedure, we just write μ instead of μ^y for this **target measure**.
- We especially seek algorithms that work well independently of $\dim \mathcal{U}$.

Markov chain Monte Carlo

Invariance, reversibility, convergence

Definition 50

A **Markov chain** on a state space $\mathcal{X}$ is a sequence $(\boldsymbol{x}_j)_{j\in\mathbb{N}}$ of $\mathcal{X}$-valued random variables such that, for each j, "$\boldsymbol{x}_{j+1}$ only depends on $\boldsymbol{x}_j$". More precisely,

$$\boldsymbol{x}_{j+1}|\boldsymbol{x}_j,\boldsymbol{x}_{j-1},\dots,\boldsymbol{x}_1 \text{ and } \boldsymbol{x}_{j+1}|\boldsymbol{x}_j \text{ are equal in distribution.}$$

For simplicity, we will assume that this distribution does not depend on j — the **time-homogeneous** case. Such a time-homogeneous Markov chain is specified by its **transition kernel** $\kappa\colon \mathcal{X}\times\mathcal{B}(\mathcal{X})\to[0,1]$,

$$\kappa(x,A) := \mathbb{P}[\boldsymbol{x}_{j+1}\in A|\boldsymbol{x}_j = x],$$
$$\kappa^m(x,A) := \mathbb{P}[\boldsymbol{x}_{j+m}\in A|\boldsymbol{x}_j = x].$$

Simply put, our aim is to construct a Markov chain for which $\kappa^m(x,\cdot)\to\mu$ as $m\to\infty$.

Definition 51

A Markov chain $(\boldsymbol{x}_j)_{j\in\mathbb{N}}$ with transition kernel $\kappa\colon \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ has $\mu \in \mathcal{P}(\mathcal{X})$ as an **invariant measure** if $\kappa\mu = \mu$, i.e.

$$(\kappa\mu)(A) := \int_{\mathcal{X}} \kappa(x, A)\, \mu(\mathrm{d}x) = \mu(A) \quad \text{for all } A \in \mathcal{B}(\mathcal{X}).$$

Further, κ is **reversible** (a.k.a. satisfies **detailed balance**) with respect to μ if $\rho^+ = \rho^- \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$, where

$$\begin{aligned}
\rho^+(\mathrm{d}x, \mathrm{d}x') &:= \kappa(x, \mathrm{d}x')\mu(\mathrm{d}x), \\
\rho^-(\mathrm{d}x, \mathrm{d}x') &:= \rho^+(\mathrm{d}x', \mathrm{d}x) = \kappa(x', \mathrm{d}x)\mu(\mathrm{d}x').
\end{aligned}$$

Theorem 52

If κ is μ-reversible, then μ is invariant under κ.

Proof of Theorem 52

Suppose that κ is μ-reversible, and let $A \in \mathcal{B}(\mathcal{X})$ be arbitrary. Then

$$
\begin{aligned}
(\kappa\mu)(A) := \int_{\mathcal{X}} \kappa(x, A)\, \mu(\mathrm{d}x) &= \int_{\mathcal{X}} \int_{A} \kappa(x, \mathrm{d}x')\, \mu(\mathrm{d}x) \\
&= \int_{\mathcal{X}} \int_{A} \kappa(x', \mathrm{d}x)\, \mu(\mathrm{d}x') && \text{(reversibility)} \\
&= \int_{A} \int_{\mathcal{X}} \kappa(x', \mathrm{d}x)\, \mu(\mathrm{d}x') && \text{(Fubini)} \\
&= \int_{A} \kappa(x', \mathcal{X})\, \mu(\mathrm{d}x') \\
&= \int_{A} 1\, \mu(\mathrm{d}x') = \mu(A).
\end{aligned}
$$

Thus, as claimed, μ is κ-invariant. $\square$

Markov chain Monte Carlo

Metropolis–Hastings on $\mathbb{R}^n$

The idea of Metropolis–Hastings

- Directly positing a kernel κ that is μ-reversible is rather tricky. Even if we had one, it might be quite complicated, so sampling from $\kappa(\boldsymbol{x}_j, \cdot)$ to generate $\boldsymbol{x}_{j+1}$ could be as hard as sampling from μ in the first place.
- The idea of the **Metropolis–Hastings** procedure is to
  - **propose** $\boldsymbol{x}'_{j+1}$ according to a (simple) **proposal kernel** $Q \colon \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$,
  - **accept or reject** this proposal and set $\boldsymbol{x}_{j+1} := \boldsymbol{x}'_{j+1}$ or $\boldsymbol{x}_{j+1} := \boldsymbol{x}_j$ respectively according to some **acceptance probability** α,
  - in such as way as Q and α together define a μ-reversible transition kernel.
- For simplicity, we first work on $\mathbb{R}^n$ and assume that μ has density ρ and Q has density q.

Algorithm 53 (Metropolis–Hastings on $\mathbb{R}^n$)

Given a target density $\rho \colon \mathbb{R}^n \to [0, \infty)$, a proposal density $q \colon \mathbb{R}^n \times \mathbb{R}^n \to [0, \infty)$, and an initial density $\rho_0 \colon \mathbb{R}^n \to [0, \infty)$, we **output** $(\boldsymbol{x}_j)_{j \in \mathbb{N}}$ as follows:

1. Draw $\boldsymbol{x}_1 \sim \rho_0$.
2. For $j \in \mathbb{N}$:
   2.1 Draw a proposal $\boldsymbol{x}'_{j+1} \sim q(\boldsymbol{x}_j, \cdot)$.
   2.2 Independently of the other random variables, draw $\boldsymbol{z}_{j+1} \sim \mathrm{Unif}([0, 1])$, and set

$$\boldsymbol{x}_{j+1} := \begin{cases} \boldsymbol{x}'_{j+1} & (\text{“accept”}) \text{ if } \boldsymbol{z}_{j+1} \leq \alpha(\boldsymbol{x}_j, \boldsymbol{x}'_{j+1}), \\ \boldsymbol{x}_j & (\text{“reject”}) \text{ otherwise,} \end{cases}$$

   where α is the acceptance probability

$$\alpha(x, x') := \min\left(1, \frac{\rho(x')q(x', x)}{\rho(x)q(x, x')}\right).$$

Example 54

The (centred, isotropic) **Gaussian random walk proposal** on $\mathbb{R}^n$ is the proposal density

$$q(x, x') \propto \exp\left(- \frac{\|x - x'\|^2}{2s^2} \right),$$

i.e. the proposal distribution $Q(x, \cdot\,) = \mathcal{N}(x, s^2 I)$, i.e.

$$\boldsymbol{x}'_{j+1} := \boldsymbol{x}_j + s\boldsymbol{\xi}_{j+1}, \qquad \boldsymbol{\xi}_{j+1} \sim \mathcal{N}(0, I),$$

where $s > 0$ is a step size parameter that can be tuned according to various criteria.

- The basic dilemma is that making s smaller promotes higher α but makes the chain immobile with high autocorrelation; choosing a large s means proposing longer-range moves, but at the cost of lower acceptance probability.
- For product measure targets μ, s should be scaled to achieve average acceptance probability ≈ 0.234 (Roberts et al., 1997).

Theorem 55

The Metropolis–Hastings transition kernel

$$\kappa(x, \mathrm{d}x') = \alpha(x, x') Q(x, \mathrm{d}x') + \left(1 - \int_{\mathbb{R}^n} \alpha(x, x'')\, Q(x, \mathrm{d}x'')\right) \delta_x(\mathrm{d}x')$$

is μ-reversible, and hence μ-invariant. If ρ and q satisfy the positivity condition

$$\rho(x') > 0 \implies (\text{for all } x \in \mathbb{R}^n, q(x, x') > 0),$$

then

$$\text{for all } x \in \mathbb{R}^n, \qquad \lim_{m \to \infty} d_{\mathrm{TV}}(\kappa^m(x, \cdot), \mu) = 0 \qquad \text{and,}$$

$$\text{for all } \nu_0 \in \mathcal{P}(\mathbb{R}^n), f \in L^1(\mu), \qquad \lim_{J \to \infty} \frac{1}{J} \sum_{j=1}^{J} f(\boldsymbol{x}_j) = \int_{\mathcal{X}} f(x)\, \mu(\mathrm{d}x) \qquad \mathbb{P}\text{-a.s.}$$

Markov chain Monte Carlo

Metropolis–Hastings on $\mathcal{H}$

Algorithm 56 (Metropolis–Hastings on a Hilbert space $\mathcal{H}$)

Given a target measure $\mu \in \mathcal{P}(\mathcal{H})$, a proposal kernel $Q \colon \mathcal{H} \times \mathcal{B}(\mathcal{H}) \to [0,1]$, and an initial distribution $\nu_0 \in \mathcal{P}(\mathcal{H})$, we **output** $(\boldsymbol{x}_j)_{j\in\mathbb{N}}$ as follows:

1. Draw $\boldsymbol{x}_1 \sim \nu_0$.
2. For $j \in \mathbb{N}$:
   2.1 Draw a proposal $\boldsymbol{x}'_{j+1} \sim Q(\boldsymbol{x}_j, \cdot)$.
   2.2 Independently of the other random variables, draw $\boldsymbol{z}_{j+1} \sim \mathrm{Unif}([0,1])$, and set

$$
\begin{aligned}
\boldsymbol{x}_{j+1} &:= \begin{cases} \boldsymbol{x}'_{j+1} & (\text{“accept”}) \text{ if } \boldsymbol{z}_{j+1} \leq \alpha(\boldsymbol{x}_j, \boldsymbol{x}'_{j+1}), \\ \boldsymbol{x}_j & (\text{“reject”}) \text{ otherwise,} \end{cases} \\
\alpha(x, x') &:= \min\left(1, \frac{\mathrm{d}\rho^-}{\mathrm{d}\rho^+}(x, x')\right), \\
\rho^+(\mathrm{d}x, \mathrm{d}x') &:= Q(x, \mathrm{d}x')\mu(\mathrm{d}x), \\
\rho^-(\mathrm{d}x, \mathrm{d}x') &:= \rho^+(\mathrm{d}x', \mathrm{d}x) = Q(x', \mathrm{d}x)\mu(\mathrm{d}x').
\end{aligned}
$$

The acceptance probability

The acceptance probability is defined through

$$\alpha(x,x') := \min\left(1, \frac{\mathrm{d}\rho^-}{\mathrm{d}\rho^+}(x,x')\right),$$

$$\rho^+(\mathrm{d}x,\mathrm{d}x') := Q(x,\mathrm{d}x')\mu(\mathrm{d}x),$$

$$\rho^-(\mathrm{d}x,\mathrm{d}x') := \rho^+(\mathrm{d}x',\mathrm{d}x) = Q(x',\mathrm{d}x)\mu(\mathrm{d}x').$$

In the familiar case of $\mathcal{H} = \mathbb{R}^n$ with densities this reduces to the usual formula:

$$\rho^+(\mathrm{d}x,\mathrm{d}x') = Q(x,\mathrm{d}x')\mu(\mathrm{d}x) = q(x,x')\rho(x)\,\mathrm{d}x'\mathrm{d}x,$$

$$\rho^-(\mathrm{d}x,\mathrm{d}x') = Q(x',\mathrm{d}x)\mu(\mathrm{d}x') = q(x',x)\rho(x')\,\mathrm{d}x\mathrm{d}x',$$

$$\frac{\mathrm{d}\rho^-}{\mathrm{d}\rho^+}(x,x') = \frac{q(x',x)\rho(x')}{q(x,x')\rho(x)}.$$

Theorem 57 (Cotter et al., 2013; Tierney, 1998)

If the Radon–Nikodym derivative $\frac{\mathrm{d}\rho^-}{\mathrm{d}\rho^+}$ *exists and we take* $\alpha(x, x') := \min\left(1, \frac{\mathrm{d}\rho^-}{\mathrm{d}\rho^+}(x, x')\right)$, *then the Metropolis–Hastings transition kernel*

$$\kappa(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + \left(1 - \int_{\mathcal{H}} \alpha(x, x'')\, Q(x, \mathrm{d}x'')\right)\delta_x(\mathrm{d}x')$$

is μ*-reversible, and hence* μ*-invariant.*

The problem is that, when $\dim \mathcal{H} = \infty$, it can be tricky to construct Q so that $\frac{\mathrm{d}\rho^-}{\mathrm{d}\rho^+}$ exists (recall the Feldman–Hájek dichotomy). In particular, $\frac{\mathrm{d}\rho^-}{\mathrm{d}\rho^+}$ does not exist for the Gaussian random walk proposal, though a simple modification can resolve this problem.

Theorem 58 (Cotter et al., 2013; Hairer et al., 2014)

Suppose that $\mathrm{d}\mu \propto e^{-\Phi}\,\mathrm{d}\mu_0$, *where* $\mu_0 = \mathcal{N}(0, C_0)$. *For any step size* $0 < s < 1$, *the* ***preconditioned Crank–Nicolson*** *proposal kernel* $Q(x, \cdot) := \mathcal{N}\big(\sqrt{1-s^2}x, s^2 C_0\big)$, *i.e.*

$$x'_{j+1} := \sqrt{1-s^2}\,x_j + s\xi_{j+1}, \qquad \xi_{j+1} \sim \mathcal{N}(0, C_0)$$

has acceptance probability

$$\alpha(x, x') = \min\big(1, \exp(\Phi(x) - \Phi(x'))\big)$$

and defines a μ*-reversible Markov chain on* $\mathcal{H}$, *and* $\kappa^m(x, \cdot) \to \mu$ *as* $m \to \infty$.

The convergence rate $\kappa^m(x, \cdot) \to \mu$ depends on a quantity called the **spectral gap**; Hairer et al. (2014) showed that pCN has a positive spectral gap on $\mathcal{H}$. As a result, although in practice we run MCMC on an n-dimensional subspace of $\mathcal{H}$, n-dimensional pCN inherits ∞-dimensional pCN's convergence rates etc. in an n-independent way.

Example 59

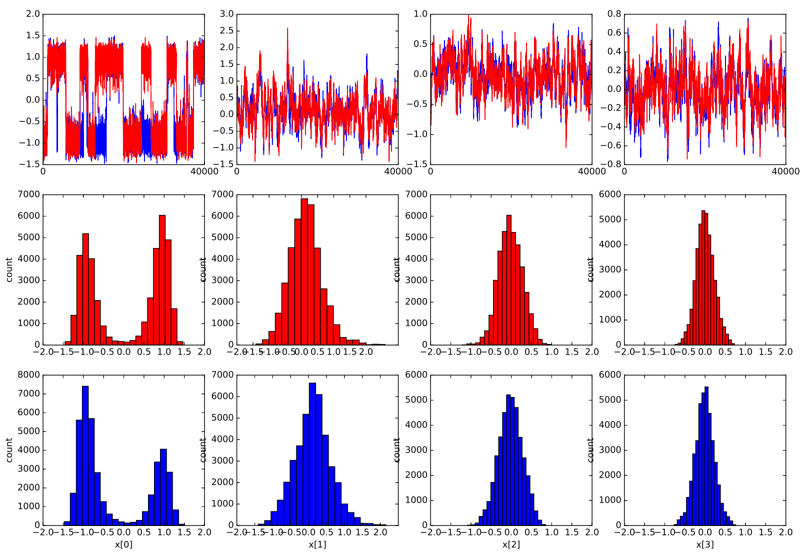
Consider the Gaussian prior $\mu_0 = \bigotimes_{n\in\mathbb{N}} \mathcal{N}(0, n^{-1}) \in \mathcal{P}(\ell^2)$ and a simple double-well potential in the first coordinate only:

$$\Phi\big((x_n)_{n\in\mathbb{N}}\big) := 4(x_1^2 - 1)^2.$$

We run GRW and pCN Metropolis–Hastings on the first $d \in \mathbb{N}$ components of ℓ^2 for dimension $1 \leq d \leq 1000$ and step size $0 < s < 1$. Each run has 10^6 iterations and the same initial random seed.

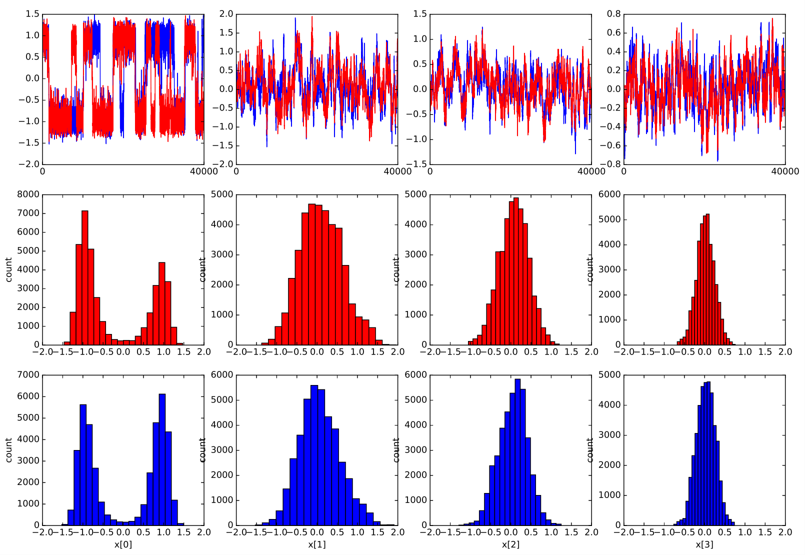
To give GRW a fighting chance, we use the prior-adapted Gaussian proposal $Q(x, \cdot) = \mathcal{N}(x, s^2 C_0)$ — the performance is even worse for the isotropic proposal $Q(x, \cdot) = \mathcal{N}(x, s^2 I)$.

Comparing GRW to pCN II



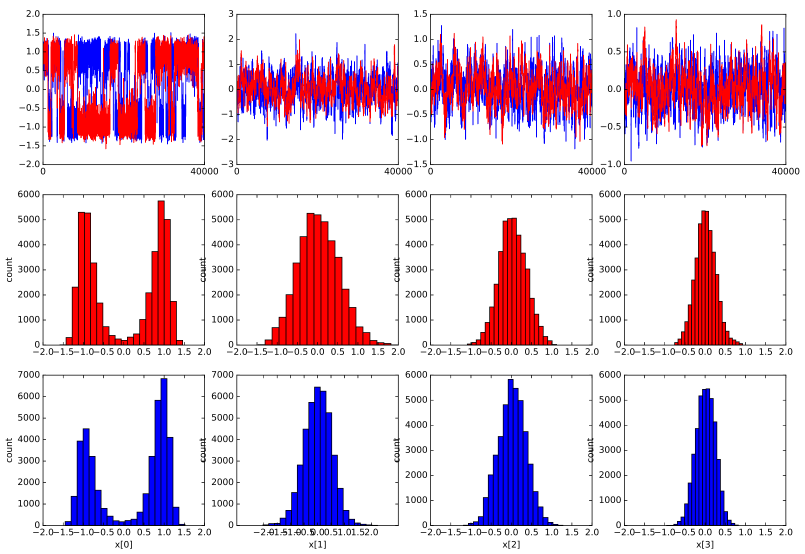
The first 40,000 iterations of GRW and pCN with step size $s = 0.1$ in dimension $d = 10$.

Comparing GRW to pCN II



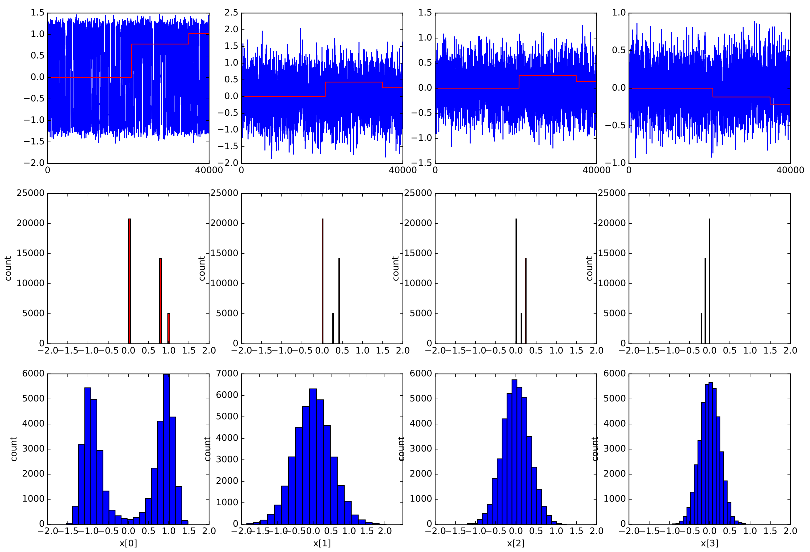
The first 40,000 iterations of GRW and pCN with step size $s = 0.1$ in dimension $d = 100$.

Comparing GRW to pCN II



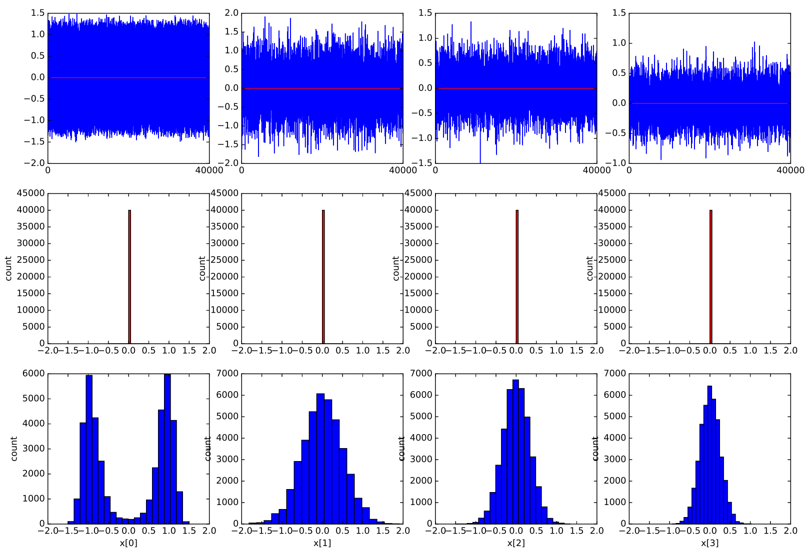
The first 40,000 iterations of GRW and pCN with step size $s = 0.2$ in dimension $d = 100$.

Comparing GRW to pCN II



The first 40,000 iterations of GRW and pCN with step size $s = 0.5$ in dimension $d = 100$.

Comparing GRW to pCN II



The first 40,000 iterations of GRW and pCN with step size $s = 0.9$ in dimension $d = 100$.

Comparing GRW to pCN III

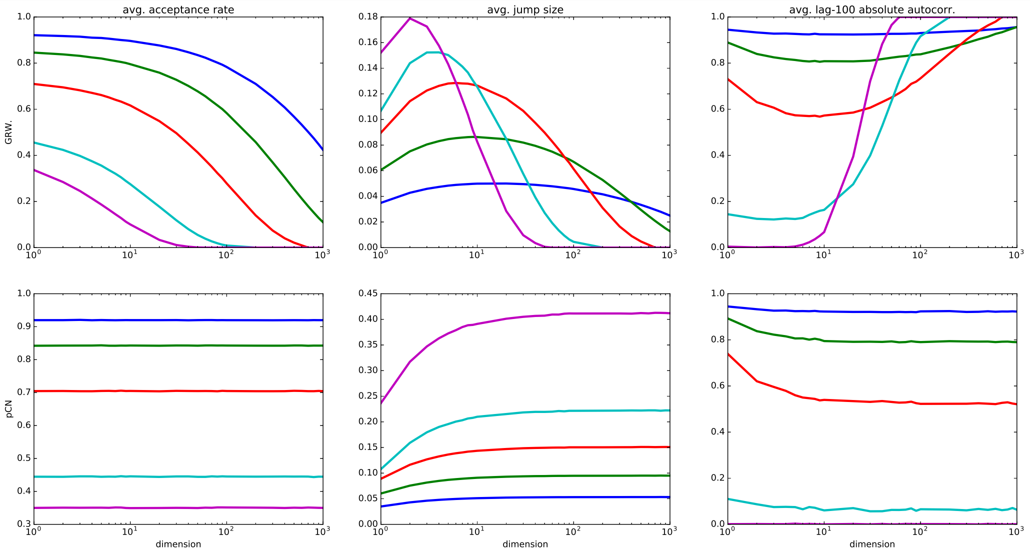
$s = 0.05$

$s = 0.1$

$s = 0.2$

$s = 0.5$

$s = 0.9$



Average acceptance rate (left), jump distance (middle), and lag-100 absolute autocorrelation (right) for GRW (top) and pCN (bottom).

Exercise 60

Let $\mathrm{d}\mu \propto e^{-\Phi}\mathrm{d}\mu_0$, $\mu_0 = \mathcal{N}(0, C_0)$, on $\mathcal{H} = \mathbb{R}^n$.

- Let q be any symmetric proposal density, i.e. $q(x, x') = q(x', x)$. Show that the Hastings ratio reduces to

$$\frac{q(x', x)\rho(x')}{q(x, x')\rho(x)} = \exp\left(\Phi(x) - \Phi(x') + \frac{1}{2}\|x\|_{C_0}^2 - \frac{1}{2}\|x'\|_{C_0}^2\right).$$

  On a hand-wavy level, what happens to the "$-\frac{1}{2}\|x'\|_{C_0}^2$" term, and hence to the acceptance probability, when $\dim \mathcal{H} = \infty$? (Hint: Cameron–Martin.)

- Consider the pCN proposal on $\mathcal{H} = \mathbb{R}^n$. Show that the Hastings ratio reduces to

$$\frac{q(x', x)\rho(x')}{q(x, x')\rho(x)} = \exp(\Phi(x) - \Phi(x')).$$

W. N. Anderson, Jr. and G. E. Trapp. Shorted operators. II. *SIAM J. Appl. Math.*, 28:60–71, 1975. doi:10.1137/0128007.

V. I. Bogachev. *Gaussian Measures*, volume 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1998. doi:10.1090/surv/062.

J. T. Chang and D. Pollard. Conditioning as disintegration. *Statist. Neerlandica*, 51(3):287–317, 1997. doi:10.1111/1467-9574.00056.

S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: modifying old algorithms to make them faster. *Statist. Sci.*, 28(3):424–446, 2013. doi:10.1214/13-STS421.

M. Dashti and A. M. Stuart. The Bayesian approach to inverse problems. In R. Ghanem, D. Higdon, and H. Owhadi, editors, *Handbook of Uncertainty Quantification*. Springer, 2017. doi:10.1007/978-3-319-12385-1_7.

M. Dashti, S. Harris, and A. Stuart. Besov priors for Bayesian inverse problems. *Inverse Probl. Imaging*, 6(2):183–200, 2012. doi:10.3934/ipi.2012.6.183.

M. M. Dunlop and A. M. Stuart. The Bayesian formulation of EIT: analysis and algorithms. *Inverse Probl. Imaging*, 10(4): 1007–1036, 2016. doi:10.3934/ipi.2016030.

J. Feldman. Equivalence and perpendicularity of Gaussian processes. *Pacific J. Math.*, 8:699–708, 1958.

X. Fernique. Intégrabilité des vecteurs gaussiens. *C. R. Acad. Sci. Paris Sér. A-B*, 270(25):A1698–A1699, 1970.

M. Hairer, A. M. Stuart, and S. J. Vollmer. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.*, 24(6):2455–2490, 2014. doi:10.1214/13-AAP982.

J. Hájek. On a property of normal distribution of any stochastic process. *Czechoslovak Math. J.*, 8 (83):610–618, 1958.

O. Kallenberg. *Foundations of Modern Probability*. Probability and its Applications. Springer-Verlag, 1997.

References ii

I. Kasanický and J. Mandel. On well-posedness of Bayesian data assimilation and inverse problems in Hilbert space, 2017. arXiv:1701.08298.

C. Kraft. Some conditions for consistency and uniform consistency of statistical procedures. *Univ. California Publ. Statist.*, 2: 125–141, 1955.

M. Lassas and S. Siltanen. Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Problems*, 20 (5):1537–1563, 2004. doi:10.1088/0266-5611/20/5/013.

M. Lassas, E. Saksman, and S. Siltanen. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Probl. Imaging*, 3(1):87–122, 2009. doi:10.3934/ipi.2009.3.87.

K. Law, A. Stuart, and K. Zygalakis. *Data Assimilation: A Mathematical Introduction*, volume 62 of *Texts in Applied Mathematics*. Springer, 2015. doi:10.1007/978-3-319-20325-6.

M. Ledoux. Isoperimetry and Gaussian analysis. In *Lectures on probability theory and statistics* (*Saint-Flour, 1994*), volume 1648 of *Lecture Notes in Math.*, pages 165–294. Springer, Berlin, 1996. doi:10.1007/BFb0095676.

M. S. Lehtinen. Modelling of priori information in time series. In *Regularisation and a priori information in retrieval of profiles. ENVISAT Kick-Off Meeting. Luosto, 23–24 March 1999*. 1999.

M. S. Lehtinen, L. Päivärinta, and E. Somersalo. Linear inverse problems for generalised random variables. *Inverse Problems*, 5(4):599–612, 1989. doi:10.1088/0266-5611/5/4/011.

H. C. Lie and T. J. Sullivan. Quasi-invariance of countable products of Cauchy measures under non-unitary dilations. *Electron. Commun. Probab.*, 23(8):1–6, 2018. doi:10.1214/18-ECP113.

H. Owhadi and C. Scovel. Conditioning Gaussian measure on Hilbert space, 2015. arXiv:1506.04208.

H. Owhadi, C. Scovel, and T. J. Sullivan. Brittleness of Bayesian inference under finite information in a continuous world. *Electron. J. Stat.*, 9:1–79, 2015. doi:10.1214/15-EJS989.

S. Reich and C. Cotter. *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge University Press, New York, 2015. doi:10.1017/CBO9781107706804.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120, 1997. doi:10.1214/aoap/1034625254.

V. Sazonov. On characteristic functionals. *Teor. Veroyatnost. i Primenen.*, 3:201–205, 1958.

C. Schillings and B. Sprungk. Inverse Probleme, 2017.

E. Somersalo, M. Cheney, and D. Isaacson. Existence and uniqueness for electrode models for electric current computed tomography. *SIAM J. Appl. Math.*, 52(4):1023–1040, 1992. doi:10.1137/0152060.

A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559, 2010. doi:10.1017/S0962492910000061.

T. J. Sullivan. *Introduction to Uncertainty Quantification*, volume 63 of *Texts in Applied Mathematics*. Springer, 2015. doi:10.1007/978-3-319-23395-6.

T. J. Sullivan. Well-posed Bayesian inverse problems and heavy-tailed stable quasi-Banach space priors. *Inverse Probl. Imaging*, 11(5):857–874, 2017. doi:10.3934/ipi.2017040.

L. Tierney. A note on Metropolis–Hastings kernels for general state spaces. *Ann. Appl. Probab.*, 8(1):1–9, 1998. doi:10.1214/aoap/1027961031.

Y. Yamasaki. *Measures on Infinite-Dimensional Spaces*, volume 5 of *Series in Pure Mathematics*. World Scientific Publishing Co., Singapore, 1985. doi:10.1142/0162.