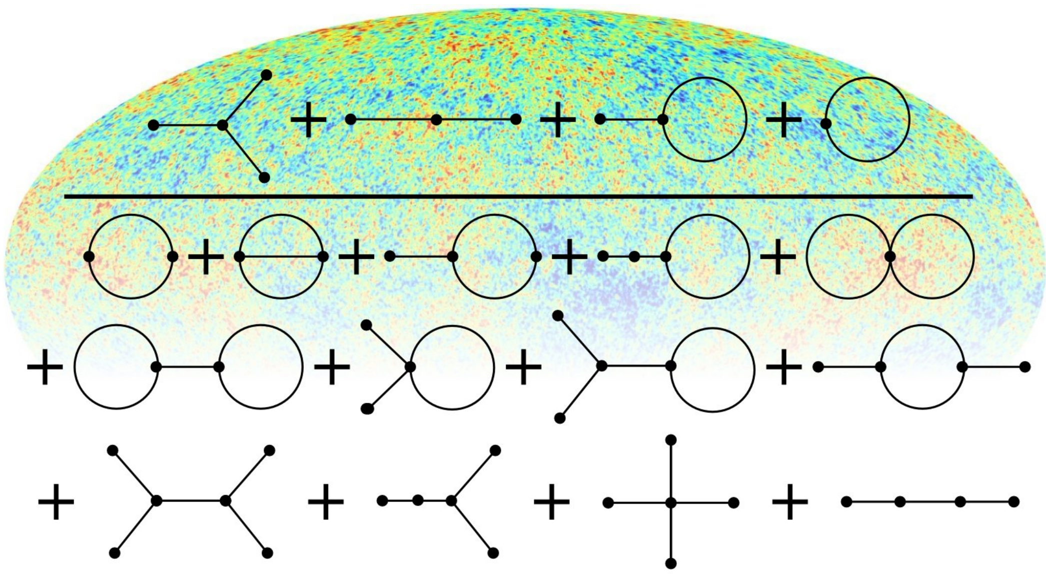
INFORMATION THEORY & INFORMATION FIELD THEORY

TORSTEN ENSSLIN



VERSION MARCH 27, 2023

CONTENTS

PREFACE

Information field theory (IFT) is information theory (IT) for fields. Fields are continuous varying functions over some space, and IT refers to logic under uncertainty, which is probabilistic reasoning. Consequently, this script introduces into IT in Part i and then extend this to IFT in Part ii.

ACKNOWLEDGMENTS

The lectures on information theory and information field (IFT) theory would not have been possible by the kind and enthusiastic support by many others. First I am grateful that the Max-Planck-Institute for Astrophysics provided an academic environment in which IFT could be developed. I have to thank many colleagues, students, postdocs, and co-authors of publications for contributions and encouragements: Mona Frommert for checking the equations and in particular the Feynman diagrams of the initial publication on IFT; Francisco-Shu Kitaura and Jens Jasche for pioneering cosmography in the spirit of IFT; Cornelius Weig for pointing the way how to use effective actions, Erwin Frey for encouraging to give a lecture course on IFT at the Ludwig-Maximilians-Universtity Munich; Henrik Junklewitz and Niels Oppermann not only for their pioneering IFT works on radioastronomical IFT applications like RESOLVE and the Faraday sky reconstruction, but also for running the first set of exercises groups of the lectures; Marco Selig for the first incarnation of Numerical Information Field Theory (NIFTy), D^3PO, and the Fermi gamma ray sky reconstruction; Michael Bell also for developing and naming NIFTy and for Faraday synthesis; Theo Steininger for version two and three of NIFTy; Martin Reineke for taking excellent custody of NIFTy since; Maksim Greiner for pioneering IFT based Galactic tomography; Sebastian Dorn for IFT based reconstructions of the primordial Universe; Vanessa Boehm for her work on IFT based gravitational lensing tomography; Philipp Arras for enhancing NIFTy with auto-differentiation and pushing RESOLVE into the VLBI domain; Margret Westerkamp for the initial typesetting of the IT/IFT Script, the slides derived from that for online teaching, and working out the relation of IFT to the Supersymmetric Theory of Stochstics by Igor Ovchinnikov; Daniel Pumpe for D^4PO and his sobering results on quasi periodic oscillations in the emission from magnetars; Natalia Porqueres for the application of IFT to the cosmic expansion history; Mahsa Ghaempanah for also sobering results on Galactic center positronium annihilation line with INTEGRAL; Fabrizia Guglielmetti for helping to get RESOLVE into usage; Sebastian Hutschenreuter and Valentina Vacca for improving the Faraday sky further; Maximilian Kurthen for pioneering Bayesian causal inference with IFT, which was followed up Matteo Guardiani in his work on Covid-19; Jakob Knollmüller for developing MGVI, which permitted IFT models to become deep, and his exploration of the relation of IFT and machine learning; Philipp Frank for geoVI, the correlated field model, and pioneering dynamical field inference; Reimar Leike for ideas everywhere and for surprising the world with his amazing 3D Galactic dust tomography maps; Lukas Platz for a beautiful updated gamma ray sky reconstruction; Gordian Edenhofer for transferring NIFTy into the JAX era; Jakob Roth for his work on correcting time and direction dependent atmospheres and ionospheres in high cadence and radio interferometric imaging; Martin Erdmann for oping doors to the machine learning world; Philipp Mertsch for showing that IFT & NIFTy can be applied by self instructors outside the IFT core team; Christoph Wellinger and Anna Nelles for applying IFT to radio signals from neutrinos; Johannes Harth-Kitzerow for pioneering Bayesian data compression; Julia Stadler for applying IFT to GRAVITY data of the Galactic Center; Nico Reeb, Philipp Zehetner, and David Outland for their work on deep sea bioluminescence detection with the ANTARES detector; Philipp Haim and Robin Dehde for exploring IFT based medical imaging; Silvan Streit and Vincent Eberle for butterfly matrix based representations of in-homogeneous point spread functions; Johannes Zacherl for exploring neural networks with a Fisher-Net architecture; Peter Biermann, Phil Kronberg, Rick Perley, and Namir Kassim for introducing me to radio astronomy, a beatiful art I am eager to transform into science; the Planck satellite mission on showing me the need for a theoretically principled signal reconstruction framework; many master and bachelor students, who helped with their enthusiasm and projects to fill IFT & NIFTy with life; many students, who took the IFT lectures and gave important feedback on the didactic of IFT. There are many more to thank here, like family, friends, and people I might have forgotten, and there will be more to come.

Thank you all!

Part I

INFORMATION THEORY

1 FROM LOGIC TO PROBABILITY

Here, we give a sketch of the Cox theorem proof [3] while following the book of Jaynes [7] and the lectures by Caticha [2]. A more rigorous proof can be found in [11].

1.1 ARISTOTELIAN LOGIC

Let A and B be statements or propositions (e.g. $A =$ "it rains" and $B =$ "there is a cloud") and $I =$ "if A is true, then B is also true" $=$ "$A \Rightarrow B$" the background information (e.g. I = "it rains only if there is a cloud").

- **strong syllogism**: $I \Rightarrow$ "if B is false then A is false" = $(\overline{B} \Rightarrow \overline{A})$
- **weak syllogism** $I \Rightarrow$ "if B is true then A is more plausible"$= J$
  - This is possible, since we can exclude the case that "B is false" which definitely would have excluded A.
- **weaker syllogism** $J \Rightarrow$ "if A is true, then B becomes more plausible"

1.2 BOOLEAN ALGEBRA

Let A and B be statements or propositions, we introduce the following relations and their notations:

- **"and"**: $AB =$ "both, A and B are true", conjunction or logical product *conjunction*
- **"or"**: $A + B =$ "at least one of the propositions A, B is true", disjunction or logical sum *disjunction*
- **"identity"**: $A = B =$ "A always has the same truth value as B", logical equivalence *logical equivalence*
- **"denial"**: $\overline{A} =$ "not A" $=$ "A is false", negation or logical complement, $A =$ "$\overline{A}$ is false", "A=$\overline{A}$" is always false *logical complement*

Notation:

- $AB + C = (AB) + C$
  - The logical product has a higher binding than the sum
- $\overline{AB} = \overline{(AB)}$= "$AB$ is false"
  - The negation of a logical product ("at least one of A and B is false") is not the product of negations $\overline{A}\overline{B}$ ("A is false and B is false" $=$ "both are false").

Boolean algebra

The Boolean algebra rests on the following axioms:

idempotency:	$AA = A$
	$A + A = A$
commutativity:	$AB = BA$
	$A + B = B + A$
associativity:	$A(BC) = (AB)C = ABC$
	$A + (B + C) = (A + B) + C = A + B + C$
distributivity:	$A(B + C) = AB + AC$
	$A + (BC) = (A + B)(A + C)$
duality:	$\overline{AB} = \overline{A} + \overline{B}$
	$\overline{A + B} = \overline{A}\,\overline{B}$
implication:	"$A \Rightarrow B$" $\equiv$ "$A = AB$" $=$
	"A and AB have the same truth value"

This set of axioms is over-complete. For example the second distributivity axioms follows from the first one and duality:

$$\begin{aligned}
\overline{A} + \overline{B}\,\overline{C} &= \overline{A} + \overline{B + C} \text{ (duality)} \\
&= \overline{A\,(B + C)} \text{ (duality)} \\
&= \overline{AB + AC} \text{ (1st distributivity)} \\
&= \overline{AB}\,\overline{AC} \text{ (duality)} \\
&= (\overline{A} + \overline{B})\,(\overline{A} + \overline{C}) \text{ (duality)},
\end{aligned}$$

which is the second distributivity axiom for $A' = \overline{A}$, $B' = \overline{B}$, and $C' = \overline{C}$.

1.3 PLAUSIBLE REASONING

Notation:

- $\pi(A|B)$ = "conditional plausibility that A is true, given that B is true" = plausibility (π) of

1.3.1 *Desiderata*

The derivation of probability from logic rests on three desiderata:

I	Degrees of plausibility are represented by real numbers.

By convention (infinitesimally) larger plausibilities are represented by (infinitesimally) larger numbers:

$$C = \text{“}A \text{ is more plausible than } B\text{”}$$
$$\Rightarrow \pi(A|C) > \pi(B|C) \;\&\; \pi(\overline{A}|C) < \pi(\overline{B}|C)$$

If information D gets updated to D' with $\pi(A|D') > \pi(A|D)$ and $\pi(B|AD') = \pi(B|AD)$:

$$\Rightarrow \pi(AB|D') \geq \pi(AB|D) \;\&\pi(\overline{A}|D') < \pi(\overline{A}|D)$$

II	Qualitative correspondence with common sense.
	1. Aristotelian logic should be included.
III	Self consistency of the plausibility value system:
	1. If a conclusion can be reasoned in several ways, their results must agree.
	2. Equivalent knowledge states are represented by equivalent plausibilities.
	3. All available information must be included in any reasoning.

1.3.2 *The product rule*

The plausibility of the product statement $AB|C$ = "A and B given C" can be decomposed in two different ways:

1. a) Decide whether B is true under C by specifying $\pi(B|C)$
   b) If this is the case, decide if A is also true $\pi(A|BC)$.
2. a) Decide whether A is true under C by specifying $\pi(A|C)$
   b) Given A, decide if B is also true $\pi(B|AC)$

From III.1 we expect both ways of reasoning to lead to the same conclusion on the plausibility of $AB|C$. This means, there must be a plausibility function $f(x,y) = z$, which fulfills

$$\pi(AB|C) = f(\pi(B|C),\, \pi(A|BC)) = f(\pi(A|C),\, \pi(B|AC)). \tag{1}$$

Furthermore, by the convention below desideratum I (or by desideratum II) we expect $f(x,y)$ to be continuous and monotonic in both x, y.
By a similar decomposition of the triple-*and* statement $ABC|D$ one can show that

$$f(f(x,y),z) = f(x, f(y,z)). \tag{2}$$

From this, Cox [3] showed that there is a new, transformed plausibility system ω in which the logical product ("and") becomes an ordinary product:

$$\omega(f(x,y)) = \omega(x)\,\omega(y) \text{ or } f(x,y) = \omega^{-1}(\omega(x)\,\omega(y)). \tag{3}$$

This leads to the product rule for the new plausibilities

$$\omega(AB|C) = \omega(A|BC)\,\omega(B|C) = \omega(B|AC)\,\omega(A|C). \tag{4}$$

1.3.3 *True and false*

1. Assume "A certain given C" = "$C \Rightarrow A$" = "$C = AC$"
   $\Rightarrow$(i)$AB|C = B|C$, because requesting A does not change the plausibility of $B|C$, since A is given under C.
   $\Rightarrow$(ii)$A|BC = A|C$
   Using (i), (ii) and the product rule we find the value for true:

$$\omega(B|C) = \omega(AB|C) = \omega(A|BC)\,\omega(B|C) = \omega(A|C)\,\omega(B|C) \Rightarrow \omega(A|C) = 1$$

2. Assume "A is impossible, given C" = "$C \Rightarrow \overline{A}$" = "$C = \overline{A}C$"
$\Rightarrow$(iii)$AB|C = A|C$
$\Rightarrow$(iv)$A|BC = A|C$
Using (iii), (iv) and the product rule we find the values for false

$$\omega(A|C) = \omega(AB|C) = \omega(A|BC)\,\omega(B|C) = \omega(A|C)\,\omega(B|C) \Rightarrow \omega(A|C) = \begin{cases} 0 \\ \infty \end{cases} .$$

In this case $-\infty$ as a solution of $\omega(A|C)$ is ruled out by the special case $A = B$.

There are two possibilities of choosing ω

- $\omega \in [0,1]$ expressing plausibilities
- $\omega' \in [1,\infty]$ expressing implausibilities

related by $\omega = \frac{1}{\omega'}$.
Convention:
$\omega \in [0,1]$ with $\omega(A|B) = 0$ expressing impossibility of A given B and
$\omega(A|B) = 1$ expressing certainty of A given B.

1.3.4 *Negation*

Aristotelian logic:

- A is either true or false
- $A\overline{A}$ is always false
- $A + \overline{A}$ is always true

The negation function $S : [0,1] \to [0,1]$ fulfilling

$$\omega(\overline{A}|B) = S(\omega(A|B)), \tag{5}$$

is monotonically decreasing with the boundary conditions $S(0) = 1$ and $S(1) = 0$. Due to consistency, this function must be of the form [3]

$$S(x) = (1 - x^m)^{1/m} \qquad x \in [0,1],\ 0 < m < \infty. \tag{6}$$

The parameter m is arbitrary and labels the different possible plausibility systems.
$\Rightarrow \omega(\overline{A}|B) = S(\omega(A|B)) = (1 - \omega^m(A|B))^{1/m}$

- sum rule: $\omega^m(\overline{A}|B) + \omega^m(A|B) = 1$
- product rule: $\omega^m(AB|C) = \omega^m(A|BC)\,\omega^m(B|C) = \omega^m(B|AC)\,\omega^m(A|C)$.

1.4 PROBABILITY

In the following, we choose a linear plausibility system with the exponent $m = 1$. These plausibilities we call probabilities

$$P(x) = \omega^m(x). \tag{7}$$

1.4.1 *Probability systems*

For probabilities, $P(A|B)$ ="probability of A given B", product and sum rule are particularly simple:

product rule and sum rule

$$\boxed{\begin{array}{rll} \text{product rule:} & P(AB|C) = P(A|BC)\,P(B|C) = P(B|AC)\,P(A|C) & (8) \\ \text{sum rule:} & P(A|B) + P(\overline{A}|B) = 1 & (9) \end{array}}$$

Probabilities can be based on

- logic (extended to uncertainty)
- relative frequencies of events (frequentist definition)

$$P(\text{specific event} \mid \text{generic event}) = \lim_{n\to\infty} \frac{n(\text{specific event})}{n(\text{generic event})}, \tag{10}$$

- set theoretical considerations (Kolmogorov system), or
- considerations on consistent bet ratios (de Finetti approach).

1.4.2 *Marginalization*

Marginalization removes the dependence of a probability $P(A, B|C)$ on the statement B. (i)

$$\begin{aligned} P(A,B|C) &= P(B|AC)\,P(A|C) \\ P(A,\overline{B}|C) &= P(\overline{B}|AC)\,P(A|C) \\ \Rightarrow\ P(A,B|C) + P(A,\overline{B}|C) &= \underbrace{\left[P(B|AC) + P(\overline{B}|AC)\right]}_{1} P(A|C) = P(A|C) \end{aligned}$$

$P(A|C) = P(A, B|C) + P(A, \overline{B}|C)$ is called the "B-marginalized probability of A". (Note the change in notation for the "and": $AB \equiv A, B$) (ii) The marginalization can be generalized to more than two options B and $\overline{B}$. Let $\{B_i\}_{i=1}^n$ be a set of n mutually exclusive ($P(B_iB_j|I) = 0$ for $i \neq j$) and exhaustive ($P(B_1 + \ldots + B_n|I) = 1$) possibilities in I, then

mutually exclusive

exhaustive

$$P(A|I) = \sum_{i=1}^{n} P(A, B_i|I) \tag{11}$$

is the B-marginalized probability of A under I.
Shortcut notation:

marginalized probability

- $P(A) = P(A|I)$
- $P(A|B) = P(A|BI)$

if the context I is either clear or unimportant.
Warning: if several contexts are present, they should be clearly marked, since otherwise confusion is guaranteed.

1.5 PROBABILISTIC REASONING

The generalized sum rule describes how two probabilities of non-exclusive and non-exhaustive statements can be added.

generalized sum rule

$$\boxed{\text{generalized sum rule:} \quad P(A+B) \quad = P(A) + P(B) - P(AB). \qquad (12)}$$

By using Bayes' theorem the "posterior" probability $P(A|B)$ of an original cause A given the observed event (the data) B can be calculated.

Bayes' theorem

$$\boxed{\text{Bayes' theorem:} \qquad P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}. \qquad (13)}$$

Here, $P(A)$ is the "prior" probability of A, $P(B|A)$ is the "likelihood" describing the forward probability of the causal process and the "evidence" $P(B)$ is just a normalization constant.

1.5.1 *Deductive logic*

Now we can check whether probabilistic reasoning already contains the syllogisms of Aristotelian logic.

- **strong syllogism:** $I =$ "$A \Rightarrow B$" $\Rightarrow$(i) $P(B|AI) = 1$ & (ii) $P(A|\overline{B}I) = 0$
  proof:
  (i) "$A \Rightarrow B$" is actually "$A = AB$" $\Rightarrow P(AB|I) = P(A|I)$

  $$P(B|AI) \quad = \quad \frac{P(AB|I)}{P(A|I)} = 1$$

  (ii)

  $$P(A|\overline{B}I) = \frac{P(A\overline{B}|I)}{P(\overline{B}|I)} = \frac{P(AB\overline{B}|I)}{P(\overline{B}|I)} = 0$$

  unless $P(\overline{B}|I) = 0$, which would turn the r.h.s. condition into an empty statement.

- **weak syllogism:** $I =$ "$A \Rightarrow B$ " $\Rightarrow P(A|BI) \geq P(A|I)$
  proof: From the strong syllogism we already know $P(B|AI) = 1$.

  $$P(A|BI) = \frac{P(B|AI)\,P(A|I)}{P(B|I)} = \frac{P(A|I)}{P(B|I)} \geq P(A|I)$$

  In the last step the triviality $P(B|I) \leq 1$ was used.

- **weaker syllogism:** $J =$ "$B \Rightarrow A$ more plausible under I" $=$ "$P(A|BI) > P(A|I)$"

claim: $J \Rightarrow J' =$ "A $\Rightarrow B$ more plausible under I"= "$P(B|AI) > P(B|I)$"
proof:

$$P(B|AI) = \underbrace{\frac{P(A|BI)}{P(A|I)}}_{>1} P(B|I) > P(B|I) \Bigg| J$$

1.5.2 *Assigning probabilities*

- I is the background information or proposition, $A_1, \ldots A_n$ is the set of mutually exclusive possibilities which exhaust I. $\Rightarrow$ "one and only one A_i with $i \in \{1, \ldots$ n$\}$ is true" and $\sum_{i=1}^{n} P(A_i|I) = 1$.
- Assume that the knowledge in I about $A_1, \ldots A_n$ is absolutely symmetric. $\Rightarrow P(A_i|I) = P(A_j|I)$ and a uniform distribution is assigned.

uniform probability distribution

$$\boxed{\text{uniform probability distribution:} \qquad P(A_i|B) = \frac{1}{n}. \qquad (14)}$$

This is Laplace's principle of the insufficient reason.
Canonical examples:

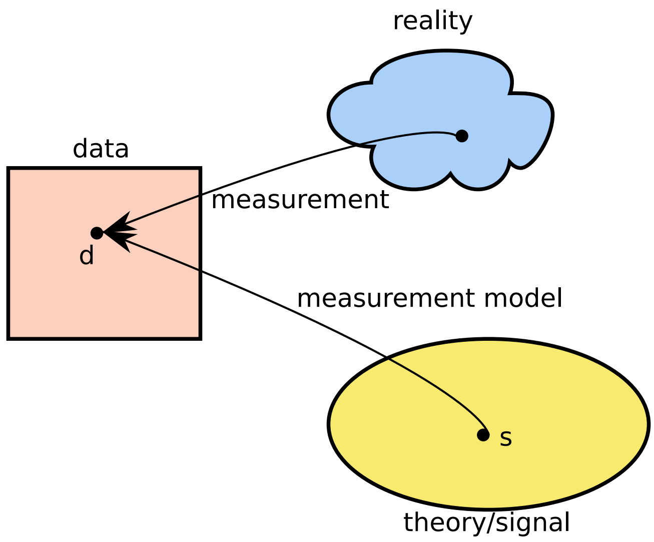
- fair die: $P(\boxdot|\text{fair die}) = \frac{1}{6}$
- loaded die: $P(\boxdot|\text{loaded die}) = \frac{1}{6}$, but $P(\boxdot|\text{previous results, loaded die})$ may differ from 1/6 depending on previous results

$\Rightarrow$ Conditional probabilities describe learning from data.

1.6 STATISTICAL INFERENCE

1.6.1 *Measurement process*

In the inference process the causality between the real physical state and the data should be inverted.



- The measurement process maps the real state into the data space.
- We only have a theory, a simplified model, to describe reality.
- The theory has unknown parameters, the signal s, which shall be determined by the data d from a measurement described via $P(d|s)$.

Potential problems:

- Theory might be insufficient to describe relevant aspects of reality.
- Measurement and theory differ too much (e.g. device was broken).
- Data is not uniquely determined (knowledge on resulting distribution is given by $P(d|s)$).
- Signal is not uniquely determined.

1.6.2 *Bayesian Inference*

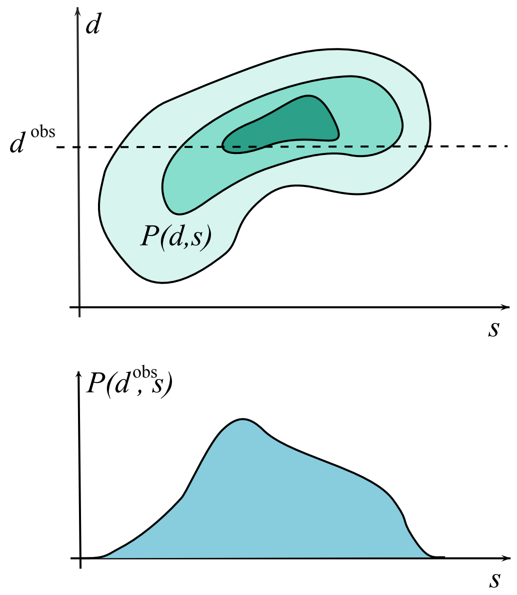
I is the background information on signal s, on measurement process producing data d. In the following I is assumed to be implicitly included among the conditionals of any probability.
Bayes' theorem permits us to construct from prior $P(s)$ and likelihood $P(d|s)$ the posterior $P(s|d)$, which describes the signal knowledge after the measurement.
Bayes' theorem:

$$P(s|d) = \frac{P(d, s)}{P(d)} = \frac{P(d|s)\,P(s)}{P(d)} \tag{15}$$

- Note the sloppy notation: $P(s)$ means the probability of the variable s to have the value s given the implicit background information I, $P(s_{\text{var}} = s_{\text{val}}|I)$, with s_{var} the unknown variable and s_{val} a concrete value.
- The joint probability $P(d, s)$ is decomposed in likelihood and prior.

- The prior $P(s)$ summarizes our knowledge on s before measuring.
- The likelihood $P(d|s)$ describes the measurement process.
- The evidence $P(d) = \sum_s P(d, s)$ serves as normalization factor.

$$\sum_s P(s|d) = \sum_s \frac{P(d,s)}{P(d)} = \frac{\sum_s P(d,s)}{\sum_{s'} P(d,s')} = 1 \tag{16}$$



- With the measurement of the data d^{obs} only the hyperplane with $d = d^{\text{obs}}$ is relevant any more. Any deduction depending on unobserved data $d^{\text{mock}} \neq d^{\text{obs}}$ is suboptimal, inconsistent, or just wrong.
- The normalization of the restricted probability $P(d = d^{\text{obs}}, s)$ is given by the area under the curve: $\sum_s P(d^{\text{obs}}, s) = P(d^{\text{obs}})$.

1.7 COIN TOSSING

1.7.1 *Recognizing the unfair coin*

$I_1 =$ "A large number of coin tosses are performed and the outcomes are stored in a data vector $d = (d_1, d_2, \ldots)$, with $d_i \in \{0, 1\}$ representing head ($d_i = 1$) or tail ($d_i = 0$) for the i-th toss. The data up to toss n is denoted by $d^{(n)} = (d_1, \ldots d_n)$. "

QUESTION 1: What is our knowledge on $d^{(1)} = (d_1)$ given I_1?
Due to symmetry in knowledgethe probability distribution is given:

$$\begin{aligned} P(d_1 = 0|I_1) &= P(d_1 = 1|I_1) && (17)\\ \Rightarrow 1 &= P(d_1 = 0|I_1) + P(d_1 = 1|I_1) && (18)\\ \Rightarrow \frac{1}{2} &= P(d_1|I_1) && (19) \end{aligned}$$

QUESTION 2: What is our knowledge about d_{n+1} given $d^{(n)}$, I_1?
With $d^{(n+1)} = (d_{n+1}, d^{(n)})$ we get

$$P(d_{n+1}|d^{(n)}, I_1) = \frac{P(d^{(n+1)}|I_1)}{P(d^{(n)}|I_1)}. \tag{20}$$

Given our knowledge I_1 we have no reason to favor any of the 2^n possible sequences $d^{(n)} \in \{0, 1\}^n$ of length n and have assign symmetric probabilities to them: $P(d^{(n)}|I_1) = 2^{-n}$

$$P(d_{n+1}|d^{(n)}, I_1) = \frac{2^{-n-1}}{2^{-n}} = \frac{1}{2}. \tag{21}$$

statistical independence

It seems that $I_1 \Rightarrow$ "All tosses are **statistically independent** of each other."
Two events A and B are statistically independent of each other under some information C if knowing B does not change the probability for A, $P(A|BC) = P(A|C)$, and vice versa. This implies, that their joint probability is just the direct product of their individual probabilities,

$$P(AB|C) = P(A|BC)\,P(B|C) = P(A|C)\,P(B|C).$$

$\Rightarrow$ Given I_1, the data $d^{(n)}$ contains no useful information on d_{n+1}. The probability has not changed. What did we miss? Something that connects the different tosses without making them explicitly dependent of each other, a shared, but hidden property.

Additional information $I_2 =$ "All tosses are done with the same coin. We highly suspect the coin to be loaded, meaning that heads occur with a frequency $f \in [0, 1]$": $\exists f \in [0, 1] : \forall i \in \mathbb{N} : \; P(d_i = 1|f, I_1, I_2) = f$. $I = I_1 I_2$, then

$$P(d_i|f, I) = \begin{cases} f & d_i = 1 \\ 1 - f & d_i = 0 \end{cases} = f^{d_i}\,(1 - f)^{1-d_i}. \tag{22}$$

QUESTION 3: What do we know about f given I and our data $d^{(n)}$ after n tosses?
We have developed probability theory so far only for discrete possibilities, but f is a continuous parameter, for which we have to extend probability theory.

1.7.2 *Probability density functions*

NOTATION: $P(f \in F|I)$ with $F \subset \Omega$. In the above case $\Omega = [0, 1]$.
We would expect the probability $P(f \in F|I)$ to be monotonically increasing with $|F| = \int_F df\, 1$, since as more possibilities are included in F the probability for it should be larger. We require $P(f \in \Omega|I) = 1$. If no value $f \in \Omega$ given I is favored, we request

$$P(f \in F|I) = \frac{|F|}{|\Omega|} = \frac{\int_F df\, 1}{\int_\Omega df\, 1}. \tag{23}$$

If a non-uniform weight distribution $w : \Omega \mapsto \mathbb{R}_0^+$ should be considered, we use $|F|_w = \int_F df\, w(f)$ and therefore

$$P(f \in F|I) = \frac{|F|_w}{|\Omega|_w} = \frac{\int_F df\, w(f)}{\int_\Omega df\, w(f)} = \int_F df\, \mathcal{P}(f|\, I) \tag{24}$$

$\mathcal{P}(f|\, I) = w(f)/|\Omega|_w$ is called probability density function (PDF).

probability density function

NORMALIZATION: $P(f \in \Omega|\, I) = \int_\Omega df\, \mathcal{P}(f|\, I) = 1$

TRANSFORMATION: A coordinate transformation $T : f \to f'$ can turn a uniform PDF $\mathcal{P}(f|\, I)$ into a non-uniform PDF $\mathcal{P}(f'|\, I)$ and vice verse. From the coordinate in-variance of the probabilities $P(f \in F|I) = P(f' \in F'|I)$ with $F' = T(F)$ it follows that

$$\int_F df\, \mathcal{P}(f|\, I) = \int_{F'} df'\, \mathcal{P}(f'|\, I) \tag{25}$$

for all sets $F \subset \Omega$, and therefore

$$\mathcal{P}(f'|\, I) = \mathcal{P}(f|\, I) \left\| \frac{df}{df'} \right\|_{f=T^{-1}(f').} \tag{26}$$

The Jacobian does not need to be uniform. Choosing a uniform prior for a PDF therefore requires first to identify the natural coordinate system.

BAYES' THEOREM: $I =$ "Let $x \in \mathbb{R}$ and $y \in \mathbb{R}$" and $\mathcal{P}(x, y|I)$ their joint PDF, *i.e.* such that $P(x \in X,\, y \in Y|I) = \int_X dx \int_Y dy\, \mathcal{P}(x, y|I)$ for any $X, Y \subset \mathbb{R}$. Then we can define the marginal and conditional PDFs, respectively,

conditional probability density function

$$\mathcal{P}(x|I) = \int dy\, \mathcal{P}(x, y|I), \tag{27}$$

$$\mathcal{P}(y|I) = \int dx\, \mathcal{P}(x, y|I), \tag{28}$$

$$\mathcal{P}(x|y, I) = \frac{\mathcal{P}(x, y|I)}{\mathcal{P}(y|I)}, \tag{29}$$

$$\mathcal{P}(y|x, I) = \frac{\mathcal{P}(x, y|I)}{\mathcal{P}(x|I)}, \tag{30}$$

such that the product rule holds,

$$\mathcal{P}(x, y|I) = \mathcal{P}(x|y, I)\mathcal{P}(y|I) = \mathcal{P}(y|x, I)\mathcal{P}(x|I), \tag{31}$$

from which Bayes' theorem for PDFs follows. It remains to be shown that the quantities defined above are indeed PDFs, that these encode the corresponding probabilities. For the y-marginalized PDF we find that this is the case,

$$\begin{aligned} P(x \in X|I) &\stackrel{?}{=} \int_X dx\, \mathcal{P}(x|I) = \int_X dx \int_{\mathbb{R}} dy\, \mathcal{P}(x, y|I) = P(x \in X, y \in \mathbb{R}|I) \\ &= P(x \in X|I) \end{aligned} \tag{32}$$

as $I \Rightarrow y \in \mathbb{R}$. Similarly, $P(y \in Y|I) = \int_Y dy\, \mathcal{P}(y|I)$. For the conditional PDF, *e.g.* for x conditioned on y (more precisely, on the statement $y_{\text{var}} = y_{\text{val}}$), we observe

$$P(x \in X|y, I) \stackrel{?}{=} \int_X dx\, \mathcal{P}(x|y, I) = \int_X dx\, \frac{\mathcal{P}(x, y|I)}{\mathcal{P}(y|I)} = \frac{\int_X dx\, \mathcal{P}(x, y|I)}{\int_{\mathbb{R}} dx\, \mathcal{P}(x, y|I)} = \frac{|X|_{\mathcal{P}(x,y|I)}}{|\mathbb{R}|_{\mathcal{P}(x,y|I)}} \tag{33}$$

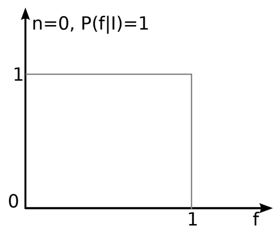
to follow exactly the spirit of a weighted measure ratio, as used above to introduced PDFs.

Note that a given PDF $\mathcal{P}(x,y)$ uniquely defines all probabilities $P(x \in X, y \in Y)$ for the continuous quantities $x, y \in \mathbb{R}$, however, the reverse is not necessary true. Any zero-measure function $\mathcal{B}(x,y)$, with $\int_X dx \int_Y dy\, \mathcal{B}(x,y) = 0$ for $\forall X, Y \subset \mathbb{R}$ can be added to $\mathcal{P}(x,y) \to \mathcal{P}'(x,y) = \mathcal{P}(x,y) + \mathcal{B}(x,y)$ without changing the resulting $P'(x \in X, y \in Y) = \int_X dx \int_Y dy\, \mathcal{P}'(x,y)$, but affecting conditional probabilities as defined in Eq. (33). *E.g.* if $\mathcal{B}(x,y) \neq 0$ for some $y = y_{\text{val}}$, but otherwise $\mathcal{B}(x,y) = 0$, $P'(x \in X | y = y_{\text{val}}, I) \neq P(x \in X | y = y_{\text{val}}, I)$.

1.7.3 *Inferring the coin load*

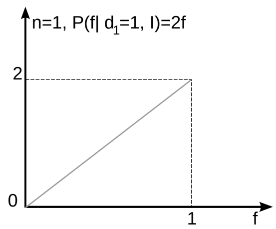
Back to question 3:

- $n = 0$: $\mathcal{P}(f, I)$ is independent of f.



- $n = 1$, $d_1 = 1$:

$$\mathcal{P}(f|d_1 = 1,\, I) = \frac{\mathcal{P}(d_1 = 1|f,\, I)\mathcal{P}(f|I)}{\int_0^1 df\, \mathcal{P}(d_1 = 1|f,\, I)\mathcal{P}(f|I)} = \frac{f}{\int_0^1 df\, f} = \frac{f}{1/2} = 2f$$



$\Rightarrow \mathcal{P}(f = 0 \mid d_1 = 1,\, I) = 0$, a coin which does not show heads ($f = 0$) can now be excluded with certainty, as a head has been observed.

- arbitrary n: $\Rightarrow$ Usage of Bayes' theorem and independence:

$$\mathcal{P}(f|d^{(n)}, I) \quad = \quad \frac{\mathcal{P}(d^{(n)}|f, I)\, \mathcal{P}(f, I)}{\mathcal{P}(d^{(n)}|I)} = \frac{\mathcal{P}(d^{(n)}, f|I)}{\mathcal{P}(d^{(n)}|I)} \tag{34}$$

$$\mathcal{P}(d^{(n)}, f|I) = \prod_{i=1}^{n} \mathcal{P}(d_i|f, I) \times 1 = \prod_{i=1}^{n} f^{d_i}\,(1-f)^{1-d_i} = f^{n_1}\,(1-f)^{n_0} \tag{35}$$

Number of heads in $d^{(n)}$: $n_1 = n_1(d^{(n)}) = \sum_{i=1}^{n} d_i$
Number of tails in $d^{(n)}$: $n_0 = n_0(d^{(n)}) = \sum_{i=1}^{n}(1 - d_i) = n - n_1$.

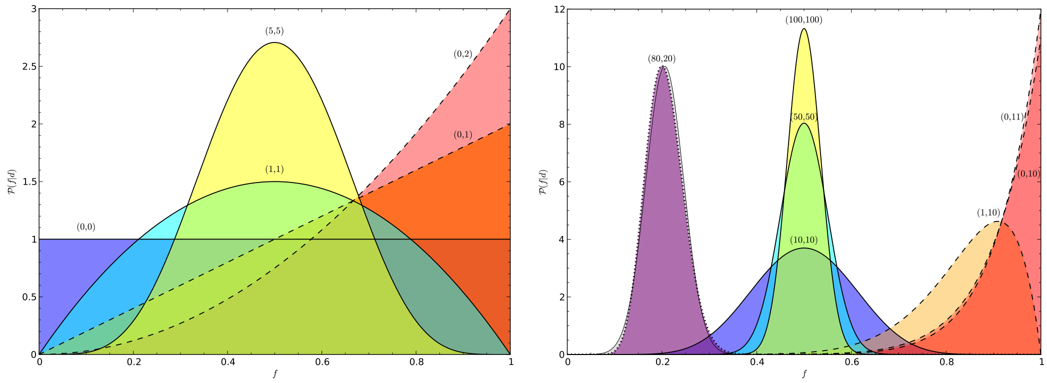


Figure 1: The posterior $\mathcal{P}(f|d^{(n)}, I)$ of the coin load parameter f for different data realizations, as marked with $(n_0, n_1) = \sum_{i=1}^{n}(1-d_i, d_i) =$ (# of tails, # of heads). **Left:** The first few tosses, with an equal number of heads and tails marked by solid lines and a preference for heads marked by dashed lines. **Right:** Situations with 10 to 200 tosses. Solid and dashed lines as before, dotted line for a case with a preference for tails. The Gaussian approximation of the posterior by (43) is shown by a thin solid line with grey filling for the case $(80, 20)$.

The prior $\mathcal{P}(f) = 1$ is uniform.
Calculate evidence of I, $\mathcal{P}(d^{(n)}|I)$, by marginalizing $\mathcal{P}(d^{(n)}, f|I)$,

$$\mathcal{P}(d^{(n)}|I) = \int_0^1 df\, \mathcal{P}(d^{(n)}, f|I) = \int_0^1 df\, f^{n_1}(1-f)^{n_0} \tag{36}$$

Integral via definition of beta function

$$\mathcal{B}(a,b) = \int_0^1 dx\, x^{a-1}(1-x)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \frac{(a-1)!\,(b-1)!}{(a+b-1)!}. \tag{37}$$

By comparison of the former equations, we derive $a = n_1 + 1$ and $b = n_0 + 1$, resulting in

$$\mathcal{P}(d^{(n)}) = \frac{n_0!\, n_1!}{(n+1)!}, \quad n = n_0 + n_1. \tag{38}$$

From Eqs. (35) & (38) we get the posterior,

$$\boxed{\mathcal{P}(f|d^{(n)}, I) \;=\; \frac{P(d^{(n)}, f|I)}{P(d^{(n)}|I)} = \frac{(n+1)!}{n_1!\, n_0!} f^{n_1}(1-f)^{n_0}. \qquad (39)}$$

$n = 2$:

$$\mathcal{P}(f|d^{(2)}) = \begin{cases} 3f^2 & , n_1 = 2 \\ 6f\,(1-f) & , n_1 = 1\, , \\ 3(1-f)^2 & , n_1 = 0 \end{cases} \tag{40}$$

The posterior density functions $\mathcal{P}(f|d^{(n)}, I)$ depending on f for different parameters n, n_1, n_2 are shown in Figure 1. The figures demonstrate that the probability density function gets more and more peaked with a growing number of tosses.

After all, what do we know about d_{n+1} given $d^{(n)}$ and I? Let's look first at the case $d_{n+1} = 1$

$$\begin{aligned}
P(d_{n+1} = 1|d^{(n)}, I) &= \int_0^1 df\, P(d_{n+1} = 1, f|\, d^{(n)}, I) \\
&= \int_0^1 df\, P(d_{n+1} = 1|f,\, d^{(n)}, I)\, \mathcal{P}(f|d^{(n)}, I) \\
&= \int_0^1 df\, f\, \mathcal{P}(f|d^{(n)}, I) \\
&= \frac{(n+1)!}{n_1!\, n_0!} \int_0^1 df\, f^{n_1+1}(1-f)^{n_0} \\
&= \frac{(n+1)!}{n_1!\, n_0!} \frac{(n_1+1)!\, n_0!}{(n+2)!} \\
&= \frac{n_1+1}{n+2}, \qquad (41) \\
P(d_{n+1} = 0|d^{(n)}, I) &= \frac{n_0+1}{n+2} \qquad (42)
\end{aligned}$$

Laplace's rule of succession

which means that the probability of the next toss being head is the mean value of the posterior $\mathcal{P}(f|d^{(n)})$, i.e., $P(d_{n+1} = 1|d^{(n)}) = \int_0^1 df\, f\, \mathcal{P}(f|d^{(n)}) \equiv \langle f \rangle_{(f|d^{(n)})}$.

1.7.4 *Large number of tosses*

Figure 1 shows that $\mathcal{P}(f|d^{(n)})$ typically looks Gaussian for a sufficiently large number of detected heads and tails (Central limit theorem). The width of this distribution gets smaller with increasing data size.

- Mean:

$$\begin{aligned}
\overline{f} &= \langle f \rangle_{(f|d^{(n)}, I)} \\
&= \int_0^1 df\, f\, \mathcal{P}(f|d^{(n)},\, I) \\
&= \frac{n_1+1}{n+2}
\end{aligned}$$

- Variance:

$$\begin{aligned}
\sigma_f^2 &= \langle (f - \overline{f})^2 \rangle_{(f|d^{(n)})} = \langle f^2 - 2\overline{f}f + \overline{f}^2 \rangle_{(f|d^{(n)})} = \langle f^2 \rangle_{(f|d^{(n)})} - \overline{f}^2 \\
&= \frac{(n_1+2)\,(n_1+1)}{(n+3)\,(n+2)} - \left(\frac{n_1+1}{n+2}\right)^2 = \frac{\overline{f}(1-\overline{f})}{n+3}.
\end{aligned}$$

$\Rightarrow$ The width/uncertainty decreases with $\sigma_f \sim 1/\sqrt{n}$.

- Gaussian approximation (only good for f, $\overline{f}$ sufficiently far away from 0 and 1:

$$\mathcal{P}(f|d^{(n)},\, I) \approx \mathcal{G}(f - \overline{f}, \sigma_f^2) = \frac{1}{\sqrt{2\pi\sigma_f^2}} \exp\left(-\frac{(f-\overline{f})^2}{2\sigma_f^2}\right) \qquad (43)$$

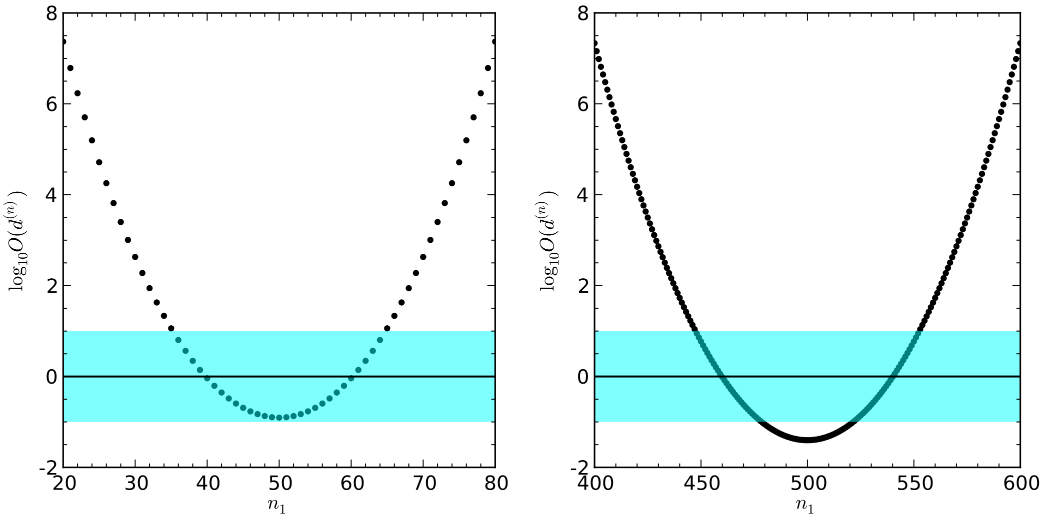


Figure 2: Odds for bets on the coin being loaded versus being fair for 100 tosses (left) and 1000 tosses (right) on a logarithmic scale as a function of the number of observed heads. The region with larger evidence for a loaded (fair) coin is above (below) the horizontal line. The undecided region around this line with odds between 1:10 and 10:1 is shaded in cyan. Less than 35 or more than 65 out of 100 tosses should be heads, before a loaded coin should be claimed (with a confidence of 10:1). A fair coin can never be claimed with such a confidence after observing only 100 tosses. However, 1000 tosses with well balanced outcomes can be sufficient for confidence in the fairness of the coin.

1.7.5 *The evidence for the load*

$I =$ "a loaded coin with, $f \in [0,1] \backslash \{\frac{1}{2}\}$"
$J =$ "a fair coin with $f = \frac{1}{2}$"
$M = I + J$
$P(I|M) = P(J|M) = 1/2$ are the hyper-priors for the hypotheses. As a discriminating quantity between the two scenarios we can regard their a posteriori odds, *odds*

$$\begin{aligned} O(d^{(n)}) &\equiv \frac{P(I|d^{(n)}, M)}{P(J|d^{(n)}, M)} \\ &= \frac{P(d^{(n)}|I, M)\, P(I|M)/P(d^{(n)}|M)}{P(d^{(n)}|J\, M)\, P(J|M)/P(d^{(n)}|M)} \\ &= \frac{P(d^{(n)}|I, M)}{P(d^{(n)}|J, M)}. \end{aligned}$$

Evidences:

- loaded coin:

$$P(d^{(n)}|I) = \frac{n_1!\, n_0!}{(n+1)!}$$

- fair coin:

$$P(d^{(n)}|J) = \frac{1}{2^n}$$

Thus, the odds of our hypotheses are

$$O(d^{(n)}) = \frac{2^n\, n_1!\, n_0!}{(n+1)!}. \tag{44}$$

Example for only heads:

$n_1 = n$	0	1	2	3	4	5	6	7	8	9	10	100	1000
$O(d^{(n)})$	1	1	4/3	2	3 1/5	5 1/3	9 1/7	16	28 4/9	51 1/5	93 1/11	$\approx 10^{28.1}$	$\approx 10^{298}$

1.7.6 *Lessons learned*

1. **Background information matters:** $P(d_{n+1}|d^{(n)}, I_1) \neq P(d_{n+1}|d^{(n)}, I_1 I_2)$, if $I_2 \not\subseteq I_1$
2. **Models are mandatory for intelligence:** E.g. model of a coin having the property of a constant head frequency f.
3. **Probability density functions** (PDFs) times a volume measure are probabilities, therefore PDFs follow Bayes theorem.
4. **Learning & forgetting:** Posterior changes with new data and usually becomes sharply peaked with large amounts of data.
5. **Sufficient statistics** are compressed data, which still gives the same information on the quantity of interest as the original data, e.g. only the number of heads and tails are relevant, but not their order: $P(f|d^{(n)}, I) = P(f|n_1, n_0, I)$.
6. **Probabilities**: Knowledge states are described by probabilities.
7. **Frequencies:** Probabilities and frequencies are in general different concepts, but in case of known frequencies they coincide, $P(d = 1|f, I) = f$.
8. **Joint probability:** All relevant information is contained in the joint probability of data and signal. Likelihood, prior, evidence, and posterior are different normalized cuts and marginalizaitons of it.
9. **Posterior:** The knowledge of the signal given the data and all model assumptions.
$$P(f|d, I) = \frac{P(d, f|I)}{\int df\, P(d, f|I)}$$
10. **Evidence:** The signal-marginalized joint probability. It is also the "likelihood" of the model.
11. **Nested models** are models where one contains the other and becomes identical to it, when some of its parameters take a specific value. The fair coin model is nested in the model of the unfair coin of unknown bias, as $f \to 1/2$ reproduces it.
12. **Occam's razor:** *Among competing hypotheses, the one with the fewest assumptions should be selected.* In a maximum likelihood comparison of nested models, however, the more complex will always win or be equal to the simpler model. The Bayesian odds ratio does not fall into this pitfall and has Occam's razor build in.

13. **Uncertainty:** The uncertainty of an inferred quantity depends in general on the data realization obtained.

1.8 ADAPTIVE INFORMATION RETRIEVAL

How to infer from adaptively taken data, in which the last outcome determines the next measurement action?

1.8.1 *Inference from adaptive data retrieval*

Data $d^{(n)} = (d_1, \ldots d_n)$ taken to infer a signal s was obtained sequentially. Let a_i, the action chosen to measure d_i via

$$d_i \hookleftarrow P(d_i|a_i, s), \tag{45}$$

depend on previously measured data through the data retrieval strategy function $A : d^{(i-1)} \to a_i$.

- A **predetermined strategy** is independent of the prior data $\Rightarrow A(d^{(i-1)}) \equiv a_i$ irrespective of $d^{(i-1)}$
- An **adaptive strategy** depends on the data: $\exists i, d^{(i-1)}, d'^{(i-1)} \; : \; A(d^{(i-1)}) \neq A(d'^{(i-1)})$

Thus, a new datum d_i depends conditionally on the previous data $d^{(i-1)}$ through strategy A,

$$P(d_i|a_i, s) = P(d_i|A(d^{(i-1)}), s) = P(d_i|d^{(i-1)}, A, s). \tag{46}$$

The likelihood of the full data set $d = d^{(n)}$ is

$$P(d|A, s) = P(d_n|d^{(n-1)}, A, s) \cdots P(d^{(1)}|A, s) = \prod_{i=1}^{n} P(d_i|d^{(i-1)}, A, s). \tag{47}$$

If we had used a different strategy B, we probably would have gotten different data, as the set of actions might have diverged.
It is however possible that the actual sequence of actions $a = (a_1, \ldots a_n)$ could have been the result of a different strategy B, e.g. the predetermined strategy $B(d^{(i)}) \equiv a_i$ that happens to coincide to A for the actual data observed (but not necessarily for other data realizations).

likelihood:

$$\begin{aligned} P(d|A, s) &= \prod_{i=1}^{n} P(d_i|A(d^{(i-1)}), s) = \prod_{i=1}^{n} P(d_i|a_i, s) && (48) \\ &= \prod_{i=1}^{n} P(d_i|B(d^{(i-1)}), s) = P(d|B, s), && (49) \end{aligned}$$

posterior:

$$\begin{aligned} P(s|d, A) &= \frac{P(d|A, s)P(s|A)}{P(d|A)} = \frac{P(d|A, s)P(s)}{P(d|A)} && (50) \\ &= \frac{P(d|A, s)P(s)}{\sum_s P(d|A, s)P(s)} = \frac{P(d|B, s)P(s)}{\sum_s P(d|B, s)P(s)} && (51) \\ &= P(s|d, B) && (52) \end{aligned}$$

Used assumption: $P(s|A) = P(s)$

⇒ Bayesian signal deduction does not depend on why some data was taken, only on how it was taken and what it was:

$$P(s|d, A) = P(s|d, B) \tag{53}$$

if A, B strategies that provide the same set of actions for the observed data: $A(d^{(i)}) = B(d^{(i)}) = a_i$ and the signal is independent of the strategy, $P(s|A) = P(s)$.

Corollary: A sequence of interdependent observations (= actions and resulting data) is open to a Bayesian analysis without knowledge of the used strategy function. A frequentist analysis, which depends on all possible data realizations, not only the observed ones, needs to fully know the strategy function, as this affects the likelihood of all possible data realizations.
⇒ A history (= record of a sequence of interrelated actions and consequences) is a valid information source for Bayesians inference, but nearly useless for frequentists analysis as it does not report what would have happened if some datum would have been different.

1.8.2 *Adaptive strategy to maximize evidence*

Can spurious evidence be created for a false hypothesis I, against the right hypothesis J? We might ask for $O(d^{(n)}) = P(I|d^{(n)}) : P(J|d^{(n)}) = 10 : 1 \gg 1$ to claim J to be proven! Can this be made more likely by tuning the strategy?
Odds:

$$\begin{aligned} O(d^{(n)}) &= \frac{P(I|d^{(n)})}{P(J|d^{(n)})} && (54) \\ &= \frac{P(d^{(n)}|I)P(I)}{P(d^{(n)}|J)P(J)} && (55) \end{aligned}$$

The expectation value of the odds against the correct hypothesis, averaged over the outcomes of possible data realizations $d = d^{(n)}$ given an observing strategy A

$$\begin{aligned} \langle O(d)\rangle_{(d|J)} &= \sum_d P(d|A, J)\, O(d) && (56) \\ &= \sum_d P(d|A, J) \frac{P(d|A, I)\, P(I)}{P(d|A, J)\, P(J)} && (57) \\ &= \frac{P(I)}{P(J)} \underbrace{\sum_d P(d|A, I)}_{=1} && (58) \\ &= \frac{P(I)}{P(J)}, && (59) \end{aligned}$$

is independent on the strategy A.

$\Rightarrow$ By tuning the strategy, no additional odds mass (expected odds) in favor of a wrong hypothesis can be generated, however, the odds mass can be redistributed. E.g. rare high odds events can be traded for an increased number of moderate odds event. Stopping a measurement sequence when a chosen significance threshold happens to be reached is such a strategy.

Does this mean we do not learn from data?
Not at all, the expected odds for the right hypothesis, $1/O$, usually increases:

$$\left\langle \frac{1}{O(d)} \right\rangle_{(d|J)} = \sum_d P(d|A,J) \frac{P(d|A,J)\,P(J)}{P(d|A,I)\,P(I)} \tag{60}$$

$$= \frac{P(J)}{P(I)} \sum_d P(d|A,I) \underbrace{\left[\frac{P(d|A,J)}{P(d|A,I)}\right]}_{\equiv r(d)}^2 \tag{61}$$

$$= \frac{P(J)}{P(I)} \left\langle r^2(d) \right\rangle_{(d|A,I)} \tag{62}$$

$$\geq \frac{P(J)}{P(I)} \left\langle r(d) \right\rangle^2_{(d|A,I)} \tag{63}$$

$$= \frac{P(J)}{P(I)} \left[\sum_d P(d|A,I) \frac{P(d|A,J)}{P(d|A,I)}\right]^2 \tag{64}$$

$$= \frac{P(J)}{P(I)} \underbrace{\left[\sum_d P(d|A,J)\right]^2}_{=1} \tag{65}$$

$$= \frac{P(J)}{P(I)}, \tag{66}$$

where we used $\left\langle r^2(d) \right\rangle_{(d|A,I)} = \bar{r}^2 + \sigma_r^2 \geq \bar{r}^2$ with

$$\bar{r} \equiv \langle r(d) \rangle_{(d|A,I)}, \tag{67}$$

$$\sigma_r^2 \equiv \left\langle \underbrace{[r(d) - \bar{r}]}_{\equiv \Delta(r)}^2 \right\rangle_{(d|A,I)} \quad \text{since} \tag{68}$$

$$\left\langle r^2(d) \right\rangle_{(d|A,I)} = \left\langle [\bar{r} + \Delta(d)]^2 \right\rangle_{(d|A,I)} \tag{69}$$

$$= \left\langle \bar{r}^2 + 2\,\bar{r}\,\Delta(d) + \Delta^2(d) \right\rangle_{(d|A,I)} \tag{70}$$

$$= \bar{r}^2 + 2\,\bar{r}\, \underbrace{\langle \Delta(d) \rangle_{(d|A,I)}}_{=0} + \underbrace{\left\langle \Delta^2(d) \right\rangle_{(d|A,I)}}_{=\sigma_r^2} \tag{71}$$

$$= \bar{r}^2 + \sigma_r^2 \qquad \square \tag{72}$$

2 DECISION THEORY

2.1 OPTIMAL RISK

Decisions should be done rationally. For example in science, given data d from a measurement of a signal s we have to decide which estimate of s we publish. For a rational decision, we need to know the possible consequences of our action.

The **loss function** $l(a, s)$ quantifies the loss associated with an action a (e.g. the number we publish) if the data was actually generated by the signal having value s. l might measure the e.g. the lost money, status, health, security, or attention.
optimal risk: The risk of an action a given the data d is the expected loss

$$r(a,\, d) = \langle l(a,\, s)\rangle_{(s|d)} = \int ds\, l(a,\, s)\mathcal{P}(s|d). \tag{73}$$

The optimal action minimizes this risk.

2.2 LOSS FUNCTIONS

- **quadratic loss**: "square error of a trying to match true s" (used often in scientific publishing)

$$l(a,\, s) = (a - s)^2 \tag{74}$$

  Calculate the best action by minimizing the optimal risk:

$$r(a,\, d) = \int ds\, (a - s)^2\mathcal{P}(s|d) \tag{75}$$

$$= \langle (a - s)^2\rangle_{(s|d)} \tag{76}$$

$$\frac{\partial r(a,\, d)}{\partial a} = \langle 2(a - s)\rangle_{(s|d)} \tag{77}$$

$$= 2a - 2\langle s\rangle_{(s|d)} \stackrel{!}{=} 0 \tag{78}$$

$$\Rightarrow a = \langle s\rangle_{(s|d)} \tag{79}$$

For the quadratic loss function the best estimator for s is the **mean** of s under the posterior distribution $\mathcal{P}(s|d)$.

- **linear loss:** "absolute loss" (often used in nummerics)

$$l(a,\, s) = |a - s| \tag{80}$$

Calculate the best action by minimizing the optimal risk:

$$\frac{\partial r(a,d)}{\partial a} = \frac{\partial}{\partial a}\int_{-\infty}^{\infty} ds\,|a-s|\,\mathcal{P}(s|d) \tag{81}$$

$$= \frac{\partial}{\partial a}\left[\int_{-\infty}^{a} ds - \int_{a}^{\infty} ds\right]\,[(a-s)\mathcal{P}(s|d)] \tag{82}$$

$$= [|_{s=a} + |_{s=a}]\,[(a-s)\mathcal{P}(s|d)] + \int_{-\infty}^{a} ds\,\mathcal{P}(s|d) - \int_{a}^{\infty} ds\,\mathcal{P}(s|d) \tag{83}$$

$$\stackrel{!}{=} 0 \tag{84}$$

$$\Rightarrow \int_{-\infty}^{a} \mathcal{P}(s|d) = \int_{a}^{\infty} \mathcal{P}(s|d) \tag{85}$$

$$\int_{-\infty}^{\infty} \mathcal{P}(s|d) = 1 \tag{86}$$

$$\Rightarrow \int_{-\infty}^{a} \mathcal{P}(s|d) = 1/2 \tag{87}$$

For the linear loss the best estimate of a signal is the **median** of its posterior distribution $\mathcal{P}(s|d)$.

- **delta loss:** For this there is the same penalty, whenever the estimate a does not exactly correspond to the true signal s, unrelated to the distance between a and s. (might be regarded as a military loss function)

$$l(a,s) = -\delta(a-s) \tag{88}$$

$$r(a,d) = -\int ds\,\delta(a-s)\,\mathcal{P}(s|d) \tag{89}$$

$$= -\mathcal{P}(a|d) \tag{90}$$

For the delta loss the best estimate of a signal is given by its **mode,** the location of the maximum of the posterior distribution $\mathcal{P}(s|d)$.

2.3 COMMUNICATION

Communication requires to decide which message $M \in \mathbb{M}$ is to be sent to the recipient. Here, we are concerned with the optimal communication of a knowledge or believe state $p(s) = \mathcal{P}(s|I)$ on some signal s in case we have to approximate it by selecting a message M from a limited set $\mathbb{M}$. Each M generates a known believe state $q(s) = \mathcal{P}(s|M)$ in the recipient, so that we can identify q and M in the following. How should we decide to select among the $M \in \mathbb{M}$ (= among the accessible qs)?

We might want to avoid the embarrassment we face by giving wrong advice. Is there a generic measure for embarrassment? In general it will depend on how much damage is done if something is assumed but something else was the case. Here, we look only at the generic case that we want to avoid to inform incorrectly, irrespective of what the different cases of s are. The goal is to assign the highest possible probability to the s that turns out the case, with the catch, that this is not known.

We follow the argumentation of [9].

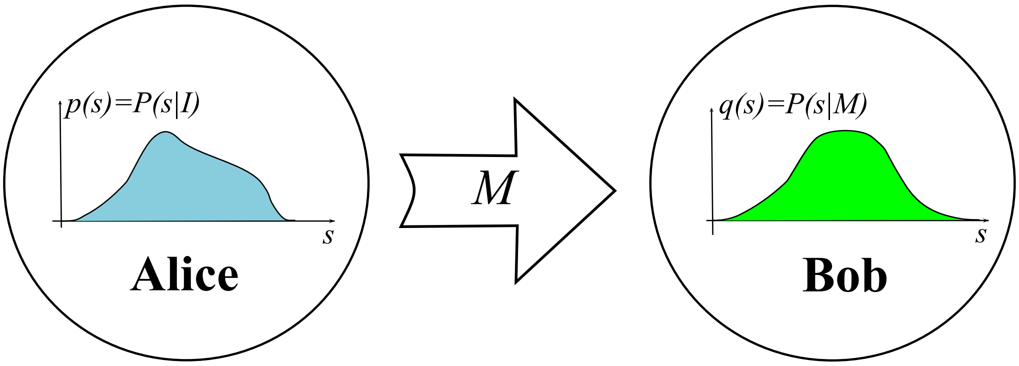


Figure 3: Communication setup: Alice wants to transfer her knowledge I to Bob, however, she is only able to select one imperfect message M from a set of possible messages $\mathbb{M}$. Which one should she send to inform Bob best? Which criteria should she use for her decision?

2.3.1 *Embarrassment – a unique loss function*

Loss function $l(q, s_0)$ to quantify embarrassment of communicating $q(s)$ in case that s turns out to be s_0.
As we do not know s_0, all we can do is to minimize the expected loss

$$r(q,p) = \langle l(q,s_0)\rangle_p = \int \mathrm{d}s_0\, l(q,s_0)p(s_0). \tag{91}$$

But which loss function is sensible?

Criterion 1. (Being local) *If $s = s_0$ turned out to be the case, l only depends on the prediction q actually made about s_0: $l(q, s_0) = \mathscr{L}(q(s_0))$*

For example frequentist's $\widetilde{p}$-values $\widetilde{p}(s_0) = \int_{s\geq s_0} ds\, \mathcal{P}(s|I)$ do not fulfill criterion (1) as they depends on the probability of counter-factual events.

Criterion 2. (Being proper) *If any belief on s can be communicated, the optimal communication should be $q = p$:*

$$\mathrm{argmin_q}\, \langle \mathscr{L}(q(s_0))\rangle_p = p \tag{92}$$

While minimizing the expected embarrassment we ensure the normalization of q via a Lagrange multiplier:

$$0 = \left(\underbrace{\frac{\mathrm{d}}{\mathrm{d}q(s)}}_{\text{minimum}} \left[\int \mathrm{d}s_0 \underbrace{\mathscr{L}(q(s_0))}_{\text{only } s_0 \text{ matters}} \underbrace{p(s_0)}_{\text{guessing } s_0} + \lambda \underbrace{\left(\int \mathrm{d}s_0\, q(s_0) - 1 \right)}_{\text{normalization}} \right] \right)_{\underbrace{q=p}_{\text{properness}}} \tag{93}$$

$$= \int \mathrm{d}s_0 \left[\mathscr{L}'(p(s_0))\, \delta(s_0 - s)\, p(s_0) + \lambda\, \delta(s_0 - s) \right] \tag{94}$$

$$= \mathscr{L}'(p(s))\, p(s) + \lambda \tag{95}$$

$$\Rightarrow \mathscr{L}'(p(s)) = -\frac{\lambda}{p(s)} \tag{96}$$

$$\Rightarrow \mathscr{L}(p(s)) = -\lambda \ln(p(s)) + \delta\,. \tag{97}$$

$\lambda > 0$ and δ are constants with respect to q, which can be chosen arbitrarily. Choosing $\lambda = 1$ and $\delta = 0$:

$$\textbf{Expected embarrassment: } r(q,p) = \langle l(q,s_0) \rangle_{p(s_0)} = -\int \mathrm{d}s_0\, p(s_0) \ln(q(s_0)) \tag{98}$$

$$\textbf{Cross entropy: } \mathcal{S}(p,q) = -\int \mathrm{d}s\, p(s) \ln(q(s)) \tag{99}$$

$$\textbf{Entropy: } \mathcal{S}(p) = -\int \mathrm{d}s\, p(s) \ln(p(s)) = \mathcal{S}(p,p) \tag{100}$$

Choosing $\delta = \int \mathrm{d}s_0\, p(s_0) \ln(p(s_0)) = -\mathcal{S}(p)$ provides a coordinate invariant measure:

Kullback-Leibler divergence

Kullback-Leibler divergence:

$$D_{\mathrm{KL}}(p||q) = \int \mathrm{d}s\, p(s) \ln\left(\frac{p(s)}{q(s)}\right) = \mathcal{S}(p,q) - \mathcal{S}(p)$$

Gibbs inequality

The **Gibbs inequality** states $D_{\mathrm{KL}}(p||q) \geq 0$ if p and q are properly normalized probabilities. Proof:

$$\begin{aligned} -D_{\mathrm{KL}}(p||q) &= \int \mathrm{d}s\, p(s) \ln\left(\frac{q(s)}{p(s)}\right) && (101) \\ &\leq \int \mathrm{d}s\, p(s) \left(\frac{q(s)}{p(s)} - 1\right) && (102) \\ &= \int \mathrm{d}s\, q(s) - \int \mathrm{d}s\, p(s) && (103) \\ &= 1 - 1 = 0\ \square && (104) \end{aligned}$$

$D_{\mathrm{KL}}(p||q) = 0$ iff (if and only if) $q(s) = p(s)$ for $\forall s$ (up to zero-measure differences, proof left for exercise).

3 INFORMATION MEASURES

Kullback-Leibler divergence

$$\mathrm{KL}_s(A,B) := D_{\mathrm{KL}}(\mathcal{P}(s|A)||\mathcal{P}(s|B)) = \int \mathrm{d}s\,\mathcal{P}(s|A)\,\ln\left(\frac{\mathcal{P}(s|A)}{\mathcal{P}(s|B)}\right) \qquad (105)$$

measures how much information on s is expected from A with respect to B. Unit is nit = nat or bits if $\log_2$ is used, conversion $1\,\mathrm{nit} = {}^1\!/_{\ln 2}\,\mathrm{bit} \approx 1.44\,\mathrm{bit}$, $1\,\mathrm{bit} = 1\,\mathrm{shannon}$.

Information (or surprise) $\mathcal{H}(s|I) = -\log\mathcal{P}(s|I)$ *information*

Product rule:

$$\begin{aligned}
\mathcal{P}(d,s|I) &= \mathcal{P}(d|s,I)\,\mathcal{P}(s|I) && (106)\\
&= \mathcal{P}(s|d,I)\,\mathcal{P}(d|I) && (107)\\
\Rightarrow \mathcal{H}(d,s|I) &= \mathcal{H}(d|s,I) + \mathcal{H}(s|I) && (108)\\
&= \mathcal{H}(s|d,I) + \mathcal{H}(d|I) && (109)
\end{aligned}$$

$\Rightarrow$ Information is additive.

Kullback-Leibler divergence

$$\mathrm{KL}_s(A,B) = \langle \mathcal{H}(s|B) - \mathcal{H}(s|A)\rangle_{(s|A)} \qquad (110)$$

measures expected information gain from B to A. Note that the averaging over $\mathcal{P}(s|A)$ is focusing on regions where $\ln\mathcal{P}(s|A) = -\mathcal{H}(s|A)$ is largest.

Example: Result vector $d^* \in \{0,1\}^n$ of n tosses of a fair coin gets known
Prior: $P(d|I) = 2^{-n}$
Posterior: $P(d|d^*,I) = \delta_{d,d^*}$

$$\begin{aligned}
\frac{\mathrm{KL}_d((d^*,I),I)}{\mathrm{bits}} &= \sum_d P(d|d^*,I)\log_2\left(\frac{P(d|d^*,I)}{P(d|I)}\right) && (111)\\
&= \sum_d \delta_{d,d^*}\log_2\left(\frac{\delta_{d,d^*}}{2^{-n}}\right) && (112)\\
&= \log_2(2^n) = n\,\log_2(2) = n && (113)
\end{aligned}$$

The result of n tosses contains exactly n bits on information on the outcome.

Optimal coding: choose message M that minimizes expected surprise

$$\mathrm{KL}_s(I,M) = \langle \mathcal{H}(s|M) - \mathcal{H}(s|I)\rangle_{(s|I)}$$

and the amount of information needed to update from M to I.

Independence:
If $\mathcal{P}(x,y|A) = \mathcal{P}(x|A)\,\mathcal{P}(y|A)$ and $\mathcal{P}(x,y|B) = \mathcal{P}(x|B)\,\mathcal{P}(y|B)$:

$$\begin{aligned}
\mathrm{KL}_{(x,y)}(A,B) &= \int \mathrm{d}x \int \mathrm{d}y\, \mathcal{P}(x,y|A)\, \ln\left(\frac{\mathcal{P}(x,y|A)}{\mathcal{P}(x,y|B)}\right) && (114)\\
&= \int \mathrm{d}x \int \mathrm{d}y\, \mathcal{P}(x|A)\,\mathcal{P}(y|A)\, \ln\left(\frac{\mathcal{P}(x|A)\,\mathcal{P}(y|A)}{\mathcal{P}(x|B)\,\mathcal{P}(y|B)}\right) && (115)\\
&= \int \mathrm{d}x \int \mathrm{d}y\, \mathcal{P}(x|A)\,\mathcal{P}(y|A)\, \left[\ln\left(\frac{\mathcal{P}(x|A)}{\mathcal{P}(x|B)}\right) + \ln\left(\frac{\mathcal{P}(y|A)}{\mathcal{P}(y|B)}\right)\right] && (116)\\
&= \int \mathrm{d}x\, \mathcal{P}(x|A)\, \ln\left(\frac{\mathcal{P}(x|A)}{\mathcal{P}(x|B)}\right) + \int \mathrm{d}y\, \mathcal{P}(y|A)\, \ln\left(\frac{\mathcal{P}(y|A)}{\mathcal{P}(y|B)}\right) && (117)\\
&= \mathrm{KL}_x(A,B) + \mathrm{KL}_y(A,B) && (118)
\end{aligned}$$

KL is additive for independent quantities.

Mutual information of I:

$$\begin{aligned}
\mathrm{MI}_{(x,y)}(I) &= D_{\mathrm{KL}}(\mathcal{P}(x,y|I)||\mathcal{P}(x|I)\mathcal{P}(y|I)) && (119)\\
&= \int \mathrm{d}x \int \mathrm{d}y\, \mathcal{P}(x,y|I)\, \ln\left(\frac{\mathcal{P}(x,y|I)}{\mathcal{P}(x|I)\mathcal{P}(y|I)}\right) && (120)\\
&= \langle \mathcal{H}(x|I) + \mathcal{H}(y|I) - \mathcal{H}(x,y|I)\rangle_{(x,y|I)} \geq 0 && (121)
\end{aligned}$$

Since

$$\frac{\mathcal{P}(x,y|I)}{\mathcal{P}(x|I)\mathcal{P}(y|I)} = \frac{\mathcal{P}(x|y,I)}{\mathcal{P}(x|I)} = \frac{\mathcal{P}(y|x,I)}{\mathcal{P}(y|I)} \qquad (122)$$

we also get

$$\begin{aligned}
\mathrm{MI}_{(x,y)}(I) &= \langle \mathcal{H}(x|I) - \mathcal{H}(x|y,I)\rangle_{(x,y|I)} && (123)\\
&= \langle \mathcal{H}(y|I) - \mathcal{H}(y|x,I)\rangle_{(x,y|I)} \geq 0 && (124)
\end{aligned}$$

The reduction of the expected surprises on one variable due to knowing the other one.
$\mathrm{MI}_{(x,y)}(I) = 0$ for independent quantities ("$x\perp y|I$"="$\mathcal{P}(x,y|I) = \mathcal{P}(x|I)\,\mathcal{P}(y|I)$").
MI used to test for relations between quantities.

Bayesian updating: $I \rightarrow (d,I)$, $\mathcal{P}(s|I) \rightarrow \mathcal{P}(s|d,I) = \frac{\mathcal{P}(d|s,I)}{\mathcal{P}(d|I)}\mathcal{P}(s|I)$

$$\begin{aligned}
\mathrm{KL}_s((d,I),I) &= \langle \mathcal{H}(s|I) - \mathcal{H}(s|d,I)\rangle_{(s|d,I)} && (125)\\
&= \int \mathrm{d}s\, \mathcal{P}(s|d,I) \ln\left(\frac{\mathcal{P}(s|d,I)}{\mathcal{P}(s|I)}\right) && (126)\\
&= \int \mathrm{d}s\, \mathcal{P}(s|d,I) \ln\left(\frac{\mathcal{P}(d|s,I)}{\mathcal{P}(d|I)}\right) && (127)\\
&= \langle \mathcal{H}(d|I) - \mathcal{H}(d|s,I)\rangle_{(s|d,I)} && (128)
\end{aligned}$$

Information gain on s by data d =how much data is less surprising if signal is known on (posterior) average.

Divergence: asymmetric distance measure (depends on direction).
Becomes symmetric for small distances:

$$\begin{aligned}
p(s) &= q(s) + \varepsilon(s) && (129)\\
\varepsilon(s) &\ll q(s), p(s)\,\forall s && (130)\\
0 &= \int \mathrm{d}s\,\varepsilon(s) && (131)
\end{aligned}$$

$$\begin{aligned}
D_{\mathrm{KL}}(p||q) &= \int \mathrm{d}s\, p(s) \log\left(\frac{p(s)}{q(s)}\right) && (132)\\
&= \int \mathrm{d}s\ (q(s)+\varepsilon(s)) \log\left(1+\frac{\varepsilon(s)}{q(s)}\right) && (133)\\
&= \int \mathrm{d}s \left\{(q(s)+\varepsilon(s))\left[\frac{\varepsilon(s)}{q(s)} - \frac{1}{2}\left(\frac{\varepsilon(s)}{q(s)}\right)^2\right] + \mathcal{O}(\varepsilon^3)\right\} && (134)\\
&= \int \mathrm{d}s \left[\varepsilon(s) + \frac{(\varepsilon(s))^2}{2q(s)} + \mathcal{O}(\varepsilon^3)\right] && (135)\\
&= 0 + \int \mathrm{d}s \frac{[p(s)-q(s)]^2}{2q(s)} + \mathcal{O}(\varepsilon^3) && (136)\\
&= \int \mathrm{d}s \frac{[p(s)-q(s)]^2}{2\sqrt{p(s)\,q(s)}} + \mathcal{O}(\varepsilon^3) && (137)
\end{aligned}$$

$1/p \approx 1/q \approx 1/\sqrt{pq}$ seems to be metric in space of probabilities
$\rightarrow$ information geometry (but be beware, original KL is not a distance!)

Probabilities are parameterized in terms of conditional parameters, $\mathcal{P}(s|\theta)$.
Expansion in terms of those leads to the **Fisher information metric:**

$$\begin{aligned}
\theta'_i &= \theta_i + \varepsilon_i && (138)\\
\mathcal{P}(s|\theta') &= \mathcal{P}(s|\theta) + \frac{\partial \mathcal{P}(s|\theta)}{\partial \theta_i}\varepsilon_i + \mathcal{O}(\varepsilon^2), \text{ sum convention} && (139)\\
\mathrm{KL}_s(\theta',\theta) &= \underbrace{\mathrm{KL}_s(\theta,\theta)}_{=0} + \underbrace{\left.\frac{\partial \mathrm{KL}_s(\theta',\theta)}{\partial \theta'_i}\right|_{\theta'=\theta}}_{=0} \varepsilon_i + \frac{1}{2}\varepsilon_i \underbrace{\left.\frac{\partial^2 \mathrm{KL}_s(\theta',\theta)}{\partial \theta'_i \partial \theta'_j}\right|_{\theta'=\theta}}_{=g^{ij}} \varepsilon_j + \mathcal{O}(\varepsilon^3) && (140)
\end{aligned}$$

where

$$
\begin{aligned}
g^{ij} &= \frac{\partial^2}{\partial\theta'_i\partial\theta'_j}\int \mathrm{d}s\,\mathcal{P}(s|\theta')\,\ln\frac{\mathcal{P}(s|\theta')}{\mathcal{P}(s|\theta)}\bigg|_{\theta'=\theta} && (141)\\
&= \frac{\partial}{\partial\theta'_i}\int \mathrm{d}s\,\left[\frac{\partial\mathcal{P}(s|\theta')}{\partial\theta'_j}\,\ln\frac{\mathcal{P}(s|\theta')}{\mathcal{P}(s|\theta)}+\frac{\partial\mathcal{P}(s|\theta')}{\partial\theta'_j}\right]\bigg|_{\theta'=\theta} && (142)\\
&= \frac{\partial}{\partial\theta'_i}\int \mathrm{d}s\,\left[\ln\frac{\mathcal{P}(s|\theta')}{\mathcal{P}(s|\theta)}+1\right]\frac{\partial\mathcal{P}(s|\theta')}{\partial\theta'_j}\bigg|_{\theta'=\theta} && (143)\\
&= \int \mathrm{d}s\,\left\{\frac{1}{\mathcal{P}(s|\theta')}\frac{\partial\mathcal{P}(s|\theta')}{\partial\theta'_i}\frac{\partial\mathcal{P}(s|\theta')}{\partial\theta'_j}+\left[\ln\frac{\mathcal{P}(s|\theta')}{\mathcal{P}(s|\theta)}+1\right]\frac{\partial^2\mathcal{P}(s|\theta')}{\partial\theta'_i\partial\theta'_j}\right\}\bigg|_{\theta'=\theta} && (144)\\
&= \int \mathrm{d}s\,\left[\frac{1}{\mathcal{P}(s|\theta)}\frac{\partial\mathcal{P}(s|\theta)}{\partial\theta_i}\frac{\partial\mathcal{P}(s|\theta)}{\partial\theta_j}+\frac{\partial^2\mathcal{P}(s|\theta)}{\partial\theta_i\partial\theta_j}\right] && (145)\\
&= \int \mathrm{d}s\,\mathcal{P}(s|\theta)\,\frac{\partial\ln\mathcal{P}(s|\theta)}{\partial\theta_i}\frac{\partial\ln\mathcal{P}(s|\theta)}{\partial\theta_j}+\underbrace{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\underbrace{\int \mathrm{d}s\,\mathcal{P}(s|\theta)}_{=1}}_{=0} && (146)\\
&= \left\langle\frac{\partial\mathcal{H}(s|\theta)}{\partial\theta_i}\frac{\partial\mathcal{H}(s|\theta)}{\partial\theta_j}\right\rangle_{(s|\theta)}, && (147)
\end{aligned}
$$

but also

$$
\begin{aligned}
g^{ij} &= \int \mathrm{d}s\,\frac{\partial\ln\mathcal{P}(s|\theta)}{\partial\theta_i}\frac{\partial\mathcal{P}(s|\theta)}{\partial\theta_j} && (148)\\
&= \frac{\partial}{\partial\theta_j}\int \mathrm{d}s\,\mathcal{P}(s|\theta)\,\frac{\partial\ln\mathcal{P}(s|\theta)}{\partial\theta_i}-\int \mathrm{d}s\,\mathcal{P}(s|\theta)\,\frac{\partial^2\ln\mathcal{P}(s|\theta)}{\partial\theta_i\partial\theta_j} && (149)\\
&= \frac{\partial}{\partial\theta_j}\int \mathrm{d}s\,\frac{\partial\mathcal{P}(s|\theta)}{\partial\theta_i}+\left\langle\frac{\partial^2\mathcal{H}(s|\theta)}{\partial\theta_i\partial\theta_j}\right\rangle_{(s|\theta)} && (150)\\
&= \underbrace{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\underbrace{\int \mathrm{d}s\,\mathcal{P}(s|\theta)}_{=1}}_{=0}+\left\langle\frac{\partial^2\mathcal{H}(s|\theta)}{\partial\theta_i\partial\theta_j}\right\rangle_{(s|\theta)} && (151)\\
&= \left\langle\frac{\partial^2\mathcal{H}(s|\theta)}{\partial\theta_i\partial\theta_j}\right\rangle_{(s|\theta)} && (152)
\end{aligned}
$$

is the (Bayesian) Fisher information metric. This measures how sensitive expected information gain is in the limit of small amounts of additional data. Used to characterize the sensitivity of future experiments with respect to parameters of interest.

4 MAXIMUM ENTROPY

4.1 DECODING A MESSAGE

Requirements on action for **optimal coding** of knowledge p with the aim to honestly inform the receiver with message M:

- Locality (possibilities not addressed by M should stay unaffected)
- Properness (if possible, $q = p$)

$\Rightarrow$ Cross entropy $\mathcal{S}(p, q) = -\int \mathrm{d}s\, p(s) \ln q(s)$ is action to choose $q(s) = \mathcal{P}(s|M)$ (up to constant in q)

- Coordinate invariance of action

$\Rightarrow$ KL divergence $D_{\mathrm{KL}}(p||q) = \mathcal{S}(p, q) - \mathcal{S}(p) = \int \mathrm{d}s\, p(s) \ln [p(s)/q(s)]$ codes same message, as $\mathcal{S}(p) = \mathcal{S}(p, p) = \mathrm{const}(q)$.

Requirements on action $\mathcal{S}[q|r]$ for **optimal decoding** of message M, where now $r(s) = \mathcal{P}(s|J)$ is initial knowledge of receiver, and $q(s) = \mathcal{P}(s|J, M)$ should be the updated state.

1. Locality:
2. Coordinate independence of result (and therefore of action)
3. Separability: "Independent systems can be equally treated jointly as well as separately."

Maximum entropy principle (Jaynes, see Sect. (4.2) for sketch of derivation)
$\Rightarrow$ Entropy $\mathcal{S}[q|r] = -D_{\mathrm{KL}}(q||r) = \mathcal{S}(r) - \mathcal{S}(q, r) = -\int \mathrm{d}s\, q(s) \ln [q(s)/r(s)]$ to be maximized w.r.t. q
$\Rightarrow$ KL divergence $D_{\mathrm{KL}}(q||r) = \mathcal{S}(q, r) - \mathcal{S}(r) = \int \mathrm{d}s\, q(s) \ln [q(s)/r(s)]$ to be minimized w.r.t. q

4.2 MAXIMUM ENTROPY PRINCIPLE

Entropy is a measure for the amount of the information, which is forcing a change in belief.
Use of entropy:

- Decide on an optimal strategy of updating (after receiving a message)
- Set up probabilities (the message was empty)
- General law to update information (the message could have come from anywhere, including nature)

Notation:
s: unknown quantity
J: initial background information
M: new information (message in form of a set of constraints)
Bayesian knowledge update of $M = \{d, \mathcal{P}(d|s, M)\}$ with $J' = JM$:

$$\mathcal{P}(s|J) \overset{M}{\to} \mathcal{P}(s|J') = \frac{\mathcal{P}(d|s, M)\mathcal{P}(s|J)}{\mathcal{P}(d)} \tag{153}$$

How does $P(s|J')$ look like under the assumption $M = \{d, \langle f(s)\rangle_{(s|M)}\}$?
$\Rightarrow$ $P(s|J')$ should carry a minimum of extra information with respect to $P(s|J)$ while being consistent with $J' = JM$.

Principle of Minimum Updating (PMU):
Beliefs must be reviewed only to the extent required by the new information.

Since we will in the following not assume a priori that this update is according to the laws of probabilities, we introduce for the following the notation $r(s) = P(s|J)$ and $q(s) = P(s|J')$ for the functions of s, that happen to describe our prior and posterior knowledge. Later we have to see whether our updating is consistent with Bayesian reasoning or not.

Entropy is a measure of the relative information of q with respect to r:

$$\begin{aligned} \text{relative entropy of } q \text{ w.r.t. } r &= \mathcal{S}[q|r] \\ &= \text{negative information gain } r \to q \end{aligned}$$

Therefore the PMU is equivalent to the Maximum Entropy Principle (MEP).

Maximum Entropy Principle (MEP)
Updating from $r(s) = P(s|I)$ to $q(s) = P(s|J' = JM)$ given some information M should maximize the entropy $\mathcal{S}[q|r]$ under the constraints of M. $\Rightarrow$

$\mathcal{S}$= action for updating, favouring the most ignorant knowledge state $P(s|J')$
In other words, $\mathcal{S}$ assigns numerical values to probability functions, such that if q_1 is preferred over q_2, then $\mathcal{S}[q_1|r] > \mathcal{S}[q_2|r]$. Following Jaynes, there are 3 criteria to construct entropy:

1. **Locality:** "Local information has only local effects."
   New information M affecting only some $\Omega' \subset \Omega = \{s\}$ in $J' = JM$ leaves the knowledge of J about $\Omega \backslash \Omega'$ unaffected

   $$\mathcal{P}(s|J', s \in \Omega\backslash\Omega') = \mathcal{P}(s|J, s \in \Omega\backslash\Omega') \tag{154}$$

   $\Rightarrow$Non-overlapping domains of s have an additive contribution to the entropy

   $$\mathcal{S}[q|r] = \int_\Omega ds\, F(q(s), r(s), s) \tag{155}$$

   F: some unknown local function

2. **Coordinate invariance:** "Chosen system of coordinates does not carry information"

Coordinate transformation:
$m(s)$: some density function
$m'(t)$: transformed density function

$$m(s)\,ds = m'(t)\,dt \tag{156}$$
$$m'(t) = m(s(t))\left|\frac{ds}{dt}\right| \tag{157}$$

This is also true for the considered probability densities q, r.

$$\mathcal{S}[q|r] = \int ds\, m_1(s)\, F'\left(\frac{q(s)}{m_2(s)}, \frac{r(s)}{m_3(s)}\right) \tag{158}$$

From 1. we know that if $\Omega = \Omega'$ and $M = \{\}$, then we require $q = r$. When there is no new information there is no reason for updating the probability density function and therefore q and r coincide.

$$\text{Jaynes shows} \Rightarrow \mathcal{S}[q|r] = \int_\Omega ds\, q(s)\, F'\left(\frac{q(s)}{r(s)}\right) \tag{159}$$

3. **Independence:** "Independent systems can be equally treated jointly as well as separately."
   Consider two independent systems:

$$s = (s_1, s_2) \tag{160}$$
$$r(s) = r_1(s_1)r_2(s_2) \tag{161}$$
$$q(s) = q_1(s_1)q_2(s_2) \tag{162}$$

New information $M = M_1 M_2$ is acquired $\Rightarrow \mathcal{S}[q|r] = \mathcal{S}[q_1|r_1] + \mathcal{S}[q_2|r_2]$
Using the results from the coordinate invariance, we get

$$\mathcal{S}[q|r] = -\int ds\, q(s) \ln\left(\frac{q(s)}{r(s)}\right) = -D_{\mathrm{KL}}(\mathrm{q}||\mathrm{r}). \tag{163}$$

Proof:

$$\begin{aligned}
\mathcal{S}[q|r] &= -\int ds\, q(s)\ln\left(\frac{q(s)}{r(s)}\right) && (164)\\
&= -\int ds_1 \int ds_2\, q_1(s_1)q_2(s_2)\ln\left(\frac{q_1(s_1)q_2(s_2)}{r_1(s_1)r_2(s_2)}\right) && (165)\\
&= -\int ds_1 \int ds_2\, q_1(s_1)q_2(s_2)\left[\ln\left(\frac{q_1(s_1)}{r_1(s_1)}\right) + \ln\left(\frac{q_2(s_2)}{r_2(s_2)}\right)\right] && (166)\\
&= -\left[\int ds_1 q_1(s_1)\ln\left(\frac{q_1(s_1)}{r_1(s_1)}\right)\right]\cdot \underbrace{\int ds_2 q_2(s_2)}_{=1} && (167)\\
&\quad -\left[\int ds_2 q_2(s_2)\ln\left(\frac{q_2(s_2)}{r_2(s_2)}\right)\right]\cdot \underbrace{\int ds_1 q_1(s_1)}_{=1} && (168)\\
&= \mathcal{S}[q_1|r_1] + \mathcal{S}[q_2|r_2] && (169)
\end{aligned}$$

Actually, the case $\mathcal{S}[q|r] = \text{const}$ for $\forall q, r$ has to be eliminated by the additional, pragmatic requirement

4. **Sensitivity:** "The gradient of $\mathcal{S}[q|r]$ should lead to the optimal q."

4.3 OPTIMAL COMMUNICATION

$$\mathcal{S}[q|r] = -D_{\mathrm{KL}}(\mathrm{q}||\mathrm{r}) \tag{170}$$

Note the different usages of MaxEnt and the KL divergence in optimal communication:
Coding a message is done via minimizing of second argument of $D_{\mathrm{KL}}(\mathrm{q}||\mathrm{r})$, to ensure that the receiver's knowledge would only need a minimal amount of information to catch up to p. **Decoding a message** is done via minimizing the first argument of $D_{\mathrm{KL}}(\mathrm{q}||\mathrm{r})$ (or maximizing $\mathcal{S}[q|r]$), to add the least amount of (spurious) information during decoding besides what the message says (maximal entropy).
Rule of thumb: The first argument (the probability averaged over) is always the more accurate one.

Optimal coding: choose message M that minimizes expected surprise

$$M = \underset{M'}{\operatorname{argmin}} \mathrm{KL}(I, M') = \underset{M'}{\operatorname{argmin}} \, \langle \mathcal{H}(s|M') - \mathcal{H}(s|I) \rangle_{(s|I)} \tag{171}$$

and the amount of information needed to update from M to I.
Optimal decoding: given initial knowledge state $r(s) = \mathcal{P}(s|J)$ and received message $M = M(p)$ about $p(s) = \mathcal{P}(s|I)$ (message here to be read as statement in the form of $M(p) = 0$), choose $q(s) = \mathcal{P}(s|J')$ consistent with M that adds the least amount of information with respect to r (leaves a maximum amount of information to be added later on):

$$q = \underset{q',\lambda}{\operatorname{argmin}} \left[D_{\mathrm{KL}}(q'||r) + \lambda \, M(q') \right] \tag{172}$$

$$J' = \underset{J'',\lambda}{\operatorname{argmin}} \left[\mathrm{KL}(J'', J) + \lambda \, M(\mathcal{P}(s|J'')) \right] \tag{173}$$

Optimal communication: optimal decoding of optimally coded message:
Sender with knowledge I codes message M to shift receiver knowledge from J to J'

$$M = \underset{M}{\operatorname{argmin}} \, \mathrm{KL}(I, J'(J, M)) \tag{174}$$

$$= \underset{M}{\operatorname{argmin}} \, \mathrm{KL} \left(I, \underset{J',\lambda}{\operatorname{argmin}} \left[\mathrm{KL}(J', J) + \lambda \, M(\mathcal{P}(s|J')) \right] \right) \tag{175}$$

$\Rightarrow$ **communication is a game**, in which sender chooses her action anticipating the reaction of the receiver.
In optimal communication, the sender wants to inform honestly (to generate minimal expected surprise for the receiver). For this, sender and receiver need to share common knowledge about coding (encoding and decoding), e.g. by having an agreement on this. Entropy singles out a coding scheme, alleviating the need to

first agree on a coding scheme.

Real communication: any of those assumption can be violated (*e.g.* sender emphasizes what she thinks is important for receiver, communicates what she wants the receiver to believe, receiver distrusts sender, sender makes wrong assumption about receiver's knowledge or ability to decode, ...)
$\Rightarrow$ really, really complicated mess, but interesting psychology

Corrective strategies:
Robust communication: Send facts! Communicate raw data instead of its interpretation! A Bayesian receivers will build up his own knowledge system.
Questions: Request for necessary, unambiguous information. Probe knowledge, assumptions, and communication strategies of communication partner.
Reputation systems: Remembering and rewarding honest and informative communications. This will encourage such and allows to identify and concentrate on trustworthy information sources.

4.4 MAXIMUM ENTROPY WITH HARD DATA CONSTRAINTS

Prior information: $I =$ "$q(d, s) = P(d, s|I)$, data d and signal s are unknown."
Updating information: $J =$"$d = d^*$ =observed data, data has a particular value."

$$p(d, s) := \mathcal{P}(d, s|I, J) = \delta_{d,d^*} \underbrace{\mathcal{P}(s|IJ)}_{=:p(s)} \tag{176}$$

$\Rightarrow p(s)$ has to be found, since if $p(s)$ is known, the updating of $p(d, s)$ is known.
Constrained entropy:

$$\begin{aligned} \mathcal{S}^*[p|q] &= -\int ds \sum_d p(d, s) \left[\ln\left(\frac{p(d, s)}{q(d, s)}\right) - \lambda \right] & (177) \\ &= -\int ds\, p(s) \left[\ln\left(\frac{p(s)}{q(d^*, s)}\right) - \lambda \right] & (178) \end{aligned}$$

For the second step we used $0 \ln 0 = 0$, since $\lim_{\epsilon \to 0} \epsilon \ln \epsilon = 0$.
The Lagrange multiplier λ enforces the normalisation of $p(s)$,

$$\frac{\partial \mathcal{S}^*}{\partial \lambda} = \int ds \sum_d p(d,s) \stackrel{!}{=} 1. \tag{179}$$

Maximizing the entropy:

$$\frac{\delta \mathcal{S}^*[p|q]}{\delta p(s')} = -\ln\left(\frac{p(s')}{q(d^*, s')}\right) + \lambda - \underbrace{\frac{p(s')}{p(s')}}_{=1} \stackrel{!}{=} 0 \tag{180}$$

$$\Rightarrow p(s) = q(d^*, s) \cdot e^{\lambda - 1} \tag{181}$$

Normalization:

$$\frac{\partial \mathcal{S}^*}{\partial \lambda} = \int ds\, p(s) = e^{\lambda-1} \underbrace{\int ds\, q(d^*, s)}_{\mathcal{Z}(d^*)} \stackrel{!}{=} 1 \tag{182}$$

$$\Rightarrow e^{\lambda-1} = \frac{1}{\mathcal{Z}(d^*)} \tag{183}$$

Merging the results from the maximization and the normalization with the partition sum $\mathcal{Z}(d^*)$ we get

$$P(s|I, J) = p(s) = \frac{q(d^*, s)}{\mathcal{Z}(d^*)} = \frac{P(d^*, s|I)}{\int ds\, P(d^*, s|I)} = P(s|d^*, I). \tag{184}$$

$\Rightarrow$ Maximum entropy embraces and extends Bayes updating!

The transition from Bayesian updating to Maximum Entropy updating has similarities to the transition from Newtonian dynamics to Lagrangian dynamics as in both cases dynamical equations containing forces (on mechanical or knowledge systems) become replaced and embraced by action principles.

4.5 MAXIMUM ENTROPY WITH SOFT DATA CONSTRAINTS

Prior information: $I =$ "$q(x) = \mathcal{P}(x|I)$, x is unknown."
Updating information: $J =$ "$d = \langle f(x)\rangle_{(x|J,I)} = \int dx\, f(x)\, \mathcal{P}(x|J, I)$" (e.g. from a perceived message)
The new information J constrains the probability density $p(x) = \mathcal{P}(x|\underbrace{J, I}_{=I'})$ (similar constraint for normalization: $\langle 1\rangle_{(x|J,I)}$=1). The constraint can be added to the entropy via a Lagrange multiplier,

$$\mathcal{S}^*[p|q] = -\int dx\, p(x) \left[\ln\left(\frac{p(x)}{q(x)}\right) - \lambda - \mu f(x)\right]. \tag{185}$$

Providing normalization, new information and maximum entropy, the following derivations are obtained:

$$\frac{\partial \mathcal{S}^*}{\partial \lambda} = \int dx\, p(x) = \langle 1\rangle_{(x|J,I)} \stackrel{!}{=} 1 \tag{186}$$

$$\frac{\partial \mathcal{S}^*}{\partial \mu} = \int dx\, p(x) f(x) = \langle f(x)\rangle_{(x|I,J)} \stackrel{!}{=} d \tag{187}$$

$$\frac{\delta \mathcal{S}^*}{\delta p(x)} = -\ln\left(\frac{p(x)}{q(x)}\right) + \lambda + \mu f(x) - \frac{p(x)}{p(x)} \stackrel{!}{=} 0. \tag{188}$$

$$\begin{aligned} \Rightarrow p(x) &= q(x) e^{\lambda-1} e^{\mu f(x)} && (189) \\ &= \frac{q(x)}{\mathcal{Z}(\mu)} e^{\mu f(x)} && (190) \\ \mathcal{Z}(\mu) &= \int dx\, q(x) e^{\mu f(x)} && (191) \end{aligned}$$

The partition sum $\mathcal{Z}(\mu)$ accounts for the normalization of the updated probability density $p(x)$. The Lagrange multiplier μ has to be chosen such that

$$d \stackrel{!}{=} \langle f(x) \rangle_{(x|J)} = \frac{\int dx\, f(x) q(x) e^{\mu f(x)}}{\mathcal{Z}(\mu)} = \frac{1}{\mathcal{Z}(\mu)} \frac{\partial \mathcal{Z}(\mu)}{\partial \mu} = \frac{\partial \ln \mathcal{Z}(\mu)}{\partial \mu}. \tag{192}$$

4.6 DIFFERENT FLAVORS OF ENTROPY

prior: $q(x) = \mathcal{P}(x|I)$
constraint $J : \langle f(x) \rangle_{(x|J,I)} = d$
posterior: $p(x) = \mathcal{P}(x|J, I)$
constraints with Lagrange multiplier:

- for normalization: $\lambda \left(\int dx\, p(x) - 1 \right)$
- for new information J: $\mu \left(\int dx\, p(x) f(x) - d \right)$

The ***constrained entropy*** $\mathcal{S}[p|q, J]$, which has to be maximized with respect to $p(x)$, λ, μ, is given by,

$$\mathcal{S}[p|q, J] = \underbrace{\underbrace{-\{\int dx\, p(x) [\ln \left(\frac{p(x)}{q(x)} \right)}_{\mathcal{S}[p|q]} - \lambda - \mu f(x)]\}}_{\mathcal{S}^*[p|q,J]} - \lambda - \mu d. \tag{193}$$

In this case $\mathcal{S}[p|q]$ represents the ***amount of relative information*** of p with respect to q in nits. $\mathcal{S}^*[p|q, J]$ denotes the ***"alternative/auxiliary entropy"***, which has to be maximized with respect to $p(x)$ and sloped by $\frac{\partial \mathcal{S}^*}{\partial \lambda} \stackrel{!}{=} 1$, $\frac{\partial \mathcal{S}^*}{\partial \mu} \stackrel{!}{=} d$.

4.7 INFORMATION GAIN BY MAXIMIZING THE ENTROPY

Instead of information gain it is more precise to talk about a loss of uncertainty associated with the relative entropy.

- relative negative information gain:

$$\mathcal{S}[p|q] = -\int dx\, p(x) \ln \left(\frac{p(x)}{q(x)} \right) \tag{194}$$

Plugging in the maximum entropy solution for $p(x)$ we obtain

$$\begin{aligned}
\mathcal{S}[p|q] &= -\underbrace{\int dx\, \frac{q(x)e^{\mu f(x)}}{\mathcal{Z}(\mu)}}_{=1,\text{ as } \int dx\, q(x)e^{\mu f(x)}=\mathcal{Z}(\mu)} \ln\left(\frac{e^{\mu f(x)}}{\mathcal{Z}(\mu)}\right) && (195)\\
&= -\left[\int dx\, \frac{q(x)e^{\mu f(x)}}{\mathcal{Z}(\mu)}\mu f(x) - \ln \mathcal{Z}(\mu)\right] && (196)\\
&= \ln \mathcal{Z}(\mu) - \mu \underbrace{\langle f(x)\rangle_{(x|J)}}_{=d} && (197)\\
\Rightarrow \mathcal{S}[p|q] &= \ln \mathcal{Z}(\mu) - \mu d && (198)
\end{aligned}$$

- auxiliary entropy:

$$\begin{aligned}
\mathcal{S}^*[p|q, J] &= \mathcal{S}[p|q] + \lambda + \mu \underbrace{\langle f(x)\rangle_{(x|J)}}_{=d} && (199)\\
&= \mathcal{S}[p|q] + \lambda + \mu d && (200)\\
&= \ln \mathcal{Z}(\mu) - \mu d + \lambda + \mu d && (201)\\
&= \ln \mathcal{Z}(\mu) + \lambda && (202)\\
&= 1 && (203)
\end{aligned}$$

For the last step we used

$$\begin{aligned}
\mathcal{Z}(\mu) &= e^{1-\lambda} && (204)\\
\lambda &= 1 - \ln \mathcal{Z}(\mu). && (205)
\end{aligned}$$

- constrained entropy:

$$\begin{aligned}
\mathcal{S}[p|q, J] &= \mathcal{S}^*[p|q, J] - \lambda - \mu d && (206)\\
&= 1 - \lambda - \mu d && (207)\\
&= \ln \mathcal{Z}(\mu) - \mu d && (208)
\end{aligned}$$

$\Rightarrow$At the maximum of $\mathcal{S}[p|q, J]$ the change in information is,

$$\mathcal{S}[p|q, J] = \mathcal{S}[p|q] = \ln \mathcal{Z}(\mu) - \mu d \qquad (209)$$

If n constraints $\langle f_i(x)\rangle = d_i$ are considered, we define

$$\begin{aligned}
d &= \begin{pmatrix} d_1\\ \vdots \\ d_n\end{pmatrix}, f(x) = \begin{pmatrix} f_1(x)\\ \vdots \\ f_n(x)\end{pmatrix}, \mu = \begin{pmatrix} \mu_1\\ \vdots \\ \mu_n\end{pmatrix}\\
d^\dagger &= (\overline{d_1}, \overline{d_2}, ..., \overline{d_n})\\
d^\dagger \mu &= \sum_{i=1}^{n} \overline{d_i}\mu_i
\end{aligned}$$

Maximizing the entropy:

- with respect to λ :
$$\frac{\partial \mathcal{S}[p|q, J]}{\partial \lambda} \stackrel{!}{=} 0$$
Condition is automatically fulfilled by normalization $1/\mathcal{Z}$.

- with respect to μ :
$$\frac{\partial \mathcal{S}[p|q, J]}{\partial \mu} = \frac{\partial \ln \mathcal{Z}(\mu)}{\partial \mu} - d \stackrel{!}{=} 0$$

The Lagrange multiplier μ is determined to,
$$\frac{\partial \ln \mathcal{Z}(\mu)}{\partial \mu} = d,$$
by maximizing the entropy ($\ln \mathcal{Z}(\mu)$ may be interpreted as the Helmholtz free energy).

Maximum Entropy Recipe:
$q(x) = \mathcal{P}(x|I)$, $J = "\langle f(x)\rangle_{(x|J,I)} = d"$, $p(x) = \mathcal{P}(x|J, I) =?$

1. calculate the partition sum: $\mathcal{Z}(\mu) = \int dx\, q(x) e^{\mu f(x)}$
2. determine μ: $\frac{\partial \ln \mathcal{Z}(\mu)}{\partial \mu} \stackrel{!}{=} d$
3. assign: $p(x) = \frac{q(x)\, e^{\mu^\dagger f(x)}}{\mathcal{Z}(\mu)}$
4. calculate the information gain (in nits=bits/ln2 $\approx$ 1.44 bits if natural logarithm is used in entropy):
$$\Delta\mathcal{I}[p|q] = -\mathcal{S}[p|q] = \mu d - \ln \mathcal{Z}(\mu)$$

4.7.1 *Coin Tossing Example*

- $I =$"$x \in \{0, 1\}$"
- $q(x) = P(x|I) = 1/2$
- $J =$"frequency of heads is f" = "$\langle x\rangle_{(x|f)} = f$"

1. calculate $\mathcal{Z}(\mu)$:
$$\mathcal{Z}(\mu) = \sum_{x \in \{0,1\}} q(x) e^{\mu x} = \frac{1}{2}(1 + e^{\mu}) \text{ for } \mu < 0 \tag{210}$$
2. determine μ:
$$\frac{\partial \ln \mathcal{Z}(\mu)}{\partial \mu} = \frac{e^{\mu}}{1 + e^{\mu}} \stackrel{!}{=} f \tag{211}$$
$$\Rightarrow e^{\mu} = \frac{f}{1 - f} \tag{212}$$
$$\Rightarrow \mu = \ln\left(\frac{f}{1 - f}\right) \tag{213}$$

Insert in $\mathcal{Z}(\mu)$:

$$\mathcal{Z}(\mu) = \frac{1}{2}\left(1 + \frac{f}{1-f}\right) = \frac{1}{2(1-f)} \tag{214}$$

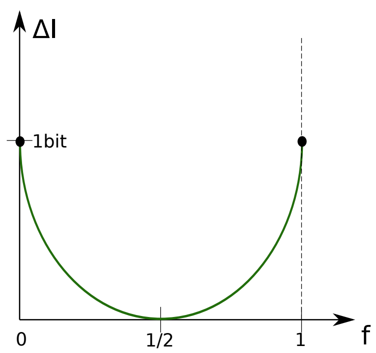
3. calculate $p(x) = \mathcal{P}(x|J, I)$:

$$\begin{aligned} p(x) &= \frac{q(x)e^{\mu x}}{\mathcal{Z}(\mu)} && (215) \\ &= \frac{1/2}{1/2(1-f)}(e^{\mu})^x && (216) \\ &= (1-f)\left(\frac{f}{1-f}\right)^x && (217) \\ &= f^x(1-f)^{1-x} && (218) \end{aligned}$$

$\Rightarrow$ The $\mathcal{P}(x|J, I)$, calculated by Maximum Entropy Principle is the same as we have used before in the coin flip example.

4. calculate the information gain ΔI:

$$\begin{aligned} \Delta\mathcal{I}[p|q] &= -\mathcal{S}[p|q] && (219) \\ &= \mu f - \ln \mathcal{Z}(\mu) && (220) \\ &= f \ln\left(\frac{f}{1-f}\right) - \ln\left(\frac{1}{2(1-f)}\right) && (221) \\ &= \ln 2 + f \ln f + (1-f)\ln(1-f) && (222) \\ &= (1 + f\, \log_2 f + (1-f)\, \log_2(1-f)) \text{ bits} && (223) \end{aligned}$$



We can get up to 1 bit of information on the next outcome by knowing that $f = 0$ or $f = 1$ as the yes/no-question "What is the outcome of the next toss?" is definitively answered. For $f = 1/2$ there is no change in information $\Delta\mathcal{I} = 0$, we are as unsure about the next outcome as before.
The amount of gained information on the sequence of the next n outcomes is n times the one of a single outcome.

4.7.2 *Positive Counts Example*

- $I =$"$n \in \mathbb{N}$"
- $q(n) = \mathcal{P}(n|I) = \text{const.} = q$
- $J =$"$\langle n \rangle = \lambda$"

1. calculate $\mathcal{Z}(\mu)$:

$$\mathcal{Z}(\mu) = q \sum_{n=0}^{\infty} e^{\mu n} = q \sum_{n=0}^{\infty} [e^{\mu}]^n = \frac{q}{1 - e^{\mu}} \text{ for } \mu < 0 \tag{224}$$

2. determine μ:

$$\frac{\partial \ln \mathcal{Z}(\mu)}{\partial \mu} = \frac{\partial}{\partial \mu} [\ln q - \ln(1 - e^{\mu})] = -\frac{1}{1 - e^{\mu}} \cdot (-e^{\mu}) \tag{225}$$

$$= \frac{e^{\mu}}{1 - e^{\mu}} \stackrel{!}{=} \lambda \tag{226}$$

$$\Rightarrow e^{\mu} = \frac{\lambda}{1 + \lambda} \tag{227}$$

Set the result in $\mathcal{Z}(\mu)$:

$$\mathcal{Z}(\mu) = \frac{q}{1 - \frac{\lambda}{1+\lambda}} = q(1 + \lambda) \tag{228}$$

3. calculate $p(n) = \mathcal{P}(n|\lambda = \langle n \rangle)$:

$$p(n) = \frac{q(n)\, e^{\mu n}}{\mathcal{Z}(\mu)} = \frac{q \cdot \left(\frac{\lambda}{1+\lambda}\right)^n}{q(1 + \lambda)} \tag{229}$$

$$= \frac{1}{1 + \lambda} \left(\frac{\lambda}{\lambda + 1} \right)^n = \lambda^n \, (1 + \lambda)^{-1-n} \tag{230}$$

Check of compliance with constraints:

$$\sum_{n=0}^{\infty} \mathcal{P}(n|\lambda) = \sum_{n=0}^{\infty} \frac{1}{1 + \lambda} \left(\frac{\lambda}{\lambda + 1} \right)^n = \frac{1}{1 + \lambda} \cdot \frac{1}{1 - \frac{\lambda}{1+\lambda}} \tag{231}$$

$$= \frac{1}{1 + \lambda} \cdot (1 + \lambda) = 1 \tag{232}$$

$$\sum_{n=0}^{\infty} n\, \mathcal{P}(n|\lambda) = \sum_{n=0}^{\infty} n \frac{1}{1 + \lambda} \underbrace{\left(\frac{\lambda}{\lambda + 1} \right)^n}_{=y^n} \tag{233}$$

$$= \frac{y}{1 + \lambda} \partial_y \sum_{n=0}^{\infty} y^n = \frac{y}{1 + \lambda} \partial_y \frac{1}{1 - y} \tag{234}$$

$$= \frac{y}{(1 + \lambda)(1 - y)^2} = \frac{\lambda}{(1 + \lambda)^2 (1 + \lambda)^{-2}} = \lambda \tag{235}$$

4.7.3 *Many Small Count Additive Processes*

Consider a total number of counts n distributed on N independent processes:

- $N =$ "number of processes"
- $n = \sum_{i=1}^{N} n_i =$"total number of counts"
- $I =$"$n_i \in \mathbb{N}$"
- $J =$"$\langle n_i \rangle_{(n_i|J)} = \delta$ for all i" $\Rightarrow$ "$\langle n \rangle_{(n|J)} = \lambda \equiv \delta N$ "

From the probability $\mathcal{P}(n_i|\delta = \langle n_i \rangle)$ known from the positive count example,

$$\mathcal{P}(n_i|\delta = \langle n_i \rangle) = \frac{1}{1+\delta}\left(\frac{\delta}{1+\delta}\right)^{n_i}, \tag{236}$$

we can calculate the probability $\mathcal{P}(n|\lambda, N)$,

$$\mathcal{P}(n|\lambda, N) = \underbrace{\sum_{n_1=0}^{\infty} \dots \sum_{n_N=0}^{\infty}}_{\equiv \sum_{\vec{n}=\vec{0}}^{\vec{\infty}}} \mathcal{P}(n, n_1, n_2, ...n_N | \underbrace{\lambda, N}_{I'=J,I}). \tag{237}$$

Assuming the independence of processes this can be decomposed to

$$\begin{aligned}
\mathcal{P}(n|I') &= \sum_{\vec{n}=\vec{0}}^{\vec{\infty}} \mathcal{P}(n|n_1, n_2, ...n_N, I')\mathcal{P}(n_1|I')\mathcal{P}(n_2|I')...\mathcal{P}(n_N|I') &(238)\\
&= \sum_{\vec{n}=\vec{0}}^{\vec{\infty}} \delta_{n,\sum_{i=0}^{N} n_i} \frac{1}{1+\delta}\left(\frac{\delta}{\delta+1}\right)^{n_1} \cdots \frac{1}{1+\delta}\left(\frac{\delta}{\delta+1}\right)^{n_N} &(239)\\
&= \left(\frac{1}{1+\delta}\right)^{N} \sum_{\vec{n}=\vec{0}}^{\vec{\infty}} \delta_{n,\sum_{i=0}^{N} n_i} \left(\frac{\delta}{1+\delta}\right)^{\sum_{i=0}^{N} n_i} &(240)\\
&= \left(\frac{1}{1+\delta}\right)^{N} \frac{N^n}{n!} \left(\frac{\delta}{1+\delta}\right)^{n}. &(241)
\end{aligned}$$

For the latter step, we used the knowledge that there are N^n possibilities to distribute n counts on N processes and $n!$ possibilities to reorder the n counts.

$$\begin{aligned}
\mathcal{P}(n|I') &= \frac{\delta^n}{(1+\delta)^{N+n}} \cdot \frac{N^n}{n!} &(242)\\
&= \frac{(\lambda/N)^n}{(1+\lambda/N)^{N+n}} \cdot \frac{N^n}{n!} &(243)\\
&= \frac{\lambda^n}{n!} \cdot \left(1+\frac{\lambda}{N}\right)^{-N} \left(1+\frac{\lambda}{N}\right)^{-n} &(244)
\end{aligned}$$

If an infinite number of processes ($N \to \infty$, $\lambda =$fixed, $\delta = \lambda/N \to 0$) is considered, the probability takes on the form of a Poisson distribution:

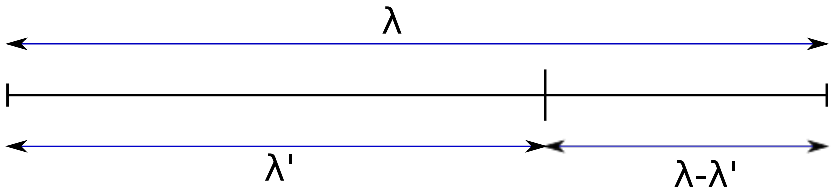
$$\mathcal{P}(n|\lambda, N \to \infty) = \frac{\lambda^n}{n!} \underbrace{\left(1+\frac{\lambda}{N}\right)^{-N}}_{\to e^{-\lambda}} \underbrace{\left(1+\frac{\lambda}{N}\right)^{-n}}_{\to 1}. \tag{245}$$

Poisson distribution:

$$\Rightarrow \mathcal{P}(n|\lambda, N \to \infty) = \frac{\lambda^n e^{-\lambda}}{n!} \tag{246}$$

The Poisson distribution is divisible:

$$\mathcal{P}(n|\lambda) = \sum_{m=0}^{n} \mathcal{P}(m|\lambda')\, \mathcal{P}(n-m|\lambda-\lambda') \tag{247}$$



If $\mathcal{P}(m|\lambda')$ and $\mathcal{P}(n-m|\lambda-\lambda')$ are Poisson distributions, $\mathcal{P}(n|\lambda)$ is a Poisson distribution as well.
Proof:

$$\begin{aligned}
\sum_{m=o}^{n} \mathcal{P}(m|\lambda')\mathcal{P}(n-m|\lambda-\lambda') &= \sum_{m=0}^{n} \frac{\lambda'^m e^{-\lambda'}}{m!} \frac{(\lambda-\lambda')^{n-m} e^{-\lambda+\lambda'}}{(n-m)!} && (248)\\
&= \frac{e^{-\lambda}}{n!} \sum_{m=0}^{n} \frac{n!}{m!(n-m)!} \lambda'^m (\lambda-\lambda')^{n-m} && (249)\\
&= \frac{e^{-\lambda}}{n!} (\lambda' + (\lambda-\lambda'))^n && (250)\\
&= \frac{e^{-\lambda}}{n!} \lambda^n && (251)\\
&= \mathcal{P}(n|\lambda) && (252)
\end{aligned}$$

In course of the proof we used the binomial identity,

$$\sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = (x+y)^n. \tag{253}$$

Additionally, it can be proven by recursion that the Poisson distribution is actually infinitely divisible.

4.8 MAXIMUM ENTROPY WITH KNOWN 1^{ST} AND 2^{ND} MOMENTS

- $I =$"$x \in \mathbb{R}$"
- $q(x) = P(x|I) = \text{const.}$
- $J =$"$\langle x \rangle_{(x|J,I)} = m,\ \langle (x-m)^2 \rangle_{(x|J,I)} = \sigma^2$"
- $P(x|J,I) = \frac{e^{\alpha x + \beta(x-m)^2}}{\mathcal{Z}(\alpha,\beta)}$

1. calculate $\mathcal{Z}(\alpha, \beta)$:

$$\begin{aligned}
\mathcal{Z}(\alpha, \beta) &= \int_{-\infty}^{\infty} dx\, e^{\alpha x+\beta(\underbrace{x-m}_{=x'})^2} && (254)\\
&= \int_{-\infty}^{\infty} dx'\, e^{\alpha x'+\alpha m+\beta x'^2} && (255)\\
\text{Completing the square:} &= e^{\alpha m}\int_{-\infty}^{\infty} dx'\, e^{\beta\left(x'^2+\frac{2\alpha x'}{2\beta}+\frac{\alpha^2}{(2\beta)^2}\right)-\frac{\alpha^2}{4\beta}} && (256)\\
&= e^{\alpha m-\frac{\alpha^2}{4\beta}}\int_{-\infty}^{\infty} dx'\, e^{\beta\left(x'+\frac{\alpha}{2\beta}\right)^2} && (257)\\
\text{Claiming } \beta<0\text{:} &= e^{\alpha m+\frac{\alpha^2}{4|\beta|}}\int_{-\infty}^{\infty} dx'\, e^{-|\beta|\left(x'-\frac{\alpha}{2|\beta|}\right)^2} && (258)\\
&= e^{\alpha m+\frac{\alpha^2}{4|\beta|}}\sqrt{\frac{\pi}{-\beta}} && (259)
\end{aligned}$$

2. determine α and β:

$$\begin{aligned}
\ln \mathcal{Z}(\alpha, \beta) &= \alpha m-\frac{\alpha^2}{4\beta}+\frac{1}{2}\ln\left(\frac{\pi}{-\beta}\right) && (260)\\
\frac{\partial \ln \mathcal{Z}(\alpha, \beta)}{\partial \alpha} &= m-\frac{\alpha}{2\beta} \stackrel{!}{=} m && (261)\\
\Rightarrow \alpha &= 0 && (262)\\
\frac{\partial \ln \mathcal{Z}(\alpha=0, \beta)}{\partial \beta} &= -\frac{1}{2\beta} \stackrel{!}{=} \sigma^2 && (263)\\
\Rightarrow \beta &= -\frac{1}{2\sigma^2} && (264)
\end{aligned}$$

Insert the result in $\mathcal{Z}(\alpha, \beta)$:

$$\mathcal{Z} = \sqrt{2\pi\sigma^2} \tag{265}$$

3. calculate $P(x|J, I)$:

$$\begin{aligned}
P(x|J, I) &= \left.\frac{e^{\alpha x+\beta(x-m)^2}}{\mathcal{Z}(\alpha, \beta)}\right|_{\alpha=0,\beta=-1/(2\sigma^2)} && (266)\\
&= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-m)^2}{2\sigma^2}} && (267)\\
&= \mathcal{G}(x-m, \sigma^2) && (268)
\end{aligned}$$

$\Rightarrow$ The Maximum Entropy PDF $P(x|J, I)$ for only known 1st and 2nd moments (and flat prior) is the Gaussian distribution.

5 GAUSSIAN DISTRIBUTION

5.1 ONE DIMENSIONAL GAUSSIAN

The Gaussian distribution is widely used, since

- it is the Maximum Entropy solution, if only 1st and 2nd moments are known.
- emerges as the distribution function of the sum of many (number $\to \infty$) independent small processes (dispersion $\to 0$, with limited high order moments) according to the central limit theorem
- it is mathematically convenient, in particular in higher dimension problems, and since it is infinitely divisible.

The Gaussian PDF with variance σ_x^2 and mean m is given by

Gaussian distribution

One dimensional Gaussian distribution:

$$\mathcal{G}(x-m,\sigma_x^2) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x-m)^2}{2\,\sigma_x^2}\right). \tag{269}$$

5.2 MULTIVARIATE GAUSSIAN

The one dimensional Gaussian distribution can be generalized to higher dimensions. Let $x = (x_1, \dots x_n)^{\mathrm{t}}$ be a vector of n **zero centered independent Gaussian distributed variables** with variances $\sigma_1^2, \dots \sigma_n^2$, respectively. Their joint probability is just the product of their individual probabilities,

$$\begin{aligned}
\mathcal{P}(x) &= \prod_{i=1}^{n} \mathcal{P}(x_i) && (270)\\
&= \prod_{i=1}^{n} \mathcal{G}(x_i,\sigma_i^2) && (271)\\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i)^2}{2\,\sigma_i^2}\right) && (272)\\
&= \frac{1}{\prod_{i=1}^{n}\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{x_i^2}{\sigma_i^2}\right) && (273)\\
&= \frac{1}{\sqrt{|2\pi X|}} \exp\left(-\frac{1}{2}x^\dagger X^{-1} x\right) && (274)
\end{aligned}$$

Multivariate Gaussian:

$$\mathcal{G}(x,X) = \frac{1}{\sqrt{|2\pi X|}} \exp\left(-\frac{1}{2}x^\dagger X^{-1} x\right), \tag{275}$$

where we introduced the diagonal covariance matrix $X = \mathrm{diag}(\sigma_1^2, \dots \sigma_n^2)$. $|X| =$

$\prod_i \sigma_i^2$ denotes the determinant of X and $x^\dagger$ the transposed and complex conjugated x.

For the former derivation of the multivariate Gaussian, we considered only independent coordinates. Dependent or correlated Gaussian variables can be obtained from this by a simple orthonormal basis transformation, denoted O,

$$y = O\,x \tag{276}$$

$$O^{-1} = O^\dagger \tag{277}$$

$$\Rightarrow \quad |O| = |O^\dagger| = |O^{-1}| = 1/|O| \tag{278}$$

$$\Rightarrow \quad |O|^2 = 1 \tag{279}$$

$$\Rightarrow \quad \|O\| = \|O^\dagger\| = 1, \tag{280}$$

in the n-dimensional space.

Conservation of probability mass:

$$\mathcal{P}(y|I)\,dy = \mathcal{P}(x|I)\,dx\big|_{x=O^\dagger y} \tag{281}$$

$$\Rightarrow \mathcal{P}(y|I) = \mathcal{G}(x,\,X)\,\underbrace{\left|\left|\frac{\partial x}{\partial y}\right|\right|}_{\frac{\partial x_i}{\partial y_j}=O^\dagger_{ij}}\Bigg|_{x=O^\dagger y} \tag{282}$$

$$= \mathcal{G}(O^\dagger y,\,X)\,\underbrace{\|O^\dagger\|}_{=1} \tag{283}$$

$$= \frac{1}{\sqrt{|2\pi X|}}\exp\left(-\frac{1}{2}\underbrace{(O^\dagger y)^\dagger}_{x^\dagger=y^\dagger O}X^{-1}\underbrace{O^\dagger y}_{x}\right) \tag{284}$$

$$= \frac{1}{\sqrt{|2\pi X|}}\exp\left(-\frac{1}{2}y^\dagger\underbrace{O\,X^{-1}O^\dagger}_{Y^{-1}}\,y\right) \tag{285}$$

$$= \frac{1}{\sqrt{|2\pi Y|}}\exp\left(-\frac{1}{2}y^\dagger Y^{-1}\,y\right) \tag{286}$$

For calculations in the last step we used,

$$\begin{aligned} |Y| &= |Y^{-1}|^{-1} \\ &= |OX^{-1}O^\dagger|^{-1} \\ &= (\underbrace{|O|}_{=\pm 1}\,|X^{-1}|\,\underbrace{|O^\dagger|}_{=\pm 1})^{-1} \\ &= |X|. \end{aligned}$$

Generic multivariate Gaussian:

$$\mathcal{P}(y) = \mathcal{G}(y, Y) = \frac{1}{\sqrt{|2\pi Y|}} \exp\left(-\frac{1}{2} y^\dagger Y^{-1} y\right) \tag{287}$$

in case Y is positive definite and symmetric (hermitian in the complex case), which is equivalent to the existence of an orthonormal transformation O that consists of the eigenvectors of Y that diagonalizes Y to a matrix $X = O^\dagger Y O$ with strictly positive values on the diagonal, the eigenvalues of Y.

Moments of the multivariate Gaussian:

$$\langle 1\rangle_{\mathcal{G}(y,Y)} = \int dy\, 1\, \mathcal{G}(y, Y) = \int dx\, 1\, \mathcal{G}(x, X) = 1 \tag{288}$$

$$1 = \frac{1}{\sqrt{|2\pi Y|}} \underbrace{\int dy\, \exp\left(-\frac{1}{2} y^\dagger Y^{-1} y\right)}_{=\sqrt{|2\pi Y|}} \tag{289}$$

$\Rightarrow$The multivariate Gaussian is properly normalized.

$$\begin{aligned}
\langle y\rangle_{\mathcal{G}(y,Y)} &= \int dy\, y\, \mathcal{G}(y, Y) &(290)\\
&= \int dy' (-y')\, \mathcal{G}(-y', Y) \| -\mathbb{1}\| &(291)\\
&= -\int dy'\, y'\, \mathcal{G}(y', Y) &(292)\\
&= -\langle y'\rangle_{\mathcal{G}(y',Y)} &(293)
\end{aligned}$$

$$\Rightarrow \langle y\rangle_{\mathcal{G}(y,Y)} = 0$$

We used the coordinate transformation $y' = -y$, $\frac{\partial y'}{\partial y} = -\mathbb{1}$.

For every odd function f of y (i.e. $f(-y) = -f(y)$) every contribution to $\langle f(y)\rangle_{\mathcal{G}(y,Y)}$ is compensated by an equally sized but oppositely directed contribution.

$$\Rightarrow \langle y\rangle_{\mathcal{G}(y,Y)} = 0 \tag{294}$$

$$\Rightarrow \langle f(y)\rangle_{\mathcal{G}(y,Y)} = 0, \text{ if } f(-y) = -f(y) \tag{295}$$

$$\begin{aligned}
\left\langle y\, y^\dagger\right\rangle_{\mathcal{G}(y,Y)} &= \int dy\, y\, y^\dagger \mathcal{G}(y, Y) &(296)\\
&= \int dx \mathcal{G}(x, X) O x x^\dagger O^\dagger &(297)\\
&= \underbrace{O\, [\int dx\, x\, x^\dagger \mathcal{G}(x, X)]}_{=X(\text{to be shown})}\, O^\dagger &(298)\\
&= O X O^\dagger = Y &(299)
\end{aligned}$$

In course of the proof we used,

$$\int dx\, x_i\, x_j\, \mathcal{G}(x,\, X) = \left[\prod_{k=1}^{n} \int dx_k \mathcal{G}(x_k, \sigma_k^2)\right] x_i\, x_j \tag{300}$$

$$= \begin{cases} \left[\int dx_i \mathcal{G}(x_i, \sigma_i^2)\, x_i\right] \left[\int dx_i \mathcal{G}(x_i, \sigma_i^2)\, x_i\right] & \text{if } i \neq j \\ \int dx_i \mathcal{G}(x_i, \sigma_i^2)\, x_i^2 & \text{if } i = j \end{cases} \tag{301}$$

$$= \begin{cases} 0 & \text{if } i \neq j \\ \sigma_i^2 & \text{if } i = j \end{cases} = \delta_{ij}\sigma_i^2 = X_{ij}. \tag{302}$$

$$\langle y y^\dagger \rangle_{\mathcal{G}(y,Y)} = Y \tag{303}$$

The expectation value of even powers a Gauss distributed random variable y can be calculated with the help of Wick's theorem.

Wick theorem (without proof):

$$\langle \prod_{j=1}^{2n} y_{i_j} \rangle_{\mathcal{G}(y,Y)} = \sum_{p \in \mathbb{P}} \prod_{(i',j') \in p} Y_{i_{i'} i_{j'}} \tag{304}$$

$\mathbb{P}$ is the set of all possible ways to partition $\{i_1, ..., i_{2n}\}$ into pairs.

Examples:

- $\langle y_{i_1} y_{i_2} \rangle_{\mathcal{G}(y,Y)} = Y_{i_1 i_2}$
- $\langle y_{i_1} y_{i_2} y_{i_3} y_{i_4} \rangle_{\mathcal{G}(y,Y)} = Y_{i_1 i_2} Y_{i_3 i_4} + Y_{i_1 i_3} Y_{i_2 i_4} + Y_{i_1 i_4} Y_{i_2 i_3}$

in particular:

- $\langle y_i^2 \rangle_{\mathcal{G}(y,Y)} = Y_{ii}$
- $\langle y_i^4 \rangle_{\mathcal{G}(y,Y)} = 3(Y_{ii})^2$
- $\langle y_i^6 \rangle_{\mathcal{G}(y,Y)} = 15(Y_{ii})^3$

$$\langle y_i^{2n} \rangle_{\mathcal{G}(y,Y)} = \frac{(2n)!}{2^n n!} (Y_{ii})^n \tag{305}$$

$$\langle y_i^{2n+1} \rangle_{\mathcal{G}(y,Y)} = 0 \tag{306}$$

5.3 MAXIMUM ENTROPY WITH KNOWN N-DIMENSIONAL 1ST AND 2ND MOMENTS

- I="unknown signal $s \in V$ =Vectorspace (e.g. $\mathbb{R}$, $\mathbb{R}^n$, $C(\mathbb{R}^n)$)"
- $q(s) = P(s|I) =$ const (to be set to 1 in calculation)

- $J =" \langle s \rangle_{(s|J,I)} = m, \, \langle (s-m)(s-m)^\dagger \rangle_{(s|J,I)} = S"$
  Constraints:

$$0 = \langle s - m \rangle = \int ds \, \mathcal{P}(s)(s-m) \tag{307}$$

$$0 = \langle (s-m)(s-m)^\dagger - S \rangle \tag{308}$$

- $p(s) = \frac{1}{\mathcal{Z}} \exp \left[\sum_{i=1}^n \mu_i (s-m)_i + \sum_{ij} \Lambda_{ij} \underbrace{\left((s-m)(s-m)^\dagger - S \right)_{ji}}_{=B_{ji}(s)} \right]$

1. calculate $\mathcal{Z}(\mu, \Lambda)$:

$$\mathcal{Z}(\mu, \Lambda) = \int ds \, \exp \left[\mu^\dagger \underbrace{(s-m)}_{s'} + \mathrm{Tr}[\Lambda B(s)] \right] \tag{309}$$

$$= \int ds' \exp \left[\mu^\dagger s' + \mathrm{Tr}[\Lambda (s' s'^\dagger - S)] \right] \tag{310}$$

$$= \int ds' \exp \left[\mu^\dagger s' + s'^\dagger \Lambda s' - \mathrm{Tr}[\Lambda S] \right] \tag{311}$$

2. determine μ and Λ:

$$\ln \mathcal{Z}(\mu, \Lambda) = -\mathrm{Tr}[\Lambda S] + \ln \left(\int ds' \, \exp(\mu^\dagger s' + s'^\dagger \Lambda s') \right)$$

$$\Rightarrow 0 \overset{!}{=} \frac{\partial \ln \mathcal{Z}(\mu, \Lambda)}{\partial \mu} \tag{312}$$

$$= \left(\frac{\partial \ln \mathcal{Z}}{\partial \mu_i} \right)_i \tag{313}$$

$$= \frac{\int ds' \, s' \exp(\mu^\dagger s' + s'^\dagger \Lambda s')}{\int ds' \, \exp(\mu^\dagger s' + s'^\dagger \Lambda s')} \tag{314}$$

$$\Rightarrow \mu = 0 \tag{315}$$

As then the integral in numerator is anti-symmetric with respect to $s' \to -s'$ and hence vanishes.

$$\Rightarrow 0 \stackrel{!}{=} \frac{\partial \ln \mathcal{Z}(\mu, \Lambda)}{\partial \Lambda} \tag{316}$$

$$= \left(\frac{\partial \ln \mathcal{Z}}{\partial \Lambda_{ij}}\right)_{ij} \tag{317}$$

$$= \underbrace{-(S_{ji})_{ij}}_{=-S} + \left(\frac{\int ds'\, s'_i s'_j \exp(s'^\dagger \Lambda s')}{\int ds'\, \exp(s'^\dagger \Lambda s')}\right)_{ij} \tag{318}$$

$$\Rightarrow S = \frac{\int ds'\, s' s'^\dagger \exp\left(-\frac{1}{2} s'^\dagger (-\frac{1}{2}\Lambda^{-1})^{-1} s'\right)}{\int ds' \exp\left(-\frac{1}{2} s'^\dagger (-\frac{1}{2}\Lambda^{-1})^{-1} s'\right)} \tag{319}$$

$$\frac{\int ds'\, s' s'^\dagger \mathcal{G}\left(s', -\frac{1}{2}\Lambda^{-1}\right)}{\int ds' \mathcal{G}\left(s', -\frac{1}{2}\Lambda^{-1}\right)} \tag{320}$$

$$= -\frac{1}{2}\Lambda^{-1} \tag{321}$$

$$\Rightarrow \Lambda = -\frac{1}{2} S^{-1} \tag{322}$$

Inserting the result in $\mathcal{Z}(\mu, \Lambda)$:

$$\mathcal{Z}(\mu, \Lambda) = \int ds' \exp\left[-\frac{1}{2} s'^\dagger S^{-1} s' + \frac{1}{2}\mathrm{Tr}[\underbrace{S^{-1} S}_{=\mathbb{1}}]\right] \tag{323}$$

$$= |2\pi S|^{1/2} e^{\frac{1}{2}\mathrm{Tr}[\mathbb{1}]} \tag{324}$$

3. calculate $P(s|J, I)$:

$$P(s|J, I) = \frac{1}{\sqrt{|2\pi S|}} \exp\left(-\frac{1}{2}(s-m)^\dagger S^{-1} (s-m)\right) \tag{325}$$

$$= \mathcal{G}(s-m, S) \tag{326}$$

$\Rightarrow$ In case only the mean $\langle s \rangle = m$ and the variance $\langle (s-m)(s-m)^\dagger \rangle = S$ are considered the safest assumption is to use a Gauss distribution $P(s|J, I) = \mathcal{G}(s - m, S)$ with this mean and variance.

Part II

INFORMATION FIELD THEORY

6 INFORMATION HAMILTONIAN

Bayes theorem:

$$\mathcal{P}(s|d,I) = \frac{\mathcal{P}(d|s,I)\mathcal{P}(s|I)}{\mathcal{P}(d|I)} \quad (327)$$

$$= \frac{\mathcal{P}(d,s|I)}{\int ds\,\mathcal{P}(d,s|I)} \quad (328)$$

$$= \frac{e^{-\mathcal{H}(d,s|I)}}{\mathcal{Z}(d)} \quad (329)$$

information Hamiltonian = "surprise or information":

$$\mathcal{H}(d,s|I) \equiv -\ln \mathcal{P}(d,s|I) \quad (330)$$

partition sum = "evidence":

$$\mathcal{Z}(d|I) \equiv \mathcal{P}(d|I) \quad (331)$$

$$= \int ds\,\mathcal{P}(d,s|I) \quad (332)$$

$$= \int ds\, e^{-\mathcal{H}(d,s|I)} \quad (333)$$

6.1 LINEAR MEASUREMENT OF A GAUSSIAN SIGNAL WITH GAUSSIAN NOISE

The simplest case of Bayesian reasoning on a continuous quantity s ($\in \mathbb{R}^n$, $\mathbb{C}^n$, or being a function) appears when prior and likelihood are Gaussians and the relation between signal and data is linear.
$I =$"$\mathcal{P}(s|I) = \mathcal{G}(s,S) = \frac{1}{\sqrt{|2\pi S|}} \exp\left(-\frac{1}{2}s^\dagger S^{-1} s\right)$, assuming that S is known. The data d depends on the signal s and the noise n via $d = Rs + n$ (either $d_i = \sum_j R_{ij} s_j + n_i$ or $d_i = \int dx\, R_{ix} s(x) + n_i$), where the response matrix R is known and the probability density of n is given by $\mathcal{P}(n|s,I) = \mathcal{G}(n,N)$ (N is known)."
So, what do we know about the possible values of s given the data d and the information I? To summarize the posterior knowledge on the signal s given data d and information I we have to calculate $\mathcal{P}(s|d,I) = \mathcal{P}(d,s|I)/\mathcal{P}(d|I)$
Calculation of the information Hamiltonian $\mathcal{H}(d,s|I)$:

- $\mathcal{H}(d,s|I) = -\ln \mathcal{P}(d,s|I) = -\ln(\mathcal{P}(d|s,I)\mathcal{P}(s|I)) = -\ln \mathcal{P}(d|s,I) - \ln \mathcal{P}(s|I) = \mathcal{H}(d|s,I) + \mathcal{H}(s|I)$
- $\mathcal{H}(s|I) = -\ln \mathcal{P}(s|I) = \frac{1}{2}s^\dagger S^{-1} s + \frac{1}{2}\ln|2\pi S|$
- $\mathcal{P}(d|s,I) = \int dn\,\mathcal{P}(d,n|s,I) = \int dn\,\underbrace{\mathcal{P}(d|s,n,I)}_{=\delta(d-(Rs+n))}\underbrace{\mathcal{P}(n|s,I)}_{=\mathcal{G}(n,N)} = \mathcal{G}(d-Rs,N)$

$$\begin{aligned}\Rightarrow \mathcal{H}(d|s,\,I) &= -\ln \mathcal{P}(d|s,\,I)\\ &= \frac{1}{2}(d-Rs)^{\dagger}N^{-1}(d-Rs)+\frac{1}{2}\ln|2\pi N|\\ &= \frac{1}{2}[d^{\dagger}N^{-1}d - s^{\dagger}\underbrace{R^{\dagger}N^{-1}d}_{\equiv j} - \underbrace{d^{\dagger}N^{-1}R}_{j^{\dagger}}s + s^{\dagger}R^{\dagger}N^{-1}Rs+\ln|2\pi N|]\\ &= \frac{1}{2}[s^{\dagger}R^{\dagger}N^{-1}Rs - s^{\dagger}j - j^{\dagger}s + d^{\dagger}N^{-1}d+\ln|2\pi N|]\end{aligned}$$

$$\begin{aligned} d &= Rs+n\\ \mathcal{P}(n,s) &= \mathcal{G}(n,N)\,\mathcal{G}(s,S)\\ \mathcal{H}(d,\,s) &= \mathcal{H}(d|s)+\mathcal{H}(s)\\ &= \frac{1}{2}[s^{\dagger}D^{-1}s - s^{\dagger}j - j^{\dagger}s]+\mathcal{H}_0,\end{aligned}$$

where we have defined a $\mathcal{H}_0$ independent of s,

$$\begin{aligned}\mathcal{H}_0 &= d^{\dagger}N^{-1}d+\ln|2\pi N|+\ln|2\pi S| \text{ and}\\ D^{-1} &= S^{-1}+R^{\dagger}N^{-1}R \text{ (information propagator)},\\ j &= R^{\dagger}N^{-1}d \text{ (information source)}.\end{aligned}$$

Quadratic completion:

We introduce the sign "$\widehat{=}$" to be the (context dependent) equality up to constant terms (with respect to the signal of the current context) and as the logarithmic brother of the proportionality sign "$\propto$".

$$\mathcal{H}(d,\,s|I) \;\widehat{=}\; \frac{1}{2}[s^{\dagger}D^{-1}s - j^{\dagger}s - s^{\dagger}j]$$

For the quadratic completion we use a trick, reading the D^{-1} as a multiplication sign and inserting the identities $\mathbb{1}=D^{-1}D$ and $\mathbb{1}=D\,D^{-1}$(exploiting that D^{-1} is invertible).

$$\begin{aligned}\mathcal{H}(d,s) &\;\widehat{=}\; \frac{1}{2}\left[s^{\dagger}D^{-1}s - j^{\dagger}D\,D^{-1}\,s - s^{\dagger}D^{-1}\,\underbrace{D\,j}_{=:\,m}\right]\\ &\;\widehat{=}\; \frac{1}{2}\left[s^{\dagger}D^{-1}s - \underbrace{(D\,j)^{\dagger}}_{=\,m^{\dagger}}\,D^{-1}\,s - s^{\dagger}D^{-1}m + m^{\dagger}D^{-1}m\right]\\ &= \frac{1}{2}\,(s-m)^{\dagger}\,D^{-1}(s-m).\end{aligned}$$

$$
\begin{aligned}
\Rightarrow \mathcal{H}(d, s|I) &= \mathcal{H}(d|s) + \mathcal{H}(s) \\
&= \frac{1}{2}(s-m)^{\dagger}D^{-1}(s-m) + \mathcal{H}_0' \\
\Rightarrow \mathcal{Z}(d) &= \int ds\, e^{-\mathcal{H}(d,s|I)} \\
&= \int ds\, e^{-\frac{1}{2}(s-m)^{\dagger}D^{-1}\underbrace{(s-m)}_{s'}-\mathcal{H}_0'} \\
&= e^{-\mathcal{H}_0'} \int ds'\, e^{-\frac{1}{2}s'^{\dagger}D^{-1}s'} \\
&= e^{-\mathcal{H}_0'}\sqrt{|2\pi D|} \\
\Rightarrow \mathcal{P}(s|d, I) &= \frac{\mathcal{P}(d, s|I)}{\mathcal{P}(d|I)} = \frac{e^{-\mathcal{H}(d,s)}}{\mathcal{Z}(d)} \\
&= \frac{e^{-\frac{1}{2}(s-m)^{\dagger}D^{-1}(s-m)-\mathcal{H}_0'}}{\sqrt{|2\pi D|}e^{-\mathcal{H}_0'}}
\end{aligned}
$$

We have therefore a Gaussian posterior,

$$\mathcal{P}(s|d, I) = \mathcal{G}(s-m, D)$$

with

- the mean

$$
\begin{aligned}
m &= \langle s \rangle_{(s|d,I)} = Dj \\
&= \left(S^{-1} + R^{\dagger}N^{-1}R\right)^{-1} R^{\dagger}N^{-1}d
\end{aligned}
$$

- the covariance

$$
\begin{aligned}
D &= \langle (s-m)(s-m)^{\dagger} \rangle_{(s|d,I)} \\
&= \left(S^{-1} + R^{\dagger}N^{-1}R\right)^{-1}
\end{aligned}
$$

- the information source

$$j = R^{\dagger}N^{-1}d.$$

The mean m is typically our best guess for the signal and the covariance describes the remaining uncertainty

$$s_x = m_x \pm \sqrt{D_{xx}} \text{ (1}\sigma\text{-range).}$$

The off-diagonal elements of D express how much the reconstruction uncertainty of two locations is correlated. The operation applied to the data to calculate $m = D\,R^{\dagger}N^{-1}d = F_{\mathrm{W}}\,d$ is called generalized Wiener filter.

7 LINEAR FILTER THEORY

7.1 OPTIMAL LINEAR FILTER

Also for a non-Gaussian, non-linear measurement situation linear correlations between data and signal can exist. Can we exploit these for signal inference?

I ="An unknown signal s is measured, yielding the data d. The covariances $\langle ss^\dagger\rangle_{(d,s)}$, $\langle ds^\dagger\rangle_{(d,s)}$ and $\langle dd^\dagger\rangle_{(d,s)}$ were determined previously."
What is the **optimal linear filter** F_L, which reconstructs the linear signal estimate via $m = F_\mathrm{L}\, d$?
Here, optimal should be a minimal expected root mean square (RMS) error E.

$$\begin{aligned}
E^2 &= \langle (s-m)^\dagger (s-m)\rangle_{(d,s)} \\
&= \sum_i \langle |s_i - m_i|^2\rangle_{(d,s)} \\
&= \langle s^\dagger s\rangle_{(d,s)} - \langle s^\dagger m\rangle_{(d,s)} - \langle m^\dagger s\rangle_{(d,s)} + \langle m^\dagger m\rangle_{(d,s)}
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \langle s^\dagger s\rangle_{(d,s)} &= \mathrm{Tr}\langle s^\dagger s\rangle_{(d,s)} = \mathrm{Tr}\langle ss^\dagger\rangle_{(d,s)} = \mathrm{Tr}S \\
\langle s^\dagger m\rangle_{(d,s)} &= \mathrm{Tr}\langle ms^\dagger\rangle_{(d,s)} = \mathrm{Tr}(F_\mathrm{L}\langle ds^\dagger\rangle_{(d,s)}) \\
\langle m^\dagger s\rangle_{(d,s)} &= \mathrm{Tr}\langle sm^\dagger\rangle_{(d,s)} = \mathrm{Tr}(\langle sd^\dagger\rangle_{(d,s)} F_\mathrm{L}^\dagger) \\
\langle m^\dagger m\rangle_{(d,s)} &= \mathrm{Tr}\langle mm^\dagger\rangle_{(d,s)} = \mathrm{Tr}(F_\mathrm{L}\langle dd^\dagger\rangle F_\mathrm{L}^\dagger)
\end{aligned}$$

For the calculations of the expectation values we exploited that the linear Filter F_L does not depend on signal s and data d.

$$\Rightarrow E^2 = \mathrm{Tr}\left[\langle ss^\dagger\rangle - F_\mathrm{L}\langle ds^\dagger\rangle - \langle sd^\dagger\rangle F_\mathrm{L}^\dagger + F_\mathrm{L}\langle dd^\dagger\rangle F_\mathrm{L}^\dagger\right]$$

The optimum is defined by a minimized error estimator $\frac{\partial E^2}{\partial F_L^\dagger} \stackrel{!}{=} 0$. For the partial derivation F_L and $F_L^\dagger$ can be regarded as independent quantities.

$$\frac{\partial E^2}{\partial F_\mathrm{L}^\dagger} = (0 - 0 - \langle sd^\dagger\rangle_{(d,s)} + F_\mathrm{L}\langle dd^\dagger\rangle_{(d,\,s)})^\dagger \stackrel{!}{=} 0$$

$$F_\mathrm{L} = \underbrace{\langle sd^\dagger\rangle_{(d,s)}}_{\text{crosscorrelation}} \underbrace{\langle dd^\dagger\rangle^{-1}_{(d,s)}}_{\text{autocorrelation matrix}}$$

The found optimal linear filter should also be correct in case of a linear measurement of a Gaussian signal and noise, and therefore we suspect,

$$F_\mathrm{L} \stackrel{?}{=} F_\mathrm{W} = \left(S^{-1} + R^\dagger N^{-1} R\right)^{-1} R^\dagger N^{-1}.$$

PROOF A:
Given a linear correlation between data and signal $d = Rs + n$ and Gaussian signal and noise $P(s,n) = \mathcal{G}(s,S)\mathcal{G}(n,N)$,

$$
\begin{aligned}
\Rightarrow \langle ss^\dagger\rangle_{(d,s)} &= \langle ss^\dagger\rangle_{(n,s)} = S \\
\langle ds^\dagger\rangle_{(d,s)} &= \langle (Rs+n)s^\dagger\rangle_{(n,s)} = R\,\underbrace{\langle ss^\dagger\rangle_{(n,s)}}_{=S} + \underbrace{\langle ns^\dagger\rangle_{(n,s)}}_{=0} = RS \\
\langle sd^\dagger\rangle_{(d,s)} &= S\,R^\dagger \qquad\qquad (\text{using } S = S^\dagger) \\
\langle dd^\dagger\rangle_{(d,s)} &= \langle (Rs+n)(Rs+n)^\dagger\rangle_{(s)} \\
&= R\langle ss^\dagger\rangle_{(n,s)}R^\dagger + R\,\underbrace{\langle sn^\dagger\rangle_{(n,s)}}_{=0} + \underbrace{\langle ns^\dagger\rangle_{(n,s)}}_{=0} R^\dagger + \langle nn^\dagger\rangle_{(n,s)} \\
&= RSR^\dagger + N
\end{aligned}
$$

$\Rightarrow$ optimal linear filter: $F_L = \langle sd^\dagger\rangle_{(d,s)}\langle dd^\dagger\rangle^{-1}_{(d,s)} = SR^\dagger(RSR^\dagger + N)^{-1}$
$\Rightarrow$ Wiener filter: $F_W = (S^{-1} + R^\dagger N^{-1}R)^{-1}R^\dagger N^{-1}$

$$
\begin{aligned}
\Rightarrow F_L &\stackrel{?}{=} F_W \\
SR^\dagger(RSR^\dagger + N)^{-1} &\stackrel{?}{=} (S^{-1} + R^\dagger N^{-1}R)^{-1}R^\dagger N^{-1} && \mid \cdot (RSR^\dagger + N) \text{ right} \\
SR^\dagger &\stackrel{?}{=} (S^{-1} + R^\dagger N^{-1}R)^{-1}R^\dagger N^{-1}(RSR^\dagger + N) && \mid (S^{-1} + R^\dagger N^{-1}R)\cdot \text{ left} \\
(S^{-1} + R^\dagger N^{-1}R)SR^\dagger &\stackrel{?}{=} R^\dagger N^{-1}(RSR^\dagger + N) \\
R^\dagger + R^\dagger N^{-1}RSR^\dagger &= R^\dagger N^{-1}RSR^\dagger + R^\dagger \quad \square
\end{aligned}
$$

PROOF B:
Consider two vector spaces (e.g. data space $\mathbb{D}$ and signal space $\mathbb{S}$) and two linear operators $A : \mathbb{D} \to \mathbb{S}$ and $B : \mathbb{S} \to \mathbb{D}$

$$
\begin{aligned}
A(BA + \mathbb{1}_\mathbb{D}) &= (AB + \mathbb{1}_\mathbb{S})A \\
\Rightarrow (AB + \mathbb{1}_\mathbb{S})^{-1}A &= A(BA + \mathbb{1}_\mathbb{D})^{-1}
\end{aligned}
$$

assuming that the above two inverses exist, as otherwise F_L and F_W are undefined. With $A = R^\dagger$ it follows:

$$
\begin{aligned}
F_L &= SR^\dagger(RSR^\dagger + N)^{-1} \\
&= SR^\dagger(N^{-1}RSR^\dagger + \mathbb{1}_\mathbb{D})^{-1}N^{-1} \\
&= S(R^\dagger N^{-1}RS + \mathbb{1}_\mathbb{S})R^\dagger N^{-1} \\
&= (R^\dagger N^{-1}R + S^{-1})^{-1}R^\dagger N^{-1} \\
&= F_W
\end{aligned}
$$

The Wiener filter is the optimal linear filter,

$$F_{\mathrm{L}} = F_{\mathrm{W}}.$$

$$F_{\mathrm{W}} = \underbrace{(S^{-1} + R^{\dagger}N^{-1}R)^{-1}}_{=D} R^{\dagger}N^{-1}$$

is the Wiener filter in the signal space and D is a signal space operation.

$$F_{\mathrm{L}} = SR^{\dagger} \underbrace{(RSR^{\dagger} + N)^{-1}}_{=\langle dd^{\dagger}\rangle^{-1}_{(d,s)}}$$

is the equivalent Wiener filter in data space and $\langle dd^{\dagger}\rangle_{(d,s)}$ is a data space operation.

However, the optimal linear filter is also defined in a non-Gaussian, non-linear measurement situation.
⇒Is it possible to define a linear response and noise in the non-Gaussian, non-linear case, as well?

- signal covariance:
$$\langle ss^{\dagger}\rangle_{(d,s)} =: S$$
- signal response:
$$\langle ds^{\dagger}\rangle_{(d,s)} =: RS$$
$$\begin{aligned} \Rightarrow R &= \langle ds^{\dagger}\rangle_{(d,s)} S^{-1} \\ &= \langle ds^{\dagger}\rangle_{(d,s)} \langle ss^{\dagger}\rangle^{-1}_{(d,s)} \end{aligned}$$

  R looks like the optimal linear filter for obtaining data d from a signal s.
- noise covariance:
$$\langle dd^{\dagger}\rangle =: RSR^{\dagger} + N$$

$$\begin{aligned} \Rightarrow N &= \langle dd^{\dagger}\rangle_{(d,s)} - RSR^{\dagger} \\ &= \langle dd^{\dagger}\rangle_{(d,s)} - \langle ds^{\dagger}\rangle_{(d,s)} \langle ss^{\dagger}\rangle^{-1}_{(d,s)} \langle ss^{\dagger}\rangle_{(d,s)} \langle ss^{\dagger}\rangle^{-1}_{(d,s)} \langle sd^{\dagger}\rangle_{(d,s)} \\ &= \langle dd^{\dagger}\rangle_{(d,s)} - \langle ds^{\dagger}\rangle_{(d,s)} \langle ss^{\dagger}\rangle^{-1}_{(d,s)} \langle sd^{\dagger}\rangle_{(d,s)} \end{aligned}$$

By construction of R, S, N we have $F_{\mathrm{L}} = F_{\mathrm{W}}$. The definition of the data $d = Rs + n$ also holds in the non-linear case, if we define the linear noise as,

$$n = d - Rs.$$

7.1.1 *Properties of the linear noise*

- correlation between linear noise and signal:

$$
\begin{aligned}
\langle ns^{\dagger}\rangle_{(d,s)} &= \langle (d-Rs)s^{\dagger}\rangle_{(d,s)} \\
&= \langle ds^{\dagger}\rangle - R\langle ss^{\dagger}\rangle \\
&= 0
\end{aligned}
$$

⇒Linear noise is by definition linearly uncorrelated to the signal.

- linear noise auto-correlation:

$$
\begin{aligned}
\langle nn^{\dagger}\rangle_{(d,s)} &= \langle (d-Rs)(d-Rs)^{\dagger}\rangle_{(d,s)} \\
&= \langle dd^{\dagger}\rangle_{(d,s)} - \langle ds^{\dagger}\rangle_{(d,s)}R^{\dagger} - R\langle sd^{\dagger}\rangle + R\langle ss^{\dagger}\rangle R^{\dagger} \\
&= (RSR^{\dagger}+N) - (RSR^{\dagger}) - (RSR^{\dagger}) + (RSR^{\dagger}) \\
&= N
\end{aligned}
$$

The linear response $R = \langle ds^{\dagger}\rangle_{(d,s)}\langle ss^{\dagger}\rangle^{-1}_{(d,s)}$ and linear additive noise $n = d - Rs$ can be defined for non-Gaussian, non-linear measurements, as well. The Wiener filter using those gives the optimal linear signal estimate. However, better non linear operations on the data may exist.

Example: $s \in \mathbb{R}$, $\mathcal{P}(s) = \mathcal{G}(s, \sigma^2)$, $d = f(s) = s^3$ is noiseless, non-linear data.
Moments: $\langle s\, s^{\dagger}\rangle_{(s)} = \sigma^2$, $\langle d\, s^{\dagger}\rangle_{(d,s)} = \langle s^4\rangle_{(s)} = 3\,\sigma^4$, $\langle d\, d^{\dagger}\rangle_{(d,s)} = \langle s^6\rangle_{(s)} = \frac{6!}{2^3 3!}\,\sigma^6 = 15\,\sigma^6$
Linear response: $R = \langle d\, s^{\dagger}\rangle_{(d,s)}\langle s\, s^{\dagger}\rangle^{-1}_{(d,s)} = 3\,\sigma^2$ increases with the signal variance probing more of the the non-linear part of the underlying non-linear response.
Noise covariance: $N = \langle d\, d^{\dagger}\rangle - \langle d\, s^{\dagger}\rangle\langle s\, s^{\dagger}\rangle^{-1}\langle s\, d^{\dagger}\rangle = 6\,\sigma^6$ increases with non-linarity.
Optimal linear filter: $F_{\mathrm{L}} = \langle s\, d^{\dagger}\rangle_{(d,s)}\langle d\, d^{\dagger}\rangle^{-1}_{(d,s)} = \frac{1}{5}\,\sigma^{-2}$ removes two signal powers from the data and scales it down.
Reconstruction error: $\langle (s - F_{\mathrm{L}} d)^2\rangle = \langle s^2\rangle - 2\,F_{\mathrm{L}}\langle s^4\rangle + F_{\mathrm{L}}^2\langle s^6\rangle = \left(1 - \frac{6}{5} + \frac{15}{25}\right)\,\sigma^2 = \frac{2}{5}\,\sigma^2$ increases with non-linearity.

Maximum Entropy perspective:
If we only know the covariances $\langle dd^{\dagger}\rangle_{(d,s)}$, $\langle ss^{\dagger}\rangle_{(d,s)}$, $\langle ds^{\dagger}\rangle_{(d,s)}$ we model $P(d,\, s)$ by a Gaussian with these constraints. The optimal signal estimate is then the Wiener filter.

7.2 SYMMETRY BETWEEN FILTER AND RESPONSE

$$
\begin{aligned}
P(n,\, s) &= \mathcal{G}(n,\, N)\mathcal{G}(s,\, S) \\
P(d,\, s) &= \int dn\, P(d,\, n\,, s) \\
&= \int dn\, \underbrace{P(d|n,\, s)}_{\delta(d-(Rs+n))}\, P(n,\, s) \\
&= \int dn\, \delta(d-(Rs+n))\mathcal{G}(n,\, N)\mathcal{G}(s,\, S) \\
&= \mathcal{G}(d-Rs,\, N)\mathcal{G}(s,\, S)
\end{aligned}
$$

- signal estimate:

$$
\begin{aligned}
\langle s\rangle_{(s|d)} &= F_W\, d = F_L d \\
&= \langle sd^{\dagger}\rangle_{(d,s)}\langle dd^{\dagger}\rangle^{-1}_{(d,s)} d
\end{aligned}
$$

- signal response:

$$\begin{aligned}\langle d\rangle_{(d|s)} &= \langle Rs+n\rangle_{(n|s)}\\ &= Rs+\underbrace{\langle n\rangle_{\mathcal{G}(n,N)}}_{=0}\\ &= Rs\\ &= \langle d^\dagger s\rangle_{(d,s)}\langle ss^\dagger\rangle^{-1}_{(d,s)}s\end{aligned}$$

There is a symmetry between filter and response by an exchange of data d and signal s,

$$\begin{aligned}\text{signal estimate} &\mathrel{\hat{=}} \text{data response}\\ \text{data estimate} &\mathrel{\hat{=}} \text{signal response.}\end{aligned}$$

We define a combined vector x,

$$x=\begin{pmatrix} d\\ s\end{pmatrix}$$

and the combined covariance X,

$$X=\langle xx^\dagger\rangle_{(x)}=\begin{pmatrix}\langle dd^\dagger\rangle_{(d,s)} & \langle ds^\dagger\rangle_{(d,s)}\\ \langle sd^\dagger\rangle_{(d,s)} & \langle ss^\dagger\rangle_{(d,s)}\end{pmatrix}.$$

The probability distribution of the combined vector x is give by a Gaussian,

$$P(x|\,X)=\mathcal{G}(x,\,X).$$

For subvectors x_a, x_b of x (e.g. $x_a=s$, $x_b=d$) the expectation value is,

$$m_a:=\langle x_a\rangle_{(x_a|x_b)}=X_{ab}(X_{bb})^{-1}x_b$$

and the probability distribution $P(x_a|x_b)$ is given by,

$$P(x_a|x_b)=\mathcal{G}(x_a-m_a,\,D_{aa})$$

with

$$D_{aa}=\left[\underbrace{X_{aa}^{-1}}_{=S^{-1}}+\underbrace{X_{aa}^{-1}X_{ab}}_{=R^\dagger}\underbrace{\left(X_{bb}-X_{ab}^\dagger X_{aa}^{-1}X_{ab}\right)^{-1}}_{=N^{-1}}\underbrace{X_{ab}^\dagger X_{aa}^{-1}}_{=R}\right]^{-1}.$$

7.3 RESPONSE

The signal s and the data d live in general in different spaces, the signal space and the data space. The response R translates between signal and data space. $R(s)$ is the image of the signal in data space.

- generic response:

$$R(s):=\langle d\rangle_{(d|s)}$$

- linear response:

$$R(s) = Rs \text{ with } R = \langle ds^\dagger \rangle_{(d,s)} \langle ss^\dagger \rangle^{-1}_{(d,s)}$$

7.3.1 *Repeated measurement of $s \in \mathbb{R}$*

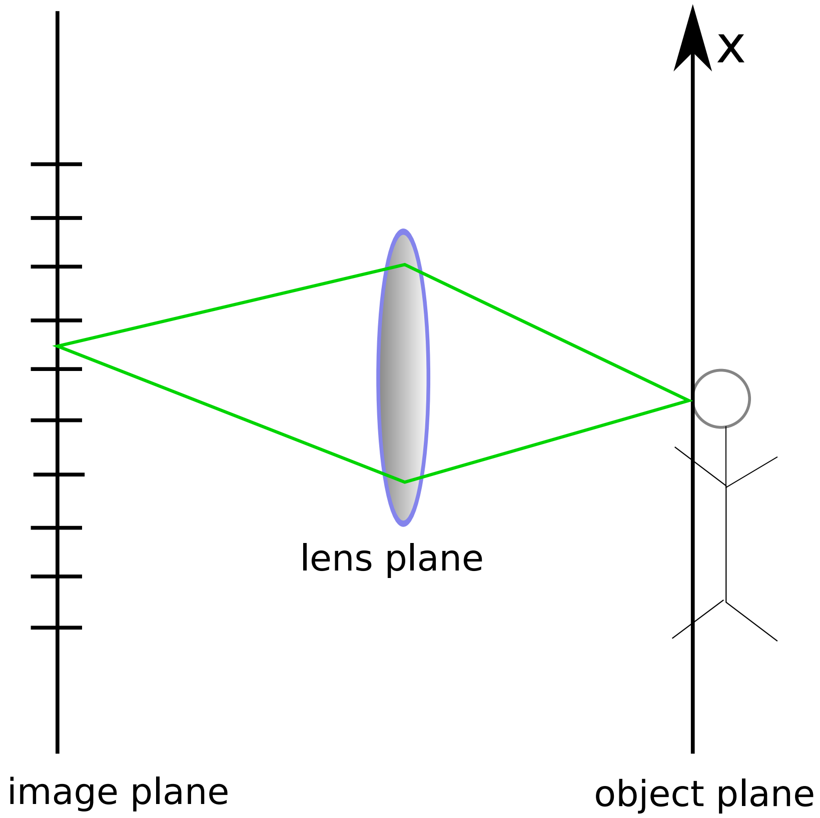
A single number $s \in \mathbb{R}$ is repeatedly measured n times . The response is the $1 \times \mathtt{n}$ matrix

$$R = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

that maps the 1-dimensional signal space to the n-dimensional data space, $\mathbb{R} \to \mathbb{R}^{\mathtt{n}}$, and specifically $s \to (s, s, \dots s)^{\mathtt{t}}$.

$$\begin{aligned} d_i &= Rs + n_i \\ d &= \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} s + \begin{pmatrix} n_1 \\ \vdots \\ n_n \end{pmatrix} \end{aligned}$$

7.3.2 *Photography*



The response translates between a 2-dimensional, continuous signal space and a m-dimensional, discrete data space,

$$R : C(\mathbb{R}^2) \to \mathbb{R}^m.$$

For example, the signal might be the brightness distribution within the image plain and an individual detector i in the focal plain of the camera measures the amount of light that is focused onto it from some small, but extended area in the image plain,

$$R_i : C(\mathbb{R}^2) \to \mathbb{R}.$$

The recorded datum d_i depending on the brightness distribution $s(x)$ and its point spread function $R_i(x)$ is then,

$$d_i = (R\,s + n)_i = \int_{\mathbb{R}^2} d^2x\, R_i(x)\, s(x) + n_i.$$

7.3.3 *Tomography*

A volume $\Omega = \mathbb{R}^u$ is probed by a set of rays of the form $x_i(t) = a_i + t\, b_i$ with $a_i \in \Omega$ the location of a detector and $b_i \in \mathcal{S}^{u-1}$ a direction. Each datum

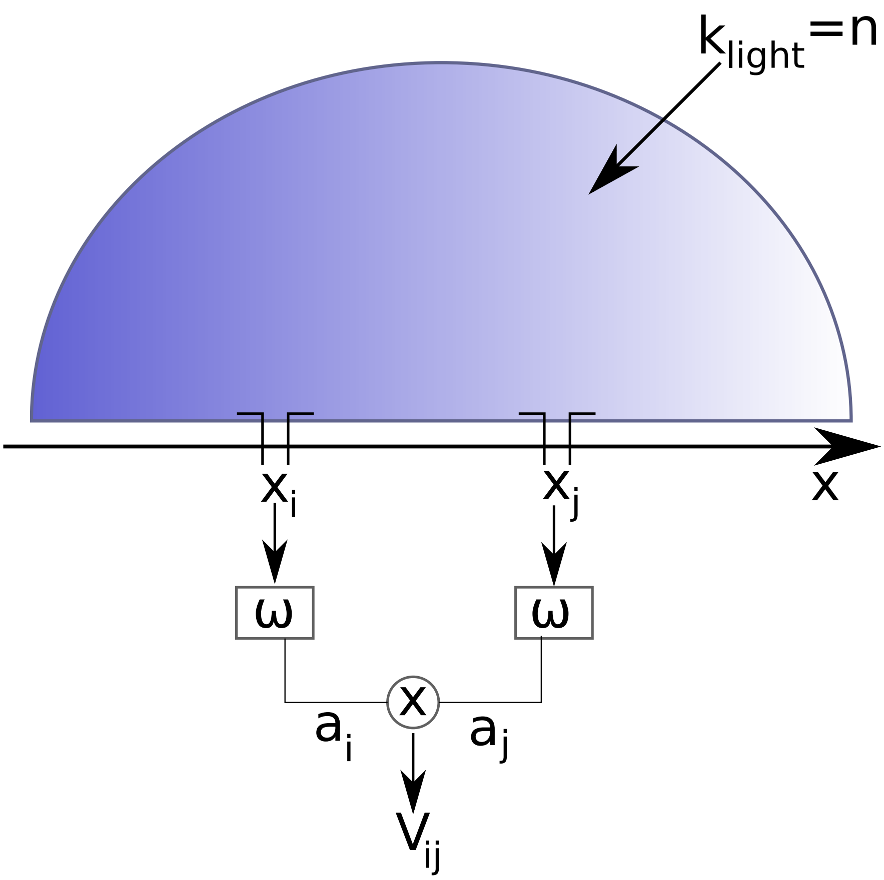
$$d_i = \int_0^{t_{\mathrm{max,i}}} dt\, s(x_i(t)) + n_i$$

contains the ray integrated signal field, e.g. the ray-integrated opacity of an absorbing medium. The response operator is therefore

$$R_{ix} = \int_0^{t_{\mathrm{max,i}}} dt\, \delta(x - x_i(t)).$$

7.3.4 *Interferometry*

An interferometer measures individual components of the Fourier transformed sky brightness $s_{\hat{n}}$ distribution by measuring the interference pattern produced by two apertures recording the electromagnetic waves of wavelength $\lambda = c/\omega$.



The measured data d_{ij} is the visibility $V_{ij} = \langle a_i a_j \rangle_{\text{time average}}$ with amplitude

$$a_i = \int_{s^2} d\hat{n}\, \sqrt{s_{\hat{n}}} \exp[i(\omega t + \varphi(\hat{n},\, t) + \frac{\omega}{c} \hat{n} \vec{x}_i)].$$

Calculate visibility:

$$
\begin{aligned}
V_{ij} &= \langle a_i a_j \rangle_t \\
&= \int d\hat{n} \int d\hat{n}' \sqrt{s_{\hat{n}} s_{\hat{n}'}} \langle e^{i(\omega t - \omega t + \varphi(\hat{n},t) - \varphi(\hat{n}',t))} \rangle \cdot e^{i\frac{\omega}{c}(\hat{n}\vec{x}_i - \hat{n}'\vec{x}_j)} \\
&= \int d\hat{n} \int d\hat{n}' \sqrt{s_{\hat{n}} s_{\hat{n}'}} \underbrace{\langle e^{i(\varphi(\hat{n},t) - \varphi(\hat{n}',t))} \rangle}_{=\delta(\hat{n}-\hat{n}')} \cdot e^{i\frac{\omega}{c}(\hat{n}\vec{x}_i - \hat{n}'\vec{x}_j)} \\
&= \int d\hat{n} \sqrt{s_{\hat{n}} s_{\hat{n}}} \exp[i \underbrace{\left(\frac{\vec{x}_i - \vec{x}_j}{\lambda}\right)}_{=\vec{k_{ij}}} \cdot \hat{n}] \\
&= \int d\hat{n}\, s_{\hat{n}} e^{i\hat{n}\vec{k}_{ij}}
\end{aligned}
$$

8 GAUSSIAN FIELDS

In Sect. 5.2 we already discussed the multivariate Gaussian,

$$\mathcal{G}(y, Y) = \frac{1}{\sqrt{|2\pi Y|}} e^{-\frac{1}{2} y^\dagger Y^{-1} y}$$

with,

$$Y = \langle y y^\dagger \rangle,\ y \in \mathbb{R}^n.$$

$\Rightarrow$All involved quantities – y, Y, $|Y|$, and Y^{-1} – can be written without specifying n, the number of degrees of freedom of y. Can one therefore take the limit $n \to \infty$?

The definition of a vector with correlated Gaussian distributed components can be generalized to a field with Gaussian statistics. Let $\varphi : \mathbb{R}^u \to \mathbb{R}$ be a field with Gaussian statistics.

Notation: We regard $\varphi = \varphi^x e_x$ (Einstein summation) as a vector in a Hilbert space with the contravariant components $\varphi^x = \varphi(x)$, where $x \in \mathbb{R}^u$. Contravariant means that if we change (e.g. scale) the unit system $e = (e_x)_x$ in which we measure the field (at location x) via $e' = Ae$ (e.g. with A a diagonal scaling matrix), the transformed field components are changed with the inverse of this, $\varphi'^x = \left(A^{-1}\right)^x_y \varphi^y$, such that the total vector stays invariant: $\varphi' = \varphi'^x e'_x = \left(A^{-1}\right)^x_y \varphi^y A^z_x e_z = \varphi^y \left(A^{-1}\right)^x_y A^z_x e_z = \varphi^y \delta^z_y e_z = \varphi^y e_y = \varphi$. One functional basis of the Hilbert space are the delta functions $e_x(y) = \delta(x-y)$. The scalar product is the integration $\psi^\dagger \varphi = \int dx\, \overline{\psi(x)}\, \varphi(x) \equiv \overline{\psi_x}\, \varphi^x$. Thus, $\varphi(x) = \varphi^y e_y(x) = \int dy\, \varphi^y \delta(x-y) = \varphi^x$.

Let us discretize φ with n pixels $X_{(n)} = \{x_1, ..., x_n\}$. Then we denote $\varphi_{(n)} = (\varphi^{x_1}, \ldots, \varphi^{x_n})^t$ the n-dimensional vector of field values at these pixel locations. The continuous field φ is said to have a Gaussian probability distribution if for any such finite subset $X_{(n)} \subset \mathbb{R}^u$ the vector $\varphi_{(n)}$ has a multivariate Gaussian distribution:

$$\mathcal{P}(\varphi_{(n)}) = \mathcal{G}(\varphi_{(n)}, \Phi_{(n)})$$

with

$$\Phi^{ij}_{(n)} = \langle \varphi^i_{(n)} \overline{\varphi^j_{(n)}} \rangle = \langle \varphi(x_i) \overline{\varphi(x_j)} \rangle.$$

Gaussian field distribution:

$$\begin{aligned}
\mathcal{G}(\varphi, \Phi) &\equiv \frac{1}{\sqrt{|2\pi\Phi|}} \exp\left(-\frac{1}{2}\varphi^\dagger \Phi^{-1}\, \varphi\right) \\
&= \frac{1}{\sqrt{|2\pi\Phi|}} \exp\left(-\frac{1}{2}\overline{\varphi^x} \left(\Phi^{-1}\right)_{xy} \varphi^y\right) \\
&\equiv \lim_{n\to\infty} \mathcal{G}(\varphi_{(n)}, \Phi_{(n)})
\end{aligned}$$

Gaussian field distribution

$$\Rightarrow \langle f(\varphi)\rangle_{(\varphi|\Phi)} = \int \mathcal{D}\varphi\, \mathcal{P}(\varphi|\Phi)\, f(\varphi)$$
$$= \lim_{n\to\infty} \left[\prod_{i=1}^{n} \int d\varphi^{i}_{(n)}\right] \mathcal{G}(\varphi_{(n)}, \Phi_{(n)})\, f(\varphi_{(n)}).$$

8.1 FIELD THEORY

Scalar Product
discrete case: $j^{\dagger}\varphi = \overline{j_i}\varphi^{i}$
continuous case: $j^{\dagger}\varphi = \int dx\, \overline{j(x)}\, \varphi(x) \equiv \overline{j_x}\varphi^{x}$
Derivative
discrete case: $\partial_{\varphi^i} j^{\dagger}\varphi = \partial_{\varphi^i}\overline{j_i}\varphi^{i} = \overline{j_i}$
continuous case: $\partial_{\varphi^x} j^{\dagger}\varphi = \frac{\delta}{\delta\varphi^x}\int dx'\, \overline{j_{x'}}\, \varphi^{x'} = \overline{j_x} \quad \Rightarrow \partial_{\varphi} j^{\dagger}\varphi = \overline{j}$
Normalisation Factors
discrete case: $|\Phi| = \prod_{i=1}^{n}\lambda_i$ (λ_i are the eigenvalues)
continuous case: $|\Phi| = \lim_{n\to\infty}\prod_{i=1}^{n}\lambda_i$ (might be undetermined)
Covariance Matrix
discrete case: $\Phi^{ij} = \langle \varphi^{i}\bar{\varphi}^{j}\rangle$
continuous case: $\Phi^{xy} = \langle \varphi^{x}\bar{\varphi}^{y}\rangle_{(\varphi)} = \left(\langle \varphi\varphi^{\dagger}\rangle_{(\varphi)}\right)^{xy}$
Inverse Covariance
discrete case: $\Phi^{-1}\Phi = \mathbb{1}$
continuous case: $\int dy\, \Phi^{-1}_{xy}\Phi^{yz} = \mathbb{1}^{z}_{x} = \delta(x-z)$
Wick Theorem
$\langle \varphi^{x}\varphi^{y}\varphi^{z}\varphi^{w}\rangle_{\mathcal{G}(\varphi,\Phi)} = \Phi^{xy}\Phi^{zw} + \Phi^{xz}\Phi^{yw} + \Phi^{yw}\Phi^{yz}$

9 WIENER FILTER THEORY

$$
\begin{aligned}
d &= Rs + n \\
d^i &= R^i_x\, s^x + n^x \text{ (R maps signal into data space)} \\
P(n,s) &= \mathcal{G}(s,\, S)\,\mathcal{G}(n,\, N) \\
\Rightarrow P(s|d) &= \mathcal{G}(s-m,\, D) \\
m &= Dj = D^{xy} j_y \\
D &= (S^{-1} + \underbrace{R^\dagger N^{-1} R}_{=M})^{-1} \\
j &= R^\dagger N^{-1} d \\
j_x &= \overline{R}^i_x \left(N^{-1}\right)_{ij} d^j
\end{aligned}
$$

9.1 STATISTICAL HOMOGENEITY

Imagine we are interested in an unknown signal over real space ($s,\, d,\, n : \mathbb{R}^u \to \mathbb{R},\, \mathbb{C}$) with known Gaussian statistics and complete data

$$
\begin{aligned}
R &= \mathbb{1} \\
d &= s + n.
\end{aligned}
$$

Furthermore, request a statistical homogeneous signal and noise. Consequently, S^{xy} and N^{xy} can not depend on an absolute location x, however, it can depend on relative distances $x - y$,

$$
\begin{aligned}
S^{xy} &= \langle s^x s^y \rangle_{(s)} = C_s(x-y) \\
N^{xy} &= \langle n^x n^y \rangle_{(n)} = C_n(x-y).
\end{aligned}
$$

The maximum of the covariance of the signal is given for $x = y$. Possible correlation functions are shown in Fig. 4.

9.2 FOURIER SPACE

There are a number of different conventions on how to define the Fourier transformation of a function $f : \mathbb{R}^u \to \mathbb{C}$. The most natural one should be symmetric between Fourier and inverse Fourier transform and is given by

$$
\begin{aligned}
f(k\!\!\!/) &= \int dx\, e^{2\pi i k\!\!\!/ x}\, f(x) \\
f(x) &= \int dk\!\!\!/\, e^{-2\pi i k\!\!\!/ x}\, f(k\!\!\!/).
\end{aligned}
$$

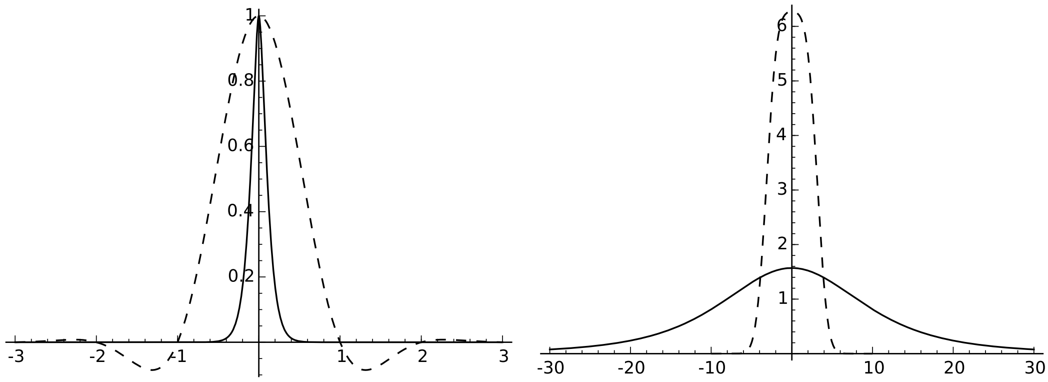


Figure 4: Possible correlation functions (left) and their Fourier space representations (right)

In physics and many other areas, however, it is convention to absorb the 2π factors in the exponential function into the variable $k = 2\pi \not{k}$. Transforming to this coordinate, the Fourier transforms read

$$\begin{aligned} f(k) &= \int dx^u \, e^{ikx} f(x) \\ f(x) &= \int \frac{dk^u}{(2\pi)^u} \, e^{-ikx} f(k). \end{aligned}$$

We regard the function as an abstract vector, and the argument is more an index in a given vector basis. Consequently, we will use also the notations $f_x \equiv f(x)$ and $f_k \equiv f(k) \equiv \int dx \, e^{ikx} f(x)$.

Fourier transformation operator:

$$F_x^k = e^{ikx},$$

which should be applied to a function by using the real space scalar product $a^\dagger b = \int dx \, \overline{a}_x b_x$, the inverse Fourier operator F^{-1} with

$$(F^{-1})_k^x = e^{-ikx},$$

which should be applied to a (Fourier space) function by using the Fourier space scalar product $a^\dagger b = \int dk/(2\pi)^u \, \overline{a}_k b^k$. Note that these are related by

$$F^{-1} = F^\dagger.$$

$\Rightarrow$ The Fourier transformation is an orthonormal transformation in function spaces, a sort of high dimensional rotation.

9.3 POWER SPECTRA

Now we can express the statistical homogeneous signal covariance matrix also in Fourier space:

$$
\begin{aligned}
S^{kk'} &= \left\langle s^k \bar{s}^{k'} \right\rangle_{(s)} = \left\langle (Fs)^k \, \overline{(Fs)}^{k'} \right\rangle_{(s)} = \left\langle (Fs)^k \, (Fs)^{\dagger k'} \right\rangle_{(s)} \\
&= \left\langle (Fs)^k \left(s^\dagger F^\dagger \right)^{k'} \right\rangle_{(s)} = \left(F \left\langle s \, s^\dagger \right\rangle_{(s)} F^\dagger \right)^{kk'} \\
&= \left(F \, S \, F^\dagger \right)^{kk'} = \left(F^k_x \, S^{xy} \;\; F^{\dagger}{}^{k'}_{y} \right)\Big|_{\text{Einstein sum}} \\
&= \int dx \, e^{ikx} \int dy \, S^{xy} \, e^{-ik'y} \\
&= \int dx \int dy \, e^{i(kx - k'y)} \, C_s(x-y) \\
&= \int dx \int dr \, e^{i(kx - k'\,(x-r))} \, C_s(r) \Big|_{y=x-r} \\
&= \underbrace{\int dx \, e^{i(k-k')\,x}}_{(2\pi)^u \delta(k-k')} \underbrace{\int dr \, e^{ik'\,r} \, C_s(r)}_{P_s(k')} \\
&= (2\pi)^u \;\; \delta(k-k') \, P_s(k),
\end{aligned}
$$

where we use k' as a second Fourier space coordinate, the Einstein notation to sum over repeated indexes (the coordinates x and y), and statistical homogenity. $P_s(k)$ is the Fourier transformed correlation function, the so called power spectrum.

9.3.1 *Units*

- $[s^k] = \int dx \, e^{ikx} s^x = V \, [s^x]$
- $[C_s(r)] = [s^x]^2$
- $[P_s(k)] = [\int dr \, e^{ikr} C_s(r)] = V [s^x]^2 = \frac{[s^k]^2}{V}$
- $[\delta(k-k')] = \left[\frac{1}{k\text{–Volume}} \right] = V$

$$
\boxed{
\begin{aligned}
\Rightarrow P_s(k) &= \frac{\langle |s^k|^2 \rangle}{V} \\
S^{kk'} &= (2\pi)^u \delta(k-k') P_s(k) \\
&= \langle s^k \bar{s}^{k'} \rangle_{(s)} \\
&= \mathbb{1}^{kk'} \frac{\langle |s^k|^2 \rangle}{V} \text{ (no summation over } k\text{!)} \\
\mathbb{1}^{kk'} &= (2\pi)^u \delta(k-k')
\end{aligned}
}
$$

9.3.2 *Wiener-Khintchin Theorem*

Wiener–Khintchin theorem

A statistical homogeneous signal s over Cartesian space with stationary auto-correlation $S^{xy} = \langle s^x \overline{s^y} \rangle_{(s)} = C_s(x-y)$ has a diagonal covariance matrix in Fourier space,

$$S^{kk'} = \langle s^k \overline{s^{k'}} \rangle_{(s)} = (2\pi)^u \delta(k-k')\, C_s(k).$$

The diagonal elements are given by the Fourier transformed auto-correlation function $C_s(k)$, which is identical to the power spectrum per volume V, $P_s(k) = \lim_{V\to\infty} \frac{1}{V} \langle | \int_V dx\, s^x e^{ikx}|^2 \rangle_{(s)} = C_s(k)$.

The Fourier space noise covariance is, since we also assume statistical homogeneous noise, similarly

$$N^{kk'} = \langle n^k \overline{n^{k'}} \rangle_{(s)} = (2\pi)^u \delta(k-k')\, C_n(k)$$

with the Fourier transformed noise covariance being identical to the noise power spectrum as well, $C_n(k) = P_n(k)$.

9.3.3 *Fourier space filter*

In order to calculate the mean $m = D\, N^{-1} d$ and variance $D = (S^{-1} + N^{-1})^{-1}$ of our Gaussian posterior $\mathcal{P}(s|d) = \mathcal{G}(s-m, D)$ we need the inverse of S, the matrix S^{-1}, which fulfills

$$\mathbb{1} = S^{-1} S.$$

In Fourier space this becomes particularly simple:

$$\begin{aligned}
\mathbb{1}_q^k &= \left(S^{-1}S\right)_q^k \iff \\
\mathbb{1}_q^k &= (2\pi)^u \delta(k-q) = S^{kk'} \left(S^{-1}\right)_{k'q} \\
&= \int \frac{dk'}{(2\pi)^u} (2\pi)^u \delta(k-k') P_s(k) \left(S^{-1}\right)_{k'q} \\
&= \left(S^{-1}\right)_{kq} P_s(k) \iff \\
\Rightarrow \left(S^{-1}\right)_{kq} &= \frac{(2\pi)^u \delta(k-q)}{P_s(k)} \\
\Rightarrow \left(N^{-1}\right)_{kq} &= \frac{(2\pi)^u \delta(k-q)}{P_n(k)} \\
\Rightarrow M_{kq} &= (R^\dagger N^{-1} R)_{kq} = (N^{-1})_{kq}
\end{aligned}$$

$$\boxed{\begin{aligned} \Rightarrow D^{kq} &= (S^{-1} + \underbrace{R^{\dagger}N^{-1}R}_{=M})^{-1\,kq} \\ &= (2\pi)^u \delta(k-q) \left([P_s(k)]^{-1} + [P_n(k)]^{-1} \right)^{-1} \end{aligned}}$$

$$\begin{aligned} \Rightarrow j_k &= (R^{\dagger}N^{-1}d)_k \\ &= \int \frac{dk'}{(2\pi)^u} (2\pi)^u \delta(k'-k) [P_n(k)]^{-1} d^{k'} \\ &= \frac{d_k}{P_n(k)} \end{aligned}$$

$$\begin{aligned} \Rightarrow m^k &= (Dj)^k = D^{kk'} j_{k'} \\ &= \int \frac{dk'}{(2\pi)^u} \frac{(2\pi)^u \delta(k-k')}{\frac{1}{P_s(k)} + \frac{1}{P_n(k)}} \frac{d^{k'}}{P_n(k')} \end{aligned}$$

$$\boxed{\begin{aligned} \Rightarrow m^k &= (Dj)^k \\ &= \underbrace{\frac{1}{1 + \frac{P_n(k)}{P_s(k)}}}_{f(k) = \text{filter function}} d^k \end{aligned}}$$

The filter function $f(k)$ reweighs all Fourier modes of the data independently and according to the ratio of the expected signal and noise power at this mode.

$$\begin{aligned} f(k) &= \frac{1}{1 + \frac{P_n(k)}{P_s(k)}} \\ &= \begin{cases} 1 & \text{if } P_s(k) \gg P_n(k) \text{ (perfect pass through)} \\ \underbrace{\frac{P_s(k)}{P_n(k)}}_{\ll 1} & \text{if } P_s(k) \ll P_n(k) \text{ (signal-to-noise weighting)} \end{cases} \end{aligned}$$

The signal reconstruction m is not only a filtered versions of the data, it is usually also a filtered version of the signal,<

$$m^k = f(k)\, d^k = f(k)\, (s^k + n^k) = \left(\frac{s+n}{1 + P_n/P_s} \right)^k$$

9.3.4 *Position space filter*

It is also instructive to investigate the Wiener filter in position space.

- reconstructed signal in position space:

$$s^x = \int \frac{dk}{(2\pi)^u} s^k e^{-ikx}$$

- reconstructed mean in position space:

$$\begin{aligned} m^x &= \int \frac{dk^u}{(2\pi)^u}\, e^{-ikx} \underbrace{m^k}_{=f(k)\,d^k} \\ &= \int \frac{dk^u}{(2\pi)^u}\, e^{-ikx}\, f^k \int dy^u\, e^{iky}\, d^y \\ &= \int dy^u \underbrace{\int \frac{dk^u}{(2\pi)^u}\, e^{-ik(x-y)}\, f(k)}_{f(x-y)}\, d^y \\ &= \int dy^u f(x-y)\, d^y = (f * d)^x \end{aligned}$$

$\Rightarrow$The posterior mean map is the data convolved with a position space kernel function given by the Fourier transformed spectral filter

$$\begin{aligned} f(r) &= \int \frac{dk^u}{(2\pi)^u}\, e^{-ikr} f(k) \\ &= \int \frac{dk^u}{(2\pi)^u}\, \frac{e^{-ikr}}{1 + {}^{P_n(k)}/_{P_s(k)}} \end{aligned}$$

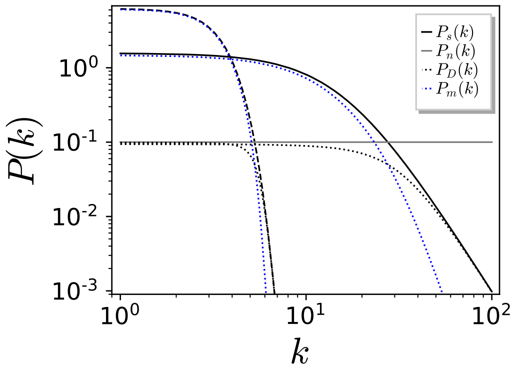
- power spectrum of the mean m:

$$\begin{aligned} P_m(k) &= \frac{1}{V}\langle |m^k|^2 \rangle_{(d,s)} \\ &= \frac{1}{V}\langle |f(k)|^2 |d^k|^2 \rangle_{(d,s)} \\ &= \frac{1}{(1 + {}^{P_n}/_{P_s})^2}\,(P_s(k) + P_n(k)) \\ &= \frac{P_s^2(k)}{P_s(k) + P_n(k)} \\ &= \frac{P_s(k)}{1 + {}^{P_n(k)}/_{P_s(k)}} \end{aligned}$$

9.3.5 *Example: large-scale signal*

- Assume white noise:

$$\begin{aligned} N^{xy} &= \langle n^x n^y \rangle_{(n)} \\ &= \delta(x-y)\sigma_n^2 \\ &= C_n(x-y) \\ N^{kq} &= \int dx \int dy\, e^{ikx} \delta(x-y) \sigma_n^2 e^{-iqy} \\ &= \sigma_n^2 \int dx\, e^{i(k-q)x} \\ &= (2\pi)^u \delta(k-q) \underbrace{\sigma_n^2}_{=P_n(k)} \end{aligned}$$



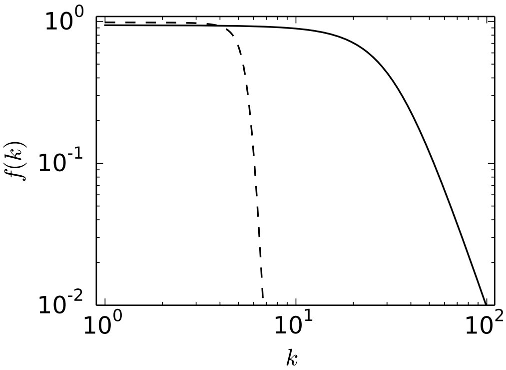


Figure 5: Left: On a log-log scale, possible spectra of signals ($P_s(k)$, black solid and dashed lines, corresponding to the ones in Fig. 4), noise ($P_n(k)$, gray horizontal line), and resulting posterior mean ($P_m(k)$, blue dotted lines) and uncertainties ($P_D(k)$, black dotted lines), each belonging to the signal spectra shown next to it. The noise spectrum is white and the relation $P_m(k) + P_D(k) = P_s(k)$ holds. On small Fourier scales or large spatial scales the signals are reconstructed accurately, up to the point of a signal to noise ratio of one, beyond which little of the signal can be recovered due to the dominating noise there. Right: Corresponding filter functions to be applied to the data to suppress the noise.

$\Rightarrow$ White noise has a constant power spectrum $P_n(k) = \sigma_n^2$.

- Assume a signal $s : \mathbb{R} \to \mathbb{R}$ with a red signal spectrum $P_s(k) = \sigma_s^2\,(k/k_0)^{-2}$

The Wiener filter is given by the spectral filter function

$$f(k) = \frac{1}{1 + P_n(k)/P_s(k)} = \frac{1}{1 + \underbrace{\frac{\sigma_n^2}{\sigma_s^2 k_0^2}}_{=q^{-2}} k^2} = \frac{q^2}{k^2 + q^2}$$

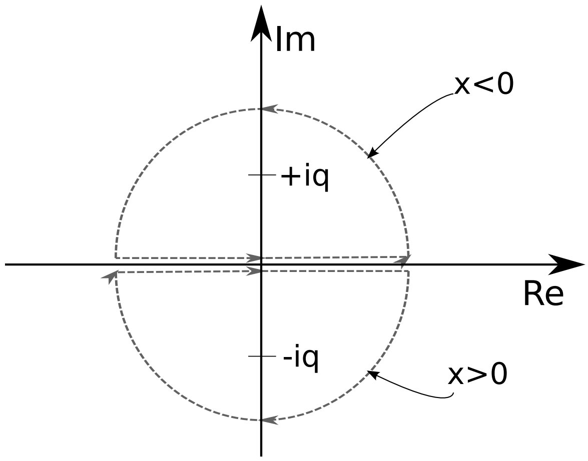
with $q = \sigma_s k_0/\sigma_n$ being the cut-off wave number of the filter. The position space Wiener filter kernel is then

$$f(x) = \int \frac{dk}{2\pi} \frac{q^2}{q^2 + k^2}\, e^{-ikx} = \frac{q^2}{2\pi} \int_{-\infty}^{\infty} dk\, \frac{e^{-ikx}}{(k + iq)\,(k - iq)}.$$

The integrand has two poles, at $k = \pm iq$, respectively. The function $f(x)$ can be calculated via Cauchy's residue theorem, which states that the integral of an analytical function $f(k)$ over a closed path γ in the complex plane is given by the sum over the residues of the function at its poles inside the path,

Cauchy's residue theorem

$$\oint_\gamma dk\, f(k) = 2\pi i \sum_{l=1}^{n} I(\gamma, k_l)\,\mathrm{Res}(f, k_l).$$



- $\{k_1, \dots k_n\}$ denotes the singularities of f inside γ
- $I(\gamma, k)$ is the winding number of the path with respect to a point k
- The residuum is usually just given by $\text{Res}(f, k_l) = f(k)\,(k - k_l)|_{k=k_l}$

For $x < 0$:

$$\begin{aligned} f(x) &= iq^2(+1)\,\frac{e^{-ikx}}{k+iq}\bigg|_{k=+iq} \\ &= \frac{iq^2 e^{-i(iq)x}}{2iq} \\ &= \frac{q}{2}e^{-qx} \end{aligned}$$

Combining the solutions of $f(x)$ for $x < 0$ and $x > 0$ we get,

$$\boxed{\begin{aligned} f(x) &= \frac{q}{2}\,e^{-q\,|x|} = \frac{1}{2\,\lambda}\,e^{-|x|/\lambda}, \text{ with} \\ q &= \frac{\sigma_s k_0}{\sigma_n} \text{ and} \\ \lambda &= \frac{1}{q} = \frac{\sigma_n}{\sigma_s k_0} \text{ the corrlelation length of } f(x)\text{, since} \\ \lambda &= \int_0^\infty dx\,\frac{f(x)}{f(0)} = \int_0^\infty dx\,e^{-|x|/\lambda}. \end{aligned}}$$

9.3.6 *Deconvolution*

I : "The measurement equation reads $d^x = \int dy\, R^x_y\, s^y + n^x$ with the translational invariant convolution kernel $R^x_y = b(x-y)$, so that $d = b * s + n$. We assume for simplicity, $P(s, n) = \mathcal{G}(s, S)\mathcal{G}(n, N)$, with known response R and known covariances S and N. In particular S, N are homogenous."

$$d^y = \int dx\, b(y-x)s^x + n^y$$

$\Rightarrow$ Fourier space:

$$
\begin{aligned}
d^k &= (b * s)^k + n^k \\
&= \int dy\, e^{iky} \left[\int dx\, b(y-x)\, s^x + n^y \right] \\
\Rightarrow (b * s)^k &= \int dx \int dy\, e^{iky} \int \frac{dk'}{(2\pi)^u}\, e^{-ik'(y-x)} b(k') \int \frac{dk''}{(2\pi)^u}\, e^{-ik''x}\, s^{k''} \\
&= \int \frac{dk'}{(2\pi)^u} \int \frac{dk''}{(2\pi)^u} \underbrace{\int dx\, e^{i(k-k')y}}_{(2\pi)^u \delta(k-k')} \underbrace{\int dy\, e^{i(k'-k'')x}}_{(2\pi)^u \delta(k'-k'')} b(k')\, s^{k''} \\
&= b(k)\, s^k \\
\Rightarrow R^k_{k'} &= (2\pi)^u \delta(k-k')\, b(k)
\end{aligned}
$$

$\Rightarrow$ The convolution response turns out to be as well diagonal in Fourier space, $R^k_{k'} = (2\pi)^u \delta(k-k')\, b(k)$, as are the signal covariance $S^{kk'} = (2\pi)^u \delta(k-k')\, P_s(k)$, the noise covariance $N_{kk'} = (2\pi)^u \delta(k-k')\, P_n(k)$, and consequently the uncertainty covariance $D^{kk'} = (2\pi)^u \delta(k-k')\, P_D(k)$.

- Calculation of the spectrum $P_D(k)$:

$$
\begin{aligned}
D &= (S^{-1} + M)^{-1} \\
M &= R^\dagger N^{-1} R
\end{aligned}
$$

  Fourier space:

$$
\begin{aligned}
M^{kq} &= \left(R^\dagger N^{-1} R \right)^{kq} \\
&= \underbrace{\left(R^\dagger \right)^{k'}_{k}}_{=(2\pi)^u \delta(k-k')\, \bar{b}(k)} \left(N^{-1} \right)_{k'q'} \underbrace{R^{q'}_{q}}_{=(2\pi)^u \delta(q-q')\, b(q)} \\
&= (2\pi)^u \delta(k-q) \underbrace{|b(k)|^2}_{P_R(k)} / P_n(k)
\end{aligned}
$$

  With this, we find

$$
\begin{aligned}
P_D(k) &= (P_S^{-1}(k) + P_M(k))^{-1} \\
&= \frac{P_s(k)}{1 + \frac{P_S(k) P_R(k)}{P_n(k)}}.
\end{aligned}
$$

- Calculation of the information source in Fourier space:

$$
j_k = (R^\dagger N^{-1} d)_k = \frac{\bar{b}(k)\, d_k}{P_n(k)}
$$

- Calculation of the Fourier components of the signal mean:

$$m^k = (D\,j)^k = \underbrace{\frac{(P_s/P_n)(k)\,\bar{b}(k)}{1+(P_sP_R/P_n)(k)}}_{f(k)}\,d^k$$

fidelity operator:

$$\begin{aligned}
Q &= SR^\dagger N^{-1}R \\
P_Q(k) &= \frac{P_sP_R}{P_n}(k) \\
f(k) &= \frac{P_Q(k)}{1+P_Q(k)}\,\frac{\bar{b}(k)}{P_R(k)} = \frac{P_Q(k)}{1+P_Q(k)}\,\frac{1}{b(k)} \\
&= \frac{1}{b(k)}\begin{cases} 1 & \text{if } P_Q(k) \gg 1 \text{ (high fidelity regime (hifi))} \\ \underbrace{P_Q(k)}_{\ll 1} & \text{if } P_Q(k) \ll 1 \text{ (low fidelity regime (lofi))} \end{cases}
\end{aligned}$$

The signal map m is not identical to the original signal. It is also shaped by the convolution, noise and deconvolution,

$$\begin{aligned}
m^k &= (fd)^k \\
&= \frac{P_Q(k)}{1+P_Q(k)}\,\frac{1}{b(k)}\,(b(k)\,s^k + n^k) \\
&= \frac{P_Q(k)}{1+P_Q(k)}\left(s^k + \frac{n^k}{b(k)}\right) \\
&= \begin{cases} s^k + \frac{n^k}{b(k)} & \text{if } P_Q(k) \gg 1 \\ P_Q(k)\left(s^k + \frac{n^k}{b(k)}\right) & \text{if } P_Q(k) \ll 1 \end{cases}
\end{aligned}$$

The power spectrum of the filtered signal map is,

$$\begin{aligned}
P_m(k) &= \frac{1}{V}\langle |m^k|^2\rangle_{(n,s)} \\
&= \frac{1}{V}\left(\frac{P_Q(k)}{1+P_Q(k)}\right)^2\left(\langle |s^k|^2\rangle + \frac{\langle |n^k|^2\rangle}{|b(k)|^2}\right) \\
&= \frac{P_Q(k)^2}{(1+P_Q(k))^2}\left(P_s + \frac{P_n}{P_R}\right)(k) \\
&= \frac{P_Q(k)^2}{(1+P_Q(k))^2}\,P_s(k)\left(\frac{P_Q(k)+1}{P_Q(k)}\right) \\
&= \frac{P_Q}{1+P_Q}(k)\,P_s(k) \\
&= \begin{cases} P_s(k) & \text{if } P_Q(k) \gg 1 \\ P_Q(k)P_s(k) & \text{if } P_Q(k) \ll 1 \end{cases}
\end{aligned}$$

9.3.7 *Missing data*

Again, we consider a deconvolution problem, but this time a part of the signal space is blocked in the area Ω. To model this, we introduce the transparency or transfer operator T

$$T^x_y = \delta(x-y)P(x \notin \Omega | x, \Omega)$$
$$P(x \notin \Omega | x, \Omega) = \begin{cases} 1 & \text{if } x \notin \Omega \\ 0 & \text{if } x \in \Omega \end{cases},$$

such that

- modified data:
$$d'^x = \underbrace{R^x_{x'} T^{x'}_y}_{=R'^y_x} s^y + n^x$$

- new information source:
$$j'_x = (R'^\dagger)^{x'}_x (N^{-1})_{x'y} d'^y$$

  $\Rightarrow$The new information source j' vanishes within Ω.

- new propagator:
$$\begin{aligned} D' &= (S^{-1} + R'^\dagger N^{-1} R')^{-1} \\ &= (S^{-1} + R^\dagger N^{-1} R - \Delta)^{-1} \end{aligned}$$

  with
$$\begin{aligned} \Delta &= R^\dagger N^{-1} R - R'^\dagger N^{-1} R' \\ &= R^\dagger N^{-1} R - T^\dagger R^\dagger N^{-1} R\, T \end{aligned}$$

  We define the complement to T the blocking operator,
$$\begin{aligned} B &= \mathbb{1} - T \\ B^x_y &= \delta(x-y)\, P(x \in \Omega | x, \Omega) \end{aligned}$$

$$\begin{aligned} \Rightarrow \Delta &= R^\dagger N^{-1} R - (\mathbb{1} - B)^\dagger R^\dagger N^{-1} R (\mathbb{1} - B) \\ &= -B^\dagger \underbrace{R^\dagger N^{-1} R}_{=M} B + B^\dagger M + M\, B \\ &= B\, M\, B \end{aligned}$$

  In the last equality we assumed for simplicity that M is local, $M \propto \delta(x-y)$.

This is the case when $R \propto \delta(x-y)$ and when the noise is white.

$$\begin{aligned} D' &= (S^{-1} + M - \Delta)^{-1} \\ &= (D^{-1} - \Delta)^{-1} \\ &= D(\mathbb{1} - \Delta D)^{-1} \end{aligned}$$

Now we can expand D' in powers of Δ using the geometrical series under the assumption that ΔD is a small expansion parameter (to be shown in Ch. 10),

$$\begin{aligned} D' &= D\,(\mathbb{1} - \Delta D)^{-1} \\ &= D\,(\mathbb{1} + \Delta D + \Delta D \Delta D + \ldots) \\ &= D + D\Delta D + D\Delta D \Delta D + \ldots, \end{aligned} \tag{334}$$

which in coordinates gives

$$D'^{xy} = D^{xy} + D^{xx'} \Delta_{x'y'} D^{y'y} + \mathcal{O}(\Delta^2).$$

In the special case of the local $M_{xy} = \delta(x-y)g(x)$ we assumed above, this reads

$$D'^{xy} = D^{xy} + D^{xz'}\, g(z')\, D^{y}_{z'}.$$

$\Rightarrow$ The information propagator/uncertainty dispersion at locations in and near Ω is increased with respect to the unblocked case.

- reconstructed signal map:
$$m'^{x} = D'^{xy}\, j'_y$$

$\Rightarrow$ The information propagation from the unblocked area $\overline{\Omega}$ into the blocked Ω is enhanced in order to compensate for the gap in the information source j' in Ω.

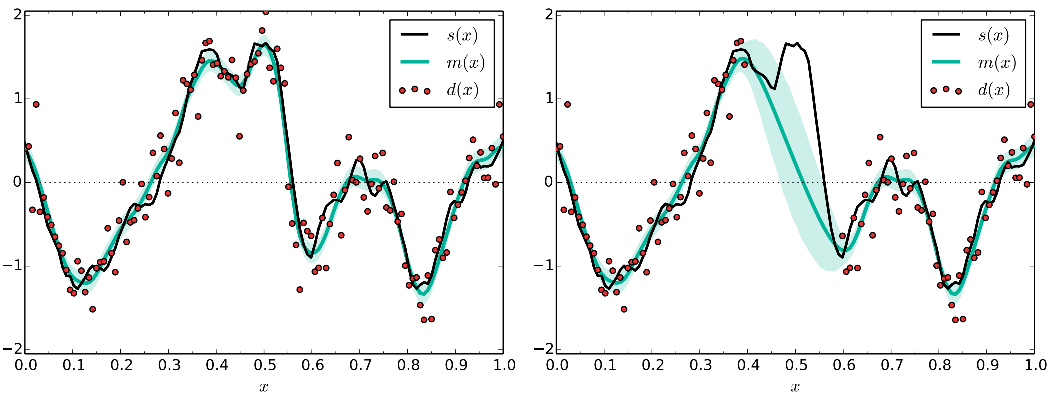


Figure 6: *Left:* A Gaussian signal s, noisy data d from signal measurements, and the Wiener filter reconstruction m including its one sigma uncertainty interval $\left[m^x - \sqrt{D^{xx}}, m^x + \sqrt{D^{xx}}\right]$. *Right:* The same but with a gap in the data for $x \in \Omega = [0.4, 0.6]$ leading to a larger reconstruction error as well as an increased uncertainty there.

10
MATRIX ALGEBRA

In the following we consider positive definite, symmetric/hermitian operators,

$$\begin{aligned} \text{hermitian: } A &= A^\dagger \\ \text{positive definite: } A &\geq 0 \\ &\to x^\dagger A x \geq 0 \, \forall x \neq 0 \\ \text{strictly positive definite: } A &> 0 \\ &\to x^\dagger A x > 0 \, \forall x \neq 0 \\ A > 0,\, B \geq 0 &\Rightarrow A + B > 0 \end{aligned}$$

- eigensystem:

  $$A = \sum_i \alpha_i a_i a_i^\dagger$$

  $\Rightarrow$ α_i are the eigenvalues of the system and a_i the corresponding orthonormal eigenvectors,

  $$\begin{aligned} A a_i &= \alpha_i a_i \\ a_i^\dagger a_j &= \delta_{ij}. \end{aligned}$$

- definition of action of a scalar function $f : \mathbb{C} \to \mathbb{C}$ on A:

  $$f(A) = \sum_i a_i a_i^\dagger f(\alpha_i)$$

Examples:

1. $f(x) = x^{1/2} \to f(A) = A^{1/2} = \sum_i \alpha_i^{1/2} a_i a_i^\dagger$
   proof:

   $$\begin{aligned} A^{1/2} A^{1/2} &= \sum_i \alpha_i a_i a_i^\dagger \sum_j \alpha_j a_j a_j^\dagger = \sum_i \sum_j \alpha_i \alpha_j a_i \delta_{ij} a_j^\dagger \\ &= \sum_i \alpha_i a_i a_i^\dagger = A, \end{aligned}$$

   where we used the definition of the scalar product of orthonormal vectors $a_i^\dagger a_j = \delta_{ij}$.

2. $f(x) = \sum_{n=0}^{\infty} \frac{1}{n!} x^n f_n \to f(A) = \sum_{n=0}^{\infty} \frac{1}{n!} f_n A^n$

$$\begin{aligned}
f(A) &= \sum_{n=0}^{\infty} \frac{1}{n!} f_n (\sum_i \alpha_i a_i a_i^\dagger)^n \\
&= \sum_{n=0}^{\infty} \frac{f_n}{n!} (\sum_{i_1} \ldots \sum_{i_n} \alpha_{i_1} \ldots \alpha_{i_n} a_{i_1} \underbrace{a_{i_1}^\dagger a_{i_2}}_{=\delta_{i_1,i_2}} \underbrace{a_{i_2}^\dagger a_{i_3}}_{=\delta_{i_2,i_3}} \ldots a_{i_n} a_{i_n}^\dagger) \\
&= \sum_{n=0}^{\infty} \frac{f_n}{n!} \sum_i \alpha_i^n a_i a_i^\dagger \\
&= \sum_i a_i a_i^\dagger \sum_{n=0}^{\infty} \frac{f_n}{n!} \alpha_i^n \\
&= \sum_i f(\alpha_i) a_i a_i^\dagger
\end{aligned}$$

3. $f(x) = x^{-1} \to f(A) = A^{-1} = \sum_i \alpha_i^{-1} a_i a_i^\dagger$

$$\begin{aligned}
f(A)A &= \sum_{ij} \underbrace{\alpha_i^{-1} \alpha_j}_{=1 \text{ if } i=j} \; a_i \underbrace{a_i^\dagger a_j}_{=\delta_{ij}} a_j^\dagger \\
&= \sum_i a_i a_i^\dagger \\
&= \mathbb{1} \\
\Rightarrow A^{-1} &= \frac{1}{A}
\end{aligned}$$

4. Missing proof for Eq. 334 using above relations:

$$\begin{aligned}
D' &= \left(D^{-1} - \Delta\right)^{-1} \\
&= \left[D^{-1/2} \left(\mathbb{1} - D^{1/2} \Delta\, D^{1/2}\right) D^{-1/2}\right]^{-1} \\
&= D^{1/2} \left(\mathbb{1} - \underbrace{D^{1/2} \Delta\, D^{1/2}}_{X}\right)^{-1} D^{1/2} \\
&= D^{1/2} \left(\mathbb{1} - X\right)^{-1} D^{1/2} \\
&= D^{1/2} \left(\mathbb{1} + X - XX + \ldots\right) D^{1/2} \\
&= D + D\,\Delta\, D + D\,\Delta\, D\,\Delta\, D + \ldots
\end{aligned}$$

as before, but we still have to show that $X < \mathbb{1}$ (all eigenvalues of X smaller than 1) so that geometric expansion is convergent:

$$\begin{aligned}
X &= D^{1/2} \Delta\, D^{1/2} = D^{1/2} B^\dagger M\, B\, D^{1/2} \\
&\leq \left(S^{-1} + M\right)^{-1/2} M \left(S^{-1} + M\right)^{-1/2} \\
&< \left(S^{-1} + M\right)^{-1/2} \left(S^{-1} + M\right) \left(S^{-1} + M\right)^{-1/2} \\
&= \mathbb{1}
\end{aligned}$$

□

11 GAUSSIAN PROCESSES

11.1 MARKOV PROCESSES

11.1.1 *Markov property*

A process $s : \mathbb{R} \mapsto \mathbb{R}^u$ (or $\mathbb{C}^u$) is Markov if any future value is independent of the past values if the present value is known,

$$f \geq t \geq p \Rightarrow \mathcal{P}(s^f|s^t, s^p) = \mathcal{P}(s^f|s^t).$$

For a Markov process, the present isolates the future from the past:

$$f \geq t \geq p \Rightarrow \mathcal{P}(s^f, s^t|s^p) = \mathcal{P}(s^f|s^t)\, P(s^t|s^p).$$

11.1.2 *Wiener process*

A Wiener process is the simplest non-deterministic stochastic process in continuous time:

$$\begin{aligned}
\dot{s}^t &= \frac{ds^t}{dt} = \sigma^t\, \xi^t, \text{ with} \\
\mathcal{P}(\xi) &= \mathcal{G}(\xi, \mathbb{1}) \text{ and } \sigma^t \text{ known.}
\end{aligned}$$

Let's assume we know s^p and want to know s^f at time $f > p$. In case ξ is known:

$$s^f = s^p + \underbrace{\int_p^f dt\, \sigma^t}_{=L_t^f}\, \xi^t$$

The linear operator L with $L_t^f = \sigma^t \mathcal{P}(p \leq t \leq f|p,\, t,\, f)$ translates $\xi \to s - s^p$ and can be inverted (in $t \in (p, \infty]$) with $\left(L^{-1}\right)_f^t = \delta(f-t)\, \frac{\partial}{\sigma^t \partial f}$.

$$\begin{aligned}
\Rightarrow \mathcal{P}(s|s^p) &= \int \mathcal{D}\xi\, \mathcal{P}(s|\xi, s^p)\, \mathcal{P}(\xi|s^p) = \int \mathcal{D}\xi\, \delta\left[s - (s^p + L\xi)\right]\, \mathcal{G}(\xi, \mathbb{1}) \\
&= \int \mathcal{D}\xi\, \frac{\delta\left[\xi - L^{-1}(s - s^p)\right]}{|L|}\, \mathcal{G}(\xi, \mathbb{1}) = \frac{\mathcal{G}(L^{-1}(s-s^p), \mathbb{1})}{|L|} \\
&= \frac{\exp\left[-\frac{1}{2}(s-s^p)^\dagger \left(L^{-1}\right)^\dagger \mathbb{1} L^{-1}(s-s^p)\right]}{|2\pi\mathbb{1}|^{1/2}\, |L|} \\
&= \frac{\exp\left[-\frac{1}{2}(s-s^p)^\dagger \left(L\, L^\dagger\right)^{-1} (s-s^p)\right]}{|2\pi L\, L^\dagger|^{1/2}} \\
&= \mathcal{G}(s - s^p,\, LL^\dagger),
\end{aligned}$$

using

$$\begin{aligned}
\left(L^{-1}\right)^{\dagger}\mathbb{1}L^{-1} &= \left(L^{\dagger}\right)^{-1}\mathbb{1}^{-1}L^{-1} = \left[L\mathbb{1}L^{\dagger}\right]^{-1} = \left[LL^{\dagger}\right]^{-1}, \\
\text{as } \left(L^{-1}\right)^{\dagger} &= \left(L^{\dagger}\right)^{-1}, \text{ since} \\
L^{\dagger}\left(L^{-1}\right)^{\dagger} &= \left(L^{-1}L\right)^{\dagger} = \mathbb{1}^{\dagger} = \mathbb{1}.
\end{aligned}$$

Now, we can calculate the remaining uncertainty dispersion of a Wiener process with data $d = s^p$,

$$\begin{aligned}
D^{tt'} &= \langle (s^t - s^p)\,(s^{t'} - s^p)\rangle_{(s|s^p)} = \left(LL^{\dagger}\right)^{tt'} \\
&= \int_{-\infty}^{\infty} dt''\,\sigma^{t''}\,P(p \leq t'' \leq t|\,p, t'', t)\sigma^{t''}\,P(p \leq t'' \leq t'|\,p, t'', t') \\
&= \int_{p}^{\min\{t,t'\}} dt''\,\left(\sigma^{t''}\right)^2. \qquad (335)
\end{aligned}$$

This implies for the **posterior uncertainty** variance of the Wiener process $D^{tt} = \langle (s^t - s^p)^2\rangle_{(s|s^p)} = \int_p^t dt'\,\left(\sigma^{t'}\right)^2$, which increases monotonically with time, and for its covariance $D^{tt'} = \min\{D^{tt}, D^{t't'}\}$. To summarize, we expect $\mathcal{P}(s|s^p, p) = \mathcal{G}(s - s^p, D)$. It should be possible to derive this result form the Wiener filter theory as well. This requires that we construct the signal prior $\mathcal{P}(s) = \mathcal{G}(s, S)$ first. This will happen in Sec. 11.2.

11.1.3 *Future expectation*

I = "$s : \mathbb{R} \mapsto \mathbb{R}$ is a Gaussian Markov process with zero mean, known **prior correlation** structure $S^{t_1 t_2} = \langle s^{t_1} s^{t_2}\rangle_{(s)}$, and known value s^t at present time t."
Question: What is the expectation of s^f for some future time $f \geq t$?
Answer: Regard s^t as data, s^f as signal, and use data space Wiener filter formula for expected signal,

$$\langle s^f\rangle_{(s^f|s^t)} = \langle s^f s^t\rangle_{(s)}\langle s^t s^t\rangle_{(s)}^{-1} s^t = \frac{S^{ft}}{S^{tt}}\,s^t.$$

The correlation structure of a zero mean Gaussian Markov process for times $f \geq t \geq p$ fulfills the relation

$$S^{fp} = \frac{S^{ft} S^{tp}}{S^{tt}}.$$

proof: using Wick's theorem

$$\langle s^f s^t s^t s^p\rangle_{(s)} = 2\,S^{ft} S^{tp} + S^{fp} S^{tt}$$

The average needs only to be performed over

$$P(s^f, s^t, s^p) = P(s^f, s^p|s^t)\,P(s^t) \overset{\text{Markov}}{=} P(s^f|s^t)\,P(s^p|s^t)\,P(s^t).$$

Figure 7: Evolution of a stock prize signal, which is known up to the present time t (red line), and afterwards being unknown. The uncertainty standard deviation $\sqrt{D^{ff}_{(t)}}$ grows with the future time f as $\propto \sqrt{f-t}$ in case of a constant volatility σ.

$$
\begin{aligned}
\Rightarrow \langle s^f s^t s^t s^p \rangle_{(s)} &= \int ds^t \int ds^f \int ds^p\, s^f\, \mathcal{P}(s^f|s^t)\, s^p\, \mathcal{P}(s^p|s^t)\, s^t s^t\, \mathcal{P}(s^t) \\
&= \langle \langle s^f \rangle_{(s^f|s^t)} \langle s^p \rangle_{(s^p|s^t)} s^t \rangle_{(s^t)} \\
&= \langle S^{ft} \left(S^{tt}\right)^{-1} s^t S^{pt} \left(S^{tt}\right)^{-1} s^t s^t s^t \rangle_{(s^t)} \\
&= \frac{S^{ft} S^{pt}}{S^{tt} S^{tt}} \langle s^t s^t s^t s^t \rangle_{(s^t)} = \frac{S^{ft} S^{pt}}{S^{tt} S^{tt}}\, 3\, S^{tt} S^{tt} = 3\, S^{ft} S^{pt} \\
&\stackrel{!}{=} 2\, S^{ft} S^{tp} + S^{fp} S^{tt}
\end{aligned}
$$

$$
\Rightarrow S^{ft} S^{tp} = S^{fp} S^{tt}
$$

11.1.4 *Example: evolution of a stock price*

- p^t: stock price at time t
- q^t: the evolution of the stock market index to which the stock belongs
- s^t: buy/sell signal indicates how much a stock over- or under-performs with respect to the market

$$
s^f = \ln \frac{p^f}{p^p} - \ln \frac{q^f}{q^p}.
$$

The trader will buy the stock if he expects s to raise, $\langle s^f \rangle_{(s^f|s^t)} > s^t$ for some $f > t > p$, with p some arbitrary reference point in the past.
The trader will sell the stock and buy other stocks if he expects s to fall, $\langle s^f \rangle_{(s^f|s^t)} < s^t$ for some $f > t$.

- no-arbitrage condition: If most traders trade this way and calculate their expectations $P(s^f|s^t)$ on a similar information basis, one can expect the no-arbitrage condition to hold: $\langle s^f \rangle_{(s^f|s^t)} = s^t$ for all $f > t$ and s to be Markov.
- If the price evolution is driven by many independent relatively small transactions, its relative changes should follow a Gaussian prior statistic $\mathcal{P}(s) = \mathcal{G}(s, S)$. The posterior after knowing s^t is $\mathcal{P}(s|s^t) = \mathcal{G}(s - s^t, D)$, with $D = D_{(t)}$ the posterior uncertainty structure.

$\Rightarrow s$ is a Gaussian Markov process with $\langle s^f \rangle_{(s^f|s^t)} = s^t$ for $f > t$, a so called martingale.

- $\langle s^f \rangle_{(s^f|s^t)} = S^{ft} \left(S^{tt}\right)^{-1} s^t \Rightarrow S^{ft} = S^{tt}$ for all $f > t$.
- $S^{ff} > S^{tt}$ is plausible $\Rightarrow \frac{d}{dt} S^{tt} \equiv \left(\sigma^t\right)^2 \geq 0$ or $S^{tt} = \int_p^t dt' \left(\sigma^{t'}\right)^2$, where σ^t is the so called volatility of the stock price

$$\Rightarrow S^{ab} = \min\{S^{aa}, S^{bb}\}$$

The trading signal is therefore a Wiener process

$$\dot{s}^t = \sigma^t \xi^t \text{ with } \mathcal{P}(\xi) = \mathcal{G}(\xi, \mathbb{1}).$$

The stock price is an exponentiated and rescaled version of this (lognormal process),

$$p^t = p^p \, \frac{q^t}{q^p} \, e^{s^t}.$$

$$\begin{aligned} \Rightarrow \langle s^f - s^t \rangle_{(s^f|s^t)} &= 0 \\ \langle e^{s^f} \rangle_{(s^f|s^t)} &= e^{s^t} \langle e^{s^f - s^t} \rangle_{(s^f|s^t)} \\ \mathcal{P}(s^f|s^t) &= \mathcal{G}(\underbrace{s^f - s^t}_{=\Delta}, \Sigma), \end{aligned}$$

with $\Sigma = D^{ff}_{(t)}$. The expectation for the stock price exceeds that of the pure market evolution, $\langle p^f / p^p \rangle_{(s^f|s^t)} \geq \left(q^f / q^p\right) \, e^{s^t}$. This is because

$$\begin{aligned} \langle e^{s^f} \rangle_{(s^f|s^t)} &= e^{s^t} \langle e^{s^f - s^t} \rangle_{(s^f|s^t)} \\ &= e^{s^t} \langle e^{\Delta} \rangle_{\mathcal{G}(\Delta,\Sigma)} \\ &= e^{s^t} \sum_{n=0}^{\infty} \frac{1}{n!} \langle \Delta^n \rangle_{\mathcal{G}(\Delta,\Sigma)} \\ &= e^{s^t} \sum_{n=0}^{\infty} \frac{1}{(2n)!} \langle \Delta^{2n} \rangle_{\mathcal{G}(\Delta,\Sigma)} \\ &= e^{s^t} \sum_{n=0}^{\infty} \frac{1}{(2n)!} \frac{(2n)!}{2^n n!} \, \Sigma^n \\ &= e^{s^t} \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{\Sigma}{2}\right)^n = e^{s^t} \, e^{\frac{1}{2}\Sigma} = e^{s^t + \frac{1}{2}\Sigma} \geq e^{s^t}. \end{aligned}$$

$\Rightarrow$ The expectation for the stock price exceeds that of the pure market evolution. The uncertainty variance $\Sigma = \int_t^f dt' \left(\sigma^{t'}\right)^2$ of the future log-price leads to a positive drift of the price itself. Thus, a refined model would be $\dot{s}_t = \sigma^t \xi^t + \mu^t$, where μ^t is a (slowly time dependent) drift rate.

11.2 STOCHASTIC CALCULUS

11.2.1 *Stratonovich's calculus*

Consider the generalized Wiener process,

$$\frac{ds^t}{dt} = \xi^t,$$

with colored Gaussian excitation ξ with Fourier space correlation $\Xi^{\omega\omega'} = \left\langle \xi^\omega \overline{\xi^{\omega'}} \right\rangle_{(\xi)} = 2\pi\delta(\omega - \omega')\, P_\xi(\omega)$ described by a bound power spectrum, $\int_{-\infty}^{\infty} d\omega\, P_\xi(\omega) < \infty$.

$$\begin{aligned}
\xi^\omega &= \int_{-\infty}^{\infty} dt\, e^{i\omega t} \xi^t = \int_{-\infty}^{\infty} dt\, e^{i\omega t} \frac{ds^t}{dt} = -\int_{-\infty}^{\infty} dt\, \frac{de^{i\omega t}}{dt}\, s^t = -i\omega \int_{-\infty}^{\infty} dt\, e^{i\omega t}\, s^t \\
&= -i\omega\, s^\omega \Rightarrow s^\omega = \frac{\xi^\omega}{-i\omega} \\
\Rightarrow S^{\omega\omega'} &= \left\langle s^\omega \overline{s^{\omega'}} \right\rangle_{(s)} = \left\langle \frac{\xi^\omega}{-i\omega}\, \frac{\overline{\xi^{\omega'}}}{i\omega} \right\rangle_{(\xi)} = \frac{\Xi^{\omega\omega'}}{\omega^2} = 2\pi\delta(\omega - \omega')\, \underbrace{\frac{P_\xi(\omega)}{\omega^2}}_{=P_s(\omega)}
\end{aligned}$$

In case of the Wiener process the noise spectrum is white with $P_\xi(\omega) \to 1$, $\Xi \to \mathbb{1}$ and $P_s(\omega) \to \omega^{-2}$.

A (non-linearly) transformed random process $f^t \equiv f(s^t)$ with $f : \mathbb{R} \mapsto \mathbb{R}$ some differentiable transformation function. The transformed process is then

$$\frac{df^t}{dt} = \frac{df(s^t)}{ds^t}\, \frac{ds^t}{dt} = f'(s^t)\, \xi^t, \tag{336}$$

with $f'(s^t) = df(s^t)/ds^t$ according to the chain rule of differential calculus. The transformed random process simply obeys $f^t = f(s^p + \int_p^t dt'\, \xi^{t'})$.

The drift of $\langle f^t \rangle_{(\xi)}$ for some small time interval Δt can be Taylor expanded in $\Delta s = s^{t+\Delta t} - s^t$,

$$\begin{aligned}
\langle \Delta f^t \rangle_{(\xi|s^t)} &= \langle f^{t+\Delta t} - f^t \rangle_{(\xi|s^t)} \\
&= \langle f(s^t + \int_t^{t+\Delta t} dt'\, \xi^{t'}) - f(s^t) \rangle_{(\xi|s^t)} \\
&= f'(s^t)\, \langle \Delta s \rangle_{(\xi|s^t)} + \frac{1}{2} f''(s^t)\, \langle (\Delta s)^2 \rangle_{(\xi|s^t)} + \frac{1}{3!} f'''(s^t)\, \langle (\Delta s)^3 \rangle_{(\xi|s^t)} \\
&\quad + \frac{1}{4!} f''''(s^t)\, \langle (\Delta s)^4 \rangle_{(\xi|s^t)} + \mathcal{O}((\Delta s)^5).
\end{aligned}$$

The required moments are

$$
\begin{aligned}
\langle \Delta s \rangle_{(\xi|s^t)} &= \int_t^{t+\Delta t} dt' \, \langle \xi^{t'} \rangle_{(\xi|s^t)} = 0, \\
\langle (\Delta s)^2 \rangle_{(\xi|s^t)} &= \int_t^{t+\Delta t} dt' \int_t^{t+\Delta t} dt'' \, \langle \xi^{t'} \xi^{t''} \rangle_{(\xi|s^t)} \\
&= \int_t^{t+\Delta t} dt' \underbrace{\int_t^{t+\Delta t} dt'' \delta(t'-t'')}_{=1} = \Delta t, \\
\langle (\Delta s)^3 \rangle_{(\xi|s^t)} &= \int_t^{t+\Delta t} dt' \int_t^{t+\Delta t} dt'' \int_t^{t+\Delta t} dt''' \, \langle \xi^{t'} \xi^{t''} \xi^{t'''} \rangle_{(\xi|s^t)} = 0, \text{ and} \\
\langle (\Delta s)^4 \rangle_{(\xi|s^t)} &= \int_t^{t+\Delta t} dt' \int_t^{t+\Delta t} dt'' \int_t^{t+\Delta t} dt''' \int_t^{t+\Delta t} dt'''' \, \underbrace{\langle \xi^{t'} \xi^{t''} \xi^{t'''} \xi^{t''''} \rangle_{(\xi|s^t)}}_{\Xi^{t't''}\Xi^{t'''t''''} + \Xi^{t't'''}\Xi^{t''t''''} + \Xi^{t't''''}\Xi^{t''t'''}} \\
&= 3\, (\Delta t)^2 ,
\end{aligned}
$$

so that

$$
\langle \Delta f^t \rangle_{(\xi|s^t)} = \frac{1}{2} f''(s^t)\, \Delta t + \frac{3}{4!} f''''(s^t) \, (\Delta t)^2 + \mathcal{O}\left((\Delta t)^3 \right).
$$

The drift rate of a non-linearly transformed Wiener process in Stratonovich's calculus is therefore

$$
\left\langle \frac{df^t}{dt} \right\rangle_{(\xi|s^t)} = \left\langle \lim_{\Delta t \to 0} \frac{\Delta f}{\Delta t} \right\rangle_{(\xi|s^t)} = \frac{1}{2} f''(s^t),
$$

where s^t is the Wiener process and $f^t = f(s^t)$ the transformation.

11.2.2 *Itô's calculus*

The transformed Wiener process in Itô's calculus is denoted by

$$
df = f'(s^t)\, ds^t + \frac{1}{2} f''(s^t)\, dt \text{ or} \tag{337}
$$

$$
\frac{df^t}{dt} = \frac{df(s^t)}{ds^t} \frac{ds^t}{dt} + \frac{1}{2} f''(s^t) = f'(s^t)\, \xi_t, + \frac{1}{2} f''(s^t) \tag{338}
$$

where s^t is the Wiener process and $f^t = f(s^t)$ the transformation. This leads as well to the drift rate

$$
\langle \frac{df^t}{dt} \rangle_{(\xi|s^t)} = \frac{1}{2} f''(s^t).
$$

Why does Itô's calculus require that the drift rate has to be added explicitly to the stochastic equation, where in Stratonovich calculus it is a simple consequence of the chain rule? The reason is that the microscopic picture of the underlying stochastic processes differ.

- Stratonovich picture: $f(s)$ acts continuously during evolution within $\Delta t \to$ drift arises automatically; excitation can have coloured spectrum.
- Itô picture: microscopic concept of time is discrete $\to$ no drift without explicit drift term; excitation should have white spectrum.

11.3 LINEAR STOCHASTIC DIFFERENTIAL EQUATIONS

Assume a generic time-independent linear stochastic differential equation of order N,

$$\sum_{n=0}^{N} a_n \frac{d^n s^t}{dt^n} = \xi^t,\ \mathcal{P}(\xi) = \mathcal{G}(\xi, \Xi),\ \Xi^{\omega\omega'} = 2\pi\delta(\omega - \omega')P_\xi(\omega).$$

In case of the white noise driven Wiener process, $s^t, \xi^t \in \mathbb{R}$, $a_n = \delta_{n1}$, and $P_\xi(\omega) = 1$.

The differential equation becomes an algebraic equation after Fourier transformation,

$$\begin{aligned}
\int_{-\infty}^{\infty} dt\, e^{i\omega t} \sum_{n=0}^{N} a_n \frac{d^n s^t}{dt^n} &= \\
\sum_{n=0}^{N} a_n \int_{-\infty}^{\infty} dt\, e^{i\omega t} \frac{d^n}{dt^n} \int_{-\infty}^{\infty} \frac{d\omega'}{2\pi} e^{-i\omega' t}\, s^{\omega'} &= \\
\sum_{n=0}^{N} a_n \int_{-\infty}^{\infty} \frac{d\omega'}{2\pi} (-i\omega')^n\, s^{\omega'} \underbrace{\int_{-\infty}^{\infty} dt\, e^{i(\omega-\omega')t}}_{2\pi\delta(\omega-\omega')} &= \\
\sum_{n=0}^{N} a_n\, (-i\omega)^n\, s^{\omega} &= \xi^{\omega},
\end{aligned}$$

$$\boxed{\begin{aligned}
\Rightarrow s^{\omega} &= \left[\sum_{n=0}^{N} a_n\, (-i\omega)^n\right]^{-1} \xi^{\omega} \\
s &= R\xi \\
R^{\omega}_{\omega'} &= 2\pi\delta(\omega - \omega') \left[\sum_{n=0}^{N} a_n\, (-i\omega)^n\right]^{-1}
\end{aligned}}$$

From this, it is obvious that

$$\begin{aligned}
\mathcal{P}(s|a, \Xi) &= \mathcal{G}(s, S),\ \text{with } S = R\Xi R^\dagger, \\
S^{\omega\omega'} &= 2\pi\delta(\omega - \omega')P_s(\omega),\ \text{and} \\
P_s(\omega) &= \frac{P_\xi(\omega)}{\left|\sum_{n=0}^{N} a_n\, (-i\omega)^n\right|^2} \equiv P_R(\omega)P_\xi(\omega).
\end{aligned}$$

11.3.1 *Example: Wiener process*

$$\begin{aligned}
\dot{s}(t) &= \xi^t \\
a_1 &= 1 \\
\Rightarrow P_R(\omega) &= \frac{1}{|a_1(-iw)^1|^2} = \frac{1}{\omega^2}
\end{aligned}$$

EXERCISE: Use this to construct a prior on s as well its posterior if the data $d = (s^p, p)$ is given! Is this consistent with the result given in Sect. 11.1.2?

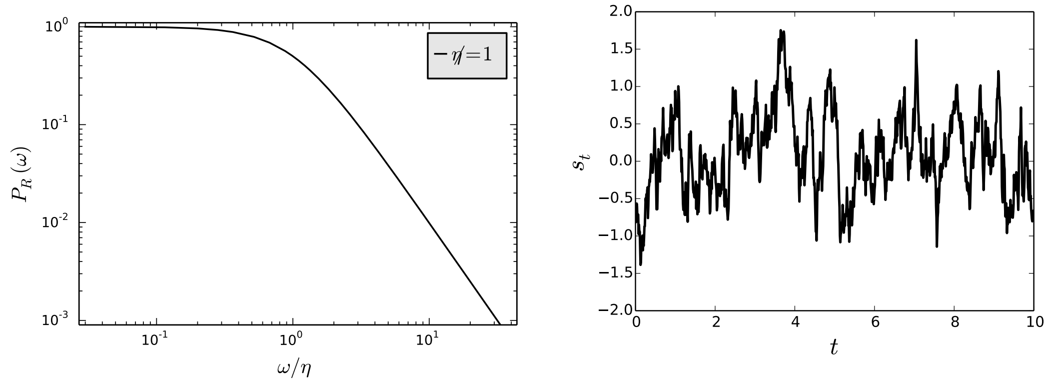


Figure 8: Left: Spectra of white noise driven Ornstein-Uhlenbeck process. Right: A signal realization for $\eta\!\!\!/ = \eta/(2\pi) = 1$.

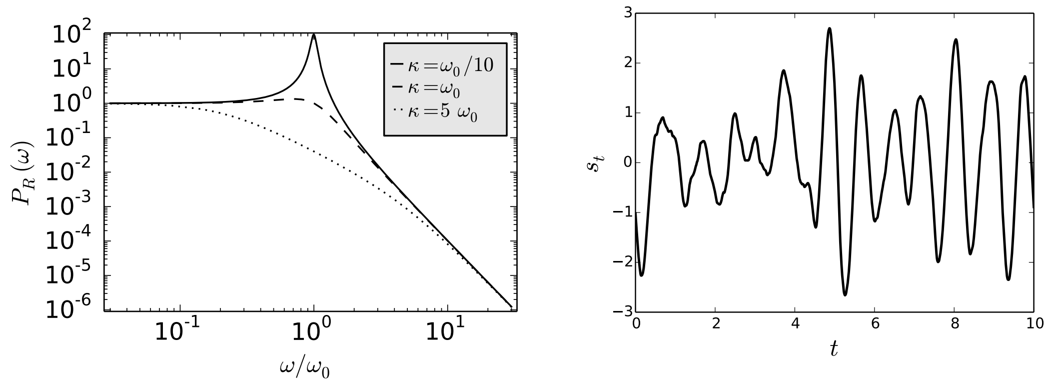


Figure 9: Left: Spectra of white noise driven harmonic oscillators for various values of the damping constant κ in term of the oscillator's eigenfrequency ω_0. A weakly damped, damped, and a strongly damped case are shown. Right: A signal realization for the weakly damped oscillator with $\nu_0 = \omega_0/(2\pi) = 1.2$.

11.3.2 *Example: Ornstein-Uhlenbeck process*

$$
\begin{aligned}
\dot{s}^t + \eta\, s^t &= \xi^t \\
a_0 &= \eta \\
a_1 &= 1 \\
\Rightarrow P_R(\omega) &= |\eta - i\omega|^{-2} = (\eta^2 + \omega^2)^{-1}
\end{aligned}
$$

For white noise, $P_\xi(\omega) = 1$,

$$P_s(\omega) = P_R(\omega) = (\eta^2 + \omega^2)^{-1}.$$

11.3.3 *Example: harmonic oscillator*

$$\ddot{s}^t + \kappa \dot{s}^t + \omega_0^2\, s^t = f\, \xi^t$$

- κ: a damping constant

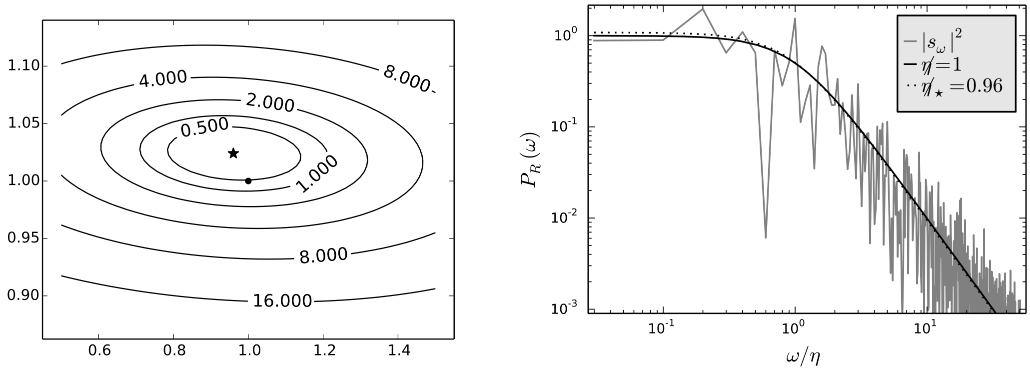


Figure 10: Parameter determination for the Ornstein-Uhlenbeck process. Left: Information Hamiltonian contours $\Delta\mathcal{H}(s,a) = \mathcal{H}(s,a) - \mathcal{H}(s,a_\star)$ as a function of $(\eta\!\!/, a_1)$ where $a_0 = \eta = 2\pi\eta\!\!/$ for the signal realization shown in Fig. 8. The minimum at $(\eta\!\!/, a_1)_\star \approx (0.96, 1.024)$ is marked by a star and the correct values at $(\eta\!\!/, a_1) = (1,1)$ by a dot. The correct value lies on the so called 1-σ contour at $\Delta\mathcal{H} = 1/2$. Rights: Power spectrum of the signal realization shown in Fig. 8, that of the original process, and that for the reconstructed parameters $a_\star$.

- ω_0: eigenfrequency of the oscillator
- f : noise coupling constant.

$$
\begin{aligned}
a_0 &= \omega_0^2 f^{-1}\\
a_1 &= \kappa f^{-1}\\
a_2 &= f^{-1}
\end{aligned}
$$

$$\Rightarrow P_R(\omega) = f^2 \left[\omega_0^4 + \left(\kappa^2 - 2\,\omega_0^2\right)\,\omega^2 + \omega^4\right]^{-1}.$$

11.4 PARAMETER DETERMINATION

In many cases, the class of the stochastic process is known, but the parameters $a = (a_0, \dots a_N)$ are unknown. Fortunately, these can be determined from signal observations.

$$\mathcal{P}(a|s) = \frac{\mathcal{P}(s|a)\mathcal{P}(a)}{\mathcal{P}(s)} = \frac{e^{-\mathcal{H}(s,a)}}{Z(s)}$$

$$
\begin{aligned}
\mathcal{H}(s,a) &= -\ln\mathcal{P}(s|a) - \ln\mathcal{P}(a)\\
&= \frac{1}{2}\left[s^\dagger S^{-1}\,s + \ln|2\pi S|\right] + \mathcal{H}(a)\\
&\widehat{=} \frac{1}{2}\int\frac{d\omega}{2\,\pi}\left[\frac{|s^\omega|^2}{P_s(\omega)} + \ln P_s(\omega)\right],
\end{aligned}
$$

where in the last step we assumed a flat prior for a.

11.5 LOGNORMAL POISSON MODEL

Events (e.g. photons, galaxies, customers) are recorded over some space (e.g. sky, universe, time). How is the according event generating process spatially structured?

- $\rho^x = \rho(x)$: event density at location x
- $d = (d^1, ..., d^n)$: number of observed events in the detector bin $i = 1, ..., n$
- $\lambda = (\lambda^1, ..., \lambda^n)$: expected number of observed events in the detector bin $i = 1, ..., n$, if $\rho(x)$ is known

$$\lambda^i = \int dx\, R^i_x \rho(x) = R^i_x \varrho^x$$

$\Rightarrow$ If the events are independent of each other:

$$\mathcal{P}(d|\lambda) = \prod_{i=1}^{n} \frac{\left(\lambda^i\right)^{d^i} e^{-\lambda^i}}{d^i!}$$
$$\mathcal{H}(d|\lambda) = \sum_{i=1}^{n} [\lambda^i - d^i \ln \lambda^i + \ln(d^i!)]$$

- for simplicity assume a local response R in the following with the exposure $\kappa(x)$ at location x :

$$\begin{aligned}\lambda^x &= \int dy\, \delta(x-y)\kappa(y)\rho(y) \\ &= (\kappa\rho)^x\end{aligned}$$

$$\begin{aligned}\Rightarrow \mathcal{H}(d|\rho) \;&\widehat{=}\; \kappa_x \rho^x - d_x \ln(\kappa\rho)^x \\ &= \kappa^\dagger \rho - d^\dagger \ln(\kappa\rho)\end{aligned}$$

In the last equation the term $d^\dagger \ln(\kappa\rho)$ is to be understood as a component-wise multiplication and function application. For the derivation of $\mathcal{H}(d|\rho)$ we neglect the $\rho-$independent term $\ln(d^i!)$ in $\mathcal{H}(d|\lambda)$.
Defining the **prior**:

- $\rho^x > 0\ \forall x$
- ρ^x can vary on logarithmic scale.
  $\Rightarrow$ Choose a more appropriate signal s^x,

$$\begin{aligned}s^x &= \ln \frac{\rho^x}{\rho_0}, \\ \rho^x &= \rho_0 e^{s^x}.\end{aligned}$$

  ρ_0 should be chosen such that $\langle s \rangle_{(s)} = 0$.

- Spatial correlations exist, described by known $S = \langle ss^\dagger \rangle_{(s)}$. Higher order corrections are ignored.

Applying the Maximum Entropy principle with known 1st and 2nd moments, we obtain the probability distribution,

$$\begin{aligned} \mathcal{P}(s) &= \mathcal{G}(s, S) \\ \mathcal{H}(s) &= \frac{1}{2}s^\dagger S^{-1} s + \frac{1}{2}\ln|2\pi S| \end{aligned}$$

From this we can calculate the **joint information Hamiltonian**,

$$\begin{aligned} \mathcal{H}(d,s) &= \mathcal{H}(d|s) + \mathcal{H}(s) \\ &\widehat{=} \frac{1}{2}s^\dagger S^{-1} s + \underbrace{\kappa^\dagger \rho_0}_{=\kappa' \to \kappa} e^s - d^\dagger \ln(\kappa^\dagger \rho_0 e^s) \\ &\widehat{=} \frac{1}{2}s^\dagger S^{-1} s + \kappa^\dagger e^s - d^\dagger s. \end{aligned}$$

In the next step we want to identify the **free and interaction Hamiltonian** in $\mathcal{H}(d,s)$. For this purpose we expand the exponential function, $e^{s^x} = 1 + s^x + \frac{1}{2}(s^x)^2 + ...$, define $\widehat{\kappa} = \mathrm{diag}(\kappa)$ and substitute,

$$\begin{aligned} \kappa^\dagger e^s &= \int dx\, \kappa(x)\,(1 + s(x) + \frac{1}{2}(s(x))^2 + ...) \\ \Rightarrow \mathcal{H}(d,s) &\widehat{=} \underbrace{\frac{1}{2}s^\dagger \underbrace{(S^{-1} + \widehat{\kappa})}_{=D^{-1}} s - \underbrace{(d-k)^\dagger}_{=j^\dagger} s}_{\text{free Hamiltonian}} + \underbrace{\kappa^\dagger \underbrace{(e^s - 1 - s - \frac{s^2}{2})}_{=\sum_{n=3}^{\infty}\frac{1}{n!}s^n}}_{\text{interaction Hamiltonian}}. \end{aligned}$$

$$\boxed{\begin{aligned} \Rightarrow \mathcal{H}(d,s) &= \frac{1}{2}s^\dagger D^{-1} s - j^\dagger s + \sum_{n=3}^{\infty} \frac{\kappa^\dagger s^n}{n!} \\ D &= (S^{-1} + \widehat{\kappa})^{-1} \\ j &= d - \kappa \end{aligned}}$$

CLASSICAL OR MAP SOLUTION

$$\begin{aligned} \frac{\partial \mathcal{H}(d,s)}{\partial s^x} &\stackrel{!}{=} 0 \\ &= \frac{\partial}{\partial s^x}\left[\left(\frac{1}{2}s^{x'} S^{-1}_{x'x''} s^{x''} + \kappa_{x'} e^{s^{x'}} - d_{x'} s^{x'}\right)\right] \\ &= \frac{1}{2}S^{-1}_{xx''} s^{x''} + \frac{1}{2}s^{x'} S^{-1}_{x'x} + (\kappa e^s)_x - d_x \\ &= \left[\frac{1}{2}S^{-1} s + \kappa e^s - d + \frac{1}{2}(s^\dagger S^{-1})^\dagger\right]^x \\ &= \left[S^{-1} s + \kappa e^s - d\right]_x \\ \frac{\partial \mathcal{H}(d,s)}{\partial s} &= S^{-1} s - d + \kappa e^s \stackrel{!}{=} 0 \\ \Rightarrow m &= S(d - \kappa e^m) \end{aligned}$$

Solving this by numerical iteration is likely unstable, as there are linear terms in m on the right hand side. An equation, which is numerically more stable under iteration, has these linear terms on the left hand side. The information propagator D and a s–dependent information source j appear thereby:

$$\begin{aligned} S^{-1}m &= d - \kappa e^{m} \\ \underbrace{(S^{-1}+\widehat{\kappa})}_{=D^{-1}} m &= \underbrace{d - \kappa(e^{m} - m)}_{=j} \\ m &= D(d - \kappa(e^{m} - m)). \end{aligned}$$

Comparison with the Wiener filter $m = (S^{-1} + R^{\dagger}N^{-1}R)^{-1}R^{\dagger}N^{-1}d'$:

- $\widehat{\kappa} \sim R^{\dagger}N^{-1}R$
- $\kappa \sim R$
- $\mathbb{1} \sim R^{\dagger}N^{-1}$

$\Rightarrow$ effective noise covariance $N \sim \widehat{\kappa}$ and response $R \sim \kappa$ are both determined by κ.

EXPANSION AROUND THE CLASSICAL SOLUTION

$$s = m + \varphi$$

Calculate the Hamiltonian:

$$\begin{aligned} \mathcal{H}(d,\, \varphi|m) &= \mathcal{H}(d,\, s = m+\varphi) \\ &\widehat{=} \frac{1}{2}(m+\varphi)^{\dagger}S^{-1}(m+\varphi) + \kappa^{\dagger}e^{m+\varphi} - d^{\dagger}(m+\varphi) \\ &\widehat{=} \frac{1}{2}\varphi^{\dagger}S^{-1}\varphi + m^{\dagger}S^{-1}\varphi + \underbrace{\kappa_m^{\dagger}}_{(\kappa_m)_x = \kappa^x e^{m_x}} e^{\varphi} - d^{\dagger}\varphi \\ &= \frac{1}{2}\varphi^{\dagger}S^{-1}\varphi - \underbrace{(d - S^{-1}m)^{\dagger}}_{=d_m^{\dagger}}\varphi + \kappa_m^{\dagger}e^{\varphi} \end{aligned}$$

The shifted problem looks like the original problem of $\mathcal{H}(d,\, s)$ with changed coefficients. We can calculate d_m,

$$\begin{aligned} d_m &= d - S^{-1}m \\ &= d - S^{-1}S(d - \kappa e^{m}) \\ &= d - d - \kappa e^{m} \\ &= \kappa_m, \end{aligned}$$

where the short notation $\kappa_m = \kappa\, e^{m}$ was introduced, indicating that the classical density $\rho_m = \rho_0 e^{m}$ could as well be absorbed into an effective exposure κ_m. Accordingly, $\mathcal{H}(d,\, \rho|m)$ can be written as,

$$
\begin{aligned}
\mathcal{H}(d,\, \varphi|m) &= \frac{1}{2}\varphi^{\dagger}(S^{-1}+\widehat{\kappa}_m)\varphi + \underbrace{(d_m-\kappa_m)^{\dagger}}_{=0}\varphi + \kappa_m^{\dagger}\left(e^{\varphi}-\varphi-\frac{\varphi^2}{2}\right) \\
&= \frac{1}{2}\varphi^{\dagger}\underbrace{(S^{-1}+\widehat{\kappa}_m)}_{=D_m^{-1}}\varphi + \kappa_m^{\dagger}\left(e^{\varphi}-\varphi-\frac{\varphi^2}{2}\right)
\end{aligned}
$$

- Noise and exposure are structured by $\kappa_m = \kappa e^m$.
- φ-field is not sourced, we are expanding around a minimum.

12 INFORMATION FIELD THEORY

Following [6, 4?]

12.1 BASIC FORMALISM

BAYES THEOREM

$$\begin{aligned}\mathcal{P}(s|d) &= \frac{\mathcal{P}(d,s)}{\mathcal{P}(d)} = \frac{e^{-\mathcal{H}(d,s)}}{\mathcal{Z}(d)} \\ \mathcal{Z}(d) &= \int \mathcal{D}s\,\mathcal{P}(d,s) = \int \mathcal{D}s\, e^{-\mathcal{H}(d,s)}\end{aligned}$$

MOMENT GENERATING FUNCTION

$$\mathcal{Z}(d,J) = \int \mathcal{D}s\, e^{-\mathcal{H}(d,s)+J^\dagger s}$$

$\Rightarrow$ Calculate moments via the generating function,

$$\langle s^{x_1}...s^{x_n}\rangle_{(s|d)} = \frac{1}{\mathcal{Z}}\frac{\delta^n \mathcal{Z}(d,J)}{\delta J_{x_1}...\delta J_{x_n}}\bigg|_{J=0}.$$

$\Rightarrow$ Calculate cumulants via cumulant-generating function,

$$\langle s^{x_1}...s^{x_n}\rangle^{\mathrm{c}}_{(s|d)} = \frac{\delta^n \ln \mathcal{Z}(d,J)}{\delta J_{x_1}...\delta J_{x_n}}\bigg|_{J=0}.$$

Examples:

$$\begin{aligned}\langle s^{x_1}\rangle^{\mathrm{c}}_{(s|d)} &= \frac{1}{\mathcal{Z}}\frac{\delta}{\delta J_{x_1}}\mathcal{Z}\bigg|_{J=0} = \langle s^{x_1}\rangle_{(s|d)} = \bar{s}_{x_1} \\ \langle s^{x_1}s^{x_2}\rangle^{\mathrm{c}}_{(s|d)} &= \frac{\delta}{\delta J_{x_2}}\left[\frac{1}{\mathcal{Z}}\frac{\delta}{\delta J_{x_1}}\mathcal{Z}\right]\bigg|_{J=0} = \frac{1}{\mathcal{Z}}\frac{\delta^2\mathcal{Z}}{\delta J_{x_1}\delta J_{x_2}} - \frac{1}{\mathcal{Z}^2}\frac{\delta\mathcal{Z}}{\delta J_{x_1}}\frac{\delta\mathcal{Z}}{\delta J_{x_2}}\bigg|_{J=0} \\ &= \langle s^{x_1}s^{x_2}\rangle_{(s|d)} - \langle s^{x_1}\rangle_{(s|d)}\langle s^{x_2}\rangle_{(s|d)} = \langle (s-\bar{s})^{x_1}(s-\bar{s})^{x_2}\rangle_{(s|d)}\end{aligned}$$

If s is Gaussian, $\mathcal{P}(s|d) = \mathcal{G}(s-m, D)$, $\mathcal{H}(s|d) \widehat{=} \frac{1}{2}(s-m)^\dagger D(s-m)$

$$\begin{aligned}\langle s\rangle_{(s|d)} &= m \\ \langle ss^\dagger\rangle^{\mathrm{c}}_{(s|d)} &= D \\ \langle ss^\dagger\rangle_{(s|d)} &= D + m\,m^\dagger \\ \langle s^{x_1}...s^{x_n}\rangle^{\mathrm{c}}_{(s|d)} &= 0 \text{ for } n \geq 3\end{aligned}$$

12.2 FREE THEORY

In the following we consider a linear response, $d = Rs + n$ (e.g. $d_i = R_i^x s_x + n_i$), with independent Gaussian signal and noise, $\mathcal{P}(s, n) = \mathcal{G}(s, S)\,\mathcal{G}(n, N)$, where $S = \langle ss^\dagger \rangle_{(s,n)}$ and $N = \langle nn^\dagger \rangle_{(s,n)}$.

$$\begin{aligned}
\mathcal{P}(d, s) &= \mathcal{G}(s, S)\,\mathcal{G}(n = d - Rs, N) \\
\mathcal{H}(d, s) &= \frac{1}{2}(d - Rs)^\dagger N^{-1}(d - Rs) + \frac{1}{2}s^\dagger S^{-1} s + \frac{1}{2}\ln(|2\pi S||2\pi N|) \\
&= \frac{1}{2}s^\dagger \underbrace{(S^{-1} + R^\dagger N^{-1} R)}_{=D^{-1}} s + s^\dagger \underbrace{R^\dagger N^{-1} d}_{=j} + \mathcal{H}_0
\end{aligned}$$

The generating function can be calculated by means of $\mathcal{H}(d, s)$,

$$\begin{aligned}
\mathcal{Z}(J) &= \int \mathcal{D}s\, e^{-\mathcal{H}(d,s) + J^\dagger s} \\
&= \int \mathcal{D}s \exp\left(-\frac{1}{2}s^\dagger D^{-1} s + \underbrace{(J + j)^\dagger}_{=j'^\dagger} s - \mathcal{H}_0 \right) \\
&= \int \mathcal{D}s \exp\left[-\frac{1}{2}\left(s^\dagger D^{-1} s - 2 j'^\dagger D\, D^{-1} s + j'^\dagger D\, D^{-1} \underbrace{D\, j'}_{=m'} \right) + \frac{1}{2} j'^\dagger D\, j' - \mathcal{H}_0 \right] \\
&= \int \mathcal{D}s \exp\left[-\frac{1}{2}\left((s - m')^\dagger D^{-1} (s - m') \right) + \frac{1}{2} j'^\dagger D\, j' - \mathcal{H}_0 \right] \\
&= |2\pi D|^{1/2} \exp\left(+\frac{1}{2}(J + j)^\dagger D\,(J + j) - \mathcal{H}_0 \right) \\
\ln \mathcal{Z}(J) &= \frac{1}{2}(J + j)^\dagger D\,(J + j) + \frac{1}{2}\ln|2\pi D| - \mathcal{H}_0
\end{aligned}$$

Actually, the variable J is not required if we instead take derivatives with respect to j.
moments:

$$\begin{aligned}
\langle s \rangle^c_{(s|d)} &= m = \frac{\delta \ln \mathcal{Z}(j)}{\delta j} = Dj \\
\langle ss^\dagger \rangle^c_{(s|d)} &= \langle (s - \bar{s})(s - \bar{s})^\dagger \rangle = \frac{\delta^2 \ln \mathcal{Z}(j)}{\delta j \delta j^\dagger} = D \\
\langle s^{x_1} ... s^{x_n} \rangle^c_{(s|d)} &= \frac{\delta^n \ln \mathcal{Z}(j)}{\delta j_{x_1} ... \delta j_{x_n}} = \frac{\delta^{n-2}}{\delta j_{x_3} ... \delta j_{x_n}} D^{x_1 x_2} = 0
\end{aligned}$$

12.3 INTERACTING FIELD THEORY

$$\mathcal{H}(d, s) = \underbrace{\frac{1}{2}s^\dagger D^{-1} s - j^\dagger s + \mathcal{H}_0}_{=\mathcal{H}_\mathcal{G}(d,s)} + \underbrace{\sum_{n=0}^{\infty} \frac{1}{n!} \Lambda^{(n)}_{x_1 ... x_n} s^{x_1} ... s^{x_n}}_{=\mathcal{H}_{\text{int}}(d,s)}$$

We aim for an expansion around the Gaussian specified by the free Hamiltonian $\mathcal{H}_\mathcal{G}(d, s)$. Thus, we want $\mathcal{H}_{\text{int}}(d, s)$ to be small. For this purpose we shift our field variable $s \to \varphi = s - t$ by subtracting a appropriately chosen t, $s = t + \varphi$ (e.g. $t = \mathrm{argmin}_s \mathcal{H}(d, s)$),

$$\begin{aligned}\mathcal{H}(d,\,\varphi|t) &= \mathcal{H}(d,\,s = t+\varphi)\\ &= \frac{1}{2}\varphi^\dagger D^{-1}\varphi - j'^\dagger\varphi + \mathcal{H}_0' + \sum_{n=0}^{\infty}\frac{1}{n!}\Lambda'^{(n)}_{x_1\ldots x_n}\varphi_{x_1}\ldots\varphi_{x_n},\end{aligned}$$

with,

$$\begin{aligned}\mathcal{H}_0' &= \mathcal{H}_0 - j^\dagger t + \frac{1}{2}t^\dagger D^{-1}t\\ j' &= j - D^{-1}t\\ \Lambda'^{(m)}_{x_1\ldots x_m} &= \sum_{n=0}^{\infty}\frac{1}{n!}\Lambda^{(m+n)}_{x_1\ldots x_{m+n}}t_{x_{m+1}}\ldots t_{x_{m+n}}.\end{aligned}$$

Exercise: Show that these formula are correct.

12.4 DIAGRAMMATIC PERTURBATION THEORY

(Following, Binney et al. [1])

$$\mathcal{H}(d,\,s) = \underbrace{\frac{1}{2}s^\dagger D^{-1}s - \underbrace{j^\dagger s}_{=\frac{1}{2}(j^\dagger s + s^\dagger j)}}_{=\mathcal{H}_\mathcal{G}(d,s)} + \mathcal{H}_0 + \underbrace{\sum_{n=0}^{\infty}\frac{1}{n!}\Lambda^{(n)}_{x_1\ldots x_n}s^{x_1}\ldots s^{x_n}}_{=\mathcal{H}_{\mathrm{int}}(d,s)}$$

$\Rightarrow$ partition function:

$$\begin{aligned}\mathcal{Z} &= \int \mathcal{D}s\, e^{-\mathcal{H}(d,s)} = \int \mathcal{D}s\, e^{-\mathcal{H}_\mathcal{G}(d,s)}\, e^{-\mathcal{H}_{\mathrm{int}}(d,s)}\\ &= \int \mathcal{D}s\, \underbrace{e^{-\mathcal{H}_\mathcal{G}(d,s)}}_{\propto\mathcal{G}(s-m,D)}\sum_{m=0}^{\infty}\frac{1}{m!}\left(\sum_{n=0}^{\infty}\frac{1}{n!}\Lambda^{(n)}_{x_1\ldots x_n}s^{x_1}\ldots s^{x_n}\right)^m\end{aligned}$$

Let us have a look at a simple case of a **local and position independent anharmonic term** first,

$$\begin{aligned}\Lambda^{(4)}_{x_1\ldots x_4} &= \delta(x_1-x_2)\delta(x_1-x_3)\delta(x_1-x_4)\lambda\\ \Rightarrow \mathcal{H}_{\mathrm{int}} &= \frac{\lambda}{4!}\int dx_1\,dx_2\,dx_3\,dx_4\,\delta(x_1-x_2)\,\delta(x_1-x_3)\,\delta(x_1-x_4)\,s^{x_1}s^{x_2}s^{x_3}s^{x_4}\\ &= \frac{\lambda}{4!}\int dx_1\delta^{x_1}_{x_2}\delta^{x_1}_{x_3}\delta^{x_1}_{x_4}\,s^{x_1}s^{x_2}s^{x_3}s^{x_4}\\ &= \frac{\lambda}{4!}\int dx_1(s^{x_1})^4\end{aligned}$$

$\Rightarrow$ partition function:

$$\mathcal{Z} = \int \mathcal{D}s\, e^{-\mathcal{H}_\mathcal{G}}\sum_{m=0}^{\infty}\frac{1}{m!}\left[-\frac{\lambda}{4!}\int dx\,\left(s^x\right)^4\right]^m$$

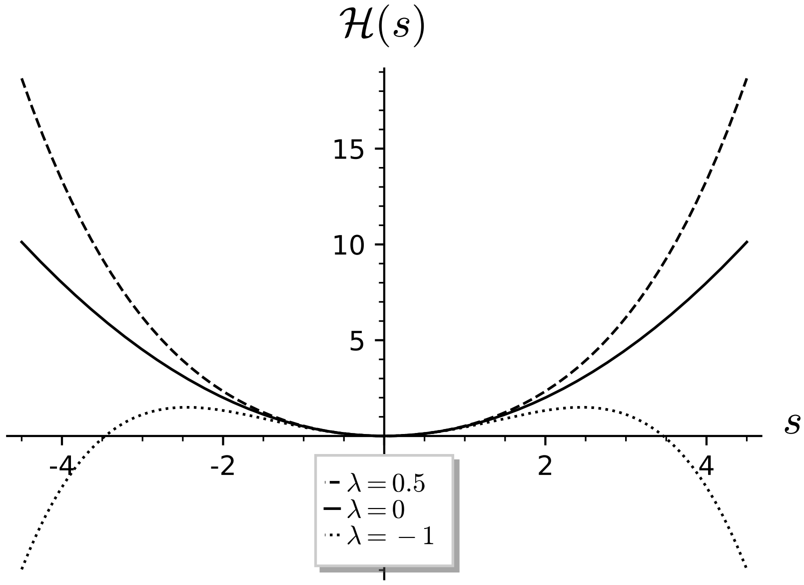


Figure 11: $\mathcal{H}(s) = \frac{1}{2}s^2 + \frac{\lambda}{4!}s^4$ for three values of λ, illustrating that next below $\lambda = 0$ the information Hamiltonian becomes unbound from below and consequently the partition function diverges. This is the reason why the expansion has a convergence radius of zero and is only an asymptotic expansion.

Using asymptotic expansion:

$$
\begin{aligned}
\mathcal{Z} &= \sum_{n=0}^{\infty} \frac{1}{n!} \int \mathcal{D}s\, e^{-\mathcal{H}_{\mathcal{G}}} \left[-\frac{\lambda}{4!} \int dx\, (s^x)^4 \right]^n \\
&= \sum_{n=0}^{\infty} \frac{1}{n!} \left\langle \left[-\frac{\lambda}{4!} \int dx\, (s^x)^4 \right]^n \right\rangle_{\mathcal{G}} \mathcal{Z}_{\mathcal{G}} \\
&= \sum_{n=0}^{\infty} \frac{1}{n!} \left(-\frac{\lambda}{4!} \int dx\, \frac{\delta^4}{\delta j_x^4} \right)^n \int \mathcal{D}s\, e^{-\frac{1}{2}s^\dagger D^{-1} s + j^\dagger s - \mathcal{H}_0} \\
&= \exp\left(-\frac{\lambda}{4!} \int dx\, \frac{\delta^4}{\delta j_x^4} \right) \mathcal{Z}_{\mathcal{G}}(j) \\
&= \exp\left[-\mathcal{H}_{\text{int}} \left(\frac{\delta}{\delta j} \right) \right] \mathcal{Z}_{\mathcal{G}}(j)
\end{aligned}
$$

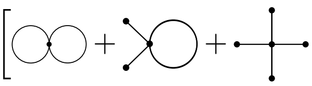
This result is also true in general and not only for our simple example. The Gaussian partition function $\mathcal{Z}_{\mathcal{G}}(j)$ is given by,

$$
\begin{aligned}
\mathcal{Z}_{\mathcal{G}}(j) &= \int \mathcal{D}s\, e^{-\frac{1}{2}s^\dagger D^{-1} s + j^\dagger s - \mathcal{H}_0} \\
&= \underbrace{e^{-\mathcal{H}_0} |2\pi D|^{1/2}}_{=\mathcal{Z}_{\mathcal{G}}(0)} e^{+\frac{1}{2}j^\dagger D j} \\
&= \mathcal{Z}_{\mathcal{G}}(0)\, e^{+\frac{1}{2}j^\dagger D j},
\end{aligned}
$$

with a j-independent prefactor $\mathcal{Z}_{\mathcal{G}}(0)$. Next, we expand $\mathcal{Z}$ to first order in λ for our considered simple case (for simplicity we only assume real j),

$$\begin{aligned}
\mathcal{Z}(j) &= \left(1 - \frac{\lambda}{4!}\int dx\, \frac{\delta^4}{\delta j_x^4} + \mathcal{O}(\lambda^2)\right) e^{\frac{1}{2}j^\dagger D j}\, \mathcal{Z}_{\mathcal{G}}(0) \\
&= \mathcal{Z}_{\mathcal{G}}(j) - \frac{\lambda}{4!}\mathcal{Z}_{\mathcal{G}}(0)\int dx\, \frac{\delta^4}{\delta j_x^4} e^{\frac{1}{2}j_y D^{yz} j_z} \\
&= \mathcal{Z}_{\mathcal{G}}(j) - \frac{\lambda}{4!}\mathcal{Z}_{\mathcal{G}}(0)\int dx\, \frac{\delta^3}{\delta j_x^3} D^{xz} j_z e^{\frac{1}{2}j^\dagger D j} \\
&= \mathcal{Z}_{\mathcal{G}}(j) - \frac{\lambda}{4!}\mathcal{Z}_{\mathcal{G}}(0)\underbrace{\int dx\, \frac{\delta^2}{\delta j_x^2}\left[D^{xx} + (D^{xz}j_z)^2\right] e^{\frac{1}{2}j^\dagger D j}}_{=A} \\
A &= \int dx\, \frac{\delta}{\delta j_x}\left[2(D^{xz}j_z)D^{xx} + (D^{xz}j_z)\left((D^{xz}j_z)^2 + D^{xx}\right)\right] e^{\frac{1}{2}j^\dagger D j} \\
&= \int dx\, \frac{\delta}{\delta j_x}\left[3(D^{xz}j_z)D^{xx} + (D^{xz}j_z)^3\right] e^{\frac{1}{2}j^\dagger D j} \\
&= \int dx\, \left[3D^{xx} + 3(D^{xz}j_z)^2 D^{xx} + 3(D^{xz}j_z)^2 D^{xx} + (D^{xz}j_z)^4\right] e^{\frac{1}{2}j^\dagger D j} \\
\Rightarrow \mathcal{Z}(j) &= \mathcal{Z}_{\mathcal{G}}(j) - \lambda\int dx\, \left[\frac{1}{8}D^{xx}D^{xx} + \frac{1}{4}D^{xx}D^{xy}j_y D^{xz}j_z + \frac{1}{4!}(D^{xz}j_z)^4\right] \mathcal{Z}_{\mathcal{G}}(j)
\end{aligned}$$

The information propagator connects different locations. In order to describe these locations and the lines between them, Feynman defined a language:

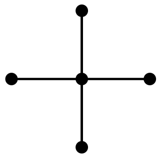
$$Z(j) = Z_{\mathcal{G}}(j) + \Bigg[$$

$$+\, \mathcal{O}(\lambda^2)\Bigg]\, Z_{\mathcal{G}}(j) \qquad (339)$$

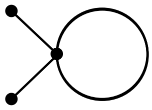
In general, one can say that $\mathcal{Z}(j)$ is the sum over all diagrams.

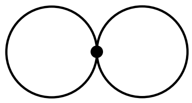
12.5 FEYNMAN RULES

- D^{xy} = line connecting x and y $\Rightarrow$
- j_y = vertex at the end of a line $\Rightarrow$
- $-\lambda$ = vertex with 4 ends $\Rightarrow$
- $-\Lambda^{(n)}_{x_1 \ldots x_n}$ = vertex with n ends
- all internal positions are intergrated over
- prefactor $= \frac{1}{\text{symmetry factor}}$, where the symmetry factor is given by the number of ways of reorderings of locations, which lead to equivalent integrals (loops account for a symmetry factor of 1/2).

EXAMPLES:

1.  $\Rightarrow -\frac{\lambda}{4!}\int dx\, D^{xz} j_z D^{xy} j_y D^{xv} j_v D^{xu} j_u = -\frac{\lambda}{4!}(Dj)^4$

2.  $\Rightarrow -\frac{\lambda}{4}\int dx\, D^{xx} D^{xy} j_y D^{xz} j_z = -\frac{\lambda}{4}(Dj)^2 \mathrm{diag}(D) = -\frac{\lambda}{4}(Dj)^2 \hat{D}$

3.  $\Rightarrow -\frac{\lambda}{8}\int dx\, D^{xx} D^{xx} = -\frac{\lambda}{8}\hat{D}^2$

Theorem: $\ln \mathcal{Z}(j)$ =sum over all connected diagrams

proof:

- define a set of all connected diagrams $\{C_i\}_i$
- define disconnected diagrams $D = D(\{n_i\})$ composed of n_i copies of C_i $\forall i$

We defined $\mathcal{Z}(j)$ as the sum over all disconnected diagrams,

$$\begin{aligned}
\mathcal{Z}(j) &= \sum_{\{n\}} D(\{n\}) \\
&= \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} ...D(\{n\}).
\end{aligned}$$

Next we use that $D(\{n\})$ is composed out of a product of numbers n_i of connected diagrams C_i,

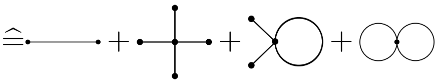
$$\begin{aligned}
\mathcal{Z}(j) &= \prod_{i=1}^{\infty}\left(\sum_{n_i=0}^{\infty}\frac{(C_i)^{n_i}}{n_i!}\right) \\
&= \prod_{i=1}^{\infty}\exp(C_i) \\
&= \exp(\sum_i C_i) \\
\ln \mathcal{Z}(j) &= \sum_i C_i.
\end{aligned}$$

12.6 DIAGRAMMATIC EXPECTATION VALUES

In our simplified example of a real ϕ^4 theory, we have a hamiltonian,

$$\mathcal{H}(\phi) = \frac{1}{2}\phi^\dagger D^{-1}\phi - j^\dagger\phi + \frac{1}{4!}\lambda^\dagger\phi^4.$$

In this case we can write the logarithm of the partition sum using Feynman rules,

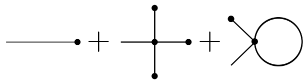
$\ln \mathcal{Z}(j) \widehat{=}$  $+ \mathcal{O}(\lambda^2)$

$$\Rightarrow \ln \mathcal{Z}(j) \widehat{=} \frac{1}{2} j^\dagger D j - \frac{\lambda}{4!}(Dj)^4 - \frac{1}{4}\lambda (Dj)^2 \widehat{D} - \frac{1}{8}\lambda \widehat{D}^2.$$

1. Calculate the expectation value of this field:

$$\begin{aligned} \langle \phi \rangle &= \frac{\delta \ln \mathcal{Z}}{\delta j} \\ &= Dj - \frac{1}{3!} D\lambda (Dj)^3 - \frac{1}{2} D\lambda (Dj)\widehat{D} + \mathcal{O}(\lambda^2) \\ \langle \phi^x \rangle &= D^{xy} j_y - \frac{\lambda}{3!}\int dy\, D^{yx}(D^{yz} j_z)^3 - \frac{\lambda}{2}\int dy\, D^{yx}(D^{yz} j_z) D^{yy} + \mathcal{O}(\lambda^2) \end{aligned}$$

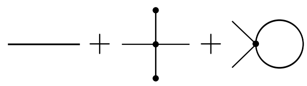
   Written in Feynman rules:

   $\langle \phi^x \rangle =$  $+ \mathcal{O}(\lambda^2)$

   $\Rightarrow$ j-derivatives can be calculated directly from diagrams by cutting end-dots/ end-vertices.

2. Calculate the covariance of the field:

$$\begin{aligned} \langle \phi^x \phi^y \rangle^{\mathrm{c}} &= \langle (\phi - \langle \phi \rangle)_x (\phi - \langle \phi \rangle)_y \rangle \\ &= \frac{\delta^2 \ln \mathcal{Z}(j)}{\delta j_x \delta j_y} \end{aligned}$$

   $\langle \phi^x \phi^y \rangle^{\mathrm{c}} =$  $+ \mathcal{O}(\lambda^2)$

   Rewrite the Feynman diagrams:

$$\begin{aligned} \langle \phi^x \phi^y \rangle^c &= D^{xy} - \frac{\lambda}{2}\int dz\, D^{zx}(D^{zu} j_u)^2 D^{zy} - \frac{\lambda}{2}\int dz\, D^{xz} D^{zz} D^{zy} + \mathcal{O}(\lambda^2) \\ \langle \phi \phi^\dagger \rangle^c &= D - \frac{\lambda}{2} D (Dj)^2 D - \frac{\lambda}{2} D(\widehat{D}) D + \mathcal{O}(\lambda^2) \end{aligned}$$

If we consider no inharmonic term in our hamiltonian ($\lambda = 0$), we get the Wiener filter solution:

$$\begin{aligned} \langle \phi \rangle &= Dj \\ \langle \phi \phi^\dagger \rangle^c &= D \end{aligned}$$

12.7 LOG-NORMAL POISSON MODEL DIAGRAMMATICALLY

The joint Hamiltonian of the log-normal Poisson model with $\rho = \rho_0 e^s$is given by,

$$
\begin{aligned}
\mathcal{H}(d,s) &\widehat{=} \frac{1}{2}s^\dagger S^{-1}s - d^\dagger s + \kappa^\dagger e^s \\
&\widehat{=} \frac{1}{2}s^\dagger(S^{-1} + \underbrace{\widehat{\kappa}}_{\widehat{\kappa}^{xy}=\kappa\delta^{xy}})s - \underbrace{(d-\kappa)^\dagger}_{=j^\dagger} s + \kappa^\dagger \sum_{n=3}^{\infty}\frac{1}{n!}s^n \\
&\widehat{=} \frac{1}{2}s^\dagger D^{-1}s - j^\dagger s + \sum_{n=0}^{\infty}\frac{1}{n!}\underbrace{\Lambda^{(n)}_{x_1\ldots x_n}}_{=\kappa_{x_1}\delta(x_1-x_2)\ldots\delta(x_1-x_n)} s^{x_1}\ldots s^{x_n}
\end{aligned}
$$

Find the mean m using the MAP:

$$
\begin{aligned}
\left.\frac{\delta H}{\delta s^\dagger}\right|_{s=m} &\stackrel{!}{=} 0 \\
&= D^{-1}s - j + \kappa\sum_{n=3}^{\infty}\left.\frac{s^{n-1}}{(n-1)!}\right|_{s=m} \\
\Rightarrow m &= D\left(j - \kappa\sum_{n=2}^{\infty}\frac{m^n}{n!}\right)
\end{aligned}
$$

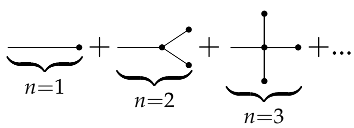
Iteration: Take the simplest guess $m_0 = 0$.

1.

$$m_1 = Dj =$$

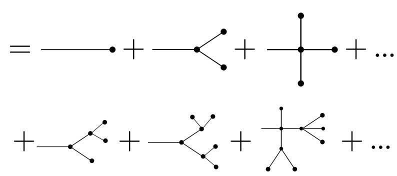
2.

$$m_2 = D\left(j - \kappa\sum_{n=2}^{\infty}\frac{(Dj)^n}{n!}\right) =$$



3.

$$m_2 = D\left(j - \kappa\sum_{n=2}^{\infty}\frac{(m_2)^n}{n!}\right) =$$



$\boxed{\Rightarrow \text{The classical/ MAP estimate } m_\infty \text{ is always given by the the sum of all tree diagrams with one external point.}}$

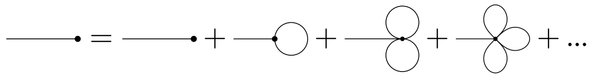
12.7.1 *Consideration of uncertainty loops*

But, $m \neq \langle s\rangle_{(s|d)}$, since $\langle s\rangle_{(s|d)}$ is the sum of all Feynman diagrams (loop and tree diagrams) with one external point,

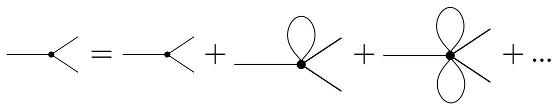
$$\langle s \rangle_{(s|d)} = \underbrace{\sum \text{tree diagrams}}_{\text{MAP}} + \underbrace{\sum \text{loop diagrams}}_{\text{uncertainty corrections}}$$

Let's try to add some loops to the MAP estimator by augmenting the considered vertices with loops, by performing the following replacements:

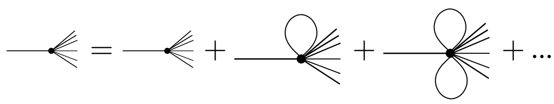
- source:



- 3-vertex:



- n-vertex:



$$\Rightarrow -\kappa_x \to -\left[\kappa_x + \frac{1}{2}\kappa_x \int dx\, D^{xx} + \frac{1}{8}\kappa_x \int dx\, D^{xx} + ... + \frac{1}{n!2^n}\kappa_x \left(\int dx\, \widehat{D}^x\right)^n\right]$$
$$-\kappa \to -\kappa\, e^{\widehat{D}/2}$$

The factor $e^{\widehat{D}/2}$ accounts for the loop corrections to the classical map,

$$m = S(d - \underbrace{\kappa e^m}_{=\kappa_m})$$

So, we can calculate the corrected map defining $\kappa_m \to \kappa_m e^{\widehat{D}/2} = \kappa_{m+\widehat{D}/2}$.

loop normalized solution:

$$m = S\left(d - \kappa_{m+\widehat{D}/2}\right)$$
$$D = \left(S^{-1} + \widehat{\kappa}_{m+\widehat{D}/2}\right)^{-1}$$
$$\kappa_t = \kappa e^t$$

Still, $m \approx \langle s \rangle_{(s|d)}$ is just an approximation, since m does not contain topological more complex diagrams like vacuum polarization diagrams.

13 THERMODYNAMICAL INFERENCE

13.1 BASICS

Following Enßlin & Weig [2010arxiv:1004.2868]
Considering a tempered posterior at $T = \frac{1}{\beta}$ with an moment generating source J,

$$\begin{aligned} \mathcal{P}(s|d, T, J) &= \frac{e^{-\beta(\mathcal{H}(d,s)+J^\dagger s)}}{Z_\beta(d, \beta, J)} \\ &= \frac{\left(\mathcal{P}(d, s)e^{-J^\dagger s}\right)^\beta}{\underbrace{\int \mathcal{D}s \left(\mathcal{P}(d, s)e^{-J^\dagger s}\right)^\beta}_{=Z(d,\beta,J)}} \end{aligned} \tag{340}$$

- $T = \beta = 1$: usual inference
- $T \to 0$, $\beta \to \infty$: enlarged contrast $\Rightarrow \mathcal{P}(s|d, T) \to \delta(s - s_{\mathrm{MAP}})$
- $T \to \infty$, $\beta \to 0$: weaker contrast $\Rightarrow \mathcal{P}(s|d, T) \to$ const.

Calculate the Boltzmann Entropy S_B with respect to a prior $q(s)$, which is equal to the Entropy we have defined as the negative information content of a system for $T = 1$ and $J = 0$:

$$S_B = -\int \mathcal{D}s\, \mathcal{P}(s|d, T, J) \ln\left(\frac{\mathcal{P}(s|d, T, J)}{q(s)}\right) \tag{341}$$

$$\begin{aligned} \Delta S_B &= \int \mathcal{D}s\, \mathcal{P}(s|d, T, J) \left[\beta(\mathcal{H}(d, s) + J^\dagger s) + \ln Z(d, \beta, J)\right] \\ &= \beta \left[\underbrace{\langle \mathcal{H}(d, s)\rangle_{(s|d,T,J)}}_{=U(d,T,J)} + J^\dagger \underbrace{\langle s\rangle_{(s|d,T,J)}}_{=m(d,T,,J)} + \underbrace{\frac{1}{\beta}\ln Z(d, \beta, J)}_{=-F(d,\beta,J)}\right] \end{aligned} \tag{342}$$

$$T\Delta S_B = U(d, T, J) + J^\dagger m(d, T, J) - F(d, \beta, J) \tag{343}$$

- internal energy: $U(d, T, J) = \langle \mathcal{H}(d, s)\rangle_{(s|d,T,J)}$
- mean field: $m(d, T, J) = \langle s\rangle_{(s|d,T,J)} = \left.\frac{\partial F}{\partial J}\right|_{J=0,\beta=1}$
- Helmholtz free energy: $F(d, \beta, J) = -\frac{1}{\beta}\ln Z(d, \beta, J)$

Inference goal:

Find $m = \langle s \rangle_{(s|d)}$ and its uncertainty $D = \langle s s^\dagger \rangle^c_{(s|d)}$. If we only get a report on m and D the maximum entropy principle requires the PDF to be Gaussian,

$$\widetilde{\mathcal{P}}\,(s|m, D) \;=\; \mathcal{G}(s - m, D). \tag{344}$$

If this is all we are aiming for, we can adopt the Gaussian PDF right from the beginning and try to infer m, D using thermodynamical methods.
Ansatz:

$$\boxed{\begin{aligned} \mathcal{P}(s|d, T, J) &\approx \widetilde{\mathcal{P}}\,(s|m, D) = \mathcal{G}(s - m, D) && (345)\\ T\Delta\,\widetilde{S_B}\,(d, T, J) &= \widetilde{U}\,(d, T, J) + J^\dagger m(d, T, J) - \widetilde{F}\,(d, \beta, J) && (346)\\ \widetilde{U}\,(d, T, J) &= \langle \mathcal{H}(d, s) \rangle_{\mathcal{G}(s-m, D)} && (347) \end{aligned}}$$

From now on we just ignore the reference probability $q(s)$ or assume $q(s) = 1$ and write S_B instead of ΔS_B.

$$\begin{aligned} \Rightarrow \widetilde{S_B}\,(d, T, J) &= -\langle \ln \widetilde{\mathcal{P}} \rangle_{\widetilde{\mathcal{P}}} \\ &= +\int \mathcal{D}s\, \mathcal{G}(s - m, D) \left[\frac{1}{2} \underbrace{(s-m)^\dagger}_{\varphi^\dagger} D^{-1}(s - m) + \frac{1}{2} \ln |2\pi D| \right] \\ &= \frac{1}{2} \left[\int \mathcal{D}\varphi \left(\mathcal{G}(\varphi, D)\, \mathrm{Tr}(\varphi\varphi^\dagger D^{-1}) \right) + \ln |2\pi D| \right] \\ &= \frac{1}{2} \mathrm{Tr} \left(\underbrace{\langle \varphi\varphi^\dagger \rangle_{\mathcal{G}(\varphi, D)}}_{=D} D^{-1} \right) + \frac{1}{2} \underbrace{\ln |2\pi D|}_{=\mathrm{Tr}(\ln |2\pi D|)} \\ &= \frac{1}{2} \mathrm{Tr}(D D^{-1}) + \frac{1}{2} \mathrm{Tr}(\ln |2\pi D|) \\ &= \frac{1}{2} \mathrm{Tr}(\mathbb{1} + \ln |2\pi D|) \\ &= \widetilde{S_B}\,(D) \end{aligned} \tag{348}$$

$$\Rightarrow \widetilde{F}\,(d, \beta, J) \;=\; \widetilde{U}\,(m_J, D_J) - T\,\widetilde{S_B}\,(D_J) + J^\dagger m_J \tag{349}$$

The solution m_J we are looking for, is in this case a function of J. We want to get rid of the dependence on J by using the Legendre transformation.

LEGENDRE TRANSFORMATION The Legendre transformation uses an ensemble of tangents on our function $F(J)$ in order to describe it.

$$F(J) \;=\; F(J_0) + \left.\frac{\partial F}{\partial J}\right|_{J_0}^\dagger (J - J_0) + \ldots \tag{350}$$

$$G \;=\; F(J_0) - \left.\frac{\partial F}{\partial J}\right|_{J_0}^\dagger J_0 \tag{351}$$

If F is convex $\Rightarrow m_J = \frac{\partial F}{\partial J}$ and F can be reconstructed from $G(m)$, if G is known for every slope m of F.

Gibbs free energy:

$$\begin{aligned} G &= F - \frac{\partial F}{\partial J}^{\dagger} J \\ &= U - TS_B + J^{\dagger}m - J^{\dagger}m \end{aligned} \tag{352}$$

$$\Rightarrow \widetilde{G}\,(d,\beta,m,D) = \widetilde{U}\,(d,\beta,m,D) - T\,\widetilde{S_B}\,(D) \tag{353}$$

Now, we can calculate the mean field m and the uncertainty dispersion D from the defined Gibbs free energy G.

mean field from minimal Gibbs free energy:

$$\frac{\delta G(d,m,D)}{\delta m} = 0 \quad \Rightarrow \quad m = \langle s \rangle_{(s|d)}\Big|_{T=1} \tag{354}$$

proof:

$$\begin{aligned} \frac{\delta G}{\delta m} &= \frac{\delta}{\delta m}\Big(F(d,J(m)) - J^{\dagger}(m)m\Big) \\ &= \frac{\delta J(m)}{\delta m}^{\dagger}\underbrace{\frac{\delta F(d,J)}{\delta J}}_{=m(J)} - \frac{\delta J^{\dagger}}{\delta m}m - J \\ &= -J \stackrel{!}{=} 0 \end{aligned}$$

$$J = 0 \quad \Rightarrow \quad m = \frac{\partial F}{\partial J}\Big|_{J=0} = \langle s \rangle_{(s|d)}$$

uncertainty dispersion:

$$\left(\frac{\delta^2 G}{\delta m \delta m^{\dagger}}\right)^{-1}\Bigg|_{m=\langle s\rangle_{(s|d)}} = \frac{-\delta^2 F}{\delta J \delta J^{\dagger}}\Big|_{J=0} = \beta D \tag{355}$$

proof:

$$\begin{aligned} \left(\frac{\delta^2 G}{\delta m \delta m^{\dagger}}\right)^{-1}\Bigg|_{m=\langle s\rangle_{(s|d)}} &= \left(-\frac{\delta J}{\delta m}\right)^{-1}\Bigg|_{m=\langle s\rangle_{(s|d)}} \\ &= -\left(\frac{\delta m(J)}{\delta J}\right)\Bigg|_{J=0} \\ &= -\frac{\delta^2 F(J)}{\delta J \delta J^{\dagger}}\Big|_{J=0} \\ &= \frac{1}{\beta}\underbrace{\frac{\delta^2}{\delta J \delta J^{\dagger}}\ln Z(d,J,\beta)}_{=\beta^2 D} \\ &= \beta D \end{aligned}$$

13.1.1 *Lognormal Poisson model*

- $\mathcal{P}(s) = \mathcal{G}(s, S)$
- $\lambda(s) = \kappa e^s$
- $\mathcal{P}(d^x|\lambda^x) = \frac{(\lambda^x)^{d^x} e^{-\lambda^x}}{d^x!}$

$$\begin{aligned}
\Rightarrow \mathcal{H}(d, s) &\widehat{=} \frac{1}{2}s^\dagger S^{-1}s - d^\dagger s + \kappa^\dagger e^s \\
\widetilde{U}(m, D) &= \langle \mathcal{H}(d, s)\rangle_{\mathcal{G}(s-m, D)}
\end{aligned}$$

$$\begin{aligned}
\langle s^\dagger S^{-1} s\rangle_{\mathcal{G}(s-m, D)} &= \mathrm{Tr}(S^{-1}\langle ss^\dagger\rangle_{\mathcal{G}(s-m, D)}) \\
&= \mathrm{Tr}(S^{-1}\langle (m+\varphi)(m+\varphi)^\dagger\rangle_{\mathcal{G}(\varphi, D)}) \\
&= \mathrm{Tr}(S^{-1}(mm^\dagger + D)) \\
&= m^\dagger S^{-1} m + \mathrm{Tr}(S^{-1}D) \\
\langle s\rangle_{\mathcal{G}(s-m, D)} &= m \\
\langle e^{s_x}\rangle_{\mathcal{G}(s-m, D)} &= \int \mathcal{D}\varphi\, \mathcal{G}(\varphi, D) e^{m_x + \varphi_x} \\
& \text{writing } j^\dagger\varphi \text{ for } \varphi_x \text{ with } j_y = \delta(y-x) \\
&= e^{m_x} \int \mathcal{D}\varphi\, \frac{\exp\left(-\frac{1}{2}\varphi^\dagger D^{-1}\varphi + j^\dagger\varphi\right)}{|2\pi D|^{1/2}} \\
&= e^{m_x} e^{\frac{1}{2}j^\dagger D j} \\
&= e^{m_x + \frac{1}{2}D_{xx}}
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \widetilde{U}(m, D) &= \frac{1}{2}m^\dagger S^{-1}m + \frac{1}{2}\mathrm{Tr}(DS^{-1}) - d^\dagger m + \kappa^\dagger e^{m+\frac{1}{2}\hat{D}} \\
\widetilde{S_B}(D) &= \frac{1}{2}\mathrm{Tr}(1 + \ln(2\pi D)) \\
\widetilde{G}(m, D) &= \widetilde{U}(m, D) - T\,\widetilde{S_B}(D) \\
&= \frac{1}{2}m^\dagger S^{-1}m + \frac{1}{2}\mathrm{Tr}(DS^{-1}) - d^\dagger m + \kappa^\dagger e^{m+\frac{1}{2}\hat{D}} - \frac{T}{2}\mathrm{Tr}(1+\ln(2\pi D))
\end{aligned}$$

mean map:

$$\begin{aligned}
0 &\stackrel{!}{=} \frac{\delta\,\widetilde{G}(m, D)}{\delta m} = S^{-1}m - d + \kappa^\dagger e^{m+\frac{1}{2}\hat{D}} \\
\Rightarrow m &= S(d - \kappa^\dagger e^{m+\frac{1}{2}\hat{D}})
\end{aligned}$$

uncertainty dispersion:

$$
\begin{aligned}
D &= T\left(\frac{\delta^2 G}{\delta m \delta m^\dagger}\right)^{-1} \\
&= T\left(\frac{\delta}{\delta m}(S^{-1}m - d + \kappa e^{m+\frac{1}{2}\hat{D}})\right)^{-1} \\
&= T\left(S^{-1} + \widehat{\kappa e^{m+\frac{1}{2}\hat{D}}}\right)^{-1}
\end{aligned}
$$

$$
\begin{aligned}
m &= S(d - \kappa e^{m+\frac{1}{2}\hat{D}}) \\
D &= T\left(S^{-1} + \widehat{\kappa e^{m+\frac{1}{2}\hat{D}}}\right)^{-1}
\end{aligned}
$$

$\Rightarrow$The loop-vertex normalized classical solution is the minimal Gibbs energy solution in a Gaussian posterior approximation.

13.1.2 *Mutual information and Gibbs free energy*

Define the KL-distance,

$$
D_{KL}[\widetilde{\mathcal{P}}, \mathcal{P}] = S[\widetilde{\mathcal{P}}, \mathcal{P}] = -\int \mathcal{D}s\ \widetilde{\mathcal{P}}\ (s|d)\ \ln \frac{\widetilde{\mathcal{P}}\ (s|d)}{\mathcal{P}(s|d)}.
$$

$$
\begin{aligned}
\Rightarrow \widetilde{G}\ (m, D) &= \langle \underbrace{\mathcal{H}(d, s)}_{=-\ln \mathcal{P}(d,s)} + \underbrace{\ln \mathcal{G}(s-m, D)}_{=-S_B} \rangle_{\mathcal{G}(s-m, D)} \\
&= \int \mathcal{D}s\, \mathcal{G}(s-m, D) \ln \frac{\mathcal{G}(s-m, D)}{\mathcal{P}(d,s)} \\
&= \int \mathcal{D}s\, \mathcal{G}(s-m, D) \ln \frac{\mathcal{G}(s-m, D)}{\mathcal{P}(s|d)} - \ln \mathcal{P}(d) \\
&\widehat{=} \int \mathcal{D}s\, \mathcal{G}(s-m, D) \ln \frac{\mathcal{G}(s-m, D)}{\mathcal{P}(s|d)} \\
&= \int \mathcal{D}s \widetilde{\mathcal{P}}(s|d) \ln \frac{\widetilde{\mathcal{P}}\ (s|d)}{\mathcal{P}(s|d)}
\end{aligned}
$$

$\Rightarrow$The Gibbs free energy describes up to a constant the mutual information,

$$
\widetilde{G}\ (m, D) = D_{KL}[\widetilde{\mathcal{P}}, \mathcal{P}].
$$

The minimal Gibbs free energy is equal to a minimal KL-distance and to a maximal mutual information of $\widetilde{\mathcal{P}}$ on $\mathcal{P}$.

13.2 OPERATOR CALCULUS FOR INFORMATION FIELD THEORY

Following Leike & Enßlin [8]
task: calculate the Gaussian average

$$\langle f(s)\rangle_{\mathscr{G}(s-m,D)}$$

for a Gauss distribution in s with mean m and covariance D.
More simple task:

$$\langle s\rangle_{\mathscr{G}(s-m,D)} = \int \mathcal{D}s\, \frac{s\, e^{(s-m)^\dagger D^{-1}(s-m)}}{|2\pi D|^{\frac{1}{2}}}$$

Observe: $\frac{\mathrm{d}}{\mathrm{d}m}\mathscr{G}(s-m,D) = D^{-1}(s-m)\,\mathscr{G}(s-m,D)$
equivalently: $(D\frac{\mathrm{d}}{\mathrm{d}m}+m)\,\mathscr{G}(s-m,D) = s\,\mathscr{G}(s-m,D)$

$$\begin{aligned}\langle s\rangle_{\mathscr{G}(s-m,D)} &= \int \mathcal{D}s\, \left(D\frac{\mathrm{d}}{\mathrm{d}m}+m\right)\mathscr{G}(s-m,D)\\ &= \left(D\frac{\mathrm{d}}{\mathrm{d}m}+m\right)\int \mathcal{D}s\,\mathscr{G}(s-m,D)\\ &= \left(D\frac{\mathrm{d}}{\mathrm{d}m}+m\right)\underbrace{\langle 1\rangle_{\mathscr{G}(s-m,D)}}_{=1} = m \qquad (356)\end{aligned}$$

Works for any moment of the Gaussian

$$\langle s^n\rangle_{\mathscr{G}(s-m,D)} = \left(D\frac{\mathrm{d}}{\mathrm{d}m}+m\right)^n 1\,. \qquad (357)$$

$\Phi := D\frac{\mathrm{d}}{\mathrm{d}m}+m$ the field operator.
vacuum vector $1:\ m\mapsto 1$ is functional that maps any field m to 1

arbitrary analytical function $f(s) = \sum_{i=0}^{\infty}\lambda_i s^i$:

$$\begin{aligned}\langle f(s)\rangle_{\mathscr{G}(s-m,D)} &= \sum_{i=0}^{\infty}\lambda_i\left\langle s^i\right\rangle_{\mathscr{G}(s-m,D)}\\ &= \sum_{i=0}^{\infty}\lambda_i\left\langle \Phi^i\right\rangle_{\mathscr{G}(s-m,D)}\\ &= \sum_{i=0}^{\infty}\lambda_i\Phi^i 1 = f(\Phi)\,1 \qquad (358)\end{aligned}$$

Instead of calculating the expectation value of $f(s)$ with respect to a Gaussian distribution we can calculate the vacuum expectation value of the operator $f(\Phi)$. We will motivate why this is useful by illustrative examples.

annihilation operator $a := D\frac{\mathrm{d}}{\mathrm{d}m}$, $a^x = D\,\frac{\mathrm{d}}{\mathrm{d}m^y}$
creation operator $a^+ := m$, $a^{+x} = m^x$
Canonical commutation relations:

$$\begin{aligned}[][a^x, a^y] &= [a^{+x}, a^{+y}] = 0\\ [a^x, a^{+y}] &= D^{xy}\,. \qquad (359)\end{aligned}$$

Strategy: Separate $\Phi = a + a^+$ and try to get the annihilation operators to the right hand side, where they annihilate on the vacuum: $a^x 1 = D^{xy} \frac{\mathrm{d}}{\mathrm{d}m^y} 1 = 0$.
Illustration 1:

$$\begin{aligned}
\langle s^x s^y \rangle_{\mathscr{G}(s-m,D)} &= \Phi^x \Phi^y 1 = \left(a^x + a^{+x}\right)\left(a^y + a^{+y}\right) 1 \\
&= \left(a^x a^y + a^{+x} a^y + a^x a^{+y} + a^{+x} a^{+y}\right) 1 \\
&= \left(0 + 0 + a^{+y} a^x + [a^x, a^{+y}] + m^x m^y\right) 1 \\
&= D^{xy} + m^x m^y \qquad (360)
\end{aligned}$$

Illustrations 2: $\left\langle e^{s^x} \right\rangle_{\mathscr{G}(s-m,D)} = e^{\Phi^x} 1 = e^{a^x + a^{+x}} 1$.
We need the **Baker-Campbell-Hausdorff** (BCH) formula (without proof):

$$e^X Y = \sum_{n=0}^{\infty} [X, Y]_n \, e^X \qquad (361)$$

with $[X, Y]_n = [X, [X, Y]_{n-1}]$ and $[X, Y]_0 = Y$.
In case $[X, [X, Y]] = 0$ we have (without proof):

$$\begin{aligned}
e^X Y &= Y e^X + [X, Y] \, e^X \qquad (362) \\
e^{X+Y} &= e^X e^Y e^{\frac{1}{2}[X,Y]}. \qquad (363)
\end{aligned}$$

In case $X = a^x$, $Y = a^{+y}$ we have $[X, Y] = [a^x, a^{+y}] = D^{xy}$, which commutes with a and a^+ such that $[X, [X, Y]] = [a^x, D^{xy}] = 0$. Consequently:

$$\begin{aligned}
e^{a^x} a^{+y} &= a^{+y} e^{a^x} + [a^x, a^{+y}] \, e^{a^x} = a^{+y} e^{a^x} + D^{xy} e^{a^x} \qquad (364) \\
e^{a^x + a^{+y}} &= e^{a^{+y}} e^{a^x} e^{\frac{1}{2}[a^x, a^{+y}]} = e^{a^{+y}} e^{a^x} e^{\frac{1}{2} D^{xy}} \qquad (365)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\left\langle e^{s^x} \right\rangle_{\mathscr{G}(s-m,D)} &= e^{a^{+x}} e^{a^x} e^{\frac{1}{2} D^{xx}} 1 \\
&= e^{m^x + \frac{1}{2} D^{xx}} \left(1 + a^x + \frac{1}{2}(a^2)^x + \ldots\right) 1 \\
&= e^{m^x + \frac{1}{2} D^{xx}} \qquad (366)
\end{aligned}$$

since D^{xx} commutes with a^x and a^{+x} and since we have $a^x 1 = 0$.

Illustrations 3:

$$\begin{aligned}
\left\langle e^{s^x} e^{s^y} \right\rangle_{\mathscr{G}(s-m,D)} &= e^{\Phi^x} e^{\Phi^y} 1 \\
&= e^{a^x + a^{+x}} e^{a^y + a^{+y}} 1 \\
&= e^{a^{+x} + \frac{1}{2} D^{xx}} e^{a^x} e^{a^{+y} + \frac{1}{2} D^{yy}} e^{a^y} 1 \\
&= e^{a^{+x} + \frac{1}{2} D^{xx} + \frac{1}{2} D^{yy}} e^{a^x} e^{a^{+y}} 1
\end{aligned}$$

Now, we need the commutator $[e^{a^x}, e^{a^{+y}}]$, which can be calculated using the BCH formula twice:

$$
\begin{aligned}
[e^{a^x}, e^{a^{+y}}] &= e^{a^x} e^{a^{+y}} - e^{a^{+y}} e^{a^x} \\
&= e^{a^x + a^{+y} + \frac{1}{2}D^{xy}} - e^{a^x + a^{+y} - \frac{1}{2}D^{xy}} \\
&= e^{a^x + a^{+y}} \left(e^{\frac{1}{2}D^{xy}} - e^{-\frac{1}{2}D^{xy}} \right) \\
&= e^{a^{+y}} e^{a^x} e^{\frac{1}{2}D^{xy}} \left(e^{\frac{1}{2}D^{xy}} - e^{-\frac{1}{2}D^{xy}} \right) \\
&= e^{a^{+y}} e^{a^x} \left(e^{D^{xy}} - 1 \right) \\
e^{a^x} e^{a^{+y}} &= e^{a^{+y}} e^{a^x} e^{D^{xy}} \\
\left\langle e^{s^x} e^{s^y} \right\rangle_{\mathscr{G}(s-m,D)} &= e^{a^{+x} + \frac{1}{2}D^{xx}} e^{D^{xy}} e^{a^{+y} + \frac{1}{2}D^{yy}} e^{a^x} 1 \\
&= e^{m^x + \frac{1}{2}D^{xx}} e^{D^{xy}} e^{m^y + \frac{1}{2}D^{yy}}
\end{aligned}
$$

Illustrations 4:

$$
\begin{aligned}
\left\langle e^{s^x} s^y \right\rangle_{\mathscr{G}(s-m,D)} &= e^{\Phi^x} \Phi^y 1 \\
&= e^{a^x + a^{+x}} \left(a^y + a^{+y} \right) 1 \\
&= e^{a^{+x} + \frac{1}{2}D^{xx}} e^{a^x} a^{+y} 1
\end{aligned}
$$

To exchange e^{a^x} and a^{+y} we use the fact that the commutator $[X, \_]$ has the algebraic properties of a derivation, meaning that it is linear and obeys the product rule

$$
\begin{aligned}
[X, YZ] &= XYZ - YZX \\
&= XYZ - YXZ + YXZ - YZX \\
&= [X, Y]\, Z + Y\, [X, Z].
\end{aligned}
$$

Together with the fact that $[a^{+y}, a^x]$ commutes with everything this implies that the commutator indeed works like taking the derivative with respect to a^{+y}. We calculate $[a^{+y}, e^{a^x}]$ step by step:

$$
\begin{aligned}
[a^{+y}, e^{a^x}] = [a^{+y}, \sum_{n=0}^{\infty} \frac{(a^n)^x}{n!}] &= \sum_{n=0}^{\infty} \frac{1}{n!} [a^{+y}, (a^n)^x] \\
&= \sum_{n=1}^{\infty} \frac{1}{n!} n [a^{+y}, a^x] (a^{n-1})^x \\
&= \sum_{n=1}^{\infty} \frac{1}{n!} n (a^{n-1})^x [a^{+y}, a^x] \\
&= -e^{a^x} D^{xy}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
e^{a^x} a^{+y} &= \left(a^{+y} + D^{xy} \right) e^{a^x} \\
\left\langle e^{s^x} s^y \right\rangle_{\mathscr{G}(s-m,D)} &= e^{a^{+x} + \frac{1}{2}D^{xx}} \left(a^{+y} + D^{xy} \right) e^{a^x} 1 \\
&= e^{m^x + \frac{1}{2}D^{xx}} \left(m^y + D^{xy} \right)
\end{aligned}
$$

Illustration 5:

$$
\begin{aligned}
\left\langle e^{s^x} e^{s^y} s^z \right\rangle_{\mathscr{G}(s-m,D)} &= e^{\Phi^x} e^{\Phi^y} \Phi^z 1 \\
&= e^{a^x + a^{+x}} e^{a^y + a^{+y}} \left(a^z + a^{+z} \right) 1 \\
&= e^{a^{+x} + \frac{1}{2} D^{xx}} e^{a^x} e^{a^{+y} + \frac{1}{2} D^{yy}} e^{a^y} a^{+z} 1 \\
&= e^{a^{+x} + \frac{1}{2} D^{xx}} e^{D^{xy}} e^{a^{+y} + \frac{1}{2} D^{yy}} (a^{+z} + D^{xz} + D^{yz}) e^{a^x} e^{a^y} 1 \\
&= e^{m^x + \frac{1}{2} D^{xx}} e^{D^{xy}} e^{m^y + \frac{1}{2} D^{yy}} \left(m^z + D^{xz} + D^{yz} \right) .
\end{aligned}
$$

14 RECONSTRUCTION WITHOUT SPECTRAL KNOWLEDGE

Following Enßlin & Weig [arxiv:1004.2868] and Enßlin & Frommert [arxiv:1002.2928]

I: A Gaussian random field s,

$$\mathcal{P}(s|S) = \mathcal{G}(s, S),$$

with unknown covariance $S = \langle ss^\dagger \rangle_{(s)}$ is observed with a linear response instrument,

$$d = Rs + n,$$

with Gaussian and signal independent noise n of known covariance $N = \langle nn^\dagger \rangle$,

$$\mathcal{P}(d, s|S) = \mathcal{G}(s, S)\mathcal{G}(d - Rs, N).$$

In the functional basis O the signal covariance S is diagonal is known .

For Example:

- statistical homogenity $\Rightarrow$Fourier basis: $O = F$
- statistical isotropy $\Rightarrow$spherical harmonics basis: $O = Y$

Strategy to estimate $m = \langle s \rangle_{(s|d, I)}$:

1. Develop theory for unknown s, S.
2. Marginalize unknown S: $\mathcal{H}(d, s) = -\ln \int \mathcal{D}S\, e^{-\mathcal{H}(d,s,S)}$
3. Solve effective theory for s.

14.1 SPECTRAL RESPRESENTATION OF S

$$\begin{aligned} S &= O^\dagger \hat{P}_s O \\ S_{xy} &= O^\dagger_{xk} \hat{P}_s(k) O_{ky} \end{aligned}$$

In the special case of $O = F$ we obtain,

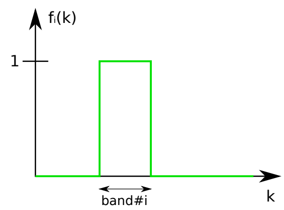
$$S_{xy} = \int e^{-ixk} P_s(k) e^{iyk} dk.$$

Now, we model the spectrum of $P_s(k)$ as a linear combinatiom of positive basis functions $f_i(k)$ with disjoint support (spectral bands) covering all of the relevant k-space,

$$P_s(k) = \sum_i f_i(k) p_i.$$

The basis functions are given by the indicator functions,

$$f_i(k) = P(x \in \text{band\#i}|k, \text{band\#i}),$$



and p_i are the corresponding spectral coefficients. We request that the bands cover completely the Fourier space and do not overlap.
Define the **spectral band matrices**:

$$\begin{aligned}
(T_i)_{xy} &= (O^\dagger \hat{f}_i O)_{xy} \\
\Rightarrow S &= O^\dagger \sum_i \hat{f}_i p_i O \\
&= \sum_i p_i O^\dagger \hat{f}_i O \\
&= \sum_i p_i T_i.
\end{aligned}$$

Besides, we claim

$$S^{-1} = \sum_i p_i^{-1} T_i.$$

proof:

$$\begin{aligned}
\mathbb{1} \overset{!}{=} SS^{-1} &= \sum_{ij} p_i p_j^{-1} T_i T_j \\
&= \sum_{ij} p_i p_j^{-1} O^\dagger \hat{f}_i \underbrace{O O^\dagger}_{=\mathbb{1}} \hat{f}_j O \\
&= \sum_{ij} p_i p_j^{-1} O^\dagger \underbrace{f_i f_j}_{=\delta_{ij}} O \\
&= \sum_i 1 O^\dagger \mathbb{1} O \\
&= \mathbb{1}
\end{aligned}$$

14.2 JOINT PDF

$$\mathcal{P}(d, s, S) = \underbrace{\mathcal{P}(d|s)}_{\text{likelihood}} \underbrace{\mathcal{P}(s|S)}_{\text{signal prior}} \underbrace{\mathcal{P}(S)}_{\text{spectral prior}}$$

LIKELIHOOD

$$\begin{aligned}
\mathcal{P}(d|s) &= \mathcal{G}(d - Rs, N) \\
\mathcal{H}(d|s) &\widehat{=} \frac{1}{2} s^\dagger \underbrace{R^\dagger N^{-1} R}_{=M} s - j^\dagger s
\end{aligned}$$

SIGNAL PRIOR

$$\begin{aligned} \mathcal{P}(s|S) &= \mathcal{G}(s, S) \\ \mathcal{H}(s|S) &= \frac{1}{2}s^\dagger S^{-1}s + \frac{1}{2}\ln|2\pi S| \end{aligned}$$

In this case, the normalization term $\frac{1}{2}\ln|2\pi S|$ can not be neglected, since S is unknown.

SPECTRAL PRIOR

I'= Spectral coefficients are positive but of unknown magnitude.
⇒Flat distribution on a logarithmic scale are estimated by the Jeffrey's prior.

$$\begin{aligned} \mathcal{P}(p_i) &\propto p_i^{-1} \\ \mathcal{P}(\tau_i) &\propto \text{const. with } \tau_i = \ln p_i \\ \mathcal{P}(p) &= \prod_i \mathcal{P}(p_i) \\ \Rightarrow \mathcal{P}(S) &= \prod_i p_i^{-1} \\ \mathcal{H}(S) &= \sum_i \ln p_i \end{aligned}$$

JOINT HAMILTONIAN

$$\begin{aligned} \mathcal{H}(d, s, S) &\widehat{=} \frac{1}{2}s^\dagger \underbrace{(S_p^{-1} + M)}_{=D_p^{-1}} s - j^\dagger s + \sum_i \ln p_i + \frac{1}{2}\ln|2\pi S_p| \\ &\widehat{=} \frac{1}{2}s^\dagger D_p^{-1}s - j^\dagger s + \sum_i \ln p_i^{(1+\rho_i/2)} \end{aligned}$$

We used $|S_p| = \prod_i p_i^{\rho_i}$ with $\rho_i = \mathrm{Tr}(T_i T_i^{-1})$ giving the number of degrees of freedom in the spectral band i.

14.3 EFFECTIVE HAMILTONIAN FROM MARGINALIZED JOINT PDF

$$\begin{aligned} \mathcal{P}(d, s) &= \int \mathcal{D}S\, \mathcal{P}(d, s, S) \\ &= \int \mathcal{D}S\, \mathcal{P}(d|s)\mathcal{P}(s|S)\mathcal{P}(S) \\ &= \mathcal{P}(d|s) \underbrace{\int \mathcal{D}S\, \mathcal{P}(s|S)\mathcal{P}(S)}_{=\mathcal{P}(s)=e^{-\mathcal{H}_{\mathrm{eff}}(s)}} \\ \mathcal{H}_{\mathrm{eff}}(s) &= -\ln \int \mathcal{D}p \, \exp\left(-\frac{1}{2}s^\dagger S_p^{-1}s - \sum_i (1 + \frac{\rho_i}{2})\ln p_i\right) \\ &= -\ln \prod_i \left[\int_0^\infty dp_i \, \exp\left(-\frac{p_i^{-1}}{2}s^\dagger T_i^{-1}s - (1 + \frac{\rho_i}{2})\ln p_i\right)\right] \\ &= -\ln \prod_i \left[\int_0^\infty dp_i \, p_i^{-(1+\frac{\rho_i}{2})} e^{-\frac{p_i^{-1}}{2}s^\dagger T_i s}\right] \end{aligned}$$

Using $t_i = \frac{1}{2}s^\dagger S^{-1}s$, $x_i = \frac{t_i}{p_i}$ and $dp_i = -\frac{t}{x_i^2}\,dx_i$ we get,

$$\begin{aligned}
\mathcal{H}_{\mathrm{eff}}(s) &= -\sum_i \ln \int_0^\infty dp_i\, p_i^{-(1+\frac{\rho_i}{2})} e^{-\frac{t_i}{p_i}} \\
&= -\sum_i \ln \left[t_i^{(1+\frac{\rho_i}{2})+1} \underbrace{\int_0^\infty dx_i\, x_i^{-2+1+\frac{\rho_i}{2}} e^{-x_i}}_{=\Gamma(\rho_i/2)} \right] \\
&= +\sum_i \frac{\rho_i}{2} \ln\left(\frac{1}{2} s^\dagger T_i^{-1} s\right)
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \mathcal{H}(d,s) &\,\widehat{=}\, \frac{1}{2}s^\dagger M s - j^\dagger s + \sum_i \frac{\rho_i}{2}\ln\left(\frac{1}{2}s^\dagger T_i^{-1}s\right) \\
M &= R^\dagger N^{-1} R \\
j &= R^\dagger N^{-1} d \\
\rho_i &= \#\text{ modes in spectral band } i
\end{aligned}$$

14.4 CLASSICAL OR MAP ESTIMATE

$$\begin{aligned}
\frac{\delta \mathcal{H}_{\mathrm{eff}}(d,s)}{\delta s} &= Ms - j + \sum_i \rho_i \frac{T_i s}{s^\dagger T_i s} \overset{!}{=} 0 \\
\Rightarrow j &= \underbrace{\left(M + \underbrace{\sum_i \underbrace{\frac{\rho_i}{s^\dagger T_i s}}_{=(p_i^*)^{-1}} T_i}_{=S_{p^*}^{-1}} \right)}_{=D_{p^*}^{-1}} s
\end{aligned}$$

$$s_{\mathrm{cl}} = m_{\mathrm{MAP}} = D_{p*} j$$

with $D_{p^*} = (S_{p^*}^{-1} + M)^{-1}$ and the power of the reconstructed map in spectral band i, $p_i^* = \frac{1}{\rho_i} s_{\mathrm{cl}}^\dagger T_i s_{\mathrm{cl}}$.

14.5 THERMODYNAMICAL APPROACH

In the following we will assume $T = 1$ for simplicity.

$$
\begin{aligned}
\Rightarrow \widetilde{G}\,(m, D) &= \widetilde{U}\,(m, D) - \widetilde{S_B}\,(D) \\
\widetilde{U}\,(m, D) &= \langle \mathcal{H}(d, s) \rangle_{\mathcal{G}(s-m, D)} \\
&= \frac{1}{2}\mathrm{Tr}(\langle ss^\dagger \rangle M) - j^\dagger \langle s \rangle + \sum_i \frac{\rho_i}{2} \underbrace{\langle \ln(s^\dagger T_i s) \rangle}_{=I_i} \\
&= \frac{1}{2}\mathrm{Tr}((mm^\dagger + D)M) - j^\dagger m + \sum_i \frac{\rho_i}{2} I_i
\end{aligned}
$$

Choose τ_i as the typical value for $s^\dagger T_i s$, around which we expand and the corresponding ansatz $\tau_i = \mathrm{Tr}((mm^\dagger + \delta D)T_i)$. In this case δ is a parameter to model the uncertainty dispersion corrections.

$$
\begin{aligned}
I_i &= \langle \ln \frac{s^\dagger T_i s}{\tau_i} + \ln \tau_i \rangle_{\mathcal{G}(s-m, D)} \\
&= \ln \tau_i + \langle \ln(1 + \frac{s^\dagger T_i s - \tau_i}{\tau_i}) \rangle_{\mathcal{G}(s-m, D)} \\
&= \ln \tau_i + \sum_{n=1}^{\infty} \frac{(-1)^n}{n \tau_i^n} \underbrace{\langle (s^\dagger T_i s - \tau_i)^n \rangle_{\mathcal{G}(s-m, D)}}_{=II_{i,n}} \\
II_{i,1} &= \mathrm{Tr}((mm^\dagger + D)T_i) - \tau_i \\
&= (1 - \delta)\mathrm{Tr}(DT_i) \\
II_{i,2} &= \ldots = II_{i,1}^2 + 4\mathrm{Tr}((mm^\dagger + \frac{1}{2}D)T_i D T_i)
\end{aligned}
$$

If we want $II_{i,1} = 0$ and $II_{i,2}$ to be minimal, then we choose $\delta = 1$. For this special case we calculate the Gibbs free energy,

$$
\begin{aligned}
G(\widetilde{m}, D) &= \frac{1}{2}\mathrm{Tr}((mm^\dagger + D)M) - j^\dagger m + \sum_i \frac{\rho_i}{2} \ln \left[\mathrm{Tr}((mm^\dagger + D)T_i) \right] \\
&\quad - \frac{1}{2}\mathrm{Tr}(1 + \ln(2\pi D)).
\end{aligned}
$$

Minimize the Gibbs free energy, in order to obtain the mean field m,

$$
\begin{aligned}
0 &\stackrel{!}{=} \frac{\delta\, \widetilde{G}}{\delta m} = Mm - j + \sum_i \frac{\rho_i}{2} \frac{2T_i m}{\mathrm{Tr}((mm^\dagger + D)T_i)} \\
\Rightarrow j &= \underbrace{(M + \sum_i (p_i^*)^{-1} T_i)}_{=D_{p^*}^{-1}}\, m \\
p_i^* &= \frac{\mathrm{Tr}((mm^\dagger + D)T_i)}{\rho_i}
\end{aligned}
$$

Unified filter formula:

$$\begin{aligned}
m &= D_p j \\
p_i &= \frac{1}{\rho_i} \mathrm{Tr}((mm^\dagger + \delta D) S_i^{-1}) \\
\delta &= 0 \ \Rightarrow \text{MAP filter} \\
\delta &= 1 \ \Rightarrow \text{critical filter}
\end{aligned}$$

15 DYNAMICAL FIELD INFERENCE

15.1 HOLISTIC PICTURE

15.1.1 *Field prior*

A dynamic field $\varphi = \varphi(t) = \varphi(x, t)$ varies in space x and time t.
Background information I:

$$\partial_t \varphi(t) = F[\varphi(t)] + \xi(t). \tag{367}$$

F possibly non-linear, possibly non-local, equal time (integro-)differential operator.
ξ a noise field summarizing uncontrolled environmental influences.
Field prior:

$$\mathcal{P}(\varphi|I) = \int \mathcal{D}\xi \, \mathcal{P}(\varphi|\xi, I)\, \mathcal{P}(\xi|I). \tag{368}$$

Field is fully determined by noise ξ and initial conditions $\varphi(0) = \varphi_0$:

$$\begin{aligned} \mathcal{P}(\varphi|\xi, I) &= \prod_{x,t} \delta \left\{ \varphi(t) - \varphi_0 - \int_0^t dt' \, \left[F[\varphi(t')] + \xi(t') \right] \right\} \\ &= \delta\left\{\dot{\varphi} - F[\varphi] - \xi\right\} \, |\partial_t - \partial_\varphi F(\varphi)|, \end{aligned} \tag{369}$$

with $\delta(\psi) = \prod_{x,t} \delta[\psi(x,t)]$ a functional delta function and $|\partial_t - \partial_\varphi F(\varphi)|$ the functional determinant of the stochastic differential equation (367).

$$\mathcal{P}(\varphi|I) = |\partial_t - \partial_\varphi F(\varphi)|\, \delta\left\{\varphi(t) - \varphi_0\right\}\, \mathcal{P}(\xi = \partial_t \varphi - F[\varphi]|I) \tag{370}$$

15.1.2 *Field posterior*

Data $d \hookleftarrow \mathcal{P}(d|\varphi)$ as resulting form field measurements, as initial field configuration of a simulation, as the data representing a simulation step, or a combination of these possibilities.
Field posterior:

$$\mathcal{P}(\varphi|d,\, I) = \frac{\mathcal{P}(d|\varphi,\, I)\, \mathcal{P}(\varphi|I)}{\mathcal{P}(d|I)}$$

Information energy:

$$\mathcal{H}(d, \varphi) = -\ln \mathcal{P}(d, \varphi|I) \tag{371}$$

15.1.3 *Partition function*

Moment generating partition function:

$$\begin{aligned} \mathcal{Z}(d,J) &= \int \mathcal{D}\varphi \, e^{-\mathcal{H}(d,\varphi|I)+J^\dagger\varphi} \\ &= \int \mathcal{D}\varphi \, |\partial_t - \partial_\varphi F(\varphi)| \, \mathcal{P}(\xi = \partial_t\varphi - F[\varphi]|I) \, e^{-\mathcal{H}(d|\varphi,I)+J^\dagger\varphi} \\ J^\dagger\varphi &= \int dx \int dt \, J^*(x,t)\,\varphi(x,t) \end{aligned} \tag{372}$$

Assume Gaussianity and linearity of measurement and driving noises: $\mathcal{P}(d|\varphi,I) = \mathcal{G}(d - R\,\varphi, N)$ and $\mathcal{P}(\xi|I) = \mathcal{G}(\xi,\Xi)$.

$$\begin{aligned} \mathcal{Z}(d,J) &= \int \mathcal{D}\varphi \, |\partial_t - \partial_\varphi F(\varphi)| \, \mathcal{G}(\partial_t\varphi - F[\varphi],\Xi) \, \mathcal{G}(d - R\varphi, N) \, e^{J^\dagger\varphi} \\ &= \int \mathcal{D}\varphi \, \frac{|\partial_t - \partial_\varphi F(\varphi)|}{|2\pi\Xi|^{1/2}|2\pi N|^{1/2}} \, e^{-\frac{1}{2}\left\{(\partial_t\varphi - F[\varphi])^\dagger\Xi^{-1}(\partial_t\varphi-F[\varphi]) + (d-R\,\varphi)^\dagger N^{-1}(d-R\,\varphi) - J^\dagger\varphi\right\}} \end{aligned}$$

15.1.4 *Linear dynamics*

Special case $F[\varphi] = F\,\varphi$, then

$$\partial_t\varphi - F[\varphi] = \underbrace{(\partial_t - F)}_{\equiv G^{-1}}\,\varphi = \xi$$

with $G = (\partial_t - F)^{-1}$ Greens function of process, such that $\varphi = G\,\xi$.
If F stationary, temporal Fourier transformation yields

$$\begin{aligned} \left(G^{-1}\right)_{\omega\omega'} &= 2\pi\delta(\omega-\omega')\,(i\omega - F) \\ G_{\omega\omega'} &= 2\pi\delta(\omega-\omega')\,(i\omega - F)^{-1} \end{aligned}$$

Partition function:

$$\begin{aligned} \mathcal{Z}(d,J) &= \frac{|G^{-1}|}{|2\pi\Xi|^{1/2}|2\pi N|^{1/2}} \int \mathcal{D}\varphi \; e^{-\frac{1}{2}\left\{\left(G^{-1}\varphi\right)^\dagger\Xi^{-1}G\,\varphi + (d-R\,\varphi)^\dagger N^{-1}(d-R\,\varphi) - j^\dagger\varphi\right\}} \\ &= \frac{|G^{-1}|\,|2\pi\,D|^{1/2}}{|2\pi\Xi|^{1/2}|2\pi N|^{1/2}} \, e^{\frac{1}{2}(J+j)^\dagger D(J+j) - \frac{1}{2}d^\dagger N^{-1}d} \\ j &= R^\dagger N^{-1} d \\ D &= \left[\Phi^{-1} + R^\dagger N^{-1} R\right]^{-1} \\ \Phi &= G\Xi G^\dagger \end{aligned}$$

Field expectations:

$$\begin{aligned} \langle\varphi\rangle_{(\varphi|d,I)} &= \left.\frac{\partial \ln \mathcal{Z}}{\partial J}\right|_{J=0} = D\,j \\ \langle\varphi\varphi^\dagger\rangle^{c}_{(\varphi|d,I)} &= \left.\frac{\partial^2 \ln \mathcal{Z}}{\partial J \partial J^\dagger}\right|_{J=0} = D \end{aligned}$$

Example: Diffusion equation, $F = \Delta$,
Greens function: $G_{(x,t)\,(x',t')} = \theta(t-t')\,\mathcal{G}(x-x', 2\,(t-t'))$
Test:

$$\begin{aligned}
(\partial_t - \Delta)_{(x,t)} G_{(x,t)\,(x',t')} &= \delta(t-t')\,\mathcal{G}(x-x',0) + \theta(t-t')\,\mathcal{G}(x-x',2\,(t-t'))\ \times 0 \\
&= \delta(t-t')\delta(x-x') = \mathbb{1}_{(x,t)\,(x',t')} \\
\varphi &= G\,\xi \\
(\partial_t - \Delta)\,\varphi &= \xi
\end{aligned}$$

In Fourier space:

$$\begin{aligned}
G_{(k,\omega)\,(k',\omega')} &= \frac{(2\pi)^{1+u}\delta(k-k')\,\delta(\omega-\omega')}{i\omega + k^2} \\
P_\varphi(k,\omega) &= \frac{P_\xi(k,\omega)}{\omega^2 + k^4}
\end{aligned}$$

15.1.5 *Noise free case*

$\partial_t \varphi(t) = F[\varphi(t)]$ and no data.

$$\mathcal{Z}(J) = \int \mathcal{D}\varphi\,|\partial_t - \partial_\varphi F(\varphi)|\,\delta\,\{\partial_t\varphi - F[\varphi]\}\ e^{J^\dagger\varphi} \tag{373}$$

Two obstacles: functional determinant and functional delta function.
Solution: introduce auxiliary fields

BOSONIC FIELD:

$$\delta\,\{\partial_t\varphi - F[\varphi]\} \propto \int \mathcal{D}\eta\, e^{i\,\eta^\dagger(\partial_t\varphi - F[\varphi])}$$

FERMIONIC FIELD:

$$\begin{aligned}
|\partial_t - \partial_\varphi F(\varphi)| &= \int \mathcal{D}\overline{\chi}\mathcal{D}\chi\, e^{\chi^\dagger(\partial_t - \partial_\varphi F(\varphi))\,\overline{\chi}} \qquad (374)\\
\chi, \overline{\chi} &\quad \text{fields of Grassmann variables}
\end{aligned}$$

Grassmann numbers:

$$\begin{aligned}
\chi, \overline{\chi} &\quad \text{two (scalar) Grassmann numbers/variables} \\
\chi\,\overline{\chi} &= -\overline{\chi}\chi, \text{ anticommuting numbers} \qquad (375)\\
\chi\,\chi &= \overline{\chi}\,\overline{\chi} = 0 \\
a\,\chi &= \chi\,a,\ \text{commutes with } a \in \mathbb{C} \qquad (376)\\
e^\chi &= 1 + \chi \\
\int d\chi\,1 &\equiv 0 \qquad (377)\\
\int d\chi\,\chi &\equiv 1 \qquad (378)\\
\partial_\chi \chi &\equiv 1, \text{ differentiation = integration} \qquad (379)\\
\int d\chi\, e^{a\chi} &= \int d\chi\,(1 + a\chi) = 0 + a = a
\end{aligned}$$

Grassmann numbers can be represented by matrices. E.g.

$$\theta_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \theta_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}$$

are Grassmann numbers as $\theta_1\theta_2 = -\theta_2\theta_1$, $\theta_1\theta_1 = 0$, and $\theta_2\theta_2 = 0$.

Determinats via Grassmann integrals:

$$\begin{aligned}
A & \quad \text{matrix of rank } m \\
\chi &= (\chi_1, \dots \chi_m)^\dagger \text{ vector of Grasmann variables} \\
\overline{\chi} &= (\overline{\chi}_1, \dots \overline{\chi}_m)^\dagger \\
d\chi &\equiv d\chi_1 \cdots d\chi_m \\
\int d\overline{\chi}\, d\chi\, e^{\chi^\dagger A \overline{\chi}} &= \int d\overline{\chi}\, d\chi\, e^{\sum_{ij=1}^m \chi_i^\dagger A_{ij} \overline{\chi_j}} \\
&= \int d\overline{\chi}\, d\chi\, \sum_{n=0}^{\infty} \frac{1}{n!} \left(\sum_{ij} \chi_i A_{ij} \overline{\chi_j} \right)^n \\
&= \int d\overline{\chi}\, d\chi \left[1 + \underbrace{\sum_{ij} \chi_i A_{ij} \overline{\chi_j} + \dots}_{\to 0 \text{ since } < m \text{ terms}} + \underbrace{\frac{1}{m!} \left(\sum_{ij} \chi_i A_{ij} \overline{\chi_j} \right)^m}_{\neq 0 \text{ since } m \text{ terms}} + \underbrace{\dots}_{=0} \right] \\
&= \frac{1}{m!} \int d\overline{\chi}\, d\chi \sum_{i_1 j_1} \dots \sum_{i_m j_m} \chi_{i_1} \overline{\chi_{j_1}} \cdots \chi_{i_m} \overline{\chi_{j_m}}\; A_{i_1 j_1} \dots A_{i_m j_m} \\
&= \sum_{\sigma \in S_n} \operatorname{sgn}\sigma \prod_{i=1}^m A_{i\sigma_i} \text{ with } \sigma \text{ permutation, } S_n \text{ Symmetric group} \\
&= |A|
\end{aligned} \tag{380}$$

by sign flips during reordering of $\chi_{i_1}\overline{\chi_{j_1}} \cdots \chi_{i_m}\overline{\chi_{j_m}}$ to match the (inverse) order of integration variables $d\overline{\chi_1} \cdots d\overline{\chi_m} d\chi_1 \cdots d\chi_m$ so that Eq. (378) can be used.

Dynamical system partition function:

$$\mathcal{Z}(J) \propto \int \mathcal{D}\varphi \mathcal{D}\eta \mathcal{D}\overline{\chi} \mathcal{D}\chi\, e^{\chi^\dagger (\partial_t - \partial_\varphi F(\varphi))\, \overline{\chi} + i\, \eta^\dagger (\partial_t \varphi - F[\varphi]) + J^\dagger \varphi} \tag{381}$$

can be regarded as the partition function of a super-symmetric field theory. There is an exchange symmetry between the bosonic (φ, η) and fermionic $(\chi, \overline{\chi})$ degrees of freedom, since both involve the same operator $\partial_t - \partial_\varphi F(\varphi)$. Details of this Parisi-Sourlas-Wu quantization are beyond the level of this course.

15.2 INFORMATION FIELD DYNAMICS

15.2.1 *Basic idea*

Computer simulations address the inference problem what is the future of a field given some initial data and a dynamical law. Fields are represented by finite data

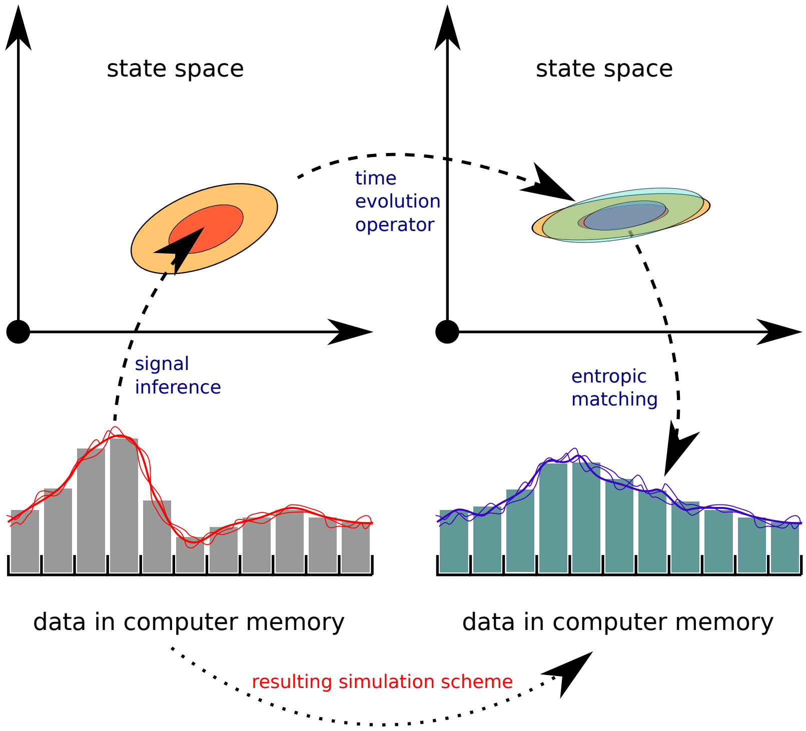


Figure 12: Illustration of the IFD concepts.

vectors, which inevitable implies information loss on sub-grid field structures. How should the differential operators of the dynamics be best represented numerically to have most accurate simulation?

Information optimal simulation schemes should be possible. Information field dyamics (IFD, [5]) is a proposal how to construct such optimal schemes. The basic idea is sketched in Fig. 12 and consists of the following steps:

1. The field to be simulated, φ, is regarded to be unknown.

2. Known is the data d in the computers memory, which is regarded to be the result of a measurement process at time t, e.g. $d = R\,\varphi + n$, with know response R, and covariances N and Φ of the noise n and the field φ, respectively.

3. This permits virtually a reconstruction of the posterior mean $m = \langle\varphi\rangle_{(\varphi|d)}$ and its uncertainty dispersion $D = \langle\varphi\,\varphi^{\dagger}\rangle_{(\varphi|d)}$, and most importantly the construction of a field posterior $\mathcal{P}(\varphi|d)$, e.g. $\mathcal{G}(\varphi - m, D)$ in case of Gaussianity and linearity.

4. The posterior $\mathcal{P}(\varphi|d)$ is probability distribution over the state space of the field. Each point in the state space represents a possible continuous field configuration. Each should evolve according to the dynamical law of the field, e.g.

$$\partial_t\varphi = F[\varphi],$$

and therefore can be time evolved to an infinitesimal future $t' = t + \delta t$ via

$$\varphi' \equiv \varphi_{t'} = \varphi + \delta t\,F[\varphi] + \mathcal{O}(\delta t^2).$$

5. Thus, the full posterior can be time evolved as well

$$\begin{aligned}\mathcal{P}(\varphi'|d) &= \mathcal{P}(\varphi|d) \left|\frac{\partial\varphi}{\partial\varphi'}\right|\Bigg|_{\varphi'=\varphi+\delta t\, F[\varphi]} \\ &= \mathcal{P}(\varphi|d) \left|\mathbb{1} - \delta t\, \frac{\partial F[\varphi]}{\partial\varphi}\right|\Bigg|_{\varphi=\varphi'-\delta t\, F[\varphi']} + \mathcal{O}(\delta t^2).\end{aligned}$$

6. Now, new data d' in computer memory has to be chosen to represent $\mathcal{P} \equiv \mathcal{P}(\varphi'|d)$ as closely as possible. If measurement process is specified, new posterior $\mathcal{P}' \equiv \mathcal{P}'(\varphi'|d')$ can be matched entropically by maximizing

$$\mathcal{S}_{\mathrm{B}}(\mathcal{P}'|\mathcal{P}) = -\int \mathcal{D}\varphi'\, \mathcal{P}'(\varphi'|d')\, \ln \frac{\mathcal{P}'(\varphi'|d')}{\mathcal{P}(\varphi'|d)}$$

with respect to d' (and if needed other parameters like R, N, Φ, ...). The resulting formula will be of the form

$$d' = \mathcal{F}[d]$$

and therefore represent a simulation scheme.

An IFD simulation scheme therefore incorporates all knowledge of the sub-grid statistics (as encoded in Φ), the relation between field φ and data d (as encoded in R and N) and the precise partial differential equation of the field evolution (as encoded in F) and tries to find a future data set d' that codes all this information optimally.

15.2.2 *Ensemble dynamics of stochastic systems*

Following [10].
$\varphi = (\varphi_1, \ldots \varphi_n)^\dagger$ finite dimensional state vector of stochastic system evolving according to

$$\partial_t\varphi = F(\varphi) + \xi$$

with white noise vectors $\xi_t \hookleftarrow \mathcal{G}(\xi_t, \Xi)$ with covariance $\langle \xi_t \xi^\dagger_{t'} \rangle_{(\xi)} = \delta(t-t')\,\Xi$.
What is the evolution of an ensemble of such systems?
Gaussian Ansatz:

$$\mathcal{P}(\varphi|t) = \mathcal{G}(\varphi - m_t, \Phi_t)$$

Follow evolution of $m_t \in \mathbb{R}^n$ and $\Phi_t \in \mathbb{R}^{n\times n}$.
Linear noise approximation:

$$\begin{aligned}\partial_t m_t &= F(m_t) \\ \partial_t\Phi_t &= \left[\frac{\partial F(m_t)}{\partial m_t}\right] \Phi_t + \Phi_t \left[\frac{\partial F(m_t)}{\partial m_t}\right]^\dagger + \Xi\end{aligned}$$

Evolution of mean m_t does not depend on Φ_t.
Entropic matching:

$$\begin{aligned}\partial_t m_t &= \langle F(\varphi)\rangle_{\mathcal{G}(\varphi-m_t,\Phi_t)} \\ \partial_t\Phi_t &= \left\langle \frac{\partial F(\varphi)}{\partial\varphi}\right\rangle_{\mathcal{G}(\varphi-m_t,\Phi_t)} \Phi_t + \Phi_t \left\langle \frac{\partial F(\varphi)}{\partial\varphi}\right\rangle^\dagger_{\mathcal{G}(\varphi-m_t,\Phi_t)} + \Xi\end{aligned}$$

Gaussian averaging couples m_t and Φ_t mutually.
Numerical experiments show that entropic matching scheme performs better than linear noise approximation in case of non-linear stochastic systems.

BIBLIOGRAPHY

[1] J.J. Binney, N.J. Dowrick, A.J. Fisher, and M.E.J. Newman, *The theory of critical phenomena*, Oxford University Press, Oxford, UK: ISBN0-19-851394-1, 1992. (Cited on page 105.)

[2] A. Caticha, *Lectures on Probability, Entropy, and Statistical Physics*, ArXiv e-prints (2008). (Cited on page 7.)

[3] R. T. Cox, *Probability, Frequency and Reasonable Expectation*, American Journal of Physics **14** (1946), 1–13. (Cited on pages 7, 9, and 10.)

[4] T. Enßlin, *Information field theory*, American Institute of Physics Conference Series (U. von Toussaint, ed.), American Institute of Physics Conference Series, vol. 1553, August 2013, pp. 184–191. (Cited on page 103.)

[5] T. A. Enßlin, *Information field dynamics for simulation scheme construction*, Phys. Rev. E**87** (2013), no. 1, 013308. (Cited on page 133.)

[6] T. A. Enßlin, M. Frommert, and F. S. Kitaura, *Information field theory for cosmological perturbation reconstruction and nonlinear signal analysis*, Phys. Rev. D**80** (2009), no. 10, 105005. (Cited on page 103.)

[7] E. T. Jaynes, *Probability Theory: The Logic of Science*, June 2003. (Cited on page 7.)

[8] R. H. Leike and T. A. Enßlin, *Operator Calculus for Information Field Theory*, ArXiv e-prints: arXiv160500660 (2016). (Cited on page 118.)

[9] ———, *Optimal Belief Approximation*, ArXiv e-prints (2016). (Cited on page 28.)

[10] T. Ramalho, M. Selig, U. Gerland, and T. A. Enßlin, *Simulation of stochastic network dynamics via entropic matching*, Phys. Rev. E**87** (2013), no. 2, 022719. (Cited on page 134.)

[11] A. Terenin and D. Draper, *Cox's Theorem and the Jaynesian Interpretation of Probability*, ArXiv e-prints (2015). (Cited on page 7.)