SYSTEMS THEORY

A Retrospective and Prospective Look

Sanjoy K. Mitter

Laboratory for Information and Decision Systems

Massachusetts Institute of Technology

July 1, 2013

IMT, Italy

Agenda of Systems Theory

- Models and their Structure
- Fundamental Limitations (Laws)
- Uncertainty and Robustness

  Robustness of performance uncertainty at different levels of granularity

- Interconnections, Architecture and Algorithms

  Architecture = organization of distributed algorithms and their implementation in hardware

Agenda of Systems Theory (cont.)

- Resource Management (Energy, Time, Space, . . .)

  A broad vision of Systems Theory aids in providing a unified conceptual framework for problems in different fields (Control, Communication, Signal Processing, Operations Research)

- Structure
- Action
- and their Interaction

History of Science in the Sense of Kuhn:

Incommensurability

Thomas Kuhn in his book *The Structure of Scientific Revolutions* distinguished between Normal Science and Revolutionary Science.

Revolutionary Science (e.g., Quantum Mechanics) arises when:

- Existing Theories fail to explain phenomena
- A new "paradigm" is needed to reconcile theory and experiment
- With the new paradigm, a new language is needed

Something like that happened in the late fifties and early sixties in the Systems and Control field.

Earlier revolution (1948):
Shannon Information Theory and
Invention of the Transistor

"The Double Big Bang," to quote Viterbi

I want to suggest that in the Systems and Control field, there was a crisis in the field in the fifties. Let me suggest as pointers three manifestations of that crises.

1. Internal Stability: Feedback Control Systems designed from an external (input/output) point of view failed to recognize the presence of these internal instabilities.

2. The approach to design of multi-input/multi-output systems was essentially a reduction to a single-input/single-output system through a decoupling procedure.

3. The attempts to deal with the Wiener filtering problem in the nonstationary situation (Zadeh–Regazzini) leading to some analog of the Wiener–Hopf equation was not very successful (no procedure analogous to Spectral Factorization was available).

It is also worth mentioning that the Mathematics that was prevalent in Linear Systems Theory at the time was Complex Function Theory and Transform Theory.

New Element

Computation and the Concept of a Solution

Solution not necessarily an analytical expression

Theories leading to Algorithms

Advent of State Space Theory

(New Paradigm)

- New Language: Algebra, Differential Equations
- Concept of State
- State Space Representation=

$$\begin{cases} \frac{\mathrm{d}x}{\mathrm{d}t} = Fx(t) + Gu(t) \\ y(t) = Hx(t) \end{cases}$$

$u =$ input, $x =$ state, $y =$ output

Extends to time-varying and nonlinear systems

Advent of State Space Theory

(New Paradigm cont.)

$$y(t) = He^{(t-t_0)F}x(t_0) + \int_{t_0}^{t} He^{(t-s)F}Gu(s)\mathrm{d}s$$

Reconciliation of Input-Output and Internal (State) Point-of-view through introduction of concepts of reachability (controllability) and observability

Natural Connection to Stability and Optimality (Calculus of Variations)

Minimize

$$J(u,x) = \int_{t_0}^{t_1} [(x(t), Qx(t)) + (u(t), Ru(t))] \mathrm{d}t$$

$$Q \geq 0 \quad , \quad R > 0$$

Behavior of optimal control

$$u(t) = K(t)x(t) \quad \text{as} \quad t_1 \to \infty$$

Role of Controllability and Observability

Deeper Aspects of Structure

Actions of semi-direct product

$$GL(n) \times \mathcal{F} \times GL(m)$$

on (F, G) controllable

$$(F, G) \mapsto (T^{-1}(F + GK)T, GL)$$

Kronecker Invariants

Transporting the algebraic variety structure of (F, G) to the quotient

Implications in System Identification

How should we think about Graphs beyond thinking about them as (V, E)?

How should we think about Systems of Coupled Differential Equations evolving over Graphs?

What are these invariants?

We should be able to distinguish between differential equations evolving over trees from differential equations evolving over graphs with loops

We need Canonical Problems

Pattern Recognition (Vision)

"Tranformation Group" acting on the space of objects is not given but needs to be identified!!

See the section on Pattern Recognition in Minsky's paper:

"Steps Towards Artificial Intelligence," *Proc. IEEE*, 1961.

Influence of Systems Theory in Coding Theory and Signal Processing

(Intersection with Behavioral View of Systems: Willems)

Linear Systems taking values in Finite Groups (Forney–Trott)

Minimality, Controllability and Observality, Duality in Signal Processing

State Space Viewpoint: Influence on Algorithms exploiting structure

Adaptive Filtering

Filtering and Stochastic Control:

Separation Principle

$$\begin{cases} \mathrm{d}X(t) &= FX(t)\mathrm{d}t + Gu(t) + J\mathrm{d}W(t) \\ \mathrm{d}Y(t) &= Hx(t)\mathrm{d} + \mathrm{d}V(t) \end{cases}$$

Choose $u(t) = \varphi(\Pi_t Y)$ to minimize

$$J(u,x) = \mathbb{E}\left[\int_{t_0}^{t_1} [(X(t), QX(t)) + (u(t), Ru(t))]\mathrm{d}t\right]$$

Solution

$$u^*(t) = K(t)\hat{X}(t)$$
$$\hat{X}(t) = \mathbb{E}(X(t)|\mathcal{F}_t^Y)$$

Separation into estimation and deterministic control

- Infinite-time

(Controllability, Observability, Stability)

- Non-linear

Smoothing (Decoding)

Compute: $\mathbb{P}(X_s, t_0 \leq s \leq t_1)|\mathcal{F}_{t_1}^Y)$

Uncertainty and Robustness

Process and Measurement Uncertainty

vs.

Model Uncertainty

Approximation of Input-Output Maps

vs.

Approximation at the State Space Representation

Two input-output maps may be close to each other but the dimensions of their corresponding state spaces may be far apart

(See: "The Legacy of George Zames,"

Mitter and Tannenbaum,

IEEE Trans. on Auto. Control)

Fundamental Problem of Control: Design of Control Systems whose performance is robust against uncertainties

For linear time-invariant, bounded, causal maps from $L^2(\mathbb{R}) \to L^2(\mathbb{R})$, which, from the Segal–Foures theorem, is in one-to-one correspondence with operators which are multiplication operators by H^∞-functions

Uncertainty in model represented by a ball in H^∞

Feedback: reduction of complexity

Deep connections to Operator Theory, in particular the work of Krein

Recent work of Y.H. Kim:

Feedback Capacity of Stationary Gaussian Channels

The computation of feedback capacity is posed as an Infinite Dimensional Variational Problem and uses Systems Theory for its solution

Interestingly, Keynes viewed the representation of "uncertainty" and how to deal with uncertainty as one of the fundamental problems of Macroeconomics

He also questioned the use of probability for certain uncertain situations (prospect of a European war is uncertain, the price of copper, rate of interest twenty years hence)

Indeed, for systems which are distributed, modeling and representation of uncertainty remains a fundamental issue

Bayesian Inference
and
Statistical Mechanics

Some Connections between Information Theory, Filtering and Statistical Mechanics

Variational Approach to Bayesian Estimation

Stochastic Control Interpretation of Nonlinear Filtering

Preliminaries

X, Y discrete random variables with joint distribution P_{XY} and marginals P_X and P_Y

$$I(X;Y) = E_{P_{XY}}\left(\log \frac{P_{XY}}{P_X \otimes P_Y}\right) \quad : \quad \text{Mutual Information}$$

Average measure of dependence of two random variables

Mutual Information is an example of the general notion of relative entropy between two measures μ and ν on some probability space $(\Omega, \mathcal{F}, P)$ (discrete for the moment)

$$h(\mu|\nu) = E_\mu \log\left(\frac{\mu}{\nu}\right)$$

Properties:

(i) $h(\mu|\nu) \geq 0$

(ii) $h(\mu|\nu) = 0 \Leftrightarrow \mu = \nu$

(iii) $h(\mu|\nu)$ jointly convex in μ, ν

(But, not symmetric). Defines a pseudo-distance between two measures μ and ν.

We will have to deal with random variables in a more general setting.

Nonlinear Dynamical Systems
forced by (scaled) white noise

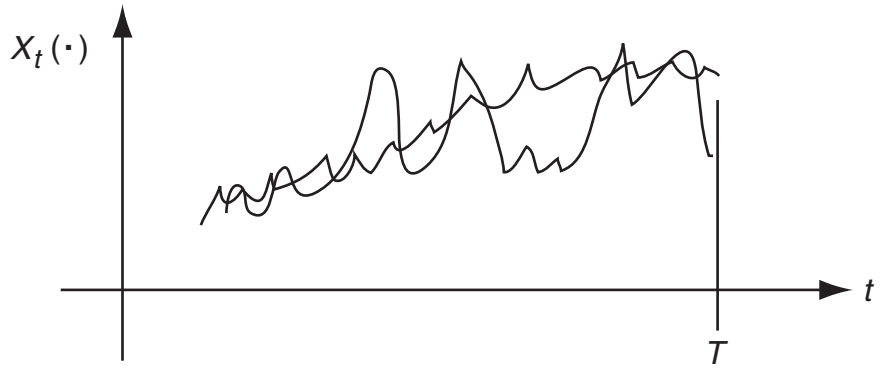
$$\frac{dx_t}{dt} = b(x_t) + \sigma(x_t)\dot{v}_t \ ,$$

where v_t: Brownian motion and $\dot{v}_t =$ white noise, formal derivative of Brownian motion

Rewrite as Integral equation

$$\begin{aligned} x_t &= x_0 + \int_0^t b(x_s)ds + \int_0^t \sigma(x_t)\dot{v}_t dt \\ &= x_0 + \int_0^t b(x_s)ds + \int_0^t \sigma(x_t)dv_t \ \leftarrow \text{Ito integral} \end{aligned}$$

We want to think of $x_{(\cdot)} := X$ as a map (random variable) from $(\Omega, \mathcal{F}, P)$ to $(\mathcal{X}, \mathcal{B}(\mathcal{X})$ where $\mathcal{X} = \mathcal{C}(0, \mathcal{T}; \mathbb{R})$ and $\mathcal{B}(\mathcal{X})$ is the Borel field associated with $\mathcal{X}$. We call the probability measure of $X \in \mathcal{P}(\mathcal{X})$ the path space measure



X is a random trajectory

Sometimes, we would want to look at these random trajectories "through" a different measure $\widehat{P}$ (instead of P) in order for it to "appear" differently, for example, trajectories of Brownian Motion.

Gibbs Measures:

Variational Characterization for Finite Systems

(H.O. Georgii: *Gibbs Measures and Phase Transitions*, Chapter 15)

Let $S =$ finite set, and $E =$ state space, finite set and let $\Omega = E^S$, finite.

Let Φ be any potential, and $H = \sum_{A \subset S} \Phi_A(w)$ be the associated Hamiltonian

The unique Gibbs measure for Φ is given by

$$\nu(\omega) \;=\; Z^{-1} \exp[-H(\omega)] \;,\; \omega \in \Omega$$

where

$$Z \;=\; \sum_{\omega \in \Omega} \exp[-H(\omega)] \;:\; \text{Partition function}$$

For each probability measure μ on Ω,

$$\mu(H) = \sum_{\omega \in \Omega} \mu(\omega) H(\omega) \text{ and } h(\mu) = -\sum_{\omega \in \Omega} \mu(\omega) \log \mu(\omega)$$

be the Energy and Entropy associated with μ

Then

$$\mu(H) - h(\mu) + \log Z = h(\mu|\nu) \geq 0$$

$$h(\mu|\nu) = 0 \Leftrightarrow \mu = \nu \qquad \square$$

$$F(\mu) = \mu(H) - h(\mu) \ : \ \text{Free Energy}$$

$$F(\nu) = -\log Z$$

Generalization of these ideas to infinite systems leads to characterization of translation-invariant Gibbs measures as minimization of Specific Free Energy. A modification of these ideas (using Exchangeability) leads to a proof of the Noisy Channel Coding Theorem (BSC).

Variational Bayes and a Problem of Reliable Communication, Part II,

N. Newton, S.K. Mitter, to appear in *J. Stat. Mech.*, 2012

$(X_t)_{t\geq 0}$ Markov Process, time homogeneous

$$P(X_t \in B | X_r, r \in [0,s]) = \pi(t-s, X_s, B) \quad 0 \leq s \leq t \leq T$$

P_t is the distribution of X_t with density p_t

$$P_t(B) = P(X_t \in B) = \int_B p_t(x)\lambda_x(dx) \quad \lambda_x : \text{ref. measure}$$

Diffusion

$$\begin{aligned}(Ap)(x) &= \frac{1}{2}\sum_{i,j}\frac{\partial^2(a_{i,j}p)}{\partial x_i \partial x_j}(x) - \sum_i \frac{\partial}{\partial x_i}(b_i p)(x) \text{ on } \mathbb{R}^d \\ X_t &= X_0 + \int_0^t b(X_s)dt + \int_0^t \sigma(X_s)dv_s \\ a &= \sigma\sigma'\end{aligned}$$

Relative Entropy

$$h(\mu|\lambda) = \int_X q(x) \log q(x) \lambda(dx) \quad \mu \text{ has density } q \text{ w.r.t. } \lambda$$
$$= +\infty \quad \text{otherwise}$$
$$\langle f, \lambda \rangle = \int_X f(x) \lambda(dx)$$

Σ_x: statistical mechanics system, associated with $(X_t)_{t \geq 0}$

P_t: state of Σ_x at time t

P_{SS}: unique invariant measure with density p_{SS}

Internal Energy $\mathcal{E}_X(P_t) = \langle H_x, P_t \rangle$

Entropy $S_x(P_t) = -h(P_t|\lambda_x)$

Free Energy $\mathcal{F}_X(P_t) = \mathcal{E}_x(P_t) - S_x(P_t)$

Energy Function $H_x(x) = -\log p_{SS}(x)$

Choice assures Energy Function is a Gibbs measure for Σ_x

Proposition:

(i) Unique minimizer of Free Energy of Σ_x is P_{SS}

(ii) $\mathcal{F}_x(P_{SS}) = 0$

(iii) Free Energy of Σ_x is non-increasing

Proof.

$$\mathcal{F}(x)(P_t) = h(P_t|P_{SS}) \Rightarrow \text{(i) and (ii)}$$

To prove (iii), $P^{(2)}_{s,t}$ = two point joint distribution

$$P^{(2)}_{s,t}(B,C) = P(X_s \in B, X_t \in C) = \int_B \pi(t-s, X, C) P_s(dx)$$

$P^{(2)}_{s,t,SS}$ = joint distribution when $P_s = P_{SS}$

Chain rule for Relative Entropy $\square$

$$
\begin{aligned}
& h(P^{(2)}_{s,t}|P^{(2)}_{s,t,SS}) \\
&\quad = \; h(P_t|P_{SS}) + \int h(\tilde{\Pi}(t,s,x,\,\cdot\,)|\tilde{\Pi}_{SS}(t-s,x,\,\cdot\,))P_t(dx) \\
&\quad \geq \; h(P_t|P_{SS})
\end{aligned}
$$

(Chain Rule)

where $\tilde{\Pi}(t,s,x,\,\cdot\,) =$ regular $(X_t = x)$-conditional distribution for X_s under the joint distribution $P^{(2)}_{s,t}$ and $\tilde{\Pi}_{SS}(t-s,x,\,\cdot\,)$ is the equivalent under the joint distribution $P^{(2)}_{s,t,SS}$.

Σ_x: one component of a two-component energy conserving system that includes a unit temperature heat bath with which Σ_x interacts

If Entropy of system = Entropy of the sum of two components then any change in this entropy resulting from the evolution of P_t = neg. of corresponding change in $\mathcal{F}_x(P_t)$

P_{SS}: unique invariant measure with density p_{SS}

Proposition: Entropy of closed system is maximized by P_{SS} and non-decreasing

Assertion (iii) in Proposition can be thought of as a Second Law of Thermodynamics for Σ_x

Observations (Interaction with Measurements)

$$Y_t = \int_0^t g(X_s)ds + W_t$$

$$E\left[\int_0^t |g(X_t)|^2 dt < \infty\right.$$

$(Z_t|t \in [0,T])$: regular conditional probability of X_t given $(Y_s|0 \leq s \leq t)$

ξ_t: density

$$\xi_t(x) = \xi_0(x) + \int_0^t (\mathcal{A}\xi_s)(x)ds + \int_0^t \xi_s(x)(g(x) - \langle g, Z_s\rangle)' d\nu_s \tag{1}$$

$$\nu_t = Y_t - \int_0^t \langle g, Z_s\rangle)ds \quad \text{Innovations}$$

We want to study the Information flow from the initial state and running observations $(Y_s|0 \leq s \leq t)$ into the regular conditional distribtution

$$P_{X_t|(Y_s,0\leq s\leq t)}\ (\cdot, y)$$

(the filter).

Is this flow, conservative, dissipative?

Information Theoretic Quantities

$$
\begin{aligned}
S(t) &= I((X_s, s \in [0,T]); Y_s, s \in [0,t]) = \text{supply} \\
C(t) &= I((X_s, s \in [t,T]); Y_s, s \in [0,t]) = \text{storage} \\
D(t) &= S(t) - C(t) = \text{dissipation}
\end{aligned}
$$

Proposition

$$
\begin{aligned}
S(t) &= C(0) + \frac{1}{2} E \int_0^t |g(X_s) - \langle g, Z_s \rangle|^2 ds \\
C(t) &= I(X_t; Z_t) = Eh(Z_t|P_t) \\
D(t) &= EI((X_s, s \in [0,t]); Y_s, s \in [0,t]|X_t)
\end{aligned}
$$

$$\dot{S}(t) = \frac{1}{2}E|g(X_t) - \langle g, Z_t\rangle|^2 \tag{2}$$

$$\dot{D}(t) = E\left(\frac{Ap_t}{p_t}\log p_t - \frac{A\xi_t}{\xi}\log \xi_t\right)(X_t) \tag{3}$$

Sensitivity of Mutual Information $C(t)$ to the randomization in the dynamics of the signal

For Diffusions

$$\dot{D}(t) = \frac{1}{2}E\nabla \log\left(\frac{\xi_t}{p_t}\right)' a\nabla \log\left(\frac{\xi_t}{p_t}\right)(X_t)$$

Rate of change of storage can be found by application of Ito's rule to

$$\xi_t \log\left(\frac{\xi_t}{p_t}\right)(X_t)$$

Equations (2) and (3) show that the supply of information is associated with the second integral in (1)

$$\int_0^t \xi_s(x)(g(x) - \langle g, Z_s \rangle)' d\nu_s$$

and the dissipation associated with the first integral in (1)

$$\int_0^t (\mathcal{A}\xi_s)(x) ds$$

$\dot{S}(t) =$ signal to noise power ratio of the observations and $\dot{D}(t) =$ measure of the rate at which X forgets its past

Notes on Proof:

$$\begin{aligned} C(t) &= I(X_t; Y_s; s \in [0,t]) = I(X_t; Z_t) \\ S(t) &= E \log M_t \ , \end{aligned}$$

where

$$\begin{aligned} M_t = \frac{dZ_0}{dP_0}(x_0) \exp \Bigg(& \int_0^t g(x_s) - \langle g, Z_s \rangle \Bigg)' dw_s \\ & + \frac{1}{2} \int_0^t |g(x_s) - \langle g, Z_s \rangle|^2 ds) \end{aligned}$$

Interactive Statistical Mechanics

The conditional distribution Z_t takes into account the partial observations available up to time t. Define an energy function for $\Sigma_{X|Z}$ in such a way that Z_t is the minimum free-energy state at time t.

Let $(\tilde{Z}_t)$ be a stochastic process that satisfies the filter equation ($\tilde{Z}_t \neq Z_0$) with density $(\tilde{\xi}_t)$.

$E\tilde{\xi}_t$ corresponds to a state of Σ_X and satisfies the Fokker–Planck equation.

Define energy function

$$\begin{aligned} H_{X|Z}(x,t) &= -\log \xi_t(x) \\ E_{X|Z}(\tilde{Z}_t,t) &= \langle H_{X|Z}(\,\cdot\,,t), \tilde{Z}_t \rangle \\ S_{X|Z}(\tilde{Z}_t) &= S_X(\tilde{Z}_t) = -h(\tilde{Z}_t|\lambda_X) \\ \mathcal{F}_{X|Z}(\tilde{Z}_t,t) &= \mathcal{E}_{X|Z}(\tilde{Z}_t,t) - S_{X|Z}(\tilde{Z}_t) \end{aligned}$$

Proposition

(i) Unique minimizer of the free energy of the conditional system $\Sigma_{X|Z}$ at time t in the state Z_t

(ii) $\mathcal{F}_{X|Z}(Z_t, t) = 0 \ \forall \ t$

(iii) If $E\mathcal{F}_{X|Z}(\tilde{Z}_t, t) < \infty$ and $h(\tilde{\Phi}_0|\Phi_0) < \infty$, where $\tilde{\Phi}_0$ and Φ_0 are the distributions of Z_0 and $\tilde{Z}_0$, then the Free Energy of $\Sigma_{X|Z}$ as state $\tilde{Z}_t$ evolves in a positive $(Y_s, s \in [0, t])$ supermartingale.

Item (iii) is like a Conditional Second Law.

We can study the statistical mechanics of the joint system (X, Z). Connection to Bayesian Inference as Free-Energy Minimization

Data Assimilation $\equiv$ Path Estimation or Filtering
or Prediction

Nonlinear Filtering: The Innovations Viewpoint

Stochastic Partial Differential Equation for the Evolution
of the Conditional Density

The Variational Viewpoint:
Information-theoretic Interpretation

Connections to Stochastic Control

Non-equilibrium Statistical Mechanics

Inference and Learning

Sanjoy K. Mitter

Laboratory for Information and Decision Systems

Massachusetts Institute of Technology

Joint work with Charles Fefferman (Princeton),

Hariharan Narayanan (U Washington),

Nigel Newton (U Essex, UK)

DARPA Meeting at Johns Hopkins Applied Physics Lab

January 15, 2013

Bayesian Inference on Topological Structures

- Abstract Framework
- Prior Measures
- Natural Observation Maps

Fitting Manifolds to Random Data

Bayesian Inference & Free Energy Minimization

(Main reference: "A Variational Approach to Nonlinear Estimation," Mitter, S.K. and Newton, N.J. , *Siam J. on Control & Optimization*, **42** 2004.)

Probability Measures on the Space of Persistence Diagrams

(Yuriy Mileyko, Sayan Mukherjee, John Harer

Duke University, Mathematics, Statistical Science)

They prove:

Theorem

Space of Persistence Diagrams with the Wasserstein metric is complete and separable. Allows us to do Bayesian Inference on Space of Persistence Diagrams.

A Variational Formulation of Bayesian Estimation

Let $(\Omega, \mathcal{F}, P)$ be a probability space, $(\mathbf{X}, \mathcal{X})$ and $(\mathbf{Y}, \mathcal{Y})$ Borel spaces, and $X : \Omega \to \mathbf{X}$ and $Y : \Omega \to \mathbf{Y}$ measurable mappings with distributions P_X, P_Y and P_{XY} on $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{X} \times \mathcal{Y}$, respectively. Suppose that:

(H1) there exists a σ-finite (reference) measure, λ_Y, on $\mathcal{Y}$ such that $P_{XY} \ll P_X \otimes \lambda_Y$. (This could be P_Y itself.)

Let $Q : \mathbf{X} \times \mathbf{Y} \to [0, \infty)$ be a version of the associated Radon-Nikodym derivative, and

$$\bar{\mathbf{Y}} = \left\{ y \in \mathbf{Y} : 0 < \int_{\mathbf{X}} Q(x, y) P_X(dx) < \infty \right\}; \qquad (1)$$

then $\bar{\mathbf{Y}} \in \mathcal{Y}$ and $P_Y(\bar{\mathbf{Y}}) = 1$. Let $H : \mathbf{X} \times \mathbf{Y} \to (-\infty, +\infty]$ be defined by

$$\begin{array}{rcll} H(x,y) & = & -\log(Q(x,y)) & \text{if } y \in \bar{\mathbf{Y}} \\ & & 0 & \text{otherwise :} \end{array} \tag{2}$$

then $P_{X|Y} : \mathcal{X} \times \mathbf{Y} \to [0,1]$, defined by

$$P_{X|Y}(A,y) = \frac{\int_A \exp(-H(x,y)) P_X(dx)}{\int_{\mathbf{X}} \exp(-H(x,y)) P_X(dx)}, \tag{3}$$

is a *regular conditional probability distribution* for X given Y; i.e.

$P_{X|Y}(\,\cdot\,, y)$ is a probability measure on $\mathcal{X}$ for each y,

$P_{X|Y}(A, \cdot\,)$ is $\mathcal{Y}$-measurable for each A, and

$P_{X|Y}(A, Y) = P(X \in A \,|\, Y)$ a.s.

Eqs. (1)–(3) constitute an 'outcome-by-outcome' abstract Bayes formula, yielding a posterior probability distribution for X for each outcome of Y.

Let $\mathcal{P}(\mathcal{X})$ be the set of probability measures on $(\mathbf{X}, \mathcal{X})$, and $\mathcal{H}(\mathbf{X})$ the set of $(-\infty, +\infty]$-valued, measurable functions on the same space. For $\tilde{P}_X, \hat{P}_X \in \mathcal{P}(\mathcal{X})$ and $\tilde{H} \in \mathcal{H}(\mathbf{X})$, we define

$$h(\tilde{P}_X \,|\, \hat{P}_X) = \begin{cases} \displaystyle\int_{\mathbf{X}} \log\left(\frac{d\tilde{P}_X}{d\hat{P}_X}\right) d\tilde{P}_X & \text{if } \tilde{P}_X \ll \hat{P}_X \text{ and the integral exists} \\ +\infty & \text{otherwise,} \end{cases} \tag{4}$$

$$i(\tilde{H}) = \begin{cases} -\log\left(\displaystyle\int_{\mathbf{X}} \exp(-\tilde{H}) dP_X\right) & \text{if } 0 < \displaystyle\int_{\mathbf{X}} \exp(-\tilde{H}) dP_X < \infty \\ -\infty & \text{otherwise,} \end{cases} \tag{5}$$

$$\langle \tilde{H}, \tilde{P}_X \rangle = \begin{cases} \displaystyle\int_{\mathbf{X}} \tilde{H} d\tilde{P}_X & \text{if the integral exists} \\ +\infty & \text{otherwise.} \end{cases} \tag{6}$$

It is well known that the relative entropy $h(\tilde{P}_X \,|\, \hat{P}_X)$ can be interpreted as the *information gain* of the probability measure $\tilde{P}_X$ over $\hat{P}_X$. In fact, any version of $-\log(d\tilde{P}_X/d\hat{P}_X)$ is a generalisation of the Shannon information for X. For almost all x, it is a measure of the 'relative degree of surprise' in the outcome $X = x$ for the two distributions $\tilde{P}_X$ and $\hat{P}_X$. Thus, $h(\tilde{P}_X \,|\, \hat{P}_X)$ is the average *reduction* in the degree of surprise in this outcome arising from the acceptance of $\tilde{P}_X$ as the distribution for X, rather than $\hat{P}_X$.

If we interpret $\exp(-\tilde{H})$ as a likelihood function for X, associated with some (unspecified) observation, then $\tilde{H}(x)$ is the 'residual degree of surprise' in that observation if we already know that $X = x$, and $i(\tilde{H})$ is the 'total degree of surprise' in that observation, i.e. the information in the unspecified observation if all we know about X is its prior P_X. In what follows we shall call $\tilde{H}(X)$ the *X-conditional information* in the unspecified observation, and $i(\tilde{H})$ the information in that observation. (Of course, $H(X, y)$ and, respectively, $i(H(\,\cdot\,, y))$ are the X-conditional information and, respectively, information in the observation that $Y = y$.)

Theorem 1

(i) $i\left((H(\,\cdot\,,y)\right) = \min_{\tilde{P}_X}[h(\tilde{P}_X|P_X) + \langle H(\,\cdot\,,y), \tilde{P}_X\rangle]$

(ii) $h(P_{X|Y}(\,\cdot\,,y)|P_X) = \max_{\tilde{H}}\left\{i(\tilde{H}) - \langle \tilde{H}, P_{X|Y}(\,\cdot\,,y)\right\}$

(iii) $P_{X|Y}(\,\cdot\,,y)$ *is the unique minimizer in* (i)

(iv) *If* H^* *is a maximizer in* (ii), *then* $\exists K \in \mathbb{R}$ *s.t.* $H^*(X) = H(\mathbf{X}, y) + K$

Conceptualization

Information Processing over and above that in prior P_X

In (i): Source of additional information is $Y = y$

Bayes Formula: Extracts info. pertinent $h(P_{X|Y}(\,\cdot\,, y)|P_X)$ and leaves *residual* $\langle H, P_{X|Y}\rangle$.

Input information is held in likelihood $\exp(-H(\,\cdot\,, y))$ and extracted information in $P_{X|Y}(\,\cdot\,, y)$

Arbitrary Information procedure that postulates $\tilde{P}_X$ as post-obs. distribution has access to additional information. Hence: the notion Apparent Information.

In (ii): Source of additional information in Posterior Distribution $P_{X|Y}(\,\cdot\,, y)$. The aim now is to postulate an observation, i.e. a likelihood function $\exp(-\tilde{H})$ which gives rise to this observation.

Input Information

$$h\left(P_{X|Y}(\,\cdot\,,y)|P_X\right)$$

is *merged* with the residual information of the postulated observation

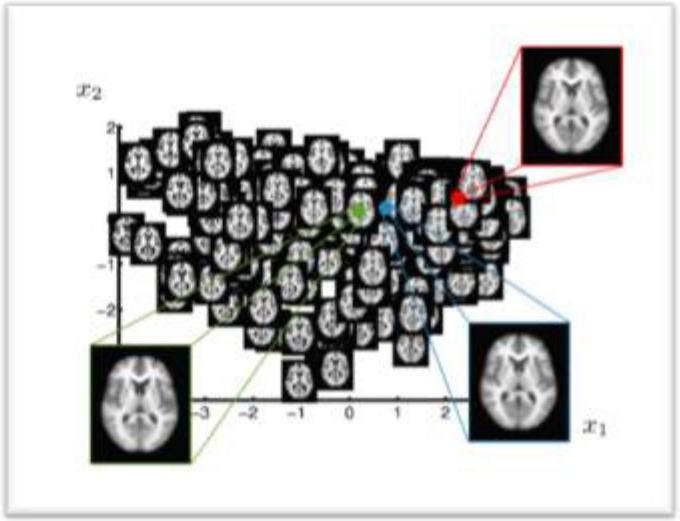
$$\begin{aligned}
\langle \tilde{H}, P_{X|Y}(\,\cdot\,,y)\rangle \quad &: \\
\text{Result} \quad &\geq \quad i(\tilde{H}) \\
\text{With equality} \quad &\Leftrightarrow \quad \text{Obs. is compatible with } P_{X|Y}
\end{aligned}$$

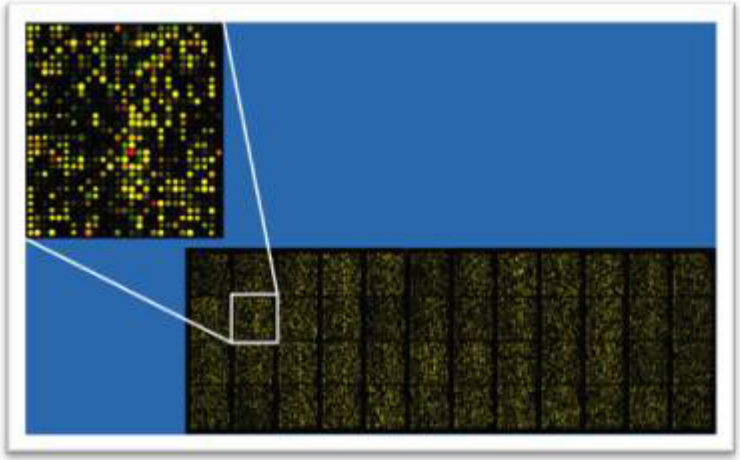
$$\begin{aligned}
i(\tilde{H}) - \langle \tilde{H}, P_{X|Y}(\,\cdot\,,y)\rangle \\
&= \text{Inf. in Postulated Obs.} \\
&\quad \text{compatible with } P_{X|Y}(\,\cdot\,,y)
\end{aligned}$$

Compatible Inf. of $\exp(-\tilde{H})$

High dimensional data

Gerber et al, On the manifold structure of the space of brain images





Number of dimensions is comparable or larger than number of samples

Curse

Sample complexity of function approximation can grow exponentially

Blessings

Concentration of measure

Asymptotic analysis

[David Donoho, AMS 2000]

Manifold learning and manifold hypothesis

Manifold learning is a collection of methodologies for analyzing data

which are motivated by the manifold hypothesis:

high dimensional data tend to lie near a low dimensional manifold

The hypothesis is a way of avoiding the curse of dimensionality

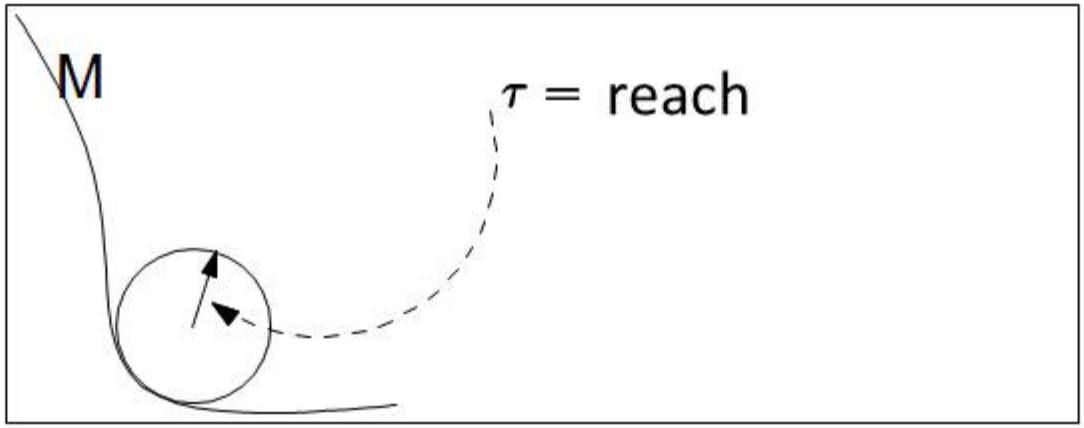
[Kambhatla-Leen'93, Tannenbaum et al'00, Roweis-Saul'00, Belkin-Niyogi'03, Donoho-Grimes'04]

When is the Manifold Hypothesis true?

Geometry may be affected by the generative process, and representation of data.

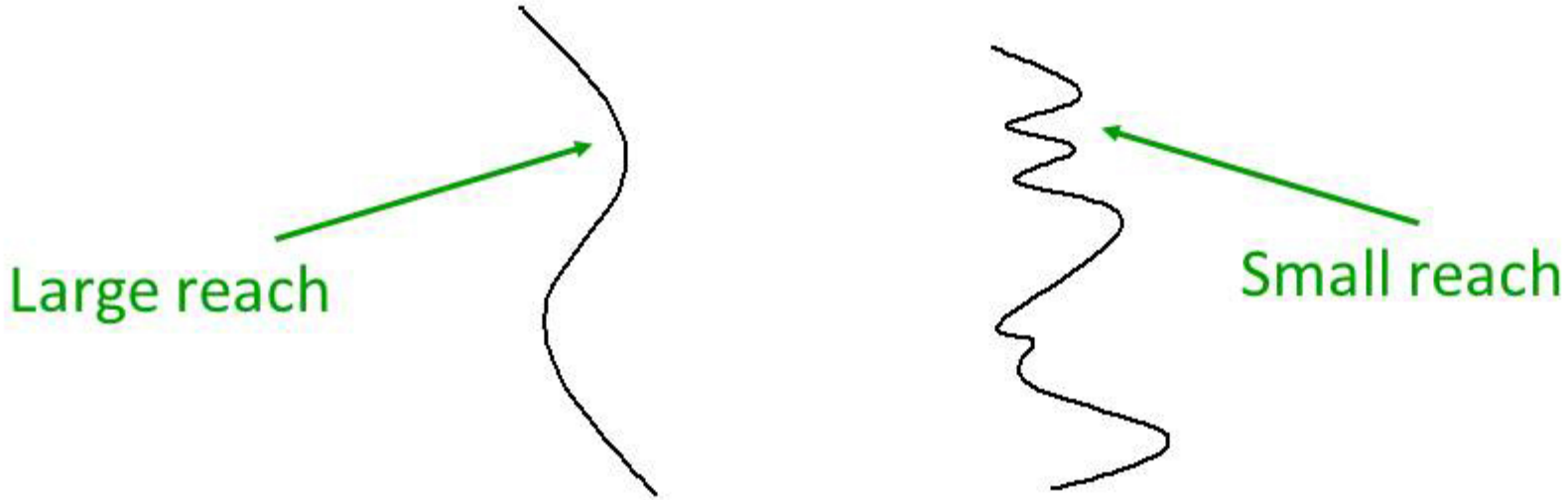
I will discuss one natural formulation of this question and its statistical and algorithmic aspects.

Reach of a submanifold of R^m



τ is the largest number such that for any $r < \tau$

any point at a distance r of $\mathcal{M}$ had a unique nearest point on $\mathcal{M}$



Low dimensional manifolds with bounded volume and reach

Let $\mathcal{G}_e = \mathcal{G}_e(d, V, \tau)$ be the family of

$d-$submanifolds of the unit ball in $\mathbb{R}^n$, with

volume $\leq V$ and reach $\geq \tau$.

Testing the Manifold Hypothesis

Suppose $\mathcal{P}$ is an unknown probability distribution
supported in the unit ball in a separable Hilbert space,
and $x_1, x_2, \ldots$ are i.i.d random samples from $\mathcal{P}$

Given error ϵ, dimension d, volume V, reach τ and confidence $1 - \delta$
is there an algorithm that takes a number of samples
depending on these parameters and outputs whether or not there is

$$\mathcal{M} \in \mathcal{G}_e = \mathcal{G}_e(d, V, \tau)$$

such that w.p $\geq 1 - \delta$, $\int \mathbf{d}(\mathcal{M}, x)^2 d\mathcal{P}(x) < \epsilon$ **?**

Sample Complexity of testing the manifold hypothesis

What is the number of samples needed for testing

the hypothesis that data lie near a low dimensional manifold?

---

the sample complexity of the task depends only on

the intrinsic dimension, volume and reach, but

not ambient dimension

Sample complexity of testing the Manifold Hypothesis

Loss

$\mathcal{L}(\mathcal{M}, \mathcal{P})$ = expected squared distance of a random point to $\mathcal{M}$

Empirical Loss

Given a set of data points $x_1, ..., x_s$

$$L_{emp}(\mathcal{M}) = \frac{\sum_i \mathbf{d}(x_i, \mathcal{M})^2}{s}$$

Sample Complexity

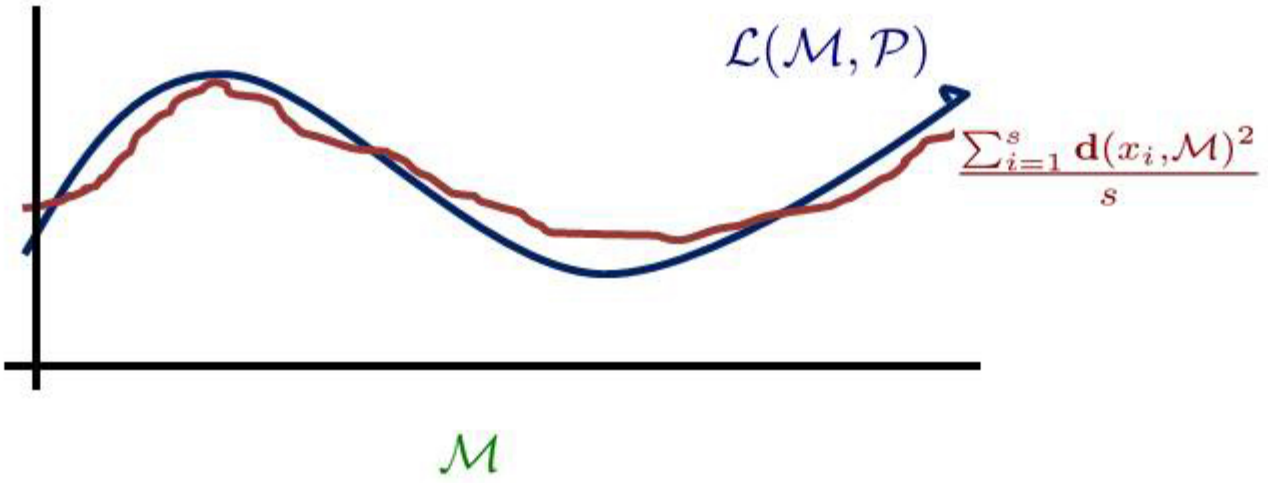
Smallest s such that $\exists$ a rule $\mathcal{A}$ given $x_1, ..., x_s$ i.i.d from $\mathcal{P}$,

$$\mathbb{P}[\mathcal{L}(\mathcal{M}_{\mathcal{A}}, \mathcal{P}) - \inf_{\mathcal{M} \in \mathcal{G}} \mathcal{L}(\mathcal{M}, \mathcal{P}) > \epsilon] < \delta$$

Empirical Risk Minimization

How large must s be to ensure

$$\mathbf{P}\left[\sup_{\mathcal{G}_e}\left|\frac{\sum_{i=1}^{s}\mathbf{d}(\mathcal{M},x_i)^2}{s}-\mathcal{L}(\mathcal{M},\mathcal{P})\right|<\epsilon\right]>1-\delta$$



Fitting manifolds

Theorem:

Let $x_1, \ldots, x_s$ be i.i.d samples from $\mathcal{P}$, a distribution supported on the ball of radius 1 in a separable Hilbert space. If

$$s \geq \frac{C\left(V\left(\frac{1}{\epsilon}+\frac{1}{\tau}\right)^{d+o(d)}+\log 1/\delta\right)}{\epsilon^2}$$

then $\mathbb{P}\left[\sup_{\mathcal{G}_e}\left|\frac{\sum_{i=1}^{s}\mathbf{d}(x_i,\mathcal{M})^2}{s}-\mathbb{E}_{\mathcal{P}}\mathbf{d}(x,\mathcal{M})^2\right|<\epsilon\right]>1-\delta.$

Proof: Approximates manifolds using point clouds and uses the uniform bound for $k-$means.

Algorithmic question

Given N points $x_1, ..., x_N$ in the unit ball in $\mathbb{R}^n$

is there a manifold $\mathcal{M} \in \mathcal{G}_e = \mathcal{G}_e(d, CV, C^{-1}\tau)$

such that $\left(\frac{1}{N}\right) \sum_{1 \le i \le N} \mathbf{d}(x_i, \mathcal{M})^2 \le C\epsilon$ **?**

Here C is some constant depending only on d.

Theorem

There is a controlled constant C depending only on d and an Algorithm that uses

$$n \exp\left((CV^2(\epsilon^{-d}\tau^{-d}))^{1+o(1)}\right) \log \frac{1}{\delta}$$

operations on real numbers such that given $x_1, \ldots, x_N \in B_n$, with probability at least $1 - \delta$, the Algorithm outputs

1. "Yes" if there exists a manifold $\mathcal{M} \in \mathcal{G}_e(d, V, \tau)$ such that

$$\sum_{i=1}^{N} \mathbf{d}(x, \mathcal{M})^2 \leq \epsilon,$$

2. "No" if there exists no manifold $\mathcal{M}' \in \mathcal{G}_e(d, CV, \tau/C)$ such that

$$\sum_{i=1}^{N} \mathbf{d}(x, \mathcal{M}')^2 \leq C\epsilon,$$

Outline

(1) Any manifold $\mathcal{M} \in \mathcal{G}_e = \mathcal{G}_e(d, V, \tau)$

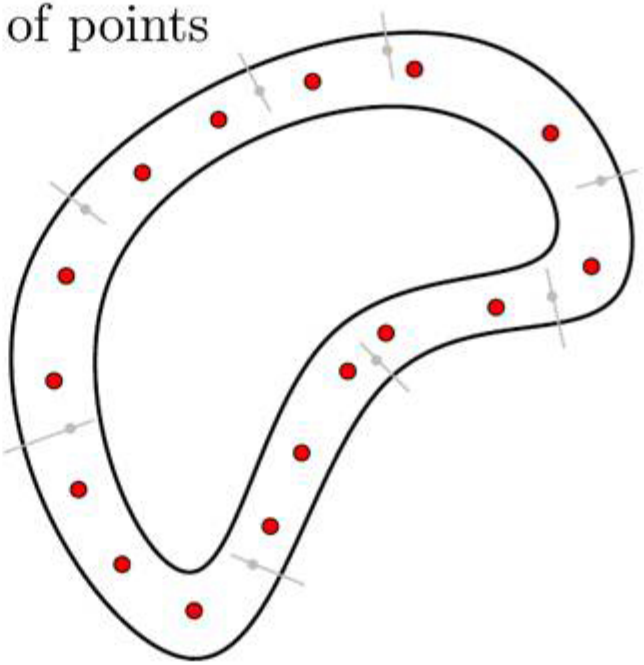
is contained in a $\epsilon-$ neighborhood of an affine subspace W of dimension

$$N_p := N_p(\epsilon)$$

This allows us to reduce the ambient dimension n to roughly N_p

Outline

(2) Reduce the problem to the question of testing whether a discrete evenly spread set of points lie on $\mathcal{M} \in \mathcal{G}_e = \mathcal{G}_e(d, V, \tau)$

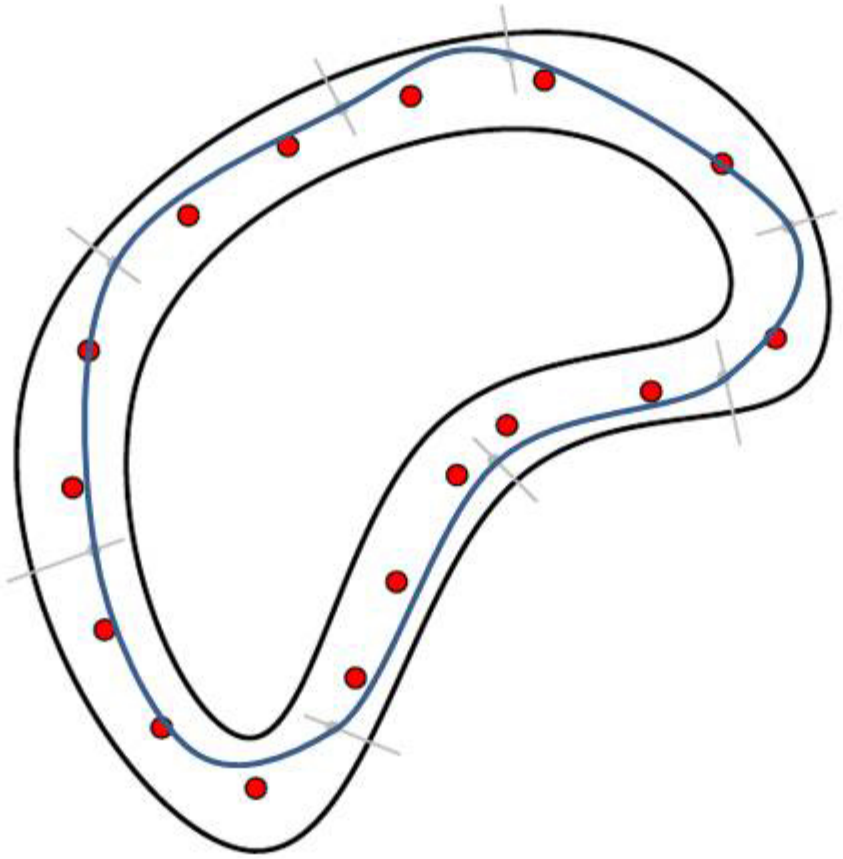


(3) Find a smooth vector bundle D^{init} defined on a tubular neighborhood of data.

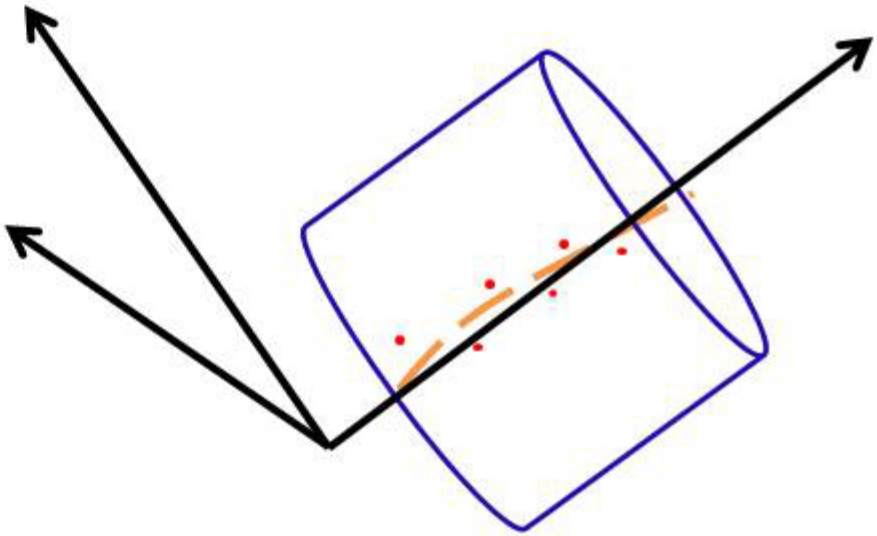
(4) Describe a putative manifold $\mathcal{M}^{init}$ as the set of zeroes of a specific section of the vector bundle

Outline

(5) Restricting the base of D^{init} to $\mathcal{M}^{init}$ we get a bundle D^{norm}.



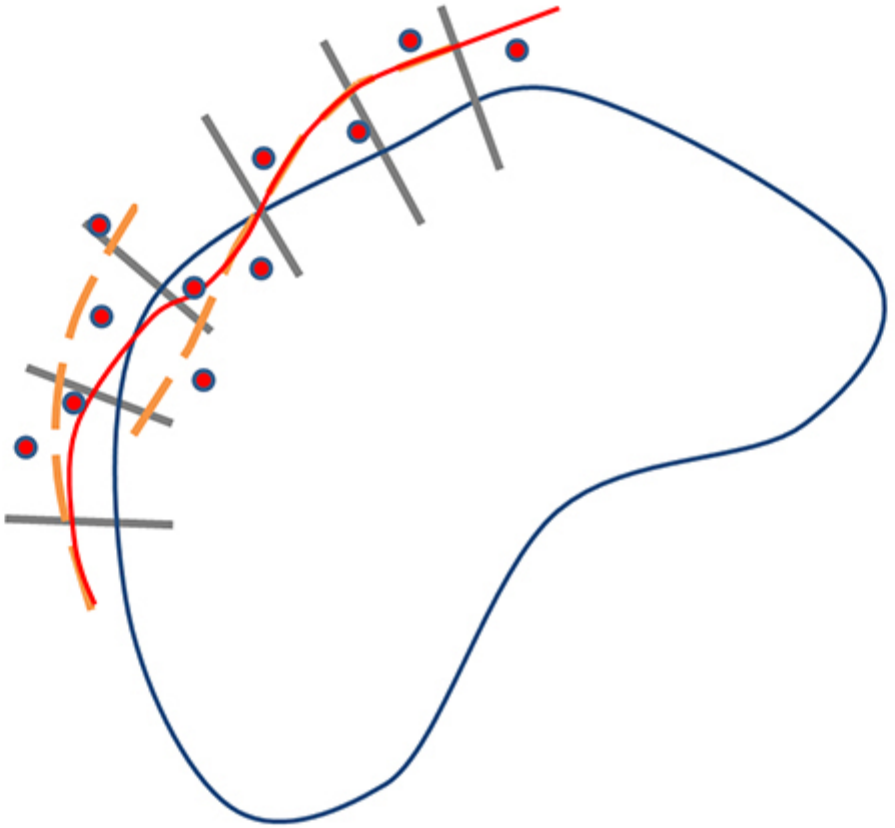
Outline



(6) Obtain individual local sections by optimizing squared loss over a space of second order ($n-d-$dimensional) jets satisfying Whitney's inequalities corresponding to $\mathcal{C}^2-$norm of τ^{-2} over cylinders "aligned to the bundle".

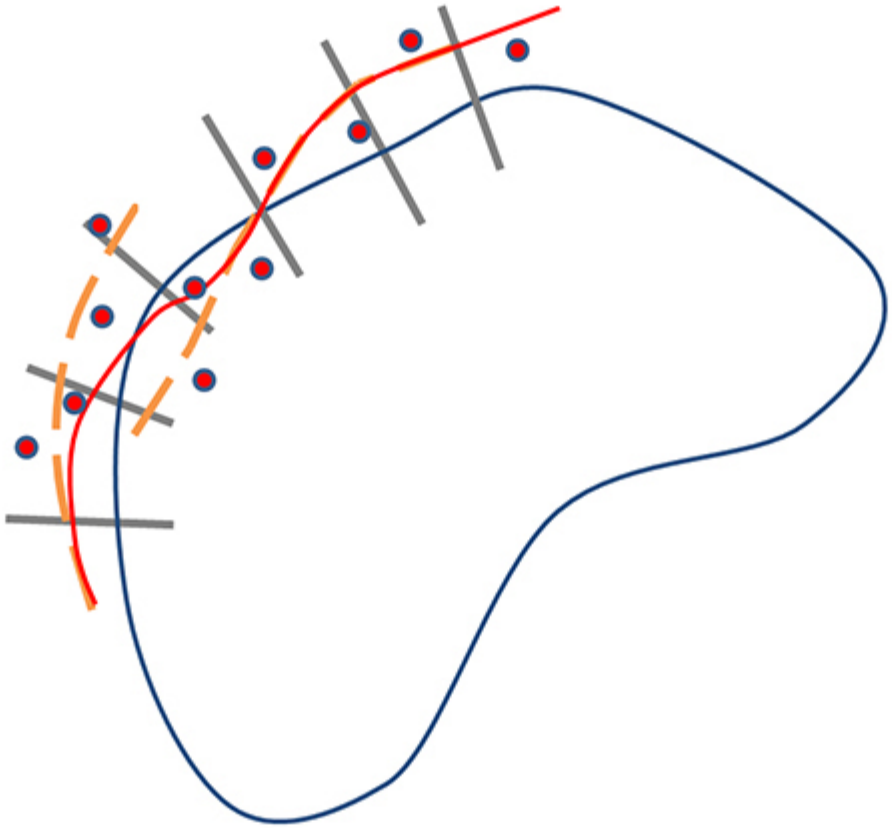
Outline

(7) Obtain a good section of D^{norm} by patching together good local sections using a partition of unity supported on $\mathcal{M}^{init}$.



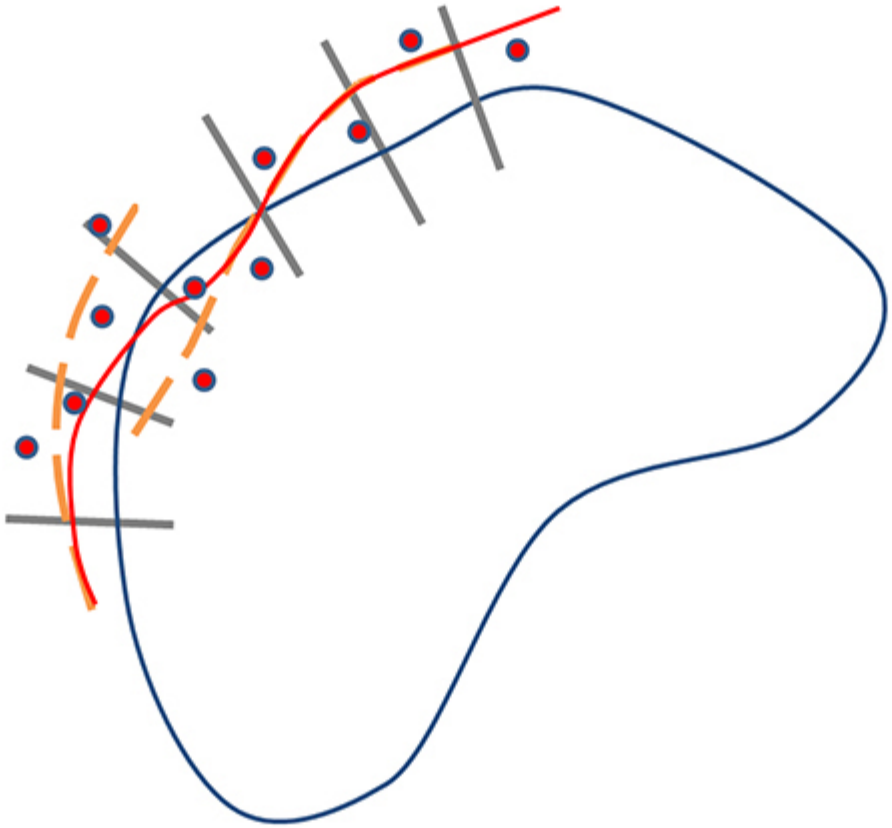
Outline

(7) Obtain a good section of D^{norm} by patching together good local sections using a partition of unity supported on $\mathcal{M}^{init}$.



Outline

(7) Obtain a good section of D^{norm} by patching together good local sections using a partition of unity supported on $\mathcal{M}^{init}$.



Concluding Remarks

- An algorithm for testing the manifold hypothesis.
- Improved sample complexity bounds for k-means.

Future directions:

(a) Make practical and test on real data

(b) Develop non-parametric versions for manifold fitting.

(c) Improve efficiency – better optimization over sections of a vector bundle?

(d) Understand the role of topology in the optimization questions involved .

Towards a Unified View of Communication and Control

Feedback communication problem

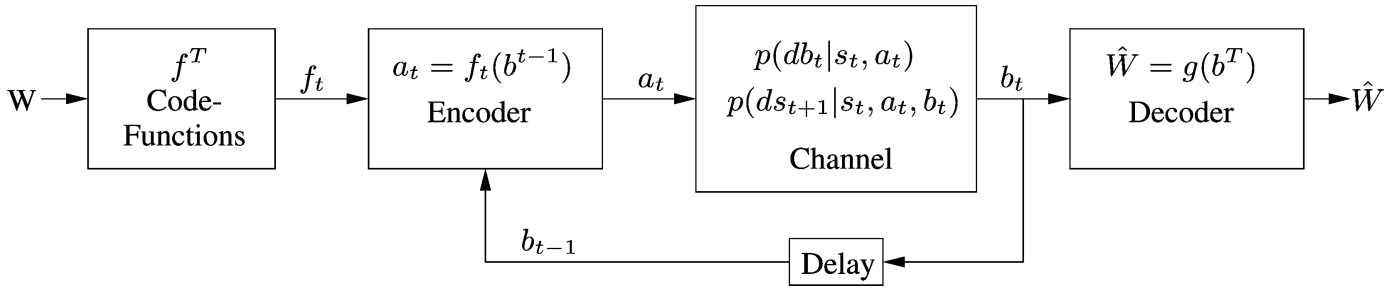


Figure 1. Interconnection

Choose encoder and decoder to transmit message over the channel to minimize the probability of error

Channel at time t: $P(db_t|a^t, b^{t-1})$ stochastic kernel

$$a^t = (a_0, \ldots, a_t)$$

$$\textit{Channel} = \text{Sequence of } \left. P(db_t|a^t, b^{t-1}) \right|_{t=1}^{t}$$

Time ordering: Message $= W, A_1, B_1, \quad , A_T, B_T, \hat{W} =$ Decoded message

$$W = (1, 2, \ldots, M)$$

Code function:

$$\mathcal{F}_t = \{f_t : B^{t-1} \to A : \text{measurable}\}$$
$$\mathcal{F}_T = \prod_{t=1}^{T} \mathcal{F}_t$$

Channel code function: $f^T = (f_1, \ldots, f_t)$

Distribution on code functions:
$P(df_t|f^{t-1})\Big|_{t=1}^{T}$

Channel code = list of M channel code functions

Code functions are introduced to reduce the feedback communication problem to a no feedback communication problem.

Average Measure of Dependence

Mutual Information

$$\begin{aligned} I(A^T; B^T) &= \mathbb{E}_{P_{A^T,B^T}} \log \left(\frac{P_{A^T,B^T}}{P_{A^T} P_{B^T}} \right) \\ &= \mathbb{E}_{P_{A^T,B^T}} \log \left(\frac{P_{B^T|A^T}}{P_{B^T}} \right) \end{aligned}$$

$$I(A^T; B^T) = \sum_{t=1}^{T} I(A^T; B_t | B^{t-1})$$

Information transmitted to the receiver depends on future $(A_{t+1}, \ldots, A_T)$.

Directed Mutual Information (Causal)

$$I(A^T \to B^T) = \sum_{t=1}^{T} I(A^t; B_t | B^{t-1})$$

To compute Mutual Information (Directed Mutual Information), need joint distribution

$$\mathbb{P}_{A^T,B^T}(da^T, db^T)$$

This can be done if we are given the channel

$$P(db^t|a^t, b^{t-1})\Big|_{t=1}^{T}$$

and channel input distributions

$$\mathcal{D}_t := \mathbb{P}(da_t|a^{t-1}, b^{t-1})\Big|_{t=1}^{T}$$

Interconnection of channel input to channel

Channel Capacity

$$C_T = \sup_{\mathcal{D}_T} \frac{1}{T} I(A^T \to B^T)$$

(Note: Optimization over original input codes, not on space of code functions.)

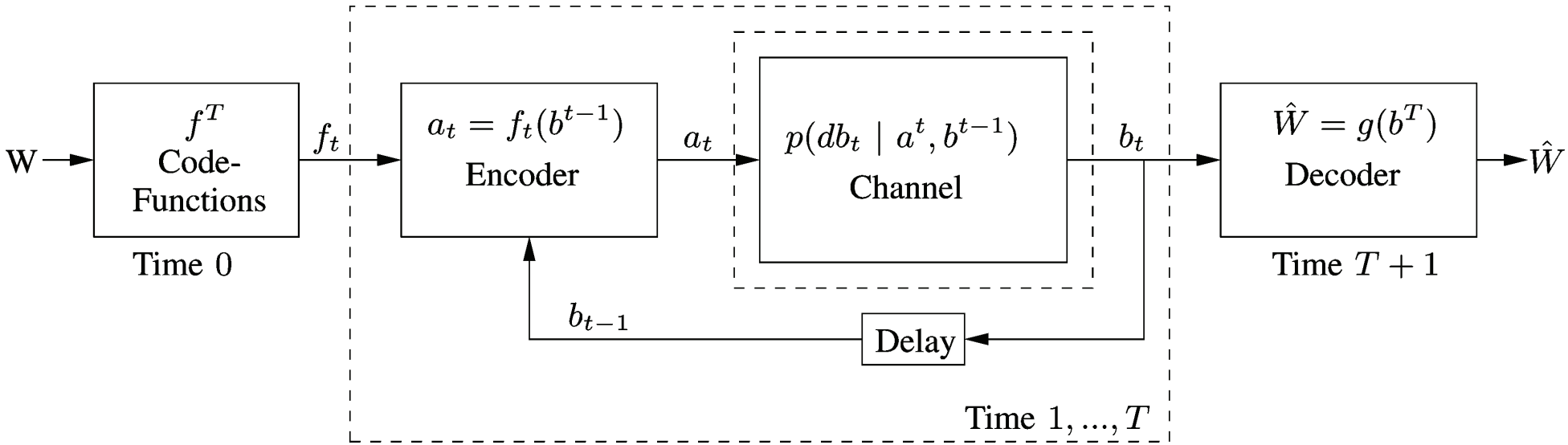


Figure 2: Markov Channels

Markov Channel

$$P(ds_{t+1}|s_t, a_t, b_t)\big|_{t=1}^{T} \quad : \quad \text{state transition}$$

$$P(db_t|s_t, a_t)\big|_{t=1}^{T} \quad : \quad \text{channel output}$$

Capacity of Markov Channels

$$(1) \qquad \sup_{\mathcal{D}_\infty} \lim_{T\to\infty} \frac{1}{T} I(A^T \to B^T)$$

It turns out that by appropriately defining sufficient statistics (π_t) (conditional distributions of the state given information from encoder to decoder) and controls $u_t(da_t|\pi_t)$, and state $X_t = (\pi_{t-1}, A_{t-1}, B_{t-1})$ and instantaneous cost $c(x_t, u_t, u_{t+1})$, (1) can be formulated as a Partially Observed Stochastic Control Problem.

In turn, this can be reformulated as a fully-observable stochastic control problem.

This problem is more like a *dual* control problem since the choice of the channel input can help the decoder identify the channel.

This is also an example where the *information pattern is nested*: The encoder has more information than the decoder.

Communication and Control

Stabilization equivalent to reliable Communication through the loop

Signaling through the loop

Open Problem

Existence of Channel Linking

Controller and Actuator

Asymmetry in Information Transfer

Problems for the Future

- Distributed Estimation and Control

Signalling: Controllers, Estimators have to communicate their actions (estimates) through the plant. There is a role for Information Theory here.

(See recent work of Sahai on Witsenhausen problem)

See: **Michael Spence (Nobel lecture)**

Signalling in Retrospect and the Information Structure of Markets

- Games as Multiple Feedback Loops

(Witsenhausen)

Related to Distributed Control

Problems for the Future (cont.)

• Connections to Statistical Mechanics and Field Theory

Information Theory of Message Passing Algorithms

(See for example: Cramer's Rule and Loop Ensembles: A. Abdesselam and D.C. Brydges)

• Interconnections and Interactions

Optimal Transportation Theory

• What is the Nature of Experimental Work in our Field?

Theory vs. Experiment

Problems for the Future (cont.)

- Systems View (Dynamical) of Economic Classifying Equilibria

(See: Global Trade and Conflicting National Interests: Ralph E. Gomory and William J. Baumol, MIT)

Concluding Remarks