Information-theoretic observations on the calculus of variations

What we know about what we know

Nisheeth Srivastava Peter Harremoës

Centrum voor Wiskunde
en Informatica

September 9, 2008

Outline

Motivating questions

- Basis for mathematical statements of physical laws
- Classical action principle

$$S[q(t)] = \int_{t_1}^{t_2} L[q(t), \dot{q}(t), t]dt \qquad (1)$$

- Several interpretational problems

Epistemologically speaking ...

- Mathematical physics uses theories to make predictions
- Learning makes predictions without domain-specific theory
- Is there a relation? Can it be made precise?

Outline

An intriguing development

- Schrodinger's equation [Frieden 1991]
- Information measures and symmetry [Vtovsky 1996]
- Quantum mechanics [Skala 2005]
- Science from Fisher information [Frieden 2005]

General statement

- Observer plays zero-sum information game with Nature
- Assumes 'bound' information J
- Observer gains information I through measurements
- EPI maximizes K = I - J

Recovering laws of physics

- Efficient measurement defined as $\kappa \doteq I/J = 1$
- Requires statement of invariance expressed as unitary transformation
- In practice, requires Fourier dual of observation space to be observable
- $$K = I[\psi(\mathbf{x})] - J[\phi(\boldsymbol{\mu})] = \mathit{extrem},$$
- Usable iff
$$I[\psi(\mathbf{x})] - J[\phi(\mathbf{x})] = \mathit{extrem}. \quad (2)$$

Example:Schrodinger's equation

- Conventional derivations make three physical assumptions
  - Energy momentum relationship $E = \frac{p^2}{2m} + V(x)$
  - Einstein's light quanta hypothesis $E = h\nu$
  - de Broglie's hypothesis $p = \frac{h}{\lambda}$
- EPI derivation dispenses with the latter two
- *Conjecture*: First assumption is entirely representational

Fisher Information

- Measures informativeness of a probability distribution p parameterized by $\boldsymbol{\theta}$,

$$I(\boldsymbol{\theta}) = \int \left(\frac{\partial \log p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^2 \; p(\mathbf{x};\boldsymbol{\theta}) \; d\mathbf{x} \tag{3}$$

- Trace of FI matrix upper bounds Stam information - understand as *capacity* of estimation procedure

Measuring physical systems

- Consider ideal V-dimensional measurement scenario

$$\mathbf{y}_n = \boldsymbol{\theta}_n + \mathbf{x}_n, \quad n = 1 \cdots N \tag{4}$$

- Assume independent observations
- Assume shift invariance
-

$$I = \int \frac{1}{p_n(\mathbf{x}_n)} \sum_n \nabla p_n(\mathbf{x}_n) \cdot \nabla p_n(\mathbf{x}_n) \tag{5}$$

Complex probability amplitudes

- Work with real probability amplitudes $p(\mathbf{x}) = q^2(\mathbf{x})$,

$$I = 4 \int \sum_n \nabla q_n(\mathbf{x}_n) \cdot \nabla q_n(\mathbf{x}_n)$$

- Define complex probability amplitudes,

$$\psi_n(x_n) = \frac{1}{\sqrt{N}} \left(q_{2n-1} + iq_{2n}\right), \quad n = 1 \cdots N/2.$$

- Then,

$$I = 4N \int d\mathbf{x} \sum_n \nabla \psi_n^* \cdot \nabla \psi_n. \tag{6}$$

Fourier duality in observation space

- Fourier duals: $\psi(x) \leftrightarrow \phi(\mu)$
- Fourier duals: $\nabla\psi(x) \leftrightarrow \imath\mu x/\hbar$
- Have introduced scaling parameter $\hbar$

Statement of symmetry

- Unitary transformation allows application of Parseval's theorem
- Restate (6) as,

$$J \equiv \frac{4N}{\hbar^2} \int d\mu \; \mu^2 \sum_n |\phi_n(\mu)|^2. \qquad (7)$$

- Physically, is simply expectation over momentum, so

$$J = \frac{8Nm}{\hbar^2} \langle E_{kin} \rangle = \frac{8Nm}{\hbar^2} \langle [W - V(x)] \rangle$$

- Can measure energy in both observational domains, hence

$$J = \frac{8Nm}{\hbar^2} \int dx \; [W - V(x)] \; \sum_n |\psi_n(x)|^2. \qquad (8)$$

Applying EPI

- EPI Lagrangian

$$\mathcal{L} = N \sum_n \int dx \ \left[4 \left| \frac{d\psi_n(x)}{dx} \right|^2 - \frac{8m}{\hbar^2} [W - V(x)] |\psi_n(x)|^2 \right]. \tag{9}$$

- Solving with Euler-Lagrange equation gives,

$$\psi_n''(x) + \frac{2m}{\hbar^2} \ [W - V(x)] \psi_n(x) = 0, \quad n = 1 \cdots N/2, \tag{10}$$

Outline

Problems in formulation

- Extremizing = finding points of least variation
- How does one derive J in a principled manner?
- What does EPI mean? Hamiltonian, path integral derivations

Problems in implementation

- Why is the Fourier transform so fundamental?
- Why is Fisher information so fundamental?
- Where do values of physical constants come from?

Outline

What is learning?

- Springs from AI algorithms of the 80s
- Mathematical formulations of cognitive processes
- Various philosophies extant
    - PAC Learning [Valiant 1984]
    - VC theory [Vapnik 1971]
    - Bayesian inference
    - Maxent learning [Berger 1996]
    - Information theoretic learning e.g. MDL [Grunwald 2007]

A disclaimer

- Link between learning theory and information optimization not formal
- Some frameworks for learning quite mathematically disjoint
- Efforts for unification continue

Model-free learning

- Define abstract information space
- Define preference relations
- Find optimality conditions

Information spaces

- Set $\mathcal{A} \leftarrow$ possible observational outcomes (rewards, states, error etc.)
- Some elements of $\mathcal{A}$ unobservable $\Rightarrow$ learning with uncertainty
- Convex subsets mathematically tractable; we restrict ourselves to these

Some notation

- X and Y are dual (conjugate) spaces of functions $x : A \to \mathbb{R}$ and $y : A \to \mathbb{R}$
- The inner product is represented as $(.,.) : X \times Y \to \mathbb{R}$, i.e.

$$(x, y) = \int_A x(a)\ y(a)\ da$$

- Hulls of convex subsets of X represented as K_X

Convex sets

- Convex sets: sets closed under convex combinations
- Hulls described by support and distance functions
- Support of convex hull K_X at $y \in Y$ is

$$F(y) = \sup\{(y, x) : x \in K_X\}. \tag{11}$$

- Distance from the center x_0 of convex hull K_X is

$$\hat{F}(x) = \inf\{D \geq 0 : x \in DK_X\}. \tag{12}$$

- Polar convex sets: support function of one is distance function for the other

Representing optimality conditions

- $x \in K_x \doteq \tilde{F}(x) \leq D < \infty$
- Dual convex functionals related as

$$F(y) = \sup_x \{(y, x) - \tilde{F}(x)\},$$
$$\tilde{F}(x) = \sup_y \{(x, y) - F(y)\}.$$

- Also satisfy the dual minimization problems,

$$D(C) = \inf\{\tilde{F}(x) : (y, x) \geq C\},$$
$$C(D) = \inf\{F(y) : (x, y) \geq D\}.$$

Optimization

- Legendre duality is statement of polar relationship between two convex hulls
- Extremizing a convex functional F(y) defined on set A gives optimal information trajectory
- Optimality conditions generalizations of Kuhn-Tucker conditions [Kuhn 1951]

Necessary conditions for extrema

Theorem

Extrema $y^* \in K_Y$ *for* $\tilde{F}(x) = \sup\{(x, y) : F(y) \leq C\} = D$ *satisfy,*

- $\beta x \in \partial F(y^*)$
- $F(y^*) = C$
- $\beta^{-1}(C) = D'(C)$

An example

- Set $y \in Y$ as probabilities and $F(y)$ as KL divergence
- Optimal function $y^* \in K_Y$ for $\tilde{F}(x) = \sup\{(x, y) : F(y) \leq C\}$
- Has the form $y_0 e^{\beta x - \gamma(\beta)}$.

Relation to statistical mechanics

- Minimizing KL divergence is precisely the principle of MDI [Kullback 1989]
- MDI is equivalent to MaxEnt in most cases e.g., distributions with finite support
- Form of solutions recovers statistical mechanics

Outline

Relation to EPI derivation of Schrodinger's equation

- Set $y \in Y$ as errors in position measurement and F(y) as Fisher Information (6)
- Informational constraint here is a symmetry property
- Symmetry expressed as statement of invariance of FI across unitary transformation (7)
- Recover EPI Lagrangian (9)

Relation to EPI derivation of Schrodinger's equation

- Set $y \in Y$ as errors in position measurement and F(y) as Fisher Information (6)
- Informational constraint here is a symmetry property
- Symmetry expressed as statement of invariance of FI across unitary transformation (7)
- Recover EPI Lagrangian (9)
- There is an error in this argument, can you spot it?

Review of limitations

- Explained significance of Fisher Information
- Significance of Fourier Transform
- EPI falls out of more general theory
- No explanation for values of physical constants
- Introduction of complex numbers is still mysterious