MaxEnt 2007

JULY 8-13, 2007

SARATOGA SPRINGS, NEW YORK USA

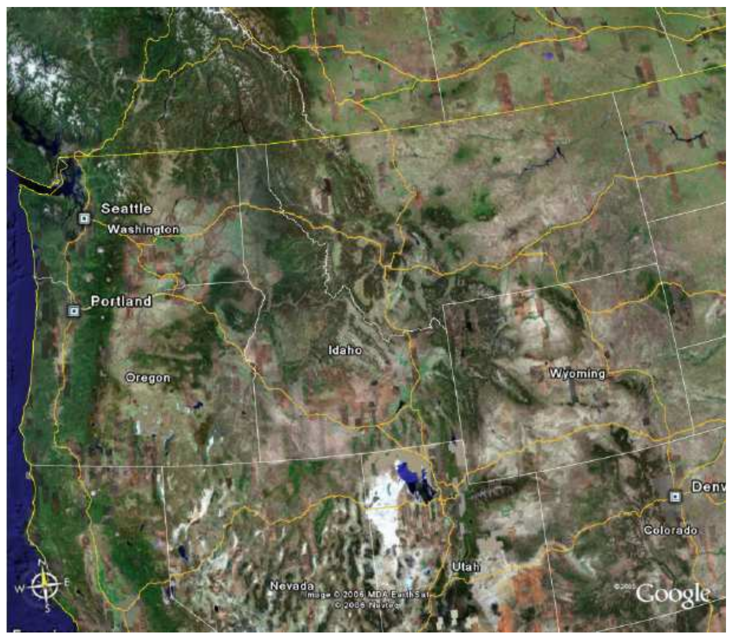
Vesslin I. Dimitrov
on
Shannon-Jaynes Entropy and Fisher Information

July 13, 2007

(154)

ON SHANNON-JAYNES ENTROPY AND FISHER INFORMATION

Vesselin I. Dimitrov
Idaho Accelerator Center
Idaho State University, Pocatello, ID
dimivess@isu.edu



The MAXENT Paradigm

- **State of knowledge:** **probability density** $p(\mathbf{x})$

- **Data:** **constraints** $\int dx p(\mathbf{x}) C_i(x) = d_i$

- **Quantification:**
  - **information content** $S[p(\mathbf{x})]$
  - **information distance** $S[p_1(\mathbf{x}), p_2(\mathbf{x})]$

- **MAXENT (constructive):** **minimize information content subject to the available data/constraints**

$$S[p(\mathbf{x})] + \lambda_i \int dx p(\mathbf{x}) C_i(\mathbf{x}) \rightarrow \min_{p(\mathbf{x})}$$

- **MAXENT (learning rule):** **minimize information distance subject to the available data/constraints**

$$S[p_1(\mathbf{x}), p_0(\mathbf{x})] + \lambda_i \int dx p_1(\mathbf{x}) C_i(\mathbf{x}) \rightarrow \min_{p_1(\mathbf{x})}$$

Motivation: Physics Laws as Rules of Inference

- **Why would universal physics laws exist? Who is enforcing them?**
  The more universal a law is, the less likely it is to actually exist. The laws of thermodynamics were once thought to have universal validity but turned out to express a little more than handling one's ignorance in a relatively honest way (Jaynes'57). Moreover, most fundamental aspects of thermodynamics laws are insensitive to the particular choice of the information distance!

- **Mechanics**:
  - ✓ Free particle (T.Toffoli'98)
  - ✓ Particle in a force field (*)
- **Electrodynamics**:
  - ✓ Free field (*)
  - ✓ Field with sources (*)
- **Thermodynamics**:
  - ✓ Entropy (E.Jaynes'1957)
  - ✓ Fisher information (Frieden, Plastino^2, Soffer'1999)
  - ✓ General criterion (Plastino^2'1997, T.Yamano'2000)
- **Quantum Mechanics**:
  - ✓ Non-relativistic (Frieden'89,?)
  - ✓ Relativistic (?)
  - ✓ Quantum Field Theory (?)

Problems with the Standard Characterizations

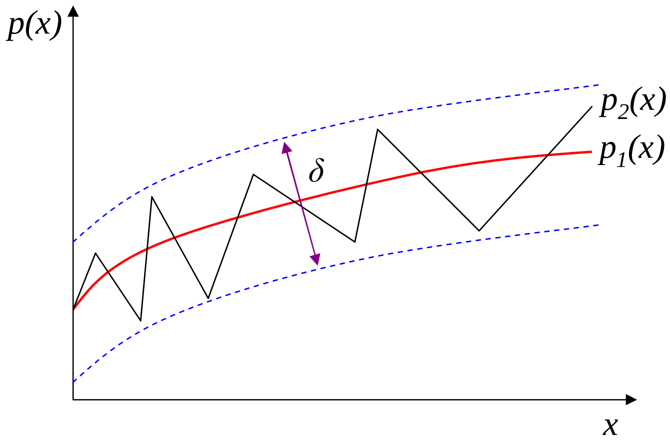
- **Shannon-Jaynes entropy**:
  - ✓ Problems with continuous variables;
  - ✓ In hindsight similar problems even with discrete variables;
  - ✓ The constraints are not likely to be in the form of true average values;
- **Shannon-Jaynes distance (Shore & Johnson):**
  - ✓ Karbelkar'86 points out that there is a problem with the constraints as average values;
  - ✓ Uffink'95 takes issue with SJ's additivity axiom; this I believe is misguided (Caticha'06);
  - ✓ The *real problem* with SJ's derivation: SJ prove that the correct form of the relevant functional is

$$H[q(\mathbf{x}), p(\mathbf{x})] = g(F[q(\mathbf{x}), p(\mathbf{x})]) \qquad F[q(\mathbf{x}), p(\mathbf{x})] = \int dx q(\mathbf{x}) f\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right)$$

    where *g(.)* is a monotonic function. Then, however, they declare *H* and *F* equivalent and require additivity for *F only*, thus excluding the possibility that *F* factorizes and $g(xy)=g(x)+g(y)$. This ignored possibility is precisely the general Renyi case!
- **Vàn's derivation:**
  - ✓ Apparently the first to consider dependence also on the derivatives of $p(\mathbf{x})$;
  - ✓ Derives the functional as a linear combination of SJ term and Fisher information term;
  - ✓ Deficient derivation – handles functionals as if they were functions.

New characterization from scratch is highly desirable!

Inadmissible and Admissible Variations



$$\delta p(x) = p_1(x) - p_2(x)$$

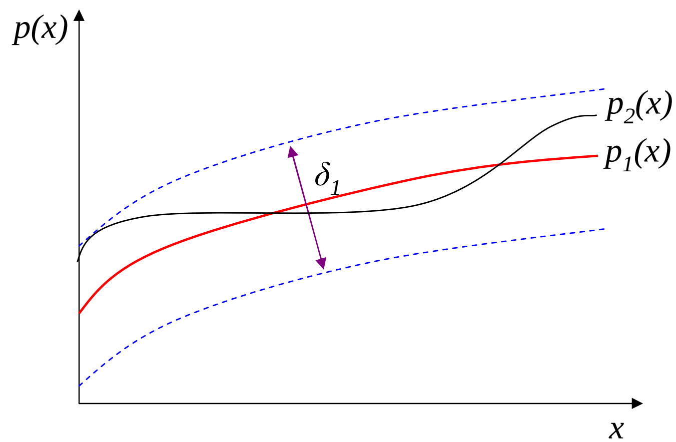
$$\forall x \quad |\delta p(x)| < \delta$$

Variations of this kind shouldn't be allowed because they are physically insignificant. Yet they are included if we write $S[p] = S[p(x)]$

$$\delta p(x) = p_1(x) - p_2(x)$$

$$\forall x \quad |\delta p(x)| < \delta_1$$

$$\left|\frac{d}{dx}\delta p(x)\right| < \delta_2$$



Excluding the former and allowing the latter variations is achieved by writing $S[p] = S[p(x), \nabla p(x)]$

Axioms & Consequences

- **Locality:** $S[p] = g\left(\int dx f(x, p, \nabla p)\right)$
- **Expandability:** $S[p] = g\left(\int dx p f(x, p, \nabla p)\right) \quad \lim_{p\to 0} p f(x, p, \nabla p) = 0$
- **Additivity (simple):** $S[p_1 p_2] = S[p_1] + S[p_2]$ $\to$ $g(x+y) = g(x) + g(y)$; $\to$ $g(xy) = g(x) + g(y)$

$$S[p] = c_1 \ln \int dx p(x) \left[\frac{p(x)}{m_1(x)}\right]^{\nu - 1} \exp\left[c_0 \left(\nabla \ln \frac{p(x)}{m_2(x)}\right)^2 \right]$$

- Information "Distance" interpretation: $m_1(x) = m_2(x) = m(x)$

$$S[p, m] = \int dc_0 dc_1 d\nu\, W(c_0, c_1, \nu)\, c_1 \ln \int dx p(x) \left[\frac{p(x)}{m(x)}\right]^{\nu - 1} \exp\left[c_0 \left(\nabla \ln \frac{p(x)}{m(x)}\right)^2 \right]$$

Axiom (ii): Conditional probabilities

$$\int dx p(x,y) = \delta(y-y_0) \qquad S \to S + \int dy \lambda(y) \int dx p(x,y)$$

$$S[p] = g\left(\int dxdy f(\frac{p}{m})\right)$$

$$\frac{\delta}{\delta p} f(\frac{p}{m}) + \lambda(y) = 0$$

$$f'(\frac{p(x,y)}{m(x,y)}) + \lambda(y) m(x,y) = 0$$

$$S[p] = g\left(\int dxdy p f(\frac{p}{m})\right)$$

$$\frac{\delta}{\delta p}[p f(\frac{p}{m})] + \lambda(y) = 0$$

$$f(\frac{p(x,y)}{m(x,y)}) + \frac{p(x,y)}{m(x,y)} f'(\frac{p(x,y)}{m(x,y)}) + \lambda(y) = 0$$

$$p(x,y) = m(x,y)\phi(y)$$

$$f(\phi) + \phi f'(\phi) + \lambda(y(\phi)) = 0 \to \phi_\lambda^*(y)$$

$$\int dx m(x,y)\phi_\lambda^*(y) \stackrel{!}{=} \delta(y-y_0) \to \phi_\lambda^*(y) = \frac{\delta(y-y_0)}{m(y)}$$

$$p(x|y) = \int dy p(x,y) = \int dy \frac{m(x,y)\delta(y-y_0)}{m(y)} = \frac{m(x,y_0)}{m(y_0)}$$

Axiom (ii): Conditional probabilities 2

$$q(x) \equiv \ln \frac{p(x)}{m(x)}$$

$$S[p] = g\left(\int dxpf\,(q,(\nabla q)^2)\right)$$

$$\delta \int dxpf\,(q,(\nabla q)^2) = \int dx\delta p\left[f + f_{,1}\right] + 2\int dxpf_{,2}\nabla q \cdot \nabla \delta q =$$
$$= \int dx\delta p\left[f + f_{,1} - 2m^{-1}\nabla \cdot (mqf_{,2}\nabla q)\right] = \int dx\delta p\left[f + f_{,1} - 2\nabla \cdot (qf_{,2}\nabla q) - 2qf_{,2}\nabla \ln m \cdot \nabla q\right] =$$

$$\frac{\delta}{\delta p}\int dxpf\,(q,(\nabla q)^2) = f + q\left[f_{,1} - 2f_{,2}\left[\frac{(\nabla q)^2}{q} + \Delta q + \underline{\nabla \ln m \cdot \nabla q}\right] - 2\nabla q \cdot \nabla f_{,2}\right]$$

The term with ln*m* can cause potential trouble with reproducing the Kolmogorov's rule unless either *m=const* or else $f_{,2}=0$, which implies $c_0=0$

General Consequences of Additivity

$$q(x) \equiv \ln\frac{p(x)}{m(x)} \qquad S[p] = g\left(\int dxpf(q,(\nabla q)^2)\right) \qquad q(x_1,x_2) = q_1(x_1) + q_2(x_2) \qquad (\nabla q)^2 = (\nabla_1 q_1)^2 + (\nabla_2 q_2)^2$$

a) $g(x+y) = g(x) + g(y) \Rightarrow g(x) = cx$

$$\int dx_1 dx_2\, p_1 p_2\, f(q_1+q_2,(\nabla_1 q_1)^2 + (\nabla_1 q_1)^2) = \int dx_1 p_1 f(q_1,(\nabla_1 q_1)^2 + \int dx_2 p_2 f(q_2,(\nabla_1 q_1)^2)$$

$$f(q_1+q_2, z_1+z_2) = f(q_1,z_1) + f(q_2,z_2) \Rightarrow f(q,z) = \alpha q + \beta z$$

$$\boxed{S[p] = c\int dxp\left[\alpha \ln\frac{p}{m} + \beta(\nabla \ln\frac{p}{m})^2\right]}$$

b) $g(xy) = g(x) + g(y) \Rightarrow g(x) = c\ln x$

$$\int dx_1 dx_2\, p_1 p_2\, f(q_1+q_2,(\nabla_1 q_1)^2 + (\nabla_1 q_1)^2) = \int dx_1 p_1 f(q_1,(\nabla_1 q_1)^2 \times \int dx_2 p_2 f(q_2,(\nabla_1 q_1)^2)$$

$$f(q_1+q_2, z_1+z_2) = f(q_1,z_1) \times f(q_2,z_2) \Rightarrow f(q,z) = \exp(\alpha q + \beta z)$$

$$\boxed{S[p] = c\ln\int dxp\left[\left(\frac{p}{m}\right)^{\alpha} \exp[\beta(\nabla \ln\frac{p}{m})^2]\right]}$$

Particular Parameter Choices

$$S[p,m]=c_1\ln\int dx\,p(x)\left[\frac{p(x)}{m(x)}\right]^{\nu-1}\exp\left[c_0\left(\nabla\ln\frac{p(x)}{m(x)}\right)^2\right]$$

- $c_0 << 1$: $$S[p,m]=c_1(1-\nu)S_R[p,m]+c_0c_1\frac{\int dx\,p(x)\left(\frac{p(x)}{m(x)}\right)^{\nu-1}\left(\nabla\ln\frac{p(x)}{m(x)}\right)^2}{\int dx\,p(x)\left(\frac{p(x)}{m(x)}\right)^{\nu-1}}$$

✓ $\nu=1$: $$S[p,m]=c_1c_0\int dx\,p(x)\left(\nabla\ln\frac{p(x)}{m(x)}\right)^2$$ (Fisher distance)

✓ $c_0=0 \quad c_1=(\nu-1)^{-1}$: $$S[p,m]=\frac{1}{\nu-1}\ln\int dx\,p(x)\left(\frac{p(x)}{m(x)}\right)^{\nu-1}$$ (Renyi distance)

✓ $c_0=b(\nu-1) \quad c_1=a(\nu-1)^{-1}$
$\nu=1$ $$S[p,m]=a\int dx\,p(x)\ln\frac{p(x)}{m(x)}+b\int dx\,p(x)\left[\nabla\ln\frac{p(x)}{m(x)}\right]^2$$

(Vàn-type distance)

Variational Equation

$$q(x) \equiv \ln \frac{p(x)}{m(x)} \qquad S[p] = g\left(\int dx p f(q, (\nabla q)^2)\right)$$

$$\boxed{f + f_{,1} - 2 f_{,2} \Delta q - 2\left(f_{,2} + q f_{,12}\right)\frac{(\nabla q)^2}{q} - 4 f_{,22} \nabla q \cdot \nabla\nabla q \cdot \nabla q - 2 f_{,2} \nabla \ln m \cdot \nabla q = \lambda C(x)}$$

$$f(q, (\nabla q)^2) = \exp\left[(\nu - 1) q + c_0 (\nabla q)^2\right] \longrightarrow \begin{array}{ll} f_{,1} = (\nu - 1) f & f_{,2} = c_0 f \\ f_{,12} = (\nu - 1) c_0 f & f_{,22} = c_0^2 f \end{array}$$

$$\nu f - 2 c_0 \left[\Delta q + \left(1 + (\nu - 1) q\right)\frac{(\nabla q)^2}{q} + 2 c_0 \nabla q \cdot \nabla\nabla q \cdot \nabla q + \underline{\nabla \ln m \cdot \nabla q}\right] f = \lambda C(x)$$

Vàn - Type Updating Rule

Vàn-type distance: $S[p_1,p_2]=2\int dx\,\psi_1^2\left[a\ln\frac{\psi_1(x)}{\psi_2(x)}+2b\left[\nabla\ln\frac{\psi_1(x)}{\psi_2(x)}\right]^2\right]$ $\quad\left(p_i(x)=\psi_i^2(x)\right)$

Vàn-type updating rule: $\delta_\psi\left[\int dx\,\psi^2\left[a\ln\frac{\psi}{\mu}+2b\left[\nabla\ln\frac{\psi}{\mu}\right]^2\right]+\lambda\int dx\,\psi^2 C(x)\right]=0$

$$\boxed{-\Delta\psi+\left[\frac{a}{b}+\frac{a}{b}\ln\left(\frac{\psi}{\mu}\right)^2+\frac{\Delta\mu}{\mu}+\lambda C(x)\right]\psi=0}$$

Non-linear Schrodinger equation of Bialynicki-Birula – Mycielski type.

If $\mu(x)$ is determined on an earlier stage of updating from a prior $\mu_0(x)$ and constraint $C_0(x)$ then

$$\frac{\Delta\mu}{\mu}=c+c\ln\left(\frac{\mu}{\mu_0}\right)^2+\frac{\Delta\mu_0}{\mu_0}+\lambda_0 C_0(x)$$

and any subsequent constraints are additive:

$$-\Delta\psi+\left[2c+c\ln\left(\frac{\psi}{\mu_0}\right)^2+\frac{\Delta\mu_0}{\mu_0}+\lambda_0 C_0(x)+\lambda C(x)\right]\psi=0$$

Fisher Updating Rule

Fisher distance: $$I[p_1, p_2] = 4\int dx\,\psi_1^2\left(\frac{\Delta\psi_2}{\psi_2} - \frac{\Delta\psi_1}{\psi_1}\right) = 4\int dx\,\psi_1\left(\frac{\psi_1}{\psi_2}\Delta\psi_2 - \Delta\psi_1\right)$$

Fisher updating rule: $$\delta_\psi\left[\int dx\,\psi\left(\frac{\psi}{\mu}\Delta\mu - \Delta\psi\right) + \lambda\int dx\psi^2 C(x)\right] = 0$$

$$\boxed{-\Delta\psi + \left[\frac{\Delta\mu}{\mu} + \lambda C(x)\right]\psi = 0}$$

Schrodinger equation where the "quantum potential" of the prior adds to the constraint.

If $\mu(x)$ is determined on an earlier stage of updating from a prior $\mu_0(x)$ and constraint $C_0(x)$ then

$$\frac{\Delta\mu}{\mu} = \frac{\Delta\mu_0}{\mu_0} + \lambda_0 C_0(x)$$

and any subsequent constraints are additive:

$$-\Delta\psi + \left[\frac{\Delta\mu_0}{\mu_0} + \lambda_0 C_0(x) + \lambda C(x)\right]\psi = 0$$

Shannon-Jaynes Updating Rule

Shannon-Jaynes distance: $$S[p_1, p_2] = \int dx p_1 \ln \frac{p_1}{p_2}$$

Shannon-Jaynes updating rule: $$\delta_p \left[\int dx p \ln \frac{p}{m} + \lambda \int dx p C(x) \right] = 0$$

$$\boxed{p(x) = m(x) \exp[-1 - \lambda C(x)]}$$

The usual exponential factor updating rule.

If $\mu(x)$ is determined on an earlier stage of updating from a prior $\mu_0(x)$ and constraint $C_0(x)$ then

$$\mu = \mu_0 \exp[-1 - \lambda_0 C_0(x)]$$

and any subsequent constraints are additive:

$$p(x) = m_0(x) \exp[2 - \lambda_0 C_0(x) - \lambda C(x)]$$

Uniqueness: Consistency with Probability Theory (i)

• Having a prior knowledge about the joint distribution of *x* and *y* in the form *m(x,y)* includes knowing the conditional distribution of *x* given *y*:

$$m(x\,|\,y) = \frac{m(x,y)}{\int dx\, m(x,y)} = \frac{m(x,y)}{m(y)}$$

• If we now obtain some piece of information about *y* e.g. *<C(y)>=d* we could proceed in two ways:

a) Start with the prior *m(y)* and the data constraint and obtain *p(y)*;

b) Apply the learning rule to the joint distribution, updating *m(x,y)* to *p(x,y)*.

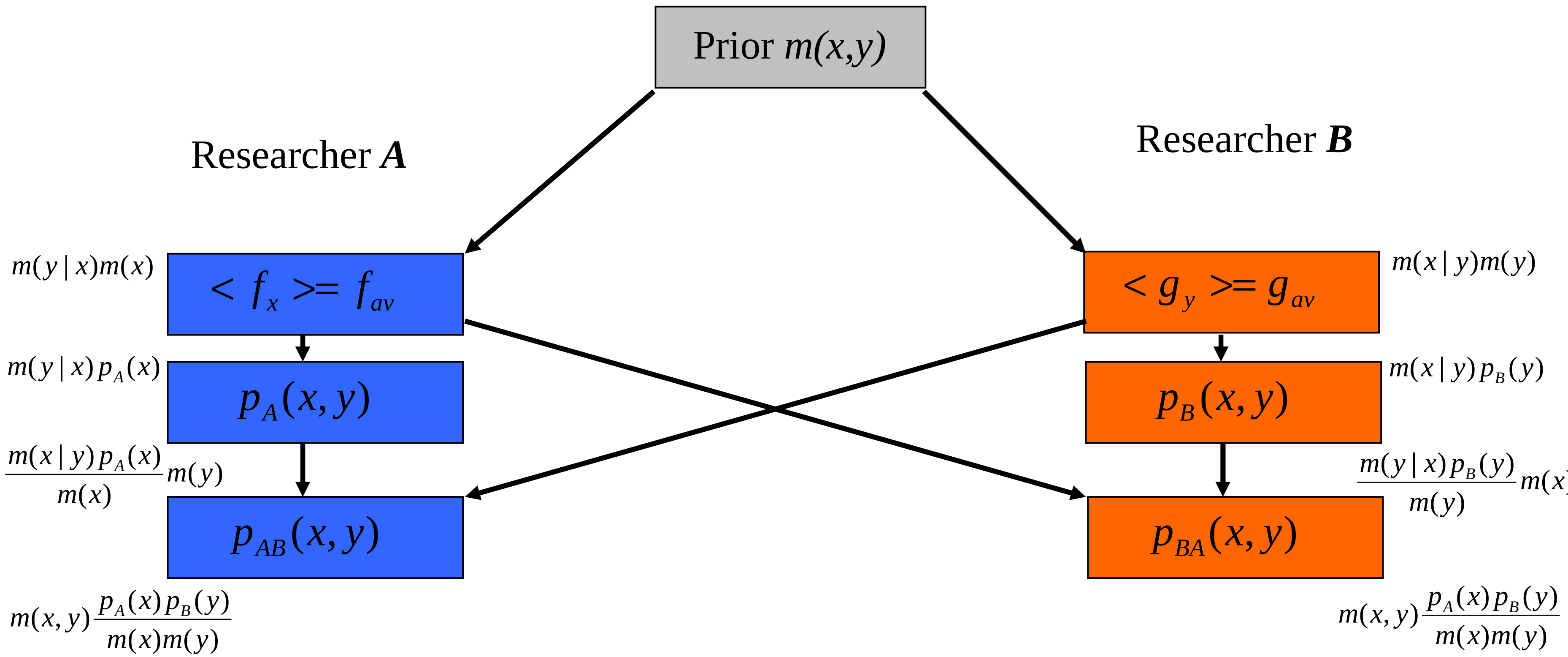
Consistency requires that the marginal of the updated joint distribution is the same as *p(y)*:

$$\int dx\, p(x,y) \overset{!}{=} p(y)$$

This is the same as requiring that the conditional distribution should remain unchanged:

$$p(x\,|\,y) = m(x\,|\,y) \qquad\qquad p(x,y) = m(x\,|\,y)\, p(y)$$

Consistency upon updating: Example 1



Consistency upon updating requires that

$$\boxed{p_{AB}(x,y) = p_{BA}(x,y)}$$

Uniqueness: Consistency with Probability Theory (ii)

- A sufficient (but not necessary) condition is the “strong additivity” of *S*:

$$S[p(x,y),m(x,y)] \overset{!}{=} S[p(y)\,|\,m(y)] + \int dy p(y) S[p(x\,|\,y), m(x\,|\,y)]$$

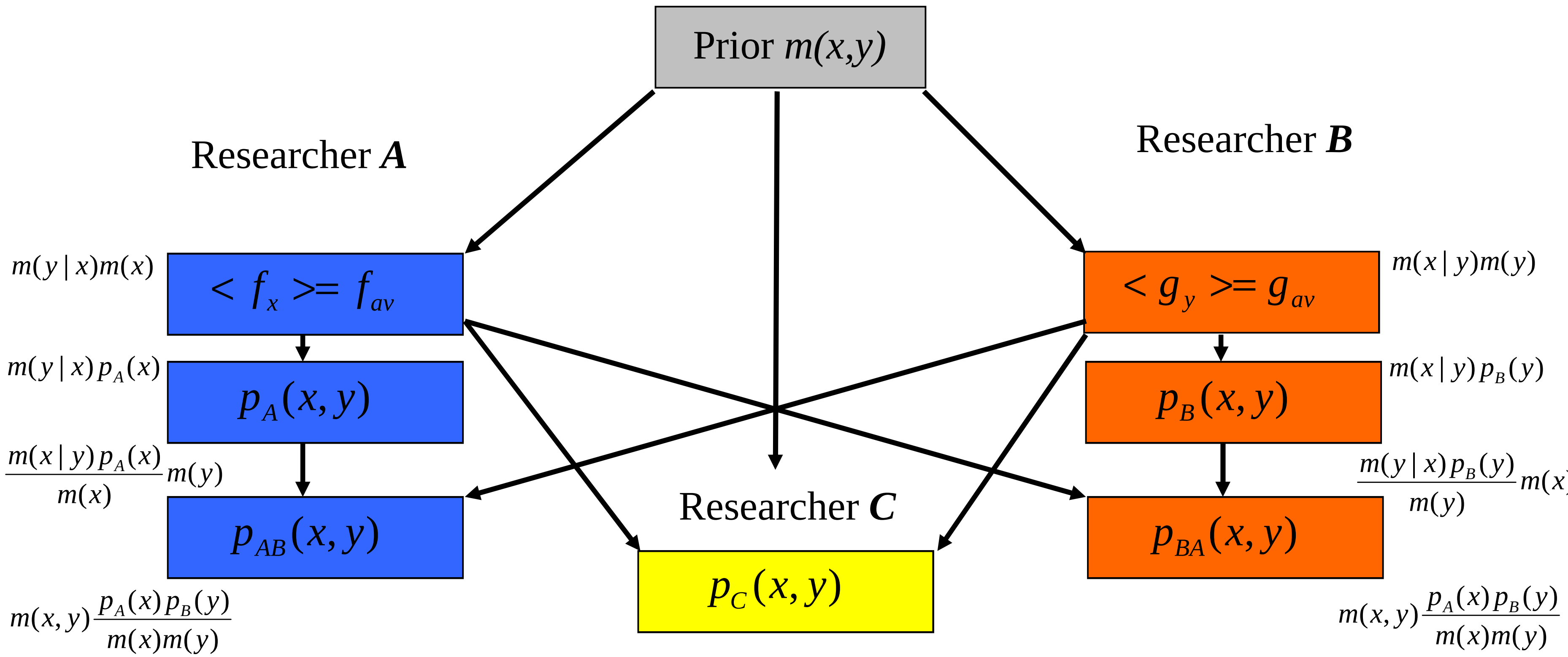
This is basically one of the Khinchin’s axioms discriminating Renyi’s $v=1$ against all other values, thus singling out the Shannon-Jaynes distance.

- A sufficient (and probably necessary) condition for the above is that the updating amounts to multiplying the prior by a function of the constraint:

$$p(x,y) = m(x,y) F(\lambda C(x,y))$$

This condition rules out the gradient-dependent terms (i.e. pins down the value of c_0 to zero), but does not restrict Renyi’s v, Idaho Accelerator Center

Consistency upon updating: Example 1



Consistency upon updating requires that

$$p_{AB}(x,y) = p_{BA}(x,y) = p_C(x,y)$$

Consistency upon updating: Example 1

$$\boxed{p_{AB}(x,y) = p_{BA}(x,y)}$$

$$p_{AB}(x,y) = p_{BA}(x,y) \quad \Longleftrightarrow \quad \begin{aligned} p_{AB}(x,y) &= m(x,y)F(\lambda_A f(x))F(\lambda_B g(y)) \\ p_{BA}(x,y) &= m(x,y)F(\lambda_B g(y))F(\lambda_A f(x)) \end{aligned}$$

Renyi with $c_0=0$: $$f(q,z) = \frac{\lambda C(x,y)}{\nu} \qquad p(x,y) = m(x,y)\left[\frac{\lambda C(x,y)}{\nu}\right]^{\frac{1}{\nu-1}}$$

$$\boxed{p_{AB}(x,y) = p_{BA}(x,y) = p_C(x,y)}$$

$$F(\lambda_A f(x))F(\lambda_B g(y)) = F(\lambda_A f(x) + \lambda_B g(y))$$

This excludes all $\nu \neq 1$ and leaves us with the Shannon-Jaynes distance!

Consistency upon updating: Example 2

$\{f_i\}\ i=1,N_1$ → $<f>=N_1^{-1}\sum_{i=1}^{N_1} f_i$ → $p_1(x)$

$<f>=w_1 N_1^{-1}\sum_{i=1}^{N_1} f_i + w_2 (N_2-N_1)^{-1}\sum_{i=N_1}^{N_2} f_i$ → $p_{12}(x)$

$w_{12}\in[0,1]$

$w_1+w_2=1$

$\{f_i\}\ i=N_1+1,N_2$ → $<f>=(N_2-N_1)^{-1}\sum_{i=N_1}^{N_2} f_i$ → $p_2(x)$

Consistency upon updating requires that

$$p_{12}(x)=w_1 p_1(x)+w_2 p_2(x)$$

Intermediate Summary

The Shannon-Jaynes relative entropy is singled out, after all, by consistency requirements even when dependence on the derivatives is allowed, as THE information distance measure to be used in a learning rule:

$$\int dx\, p(x) \ln \frac{p(x)}{m(x)} + \int dx\, p(x) [\lambda_0 + \lambda C(x)] \rightarrow \min$$

$$\lambda_0 : \int dx\, p(x) = 1 \quad \lambda : \int dx\, p(x) C(x) = d$$

In practical applications, however, one rarely has prior knowledge in the form of a prior distribution; it is more likely to be in the form of data. In such cases the updating rule is useless but one can still utilize the derived unique information distance for identifying least-informative priors.

Least-Informative Priors – Constructive Approach

- Identify a generic operation invoking "information loss";
- Notice that the information loss should be (at least for small information content) proportional to the information content unless negative information is tolerated;
- Look for the least-informative prior (under given data constraints) as the one least sensitive to the operation above. The sensitivity should be defined consistently with the adopted information distance.

- The generic operation is coarse-graining:

$$p_\sigma(x) = \frac{1}{\sigma}\int dx'\, p(x-x')\, f\left(\frac{x'}{\sigma}\right)$$

- The information distance is the Shannon-Jaynes one:

$$S[p, p_\sigma] = \int dx\, p \ln\frac{p(x)}{p_\sigma(x)} \xrightarrow{!} \min \quad \text{(subject to data constraints)}$$

Coarse-Graining = Adding Noise

- Coarse-graining is equivalent to adding noise: $p(x);\qquad x \rightarrow y = x + \sqrt{\sigma} z:\quad p(x) \rightarrow p(y,z)$

- "Distance" between the original and the "noisy" probability distributions:

$$S[p_0(y,z), p_\sigma(y,z)] = \underbrace{S[p_0(z), p_\sigma(z)]}_{0} + \int dz p_\sigma(z) S[p_0(y\,|\,z), p_\sigma(y\,|\,z)]$$

$$p_\sigma(y\,|\,z) = p(x + \sqrt{\sigma} z)$$

- For the Shannon-Jaynes distance:

$$S[p(x), p(x+\sqrt{\sigma}z)] = -\int dx p(x) \ln\left[1 + \sum_{n=1}^{\infty} \frac{\sigma^{n/2} z^n}{n!} \frac{1}{p(x)} \frac{d^n p(x)}{dx^n}\right] =$$

$$= z^2 \frac{\sigma}{2} \int dx \frac{1}{p(x)} \left[\frac{dp(x)}{dx}\right]^2 + O(\sigma^{3/2})$$

(just another instance of De Bruijn identity)

Essentially the same result holds for general Renyi ν as well as for symmetrized SJ distance or the symmetric Bernardo-Rueda "intrinsic discrepancy".

$$\boxed{S[p_0(y,z), p_\sigma(y,z)] = \frac{< z^2 >}{2} \sigma \int dx \frac{1}{p(x)} \left[\frac{dp(x)}{dx}\right]^2 + O(\sigma^2)}$$

Least-Informative Priors: Constructive MAXENT

Least-informative prior:

The probability density least-sensitive to coarse-graining subject to all available data constraints:

$$\boxed{\int dx\, p(x)\left[\frac{\nabla p(x)}{p(x)}\right]^2 + \lambda \int dx\, p(x) C(x) \rightarrow \min}$$

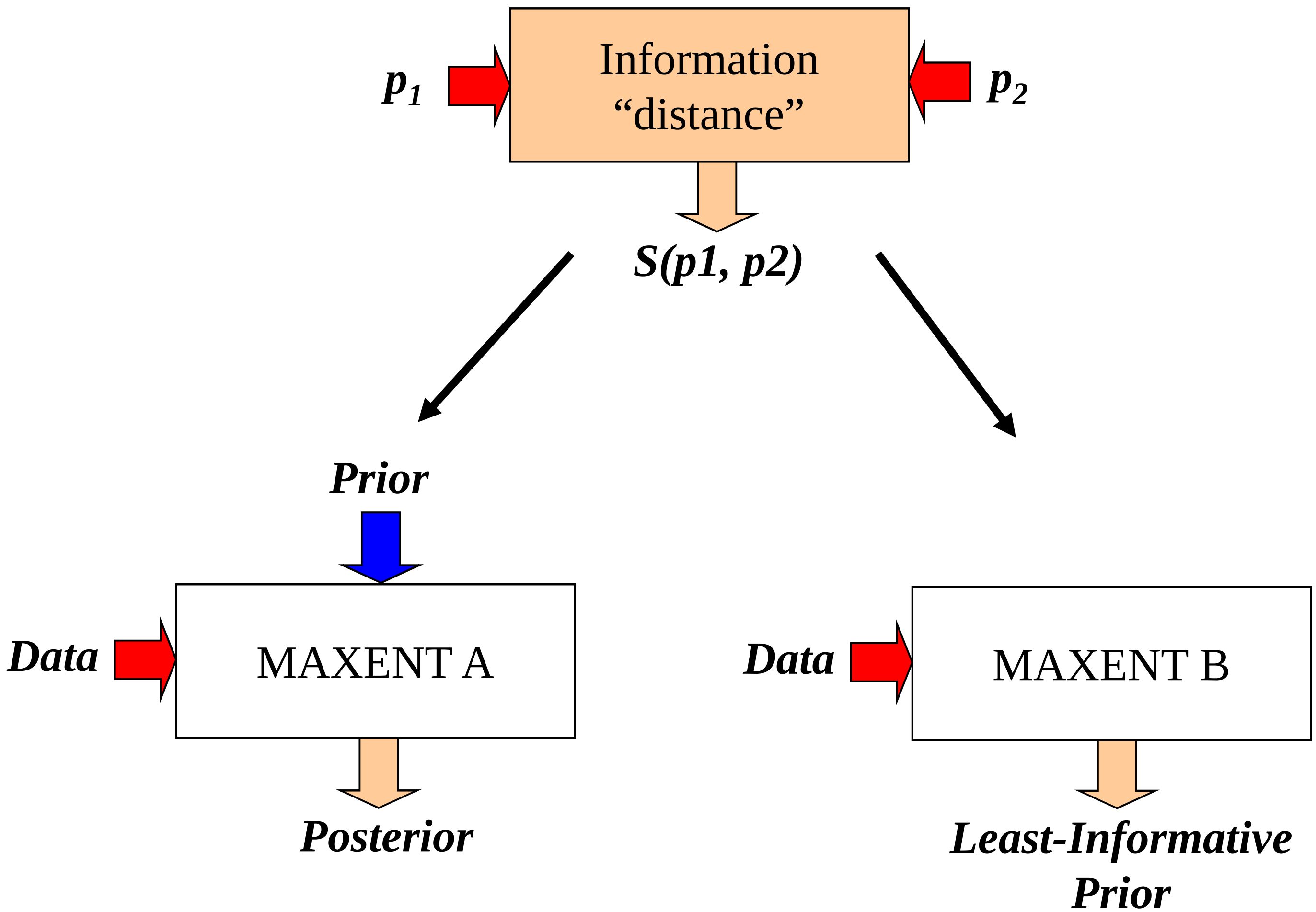
Putting here $p(x) = \psi^2(x)$ and varying ψ we obtain theEuler-Lagrange equation

$$\boxed{-\Delta\psi(x) + \frac{\lambda}{4} C(x)\psi(x) = 0}$$

$$\Delta\psi(x) - \frac{2m}{\hbar^2}[E - U(x)]\psi(x) = 0$$

This is a Schrodinger equation; non-relativistic Quantum Mechanics appears to be based on applying the constructive MAXENT to a system with given average kinetic energy, i.e. temperature!

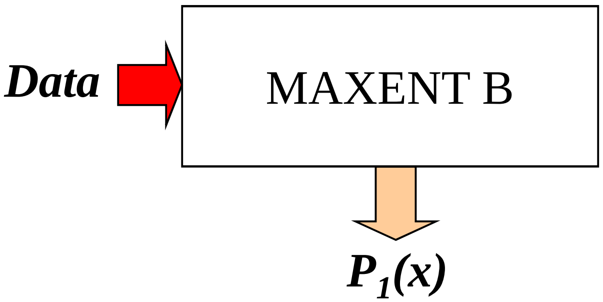
Least-Informative Priors: Consistency Condition

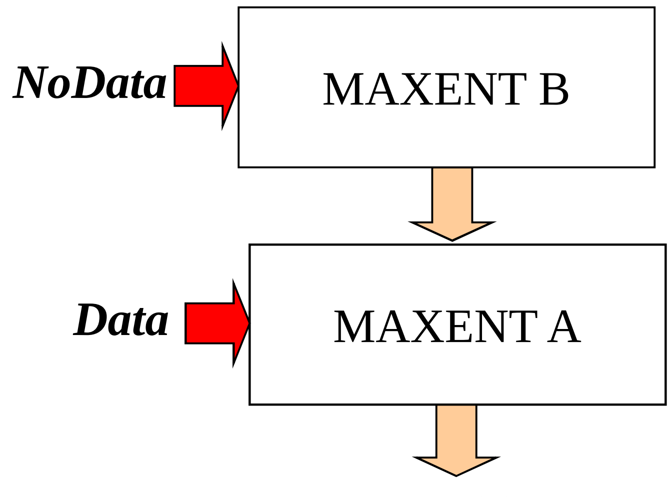


The posterior is the closest distribution to the prior subject to the data constraints. "Closest" is with regard to the information distance.

The LI prior is the one least sensitive to coarse-graining subject to the data constraints. "Least sensitive" is with regard to the information distance.

Least-Informative Priors: Consistency Condition





Consistency condition:

$$\boxed{P_1(x) = P_2(x)}$$

Checks o.k.? In some cases, maybe! In general – not!
MAXENT B with no data = uninformative prior=problem!

Conclusion: The MAXENT Approach

Updating rule:

$$m(x) \downarrow$$

$$C(x) \rightarrow \boxed{\text{MAXENT A}} \rightarrow p(x)$$

$$\delta_p\left[\int dxp\ln\frac{p}{m}+\lambda\int dxpC(x)\right]=0$$

$$p(x)=m(x)\exp[-1-\lambda C(x)]$$

Based on Shannon-Jaynes relative entropy

Constructive rule:

$$C(x) \rightarrow \boxed{\text{MAXENT B}} \rightarrow p(x)$$

$$\delta_p\left[\int dxp(\nabla\ln p)^2+\lambda\int dxpC(x)\right]=0$$

$$p(x)=\psi^2(x)$$

$$-\Delta\psi(x)+\lambda C(x)\psi(x)=0$$

Based on Fisher information for the "location" parameter as a measure for sensitivity to coarse-graining

Summary

- **Axiomatic characterization of both constructive and learning MAXENT is possible, based on the notion of information distance;**

- **The usual requirements for smoothness, consistency with probability theory and additivity narrow down the possible form of the information distance to Renyi – type one with an exponential Fisher-type modification factor;**

- **Additional requirement that different ways of feeding the same information into the method produce the same results pins down the value of the Renyi's parameter to 1 and excludes the derivative terms singling out the Shannon-Jaynes distance as the correct one;**

- **Since the information distance does not measure information content, I propose to consider the information content proportional to the sensitivity of the probability distribution to coarse-graining. If this is accepted, a unique constructive procedure for priors subject to available information results, involving the minimization of Fisher information under constraints.**

- **The constructive procedure, if taken seriously, has far reaching implications for the nature of those physics laws which can be formulated as variational principle. They may turn out to have very little to do with physics *per se*.**

Thoughts

If we agree to use probability distributions (wave functions) to encode all knowledge we have about a system, there is no way of treating new information in the form of a constraint in any exclusive way: once the PD gets updated all mean values it predicts are treated on the same footing. Then, obtaining new information basically means feeding in generally inconsistent (with what the PD already predicts for the particular observable) data. The behavior of the MAXENT makes sense – the rule gives preference to the most up-to-date “data”. This would be an insurmountable conceptual difficulty if we could actually obtain information in the form of mean values, which we can’t.

Fisher Distance Proper and Quantum Mechanics

QM connection: $p(x) = \psi^2(x) \quad \psi(x) = \sum \theta_i \phi_i(x) \qquad \int dx \phi_i(x)\phi_j(x) = \delta_{ij} \quad \sum \phi_i(x)\phi_i(x') = \delta(x - x')$

$$< \psi_1 | \psi_2 >= \int dx \psi_1(x)\psi_2(x) = \sum \theta_i^{(1)}\theta_i^{(2)}$$

$$(I_F)_{ij}[p] = \int dxp \frac{\partial \ln p}{\partial \theta_i}\frac{\partial \ln p}{\partial \theta_j} = 4\int dx \frac{\partial \psi}{\partial \theta_i}\frac{\partial \psi}{\partial \theta_j} = 4\int dx \phi_i \phi_j = 4\delta_{ij}$$

Distance based on Fisher metrics:

$$D_F^2(p_1, p_2) = 4\sum(\theta_i^{(1)} - \theta_i^{(2)})^2 =$$
$$= 8(1 - \sum \theta_i^{(1)}\theta_i^{(2)}) = 8(1 - < \psi_1 | \psi_2 >)$$

Minimizing the Fisher metrics distance = maximizing the QM overlap

$$\int dx \sqrt{p(x)m(x)} = \int dxp(x)\left[\frac{p(x)}{m(x)}\right]^{-1/2} \sim \exp[S_R^{\nu=1/2}(p,m)]$$

QM inference based on Hilbert space metrics (same as Fisher metrics) is equivalent to using MAXENT with ν=1/2 Renyi distance, and thus it is guaranteed to eventually produce inconsistencies, i.e. paradoxes!

Variation for General Functional Form

$$q(x) \equiv \ln \frac{p(x)}{m(x)} \qquad S[p] = g\left(\int dx p f(q, (\nabla q)^2)\right)$$

$$\delta \int dx p f(q, (\nabla q)^2) = \int dx \delta p \left[f + f_{,1}\right] + 2\int dx p f_{,2} \nabla q \cdot \delta \nabla q$$

$$p \delta \nabla q = \nabla \delta p - \delta p \nabla \ln p = \nabla \delta p - \delta p \nabla \ln q - p \nabla \ln m$$

$$\delta \int dx p f(q, (\nabla q)^2) = \int dx \delta p \left[f + f_{,1} - 2 f_{,2} \nabla q \cdot (\nabla \ln q + \nabla \ln m) - 2\nabla \cdot (f_{,2} \nabla q)\right]$$

$$\nabla \cdot (f_{,2} \nabla q) = f_{,2} \Delta q + (\nabla q)^2 f_{,21} + 2 f_{,22} \nabla q \cdot \nabla\nabla q \cdot \nabla q$$

$$\frac{\delta}{\delta p} \int dx p f(q, (\nabla q)^2) =$$

$$= f + f_{,1} - 2 f_{,2} \Delta q - 2\left(f_{,2} + q f_{,12}\right) \frac{(\nabla q)^2}{q} - 4 f_{,22} \nabla q \cdot \nabla\nabla q \cdot \nabla q - 2 f_{,2} \nabla \ln m \cdot \nabla q$$

$$\boxed{f + f_{,1} - 2 f_{,2} \Delta q - 2\left(f_{,2} + q f_{,12}\right) \frac{(\nabla q)^2}{q} - 4 f_{,22} \nabla q \cdot \nabla\nabla q \cdot \nabla q - 2 f_{,2} \nabla \ln m \cdot \nabla q = \lambda C(x)}$$