Information Costs and Sequential Information Sampling

Benjamin Hébert (Stanford)  Michael Woodford (Columbia)

November 27, 2018
Decision Maker (DM) chooses action \( a \in A \)

State of nature \( x \in X, \ |X| \geq |A| \)

Payoff \( u_{a,x} \) depends on action, state

But \( x \in X \) not known, prior \( q_0 \in \mathcal{P}(X) \)

DM can purchase a signal structure

- Signal alphabet \( S, \ |S| \times |X| \) matrix \( p = \{p_x \in \mathcal{P}(S)\}_{x \in X} \)
- Unconditional signal probability \( \pi \in \mathcal{P}(S) \)
- Bayes posteriors \( q_s \in \mathcal{P}(X) \) for each \( s \in S \)
The Optimization Problem

- Payoff from taking best action given beliefs:
  \[ \hat{u}(q) = \max_{a \in A} u_a^T \cdot q \]

- The problem: choose signal alphabet \( S \), signal structure \( p \), to maximize expected utility less cost \( C(p, q_0; S) \),
  \[ \sum_{s \in S} \pi_s \hat{u}(q_s) - \theta C(p, q_0; S), \]
  where \( \theta > 0 \) is positive constant.

- Sims 2003, many others: \( C(\cdot) \) is Shannon’s mutual information
Mutual Information

- MI related to how many bits required to transmit or store information
  - not obviously relevant for economic decisions
- MI has some undesirable properties:
  - Caplin et al. [2018], Caplin and Dean [2013], Dewan and Neligh [2017], Woodford [2012]
  - related to 1) curvature and 2) jumps in payoffs
  - but reasonable in, e.g., linear-quadratic-Gaussian settings (Sims [2010])
- Choice of cost function can matter:
  - Security design (Yang [2017]): what alternatives to consider?
  - Global games (Morris and Yang [2016]): what costs satisfy conditions?
  - Perceptual experiments (Dean and Neligh [2018])
- What we do: propose alternative that fixes these issues
  - and fits perceptual experiments of Dean and Neligh [2018]
An Example: Buying a Security

- Consider an example based on Yang [2017]
- DM is offered a security $s$ at price $K$
  - Security payoff $s_x$ depends on an asset value $x \in X \subset \mathbb{R}_+$
  - Actions $A = \{L, R\}$, accept ($L$, “like”) or reject ($R$)
  - Utility is $u_{R,x} = 0$, $u_{L,x} = s_x - K$
- DM chooses signal structure and updates beliefs before accept/reject
- Security $s$ could be almost any security
  - have to consider behavior for all securities to find optimal one
- example: “digital put”: $s_x = \begin{cases} \bar{s} & x \leq 10 \\ 0 & x > 10 \end{cases}$
Another Example: Perceptual Experiment

“Motion strength” experiment (Shadlen et al. [2007])
Field of dots on the screen, each moving left or right
\( x \in X \) is number of dots moving right
Actions \( A = \{ L, R \} \) picks direction of motion
\( u_{L,x} = 1(x \leq \frac{1}{2}|X|), \ u_{R,x} = 1(x > \frac{1}{2}|X|) \)
- Note similarity to “digital put” security example
Results with Mutual Information

Suppose $C(\cdot)$ is MI
Evidence from Perceptual Experiments
DDM Models

- How does the psychology and neuroscience literature fit this data?
- DDM: drift-diffusion models
- Brownian state, boundaries for left/right, drift towards correct answer
  - e.g. Fehr and Rangel [2011], Krajbich et al. [2014]
- Optimizing models: Fudenberg et al. [forthcoming], Moscarini and Smith [2001], Woodford [2014]
- Limitations: usually two actions/states, tricky to solve
- “Sequential Evidence Accumulation”
Introduction

Properties of a Good Cost Function

- Property 1: notion of states being easier or harder to discriminate
  - Shorthand: “perceptual distance” between states
  - Should address “digital option” issue
- Property 2: it should summarize sequential evidence accumulation
  - More accurately represent process of acquiring information
- These two properties are connected
  - We link costs of info. in sequential model to perceptual distance
Our Alternative

- We introduce the “neighborhood-based” cost functions
  - for the continuous state case, the “Fisher information” cost function
- We show that they satisfy both property #1 and property #2
  - MI satisfies property #2, but not #1 (see also Morris and Strack [2017])
- We show in applications that neighborhood-based cost functions fix some of the problems with mutual information, while making the same predictions in the LQG setting
Outline

1. Continuous Time Sequential Evidence Accumulation
   - Model derived from DT RI setting in Hébert and Woodford [2018]
   - alternative with jumps in appendix (Che and Mierendorff [2017], Zhong [2018])

2. Uniformly posterior-separable cost functions as solutions

3. The neighborhood based cost function

4. Applications
Sequential Evidence Accumulation

- States $X$, actions $A$, payoff $u_{a,x}$, function $\hat{u}(q)$ as in static model
- Accumulate information and then make a decision
- Beliefs $q_t$, where $t$ is time
- Cost of delay: $\kappa$ per unit time
- Choose optimal stopping time $\tau$
- Value function: $V(q_t) = \sup_{\ldots} E_t[\hat{u}(q_\tau) - \kappa \tau]$
Dynamics of Beliefs

- Beliefs are martingales:

\[ dq_t = \text{Diag}(q_t)\sigma_t dB_t, \]

- \text{Diag}(q_t) is diagonal matrix with \( q_t \) on diagonal
- \( B_t \) is an \( |X| \)-dimensional Brownian motion
- \( \sigma_t \) is our control variable (an \( |X| \times |X| \) matrix)
- Beliefs stay in simplex: \( q_T^T \sigma_t = \vec{0} \)
  - \( M(q_t) \): set of \( \sigma_t \) s.t. \( q_T^T \sigma_t = \vec{0} \)
Variance of beliefs is limited:

$$\frac{1}{2} tr[\sigma_t \sigma_t^T k(q_t)] \leq \chi$$

- $tr[\cdot]$ is the trace, $\chi > 0$ is a positive constant
- $k(q_t)$ is the "information cost matrix function"
  - having volatile beliefs means acquiring info rapidly
  - $k(q_t)$ describes the cost of a small amount of information
Info. Cost Matrix Function Properties

- $k(q_t)$ is PSD, null space is vectors constant in support of $q$
- Diagonal: how hard is it to learn about a state
- Off-diagonal: how hard to discriminate between states
  - more negative means harder to discriminate
- Example 1: Inverse Fisher Info. Matrix

$$k(q) = \begin{bmatrix}
q_1(1 - q_1) & -q_1 q_2 & \cdots & -q_1 q_{|X|} \\
-q_1 q_2 & q_2(1 - q_2) & \cdots & -q_2 q_{|X|} \\
\vdots & \vdots & \ddots & \vdots \\
-q_1 q_{|X|} & -q_2 q_{|X|} & \cdots & q_{|X|}(1 - q_{|X|})
\end{bmatrix}$$

- Note symmetry of off-diagonal elements
- This will be related to mutual information
Example 2: A “neighborhood-based” info cost matrix function

\[ k(q) = \begin{bmatrix}
\frac{q_1 q_2}{q_1 + q_2} & -\frac{q_1 q_2}{q_1 + q_2} & 0 & \cdots & 0 \\
-\frac{q_1 q_2}{q_1 + q_2} & \frac{q_1 q_2}{q_1 + q_2} + \frac{q_2 q_3}{q_2 + q_3} & -\frac{q_2 q_3}{q_2 + q_3} & \ddots & \vdots \\
0 & -\frac{q_2 q_3}{q_2 + q_3} & \ddots & \ddots & 0 \\
\vdots & \ddots & \ddots & \frac{q|x| - 1 q|x| - 2}{q|x| - 2 + q|x| - 1} + \frac{q|x| q|x| - 1}{q|x| - 1 + q|x|} & -\frac{q|x| - 1 q|x|}{q|x| + q|x| - 1} \\
0 & \cdots & 0 & -\frac{q|x| - 1 q|x|}{q|x| + q|x| - 1} & \frac{q|x| + q|x| - 1}{q|x| - 1 q|x|} \\
\end{bmatrix} \]

- Note “local” nature of costs: zeroes for non-adjacent states
- Meaning: hard to discriminate between adjacent states
Summary of CT Model

- Notation: Extend $V$ to measures, $V(\alpha q_t) = \alpha V(q_t)$
- Assuming twice-differentiability, HJB is

$$0 = \max_{\sigma \in M(q_t)} \frac{1}{2} \text{tr}[\sigma^T \text{Diag}(q_t)V_{qq}(q_t)\text{Diag}(q_t)\sigma]dt - \kappa dt,$$

subject to $\frac{1}{2} \text{tr}[\sigma^T k(q_t)\sigma] \leq \chi$.
- no-discounting key to tractability, decision times often short
- Result: constraint binds
Assumption

There exists an $H : \mathbb{R}^{|X|}_{+} \rightarrow \mathbb{R}$ s.t.

$$\text{Diag}(q_t)^{-1} k(q_t) \text{Diag}(q_t)^{-1} = H_{qq}(q_t)$$

- Covers inverse Fisher matrix, neighborhood-based cost function
- Does not cover constant $k(q)$ matrix
- “integrability” assumption, made for tractability
- $H(\cdot)$ is convex
Given any convex \( H \) function (i.e. from assumption 1), there is a “Bregman divergence,”

\[
D_H(q' \| q) = H(q') - H(q) - (q' - q)^T \cdot H_q(q).
\]

This defines a “uniformly posterior-separable cost function” (Caplin et al. [2018]),

\[
C_H(p, q_0; S) = \sum_{s \in S} \pi_s D_H(q_s \| q_0).
\]

MI: \( H \) is negative Shannon’s entropy, \( D_H \) is KL divergence, \( C_H \) is MI.
Result

Theorem

Under Assumption 1, the value function that solves the continuous time sequential evidence accumulation problem is

$$V(q_0) = \max_{\pi \in \mathcal{P}(A), \{q_a \in \mathcal{P}(X)\}_{a \in A}} \sum_{a \in A} \pi(a)(u_a^T \cdot q_a) - \theta \sum_{a \in A} \pi(a)D_H(q_a || q_0),$$

subject to the constraint that $\sum_{a \in A} \pi(a)q_a = q_0$, where $D_H$ is the Bregman divergence associated with the entropy function $H$ that is defined by Assumption 1, and $\theta = \frac{\kappa}{\chi}$.

There exist maximizers $\pi^*$ and $q_a^*$ such that $\pi^*$ is the unconditional probability, in the continuous time problem, of choosing a particular action, and $q_a^*$, for all $a$ such that $\pi^*(a) > 0$, is the unique belief the DM will hold when stopping and choosing that action.
“Uniformly posterior-separable” cost functions
- generate behavior in static RI model that is equivalent to behavior in sequential evidence accumulation problem
- Fit other evidence (Caplin et al. [2018])

UPS includes mutual information
- Associated $k(q_t)$ is the inverse Fisher info matrix
- but this $k(q_t)$ matrix lacks notion of “perceptual distance”

Next: connection between $k(q_t)$ and comparative statics
Information Costs and Comparative Statics

- FOC of Theorem 1, $A = \{L, R\}$:
  \[ u_L - u_R = \theta H_q(\pi^*_L q^*_L) - \theta H_q(q_0 - \pi^*_L q^*_L) \]

- Assume $\pi_L, q_L$ interior, $q_L \neq q_0$, consider perturbation to state $x \in X$ payoffs, $u_L(\epsilon) = u_L + \epsilon e_x$.

- Recall $H_{qq}(q) = \text{Diag}(q)^{-1} k(q) \text{Diag}(q)^{-1}$

- Define $H_{qq}(\pi^*_L q^*_L) + H_{qq}(q_0 - \pi^*_L q^*_L) = H_{diag} - H_{off}$. For any $x' \in X$,
  \[ e_{x'}^T \frac{d(\pi^*_L(\epsilon) q^*_L(\epsilon))}{d\epsilon} \bigg|_{\epsilon=0} = \theta^{-1} e_{x'}^T [H_{diag}^{-1} \sum_{j=0}^{\infty} (H_{diag}^{-1} H_{off})^j] e_x \]
Perceptual Distance and Complementarity

\[
e_{x'}^T d\left( \frac{\pi^*_L(\epsilon) q^*_L(\epsilon)}{d\epsilon} \right) \bigg|_{\epsilon=0} = \theta^{-1} e_{x'}^T \left[ H^{-1}_{\text{diag}} \sum_{j=0}^{\infty} \left( H^{-1}_{\text{diag}} H_{\text{off}} \right)^j \right] e_x
\]

- “zero-round” effect only if \( x' = x \)
- “first-round” effect if hard to discriminate between \( x', x \)
  - more negative \( k_{x, x'}(q) \) \( \Rightarrow \) more positive \( H_{\text{off}, x, x'} \)
  - MI: \( H_{\text{off}} \) constant for all \( x' \neq x \)
    - neighborhood example: only if \( x', x \) adjacent
- higher-round effects
We now introduce the “neighborhood-based” cost functions. These allow us to specify “neighborhoods” that describe which states are easier or harder to discriminate. The neighborhoods will determine the $k(q_t)$ matrix, which (by Assumption 1) determines the $H(q_t)$ function, and thus by Theorem 1 a Bregman divergence and UPS cost function.

Special case of particular interest: states ordered on a line. neighborhoods are pairs of adjacent states.
Neighborhoods

- Suppose $X$ is union of finite collection of neighborhoods $\{X_i\}_{i \in I}$
  - $\{X_i\}$ overlap, not disjoint sets, “connectedness” between any $x, x'$
  - $|X_i| \times |X|$ matrix $E_i$ selects states in $X_i$, $\bar{q}_i = \iota_{|X_i|} E_i q$

- We define the neighborhood-based information cost matrix function:

$$k_N(q; \rho) = \begin{cases} \sum_{i \in I} c_i \bar{q}_i |X_i|^{1-\rho} E_i^T (g^+(q_i))^{2-\rho} E_i & \rho \neq 2, \\ \sum_{i \in I} c_i \bar{q}_i |X_i|^{-1} E_i^T (I - q_i \iota^T)(I - \iota q_i^T)E_i & \rho = 2, \end{cases}$$

- $g^+(\cdot)$ is inverse Fisher matrix, $\rho$ controls curvature, $c_i$ weights

- Special case: $\rho = 1$, one neighborhood with all states, $k_N = g^+$
Generalized Entropy

- $H^{\text{Gen}}(q_i; \rho)$ is generalized entropy index of Shorrocks [1980] on the neighborhood $i \in I$, defined interior $q_i$:

$$
H^{\text{Gen}}(q_i; \rho) = \begin{cases} 
\frac{1}{|X_i|} \frac{1}{(\rho-2)(\rho-1)} \sum_{x \in X_i} \{(|X_i| e_x^T q_i)^{2-\rho} - 1\} & \rho \notin \{1, 2\} \\
- \frac{1}{|X_i|} \sum_{x \in X_i} \ln(e_x^T q_i) & \rho = 2 \\
\sum_{x \in X_i} e_x^T q_i \ln(e_x^T q_i) & \rho = 1.
\end{cases}
$$

- $\rho = 1$: negative of Shannon’s entropy
- $\rho > 1$: more curved, still minimized at uniform
Uniformly Posterior-Separable Neighborhood Cost

Lemma

The $H_N(q)$ function associated with $k_N(q; \rho)$ is

$$H_N(q; \rho) = \sum_{i \in I} c_i \bar{q}_i H^\text{Gen}(q_i; \rho),$$

and is defined on the boundary by continuity for $\rho < 2$ and as infinity for $\rho \geq 2$.

- Weighted (by $c_i$ weights and $\bar{q}_i$ probabilities) generalized entropy indices for each neighborhood
- Use of $\rho \neq 1$ inspired by Dean and Neligh [2018]
  - neighborhoods distinct concept from cost within each neighborhood
Uniformly Posterior-Separable Neighborhood Cost

Corollary

Consider a rational inattention problem with a neighborhood-based information-cost function, and let \( x, x' \) be two states with the property that (i) \( u_{a,x} = u_{a,x'} \) for all actions \( a \in A \), (ii) \( q_{0,x} = q_{0,x'} \), and (iii) the set of neighborhoods \( \{X_i\} \) such that \( x \in X_i \) is the same as the set such that \( x' \in X_i \). Then under the optimal policy, \( p^*_x = p^*_x' \). If \( \rho = 1 \), this result holds even if \( q_{0,x} \neq q_{0,x'} \).

- Recall example of perceptual experiment/digital put
  - \( u_{L,x} = u_{L,x'} \) if \( x, x' \) on same side of 10
Result with MI

Suppose $C(\cdot)$ is MI:
Suppose $\rho = 1$, neighborhoods are adjacent states, equally costly

- $c_i = 1$, $X_1 = \{x_1, x_2\}$, $X_2 = \{x_2, x_3\}$, ...

![Graph showing response probabilities with neighborhood cost function]
States on a Line

- Suppose neighborhoods are only consecutive states, equally costly
  \[ c_i = 1, \ X_1 = \{x_1, x_2\}, \ X_2 = \{x_2, x_3\}, \ldots \]
- Within each neighborhood, approximate \( H_{\text{Gen}}(q_i; \rho) \) near uniform:
  \[
  H_N(q; \rho) \approx \frac{1}{4} \sum_{j=0}^{|X|-1} \frac{((e_j^T - e_{j+1}^T)q)^2}{\frac{1}{2}(e_j^T + e_{j+1}^T)q}
  \]
  - same for all \( \rho \), exact for \( \rho = 0 \)
- In the limit as the number of states \(|X|\) goes to infinity:
  - \( H_N(q; 1) \) converges to the “Fisher information,”
  \[
  H_N(q; 1) \to \frac{1}{4} \int_{\text{supp}(q)} q(x)\left(\frac{q'(x)}{q(x)}\right)^2 dx
  \]
Continuous State RI Problem

\[
V_N(q_0) = \sup_{\pi \in \mathcal{P}(A), \{q_a \in \mathcal{P}_{\text{LipG}}\}} \sum_{a \in A} \pi(a) \int_{\text{supp}(q)} u_a(x) q_a(x) dx
\]

\[
- \frac{\theta}{4} \sum_{a \in A} \left\{ \pi(a) \int_{\text{supp}(q)} \frac{(q'_a(x))^2}{q_a(x)} dx \right\}
+ \frac{\theta}{4} \int_{\text{supp}(q)} \frac{(q'_0(x))^2}{q_0(x)} dx,
\]

subject to \( \sum_{a \in A} \pi(a) q_a(x) = q(x) \).

- \( \mathcal{P}_{\text{LipG}} \) is differentiable PDFs with Lipschitz-continuous derivatives
  - Assumption is \( q_0 \in \mathcal{P}_{\text{LipG}} \), result is \( q_a \in \mathcal{P}_{\text{LipG}} \)
- Can be written as choice of \( p_a(x) \) (prob. a given \( x \)) instead
- A single-parameter (\( \theta \)) cost function, just like MI, for use whenever states ordered on a line
  - Interpretation: Cramér-Rao related to estimating location parameter with given local uncertainty (Cover and Thomas [2012])
Applications

1. Security design (Yang [2017])
   - Security buyer example part of this problem
2. Perceptual Experiments (Dean and Neligh [2018])
   - Smooth response to discontinuous payoffs
3. Global Games (Morris and Yang [2016])
   - Uniqueness requires smooth response to discontinuous payoffs
4. Linear-Quadratic-Gaussian tracking problems (Sims [2010])
   - Fisher info. cost same as mutual information!
Security Design

- **Buyer’s problem:**
  \[
  V(q_0; s, K) = \max_{\pi_L \in [0,1], q_L, q_R \in \mathcal{P}(X)} \pi_L q_L^T (s - K \iota) \\
  - \theta \pi_L D_H(q_L \| q_0) - \theta (1 - \pi_L) D_H(q_R \| q_0),
  \]
  subject to \( \pi_L q_L + (1 - \pi_L) q_R = q_0 \).

- **Seller’s problem:**
  \[
  \max_{s, K \geq 0} \pi_L(s, K) q_L(s, K)^T (K \iota - \beta s)
  \]
  subject to limited liability: \( 0 \leq s_x \leq x \).

- \( \beta < 1 \) motivates trade.
Applications

\( H(\cdot) \) Functions Considered

1. Shannon’s entropy (MI): Yang [2017] proves debt is optimal
2. “weighted” Shannon’s entropy:
   \[ H_w(q) = \sum_{x \in \mathcal{X}} (e_x^T w)(e_x^T q) \ln\left( \frac{e_x^T q}{\ell^T q} \right), \]
3. Generalized entropy (\( \rho = 13 \), only one neighborhood)
4. Neighborhoods (ordered on a line, \( \rho = 13 \), \( c_i = 1 \))
Numerical experiment, $|X| = 21$, values 0 to 10

In Yang [2017], two cases: $\pi_L^* = 1$ or $\pi_L^* < 1$
  - Choose parameters so that $\pi_L^* < 1$

Choose weights and prior to illustrate differences:
  - $w = \frac{3}{2} + \frac{x}{10}$ (harder to learn about better states, leads to equity)
  - $q_0$ 50/50 mix of uniform and binomial (hump-shape, away from zero)
  - $\beta = \frac{1}{2}$ (far from 1 so that $w$ matters)

Choose $\theta$ for each to make securities similar size
  - $\theta$ not comparable across $H(\cdot)$ functions
Applications

Optimal Securities

\[ s(x) \]

- Selling Everything
- Shannon Entropy
- Weighted S. Entropy
- Generalized Entropy
- Neighborhood

\[ x \]
LQG Tracking Application

- Assume action space is real line
- Gaussian prior $N(\mu, \sigma^2)$, Fisher info. cost
- Goal is to track action with state
- Formulate problem as choice of $p(a|x)$:

$$\max_{p(a|x)} \int_{-\infty}^{\infty} q(x) \int_{-\infty}^{\infty} \left[ p(a|x)(a - x)^2 + \frac{\theta}{4} \frac{p_x(a|x)^2}{p(a|x)} \right] da dx$$
LQG Results

- If $\theta < 4\sigma^2$: optimal policy equivalent to receiving signal normal signal
  - $s = x + \epsilon$, $\epsilon \sim N(0, \nu^2)$
  - $a = (1 - \beta)\mu + \beta s$, $\beta = \frac{\sigma^2}{\sigma^2 + \nu^2}$
  - $\nu^2 = \sigma^2[2\sigma^2 \theta^{-\frac{1}{2}} - 1]^{-1}$

- If $\theta \geq 4\sigma^2$: optimal policy is to gather no information
  - $a = \mu$

- Results closely mirror LQG tracking with MI
Neighborhood-based cost functions can be used in place of mutual information in rational inattention problems

They capture notions of “perceptual distance” and summarize a sequential evidence accumulation process

They are particularly convenient when states are ordered on a line

In several applications (security design, lab experiments, global games) perceptual distance matters

- Having a concrete alternative should help this list grow

In linear-quadratic-Gaussian settings, predictions are the same as MI

Consider using them!


