Coarse Decision Making

Nabil Al-Najjar & Mallesh Pai

Northwestern University,
MEDS

Decentralization Conference
April 2009

To obtain a copy of this paper and other related work, visit:
http://www.kellogg.northwestern.edu/faculty/alnajjar/htm/index.htm
The 30 second ‘elevator pitch’

1. We provide a model of **coarse decision making**:
   “individuals choose not to optimize over all technologically and informationally feasible decision rules”

2. Coarse decision making is **behaviorally important**:
   - Simplicity-biased behavior
   - Heuristics, rules of thumb, categorization, linear orders, and satisficing
   - Concern for robustness

3. Our model gives insights **why coarse decision making works the way it does**, and not just rationalize its existence
Examples and Background

- **Psychology**: categorization, concepts, analogies ...
- **Style investing**: Sharpe 92, Barberis-Shleifer 03
- **Discrimination**: Fryer-Jackson 08 among many others
- **Organizations and corporate culture**: Kreps 90, Cremer-Garicano-Prat 07
- **Analogies, similarity**: Samuelson, Gilboa-Schmeidler..
- **Linear orders**: Rubinstein, G. Kalai..
Typical approaches

- Computational complexity
- Costly communication
- Costly introspection
- Behavioral biases

**Goal:** offer a unified explanation of CDM based on difficulties caused by statistical inference
Our methodology: *Back to the future.*

Savage (1951) wrote:

“The central problem of statistics is [...] to make reasonably secure statements on the basis of incomplete information.”

What applies to statisticians ought to apply just as well to economic actors.

Model decision makers as *frequentist statisticians* concerned about obtaining *robust, distribution-free* inferences
Basic idea

Two-stage decision problem:

1. **Select a decision frame, or model**: a set of contingent decision rules $\mathcal{F}$ from outcomes to actions

2. **Inference and decision making**: A sample is observed and a rule $f \in \mathcal{F}$ is selected as a function of the data.

A Bayesian will say: *But surely, you must be joking!*

Why would a decision maker in his right mind do this? ..willingly restricting his set of options?
Analogy with classical statistics: **OLS**

Typical “regression” question:

“I would like to fit a bunch of points with a curve”

My predicament as a classical statistician:

- The more freedom I allow myself, the better I can **fit**
- But the more freedom I have the more likely I will **over-fit**

Example of decision frame $\mathcal{F}$ in this case is: “All linear regression equations with a particular set of regressors”

then use OLS to select the best regression $\hat{f} \in \mathcal{F}$.
Two interpretations of our model

- As a model in K-T’s “heuristics & biases” tradition
  - What might explain the heuristics we actually observe?

- As normative model of rational decision making
Where we are at in this project

- Al-Najjar: “Decision Makers as Statisticians.. ” Ecma
- This paper
- Rationalization and framing
- Axiomatic foundations
1 Introduction

2 Inference
   - Model
   - Uniform learning

3 VC theory

4 Behavioral Consequences
   - Categorization
   - Linear Orders
   - Satisficing

5 Applications

6 Conclusions
Outcomes, decision rules and payoffs

- **Finite** space of observables $X$, actions $A$ and outcomes $Y$.

- The Decision Maker has a utility function
  \[ u : Y \times A \rightarrow [0, 1]. \]

- Observables $X$ do not directly influence payoff.

- A **Decision Rule** is a function from observables to action
  \[ f : X \rightarrow A \]

- Set of probability distributions $\mathcal{P} = \Delta(X \times Y)$

- Expected payoff given the true distribution:
  \[ E_{\mathcal{P}}(F) = E_{\mathcal{P}}u(x, y, f(x)). \]
The decision maker observes a sample:

\[ s^t = ((x_1, y_1), \ldots, (x_t, y_t)) \]

drawn i.i.d. from an unknown distribution \( P \).

Decision makers are *frequentist statisticians* who learn about performance of a \( f \) from the *empirical performance* \( \nu(s^t)(A) = \frac{\text{Number of observations in the sample in } A}{t} \).

The empirical performance of a rule \( f \) on the sample \( x^t \):

\[ E_{\nu(s^t)} f = \frac{1}{t} \sum_{i=1}^{t} u(y_i, f(x_i)). \]
Learning from i.i.d. Data

- $E_{\nu(s^t)} f$ is subject to sampling error; so the decision maker will also be concerned with the **empirical discrepancy**:

$$\Delta_t(f) = \sup_P \int_{s^t} |E_{\nu(s^t)} f - E_P f| \, dP^t,$$

- Frequentist “believes” that, with suitably large $t$, he is likely to observe a **representative sample**
  - Empirical frequencies are ‘close’ to true probabilities.

- Given $\epsilon$, one can find $t$ such that

$$\Delta_t(f) < \epsilon$$

for each and every $f$
Being a bad classical statistician

Here is how a naive statistician would proceed:

- **F** is the set of all feasible rules
- If there is enough data that
  \[ \sup_{f \in F} \Delta_t(f) < \epsilon \]
  ..then choose
  \[ f^*_s \in \arg\max_{f \in F} E_{\nu(s)} f \]
- because this will be \( \epsilon \) close to the true optimal choice
  \[ f^*_p \in \arg\max_{f \in F} E_P f \]

It is important to understand why this logic is flawed..
Let’s examine what this means more closely...

- Fix an event $A$
- The blue ball represents the set of representative samples
  - *i.e.*, ones where the empirical freq. of $A$ is a good estimate of the true probability of $A$
- The weak LLN says that the set of samples representative for $A$ has high probability

$X^t$ with distribution $P^t$

$$Good_A = \{x^t: |v(A;x^t) - P(A)| < \epsilon \}$$
The set of samples representative for \( B \) also has high probability.

..but the set of samples representative for both \( A \) and \( B \) is potentially smaller.
Add yet another event $C$ and the set of samples representative for any *single event* still has high probability.

..but the set of samples representative for *all three events* is the intersection.

This can shrink quickly as one adds more and more events..
The sets of representative samples may stack on top of each other.

In this case, one can add events while ensuring the set of jointly representative samples still has high probability.

Whether this happens or not is key...
The need for uniform learning

I can guarantee $\sup_{f \in F} \Delta_t(f) < \epsilon$ for each $f$ with ‘reasonable’ amount of data.

..but if I choose $f^*_st \in \arg\max_{f \in F} E_{\nu(st)} f$

..I ensure this is a good choice, I need to know:

$f^*_st \approx \arg\max_{f \in F} E_P f$

That is, I need

$$\sup_P \int_{s^t} \sup_{f \in F} \left| E_{\nu(st)} f - E_P f \right| dP^t < \epsilon.$$
Uniform learning

Contrast with non-uniform learning

\[ \mathcal{F} \subset \mathbf{F} \text{ is } \epsilon\text{-learnable with data } t \text{ if} \]

\[ \sup_{P} \int_{s^t} \sup_{f \in \mathcal{F}} \left| E_{\nu(s^t)} f - E_{P} f \right| dP^t < \epsilon. \]

\[ \Delta_t(\mathcal{F}) \]

Contrast this with non-uniform (WLLN) learning one rule at a time:

\[ \sup_{f \in \mathcal{F}} \sup_{P} \int_{s^t} \left| E_{\nu(s^t)} f - E_{P} f \right| dP^t < \epsilon. \]

\[ \Delta_t(f) \]
So many theories, so few facts

Uniform and non-uniform learning correspond to different statistical experiments:

- **Bounding**
  \[
  \sup_P \int_{s^t} \left| E_{\nu(s^t)} f - E_P f \right| dP^t \\
  \Delta_t(f)
  \]
  corresponds to an experiment with fresh sample taken for each rule \( f \)

- \[
  \sup_P \int_{s^t} \sup_{f \in \mathcal{F}} \left| E_{\nu(s^t)} f - E_P f \right| dP^t < \epsilon.
  \]
  \[\Delta_t(\mathcal{F})\]
  corresponds to an experiment where the decision maker gets one shot at the data
A *model or decision frame* is a pair \((\mathcal{F}, \epsilon)\) where \(\mathcal{F} \subseteq \mathbf{F}\) and \(\epsilon > 0\) such that:

1. \(\Delta_t(\mathcal{F}) \leq \epsilon\);
2. Given the data \(s^t\), he selects the empirically best performing rule in \(\mathcal{F}\):

\[
f_{s^t}^{\mathcal{F}} \in \arg\max_{f \in \mathcal{F}} E_{\nu(s^t)} f.
\]
Coarse decision making

Unless $F$ is small, e.g., $X$, $A$ are small (urns and coins)

ensuring that

$$\Delta_t(F) < \epsilon$$

would require obscene amounts of data... orders of magnitudes larger than what is needed for $\sup_{f \in F} \Delta_t(f) < \epsilon$

So when the set of rules to choose from is large relative to available data, $F$ is not a appropriate decision frame.

The solution is Coarse decision making: restrict to $F \subsetneq F$

This may look like so-called “bounded rationality” but it is not!
Example: *So many theories, so few facts*

Best known bounds to evaluate all rules $\mathcal{F}$ within $\epsilon = 0.01$ accuracy

<table>
<thead>
<tr>
<th>$#X$</th>
<th>$\epsilon$</th>
<th># of observations needed</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.01</td>
<td>27,188,099 observations</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>53,804,950 observations</td>
</tr>
</tbody>
</table>

These are, respectively, approximately 1000 and 2000 times the amount of data needed for learning any single rule.
Consider the outcome space:

\[(\text{Individual attributes}) \times (\text{diet}) \times (\text{health conseq.})\]

Say there are 20 binary individual attributes (weight, age .. )

20 relevant binary attributes of diet

10 relevant binary attributes of health consequences

Then there are \(2^{50}\) outcomes

The \textit{minimum} amount of data needed to evaluate the probabilities of all events within 0.01-confidence is:

\[7,036,874,417,766,400 \text{ OR } 7.04 \times 10^{15}\]
Fitting vs. overfitting

We can write:

$$\Delta_t(\mathcal{F}) = \sup_P \left[ (E_P f^*_P - \max_{f \in \mathcal{F}} E_P f) + \int_{s^t} (\max_{f \in \mathcal{F}} E_P f - E_P f_{x^t}) dP^t \right].$$

A: measures fit
B: measures over-fit

A model $\mathcal{F}$ with small $\Delta_t(\mathcal{F})$ balances two conflicting criteria:

1. Term A: Fit improves as $\mathcal{F}$ becomes ‘large.’
2. Term B: But as $\mathcal{F}$ becomes ‘too large,’ the selected rule $f_{x^t}$ will tend to track the data too closely and thus over-fit.
Figure: $F'$ leads to worse fit than $F$, but has smaller over-fit
Contrast with Bayesian Decision Making

- prior belief $\pi$ on $\Delta(X)$
- $\pi$ is updated via Bayesian rule

Think of our model as *descriptive*: just look at
- “99%” of all applied work, in all fields
- how you and your applied colleagues learn from data
- how you teach statistics to students

But there may also be *normative reasons* why decision makers shun Bayesianism. For instance:

“Unfortunately, in high-dimensional problems, arbitrary details of the prior can really matter; indeed, the prior can swamp the data, no matter how much data you have.” (Diaconis and Freedman, 1986, p. 15)
1. Introduction
2. Inference
   - Model
   - Uniform learning
3. VC theory
4. Behavioral Consequences
   - Categorization
   - Linear Orders
   - Satisficing
5. Applications
6. Conclusions
Vapnik-Chervonenkis theory

Published in 1971 in English; introduced first in Russian a decade earlier

This is a major statistical tool used in non-parametric estimation, pattern recognition, statistical learning theory...

Provides a *uniform law of large numbers* for a class of events as a function of a combinatorial property called the *VC-dimension*

Historically, it is a massive generalization of *the Glivenko-Cantelli Theorem*

Ties in closely with the theories of *empirical processes* and *large deviations*
Of course, you ARE familiar with an example of learning uniformly over a family of events: “The empirical distributions converge uniformly to the true distribution almost surely.”

- “The Fundamental Theorem of Statistics”
  - Of course, Frequentist Statistics
- The Glivenko-Cantelli Theorem
- Generalized to arbitrary families of events by Vapnik and Chervonenkis.
Over-fitting as failure of uniform learning

Typical “regression” question:

“I would like to fit a bunch of points with a curve”

My predicament as a frequentist is:

- The more freedom I allow myself in choosing the curve, the better I can fit the sample
- But the more freedom I have the more likely I will over-fit
Over-fitting as failure of uniform learning

So what makes a family of functions simple or complex?

Is the following one-parameter family simple or complex?

\[ \{ n : \sin nx \} \]

.. and what does this mean in the first place? This family can be made to fit any set of data

(Graphs taken from Harman&Kulkarni’s book, which provides excellent discussion)
Formal definition

Definitions

- \( C \) shatters a subset \( \{x_1, \ldots, x_l\} \subset X \) if

\[
2^{\{x_1, \ldots, x_l\}} = \{\{x_1, \ldots, x_l\} \cap A : A \in C}\.
\]

- \( V_C \) is the largest integer \( l \) for which there is a subset of size \( l \) that can be shattered by \( C \).

VC Theorem

A class of sets is uniformly learnable iff it has a finite VC-dimension.
The Glivenko-Cantelli Theorem:

- $X = [0, 1]$;
- $\Delta(X)$ set of all probability distributions on Borel sets
- The Glivenko-Cantelli Theorem: The empirical distributions converge uniformly to the true distribution.
So why is learning so easy for the Glivenko-Cantelli class of events?

Because all you need for distribution functions is:

\[ C = \{ [0, t], \ t \in [0, 1] \text{ and their complements } \} \]

But: \( V_C = 2 \)

Implication: linear ordering are inherently appealing from a learning standpoint.
Example 2

The algebra generated by the Glivenko-Cantelli class:

- \( X = [0, 1] \);
- \( \mathcal{C} \) is the Glivenko-Cantelli class of half intervals
- \( \tilde{\mathcal{C}} \) is the algebra generated by \( \mathcal{C} \).
- \( V_{\tilde{\mathcal{C}}} = \infty \)

VC Theorem \( \implies \) Uniform learning is impossible.

Implication: Closing under algebraic operations is not innocuous when learning is taken seriously.
Example 3: *Orthogonal lenses*

- $X \times Y \equiv [0, 1] \times [0, 1]$;
- $C_1 = \text{all subsets of the form } [0, t] \times Y \text{ and their complements}$.
- $C_2 = \text{all subsets of the form } X \times [0, t] \text{ and their complements}$.
- $V_{C_1} = V_{C_2} = 2$
- Define $C = C_1 \cup C_2$
- Then $V_C = 3 \implies \text{Learning is harder}$.
- In fact, $V_{\mathcal{H}} = 3$, where $\mathcal{H}$ is half-spaces in $R^2$

**Implication:** Combining “models” is not innocuous when learning from finite data is to be taken seriously.
1 Introduction

2 Inference
   - Model
   - Uniform learning

3 VC theory

4 Behavioral Consequences
   - Categorization
   - Linear Orders
   - Satisficing

5 Applications

6 Conclusions
Central to cognitive psychology

appears in several economic models

Given our formalism, a categorization-based model or partition model consists of:

1. A categorization function \( \kappa : X \rightarrow \{1, 2, \ldots, K\} \).
2. The set of decision rules:

\[
\mathcal{F}_\kappa = \{ f \mid f = g \circ \kappa, \text{ for some } g : \{1, 2, \ldots, K\} \rightarrow A \}.
\]
Theorem 1

For every $t$ and $\epsilon > 0$, there exists an integer $k^+$, depending only on $\epsilon$ and the amount of available data $t$, such that for every categorization function $\kappa$, with number of categories $K$,

$$\Delta_t(\mathcal{F}_\kappa) < \epsilon \implies K \leq k^+.$$

For every integer $k^- \leq \#X$ there exists $T$ such that for every $t \geq T$ there is a categorization rule $\kappa$ with

$$K = k^- \quad \text{and} \quad \Delta_t(\mathcal{F}_\kappa) \leq \epsilon.$$
If there are no constraints on the amount of available data, no coarse categorization arises. The decision maker can simply treat each singleton \( \{x\} \) as separate category (thus, setting \( k^- = \#X \)), in which case \( F_\kappa \) coincides with the set of all rules \( F \), and still ensure that \( \Delta_t(F) \) is small.

The theorem has a bite when data is 'scarce'. In particular, when \( k^+ \ll \#X \), the decision maker must coarsely categorize. When data is relatively scarce, an overly fine categorization will result in overfitting. A decision maker concerned with robustness will, as a result, choose a coarser categorization.
The bulk of the existing literature has viewed coarse decision making as an expression of bounded rationality.

The amount of data he has available limit him to choosing partitional rule with a small number of partitions.

This is done as function of the amount of data available, with categories getting finer there more data is available.

there is no *direct* connection between coarseness of categorization to things like the monetary stakes agents have in the outcome, their depth of introspection or contemplation costs.
Another heuristic that decision makers seem to use is organizing the space of observables according to some linear order, e.g.:

- In Management: balanced scorecards (Kaplan and Norton (92,96)).
- In Investing: ranking companies by P/E, Price-to-book value etc.
- In language: Rubinstein (96,00), Kalai (03)
Formally, in our model: A linear attribute model $(v, w)$ is a pair of functions:

$$v : X \rightarrow \mathcal{R}^n \text{ and } w : A \rightarrow \mathcal{R},$$

and the collection of all monotone decision rules:

$$\mathcal{F}_{v, w} = \left\{ f \mid v(x') \geq v(x) \implies w(f(x')) \geq w(f(x)) \right\}.$$

In other words, the decision maker orders both the space of observables and the space of actions, and considers rules that take a ‘larger’ action on a ‘larger’ observable.
Theorem 2

For every $t$ and $\epsilon > 0$, there is an integer $n^+(\epsilon, t)$ such that for any linear attribute model $(v, w)$, $\Delta_t(F_{v,w}) < \epsilon$, implies

$$n \leq n^+(\epsilon, t).$$

The interpretation is similar to Theorem 1. If data is scarce, then a decision maker concerned with overfitting must organize the observables along a small number of dimensions.
Simon (55) proposed the idea of *satisficing* whereby a decision maker uses a plan which, while suboptimal, represents an attempt to do ‘reasonably well.’ He proposes computational complexity and cost of information gathering as possible motivations for this behavior.

We study similar behavior, where a decision maker worried about overfitting *coarsens* the set of actions $A$ available to him.
Satisficing (continued)

Formally, suppose:

\[ A = Y = \left\{ 0, \frac{1}{k}, \frac{2}{k}, \ldots, 1 \right\}, \]

and the payoff function given by the usual distance
\[ u(y, a) = -|y - a|. \]

We consider a linear attribute model, with a one-dimensional
linear order \( v \) on \( X \), and the standard ordering on \( Y \).
\[ \mathcal{F} = \{ f \mid v(x') \geq v(x) \implies f(x') \geq f(x) \}. \]

If decision maker considers a coarser set of actions \( A' \subset A \),
\[ |A'| = k' < k. \]
\[ \mathcal{F}_{A'} = \{ f \mid f : X \to A'; v(x') \geq v(x) \implies f(x') \geq f(x) \}. \]
Satisficing (contd.)

Theorem 3

For every $t$ and $\epsilon > 0$, there is an integer $k^+ (\epsilon, t)$ such that for any $k$, and any satisficed model $\mathcal{F}_{A'}$ ($|A'| \leq k$): $\Delta_t (\mathcal{F}_{A'}) < \epsilon$ only if:

$$|A'| \leq k^+ (\epsilon, t).$$

In particular, for $k$ large enough, $\Delta_t (\mathcal{F}) > \epsilon$.

In other words, a decision maker with limited data will prefer to consider a smaller set of actions, and compute the best plan with respect to those, to prevent overfitting.

For example consider a firm that is considering contingent production plans, and can make anywhere up to a million units. It may be better served considering production plans that produce lots of (say) 5000 units.
1. Introduction

2. Inference
   • Model
   • Uniform learning

3. VC theory

4. Behavioral Consequences
   • Categorization
   • Linear Orders
   • Satisficing

5. Applications

6. Conclusions
Kreps 90 views “corporate culture” as a principle that facilitates coordination and learning when it is difficult to specify everything contractually.

He adds:

“Consistency and simplicity being virtues, the culture/principle will reign even when it is not first best . . . will be taken into areas where it serves no purpose except to communicate or reinforce itself.”
large finite space of business situations \( X = \{x_1, \ldots, x_N\} \)
each arises with equal probability
given \( x \in X \), one of two actions \( \{0, 1\} \) can be taken, \( A^*(y) \) is principal's preferred action.
Two agents
If both agents take action \( A^*(y) \), the principal gets a payoff of 1, otherwise 0
Agents do not know principal's preference, see past \( t \) past business situations and (correct) actions for them.
They both have models which are partitions \( C_1, C_2 \) of \( X \).
Proposition 1

Suppose the principal can change his preferred action for any business situation \( x \), but if he does then he only gets a benefit \( \alpha < 1 \) if agents co-ordinate correctly on it. There exist environments, i.e. a preferred action for the principal \( A^* \), partitions for the two agents, \( C_1 \) and \( C_2 \), and a distribution over the business situations \( \pi \); such that for any \( \alpha > 0 \), the principal would prefer to change his preferred actions to an \( A' : X \to \{0, 1\} \) where \( A' \) is measurable with respect to \( C_1 \land C_2 \).

In other words, the principal, even at a cost, would prefer to ‘water down’ his preferences so that they are simple enough for both agents to learn.
Proposition 2

Suppose the environment is as described above. Further, suppose the principal can costlessly refine the partitions employed by the two agents to $C^*$, the finest possible partition. Fix an environment i.e. $\pi$, the partitions used by the two agents $C_1$ and $C_2$, and the preferences of the principal $A^*$. There exist environments such that there exists an integer $n^-(t)$ which depends on the amount of data available, such that if the number of business situations $N \geq n^-$, the principal would strictly prefer not to refine the partitions.

In other words, the principal may not want to refine his agents’ partitions, even if that was costless, since it would lead his agents to overfit.
Thirty years ago, Lucas wrote:

“To the journalist, each year brings unprecedented new phenomena, calling for unprecedented new theories (where ‘theory’ amounts to a description of the new phenomena together with the assertion that they are new).”

Like Lucas, we believe that: “it is in our interest to take exactly the opposite viewpoint.”

There is no way a unified model would ever fit data better than a collection of disparate models, each tailor made to fit a particular instance of the problem.

But the latter would be at best a descriptive account, a first pass at organizing raw evidence, not an explanation of anything.
Concluding remarks

Private information matters... but not everything reduces to an informational problem

*Understanding “biases” and “bounded rationality:”* some behavior may look anomalous or irrational (e.g., categorization) may be a good response to statistical complexity and learning problems

*Understanding Diversity* as choice of different frames

*Framing, persuasion, and rationalization...*