The Game of Twenty Questions with noisy answers. Applications to Fast face detection, micro-surgical tool tracking and electron microscopy

Graduate Summer School: Computer Vision
July 22 - August 9, 2013

Bruno M. Jedynak

Dept. of Applied Mathematics and Statistics
Johns Hopkins University

July 31, 2013
The game of 20 questions

$\mathcal{X}$ is the parameter space ($\{1, \ldots, 1,000,000\}, \{A, B\}, \text{unit interval}, \text{unit square}, \text{object pose}$).

$X_1 \in \mathcal{X}, X_2 \in \mathcal{X}, \ldots, X_K \in \mathcal{X}$ are the parameters, or targets, or object poses.

Ask $N$ questions, $0 \leq n \leq N - 1$: select $A_n \subset \mathcal{X}$.

The answer is $Z_n = 1_{X_1 \in A_n} + \ldots + 1_{X_K \in A_n}$ is the number of targets within $A_n$.

$Z_n$ is not observed.

Instead, we observe $Y_{n+1}$, (a noisy version of $Z_n$)

$$Y_{n+1} = \begin{cases} \sim f_1 & \text{if } Z_n = 1 \\ \sim f_0 & \text{if } Z_n = 0 \end{cases}$$

where $f_0$ and $f_1$ are probability distributions.

The goal is to choose $N$ questions such that $(X_1, \ldots, X_K)$ can be estimated as accurately as possible after observing the answers.
Past and current research projects (applications)

- Road tracking [1996] (Donald Geman, BJ);
- Face detection [2010], surgical tool tracking [2012], electron microscopy [2013] (BJ, Raphael Sznitman);
- Table setting analysis [current] (Donald Geman, Yoruk Erdem, Ehsan Jahangiri, BJ, Laurent Younes);
- Root finding using noisy measurements [current] (Peter Frazier, Shane Henderson, BJ);
- Stochastic simulations: screening [current] (Peter Frazier, BJ);
- Experiments in visual perception [current] (Jonathan Flombaum, Heeyeon Im, BJ)
Road Tracking from Remote Sensing Images
Face detection

\[
\frac{7150}{577729} = 0.012
\]

\[
\frac{16006}{1000000} = 0.016
\]
Tool tracking, electron microscopy
Table setting
Past and current research projects (theory)

- 20 questions as a model for perception [1995] (Donald, Geman, BJ)
- Metric loss functions [current] (Peter Frazier, Shane Henderson, Rolf Waeber, Avner Bar-Hen, BJ)
- Multiple objects. Entropy Loss. [current] (Peter Frazier, Weidong Han, BJ)
Today’s talk:

- Specify the mathematical game
- Review a Frequentist result
- Bayesian pont of view with entropy loss
- Play the game
- Probabilistic bisection policy
- Dyadic policy in 1 dimension
- Detecting multiple objects simultaneously
- Application to face detection
- Application to tool tracking
References


4. Sznitman, Raphael; Richa, Rogerio; Taylor, Russell; jedynak, bruno; Hager, Gregory D. ”Unified detection and tracking of instruments during retinal microsurgery”, *IEEE PAMI*, 2012 Oct 1

Unidimensional Binary Symmetric Noise

\(X \in [0; 1]\) is the parameter of interest. (1D case)

Ask questions: At time \(n, 0 \leq n \leq N - 1\), select \(x_n, 0 \leq x_n \leq 1\)

\(Z_n = 1_{X \leq x_n}\) is the "true" answer

\(Z_n\) is not observed

Instead, we observe \(Y_{n+1}\), a noisy version of \(Z_n\)

\[
Y_{n+1} = \begin{cases} 
Z_n & \text{with probability } 1 - \epsilon \\
1 - Z_n & \text{with probability } \epsilon 
\end{cases}
\]

with \(0 \leq \epsilon \leq 1\)

\(Y_{n+1}\) is Binary and the noise is symmetrical
Frequentist analysis

Choose $N$ questions $x_0, \ldots, x_{N-1}$ and propose an estimator $\hat{X}_N$ of $X$ in order to minimize

$$\sup_{X \in [0;1]} E|\hat{X}_N - X|$$

**Valid (or adapted) policy:** For each time $i$, the sample locations $x_i$ depends only on the available information at time $i$.

**Non-Adaptive policy:** Valid policies for which the sample locations $x_i$ can be chosen simultaneously.
Frequentist analysis

noseless case $\epsilon = 0$

1. restricted to non-adaptive policies,

$$\sup_{X \in [0;1]} E |\hat{X}_n - X| \leq \frac{1}{2(n+1)}$$

achieved by choosing $x_0, \ldots, x_{N-1}$ regularly spaced over $[0;1]$

$$\left\{ \frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1} \right\}$$

2. however,

$$\sup_{X \in [0;1]} E |\hat{X}_n - X| \leq \frac{1}{2n+1}$$

achieved by choosing the dichotomy policy which is adaptive.
Frequentist analysis

**noisy case** \( 0 < \epsilon < 1 \)

\[
\sup_{X \in [0;1]} E|\hat{X}_n - X| \leq 2\left(\frac{1}{2} + \sqrt{\epsilon(1 - \epsilon)}\right)^{n/2}
\]

using the **probabilistic bisection policy**. [Please wait a few slides :-)]

Exponential error decay behavior

Is this policy optimal for this criterium?

Ref: *Active Learning and Sampling*, Chapter 8, Rui Castro and Robert Nowak.
Bayesian analysis

$X$ is random, with density $p_0$ over $[0; 1]$
Consider more general questions of the form $X \in A$? where $A$
is a Lebesgue measurable subset of $[0; 1]$
More general answers. $f_0$ and $f_1$ are point mass functions or
densities

$$Y_{n+1} \sim \begin{cases} f_1 & \text{if } X \in A_n \\ f_0 & \text{if } X \notin A_n \end{cases}$$

Previously $A_n = [0; x_n]$, $f_1$ is $Bernoulli(1 - \epsilon)$ and $f_0$ is
$Bernoulli(\epsilon)$
Bayes rule

If at time $n$, the history of answers is $B_n = \{ Y_1 = y_1, \ldots, Y_n = y_n \}$, the posterior is $p_n$, the question is "$X \in A_n$?" and the answer is $Y_{n+1} = y_{n+1}$ then $p_{n+1}$ is the conditional distribution of $X$ at time $n+1$. Using Bayes rule,

$$p_{n+1}(x) = p(x|Y_{n+1} = y_{n+1}, B_n)$$

$$\propto P(Y_{n+1} = y_{n+1}|X = x, B_n)p(x|B_n)$$

$$\propto P(Y_{n+1} = y_{n+1}|X = x)p_n(x)$$

$$\propto p_n(x) \begin{cases} f_1(y_{n+1}) & \text{if } x \in A_n \\ f_0(y_{n+1}) & \text{if } x \notin A_n \end{cases}$$
Controlled Markov chain

As questions are asked and answered, the density of $X$ "evolves" becoming $p_1, p_2, \ldots$ successive posterior densities.

At time $n$,

The state is the density $p_n$.

The control is $A_n$. It affects the transition probability from $p_n$ to $p_{n+1}$.

The Markov property comes from the fact that the noise is memoryless.

The functional of interest for today is the Shannon Differential Entropy

$$H(p_n) = -\int_0^1 p_n(x) \log_2 p_n(x) \, dx$$

If $p_n(x) = \frac{1}{b-a}, a \leq x \leq b$, then $2^{H(p_n)} = b - a$
Let’s play the game ...

Define the **value function** as

\[ V(p, n) = \inf_{\pi} E^{\pi}[H(p_N)|p_n = p], n = 0, \ldots, N \]

where the infimum is taken over all valid policies \( \pi \).

Let’s play
Bellman optimality principle

Define the **value function** as

\[ V(p, n) = \inf_{\pi} \mathbb{E}^\pi [H(p_N) | p_n = p], \quad n = 0, \ldots, N \]

where the infimum is taken over all valid policies \( \pi \).

**Principle of optimality:** ”An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.” Richard Bellman.

**Bellman recursion**

\[ V(p, n) = \inf_{A_n} \mathbb{E} [V(p_{n+1}, n + 1) | A_n, p_n = p], \quad n = 0, \ldots, N \]

and a policy which chooses an \( x_n \) attaining the minimum above is optimal.
Greedy policy

Consider minimizing the value function at time $N-1$.

$$V(p, N-1) = \inf_{A_{N-1}} E[V(p_N, N)|A_{N-1}, p_{N-1} = p]$$
$$= \inf_{A_{N-1}} E[H(p_N)|A_{N-1}, p_{N-1} = p]$$
$$= \inf_{A_{N-1}} (H(p_{N-1}) - I(X, Y_N|A_{N-1}, p_{N-1}))$$
$$= H(p_{N-1}) - \sup_{A_{N-1}} I(X, Y_N|A_{N-1}, p_{N-1})$$

where $I(X, Y_N|A_{N-1}, p_{N-1})$ denotes the **Mutual Information** between $X$ and $Y_N$ when the density of $X$ is $p_{N-1}$ and the control is $A_{N-1}$
notate \( p(A) = \int_A p(x)dx \) and \( u = p_{N-1}(A_{N-1}) \)

\[
I(X, Y_N|A_{N-1}, p_{N-1}) = H(Y_N|A_{N-1}, p_{N-1}) - H(Y_N|X, A_{N-1}, p_{N-1})
\]
\[
= H(uf_1 + (1 - u)f_0) - uH(f_1) - (1 - u)H(f_0)
\]
\[
= \phi(u)
\]

\( \phi \) is concave, \( \phi(0) = \phi(1) = 0 \) and

The channel capacity is

\[
C = C(f_0, f_1) = \max_u \phi(u)
\]
Optimal policy

**Theorem:** The policy for which $A_n$ is such that

$$p_n(A_n) = u^* = \arg \max_u \phi(u)$$

is optimal and the value function is

$$V(p_n, n) = H(p_n) - (N - n)C$$

Proof: check that this policy verifies Bellman’s recursion. The main point is that the **expected** gain in Entropy $C$ realized at each step is independent of the state $p_n$.

When $\phi(u) = \phi(1 - u)$ then $u^* = 1/2$.

When moreover $A_n = [0; x_n]$, $x_n$ is the **median** of $p_n$. This policy is the **probabilistic bisection policy** used in the frequentist analysis.
Simulation of the probabilistic bisection policy

The process $H(\rho_n)$ for the binary symmetric channel. $\rho_0$ is Uniform([0; 1]). **Left:** $\epsilon = 0.2$ $C = 0.28$ **Right:** $\epsilon = 0.4$ $C = 0.03$
Illustration of the dyadic policy when $p_0$ is uniform on $[0, 1]$ and $u^* = 5/8$. The prior is displayed on top. Below, the sets $A_{n,k}$ are illustrated for $n = 0, 1, 2$. Each question $A_n$ is the union of the dark grey subsets $A_{n,k}$ for that value of $n$. 
Simulation of the dyadic policy

The process $H(p_n)$ $p_0$ is Uniform([0; 1]). **Left:** Binary symmetric channel $\epsilon = 0.2$ $C = 0.28$ **Right:** Normal channel $C = 0.47$
Properties of the dyadic policy

\[ H(p_{n+1}) - H(p_n) = -D \left( B \left( \frac{u^* f_1(Y_{n+1})}{u^* f_1(Y_{n+1}) + (1 - u^*) f_0(Y_{n+1})} \right), B(u^*) \right) \]

(1)

where \( Y_n \) is a sequence of i.i.d. random variables with pmf or density the mixture \( u^* f_1 + (1 - u^*) f_0 \)

\[ \frac{H(p_n)}{n} \rightarrow -C \text{ a.s.} \]

(2)

and

\[ \frac{H(p_n) + nC}{\sqrt{n}} \rightarrow N(0, \sigma^2) \text{ in distribution,} \]

(3)
Playing in 2 dimensions

\[ X^* = (X_1^*, X_2^*). \text{ Pitfall with minimizing } E[H(p_N)] \]

Ex: \( X^* \sim U([0; s_1] \times [0; s_2]) \), then \( H(X^*) = \log(s_1) + \log(s_2) \)
which can be arbitrarily small with, say, \( s_1 = 1 \).
Instead, notate \( H_1(p_N) = H(\int p_N(.., u_2)du_2) \), similarly for \( H_2(p_N) \)

\[ \inf_{\pi} E^\pi[\max(H_1(p_N), H_2(p_N))|p_0 = p] \]

Optimal policy seems out of reach. Instead,

\[ V(p) = \inf_{\pi} \lim_{N \to +\infty} \frac{1}{N} E^\pi[\max(H_1(p_N), H_2(p_N))|p_0 = p] \]
Playing in 2 dimensions with questions on the marginals

For further simplification, we assume that questions concern only one coordinate. That is, the sets queried are either of type 1, \( A_n = B \times \mathbb{R} \) where \( B \) is a finite union of intervals of \( \mathbb{R} \), or alternatively of type 2, \( A_n = \mathbb{R} \times B \). In each case, we assume that the response passes through a memoryless noisy channel with densities \( f_0^{(1)} \) and \( f_1^{(1)} \) for questions of type 1, and \( f_0^{(2)} \) and \( f_1^{(2)} \) for questions of type 2. Let \( C_1 \) and \( C_2 \) be the channel capacities for questions of type 1 and 2 respectively. We also assume that \( p_0 \) is a product of its marginals. This guarantees that \( p_n \) for all \( n > 0 \) remains a product of its marginals and that only one marginal distribution is modified at each point in time.
Playing in 2 dimensions with questions on the marginals

The following policy:
At step $n$, choose the type of question at random, choosing type 1 with probability $\frac{C_2}{C_1 + C_2}$ and type 2 with probability $\frac{C_1}{C_1 + C_2}$.
Then, in the dimension chosen, choose the subset to be queried according to the 1-dimensional dyadic policy.
Is optimal. Moreover, it verifies a CLT:
$$\lim_{n \to \infty} \frac{1}{\sqrt{n}} \left[ \max(H_1(p_n), H_2(p_n)) + \frac{C_1 C_2}{C_1 + C_2} n \right] = D$$
$$\max \left( \sigma_1 \sqrt{C_2} Z_1, \sigma_2 \sqrt{C_1} Z_2 \right) \frac{1}{\sqrt{C_1 + C_2}}.$$

Here, $Z_1$ and $Z_2$ are independent standard normal random variables, and $\sigma_i^2$ is the variance of the increment of $H_i(p_{n+1}) - H_i(p_n)$ when measuring type $i$. 
Character localization

Figure: From left to right: Example of an image containing the character “T”. Examples of subset-based questions. In each image, we show the queried region by the gray area.
Character localization

Figure: Pixel Reordering: (top) Example images from the test set. (bottom) Corresponding $\ell$-images. Dark regions indicate pixels more likely to contain the character, while light regions are less likely.
Character localization

Figure: Noise-free evaluations and convergence in entropy. (a) The distribution of number of noise-free evaluations needed to locate the target character. (b) Plot of $H(p_n)/n$ as a function of $n$. Each line corresponds to one image, with $H(p_n)/n$ plotted over $n = 1, \ldots, 16$. 
Central Limit Theorem: (a) Distribution of \( \frac{H(p_n) - (H(p_0) - nC)}{\sqrt{n}} \), with mean -0.01. The distribution is close to Gaussian as the Q-Q plot (b) shows.
Detecting two objects

Assume two targets $X = \{X_1, X_2\}$, are independent. We also assume that they have the same prior distribution, notated $p_0$. For specificity, let’s assume that both $X_1$ and $X_2$ belong to the interval $[0; 1]$. A series of $N$ questions are asked to locate $X_1, X_2$. The first question is indexed by the set $A_0$. The first answer is

$$Y_1 = 1_{A_0}(X_1) + 1_{A_0}(X_2)$$

The $n^{th}$ question is indexed by $A_{n-1}$ and has answer

$$Y_n = 1_{A_{n-1}}(X_1) + 1_{A_{n-1}}(X_2)$$
Policies

We define the value function in the usual way

\[ V(p, n) = \inf_\pi E^\pi [H(p_N)|p_n = p], \quad n = 0, \ldots, N \]  

(4)

where \( p_n \) is the posterior distribution over \( \{X_1, X_2\} \) after observing the answers to the first \( n \) questions. We also define the greedy policy \( \pi_G \). It is a sequential policy which we define iteratively.

\[ A_0 = \arg \min_A E[H(p_1)|p_0, A_0 = A] \]  

(5)

If the first \( n \) questions are indexed by \( A_0, \ldots, A_{n-1} \) and the answers are \( Y_1, \ldots, Y_n \), then \( A_n \) is chosen such that

\[ A_n = \arg \min_A E[H(p_{n+1})|p_n, A_n = A] \]  

(6)

Finally, we note \( \pi_D \) the diadic policy.
Result for the noiseless case

\[ H(p) - \log_2(3)N < V(p, 0) = \inf_{\pi} E^{\pi}[H(p_N)|p_0 = p] \]
\[ V(p, 0) \leq E^{\pi G}[H(p_N)|p_0 = p] \]
\[ E^{\pi G}[H(p_N)|p_0 = p] \leq E^{\pi D}[H(p_N)|p_0 = p] \]
\[ E^{\pi D}[H(p_N)|p_0 = p] = H(p) - 1.5N \]

Interpretation: Learning 20 bits from each object requires only about \( \frac{40}{1.5} \sim 27 \) questions, compared to 40 if we were to learn first about \( X_1 \) and then about \( X_2 \).
Multiple objects. No noise.
From Idealistic Setting to Computer Vision

- Finite number of identifiable poses (i.e. finite number of images)
- Finite set of questions
- Very strong Classifiers available (i.e. Oracles)
- Only available when evaluating very small set of poses (e.g. single pixel).
- Can create weak classifiers (i.e. providing noisy answers for a collection of poses)
Face Localization and Detection

Let us find the center of a face in an image:

- Let \( X^* = (X_1^*, X_2^*) \) be a discrete random variable that defines the face center.
- Let \( p_0 \sim U([0, M] \times [0, N]) \)

---

1 Sznitman, Jedynak. **Active Testing for face localization and detection.** *PAMI*, 2010.
Let $\Lambda$ be a regular decomposition of the pose space into quadrants, such that

$$\Lambda = \{\Lambda_{i,j}, i = 0, \ldots, d, j = 0, \ldots, 4^{i-1} - 1\}$$
Face Questions

- $K = 29$ questions available at each node

Search space queried: $\Lambda_{i,j}$

Proportion of edges in region

Proportion of oriented edges in region

Viola Jones face detector
Active Testing Algorithm

1. Initialize node to query: $i = 0, j = 0$
2. Initialize question type: $k = 0$
3. Repeat
   3.1 Test: $y = X_{i,j}^k$
   3.2 Update $p_{t+1}(\cdot)$ from $y$ and $p_t(\cdot)$
   3.3 Choose next Question:
   
   $$\{i, j, k\} = \arg\max_{i,j,k} \text{MI}(i, j, k)$$

4. Until $H(p_{t+1}) < \epsilon$ or a fixed number of iterations.
MIT+CMU Face Dataset:

7150 / 577729 = 0.012

16006 / 100000 = 0.16

13391 / 307200 = 0.043

32324 / 881600 = 0.036

5905 / 426429 = 0.013

18167 / 372960 = 0.048
MIT+CMU: Performance and Iterations

![Graph showing performance and iterations](image-url)
Application: 
*Instrument detection and Tracking in Retinal Microsurgery*
Application:
*Instrument detection and Tracking in Retinal Microsurgery*
Let us find the tool pose in an image:\footnote{Sznitman et al. \textit{Unified Detection and Tracking in Retinal Microsurgery.} \textit{MICCAI}, 2011.}:

- Let $X^* = (X_1^*, X_2^*, X_3^*)$ be a discrete random variable that defines the tool pose.
- Let the space of possible tool locations be:
  $$S = [0, P] \times [-\pi/2, \pi/2] \times [\delta, L]$$
- Let $p_0 \sim U(S \cup \{\square\})$
Let us find the tool pose in an image \(^1\):

- Let \(X^* = (X_1^*, X_2^*, X_3^*)\) be a discrete random variable that defines the tool pose.

- Let the space of possible tool locations be:
  \[
  \mathcal{S} = [0, P] \times [-\pi/2, \pi/2] \times [\delta, L]
  \]

- Let \(p_0 \sim U(\mathcal{S} \cup \{\Box\})\)

- More complicated density: need a way to organize the search space for efficiency.

---

Search Space Decomposition

Let \( \Lambda \) be a regular decomposition of the pose space, such that

\[
\Lambda = \{ \Lambda_{i,j}, \ i = 0, \ldots, d, \ j = 0, \ldots, 2^{i-1} - 1 \}
\]

- Will represent \( p_n \) via \( \Lambda \).
At each node $\Lambda_{i,j}$, can evaluate a question type: $k = 1, \ldots, \mathcal{K}$
At each node $\Lambda_{i,j}$, can evaluate a question type: $k = 1, \ldots, K$

A question $X^k_{i,j}$ asks: “is $X^* \in \Lambda_{i,j}$” by computing a function $k$ of the image:

$$X^k_{i,j} : I_{\Lambda_{i,j}} \rightarrow R$$
At each node $\Lambda_{i,j}$, can evaluate a question type: $k = 1, \ldots, K$

A question $X^k_{i,j}$ asks: “is $X^* \in \Lambda_{i,j}$” by computing a function $k$ of the image:

$$X^k_{i,j} : I_{\Lambda_{i,j}} \mapsto \mathbb{R}$$

Answer $Y^k_{i,j}$ is random,

$$Y^k_{i,j} = \begin{cases} 
    f_1(\cdot; i, j) & \text{if } X^* \in \Lambda_{i,j} \\
    f_0(\cdot; i, j) & \text{if } X^* \notin \Lambda_{i,j}
\end{cases}$$

$(f_1, f_0)$ are estimated from labeled training data.
Noisy Tool Questions:
Noisy Tool Questions:
Noisy Tool Questions:
Noisy and Oracle Tool Question:

Oracle Questions
Active Testing Algorithm

1. Initialize node to query: \( i = 0, j = 0 \)
2. Initialize question type: \( k = 0 \)
3. Repeat
   3.1 Test: \( y = X_{i,j}^k \)
   3.2 Update \( p_{t+1}(\cdot) \) from \( y \) and \( p_t(\cdot) \)
   3.3 Choose next Question:
   \[
   \{i, j, k\} = \arg \max_{i,j,k} \text{MI}(i, j, k)
   \]
4. Until \( H(p_{t+1}) < \epsilon \) or a fixed number of iterations.
Tool Tracking by Active Testing Filtering

- Given an image sequence, $\mathcal{I} = (I^1, \ldots, I^T)$ and a tool dynamics model $P(X_t^*|X_{t-1}^*)$, we can perform Bayesian Filtering:

1. Initialize: $p_0(X^*)$

2. Repeat
   
   2.1 $P_t(X^*|\mathcal{I}^{t-1}) = \int P(X_t^*|X_{t-1}^*)p_{t-1}(X^*)dX^{t-1}$
   
   2.2 $P_t(X^*|\mathcal{I}^t) = \text{ActiveTesting}(I^t, P_t(X^*|\mathcal{I}^{t-1}))$
The "20 questions with noise" model offers a framework for "machine perception". It is amenable to mathematical analysis through the use of information theory, control theory and probabilities.