Strong Entropy Concentration, Game Theory and Algorithmic Randomness

Peter Grünwald
CWI and EURANDOM
www.cwi.nl/~pdg

Overview

1. Strong Entropy Concentration
   - The Maximum Entropy Principle
   - Jaynes' Concentration Phenomenon
   - Cover/Campenhout's Conditional limit theorem
   - The Strong Concentration Phenomenon
2. Applications
   - Universal Models (MDL)
   - Game Theory / Log-Loss Prediction
   - Algorithmic Randomness / General Prediction

Setting

$\mathcal{X}$	Sample Space (finite, or countable, or $\mathbb{R}^m$)
$E_P[\phi(X)] = t$	'Constraint' for distributions over $\mathcal{X}$, where $\phi(X) = (\phi_1(X), \ldots, \phi_k(X))$
ϕ_i	random variable • 'lattice type' (if $\mathcal{X}$ finite/countable) • continuous (if $\mathcal{X}$ real-valued)
$\mathbf{H}$	Entropy

Maximum Entropy Principle

Jaynes 1957

Suppose we only know that

$$X \sim P \; ; \; E_P[\phi(X)] = t$$

We are asked to make probabilistic predictions/ decisions about X

According to 'MaxEnt', we should predict using the $\tilde{P}$ that maximizes entropy under the constraint:

$$\tilde{P} = \arg\max_{P: E_P[\phi(X)]=t} \mathbf{H}(P)$$

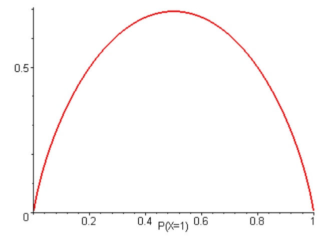
$$\tilde{P} = \arg\max_{P: E_P[\phi(X)]=t} \mathbf{H}(P)$$

- where, if $\mathcal{X}$ is finite,

$$\mathbf{H}(P) := E_P[-\ln P(X)] = -\sum_{x\in\mathcal{X}} P(x)\ln P(x).$$

- Under mild conditions on $\phi(X)$ and t, a unique MaxEnt $\tilde{P}$ is guaranteed to exist.

Example 1: if there are no constraints, then $\tilde{P}$ is uniform



MaxEnt generalizes Laplace's (1812) Principle of Indifference

Example 2: Brandeis Dice
(Jaynes 1962)

$$\mathcal{X} = \{1,2,3,4,5,6\}$$
$$E_P[X] = 4.5$$
$$\tilde{P}(X=x) = \frac{1}{Z(\beta)}e^{-\beta x}$$
$$Z(\beta) = \sum_{x\in\mathcal{X}} e^{-\beta x}$$
$$\beta = -0.345\ldots$$

Example 2: Brandeis Dice,continued
(Jaynes 1962)

In *practice*, given $X_1, X_2, \ldots, X_n$

Observe *empirical averages of some function(s) of* X :

$$\frac{1}{n}\sum_{i=1}^{n}\phi(X_i) = t$$

in dice case:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = 4.5$$

Motivation

Rule of Thumb: as symmetric, uniform as possible

Prime Motivation: the MaxEnt distribution for a constraint is the least committal, most `random', most *inherently uncertain* distribution, making the smallest number of additional assumptions beyond what is known etc.

Concentration Phenomenon

- In what sense is $\tilde{P}$ `most random distribution'?
- Let $\mathcal{X}$ be finite. Jaynes' Concentration Phenomenon says that

Nearly all sequences satisfying the constraint have empirical *frequencies* extremely close to Maximum Entropy *probabilities.*

Concentration Phenomenon

$\mathbb{P}^{(n)}(x)$ empirical frequency of $x \in \mathcal{X}$ in $(x_1, \ldots, x_n)$

$\mathcal{C}^{(n)} \equiv \{(x_1, \ldots, x_n) \in \mathcal{X}^n \ : \ |\frac{1}{n} \sum_{i=1}^{n} \phi(x_i) = t\}$

For all $\epsilon > 0$ there exists $c_\epsilon > 0$ such that

$$\frac{\#\{(x_1,\ldots,x_n) \in \mathcal{C}^n \ : \ \exists x \in \mathcal{X} \ |\mathbb{P}^{(n)}(x) - \tilde{P}(x)| > \epsilon\}}{\#(\mathcal{C}^n)} = O(e^{-c_\epsilon n})$$

Nearly all sequences satisfying the constraint have empirical *frequencies* extremely close to Maximum Entropy *probabilities.*

Concentration Phenomenon

Dice Example:

Sequences consisting of 50% 4's and 50% 5's ($\mathbb{P}^{(n)}(4) = \mathbb{P}^{(n)}(5) = 0.5$) satisfy the constraint but are **extremely** rare!

Minimum Relative Entropy Principle

GIVEN a *prior* distribution Q over $\mathcal{X}$ and a constraint

$$E_P[\phi(X)] = t$$

Among all distributions satisfying the constraint, choose the one `closest' to Q in relative entropy sense:

$$\tilde{P} = \arg \inf_{P : E_P[\phi(X)] = t} D(P \| Q)$$

Concentration and Conditioning

- If Q uniform, then MinRelEnt becomes MaxEnt
- Concentration phenomenon can be restated as:

For all $\epsilon > 0$ there exists $c_\epsilon > 0$ such that

$$Q^n(\text{there exists } x \in \mathcal{X} : |\mathbb{P}^{(n)}(x) - \tilde{P}(x)| > \epsilon \mid \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) = t) \leq O(e^{-c_\epsilon n})$$

Concentration and Conditioning

- If Q uniform, then MinRelEnt becomes MaxEnt
- Concentration phenomenon can be restated as:

For all $\epsilon > 0$ there exists $c_\epsilon > 0$ such that

$$Q^n(\text{there exists } x \in \mathcal{X} : |\mathbb{P}^{(n)}(x) - \tilde{P}(x)| > \epsilon \mid \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) = t) \leq O(e^{-c_\epsilon n})$$

Note, by Chernoff bounds:

$$\tilde{P}^n(\text{there exists } x \in \mathcal{X} : |\mathbb{P}^{(n)}(x) - \tilde{P}(x)| > \epsilon) \leq O(e^{-c_\epsilon n})$$

The Clue

- Hence, if Q uniform, $\tilde{P}^n$ and $Q^n(\cdot \mid \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) = t)$ assign approximately the same probability to the event $|\mathbb{P}^{(n)}(x) - \tilde{P}(x)| > \epsilon$
- May conjecture that more generally, for arbitrary Q and almost all sets $\mathcal{A}$ we will ever be interested in:

$$\tilde{P}^n(\mathcal{A}) \approx Q^n(\mathcal{A} \mid \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) = t)$$

Theorem 1. (the concentration phenomenon for typical sets, lattice case) Assume we are given a constraint $E_P[\phi(X)] = t$ and a prior Q such that

1. ϕ is a k-dimensional lattice random vector $\phi(x) = (\phi_1(x), \ldots, \phi_k(x))$ with span $h = (h_1, \ldots, h_k)$;
2. t is in the interior of the convex hull of the range of ϕ;
3. a Minimum Relative Entropy $\tilde{P}$ for the constraint exists and has invertible covariance matrix Σ,

Then there exists a sequence $\{c_i\}$ satisfying

$$\lim_{n\to\infty} c_n = \frac{\prod_{j=1}^{k} h_j}{\sqrt{(2\pi)^k \det \Sigma}}$$

such that the following holds:

Let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be an arbitrary sequence of sets with $\mathcal{A}_i \subset \mathcal{X}^i$. For all n with $Q(T_n = \tilde{t}) > 0$, we have:

$$\tilde{P}(\mathcal{A}_n) \geq n^{-k/2} c_n Q(\mathcal{A}_n \mid \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) = t).$$

Corollary: Strong Concentration Phenomenon, Part I

Suppose $\mathcal{B}_1, \mathcal{B}_2, \ldots$ is a sequence of sets with $\mathcal{B}_i \subset \mathcal{X}^i$ that are `typical' in the sense that the probability $\tilde{P}(\mathcal{B}_n)$ tends to 1 `fast enough', that is:

$$1 - \tilde{P}(\mathcal{B}_n) = O(f(n)n^{-k/2})$$

for some function $f : \mathbf{N} \to \mathbf{R}$; $f(n) = o(1)$.

Then $Q(\mathcal{B}_n | \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) = t)$ tends to 1 in the sense that $1 - Q(\mathcal{B}_n | \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) = t) = O(f(n))$.

Corollary: Strong Concentration Phenomenon, Part I: typical sets

- Our bound is tight.
- Proof technique uses `local' central limit theorem for lattice random vectors; can be extended to real-valued continuous random vectors
- Previous, similar results made use of Stirling's approximation
  - get bound of form $\tilde{P}(\mathcal{A}_n) \geq n^{-|\mathcal{X}|} c_n Q(\mathcal{A}_n | \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) = t)$
  - Not tight; applicable only to finite sample spaces (cardinality of sample space has nothing to do with the phenomenon)

Strong Concentration Phenomenon, Part II: arbitrary (measurable) sets

Theorem 2. **Strong Concentration Phenomenon/ Strong Conditional Limit Theorem**
Assume we are given a prior distribution Q and a constraint $E_P[\phi(X)] = t$ such that

1. ϕ is a lattice random vector or a continuous function $\phi : \mathcal{X} \to \mathbf{R}^k$;
2. t is in the interior of the convex hull of the range of ϕ;
3. A minimum relative entropy $\tilde{P}$ exists.

Let $\{m_i\}$ be an increasing sequence with $m_i \in \mathbf{N}$, such that $\lim_{n\to\infty} m_n/n = 0$.

Then as $n \to \infty$, $Q^{m_n}(\cdot \mid \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) = t)$ converges to $\tilde{P}^{m_n}(\cdot)$ (in the sense of weak convergence).

Strong Concentration Phenomenon, Part II: arbitrary (measurable) sets

- Note *m* can grow quite fast as *n* tends to infinity, e.g. $m = \lceil n/\log n \rceil$ will do.
- Generalizes Van Campenhout and Cover's (1981) Conditional Limit Theorem
  (they only consider fixed m as n tends to infinity)
- Relation to Large Deviations (Sanov's Thm.)

Applications

- Universal Codes/Models for exponential families (MDL)
  - Use Theorem 1 to construct 2-part codes achieving the Shtarkov-Rissanen minimax ('normalized maximum likelihood') code lengths
- Game-Theoretic Characterization of MaxEnt
  - Sequential prediction wrt log loss
- MaxEnt and Algorithmic Randomness
  - Sequential prediction wrt general loss

Consequences for Sequential Prediction

- Let $x_1, \ldots, x_n$ be *any* sequence satisfying the constraint. Then sequential prediction of the x_i based on MaxEnt $\tilde{P}$ is worst-case optimal
  if prediction error is measured using log-loss.
- Let $x_1, x_2, \ldots$ be a sequence that is algorithmically random with respect to the constraint. Then sequential prediction of the x_i based on $\tilde{P}$ is 'almost' optimal for *every* loss function.

Game-Theoretic Characterization of MaxEnt

Theorem 3. Let $\mathcal{X}$ be a countable sample space. Assume we are given a constraint $E_P[\phi(X)] = t$ such that ϕ is a lattice random vector and t is in the interior of the convex hull of the range of ϕ. Let $\mathcal{C}^{(n)} = \{(x_1, \ldots, x_n) \mid \frac{1}{n}\sum_{i=1}^n \phi(x_i) = t\}$.

Let $\tilde{P}$ be the distribution minimizing $D(P\|Q)$ (over P). Then the in¯mum in

$$\inf_{P \in \mathcal{P}(\mathcal{X}^\infty)} \sup_{\{n \,:\, \mathcal{C}^{(n)} \neq \emptyset\}} \sup_{x^{(n)} \in \mathcal{C}^{(n)}} -\frac{1}{n} \log \frac{P(x_1, \ldots, x_n)}{Q^n(x_1, \ldots, x_n)}$$

is achieved by the distribution $\tilde{P}$, and is equal to $\mathbf{H}(\tilde{P})$.

Game-Theoretic Characterization of MaxEnt

- Generalizes previous game-theoretic justification/characterization of MaxEnt as minimax-optimal prediction strategy over all *distributions* satisfying constraint…
  - Topsoe 1979, Grunwald 1998
- …to minimax-optimal prediction strategy over all *sequences* satisfying constraint
  - more `COLT-style'

MaxEnt and Algorithmic Randomness

`If the information incorporated into the maximum-entropy analysis includes ***all the constraints actually operating in the random experiment***, then the distribution predicted by maximum entropy is overwhelmingly the most likely to be observed experimentally' - Jaynes, 1996.

MaxEnt and Algorithmic Randomness

`If the information incorporated into the maximum-entropy analysis includes ***all the constraints actually operating in the random experiment***, then the distribution predicted by maximum entropy is overwhelmingly the most likely to be observed experimentally' - Jaynes, 1996.

What the … does this mean?

MaxEnt and Algorithmic Randomness

Using Theorem 1, we can make Jaynes' statement precise:

Suppose $x_1, x_2, \ldots$ is algorithmically random with respect to constraint $\mathcal{C}^{(n)} \equiv \{(x_1, \ldots, x_n)\} | \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) = t\}$ in the sense that $K((x_1, \ldots, x_n) | \mathcal{C}^{(n)}) = |\log \#(\mathcal{C}^{(n)})| + O(1)$, and:

Suppose $\mathcal{B}_1, \mathcal{B}_2, \ldots$ is a sequence of sets with $\mathcal{B}_i \subset \mathcal{X}^i$ such that $K(\mathcal{B}_n | n) = O(1)$ and such that the $\mathcal{B}_i$ are `typical' in the sense that the probability $\tilde{P}(\mathcal{B}_n)$ tends to $1$ `fast enough', that is:

$$1 - \tilde{P}(\mathcal{B}_n) = O(f(n) n^{-k/2})$$

for some function $f : \mathbf{N} \rightarrow \mathbf{R}$; $f(n) = o(1)$.

Then for all large n, $(x_1, x_2, \ldots, x_n) \in \mathcal{B}_n$.

Consequences for Sequential Prediction

- Let $x_1, \ldots, x_n$ be *any* sequence satisfying the constraint. Then sequential prediction of the x_i based on MaxEnt $\tilde{P}$ is worst-case optimal
  if prediction error is measured using log-loss.
- Let $x_1, x_2, \ldots$ be a sequence that is algorithmically random with respect to the constraint. Then sequential prediction of the x_i based on $\tilde{P}$ is 'almost' optimal for *every* loss function.

Thank you for your attention!