From Decision Making under Uncertainty to Probabilistic Graphical Models

Contact: andreas.holzinger AT tuwien.ac.at

What will we learn today?

- 00 Reflection from last lecture
- 01 Decision Making under uncertainty
- 02 Some Basics of Graphs/Networks
- 03 Bayesian Networks (BN)
- 04 Markov Chain Monte Carlo (MCMC)
- 05 Metropolis Hastings Algorithm (MH)
- 06 Probabilistic Programming (PP)
00 Reflection
Warm-up Quiz
On top-level - which machine learning approaches do we know?

- **Symbolic ML**
  - First order logic, inverse deduction, knowledge composition
  - Tom Mitchell, Steve Muggleton, Ross Quinlan, ...

- **Bayesian ML**
  - Statistical learning, probabilistic inference
  - Judea Pearl, Michael Jordan, David Heckermann, ...

- **Cognitive ML**
  - Analogisms from Psychology, Kernel machines
  - Vladimir Vapnik, Peter Hart, Douglas Hofstaedter, ...

- **Connectionist ML**
  - Neuroscience, Backpropagation
  - Geoffrey Hinton, Yoshua Bengio, Yann LeCun, ...

- **Evolutionary ML**
  - Nature-inspired concepts, genetic programming
  - John Holland (1929-2015), John Koza, Hod Lipson, ...
Why is this important for us?


A short historical digression: whom do you see here?

Both Images are in the public domain
Remember: What is “personalized medicine”?

01 Decision Making under uncertainty

How to go from System 1 deep learning to System 2 deep learning?


https://www.youtube.com/watch?v=T3sxeTgT4qc
What is medical action?

... permanent decision making under uncertainty!
Decision Making: Learn good policy for selecting actions

Goal: Learn an **optimal policy** for selecting best actions within a given **context**

For $t = 1, \ldots, T$

1) The world produces a “context” $x_t \in X$

2) The learner selects an action $a_t \in \{1, \ldots, K\}$

3) The world reacts with a reward $r_t(a_t) \in [0,1]$
Why is decision making so difficult?

How does a medical doctor make a decision?

3 July 1959, Volume 130, Number 3366

Reasoning Foundations of Medical Diagnosis

Symbolic logic, probability, and value theory aid our understanding of how physicians reason.

Robert S. Ledley and Lee B. Lusted

The purpose of this article is to analyze the complicated reasoning processes inherent in medical diagnosis. The importance of this problem has received recent emphasis by the increasing interest in the use of electronic computers as an aid to medical diagnostic processes fitted into a definite disease category, or that it may be one of several possible diseases, or else that its exact nature cannot be determined." This, obviously, is a greatly simplified explanation of the process of diagnosis, for the physician might also comment that after seeing a once are the ones who do remember and consider the most possibilities."

Computers are especially suited to help the physician collect and process clinical information and remind him of diagnoses which he may have overlooked. In many cases computers may be as simple as a set of hand-sorted cards, whereas in other cases the use of a large-scale digital electronic computer may be indicated. There are other ways in which computers may serve the physician, and some of these are suggested in this paper. For example, medical students might find the computer an important aid in learning the methods of differential diagnosis. But to use the computer thus we must understand how the physician makes a medical diagnosis. This, then, brings us to the subject of our investigation: the reasoning foundations of medical diagnosis and treatment.

Medical diagnosis involves processes that can be systematically analyzed, as well as those characterized as "intangible." For instance, the reasoning foundations of medical diagnostic procedures...
What about the accuracy and uncertainty in medicine?

- Medical (clinical) data are defined and detected disturbingly “soft” ...
- ... having an obvious degree of **variability** and **inaccuracy**.
- Taking a medical history, the performance of a physical examination, the interpretation of laboratory tests, even the definition of diseases ... are surprisingly **inexact**.
- Data is defined, collected, and interpreted with a degree of variability and inaccuracy which falls far short of the standards **which engineers do expect from most data**.
- Moreover, standards might be **interpreted variably** by different medical doctors, different hospitals, different medical schools, different medical cultures, ...

Why is the patient-doctor dialogue so important?

- Clinician influences patient's giving of information
  - Patient influences clinician's reception of information
    - Patient provides information
      - Patient seeks medical care
    - Patient responds with certain feelings, and decides to take those therapeutic actions recommended by clinician
      - Patient's response influences clinician's decisions
        - Clinician's description/explanation influences patient's response
          - Outcome of medical care

How can we learn and infer from data?

\[ \mathcal{H} \ldots \{h_1, h_2, \ldots, h_n\} \]

\[ h \ldots \text{hypotheses} \]

\[ p(h|d) = \frac{p(d|h) \cdot p(h)}{\sum_{h \in \mathcal{H}} p(d|h^\prime) \cdot p(h^\prime)} \]

Prior Probability

Likelihood

Posterior Probability

Problem in \( \mathbb{R}^n \rightarrow \text{complex} \)
How do humans make decisions under uncertainty?

Image by Christopher D. Wickens 1984, modified by Andreas Holzinger 2004
02 Graphs = Networks
What are Probabilistic Graphical Models (PGM)?

- PGM can be seen as a combination between Graph Theory + Probability Theory + Machine Learning

- One of the most exciting advancements in AI in the last decades – with enormous future potential

- Compact representation for exponentially-large probability distributions

- Example Question: “Is there a path connecting two proteins?”

\[
\text{Path}(X,Y) := \text{edge}(X,Y) \\
\text{Path}(X,Y) := \text{edge}(X,Y), \text{path}(Z,Y)
\]

This cannot be expressed in first-order logic.
We start in 1736

Leonhard Euler 1741. Solutio problematis ad geometriam situs pertinentis.
Commentarii academiae scientiarum Petropolitanae, 8, 128-140.
275 years later ... the “Nobel-prize in Computer Science”

http://amturing.acm.org/vp/pearl_2658896.cfm
Nobel Prize in Chemistry 2013

http://news.harvard.edu/gazette/story/2013/10/nobel_prize_awarded_2013

http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013
First Question: Where do graphs come from?

- Graphs as models for networks
  - given as direct input (point cloud data sets)
  - Given as properties of a structure
  - Given as a representation of information (e.g. Facebook data, viral marketing, etc., ...)

- Graphs as nonparametric basis
  - we learn the structure from samples and infer
  - flat vector data, e.g. similarity graphs
  - encoding structural properties (e.g. smoothness, independence, ...)
What do you see here?

NGC 5139 Omega Centauri by Edmund Halley in 1677, ESO, Atacama, Chile (2011)
Why is time and space important in health informatics?

- **Time**
  - e.g. Entropy

- **Space**
  - e.g. Topology

Dali, S. (1931) *The persistence of memory*

Bagula & Bourke (2012) *Klein-Bottle*
Why is the Complexity Problem: Time versus Space affecting us?

P versus NP and the Computational Complexity Zoo, please have a look at https://www.youtube.com/watch?v=YX40hbAHx3s
Why are protein structures so important?

First yeast protein-protein interaction network

Nodes = proteins
Links = physical interactions (bindings)
Red Nodes = lethal
Green Nodes = non-lethal
Orange = slow growth
Yellow = not known

First human protein-protein interaction network

Light blue = known proteins
Orange = disease proteins
Yellow ones = not known yet

Non-Natural Network Example: Blogosphere

03 Bayesian Networks
“Bayes’ Nets”
Book recommendations 1


http://www.cs.ucl.ac.uk/staff/d.barber/brml/
Book recommendations 2


Chapter 8 Graphical Models is as sample chapter fully downloadable for free

https://goo.gl/6a7rOC

http://bayes.cs.ucla.edu/BOOK-2K/
What are the rules of probability?

\[ P(x) = \sum_y P(x, y) \]

\[ P(x, y) = P(y|x)P(x) \]

\[ P(y|x) = \frac{P(x|y)P(y)}{P(x)} \]

\[ P(x) = \sum_y P(x|y)P(y) \]
Digression:
Markov Processes in Machine Learning
Why are Markov decision processes so important?

- Markov decision processes (MDP) are ...
- random processes in which the future, given the present, is independent of the past!
- one of the most important classes of random processes!

The Markov Process in Medical Prognosis

*J. Robert Beck, M.D., and Stephen G. Pauker, M.D.*

The physician's estimate of prognosis under alternative treatment plans is a principal factor in therapeutic decision making. Current methods of reporting prognosis, which include five-year survivals, survival curves, and quality-adjusted life expectancy, are crude estimates of natural history. In this paper we describe a general-purpose model of medical prognosis based on the Markov process and show how this simple mathematical tool may be used to generate detailed and accurate assessments of life expectancy and health status. (Med Decis Making 3:419–458, 1983)
How can MDP be useful for machine learning?

From where do we know such behaviour?

**What is the result of the Expected Utility Theory \( E(U|d) \)?**

For a single decision variable an agent can select \( D = d \) for any \( d \in \text{dom}(D) \). The expected utility of decision \( D = d \) is

\[
E(U | d) = \sum_{x_1, \ldots, x_n} P(x_1, \ldots, x_n | d)U(x_1, \ldots, x_n, d)
\]

An optimal single decision is the decision \( D = d_{\text{max}} \) whose expected utility is maximal:

\[
d_{\text{max}} = \arg \max_{d \in \text{dom}(D)} E(U | d)
\]

initialize $V(s)$ arbitrarily

loop until policy good enough
  loop for $s \in S$
    loop for $a \in A$
      $Q(s, a) := R(s, a) + \gamma \sum_{s' \in S} T(s, a, s') V(s')$
      $V(s) := \max_a Q(s, a)$
    end loop
  end loop
end loop

What has an RL-agent to do with MDP?

Intelligent behavior arises from the actions of an individual seeking to maximize its received reward signals in a complex and changing world.

RL – Types of Feedback (crucial!)

- **Supervised**: Learner told best \( a \)
- **Exhaustive**: Learner shown every possible \( x \)
- **One-shot**: Current \( x \) independent of past \( a \)

Problem Formulation in a MDP

- Markov decision processes specify setting and tasks
- Planning methods use knowledge of $P$ and $R$ to compute a good policy $\pi$
- Markov decision process model captures both sequential feedback and the more specific one-shot feedback (when $P(s'|s, a)$ is independent of both $s$ and $a$)

Agent observes environmental state at each step $t$

- 1) Overserves
- 2) Executes
- 3) Receives Reward

Executes action $A_t$:

$O_t = sa_t = se_t$

Agent state = environment state = information state

Markov decision process (MDP)

*Image credit to David Silver, UCL*
Environmental State is the current representation

- i.e. whatever data the environment uses to pick the next observation/reward
- The environment state is not usually visible to the agent
- Even if $S$ is visible, it may contain irrelevant information
- A State $S_t$ is Markov iff:

$$\mathbb{P}[S_{t+1} | S_t] = \mathbb{P}[S_{t+1} | S_1, \ldots, S_t]$$
Agent State is the agents internal representation

- i.e. whatever information the agent uses to pick the next action
- it is the information used by reinforcement learning algorithms
- It can be any function of history:
- \[ S = f(H) \]
- \[ H_t = O_1, R_1, A_1, \ldots, A_{t-1}, O_t, R_t \]
Components of RL Agents and Policy of Agents

- **RL agent components:**
  - **Policy:** agent's behaviour function
  - **Value function:** how good is each state and/or action
  - **Model:** agent's representation of the environment

- **Policy as the agent's behaviour**
  - is a map from state to action, e.g.
  - Deterministic policy: \( a = (s) \)
  - Stochastic policy: \( \pi(s) = P[A_t = a | S_t = s] \)

- **Value function is prediction of future reward:**

\[
\nu_\pi(s) = \mathbb{E}_\pi \left[ R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \mid S_t = s \right]
\]
Partial observability: when agent only indirectly observes environment

Formally this is a Partially Observable Markov Decision Process (POMDP):
- Agent must construct its own state representation $S$
  for example:

  - Complete history: $S^a_t = H_t$
  - Beliefs of environment state: $S^a_t = (P[S^e_t = s^1], ..., P[S^e_t = s^n])$
  - Recurrent neural network: $S^a_t = \sigma(S^a_{t-1} W_s + O_t W_o)$
Back to Bayesian Networks
Three types of Probabilistic Graphical Models

**Undirected:** Markov random fields, useful e.g. for computer vision (Details: Murphy 19)

\[
P(X) = \frac{1}{Z} \exp \left( \sum_{ij} W_{ij} x_i x_j + \sum_i x_i b_i \right)
\]

**Directed:** Bayes Nets, useful for designing models (Details: Murphy 10)

\[
p(x) = \prod_{k=1}^{K} p(x_k | pa_k)
\]

**Factored:** useful for inference/learning

\[
p(x) = \prod_{s} f_s(x_s)
\]
So, what is a directed Bayesian Network (BN)?

- is a **probabilistic model**, consisting of two parts:
  1) a dependency structure and
  2) local probability models.

\[ p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid Pa(x_i)) \]

Where \( Pa(x_i) \) are the parents of \( x_i \)

BN inherently model the **uncertainty in the data**. They are a successful marriage between probability theory and graph theory; allow to model a multidimensional probability distribution in a sparse way by searching independency relations in the data. Furthermore this model allows different strategies to integrate two data sources.

Example: Directed Bayesian Network with 7 nodes

\[ p(X_1, \ldots, X_7) = \]
\[ p(X_1) p(X_2) p(X_3) p(X_4 | X_1, X_2, X_3) \cdot \]
\[ p(X_5 | X_1, X_3) p(X_6 | X_4) p(X_7 | X_4, X_5) \]
Clinical Case Example

What is important in clinical decision making?

- The prediction of the future course of a disease conditional on the patient’s history and a projected treatment strategy.
- Danger: probable Information!
- Therefore valid prognostic models can be of great benefit for clinical decision making and of great value to the patient, e.g., for notification and quality of-life decisions.

How can one predict the future on past data and present status?

Example: Breast cancer - Probability Table

<table>
<thead>
<tr>
<th>Category</th>
<th>Node description</th>
<th>State description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis</td>
<td>Breast cancer</td>
<td>Present, absent.</td>
</tr>
<tr>
<td>Clinical history</td>
<td>Habit of drinking alcoholic beverages and smoking</td>
<td>Yes, no.</td>
</tr>
<tr>
<td></td>
<td>Taking female hormones</td>
<td>Yes, no.</td>
</tr>
<tr>
<td></td>
<td>Have gone through menopause</td>
<td>Yes, no.</td>
</tr>
<tr>
<td></td>
<td>Have ever been pregnant</td>
<td>Yes, no.</td>
</tr>
<tr>
<td></td>
<td>Family member has breast cancer</td>
<td>Yes, no.</td>
</tr>
<tr>
<td>Physical findings</td>
<td>Nipple discharge</td>
<td>Yes, no.</td>
</tr>
<tr>
<td></td>
<td>Skin thickening</td>
<td>Yes, no.</td>
</tr>
<tr>
<td></td>
<td>Breast pain</td>
<td>Yes, no.</td>
</tr>
<tr>
<td></td>
<td>Have a lump(s)</td>
<td>Yes, no.</td>
</tr>
<tr>
<td>Mammographic findings</td>
<td>Architectural distortion</td>
<td>Present, absent.</td>
</tr>
<tr>
<td></td>
<td>Mass</td>
<td>Score from one to three, score from four to five, absent</td>
</tr>
<tr>
<td></td>
<td>Microcalcification cluster</td>
<td>Score from one to three, score from four to five, absent</td>
</tr>
<tr>
<td></td>
<td>Asymmetry</td>
<td>Present, absent.</td>
</tr>
</tbody>
</table>

Breast cancer – big picture – state of 1999

10 years later: Integration of microarray data

- Integrating microarray data from multiple studies to increase sample size;
- = approach to the development of more robust prognostic tests

Example: Bayes Net with four binary variables

First the structure is learned using a search strategy.

Since the number of possible structures increases super exponentially with the number of variables,

the well-known greedy search algorithm $K2$ can be used in combination with the Bayesian Dirichlet (BD) scoring metric:

$$p(S|D) \propto p(S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left[ \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \right]$$

$N_{ijk}$ ... number of cases in the data set $D$ having variable $i$ in state $k$ associated with the $j$-th instantiation of its parents in current structure $S$.

$n$ is the total number of variables.
Next, $N_{ij}$ is calculated by summing over all states of a variable:

$$N_{ij} = \sum_{k=1}^{r_i} N_{ijk} \cdot N'_{ijk}$$

and $N'_{ij}$ have similar meanings but refer to prior knowledge for the parameters.

When no knowledge is available they are estimated using $N_{ijk} = N/\left(r_iq_i\right)$

with $N$ the equivalent sample size,

$r_i$ the number of states of variable $i$ and

$q_i$ the number of instantiations of the parents of variable $i$.

$\Gamma(.)$ corresponds to the gamma distribution.

Finally $p(S)$ is the prior probability of the structure.

$p(S)$ is calculated by:

$$p(S) = \prod_{i=1}^{n} \prod_{l_i=1}^{pi} p(l_i \rightarrow x_i) \prod_{m_i=1}^{oi} p(m_i x_i)$$

with $p_i$ the number of parents of variable $x_i$ and $o_i$ all the variables that are not a parent of $x_i$.

Next, $p(a \rightarrow b)$ is the probability that there is an edge from $a$ to $b$ while $p(ab)$ is the inverse, i.e. the probability that there is no edge from $a$ to $b$
Parameter learning -> second step

• Estimating the parameters of the local probability models corresponding with the dependency structure.
• CPTs are used to model these local probability models.
• For each variable and instantiation of its parents there exists a CPT that consists of a set of parameters.
• Each set of parameters was given a uniform Dirichlet prior:

\[ p(\theta_{ij}|S) = \text{Dir}(\theta_{ij}|N'_{ij1}, \ldots, N'_{ijk}, \ldots, N'_{ijr_i}) \]

Note: With \( \theta_{ij} \) a parameter set where \( i \) refers to the variable and \( j \) to the \( j \)-th instantiation of the parents in the current structure. \( \theta_{ij} \) contains a probability for every value of the variable \( x_i \) given the current instantiation of the parents. \( \text{Dir} \) corresponds to the Dirichlet distribution with \((N'_{ij1}, \ldots, N'_{ijr_i})\) as parameters of this Dirichlet distribution. Parameter learning then consists of updating these Dirichlet priors with data. This is straightforward because the multinomial distribution that is used to model the data, and the Dirichlet distribution that models the prior, are conjugate distributions. This results in a Dirichlet posterior over the parameter set:

\[ p(\theta_{ij}|D,S) = \text{Dir}(\theta_{ij}|N'_{ij1} + N_{ij1}, \ldots, N'_{ijk} + N_{ijk}, \ldots, N'_{ijr_i} + N_{ijr_i}) \]

with \( N_{ijk} \) defined as before.
Predicting the prognosis of breast cancer (integrated a.)

For certain cases it is tractable if:

- Just one variable is unobserved
- We have singly connected graphs (no undirected loops -> belief propagation)
- Assigning probability to fully observed set of variables

Possibility: Monte Carlo Methods (generate many samples according to the Bayes Net distribution and then count the results)

Otherwise: approximate solutions ...
Often it is better to have a good solution within time – than an perfect solution too late …
Digression: Graphical Models and Decision Making

\[ \mathcal{D} \equiv \{ X_1^{(i)}, X_2^{(i)}, \ldots, X_n^{(i)} \}_{i=1}^N \]
What Classes of Graphical Models do we know?

Naïve Bayes classifier as DGM (single/nested plates)

\[ p(\theta, \mathcal{D}) = p(\theta) \prod_{i=1}^{N} p(x_i | \theta) \]

Regulatory > Metabolic > Signaling > Protein > Co-expression

Image credit to Anna Goldenberg, Toronto
- Medicine is an extremely complex application domain – dealing most of the time with uncertainties -> **probable information**!
- When we have big data but little knowledge automatic ML can help to gain insight:
  - **Structure learning and prediction in large-scale biomedical networks with probabilistic graphical models**
- If we have little data and deal with NP-hard problems we still need the human-in-the-loop
\- Hypothesis: most biological functions involve the interactions between many proteins, and the complexity of living systems arises as a result of such interactions.

\- In this context, the problem of inferring a global protein network for a given organism,

\- using all (genomic) data of the organism,

\- is one of the main challenges in computational biology

Important for health informatics: Discovering relationships between biological components

Unsolved problem in computer science:

Can the graph isomorphism problem be solved in polynomial time?

- So far, no polynomial time algorithm is known.
- It is also not known if it is NP-complete
- We know that subgraph-isomorphism is NP-complete
Finally a practical example

04 Markov Chain Monte Carlo (MCMC)
Monte Carlo Method (MC)
Monte Carlo Sampling
Markov Chains (MC)
MCMC
Metropolis-Hastings
Often we want to calculate characteristics of a \textbf{high-dimensional} probability distribution ... 

\[ p(h|d) \propto p(D|\theta) \times p(h) \]

Posterior integration problem: (almost) all statistical inference can be deduced from the posterior distribution by calculating the appropriate sums, which involves an integration:

\[ J = \int f(\theta) \times p(\theta|D) d\theta \]
What is the problem of learning and inference?

- **Statistical physics:** computing the partition function – this is evaluating the posterior probability of a hypothesis and this requires summing over all hypotheses ... remember:

\[
\mathcal{H} = \{H_1, H_2, \ldots, H_n\} \quad \forall (h, d)
\]

\[
P(h|d) = \frac{P(d|h) \ast P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}
\]
What was the origin of MCMC?
Summary: What are Monte Carlo methods?

- Class of algorithms that rely on repeated random sampling
- Basic idea: using randomness to solve problems with high uncertainty (Laplace, 1781)
- For solving multidimensional integrals which would otherwise intractable
- For simulation of systems with many dof
- e.g. fluids, gases, particle collectives, cellular structures - see our last tutorial on Tumor growth simulation!
- for solving problems of probabilistic inference involved in developing computational models
- as a source of hypotheses about how the human mind might solve problems of inference
- For a function $f(x)$ and distribution $P(x)$, the expectation of $f$ with respect to $P$ is generally the average of $f$, when $x$ is drawn from the probability distribution $P(x)$

$$\mathbb{E}_{p(x)}(f(x)) = \sum_{x} f(x)P(x)dx$$
- Solving intractable integrals
- Bayesian statistics: normalizing constants, expectations, marginalization
- Stochastic Optimization
- Generalization of simulated annealing
- Monte Carlo expectation maximization (EM)
Physical simulation via MC

- Physical simulation
- estimating neutron diffusion time
- Computing expected utilities and best responses toward Nash equilibria
- Computing volumes in high-dimensions
- Computing eigen-functions and values of operators (e.g. Schrödinger)
- Statistical physics
- Counting many things as fast as possible
- Expectation of a function $f(x, y)$ with respect to a random variable $x$ is denoted by $\mathbb{E}_x[f(x, y)]$.

- In situations where there is no ambiguity as to which variable is being averaged over, this will be simplified by omitting the suffix, for instance $\mathbb{E}_x$.

- If the distribution of $x$ is conditioned on another variable $z$, then the corresponding conditional expectation will be written $\mathbb{E}_x[f(x)|z]$.

- Similarly, the variance is denoted $\text{var}[f(x)]$, and for vector variables the covariance is written $\text{cov}[x, y]$. 
Global optimization: What is the main problem?

\[ \text{argmax} \ f(x) \]

Normalization:  
\[ p(x|y) = \frac{p(y|x) \times p(x)}{\int_X p(y|x) \times p(x) \, dx} \]

Marginalization:  
\[ p(x) = \int_Z p(x, z) \, dz \]

Expectation:  
\[ \mathbb{E}_{p(x)}(f(x)) = \int_X f(x)p(x) \, dx \]
Finally a practical example

05 Metropolis-Hastings Algorithm
THE MONTE CARLO METHOD

Nicholas Metropolis and S. Ulam
Los Alamos Laboratory

We shall present here the motivation and a general description of a method dealing with a class of problems in mathematical physics. The method is, essentially, a statistical approach to the study of differential equations, or more generally, of integro-differential equations that occur in various branches of the natural sciences.

Already in the nineteenth century a sharp distinction began to appear between two different mathematical methods of treating physical phenomena. Problems involving only a few particles were studied in classical mechanics, through the study of systems of ordinary differential equations. For the description of systems with very many particles, an entirely different technique was used, namely, the method of statistical mechanics. In this latter approach, one does not concentrate on the individual particles but studies the properties of sets of particles. In pure mathematics an intensive study of the properties of sets of points was the subject of a new field. This is the so-called theory of sets, the basic theory of integration, and the twentieth century development of the theory of probabilities prepared the formal apparatus for the use of such models in theoretical physics, i.e., description of properties of aggregates of points rather than of individual points and

Image Source:
http://www.manhattanprojectvoices.org/or-al-histories/nicholas-metropolis-interview
Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ALANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND ANGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

AND

EDWARD TELLER,* Department of Physics, University of Chicago, Chicago, Illinois
(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for systems consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.

I. INTRODUCTION

THE purpose of this paper is to describe a general method, suitable for fast electronic computing machines, of calculating the properties of any substance which may be considered as composed of interacting individual molecules. Classical statistics is assumed, only two-body forces are considered, and the potential field of a molecule is assumed spherically symmetric. These are the usual assumptions made in theories of liquids. Subject to the above assumptions, the method is not restricted to any range of temperature or density. This paper will also present results of a preliminary two-dimensional calculation for the rigid-sphere system. Work on the two-dimensional case with a Lennard-Jones potential is in progress and will be reported in a later paper. Also, the problem in three dimensions is being investigated.

II. THE GENERAL METHOD FOR AN ARBITRARY POTENTIAL BETWEEN THE PARTICLES

In order to reduce the problem to a feasible size for numerical work, we can, of course, consider only a finite number of particles. This number \( N \) may be as high as several hundred. Our system consists of a square† containing \( N \) particles. In order to minimize the surface effects we suppose the complete substance to be periodic, consisting of many such squares, each square containing \( N \) particles in the same configuration. Thus we define \( d_{AB} \), the minimum distance between particles \( A \) and \( B \), as the shortest distance between \( A \) and any of the particles \( B_i \) of which there is one in each of the squares which comprise the complete substance. If we have a potential which falls off rapidly with distance, there will be at most one of the distances \( AB \) which can make a substantial contribution; hence we need consider only the minimum distance \( d_{AB} \).

* Now at the Radiation Laboratory of the University of California, Livermore, California.
† We will use the two-dimensional nomenclature here since it is easier to visualize. The extension to three dimensions is obvious.

Monte Carlo sampling methods using Markov chains and their applications

BY W. K. HASTINGS
University of Toronto

SUMMARY

A generalization of the sampling method introduced by Metropolis et al. (1953) is presented along with an exposition of the relevant theory, techniques of application and methods and difficulties of assessing the error in Monte Carlo estimates. Examples of the methods, including the generation of random orthogonal matrices and potential applications of the methods to numerical problems arising in statistics, are discussed.

1. Introduction

For numerical problems in a large number of dimensions, Monte Carlo methods are often more efficient than conventional numerical methods. However, implementation of the Monte Carlo methods requires sampling from high dimensional probability distributions and this may be very difficult and expensive in analysis and computer time. General methods for sampling from, or estimating expectations with respect to, such distributions are as follows.

(i) If possible, factorize the distribution into the product of one-dimensional conditional distributions from which samples may be obtained.

(ii) Use importance sampling, which may also be used for variance reduction. That is, in order to evaluate the integral

\[ J = \int f(x)p(x)dx = E_x(f), \]

where \( p(x) \) is a probability density function, instead of obtaining independent samples \( x_1, \ldots, x_N \) from \( p(x) \) and using the estimate \( \hat{J} = \frac{\sum f(x_i)}{N} \), we instead obtain the sample from
So what is the MH-algorithm doing?

1. Choose a starting point $x^1$.
2. for $i = 2$ to $L$ do
3. Draw a candidate sample $x^{cand}$ from the proposal $\tilde{q}(x'|x^{l-1})$.
4. Let $a = \frac{\tilde{q}(x^{l-1}|x^{cand})p(x^{cand})}{\tilde{q}(x^{cand}|x^{l-1})p(x^{l-1})}$
5. if $a \geq 1$ then $x^l = x^{cand}$
6. else
7. draw a random value $u$ uniformly from the unit interval $[0, 1]$.
8. if $u < a$ then $x^l = x^{cand}$
9. else
10. $x^l = x^{l-1}$
11. end if
12. end if
13. end for
Importance sampling is a technique to approximate averages with respect to an intractable distribution $p(x)$.

The term ‘sampling’ is arguably a misnomer since the method does not attempt to draw samples from $p(x)$.

Rather the method draws samples from a simpler importance distribution $q(x)$ and then reweights them such that averages with respect to $p(x)$ can be approximated using the samples from $q(x)$. 
Why is Gibbs Sampling important?

- The Gibbs Sampler is an interesting special case of MH:

![Gibbs Sampling Diagram](image-source)

Image Source: Peter Mueller, Anderson Cancer Center
How to learn modular structures from Network Data?

Algorithm 1 RJMCMC for sampling parameters

Inputs:
Node Variables Data X
Network Data B

for iterations $j = 1$ to $J$ do
    Sample $A^{(j+1)}$ given $A^{(j)}$ using Alg 2 in (Azizi et al., 2014)
    Sample $S^{(j+1)}$ given $S^{(j)}$ using Alg 3 in (Azizi et al., 2014)
    for modules $k = 1$ to $K^{(j)}$ do
        Propose $w^{(j+1)}_k \sim \mathcal{N}(w^{(j)}_k, I)$
        Accept with probability $P_{mh}$; update $S^{(j+1)}$
        for parents $r = 1$ to $R_k$ do
            Propose $z^{(j+1)}_k \sim \mathcal{N}(z^{(j)}_k, I)$; accept with $P_{mh}$
            Propose $\pi^{(j+1)}_k \sim \mathcal{N}(\pi^{(j)}_k, I)$; accept with $P_{mh}$
        end for
    end for
    for condition $c = 1$ to $C$ do
        Propose $\mu^{(j+1)}_c \sim \mathcal{N}(\mu^{(j)}_c, I)$; accept with $P_{mh}$
        Propose $\gamma^{(j+1)}_c \sim \mathcal{N}(\gamma^{(j)}_c, I)$; accept with $P_{mh}$
    end for
end for

An alternative approach

Discrete-time stochastic epidemic model of COVID-19

- $B_{11}(t)$ is the number of susceptible individuals who become newly infected;
- $B_{12}(t)$ is the number of quarantined susceptible individuals who have contact with infected individuals but are not infected;
- $B_{21}(t)$ is the number of new cases with symptom onset;
- $B_{31}(t)$ is the number of new confirmed and admitted patients;
- $B_{32}(t)$ is the number of new death from infected individuals;
- $B_{33}(t)$ is the number of newly recovered from infected individuals;
- $B_{41}(t)$ is the number of people released from quarantine;
- $B_{61}(t)$ is the number of people admitted to hospital (also isolated);
- $B_{62}(t)$ is the number of newly recovered from hospitalized cases;
- $B_{63}(t)$ is the number of new death from hospitalized cases.

https://www.aimspress.com/MBE/2020/4/2792
(Online open available)

$L(B_{11}(t), B_{12}(t), B_{21}(t), B_{31}(t), B_{32}(t), B_{33}(t), B_{41}(t), B_{51}(t), B_{61}(t), B_{62}(t)|\Theta) = \prod_{t=0}^{T^n} g_{i,j}(B_{ij}(t)|\Theta)$
06 Probabilistic Programming
Book recommendations

Avi Pfeffer 2016. Practical probabilistic programming, Shelter Island (NY), Manning.


Probabilistic thinking is a valuable tool for decision making
Overcoming uncertainties is the huge success currently in machine learning (and for AI ;-) 
Probabilistic reasoning is a versatile tool
PPLs are domain specific languages that use probabilistic models and the methods to make inferences in those models
The “magic” is in combining “probability methods” with “representational power”
- C → Probabilistic-C
- Scala → Figaro
- Scheme → Church
- Excel → Tabular
- Prolog → Problog
- Javascript → webPP
  → Venture
- Python → PyMC
**Medical Example**

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTAAGAAGATACATGATGAAAGAAACCT</td>
<td>X</td>
</tr>
<tr>
<td>GCCCTTGTGACTACTTTACTCTACAGCTCA</td>
<td>X</td>
</tr>
<tr>
<td>TTAATAAGAGAGAGACTCCTACGATAC</td>
<td>Y</td>
</tr>
<tr>
<td>CCAAGAGCCTCTGTAATTAGATTTGCAATA</td>
<td>Y</td>
</tr>
<tr>
<td>TATGACUTCTGTGCAGTAGATTTUCTTT</td>
<td>X</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

1. Simple example: Nucleotide “A” may follow nucleotide “T” in the sequences more frequently for outcome X than for outcome Y.

\[ P(A|T, X) > P(A|T, Y) \]

2. Specify the prior distribution:

\[ P(\theta | D) = \frac{P(D | \theta) \cdot P(\theta)}{P(D)} \]

3. Specify the experimental data:

\[ P(\theta | D) = \frac{P(D | \theta) \cdot P(\theta)}{P(D)} \]

4. Experimental Data:

<table>
<thead>
<tr>
<th>Observation</th>
<th>Nucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. Prior Distribution of the Nucleotides

6. Posterior Distribution of the Nucleotides

Image Source: Dan Williams, Life Technologies, Austin TX
Digression on Concept Learning
Why is decision making so hard for machines?

You are talking to your colleague and want to refer to the middle object – which wording would you prefer: circle or blue?

Recursive reasoning: a case for probabilistic programming

```javascript
var literalListener = function(property){
    Infer(function(){
        var object = refPrior(context)
        condition(object[property])
        return object
    })
}

var speaker = function(object) {
    Infer(function(){
        var property = propPrior()
        condition(
            object ==
        )
    })
}

var listener = function(property) {
    Infer(function(){
        var object = refPrior(context)
        condition(utterance ==
            sample(speaker(object)))
        return object
    })
}
```

Why do we need concepts?
- can be relational and abstract
- category = set of objects that have commonalities
- concept = mental representation of categories
- concepts can be defined, e.g. triangle = a polygon with three sides, a gland = group of cells

Example: How do human pathologists make diagnoses?

What is ground truth? Where is the ground truth?

- := information provided by direct observation (empirical evidence) in contrast to information provided by inference
  - *Empirical evidence* = information acquired by observation or by experimentation in order to verify the truth (fit to reality) or falsify (non-fit to reality).
  - *Empirical inference* = drawing conclusions from empirical data (observations, measurements)
  - *Causal inference* = drawing conclusions about a causal connection based on the conditions of the occurrence of an effect
  - *Causal machine learning* is key to ethical AI in health to model explainability for bias avoidance and algorithmic fairness for decision making

Wassily Kandinsky (1866 – 1944)

Note: Image is in the public domain and is used according UrhG §42 lit. f Abs 1 as “Belegfunktion” for discussion with students.

Hubel & Wiesel (1962): Our world is compositional!


Bruner, Goodnow, and Austin (1956) published “A Study of Thinking”, which became a landmark in cognitive science and has much influence on machine learning.

- Rule-Based Categories
- A concept specifies conditions for membership

which is highly relevant for ML research, concerns the factors that determine the subjective difficulty of concepts:

- Why are some concepts psychologically extremely simple and easy to learn,
- while others seem to be extremely difficult, complex, or even incoherent?
- These questions have been studied since the 1960s but are still unanswered ...

How can we model basic cognitive capacities as intuitive Bayes?

- Similarity
- Representativeness and evidential support
- Causal judgement
- Coincidences and causal discovery
- Diagnostic inference
- Predicting the future

\[ P(h|x, T) = \frac{P(x|h, T)P(h|T)}{\sum_{h' \in H_T} P(x|h', T)P(h'|T)} \]

How does our mind get so much out of it?

How can we learn words for objects – concepts from examples

How do we understand our world?

What is probabilistic program induction?

What is the difference between deduction, induction, abduction?

- **Deductive Reasoning** = Hypothesis > Observations > Logical Conclusions (general → specific – proven correctness)
  - DANGER: Hypothesis must be correct! DR defines whether the truth of a conclusion can be determined for that rule, based on the truth of premises: A=B, B=C, therefore A=C
- **Inductive reasoning** = makes broad generalizations from specific observations (specific → general – not proven correctness)
  - DANGER: allows a conclusion to be false if the premises are true
  - generate hypotheses and use DR for answering specific questions
- **Abductive reasoning** = inference = to get the best explanation from an incomplete set of preconditions.
  - Given a true conclusion and a rule, it attempts to select some possible premises that, if true also, may support the conclusion, though not uniquely.
  - Example: "When it rains, the grass gets wet. The grass is wet. Therefore, it might have rained." This kind of reasoning can be used to develop a hypothesis, which in turn can be tested by additional reasoning or data.
Drawn by Human or Machine Learning Algorithm?

What can a Bayesian program learning (BPL) framework do?

A Bayesian program learning (BPL) framework, capable of learning a large class of visual concepts from just a single example and generalizing in ways that are mostly indistinguishable from people.

What can we do with graphical models?

**Principles**

- Classes: \{R, D, S\} (Risks, Diseases, Symptoms)
- Causal laws: R \rightarrow D, D \rightarrow S

**Structure**

A network diagram showing relationships between factors such as high fat diet, stressful lifestyle, working in factory, smoking, heart disease, lung cancer, bronchitis, flu, chest pain, coughing, headache, and fever.

**Data**

- Patient 1: Stressful lifestyle
  - Chest Pain
- Patient 2: Smoking
  - Coughing
- Patient 3: Working in factory
  - Chest Pain
  
  ...
- Cognition as probabilistic inference
  - Visual perception, language acquisition, motor learning, associative learning, memory, attention, categorization, reasoning, causal inference, decision making, theory of mind
- Learning concepts from examples
- Learning causation from correlation
- Learning and applying intuitive theories (balancing complexity vs. fit)
Thank you!
Appendix