CPS615/713

Computational Science Simulation Track
Some Practical Optimization Methods

Geoffrey Fox
NPAC
Syracuse University
111 College Place
Syracuse, New York 13244-4100

Abstract of Optimization Presentation

- We review problems including TSP, linear programming, and data analysis (χ^2).
- We discuss relevance of central limit theorem, least squares, maximum likelihood, histograms, and scatterplots.
- We use simple length measurement, as well as full scale physics experiment to discuss χ^2, maximum likelihood, and their contrast.
- Computer implementations are discussed for Nonlinear Minimization, linear programming, and histogramming.
- The foils end with notes for a set of CPS713 projects in this area.

Optimization Examples—I

- Given $f(\underline{x})$, find minimum value of f as a function of $\underline{x}$—for example,
  - $f(\underline{x}) = x_1^2 + x_2^2$ (continuous variables)
    minimum $x_1 = x_2 = 0$

Travelling Salesman Problem

- Let

$$
\begin{aligned}
i = 1 \ldots N \quad & \text{label cities} \\
\eta_{ip} = 1 \quad & \text{if traveller is at city } i \text{ at time step} \\
& p = 1 \ldots N \\
d_{ij} = \quad & \text{distance from city } i \text{ to city } j.
\end{aligned}
$$

- Minimize

$$D = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{p=1}^{N-1} d_{ij}\, \eta_{ip}\, \eta_{jp+1}$$

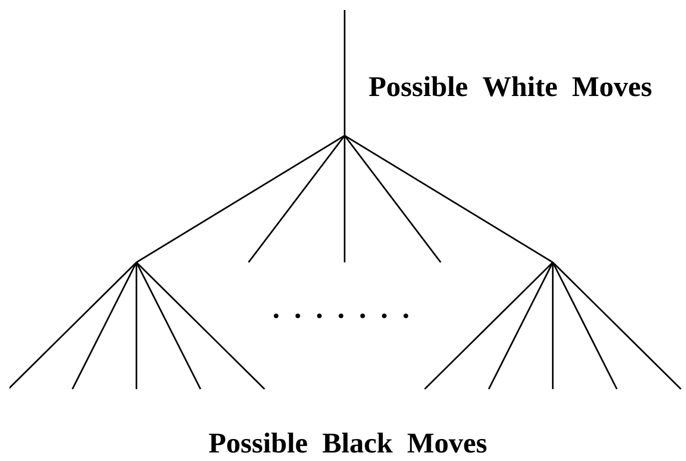
This is a (NP Complete) Discrete Variables example. The concept of NP completeness, discussed in detail by computer science, appears of little practical significance.

Examples – II

- Given a collection of molecules, find its ground state (minimize its energy as a function of atomic positions) and see if it will make a nice drug.
- Given a set of pixels, forming an image (as seen by your eyes), optimize your life based on information—pattern recognition/(artificial) intelligence (e.g., if image contains a lion, run away; while if image contains football thrown by Jim Kelly, catch it and run to goal line).
- Given a set of ELINT (electronic intelligence = signals as a function of position), decide on best steps to win the battle—spatial reasoning or situation assessment.
- Minimize $\int (\nabla\phi)^2 \, d^3x$ which is variational principle that derives Finite Element method

Examples – III

- Given position on chess board, decide on best move. Write down a tree



and expand several "plies" (two plies—one by a white, and one by block make a move). Assign values to final position. Chosen move is one that leads to highest (best) value of worst possible scenario—minimax problem.

Examples IV – Linear Programming

Transportation

- Take M warehouses and N retail outlets. Consider one product. We wish to minimize cost of shipping products from warehouses to outlets.
  - a_i units at warehouse i
  - b_j units needed at outlet j
  - x_{ij} units shipped from i to j at cost c_{ij} each

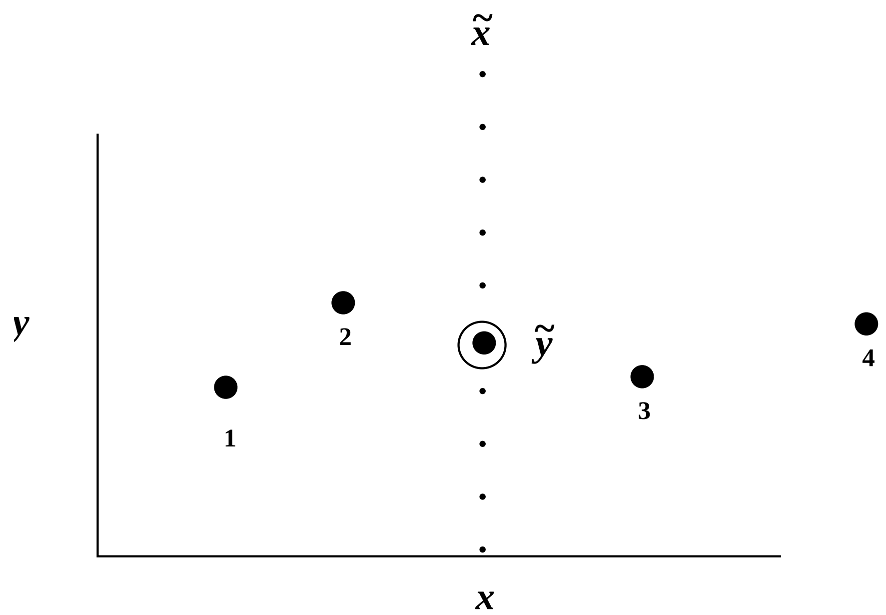
$$\text{constraints}\begin{cases} \sum\limits_{j=1}^{N} x_{ij} \leq a_i \\ \\ \sum\limits_{i=1}^{M} x_{ij} = b_j \end{cases}$$

- Minimize $\sum\limits_{i=1}^{N} \sum\limits_{j=1}^{M} c_{ij}\, x_{ij}$ objective function.
- This is Linear Programming—objective function and constraints are linear.

Optimization Examples – V

Interpolation

- Given a set of values $y = y_i$ at $x = x_i$, find the best estimate of $y = \tilde{y}$ at $x \geq \tilde{x}$.



- Typically do by assuming $y(x)$ is a polynomial—maybe of degree 3 in this example.
- Note that interpolation, like many optimization problems, is mathematically undefined. One can find polynomials of degree 4 taking any given values of x_1, x_2, x_3, x_4, and $\tilde{x}$. We can only determine value at $\tilde{x}$ by <u>assuming</u> polynomial of degree 3 (or other limiting assumption).

A Collection of Methods for Optimization

Maximum Likelihood
χ^2
Maximum Entropy
Combinational Optimization
α-β Pruning
Branch and Bound
Branch and Cut
Linear Programming
Nonlinear Programming
Integer Programming
Heuristic
Greedy Algorithm
Conjugate Gradient
Simulated Annealing
Simulated Tempering
Neural Networks
Deterministic Annealing
Genetic Algorithms

We will start with Maximum Likelihood and χ^2 illustrated by data analysis problem

Optimization Example – V: Analysis of Experimental Data

- Interpolation involves finding function values given function values at other points. This naturally extends to case where given function values have errors.
- We consider the general case where we have a set of observations $x_1, x_2 \ldots x_n$ and we wish to deduce something from it.
  - a) x_i are separate measurements of the length of a rod which differ due to uncertainities in measurement process (measurer drunk, rod quivering in earthquake).
    Goal: What is best estimate for length of rod and error therein.

Analysis of Experimental Data (cont.)

- b) x_i are measurements of particle positions in a detector at the proposed SSC—Superconducting Supercollidor

  Goal: Deduce which is the correct theory of the fundamental interactions. What particles exist?

- c) x_i could be number of people claiming they will vote for Clinton and i could label either geographic position or time sample taken.

  Goal: Deduce who will win Presidential election and what strategy changes are appropriate to affect result.

- In each case, the x_i's are related to "goal" (which can be qualitative or quantitative) by probability distributions.

Data Analysis Examples—I

- Consider three examples defined already. Their characteristics are:
  - (a) Length of Rod
    This is simplest case. Here, the length is well defined and uncertainties come from variations in measurement process itself.
  - (b) High Energy Physics
    The events recorded are fundamentally probabilistic because quantum theory describes events this way, <u>not</u> deterministically. There are no “hidden variables,” such as $\underline{\alpha}$ in example (c).

Data Analysis Examples—II

- (c) Voting

  - We can associate with the voting process, a function V

$$
\begin{aligned}
V &= 0 \qquad \text{doesn't vote} \\
&= 1 \qquad \text{votes for Clinton} \\
&= 2 \qquad \text{votes for Bush} \\
&= 3 \qquad \text{votes for Perot, etc.}
\end{aligned}
$$

  - Each person is described by a set of attributes $\underline{\alpha}$
    $\alpha_1 = \text{age}, \ \alpha_2 = \text{income}, \ \alpha_3 = \text{place they live, etc.}$
    with a big enough size vector $\underline{\alpha}$

$$V \equiv V\left(\underline{\alpha}\right)$$

  - Each person $\rightarrow \underline{\alpha}_i$ a particular vector.
  - If we were mathematicians, we could find $V\left(\underline{\alpha}\right)$ from survey (determine parameters in it) and so predict election result.
  - But society is a "complex system." Impossible to write down some function V even though actions might be described by some sort of function.

$$\Rightarrow \text{use probabilistic approach.}$$

The Computational and Numerical Issues

- 1. The underlying mathematics is that of random variables. The key basic theorem is the
  - CENTRAL LIMIT THEOREM—this is at the level of CPS615.
- 2. There is a set of methods where we believe that we KNOW something precise about associated probability distributions.
  - LEAST SQUARES METHOD
  - MAXIMUM LIKELIHOOD METHOD
- 3. There are a set of techniques for data exploration where we search a large dataset for expected or unexpected correlations and clusterings which are qualitative and usually not describable by mathematical functions.
  - HISTOGRAMS
  - SCATTER PLOTS

<u>SSC Scenario—I</u>

- **Data Analysis at the "SSC"**

  1. A typical problem which is rich enough to exhibit all these issues.

$$\begin{array}{ccl} p \quad \overline{p} & & \rightarrow N \text{ particles} \\ \rightarrow \quad \leftarrow & & N = 2 \rightarrow 1000? \end{array}$$

<u>Particles labelled by</u>

type,		π, k, p, $\overline{p}\ldots$
momentum,		$\underline{p}$, E
and energy, mass^2	$=$	$E^2 - \underline{p}^2$

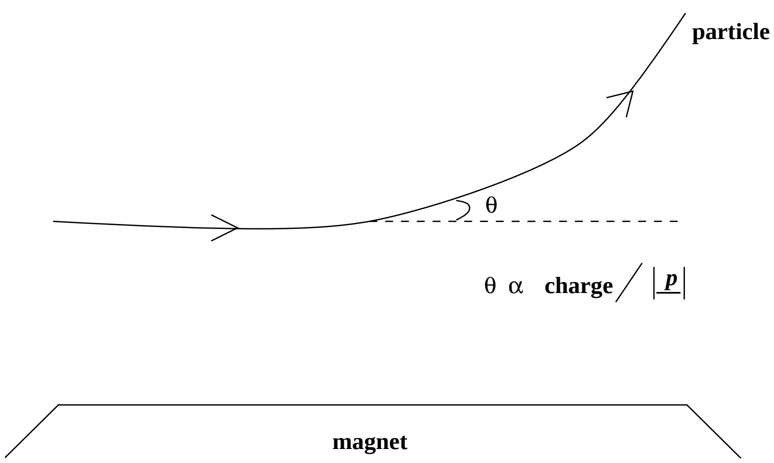
$\uparrow$

essentially determined
by type

- A typical experiment might record 10 million such events on some storage medium (e.g., ~ 3 events per second for a year).
- The events might be "unbiased", i.e., all possible events recorded or they may be selected by a "trigger."

SSC Scenario II

- The events are specified by a set of parameters.
- N = number of particles, and for each particle, parameters defining them. This is a vector $\underline{y}$.
- Each event described by $\underline{y}_i$ produced with probability $p\left(\underline{y}_i\right)$.
- Any realistic detector is imperfect. "Image Processing" is used to find tracks of charged particles. Magnets used to bend charged particles where size of bend $\Rightarrow$ sign of charge, magnitude of momentum.



Uncertainties in SSC Data Analysis

- Flaws in Image Processing

  cf. SDI problem of tracking a bunch of missiles

  "Mathematically ill-posed problem" use neural networks, etc.

- Geometric problems

  $\Rightarrow$ In some regions of momentum $\underline{p}$ particles are "lost" with some probability.

- Neutral particles only make tracks in dense material $\Rightarrow$ harder to measure and "destructive"

  - So first measure charged particles
  - then measure photons (π^0)
  - then measure $\mu^\pm$
  - cannot measure neutrinos

SSC Data Analysis Problem Statement

- What you see = probability from Science convoluted (integrated) with probability from measurement probability from trigger.
- Probability from Science contains functional dependences which are understood together with those that are not. ⇒ this is also a "complex system".
- The observed features =
  $P\left(\underline{y}_i\right) = \int$ known dependencies

  convoluted with unknown dependencies

  together with totally unknown probabilities.
- Can MODEL dependencies and probabilities (use suitably inspired = theoretically motivated "guesswork") ⇒ PARAMETERIZE ignorance or use approaches which do NOT involve parameterization.

Maximum Likelihood Principle—Example

Rod Quivering in an Earthquake—I

- x_1, $x_2 \ldots x_N$ are N measurements of the length of a rod.
- Owing to a Gaussian earthquake (or a Gaussian i.e., normal drunkenness) occurring during measurement, the x_i are scattered about the true value:
- You visit your favorite theorist (or do yourself after sobering up) who <u>believes</u> that
- the probability distribution of x_i is Gaussian with mean α_0 = true length and standard deviation σ_0.
- We may write down the (relative) probability of each measurement x_i as:

$$P_i\left(x_i\right) = c \, \exp\left\{-\frac{1}{2\sigma_0^2}\left(x_i - \alpha_0\right)^2\right\}$$

Rod Quivering in an Earthquake–II

- The Likelihood function for whole experiment of N events is defined as:

$$\begin{aligned} L\left(x_1 \ldots x_N\right) &= \prod_{i=1}^{N} c \exp\left\{-\frac{1}{2\sigma_0^2}\left(x_i-\alpha_0\right)^2\right\} \\ &= \tilde{c} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_i\left(x_i-\alpha_0\right)^2\right\} \end{aligned}$$

- This is—up to a constant—the probability of the x_i given the true length α_0 and certain theoretical assumptions we represent as T.

$$L\left(x_1 \ldots x_N\right) = \Pr\left\{x_1 \ldots x_n | \alpha_0 , T\right\}$$

  in conditional probability notation.

- T is, for instance, the notion that distribution Gaussian and the value of σ_0 taken (the latter being calculable from the strength of earthquake or amount of beer drunk).

Bayes Formula

- Use of Bayes Formula

  Now what we really want is the probability distribution of an estimate α of α_0. We can formally get this by use of Bayes law:

- First extend $P(x_i) = \exp\left\{-\frac{1}{2\sigma_0^2}(x_i - \alpha)^2\right\}$ to be defined for any value of α: not necessarily $\alpha =$ true value α_0. Introduce:

- $P(\alpha | x_i, T) =$ Probability true length is α given measurements x_i and predjudice T. Then Bayes law of conditional probability is

$$P(\alpha | x_i, T) = P(x_i | \alpha, T) \frac{P(\alpha | T)}{P(x_i | T)} \qquad (*)$$

- This is the essence of the maximum likelihood principle. $(*)$ converts the likelihood $L(x_1 \ldots x_n) = P(x_i | \alpha, T)$ which is a function probability of x_i for fixed α, T into $P(\alpha | x_i, T)$ which is the desired probability distribution of α for given measurements x_i and predjudice T.

Interpretation of Bayes Relation – I

- Interpretation of equation $(*)$ on previous page
- (α) We can ignore $P\left(x_i|T\right)$ because it is independent of α and fixed for our given x_i and T. It has the practical point that Likelihood is never normalized sensibly.
- (β) $P(\alpha|T)$ philosophically represents a priori knowledge of α before experiment (measurement of x_i) performed.
- This is shaky—not because we don't really know how to define what $|T$ ("given T") means—but, for instance, suppose we wish to parameterize "complete ignorance" of α before experiment performed.

Interpretation of Bayes Relation – II

- Natural is $P(\alpha|T) = \text{constant}$
- But suppose we put $\beta = \alpha^2$ and then β has a priori probability $\varphi(\beta|T)$

$$\varphi(\beta|T)d\beta = P(\alpha|T)d\alpha$$

  i.e., $\varphi(\beta|T) = \frac{1}{2\sqrt{\beta}} \neq$ constant.
- So the concept of "complete ignorance" is not well defined. Jeffrey (a big bad English-Oxford-book) spends 100's of pages trying to avoid this difficulty. His efforts are so much garbage.
- Fortunately, there is no trouble for large n because of both an obvious numerical reason (see next section (i) (d)) and a not so obvious mathematical reason.

Basic Likelihood Formula

- The Likelihood Maximizes

  Consider, again, the explicit form of the Likelihood of a rod quivering in the earthquake.

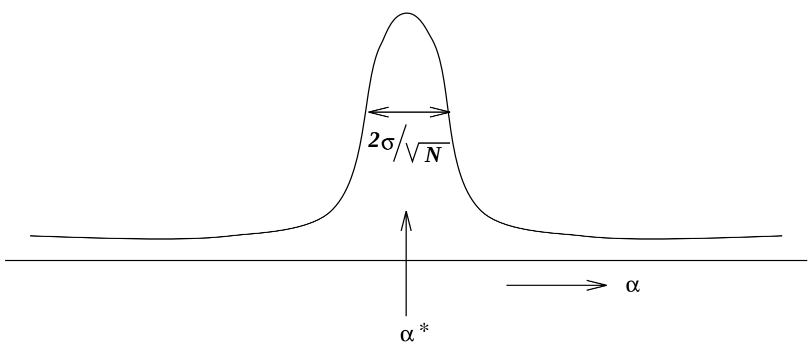
$$L\left(x_1 \ldots x_N\right) = c \, \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(x_i - \alpha\right)^2\right\}$$

- According to Bayes Principle (*), we now regard L as a function of α and so rewrite it:

$$\begin{aligned} L(\alpha) \;=\; c \, \exp\Bigg\{ & -\frac{N}{2\sigma^2}\left(\alpha - \frac{1}{N}\sum_{i=1}^{n} x_i\right)^2 \\ & - \frac{1}{2\sigma^2}\left[\sum_i x_i^2 - \frac{1}{N}\left(\sum_i x_i\right)^2\right]\Bigg\} \end{aligned}$$

Pictorial View of Maximum Likelihood

- The last term is a constant independent of α, and so as a function of α, L is just a Gaussian which maximizes at $\alpha = \frac{1}{N}\sum_{i=1}^{N} x_i$ and has width $\frac{\sigma}{\sqrt{N}}$.



- So as $L = \Pr(\alpha | x_i, T)$ we say that maximum likelihood estimate of α is $\alpha^* = \frac{1}{N}\sum_{i=1}^{N} x_i$ and its standard deviation is $\sigma/\sqrt{N}$.
- Note that whatever $P(\alpha)$ you put in the shape of $LP(\alpha|T)$ is essentially independent of P as $N \to \infty$ and we get a Gaussian peaked at α^* riding on top of constant value $P(\alpha^*|T)$.

General Maximum Likelihood Method

- The above is glib in many ways—especially in its treatment of errors. We will repair this later with real live theorems. First, we state the obvious generalization.
- The general maximum likelihood method is to put $P(\alpha|T) = 1$ in Bayes Principle (*) and form $L(\alpha) = \Pi_{i=1}^{N} P\left(x_i\,, \alpha\right)$ for any measurement x_i whose probability $P\left(x_i\,, \alpha\right)$ depends on value x_i and theoretical parameter(s) α.
- The maximum likelihood estimate of α is $\alpha = \alpha^*$ where $L\left(\alpha^*\right)$ is a maximum. The error $\Delta\alpha$ is the standard deviation of the distribution of L as a function of α, i.e.,

$$L(\alpha) = \exp\left\{\text{const.} - \frac{1}{2\Delta\alpha^2}\left(\alpha - \alpha^*\right)^2\right\}$$

- Note one usually manipulates

$$-\ln L = -\sum_{i=1}^{N} \ln P\left(\underline{x}_i,\, \underline{\alpha}\right)$$

Lifetime Example for Maximum Likelihood—I

- Physics texts give the example of determining the lifetime τ_0 of a particle from the observation of decay times $t_1 \ldots t_N$ of N decays.
- The probability distribution is Poisson:

$$p(t)dt = \frac{1}{\tau} \exp\left(-\frac{t}{\tau}\right) dt$$

- Now dt is meaningless—it is independent of τ and only affects normalization of likelihood. The unknown $P(\alpha|T)$, $P\left(x_i|T\right)$ terms in (B1) made this arbitrary anyway!
- Likelihood $L(\tau) \propto \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \ldots \frac{1}{\tau} e^{-\frac{t_N}{\tau}}$ or $L(\tau) = \exp\left[-\frac{1}{\tau} \sum_{i=1}^{N} t_i - N \ln \tau\right]$ where we lazily convert proportionality into an = sign. Maximum is

$$\begin{aligned} \frac{dL}{d\tau} &= 0 \qquad \text{or} \\ \frac{1}{\tau^{*2}} \sum_{i=1}^{N} t_i - \frac{N}{\tau^*} &= 0 \\ \underline{\text{or}} \qquad \tau^* &= \frac{1}{N} \sum_{i=1}^{N} t_i \end{aligned}$$

Lifetime Example for Maximum Likelihood – II

- i.e., the maximum likelihood estimate is just the mean: such a simple result is only true because of particular Poisson example.
- In this special case, we can illustrate that maximum likelihood method does give results which "converge" as $N \to \infty$.
- According to the central limit theorem, the variable $y = \frac{1}{N}\sum_{i=1}^{N} t_i$ is a sum over independent random variables and has mean $\langle y \rangle = \tau_0$ and standard deviation σ

$$\begin{aligned}
\sigma^2 &= \frac{1}{N}\left\{\left\langle t_i^2 \right\rangle - \left\langle t_i \right\rangle^2\right\} \\
&= \frac{1}{N}\left\{\frac{1}{\tau_0}\int t^2 e^{-\frac{t}{\tau_0}} dt - \tau_0^2\right\} \\
\text{or} \quad \sigma &= \frac{\tau_0}{\sqrt{N}}
\end{aligned}$$

Multiplication of Experiments

- If we have two experiments that are independent but depend on same theoretical parameter α, then
  - $L_1 = \Pi_{i=1}^{N_1} p\,(x_i\,, \alpha)$ is likelihood of experiment #1
  - $L_2 = \Pi_{i=1}^{N_2} q\,(y_i\,, \alpha)$ is likelihood of experiment #2
- and clearly the combined likelihood is

$$L = L_1 \; L_2$$

- i.e., one should multiply likelihoods when combining experiments. We will apply this in the next section (χ^2) to case when each L_i is Gaussian in α.

General Maximum Likelihood

- Note that one can prove that Maximum Likelihood is asymptotically the "best algorithm," i.e., in limit $N \to \infty$ (N total events, there is NO method which will give a smaller standard deviation).
- Also, Maximum Likelihood method will give one the "right answer," i.e., Bias = Correct Estimate—Maximum Likelihood Estimate $\to 0$ (like $1/N$) as $N \to \infty$.

χ^2 Method

- This is standard and fully, correct (best) method for combining results from several independent experiments. It is in fact used (because it's easy conceptually and for computer) even when it's inapplicable (e.g., if there are correlations between supposed "independent" random variables or if the basic experiments have too few events to be Gaussian).
- The problem is: Given N experimental measurements x_i and errors σ_i plus estimates ξ_i of the x_i which are functions of N theoretical parameters. Then the probability of each observation is

$$P_i = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{ -\frac{(x_i - \xi_i)^2}{2\sigma_i^2} \right\}$$

χ^2 Method Derived from Likelihood

- The Likelihood

$$L = \prod_{i=1}^{N} P_i \propto \exp\left\{ - \sum_{i=1}^{N} \frac{(x_i - \xi_i)^2}{2\sigma_i^2} \right\}$$

- so, the principle of maximizing the likelihood is equivalent to minimizing χ^2 where

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - \xi_i)^2}{\sigma_i^2}$$

  with respect to N theoretical parameters in ξ_i.

- (i) Notice that χ^2 is the sum $\sum_{i=1}^{N} v_i^2$ where v_i are independent Gaussian random variables of unit standard deviation and zero mean.
- We <u>can</u> calculate χ^2 distribution exactly, but in practice this is rarely necessary. Thus, we can easily show that—for fixed ξ_i—that

$$\begin{aligned} \langle \chi^2 \rangle &= N \\ \sigma_{\chi^2}^2 &= 2N \ \ (**) \end{aligned}$$

- and this is sufficient for large N as central limit theorem says χ^2 will be Gaussian with above mean and standard deviation.

χ^2 Standard Deviation

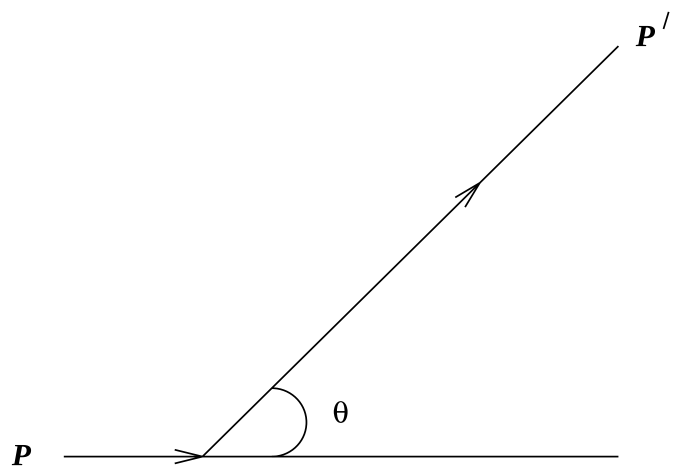
(ii) Now $(**)$ is not quite correct because if we determine the theoretical parameters to minimize χ^2 then implicitly ξ_i are also random variables that are functions of x_j.

- It is not too hard to show that in this case, one should just replace N by $N - n$ (number of degrees of freedom) in $(**)$.
- In any case, $(**)$ is very important because it allows an easy criterion for the goodness of fit.
- Thus, if value of χ^2 is many (say > 2) standard deviations away from its mean, then the theoretical form is in doubt.
- The usual maximum likelihood method does not have this advantage:
- Remember, L was unnormalized and so <u>its</u> value had no significance for the goodness of fit. This is quite a difficult problem.

Scattering Experiment Example

Examples:

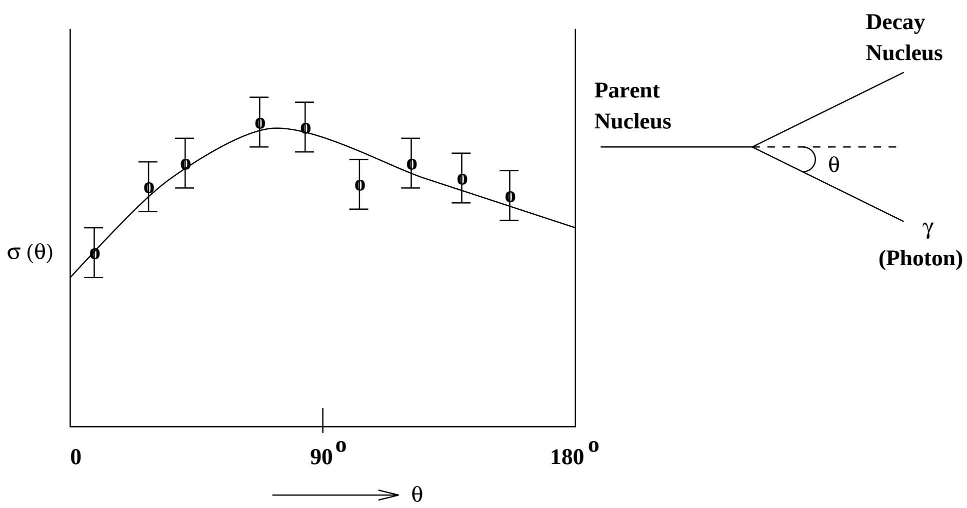
1) A physics experiment measures a scattering cross-section σ as a function of angle θ.



In the case of decay (of nucleus), we know rigorously that angular distribution $\sigma(\theta)$ is a polynomial in $\cos\theta$ with highest degree $\leq 2J$ (spin of nucleus)

$$\sigma(\theta) = a_0 + a_1 \cos\theta + a_2 \cos^2\theta + a_3 \cos^3\theta$$

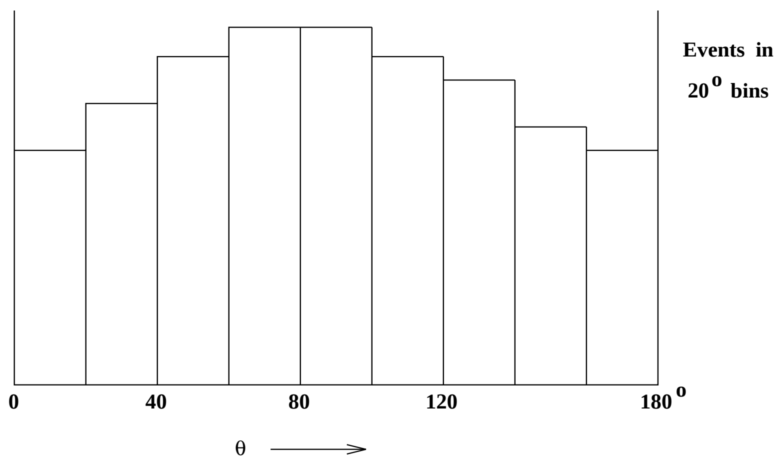
Individual Data Points



- There are two typical forms of data—firstly, isolated points.
- The measurements x_i's are cross sections and the theoretical predictions ξ_i are linear in unknown parameters a_j.

$$\xi_i = \sum_{m=1}^{n} c_{im}\, a_m$$

Histogrammed Data



- Now some experiments measure as suggested in previous picture at a set of discrete angles.
- More common is to measure a more or less complete range of θ but bin these, so that as in picture, basic measurement is number of events N_i in i'th "bin."
- Here, we have nine bins in $0 \leq \theta \leq 180$.

$$x_i = N_i \text{ (up to constant)}$$
$$\text{error in } x_i \text{ is } \sqrt{N_i}$$

Scattering Experiment—χ^2 Minimization

- Under these circumstances, the χ^2 method becomes:

$$
\begin{aligned}
\text{minimize:} \quad & \sum_{i=1}^{N} \frac{(x_i - \xi_i)^2}{\sigma_i^2} \qquad - (\chi^2) \\
x_i &= \text{const. } N_i \\
\sigma_i &= \text{const. } \sqrt{N_i} \qquad \sigma_i = \frac{x_i}{\sqrt{N_i}}
\end{aligned}
$$

  where ith bin has N_i events. Note estimated error comes from measurement itself.

- Minimizing (χ^2) with respect to parameters a_m gives

$$
\begin{aligned}
\sum_{i=1}^{N} \frac{(x_i - \xi_i)}{\sigma_i^2} \frac{\partial \xi_i}{\partial a_m} &= 0 \qquad \text{or} \\
\sum_l A_{ml}\, a_l^* &= b_m
\end{aligned}
$$

  where * denotes χ^2 (or maximum likelihood) estimate.

χ^2 Formulae

- The χ^2 matrix

$$a_{ml} = \sum_i \frac{C_{im} C_{il}}{\sigma_i^2}$$

- which is a Positive Symmetric Matrix

$$= \sum_i \frac{(\partial \xi_i / \partial a_m) \; (\partial \xi_i / \partial a_l)}{\sigma_i^2}$$

- The right hand side

$$b_m = \sum_i \frac{x_i C_{im}}{\sigma_i^2} = \sum_i \frac{x_i}{\sigma_i^2} \frac{\partial \xi_i}{\partial a_m}$$

- Therefore, $\underline{a}^* = A^{-1} \underline{b}$ is our best estimate for the parameters $\underline{a}$.

Errors in χ^2

- We can find errors in $\underline{a}$ in either of two ways:
  - 1. regard $\underline{a}^*$ as a function of x_i and calculate standard deviation from central limit theorem.
  - 2. shape of likelihood

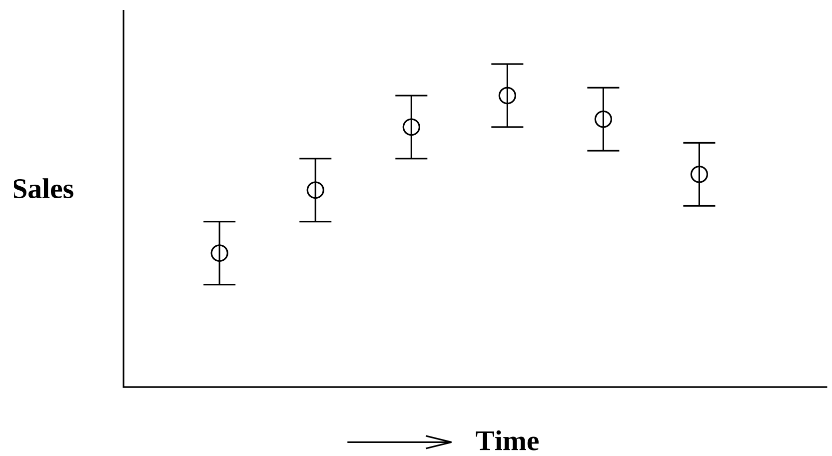
$$L = \exp\left(-\frac{1}{2}\chi^2\right)$$

- We have $-$ as far as shape of L is concerned.

  $L = \exp$ (Independent of a plus no terms linear in $a - a^*$ due to "maximum" condition $-1/2\,(\underline{a} - \underline{a}^*)^T A\,(\underline{a} - \underline{a}^*)$).
- i.e., this above formula implies that A is error matrix for a.

Example II – Regression Analysis

- In typical regression analysis the ξ_i are also linear in the unknowns, e.g.,



$$\text{Sales}(t) = \sum_{i=1}^{m} x_i P_i(t)$$

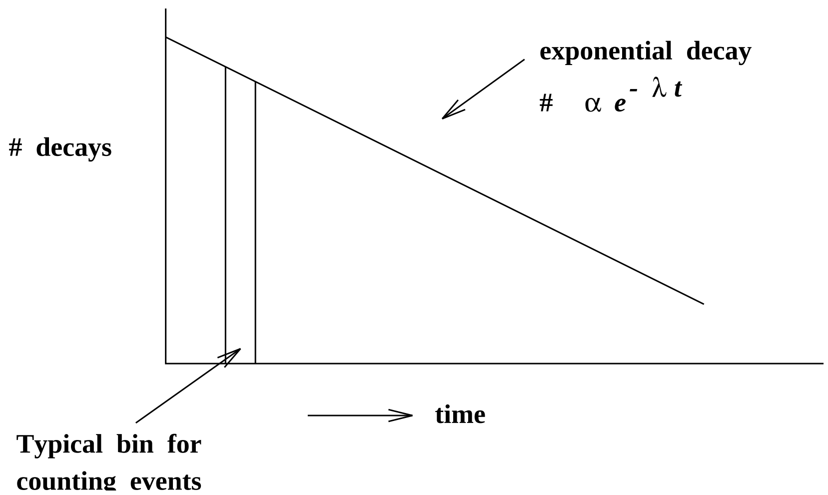
- where x_i is unknown and P_i is known expansion function.

$$\begin{aligned} \text{e.g.,}\quad P_1(t) &= 1 \\ P_2(t) &= (t - 1992) \\ P_3(t) &= (t - 1992)^2, \text{ etc.} \end{aligned}$$

- $P_k(t)$ isn't linear in t, but Sales(t) is linear in $P_k(t)$ as this is what counts!

Example III – Nuclear Decay

- However, we consider nuclear decay



- e_i = number of decays observed in time interval $t_0 + (i-1)\delta t$ to $t_0 + i\,\delta t$.

$$\xi_i = c\,\exp\left(-\lambda[t_o + (i - 1/2)]\delta t\right)$$

  is a function of two unknowns, c and λ.

- ξ is linear in c but nonlinear in λ.
- Note we can analyze nuclear decays in two ways—maximum likelihood or χ^2. Let us compare these.

Maximum Likelihood and χ^2 for Experimental Data Analysis

- a) **Maximum Likelihood**

  i labels events $\quad i = 1 \ldots n_{ev}$

  - Minimize—$\ln L$

$$L = \prod_{i=1}^{n_{ev}} \frac{1}{\tau} c \exp\left(-\lambda t_i\right)$$

- b) χ^2

  i labels bins $\quad (i = 1 \ldots N_{\mathrm{bin}})$ in time

  - Minimize χ^2

$$\chi^2 = \sum_{i=1}^{N_{\mathrm{bin}}} \frac{\left(N_i - \xi_i\right)^2}{\sigma_i^2}$$

  - where $\sigma_i \sim \sqrt{N_i}$ and N_i is number of events in i th bin $(t_0 + (i-1)\delta t \leq t \leq t_0)$.
  - If N_i large, N_i will be Gaussian, whatever distribution of individual events.

When Should You Use χ^2 and When Maximum Likelihood?

- Which is best? Likelihood method is "best," but χ^2 method is computationally easier as $N_{\text{bin}} << N_{\text{ev}}$.
- Typically, as number of events in each bin must be large
$$N_{\text{bin}} \sim \frac{N_{\text{ev}}}{50}\,.$$
- χ^2 method will give an estimate with a larger standard deviation, but the difference will not be large if $\lambda\delta t$ is small, so probabilities $\exp(-\lambda t)$ do not vary across bin. So, can choose correctly
  - Small total number of events—no good method.
  - Large number of events–Maximum Likelihood works and is best.
  - Very large number of events—Maximum Likelihood and χ^2 works, but χ^2 will give a good answer with less computational complexity.

Least Squares

- Generally, we analyze

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{e_i - \xi_i}{\sigma_i} \right)^2$$
$$= \sum_{i=1}^{N} \text{res}_i^2 \qquad \leftarrow \text{ scaled residual}$$

which shows why χ^2 is also called least squares fit. As in above examples, expand ξ_i about $\underline{x}_0$

$$\xi(\underline{x}) = \xi(\underline{x}_0) + (\underline{x} - \underline{x}_0)^T \frac{\partial \xi}{\partial \underline{x}} + \text{Higher order in } (\underline{x} - \underline{x}_0)$$

- In this approximation, we now get "standard linear formulae."

$$\chi^2(\underline{x}) = \chi^2(\underline{x}_0) + \frac{1}{2}(\underline{x} - \underline{x}_0)^T A (\underline{x} - \underline{x}_0) - (\underline{x} - \underline{x}_0)^T \underline{b} + \text{corrections}$$

- Minimize χ^2 ignoring corrections gives

$$\underline{\underline{A}}(\underline{x} - \underline{x}_0) = \underline{b}$$

which is same equation we found for iterative solution of P.D.E. **BUT** only exact when ξ_i LINEAR in unknown parameters $\underline{x}$.

Nonlinear χ^2 Minimization

- $\underline{\underline{A}}$ is symmetric positive definitive

$$
\begin{aligned}
\xi_i &= \xi_i(\underline{x}_0) + \sum_{d=1}^{m} t_{id}(x - x_0)_d \\
\text{where} \quad t_{id} &= \frac{\partial \xi_i}{\partial x_d} \\
A_{dd'} &= \sum_i t_{id} t_{id'}
\end{aligned}
$$

- Basic idea:
  - Think of initial $\underline{x}_0$ and establish recursive method
$$\underline{\underline{A}}(\underline{x}_{k+1} - \underline{x}_k) = \underline{b}$$
  - Calculate $\underline{\underline{A}}$ and $\underline{b}$ using $\underline{x} = \underline{x}_k$.

$$\underline{x}_{k+1} = \underline{x}_k + A^{-1} b \qquad \text{(NL.1)}$$

- As A positive definite, inverse always exists.

Validity of Linearization When Does It Work?

- Why doesn't this work?

Diagonalize **Transformed**

$$\begin{bmatrix} \lambda_1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} \delta\tilde{x}_1 \\ \\ \\ \delta\tilde{x}_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \\ \\ b_n \end{bmatrix}$$

- $\lambda_1 > \ldots > \lambda_n$.
- My experience at $n \sim 30$, $\lambda_n \sim 10^{-8}\lambda_1$.

  But we made approximation and so λ_n term swamped by corrections neglected in expanding ξ_i linearly. Equation (NL.2) gives big changes

$$\delta\tilde{x}_m = \lambda_m^{-1} b_m \qquad \text{(NL.2)}$$

  Better is to use

$$\delta\tilde{x}_m = 0 \qquad \text{(NL.3)}$$

  And, use of (NL.2) usually gives $\chi^2(\underline{x}_{k+1}) >> \chi^2(\underline{k}_n)$. Not $\chi^2(x_{k+1}) < \chi^2(\underline{x}_k)$

Nonlinear Minimization That Works!

- Some solutions to this problem are
- 1) **Conjugate Gradient**, but note matrix no longer sparse and so iterative approach not so natural.
- 2) **Marquardt's method** changes A in ad-hoc fashion, solving (NL.1) with the replaced value

$$\underline{\underline{A}} \rightarrow \underline{\underline{A}} + QI$$

where Q is Marquardt's Parameter

$$\lambda_i \rightarrow \lambda_i + Q$$

- Suppose, $Q \sim \sqrt{\lambda_1 \lambda_n}$. Then shifts $\delta\tilde{x}_d$ are unchanged for large λ_d, small for small λ_d.
- Increase Q if χ^2 increases when you use (NL.1).
- Decrease Q if χ^2 decreases when you use (NL.1).

Linear Programming

- Minimize $z = \Sigma_{j=1}^{r} c_j x_j$ where $x_j \geq 0$ and there are Linear Constraints

$$\sum_{j=1}^{r} a_{ij}\, x_j \qquad \leq, \geq, \text{or} = b_i \qquad i = 1 \ldots m$$

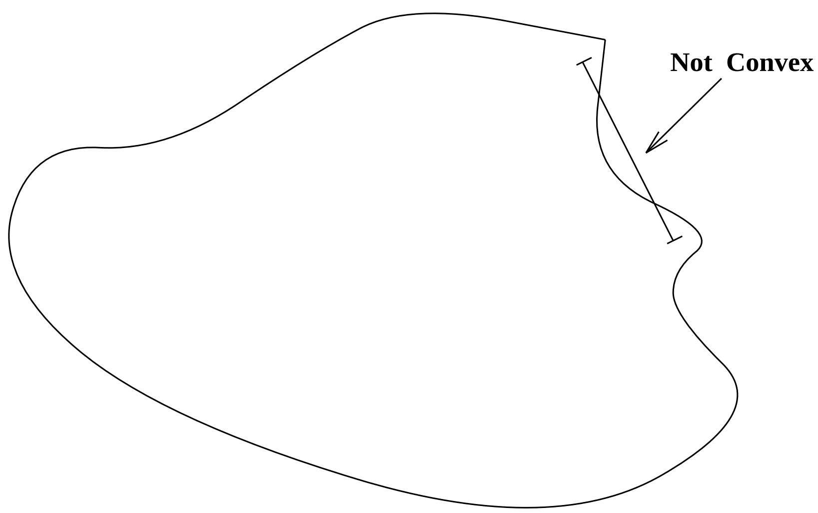
- Introduce new variables to convert **all** constraints to equalities.
- Suppose $i = 1$ is $\geq$. Write $i = 1$ constraint as

$$x_{r+1} = \sum_{j=1}^{r} a_{ij} x_j - b_1 \qquad \geq 0$$

$$\begin{aligned} \sum_{j=1}^{r} a_{ij} x_j - x_{r+1} &= b_1 \\ a_{1r+1} &= -1 \\ c_{r+1} &= 0 \end{aligned}$$

- So linear programming becomes:
  - Minimize $z = \underline{cx}$
  - Subject to $\underline{\underline{A}}\underline{x} = \underline{b}$ as well as that all components of $\underline{x}$ are positive.

Convex Regions and Linear Programming



- Set of $\underline{x}$ satisfying constraints are convex region.
- Convex says that if $\underline{x}_1$ and $\underline{x}_2$ in a region, $\lambda \underline{x}_1 + (1-\lambda)\underline{x}_2$ in region.
- Extreme point of convex region is one where

$$x_{\text{extreme}} \neq \lambda \underline{x}_1 + (1-\lambda)\underline{x}_2$$

  where $\lambda \neq 0, 1$.
- One can show that the optimal solution (if it exists) is an extreme point of the convex region.

Matrix Formulation of Linear Programming

- Consider $\underline{\underline{A}}\underline{x} = \underline{b}$. $\underline{\underline{A}}$ is $m \times r$ matrix. Usually $r >> m$.
- Can show that extreme points satisfy

  m nonzero elements in $\underline{x}$

  $n - m$ zero elements in $\underline{x}$
- Choose columns so non-zero elements are first m entries in $\underline{x}$

$$\underline{\underline{A}}^{\text{reduced}}\underline{x} = \underline{b}$$

  where $\underline{\underline{A}}^{\text{reduced}}$ is $m \times m$.
- Simplex method is a strategy for moving from one solution $\underline{x}^{\text{extreme}}$ to another $\underline{x}^{\text{extreme}}$ by changing one column of $\underline{\underline{A}}^{\text{reduced}}$ to another that is not being used.
- There is another class of methods—associated with Karmarkar at AT and T which are so called interior point methods. That is, you do not go from vertex to vertex of simplex but rather wander around inside simplex. Both simplex and interior point methods are important.

Computational Issues in χ^2 and Maximum Likelihood

- Take Maximum Likelihood for definiteness
  - 1. Complex nonlinear (in dependence of parameters) function to maximize
  - 2. Each function value

    $$-\ln L = -\sum_{i=1}^{N} \ln p\left(\underline{x}_i, \underline{\alpha}\right)$$

    (where $\underline{x}_i$ determines events and $\underline{\alpha}$ are unknown parameters) is sum of N independent calculations which is natural form of parallelism
  - 3. Parallelism in optimization (in space of $\underline{\alpha}$) depends on particular minimization approach adopted. Also depends on complexity of calculation

    $$\ln p\left(\underline{x}_i, \underline{\alpha}\right)$$

    $$\frac{\partial}{\partial \underline{\alpha}} \ln p\left(\underline{x}_i, \underline{\alpha}\right)$$

    versus complexity of manipulation of $-LnL\left(\underline{\alpha}\right)$ as a function of $\underline{\alpha}$
  - 4. χ^2 and Likelihood are similar in aspects of analysis.

Matrices in χ^2 Maximum Likelihood

- The arithmetic in "$\underline{\alpha}$ world" is that of full matrices which are $n \times n$ where n is number of parameters.
- In "linear" case (ξ_i linear in components of $\underline{\alpha}$, χ^2 quadratic in components), minimization requires single solution of equations

$$\underline{\underline{M}}\alpha = \underline{b} \qquad (**)$$

- In "nonlinear" case (ξ_i nonlinear in α), one must iteratively solve $(**)$.
- Note that one expands ξ_i up to linear term which gives a quadratic approximations to χ^2.
- Thus, need to calculate

$$\text{“value,”} \quad \frac{\partial}{\partial \alpha_k} \text{ “value”} \qquad \text{NOT} \quad \frac{\partial^2}{\partial \alpha_k \, \partial \alpha_j} \text{ “value”}$$

- where "value"

$$\begin{aligned} &= \ln p\left(\underline{x}_i\right) \text{for Likelihood method} \\ &= \xi_i\left(\underline{x}_i\right) \text{for } \chi^2 \text{method} \end{aligned}$$

Basic Software Structure for Derivative Calculation

Often, values are formed inside loops. Inside loop over events

```
      DO 1    iparm = 1, nparm
 1            deriv(iparm) = 0
              value = 0
      DO 2    ipart = 1

              value = value + part (ipart)

      DO 2    iparm = 1, nparm
 2    deriv(iparm) = deriv(iparm) +  ∂ / ∂ α_iparm
 Inside loop over events
```

This code can get rather tedious. `ADIFOR` formally differentiates "existing" FORTRAN code and uses "chain rule" to generate derivative calculations.

Computational Complexity

- Maximum Likelihood VERY VERY expensive as

$$\sum_{i=1}^{\#\ \text{events}}$$

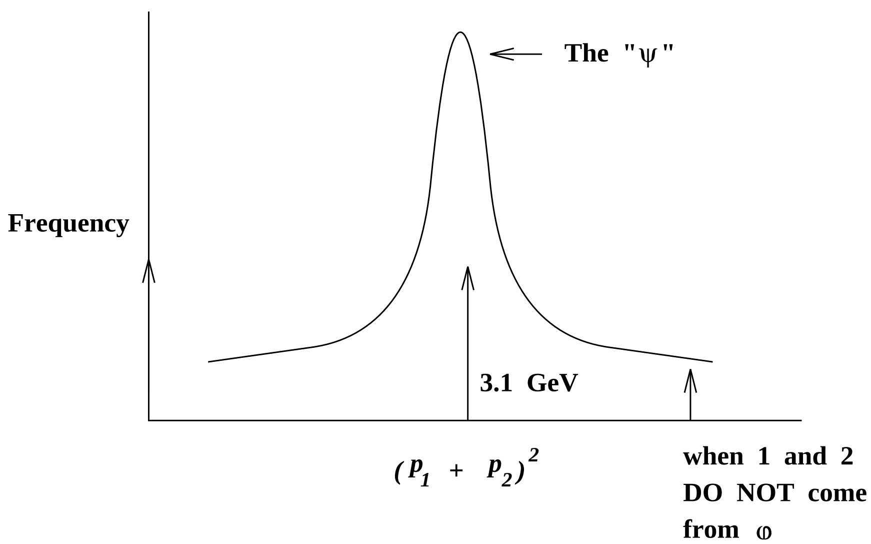
- Remember that the number of events could be 10^7 at "SSC". χ^2 is cheaper as

$$\sum_{i=1}^{\#\ \text{experiments or bins}}$$

- Usually, number of experiments (bins) measured in $10 \rightarrow 10^3$ range.

Histograms

- High energy physicists like to find new particles.
- For example, Nobel prize was given for the discover of ψ particle which is formed from "charm" and "anticharm" quarks.
- Mass of $\varphi = 3.1$ GeV $= (p_1 + p_2)^2$ where ψ decays into particle 1 plus particle 2 where particle i has momentum $\underline{p}_i$.
- We measure p_1 and p_2 as well as lots of other things and plot



- There are an accumulation of events with $(p_1 + p_2)^2$ near 3.1 GeV.

Histogram Formalism I

- We can see with a histogram. Let

$$M_0^2 \leq (p_1 + p_2)^2 \leq M_1^2$$

be a region of interest.

- We have B Bins where each has width ΔM^2

$$\Delta M^2 = \frac{\left(M_1^2 - M_0^2\right)}{B}$$

- Let l'th bin lie between $M_0^2 + (l-1)\,\Delta M^2$ and $M_0^2 + l\Delta M^2$. $l = 1 \ldots B$.
- Define Random Variable

$$\begin{aligned} \eta_l\ (\text{event}) &= 1 \quad \text{event lies in } l\,\text{th bin} \\ &= 0 \quad \text{event lies outside } l\,\text{th bin} \end{aligned}$$

- Let $\langle \eta_l \rangle = \mu_l =$ probability that an event will fall in bin.

Histogram Formalism II

- The means and standard deviations

$$\begin{aligned}
\langle\eta_l\rangle &= (1-\mu_l)\cdot \mathrm{O} + \mu_l \cdot 1 = \mu_l \\
\langle\eta_l^2\rangle &= (1-\mu_l)\cdot \mathrm{O}^2 + \mu_l \cdot 1^2 = \mu_l \\
\langle\eta_l^2\rangle - \langle\eta_l\rangle^2 &= \mu_l - \mu_l^2 \qquad\qquad \text{(Fulle)} \\
&= \mu_l \quad \text{if} \quad \mu_l << 1 \qquad \text{(Approxe)}
\end{aligned}$$

- Standard deviation

$$\sigma_l = \sqrt{\mu_l} = \sqrt{\text{mean}_l}$$

- Define

$$\begin{aligned}
N_l &= \sum_{i=1}^{\#\ \text{events}} \eta_l \;\; (\text{event } i) \\
&= \#\ \text{events in } l\text{th bin}
\end{aligned}$$

- The central limit theorem applies to N_l or to

$$f_l = \frac{N_l}{N_{\text{total}}} \leftarrow \text{Total \# events}$$

- This gives $\langle f_l\rangle = \mu_l$ and f_l is Gaussianly distributed with error

$$\langle\sigma_{\eta_l}\rangle = \frac{\langle\sigma_{\eta l}\rangle}{\sqrt{N_{\text{total}}}}$$

Errors of Histograms – I

- Note this implies that N_l:

  a) is GAUSSIAN

  b) has Standard deviation

$$= \sqrt{N_{\text{total}}} \langle \sigma_{\eta_l} \rangle \;=\; \sqrt{N_{\text{total}} \text{ mean } (\eta_l)} \;=\; \sqrt{\langle N_l \rangle}$$

- This is a very very important rule in all SAMPLING EXPERIMENTS.

  e.g.,

  $l = 1$ vote for Clinton

  $l = 2$ vote for Perot, etc.

Errors of Histograms – II

- In a poll of 1,000 people, one finds that 25% are in lth bin (i.e., will vote for xyz).

$$\begin{aligned} N_l &= 250 \\ \sqrt{N_l} &\approx 16 \end{aligned}$$

- Thus, error in estimate of fraction voting for xyz is <u>16%</u>—this is STATISTICAL error only.
- The systematic bias (is "1,000" chosen people representative) is a harder problem!
- As $0.25 = \mu_l$ not that small, should reduces statistical error by

$$\sqrt{\frac{0.25 - 0.25^2}{0.25}} = \sqrt{\frac{3}{4}} = 0.86$$

using equations (fulle), not (approxe).

Histograms as Robust Data Exploration

- Histograms are unbiased ways of analyzing data. One does NOT need to know dependence of data on $(p_1 + p_2)^2$.
- General problem is data exploration. We have a set of events in a space of many dimensions
- High Energy Physics has momenta

	p_1	p_2	p_3	p_4	p_5	... and
Particle Type	1	2	3	4	5	etc.

Political Polls

- (Unknown) parameters affecting a person's choice for president.

General Histogram

1. Select subspace

   e.g., all events with > 30 particles or all events with a lepton.

   e.g., all people in phone book whose surname starts with Kr.

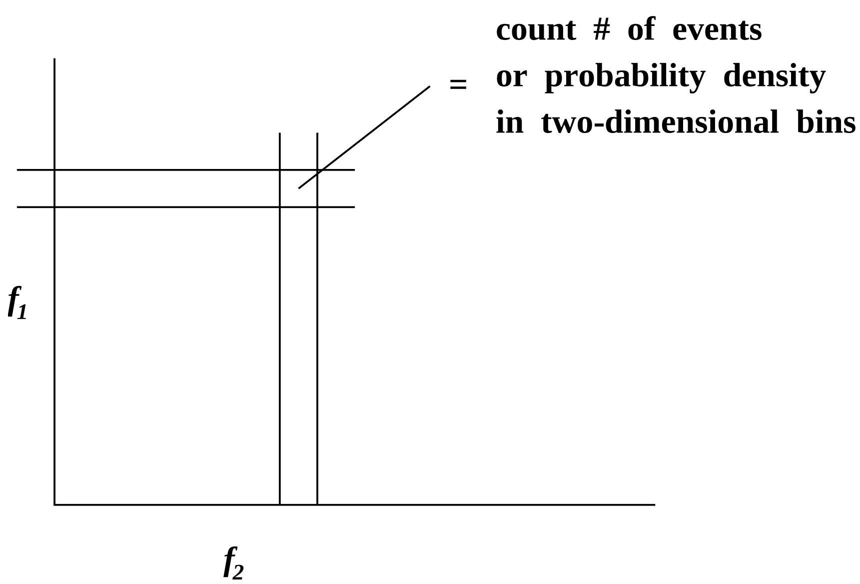
2. Choose some d functions f (event).

$$
\begin{aligned}
i &= 1 \ldots d \\
d &= 1 \\
f_1 &= (p_1 + p_2)^2 \\
\text{or} \quad f_1 &= \text{choice for President}
\end{aligned}
$$

3. Set up set of bins—conventionally uniform in possible values for each f_i.

Scatterplots

- Histogram can be used if $d = 1$.
- If $d = 2$, we find a scatterplot, such as



- where we set up bins in two-dimensional space.
- e.g., f_1 = Religion and f_2 = # of z's in name would be correlated (Catholic Church strong in Poland?)
- e.g., f_2 = Maximum transverse momentum and $f_1 = (p_1 + p_2)^2$ would be correlated.
- as ψ's tend to be produced in high transverse momentum (violent) collisions.
- $d \geq 3$. The future with virtual reality exploration.

Possible Projects I—Maximum Likelihood

I) Produce a general χ^2 or maximum likelihood data analysis system on a parallel machine in either explicit message passing or HPF. Apply to a simple example. You can use Marquardt's method for minimization (see gcf records (my program "Manxcat" performed successfully $10^6 \chi^2$ fits), Numerical Recipes).

- Basic Structure:

Initialize Parameters

For each parameter choice
Initialize data structures
Loop over events
CALL USER1 (I'th event)
accumulate in data structures
Find new parameter values
Are we finished?

Possible Projects II

Separately or as follow-on to I.

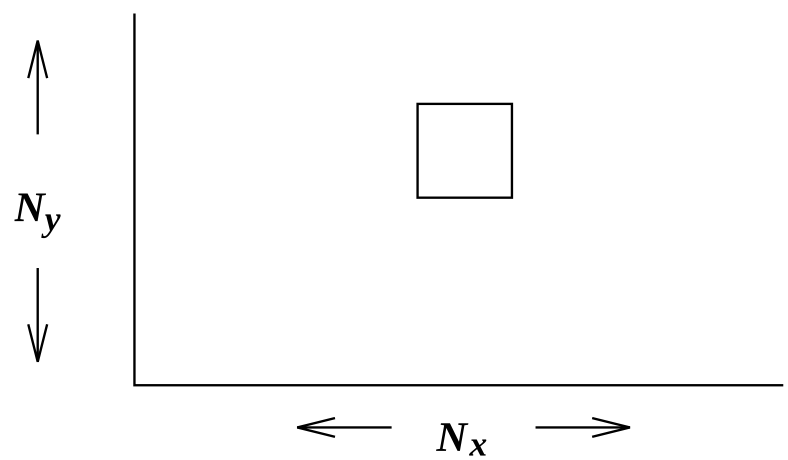
II) Perform fit to specific dataset. Simplest: Generate "fake" data, e.g.,

$$f(x, y) = A \exp\left[-\frac{(x-\alpha)^2}{2\gamma^2} - \frac{(y-\beta)^2}{2\delta^2}\right]$$

- Fixed Parameters are A, α, β, γ, δ
- events—choose x_i, y_i randomly.

$$P\left(\underline{x}\right) \propto F(x, y)$$

- χ^2 — divide two-dimensional plane into subsquares with (see figure) $N_x N_y$ bins. Label bins with i.



Possible Projects II (cont.)

- Number of events in each bin is

$$\begin{aligned} N_{\text{expected}} &\propto \int_{\text{bin}} f(x,y)\,dx\,dy \\ &\sim f\left(x_{\text{central}}, y_{\text{central}}\right) dx\,dy \end{aligned}$$

$$\text{with error} \qquad \sqrt{N_{\text{expected}}}$$

- Assume Total number of events large.
- Generate "fake" experimental measurement as what you get from Gaussian distribution with

  mean $\quad N_{\text{expected}}$

  error $\quad \sqrt{N_{\text{expected}}}$

  using chosen values A, α, β, γ, δ.
- Fit to general form and see that you find "chosen" values for these five parameters.

More Possible Maximum Likelihood Projects

III) Find an interesting data set

— High Energy Physics

— Stock Option Pricing tapes from NPAC

Simplest: each event/χ^2 bin on a separate node of a MIMD parallel machine.

Complex: allow parallelism in calculation of USER1, as in parallel stock pricing model

IV) Compare χ^2 with maximum likelihood

Performance vs. Accuracy Tradeoff

V) Integrate into AVS

VI) Evaluate High Performance Fortran

Projects—Data Exploration

I) Produce a generic histogram/scatterplot/other "robust" data analysis methods package on a parallel machine.

- This should be structured

Initialize Datastructures
(User defines histograms, etc.)

Loop over events

CALL USER2

This has well defined ways of storing data in histograms/scatterplots

e.g., CALL HIST (x, Fred) deposits value x in histogram whose name is Fred

End Loop

Print out histograms

- A Parallel Histogram is a simple database problem. Replicate in each node and combine at end.
- Scatterplots may be too big?
- Implement parallelism for histograms and parallelism for event generation.

More Data Exploration Projects

II) Test approach with some examples

- "Fake data" as with $f(x, y)$ before
- Real data as before
- Theoretical calculation as in "Quantum Monte Carlo" (Cornell expertise) in Chemistry and Physics.

III) Evaluate High Performance Fortran

IV) Integrate into AVS

V) Integrate with visual data exploration—user is walking through a multi-dimensional space looking for expected or unexpected structure.

- Histograms/Scatterplots are one or two dimensions. There is a natural 3D extension which deserves consideration.
- What world to be explored has much higher dimensionality?

VI) Use Mathematica as front end to define selection of subspaces.