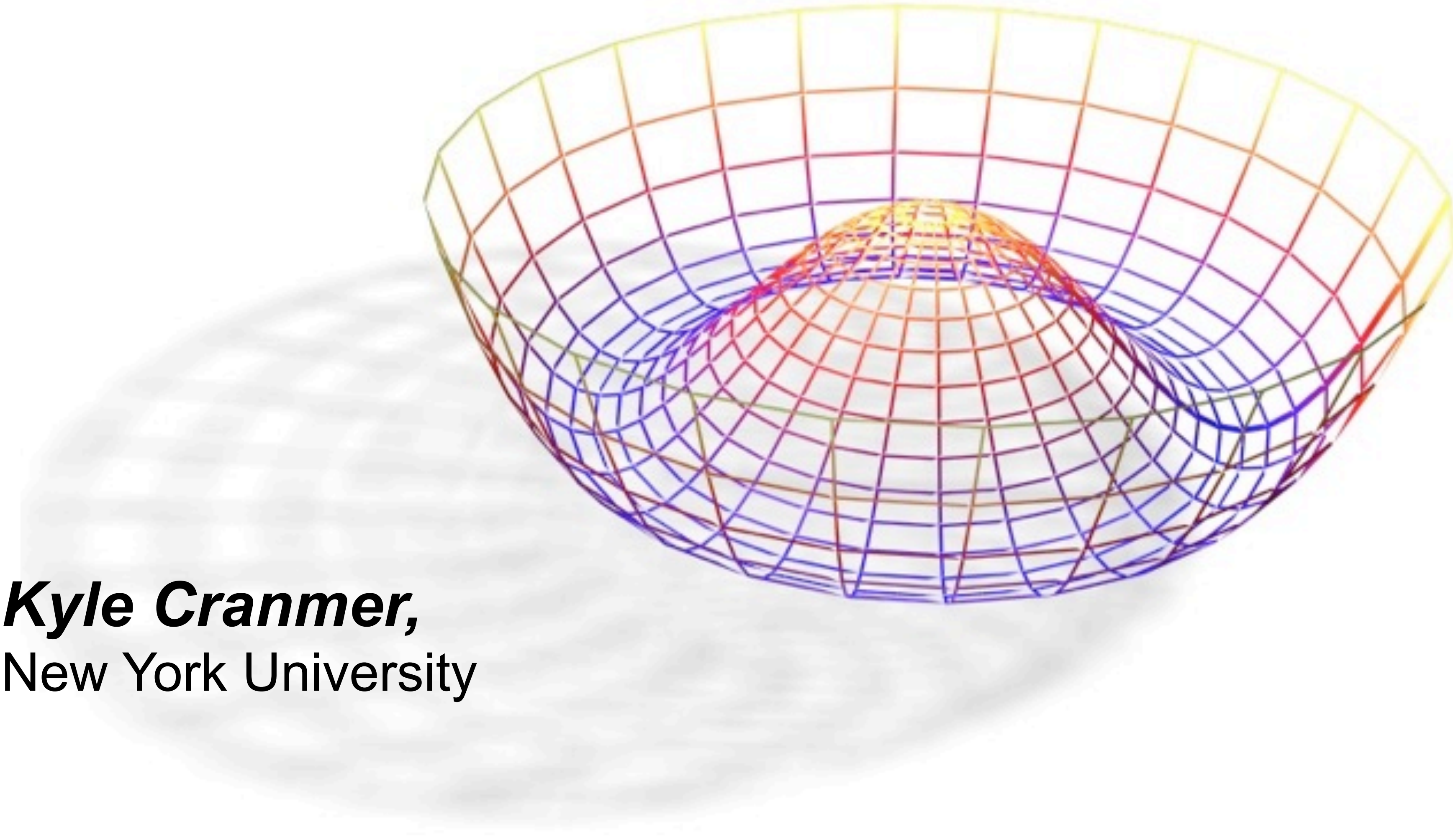
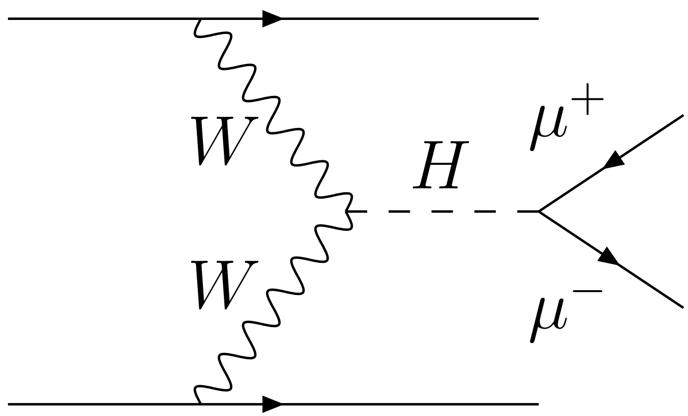
Thoughts on the Matrix Element Method



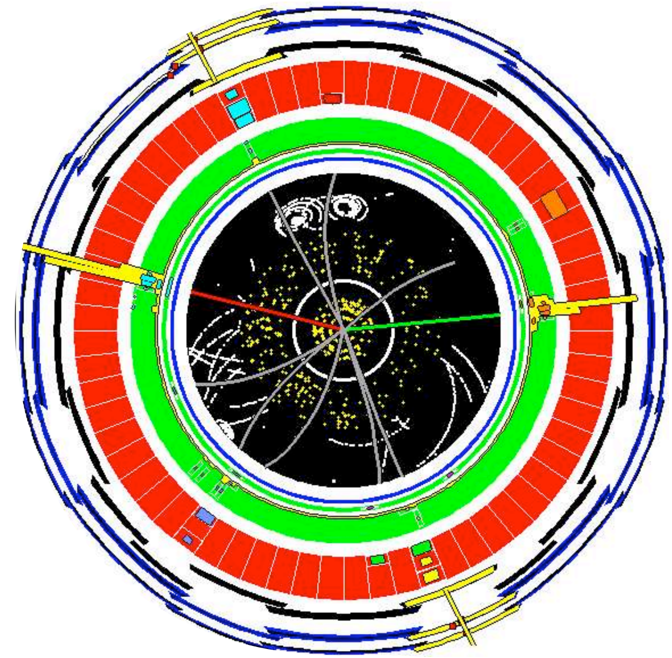
Kyle Cranmer,
New York University

MEM in pictures

$L(x|H_0) =$



$\oplus$



Maximum Significance

In [hep-ph/0605268] Tilman and I used the Neyman-Pearson lemma to establish a formal maximum expected significance using MEM.

- region of the data that maximizes power of a **simple hypothesis test** is given by the likelihood ratio

$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

Expected significance: you don't need to match specific observations $\{x_i\}$.

- the MC integration is always "forward" [generate ϕ, smear via $W(x|\phi)$]

What we really care about computing is the distribution of this ratio, not the numerator or the denominator

- **theme**: instead of computing a cross-section, we compute a formal statistical quantity at some order in perturbation theory

Today: some generalizations of this idea

Marked Poisson Process

Channel: a subset of the data defined by some selection requirements.

- eg. all events with 4 electrons with energy > 10 GeV
- n: number of events observed in the channel
- ν: number of events expected in the channel

Discriminating variable: a property of those events that can be measured and which helps discriminate the signal from background

- for MEM, this is observed kinematics and particle ID information
- $f(x)$: the p.d.f. of the discriminating variable x, ie. $\int d\phi\, |M|^2\, W(x|\phi)$

$$\mathcal{D} = \{x_1, \ldots, x_n\}$$

Marked Poisson Process:

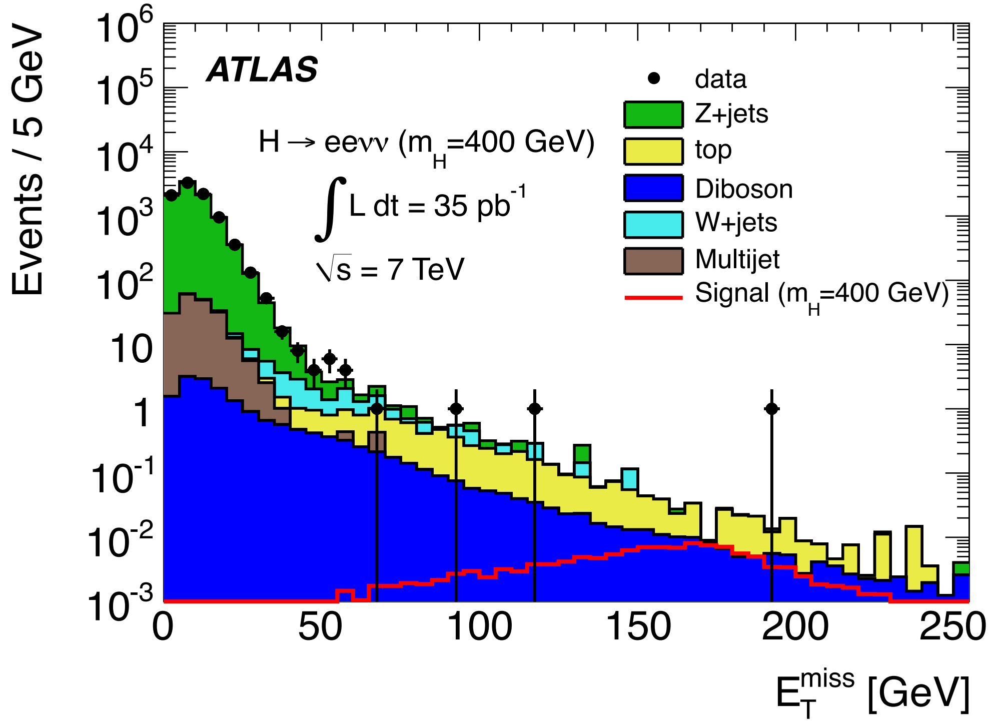
$$\mathbf{f}(\mathcal{D}|\nu) = \mathrm{Pois}(n|\nu) \prod_{e=1}^{n} f(x_e)$$

Mixture model

Sample: a sample of simulated events corresponding to particular type interaction that populates the channel.

- statisticians call this a mixture model

$$f(x) = \frac{1}{\nu_{\rm tot}} \sum_{s \in \rm samples} \nu_s f_s(x) \;, \qquad \nu_{\rm tot} = \sum_{s \in \rm samples} \nu_s$$



Note, *f(x)* is a normalized pdf, so all rate information due to acceptance & tagging encoded in ν

$$\nu = L\sigma = L \int d\phi |\mathcal{M}(\phi)|^2 W(x|\phi)$$

$$f(x) = \frac{1}{\sigma} \int d\phi |\mathcal{M}(\phi)|^2 W(x|\phi)$$

What to do for reducible backgrounds, where M, W uncertain?

Parametrizing the model $\boldsymbol{\alpha} = (\mu, \boldsymbol{\theta})$

Center for Cosmology and Particle Physics

Parameters of interest (*μ*): parameters of the theory that modify the rates and shapes of the distributions, eg.

- the mass of a hypothesized particle
- the "signal strength" μ=0 no signal, μ=1 predicted signal rate

Nuisance parameters (*θ* or α_p): associated to uncertainty in:

- response of the detector (calibration)
  - typically ignored in MEM, need $W(x \mid \phi) \rightarrow W(x \mid \phi, \theta)$
- theoretical uncertainties

Lead to a parametrized model: $\nu \to \nu(\boldsymbol{\alpha}), f(x) \to f(x|\boldsymbol{\alpha})$

$$\mathbf{f}(\mathcal{D}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|\nu(\boldsymbol{\alpha})) \prod_{e=1}^{n} f(x_e|\boldsymbol{\alpha})$$

Simultaneous multi-channel model

Control Regions: Some channels are not populated by signal processes, but are used to constrain the nuisance parameters

Constraint Terms: Often auxiliary measurements for certain nuisance parameters summarized / idealized as

$$f_p(a_p|\alpha_p) \qquad \text{for} \;\; p \in \mathbb{S}$$

Simultaneous Multi-Channel Model: Several disjoint regions of the data are modeled simultaneously. Identification of common parameters across many channels requires coordination between groups such that meaning of the parameters are really the same.

$$\mathbf{f}_{\rm tot}(\mathcal{D}_{\rm sim}, \mathcal{G}|\boldsymbol{\alpha}) = \prod_{c\in\rm channels}\left[\mathrm{Pois}(n_c|\nu_c(\boldsymbol{\alpha}))\prod_{e=1}^{n_c} f_c(x_{ce}|\boldsymbol{\alpha})\right]\cdot\prod_{p\in\mathbb{S}} f_p(a_p|\alpha_p)$$

where

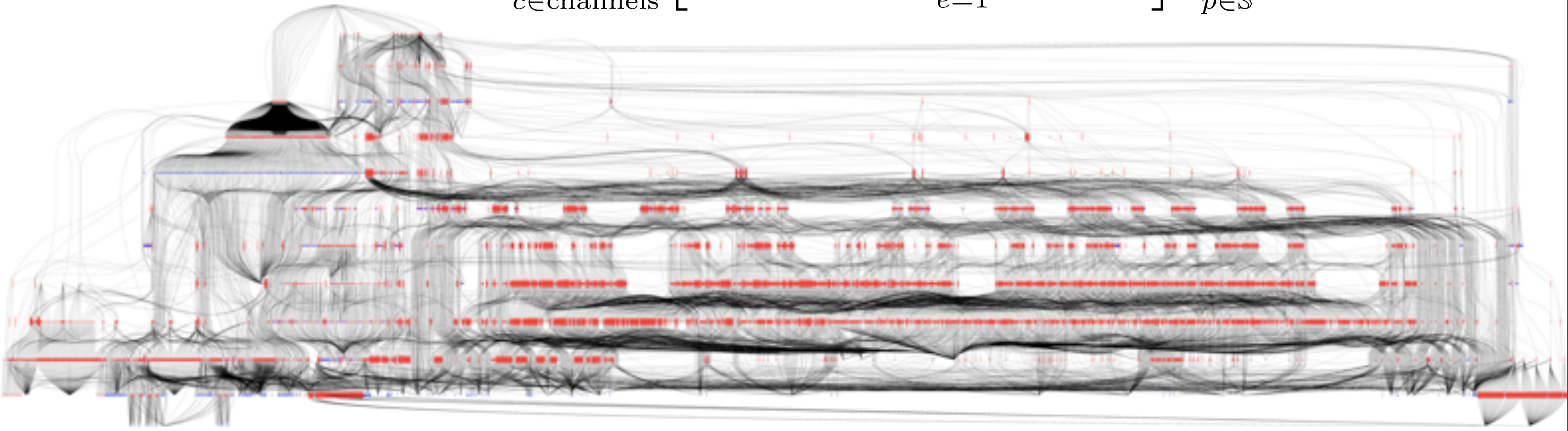
$$\mathcal{D}_{\rm sim} = \{\mathcal{D}_1, \ldots, \mathcal{D}_{c_{\max}}\}, \quad \mathcal{G} = \{a_p\} \;\; \text{for} \;\; p \in \mathbb{S}$$

Visualizing the combined model

RooFit / RooStats: is the modeling language (C++) which provides technologies for collaborative modeling

- provides technology to publish likelihood functions digitally
- and more, it's the full model so we can also generate pseudo-data

$$\mathbf{f}_{\mathrm{tot}}(\mathcal{D}_{\mathrm{sim}}, \mathcal{G}|\boldsymbol{\alpha}) = \prod_{c \in \mathrm{channels}} \left[\mathrm{Pois}(n_c|\nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce}|\boldsymbol{\alpha}) \right] \cdot \prod_{p \in \mathbb{S}} f_p(a_p|\alpha_p)$$



To incorporate MEM approaches directly into common statistical machinery (used for Higgs, SUSY) need interface to RooFit/RooStats

- specifically, need a class that inherits from RooAbsPdf

Some confusion

Matrix Element Method

The Matrix Element Method consist in minimizing a likelihood.

The likelihood for N events is defined as $L(\alpha) = e^{-N\int \bar{P}(x,\alpha)dx} \prod_{i=1}^{N} \bar{P}(x_i;\alpha)$

The best estimate of the parameter α is obtained through a maximisation of the likelihood. It is common practice to minimize $-ln(L(\alpha))$ with respect to α,
$-ln(L) = -\sum_{i=1}^{N} ln(\bar{P}(x_i;\alpha)) + N\int \bar{P}(x,\alpha)dx$

In general, the probability that an event is accepted depends on the characteristics of the measured event, and not on the process that produced it. The measured probability density $\bar{P}(x,\alpha)$ can be related to the produced probability density $P(x,\alpha)$:

I don't actually understand this... $\int P(x,\alpha)\, dx=1$ *if P* is a pdf, and Poisson is not e^{-N}, it is
$$\mathrm{Pois}(n|\nu) = \frac{\nu^n e^{-\nu}}{n!}$$

I would write:
$$\mathbf{f}(\mathcal{D}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|\nu(\boldsymbol{\alpha})) \prod_{e=1}^{n} f(x_e|\boldsymbol{\alpha})$$

$$-\ln L(\alpha) = \underbrace{\nu(\alpha) - n\ln\nu(\alpha)}_{\text{extended term}} - \sum_{e=1}^{n} \ln f(x_e) + \underbrace{\ln n!}_{\text{constant}}$$

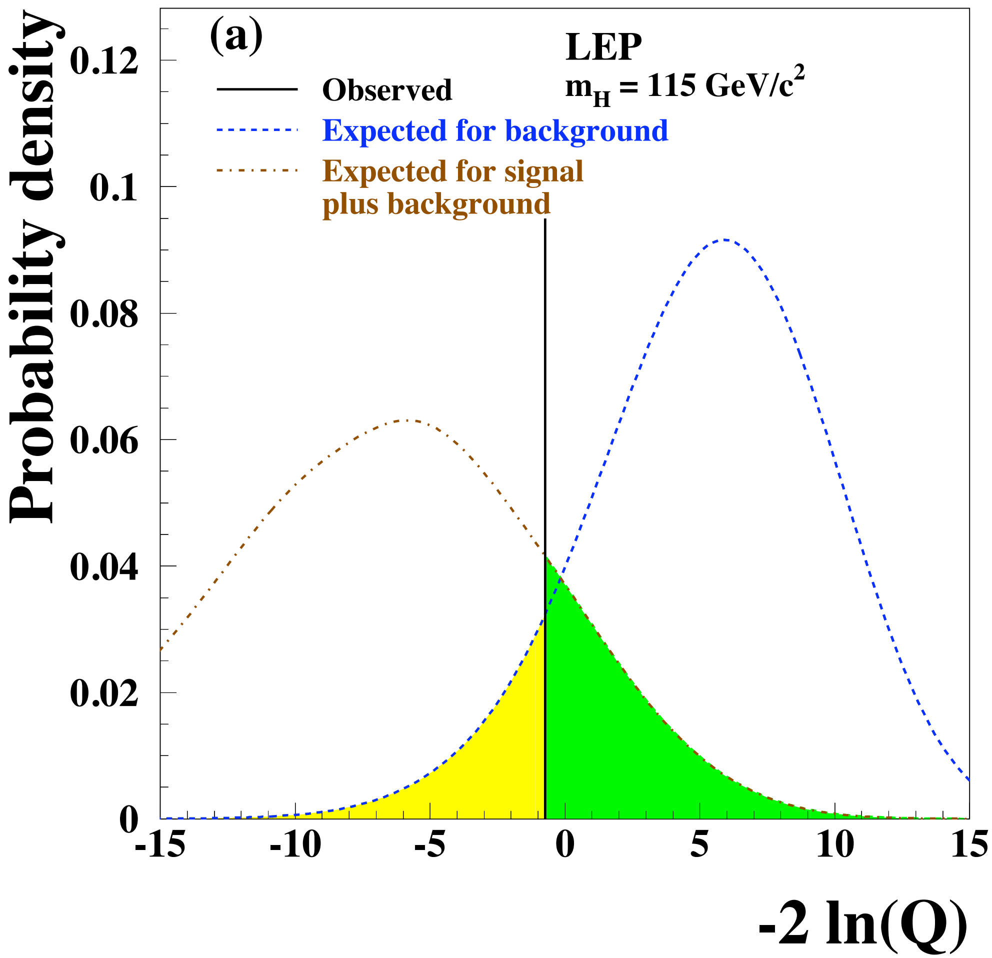
The simple hypothesis test case

Center for Cosmology and Particle Physics

Special case of the general probability model (no nuisance parameters)

$$Q = \frac{L(x|H_1)}{L(x|H_0)} = \frac{\prod_i^{N_{chan}} Pois(n_i|s_i+b_i) \prod_j^{n_i} \frac{s_i f_s(x_{ij}) + b_i f_b(x_{ij})}{s_i+b_i}}{\prod_i^{N_{chan}} Pois(n_i|b_i) \prod_j^{n_i} f_b(x_{ij})}$$

$$q = \ln Q = -s_{tot} + \sum_i^{N_{chan}} \sum_j^{n_i} \ln\left(1 + \frac{s_i f_s(x_{ij})}{b_i f_b(x_{ij})}\right)$$



Instead of simply counting events, the optimal test statistic is equivalent to adding events **weighted** by

ln(1 + signal/background)

The test statistic is a map q:data $\to \mathbb{R}$

By repeating the experiment many times, you obtain a distribution for q

The simple hypothesis test case

There is a clever trick for bootstrapping from distribution of q for a single event to the distribution for an experiment with N events

$$Q = \frac{L(x|H_1)}{L(x|H_0)} = \frac{\prod_i^{N_{chan}} Pois(n_i|s_i+b_i) \prod_j^{n_i} \frac{s_i f_s(x_{ij}) + b_i f_b(x_{ij})}{s_i+b_i}}{\prod_i^{N_{chan}} Pois(n_i|b_i) \prod_j^{n_i} f_b(x_{ij})}$$

$$q = \ln Q = -s_{tot} + \sum_i^{N_{chan}} \sum_j^{n_i} \ln\left(1 + \frac{s_i f_s(x_{ij})}{b_i f_b(x_{ij})}\right)$$

For N events, use Fourier transform to perform N convolutions

$$\rho_{N,i}(q) = \underbrace{\rho_{N,i}(q) \oplus \cdots \oplus \rho_{N,i}(q)}_{N \text{ times}} = \mathcal{F}^{-1}\left\{\left[\mathcal{F}\left(\rho_{1,i}\right)\right]^N\right\}$$

To include Poisson fluctuations on N for a given luminosity, one can exponentiate

$$\rho_i(q) = \sum_{N=0}^{\infty} P(N; L\sigma_i) \cdot \rho_{N,i}(q) = \mathcal{F}^{-1}\left\{e^{L\sigma_i\left[\mathcal{F}(\rho_{1,i}(q))-1\right]}\right\}$$

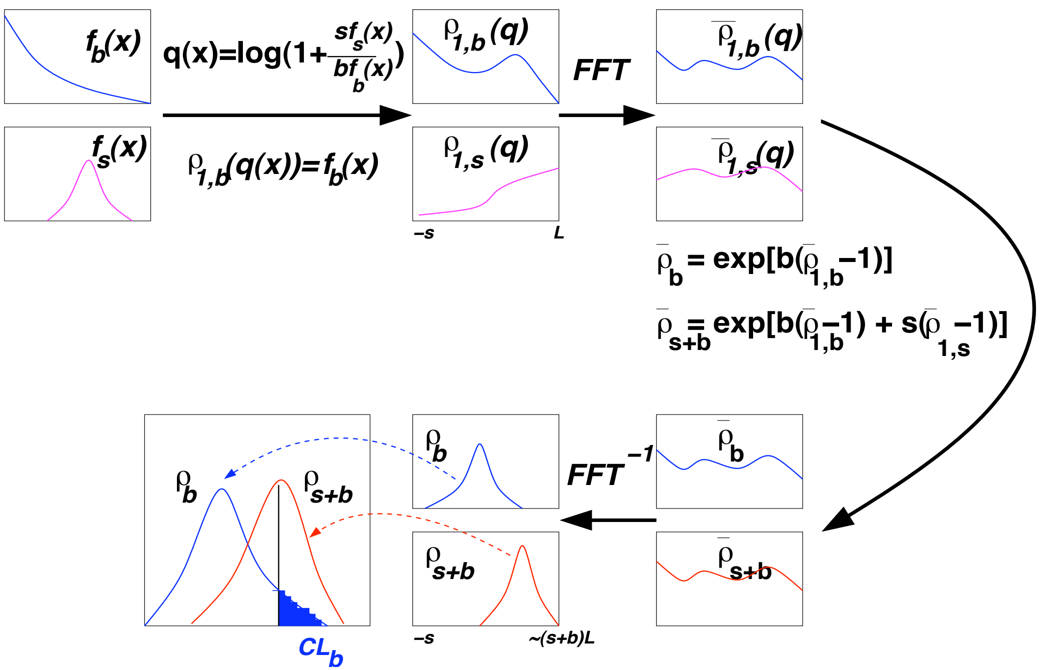
K.C., T. Plehn, hep-ph/0605268

The simple hypothesis test case

There is a clever trick for bootstrapping from distribution of q for a single event to the distribution for an experiment with N events

$$Q = \frac{L(x|H_1)}{L(x|H_0)} = \frac{\prod_i^{N_{chan}} Pois(n_i|s_i+b_i) \prod_j^{n_i} \frac{s_i f_s(x_{ij}) + b_i f_b(x_{ij})}{s_i+b_i}}{\prod_i^{N_{chan}} Pois(n_i|b_i) \prod_j^{n_i} f_b(x_{ij})}$$

$$q = \ln Q = -s_{tot} + \sum_i^{N_{chan}} \sum_j^{n_i} \ln\left(1 + \frac{s_i f_s(x_{ij})}{b_i f_b(x_{ij})}\right)$$



Hu and Nielsen's CLFFT used Fourier Transform and exponentiation trick to transform the log-likelihood ratio distribution for one event to the distribution for an experiment

K.C., T. Plehn, hep-ph/0605268

Profile likelihood

When we go beyond simple hypothesis tests to **parametrized families** of distributions, there is no **uniformly most powerful** test in general

- The most common generalization of the likelihood ratio test statistic is to keep null in numerator and best fit in denominator [Feldman-Cousins]
- In the presence of nuisance parameters, it is the profile likelihood ratio

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})} = \frac{f(\mathcal{D}|\mu, \hat{\hat{\theta}}(\mu;\mathcal{D})\,)}{f(\mathcal{D}|\hat{\mu}, \hat{\theta})}$$

- The Fourier exponentiation trick doesn't work anymore, but the asymptotically the distributions are known

G. Cowan, K. C., E. Gross, O. Vitells. Eur. Phys. J., C71 2011. arXiv:1007.1727

Specifically, I'd like to incorporate experimental uncertainty into the transfer functions: $W(x\,|\,\phi) \rightarrow W(x\,|\,\phi,\theta)$

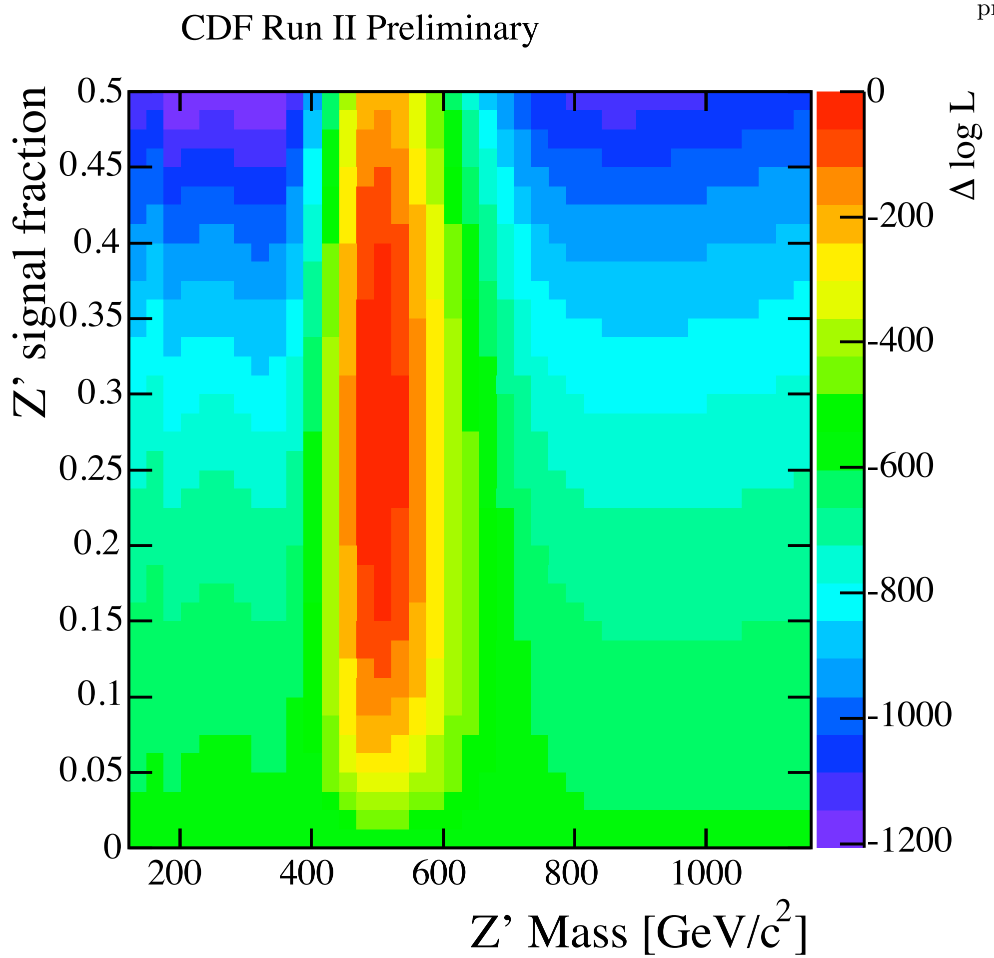
CDF Z' MEM Analysis

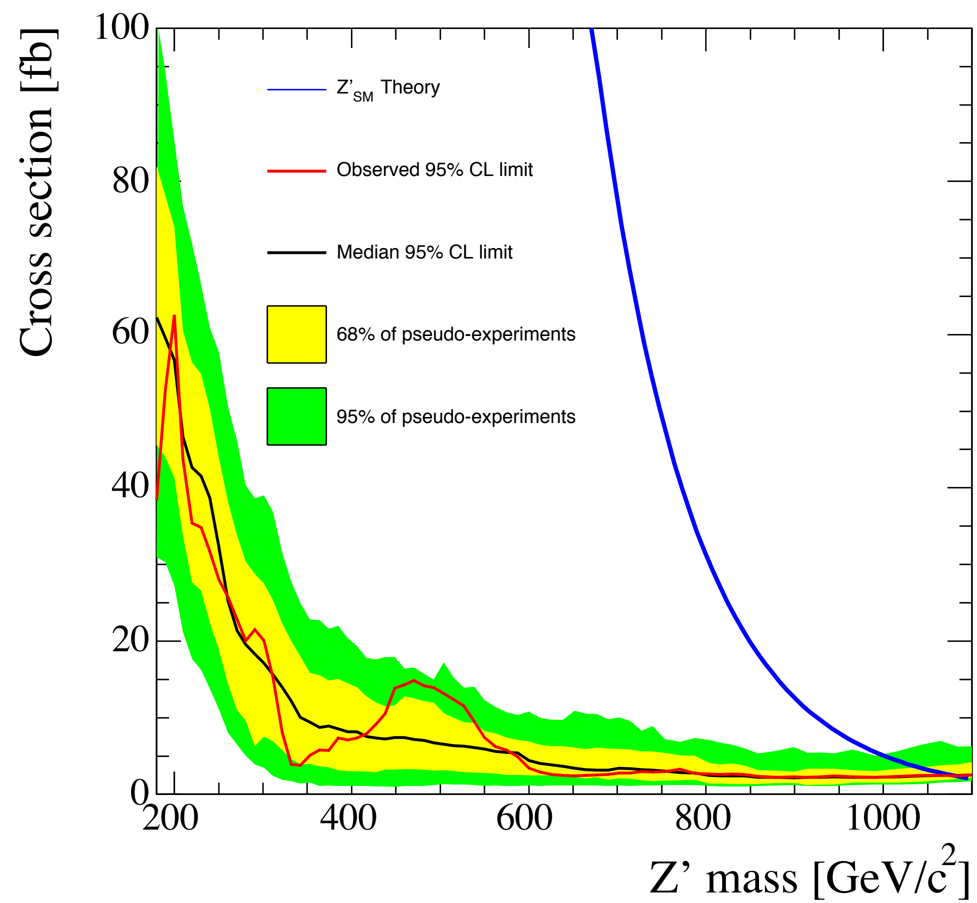
We directly integrated MEM Likelihood into limit-setting procedure

- Included interference of Z' and Z/γ
- 2-d Feldman-Cousins instead of "raster scan"

CDF Collaboration Z'→ μμ
Phys.Rev.Lett. 106 (2011)
arXiv:1101.4578

We present a search for a new narrow, spin-1, high mass resonance decaying to $\mu^+\mu^- + X$, using a matrix element based likelihood and a simultaneous measurement of the resonance mass and production rate. In data with 4.6 fb^{-1} of integrated luminosity collected by the CDF detector in





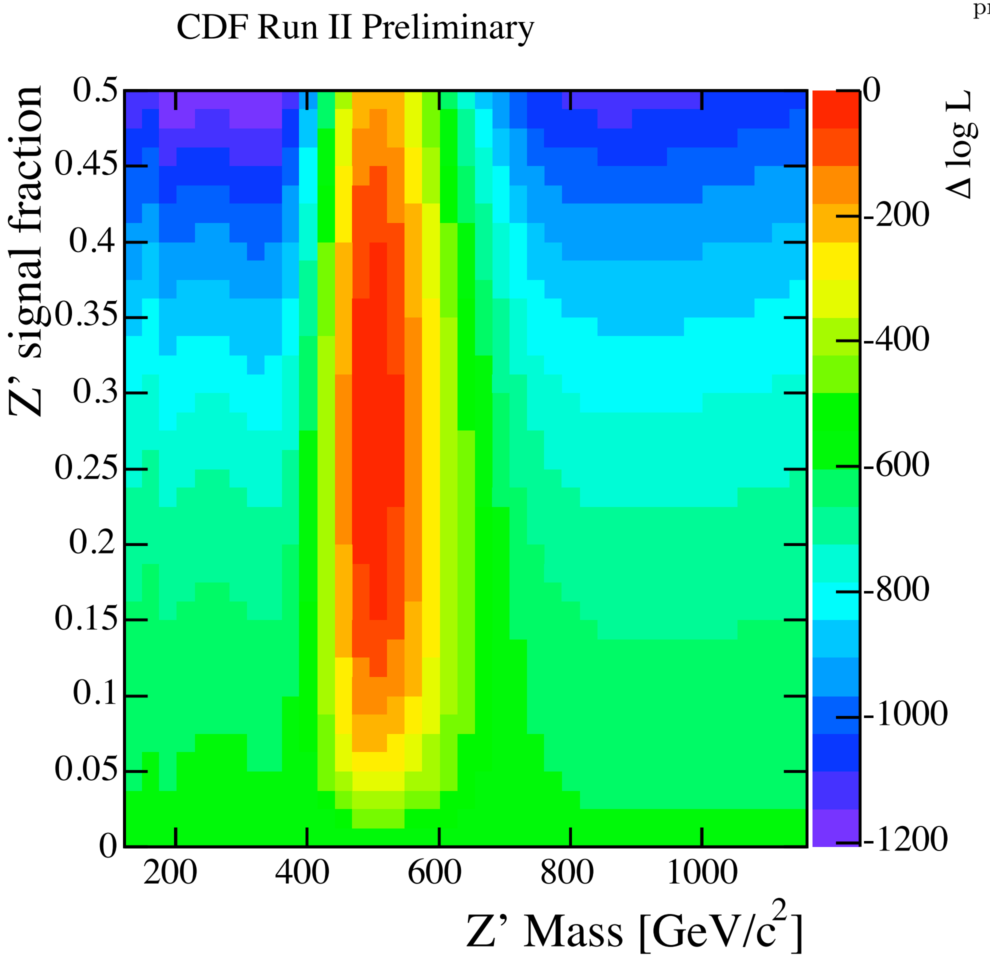
CDF Z’ MEM Analysis

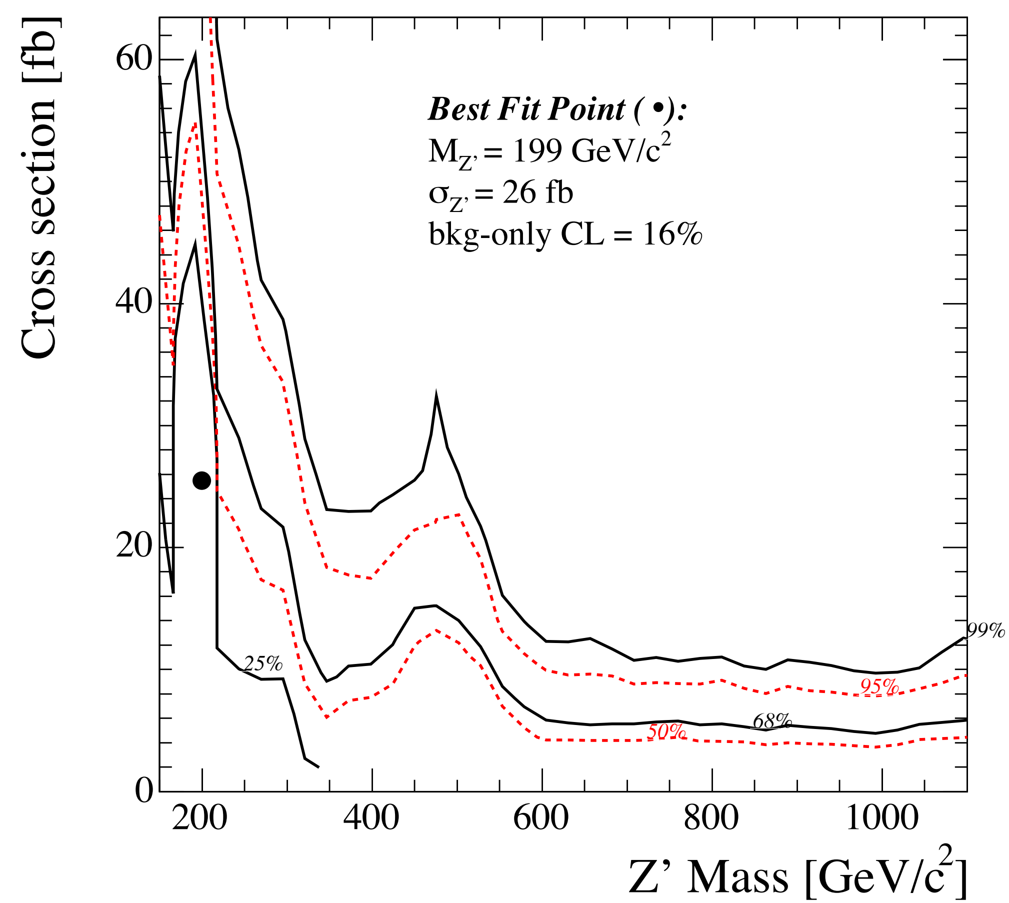
We directly integrated MEM Likelihood into limit-setting procedure

- Included interference of Z' and Z/γ
- 2-d Feldman-Cousins instead of “raster scan”

CDF Collaboration $Z' \rightarrow \mu\mu$
Phys.Rev.Lett. 106 (2011)
arXiv:1101.4578

We present a search for a new narrow, spin-1, high mass resonance decaying to $\mu^+\mu^- + X$, using a matrix element based likelihood and a simultaneous measurement of the resonance mass and production rate. In data with 4.6 fb^{-1} of integrated luminosity collected by the CDF detector in





Cramér-Rao & Fisher Information

Similar to the Neyman-Pearson lemma for simple hypothesis tests is the Cramér-Rao bound for the covariance of an (unbiased) estimator

$$\mathrm{cov}[\hat{\boldsymbol{\alpha}}|\boldsymbol{\alpha}] \geq I^{-1}_{\mu\nu}(\boldsymbol{\alpha})$$

where $I_{\mu\nu}$ is the Fisher Information matrix

$$I_{\mu\nu}(\boldsymbol{\alpha}) = \int p(\mathbf{x}|\boldsymbol{\alpha}) \frac{\partial \ln p(\mathbf{x}|\boldsymbol{\alpha})}{\partial \alpha_\mu} \frac{\partial \ln p(\mathbf{x}|\boldsymbol{\alpha})}{\partial \alpha_\nu} d\mathbf{x} = E\left[\partial_\mu \ln L(\boldsymbol{\alpha}) \partial_\nu \ln L(\boldsymbol{\alpha}) | \boldsymbol{\alpha}\right]$$

In the case of our Marked Poisson model, this is given by

$$I_{\mu\nu}(\boldsymbol{\alpha}) \to \int dx \frac{\partial\ \nu(\boldsymbol{\alpha}) f(x|\boldsymbol{\alpha})}{\partial \alpha_\mu} \frac{\partial\ \nu(\boldsymbol{\alpha}) f(x|\boldsymbol{\alpha})}{\partial \alpha_\nu} \frac{1}{\nu(\boldsymbol{\alpha}) f(x|\boldsymbol{\alpha})}$$

B. Allanach, K.C. [in prep.]

The integral through the transfer function is easy in the "forward" direction

- Evaluating derivative would be aided by importance sampling

Integrate or maximize?

Bayesian / Frequentist often comes down to integrate vs. maximize

- true momenta ϕ plays role of "nuisance parameters"
- Lorentz-invariant phase space $d\phi$ plays role of prior [w/ frequency interpretation]

Perhaps the "Profiled" MEM is even more powerful?

- note, similarity to constrained fit, but also use $|M(\phi)|^2$

	Likelihood	
	$\lvert M(\phi)\rvert^2 W(x\mid\phi)$	$W(x\mid\phi)$
$\int d\phi$	Typical Matrix Element Method	N/A
$\sup_\phi$	"Profiled" MEM	Constrained fit (two-stage: $x \rightarrow \varphi \rightarrow \alpha$)

Warm-up for "Profiled" MEM

Consider a simple case where some interaction characterized by *M* produces particles of energy e_i

- the matrix element is represented by Gaussian: $G(e|M,\sigma_m)$
- the transfer function is a simple Gaussian: $G(x|e,\sigma_e)$

$$P(\{x_i\}|M,\{e_i\}) = \prod_i G(e_i|M,\sigma_m)\, G(x_i|e_i,\sigma_e)$$

One can find the maximum likelihood estimators

$$\hat{e}_i = x_i \qquad\qquad \hat{M} = \frac{1}{n}\sum_i \hat{e}_i = \bar{x}$$

and the estimators are consistent [as n→∞, expectation = true value]

$$E[\hat{M}] = M$$

... so far so good.

Neyman-Scott Phenomena

Center for Cosmology and Particle Physics

Consider a simple case where some interaction characterized by *M* produces particles of energy e_i

- the matrix element is represented by **falling exponential**
- the transfer function is a simple Gaussian: $G(x|e,\sigma_e)$

$$P(\{x_i\}|M,\{e_i\}) = \prod_i \frac{1}{M} e^{-e_i/M}\, G(x_i|e_i,\sigma_e)$$

One can find the maximum likelihood estimators

$$\hat{M} = \frac{\bar{x} + \sqrt{\bar{x}^2 - 4\sigma_e^2}}{2}$$

but the estimator is ***inconsistent!***

$$E[\hat{M}] \neq M$$

This is a general problem if you add more parameters as you add more data, the estimator can be biased even in limit of infinite data!

MEM → MEPSM

Jet-levels: Parton → Hadron → Reconstructed

- it may be benificial to factorize these stages for transfer function
  - W(Reco|Parton) → W(Reco|Hadron) W(Hadron | Parton)

To deal with extra jet radiation, will need to deal with ME-PS matching

- "Poor-man's MEM":
  - store large sample at hadron level, only apply W(reco|hadron)
    - implementation is trivial, but phase space integration is inefficient
- MLM Matching
  - basically requires N_{jet} @ hadron-level = N_{parton} defined at some scale
  - alignment of reco jet algorithm with matching procedure would mean Njet=Nparton
  - if W(reco|parton) encodes jet reconstruction inefficiencies, then $\sum d\phi\ |M_n|^2$ for $n \geq n_{jet}$

A phase space integration idea

The biggest practical issue with the matrix element method is that it is very computationally intensive.

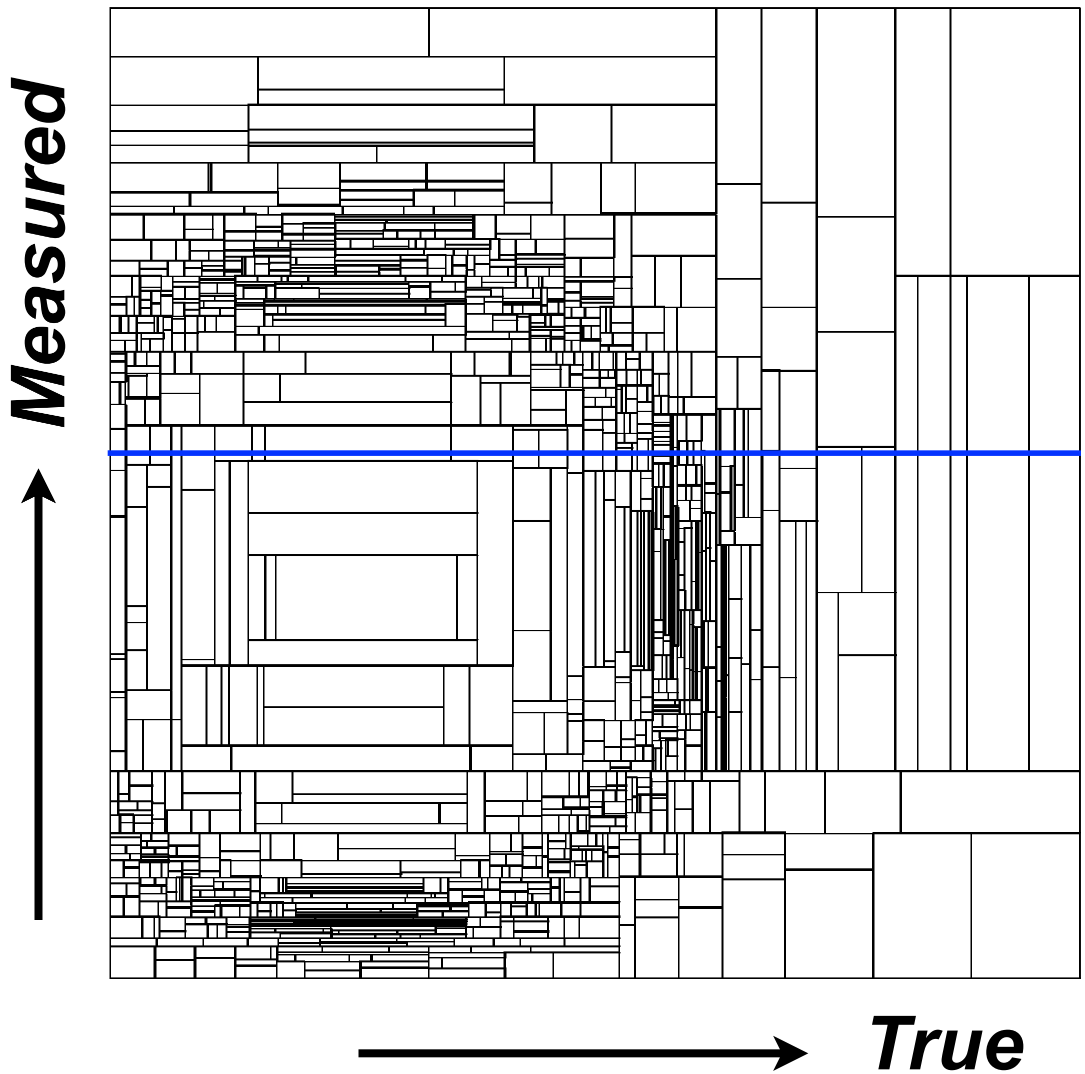
Normally, integration over degrees of freedom in matrix element requires a new Vegas grid for each measurement!

Instead, integrate the joint distribution

- save joint grid

Then for each measurement

- take a slice through the grid
- induced importance sampling



Nested Sampling

In Bayes's theorem, $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, the normalization $P(B)$ often called "evidence". Similar to MEM with $B \rightarrow x, A \rightarrow \phi$

$$P(x) = \frac{1}{\sigma}\int d\phi |\mathcal{M}(\phi)|^2 W(x|\phi)$$

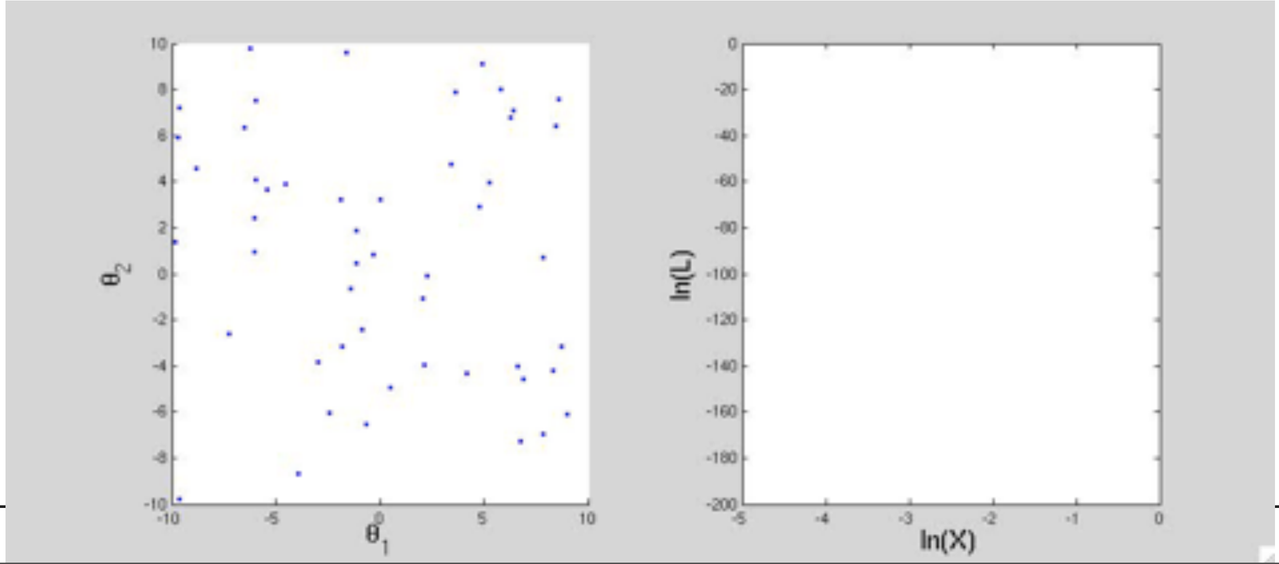
Nested sampling:

An algorithm originally aimed primarily at the Bayesian evidence computation (Skilling, 2006):

$$X(\lambda) = \int_{\mathcal{L}(\theta)>\lambda} P(\theta) d\theta$$

$$P(d) = \int d\theta \mathcal{L}(\theta) P(\theta) = \int_0^1 X(\lambda) d\lambda$$

Feroz et al (2008), *arxiv: 0807.4512*, Trotta et al (2008), *arxiv: 0809.3792*



(animation courtesy of David Parkinson)

Conclusions

MEM natural procedure that provides most powerful test in case of simple hypothesis tests

- In that case, what we want to integrate is a ratio
- noted difficulty when there are irreducible backgrounds

For parametrized model, Cramér-Rao bound is similar to Neyman-Pearson

- Showed explicit form of what we need to calculate in that case
- Showed CDF Z' example for MEM embedded in Feldman-Cousins including interference effects

To include experimental uncertainties, parametrize transfer functions!

- MEM codes should provide interfaces to RooFit/RooStats

Considered "Profiled" MEM as alternative to traditional MEM

- leads to inconsistent estimators and Neyman-Scott phenomena

Some thoughts on "MEPSM" for matching partons

Two thoughts on PS integration: "induced grid" & nested sampling

Neyman-Scott backup

The Gaussian case:

```
f2[x1_, x2_, M_, e1_, e2_] :=
 Exp[-(e1 - M)^2 / (2 sm^2)] / (Sqrt[2 Pi] sm) * Exp[-(e2 - M)^2 / (2 sm^2)] / (Sqrt[2 Pi] sm) *
  Exp[-(x1 - e1)^2 / (2 se^2)] / (Sqrt[2 Pi] se) * Exp[-(x2 - e2)^2 / (2 se^2)] / (Sqrt[2 Pi] se)

Solve[D[Log[f2[x, y, M, e1, e2]], e1] == 0 && D[Log[f2[x, y, M, e1, e2]], e2] == 0 &&
  D[Log[f2[x, y, M, e1, e2]], M] == 0, {M, e1, e2}]
```

$$\left\{\left\{M \to \frac{x+y}{2},\ e1 \to -\frac{-se^2\, x - 2\, sm^2\, x - se^2\, y}{2\left(se^2 + sm^2\right)},\ e2 \to -\frac{-se^2\, x - se^2\, y - 2\, sm^2\, y}{2\left(se^2 + sm^2\right)}\right\}\right\}$$

The exponential case:

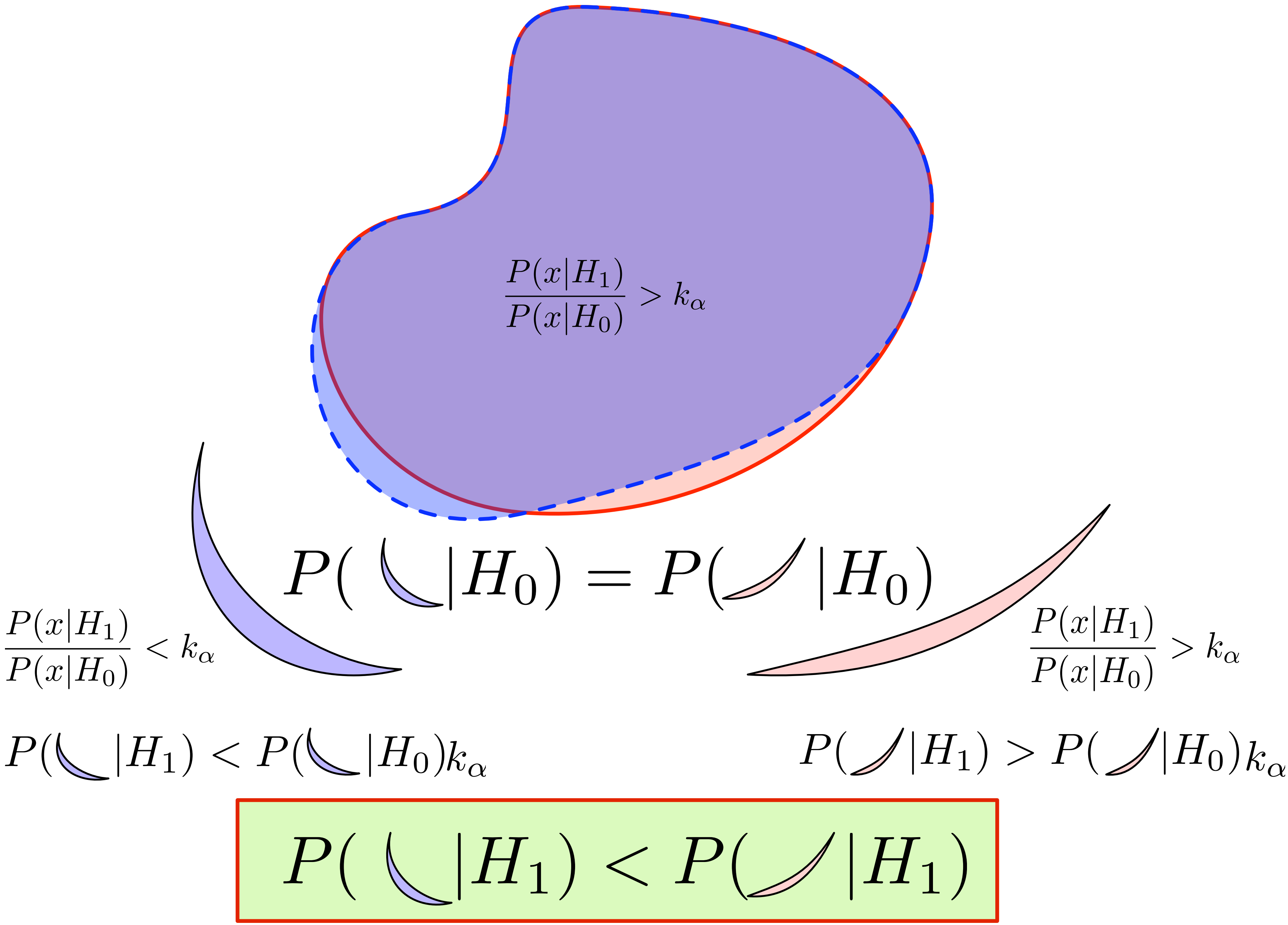
```
In[23]:= g[x1_, x2_, M_, e1_, e2_] :=
  Exp[-e1 / M] / M  * Exp[-e2 / M] / M  * Exp[-(x1 - e1)^2 / (2 se^2)] / (Sqrt[2 Pi] se) *
   Exp[-(x2 - e2)^2 / (2 se^2)] / (Sqrt[2 Pi] se)

In[34]:= Solve[D[Log[g[x, y, M, e1, e2]], e1] == 0 && D[Log[g[x, y, M, e1, e2]], e2] == 0 &&
  D[Log[g[x, y, M, e1, e2]], M] == 0, {M, e1, e2}]
```

Out[34]=

$$\left\{\left\{M \to \frac{1}{4}\left(x + y - \sqrt{-16\, se^2 + x^2 + 2\, x\, y + y^2}\right),\ e1 \to \frac{1}{2}\left(\frac{3\, x}{2} - \frac{y}{2} - \frac{1}{2}\sqrt{-16\, se^2 + x^2 + 2\, x\, y + y^2}\right),\right.\right.$$
$$\left. e2 \to \frac{1}{2}\left(-\frac{x}{2} + \frac{3\, y}{2} - \frac{1}{2}\sqrt{-16\, se^2 + x^2 + 2\, x\, y + y^2}\right)\right\},\ \left\{M \to \frac{1}{4}\left(x + y + \sqrt{-16\, se^2 + x^2 + 2\, x\, y + y^2}\right),\right.$$
$$\left.\left. e1 \to \frac{1}{2}\left(\frac{3\, x}{2} - \frac{y}{2} + \frac{1}{2}\sqrt{-16\, se^2 + x^2 + 2\, x\, y + y^2}\right),\ e2 \to \frac{1}{2}\left(-\frac{x}{2} + \frac{3\, y}{2} + \frac{1}{2}\sqrt{-16\, se^2 + x^2 + 2\, x\, y + y^2}\right)\right\}\right\}$$

A short proof of Neyman-Pearson



The new region region has less power.