Efficient Neutrino Oscillation Parameter Inference with Gaussian Process

Lingge Li, Nitish Nayak, Jianming Bian, Pierre Baldi

UC-Irvine

PhystatNu - 2019

Neutrino Oscillations

- Neutrinos : 2 kinds of states, each of which come in 3 types
  - Interacting, i.e what we observe $\rightarrow$ flavor states (ν_e, ν_μ, ν_τ)
  - Propagating, i.e in between observations $\rightarrow$ mass eigenstates (ν_1, ν_2, ν_3)
- Principle of superposition connects them via 3×3 unitary matrix (U_{PMNS}), i.e.

$$\begin{bmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{bmatrix} = U_{PMNS} \begin{bmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{bmatrix}$$

- Via QM, neutrinos starting out as one flavor can be observed as another ("Oscillations").
- Well defined probability which depends on :
  - Energy of neutrino, E_ν and length of propagation, L
  - mass-squared splittings, Δm^2_{32}, Δm^2_{21}, i.e $\Delta m^2_{ij} = m^2_i - m^2_j$
  - U_{PMNS}

- For neutrino propagation in vacuum, the oscillation probability in all its glory:

$$P(\nu_\alpha \rightarrow \nu_\beta) = \delta_{\alpha\beta} - 4 \sum_{i>j}^{3} \Re(U^*_{\alpha i} U_{\beta i} U_{\alpha j} U^*_{\beta i}) \sin^2(\frac{\Delta m^2_{ij} L}{4E_\nu}) + 2 \sum_{i>j}^{3} \Im(U^*_{\alpha i} U_{\beta i} U_{\alpha j} U^*_{\beta i}) \sin(\frac{\Delta m^2_{ij} L}{4E_\nu})$$

Neutrino Oscillations Contd..

- U_{PMNS} commonly parameterized as

$$U_{PMNS} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos\theta_{23} & sin\theta_{23} \\ 0 & -sin\theta_{23} & cos\theta_{23} \end{bmatrix} \begin{bmatrix} cos\theta_{13} & 0 & sin\theta_{13}e^{-i\delta_{CP}} \\ 0 & 1 & 0 \\ -sin\theta_{13}e^{i\delta_{CP}} & 0 & cos\theta_{13} \end{bmatrix} \begin{bmatrix} cos\theta_{12} & sin\theta_{12} & 0 \\ -sin\theta_{12} & cos\theta_{12} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

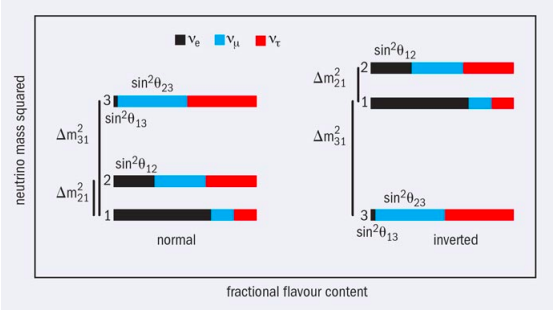
- Physics program entails measuring $P(\nu_\alpha \rightarrow \nu_\beta)$ to infer U_{PMNS} and Δm^2_{ij} parameters
- Broadly, solar experiments give handle on (21) parameters, reactor experiments for θ_{13}

- Long baseline (LBL) experiments (this talk) gives handle on (32).
  - $P(\nu_\mu \rightarrow \nu_\mu)$ sensitive to $sin^2(2\theta_{23})$ and $|\Delta m^2_{32}|$
  - Non-zero θ_{13} opens up $P(\nu_\mu \rightarrow \nu_e)$ channel, sensitive to δ_{CP}, θ_{23} octant and $sgn(\Delta m^2_{32})$

Physics Implications

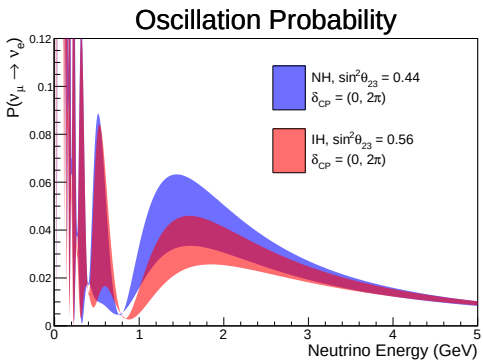
In the LBL context, we want to know if :

- $\Delta m^2_{32} > 0$ or < 0? (Normal or Inverted)
  - Identifying mass hierarchy (NH or IH) has implications for neutrino mass measurements
- Octant of θ_{23} or $\theta_{23} = 45°$?
- $sin\delta_{CP} \neq 0$?
  - Lepton sector CP-violation. Gives us a clue towards explaining matter-antimatter asymmetry



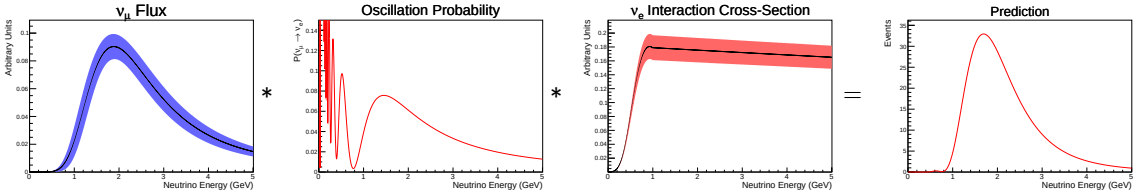
Statistical Issues

- Oscillation Parameters are typically measured via MLE using the underlying PMNS model and comparing it to observation
- However, experiments collect only a handful of statistics. $\mathcal{O}(10-100)$ over years of operation for the $\nu_\mu \to \nu_e$ channel
- Oscillation probabilities have complicated dependence on multiple parameters $\implies$ difficult to delineate
- Confidence Intervals are hard to find as Likelihood ratios don't satisfy asymptotic properties.
- Let's illustrate this with a toy experiment..



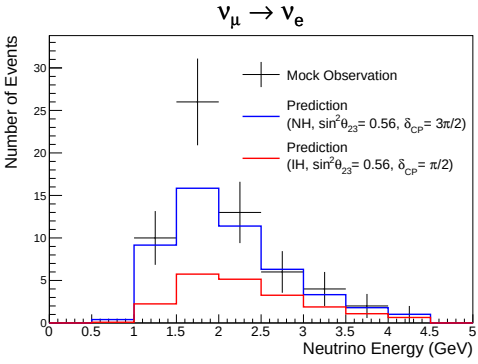
Toy Experiment

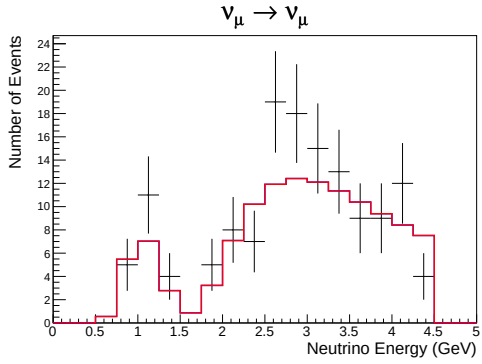
- Modelled on NOvA. Baseline, $L = 810$km with ν_μ flux peaking at 2GeV
- $\nu_\mu \to \nu_e$ by multiplying toy shapes for flux, cross-section and oscillation probability.
- 10% normalisation errors on flux and xsec model



- $P(\nu_\mu \to \nu_e)$ using 3-flavor PMNS with MSW corrections added for matter propagation.
- Similar setup for $\nu_\mu \to \nu_\mu$ to constrain $sin^2(2\theta_{23})$ and $|\Delta m^2_{32}|$ but with 2-flavor approximation
- $P(\nu_\mu \to \nu_\mu) \sim 1 - sin^2(2\theta_{23})sin^2(\Delta m^2_{32} L/4E)$

Toy Experiment





- Toy data ($\vec{x}$) from Poisson variations at some chosen oscillation parameters.
- With (θ, δ) denoting list of oscillation and nuisance (flux and xsec errors) parameters,
- Best-fit $(\hat{\theta}, \hat{\delta})$ found by minimizing negative log-likelihood over energy bins, i

$$-2\log L(\theta,\delta) = -2\sum_{i\in I}\log Pois(x_i; v(\theta,\delta)_i) - \sum_{i\in I} x_i + \sum_{i\in I} v(\theta,\delta)_i + \delta^2$$

Confidence Intervals

- θ_0 included in the $1-\alpha$ confidence contour if we fail to reject the null ($\theta = \theta_0$) at α level
- Use an Inverted Likelihood Ratio Test (LRT)
- Neyman-Pearson Lemma : Likelihood Ratio (LR) is the most powerful test statistic

Table 38.2: Values of $\Delta\chi^2$ or $2\Delta \ln L$ corresponding to a coverage probability $1-\alpha$ in the large data sample limit, for joint estimation of m parameters.

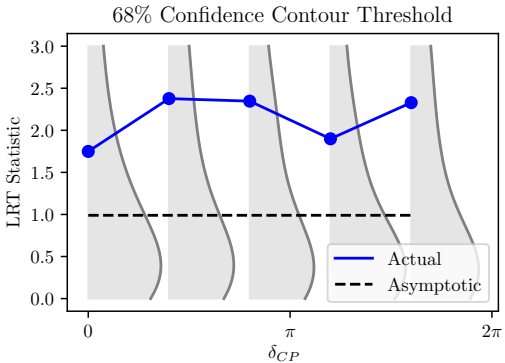
$(1-\alpha)$ (%)	$m=1$	$m=2$	$m=3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

From the PDG Review on Statistics

- Easy to estimate in the asymptotic case as LR is a χ^2 distribution. (Wilks Theorem)
- However, that's not the case here!
- Proceed via Unified approach (Feldman-Cousins, 1998)

Feldman-Cousins

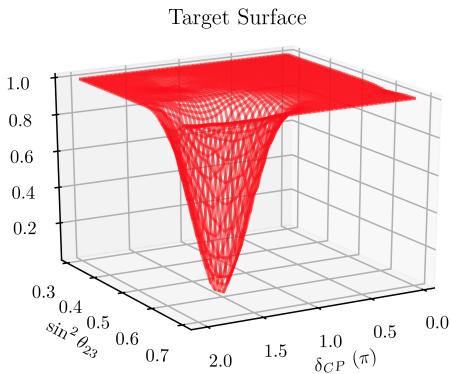
- Seminal result giving an ordering principle for confidence intervals in non-asymptotic cases
- For given θ_0, explicitly simulate distribution of test statistic, LR via Monte-Carlo experiments at θ_0



- 68% confidence interval for δ_{CP} : All δ_{CP} values for which LR for observed data (critical value) lies within threshold
- Confidence of rejecting given $\delta_{CP} = \delta_0$ given by percentile of $crit(\delta_0)$
- Gives us the "correct" confidence interval in the frequentist sense by construction, since its essentially a grid search over the entire parameter space.

A more efficient FC

- Grid search across multi-dimensional parameter space $\Longrightarrow$ extremely intense computational demands
- It'd be nice to be able to come up with a more refined search algorithm.



- We can expect intuitively :
  - Given a point in parameter space that is rejected at high confidence, it is likely that points near it will also be rejected
  - Further, the variation in the LR percentiles ought to be smooth.
- An efficient search would therefore :
  - Learn local features in the LR percentile surface to guide the search
  - Favor simulating the LR test statistic distribution near the edge of the desired confidence contour than further out.

Bayesian Supervised Learning

- Our goal is to approximate the FC percentile surface non parametrically using only a fraction of the grid points.

- Classical supervised learning $\rightarrow$ training data to get best-fit model.
- Predictions for new data are best-guess

- A Bayesian approach can assume a model itself to be a random variable with a certain probability distribution.
- Training data updates your priors about the model distribution
- Predictions for new data is a posterior distribution in model space.
- Quantifies uncertainty in model estimates. Gets smaller with more training data
- Can be pretty non-parametric

Gaussian Process

- Special case of Bayesian Learning. Model distribution is an extension of multivariate gaussians to function space.
- Technically, its a probability measure defined over ∞-dim function space parameterized only by a mean function, $\mu(x)$ and a covariance function (kernel), $k(x, x')$
- We say, $f \sim \mathcal{GP}(\mu, k(\cdot, \cdot))$ if

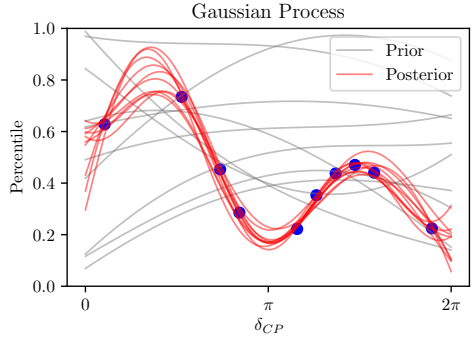
$$\begin{pmatrix} f(x) \\ f(x') \end{pmatrix} \sim \mathcal{N}(\begin{bmatrix} \mu(x) \\ \mu(x') \end{bmatrix}, \begin{bmatrix} k(x,x) & k(x,x') \\ k(x,x') & k(x',x') \end{bmatrix}).$$

- Intuitively, we can picture each draw from a $\mathcal{GP}(\mu, k(\cdot, \cdot))$ giving us a different $f(x)$ with the average result being $\mu(x)$
- The kernel encodes the correlation between nearby points. A commonly used kernel is the radial basis function, $k(x, x') = \exp(-(x - x')^2/l^2)$
- A RBF kernel tells us that $\mathcal{GP}$ results at nearby points are highly influenced by observations at a given point while further out, they aren't.

Why $\mathcal{GP}$s?

- Enormously flexible! Can basically approximate any well behaved function with an appropriate choice of the kernel.
- Predictions at new data points are computationally tractable with basic linear algebra, i.e for $\mathcal{GP}(\mathbf{0}, k(\cdot,\cdot))$:

$$f(x')|f(x) \sim \mathcal{N}(\frac{k(x,x')}{k(x,x)}f(x), k(x',x') - \frac{k(x,x')^2}{k(x,x)})$$



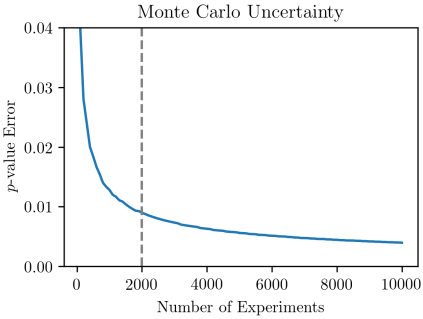
- Kernel hyperparameters can be learned via maximising the likelihood of current set of observations marginalised over the function distribution, f

$\mathcal{GP}$s in Literature

- $\mathcal{GP}$s in HEP : arXiv:1709.05681, M. Frate, K. Cranmer et al. Using $\mathcal{GP}$s to describe background spectra in dijet resonance searches at the LHC non-parametrically.
- Used in Astrophysics for modelling stochasticity of light yields in stars, active galactic nuclei etc
- Many other fields!

$\mathcal{GP}$s for FC

- Fitting a GP to target percentile surface for a given contour. (Stochasticity of the target surface)
- "Observation" at a given point in parameter space, θ means simulating the LRT distribution and finding the percentile of $crit(\theta)$
- Choose a RBF Kernel with an additional term incorporating variance of percentile estimate at θ.



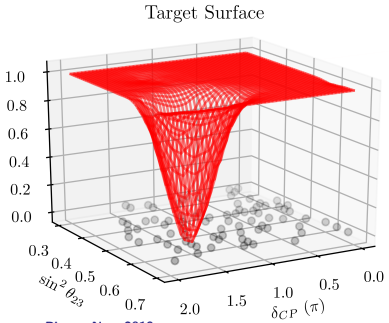
- $k(\cdot,\cdot) = k_{RBF}(\cdot,\cdot) + \sigma_p^2 I$
- The additional variance encodes the binomial error resulting from throwing finite number of experiments to simulate the LRT distribution at θ
- Allows us to incorporate varying number of experiments thrown into the CI search, reducing computational burden further.

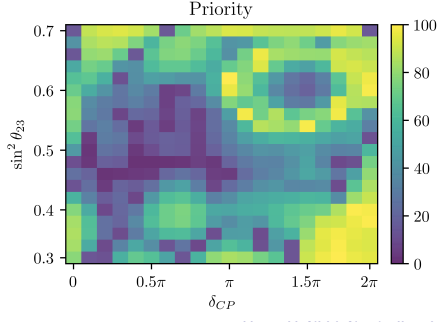
Optimised Confidence Interval Search

- Use an acquisition function that proposes new points in θ-space to explore based on $\mathcal{GP}$ approximated percentile surface.

$$a(\theta) = \sum_{\alpha_i} \left| \frac{\hat{q}(\theta) - \alpha_i}{\sigma_{\hat{q}(\theta)}} \right|^{-1}$$

- Here, $\hat{q}(\theta)$ is $\mathcal{GP}$ mean, $\sigma_{\hat{q}(\theta)}$ is $\mathcal{GP}$ std-dev, α_i is chosen to be $(0.68, 0.90)$
- $a(\theta)$ balances between exploration, i.e MC experiments at new points and exploitation, i.e reducing $\mathcal{GP}$ error



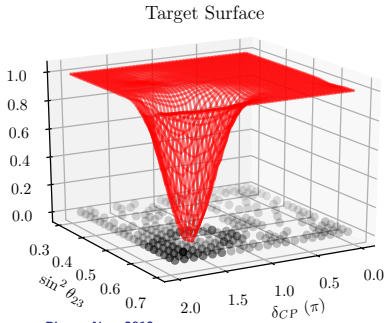


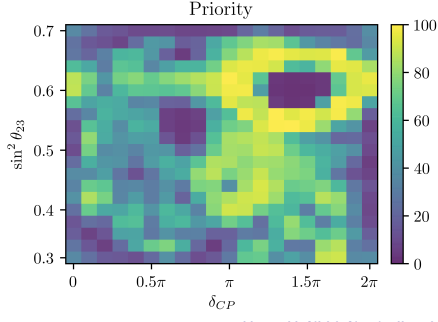
Optimised Confidence Interval Search

- Use an acquisition function that proposes new points in θ-space to explore based on $\mathcal{GP}$ approximated percentile surface.

$$a(\theta) = \sum_{\alpha_i} \left| \frac{\hat{q}(\theta) - \alpha_i}{\sigma_{\hat{q}(\theta)}} \right|^{-1}$$

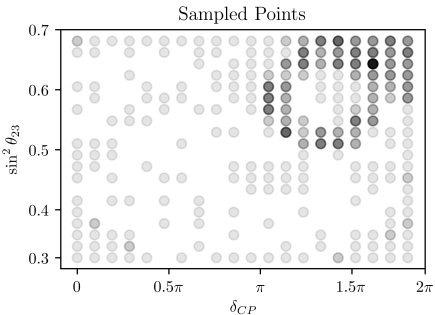
- Here, $\hat{q}(\theta)$ is $\mathcal{GP}$ mean, $\sigma_{\hat{q}(\theta)}$ is $\mathcal{GP}$ std-dev, α_i is chosen to be $(0.68, 0.90)$
- $a(\theta)$ balances between exploration, i.e MC experiments at new points and exploitation, i.e reducing $\mathcal{GP}$ error

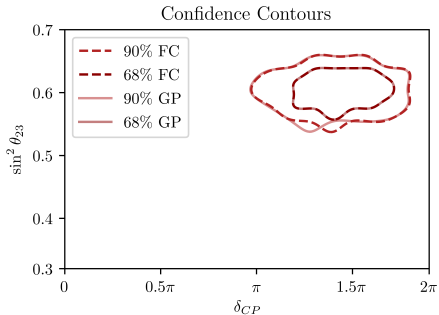




Results

- "Real" data similar to latest best-fit estimate from NOvA. ($sin^2\theta_{23} = 0.56$, $\Delta m^2_{32} = 2.44 \times 10^{-3} eV^2$, $\delta_{CP} = 1.5\pi$)
- $sin^2\theta_{23} - \delta_{CP}$ 68% and 90% CI for IH after 5 iterations

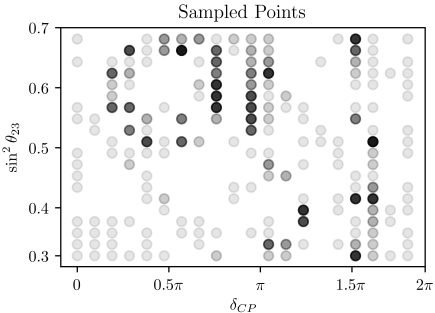


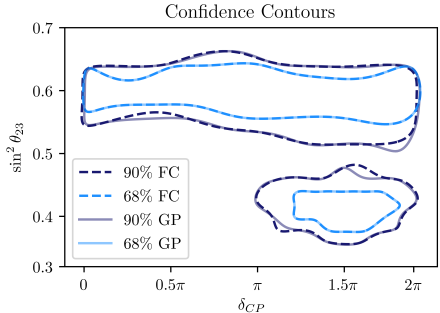


- Grayscale denotes number of experiments thrown in relation to FC (2000)
- Algorithm does a good job of finding the FC contour edge!

Results

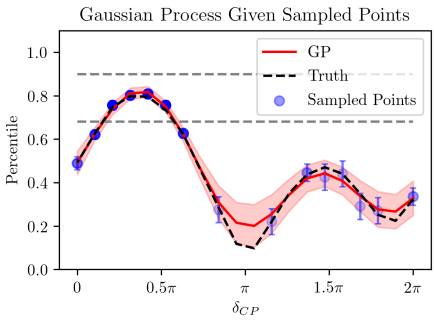
- "Real" data similar to latest best-fit estimate from NOvA. ($sin^2\theta_{23} = 0.56$, $\Delta m^2_{32} = 2.44 \times 10^{-3} eV^2$, $\delta_{CP} = 1.5\pi$)
- $sin^2\theta_{23} - \delta_{CP}$ 68% and 90% CI for NH after 5 iterations

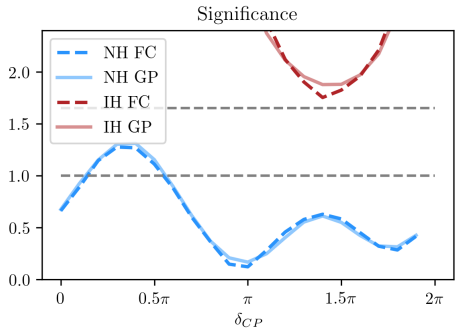




Results

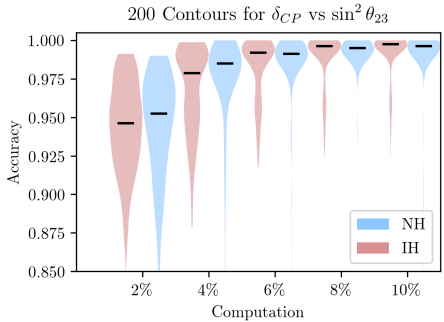
- "Real" data similar to latest best-fit estimate from NOvA. ($sin^2\theta_{23} = 0.56$, $\Delta m^2_{32} = 2.44 \times 10^{-3} eV^2$, $\delta_{CP} = 1.5\pi$)
- Significance of rejecting δ_{CP} only after 5 iterations. (Percentile converted to Z-score significance)

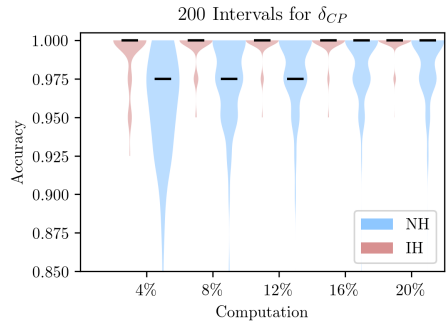




Results

- 200 different runs for "real" data at the same point as before.
- Use classification accuracy of all grid points, taking FC result as truth, to evaluate performance.
- Progress shows the search algorithm converges to the FC value $\sim 10\times$ faster for 2D case and $\sim 5\times$ for 1D case





- Median Accuracies for 1D is 100%, for 2D is $> 99.5\%$ (both NH, IH)
- Mean Accuracies for 1D is 98.5% (99.8%) for NH (IH), for 2D is $> 99\%$ (both NH, IH)

Summary and Conclusions

- Neutrino oscillation experiments provide interesting test case for estimating frequentist confidence intervals
- LBL experiments typically proceed via Feldman-Cousins
- However, simulating LRT distributions across multi-dimensional parameter space requires huge computational resources
- We've studied a Bayesian approach using Gaussian processes on a toy LBL set-up
- Helps us estimate frequentist contour edges to quite a high accuracy without having to sample the entire parameter space!
- Order of magnitude gain in computation!
- All code with illustrative notebooks here : `https://github.com/nitish-nayak/ToyNuOscCI`, maintained by Lingge (linggeli7@gmail.com) and myself (nayakb@uci.edu)

Backup

$\mathcal{GP}$ Technical Details

- Rasmussen and Williams has a good discussion about convergence to true functions in regression settings (typically using squared loss functions) :
  `http://www.gaussianprocess.org/gpml/chapters/RW7.pdf`
- Well behaved $\implies$ expressible as a generalised fourier series of kernel eigenfunctions
- If kernel is non-degenerate, approximation is guaranteed to converge to true function
- If degenerate, convergence towards an L_2 approximation of the true function
- Rates of convergence typically depends on mean and kernel smoothness as well as smoothness of the true function

- Hyperparameters ($\mathbf{w}$) learned via maximising log marginal likelihood :

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \int p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{f}) p(\mathbf{f}|\mathbf{X}, \mathbf{w}) d\mathbf{f}$$

- Clearly,

$$\mathbf{f}|\mathbf{X}, \mathbf{w} \sim \mathcal{N}(\mathbf{0}, K(\mathbf{X}, \mathbf{w}))$$

- Some algebra gives us :

$$-2 \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \mathbf{y}^T K^{-1} \mathbf{y} + \log |K| + n \log 2\pi$$

- Minimising above equation gives us a good choice for $\mathbf{w}$
- $log|K|$ acts as a penalty term for complexity and therefore reduces overfitting to data

$\mathcal{GP}$ for FC

- "Gaussian" not a statement of the underlying distribution of the test statistic, which can still be heavily non-Gaussian
- Rather, "Gaussianity" for a stochastic process generating the test statistic distributions. Stochasticity mostly from finite FC grid resolution or finite number of MC experiments for simulating the test statistic distribution
- Assumption we're making for this stochasticity is that it can be parameterised by a kernel describing the relationship between the distributions at neighbouring points $\implies$ multi-variate gaussian

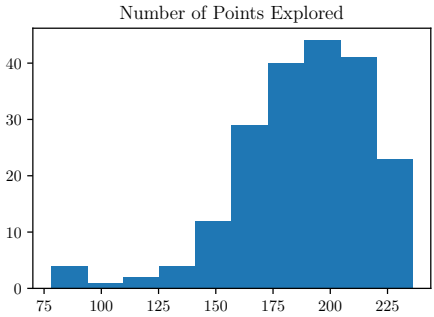
- Also important to note, no real statement about FC coverage or handling of nuisance parameters. Assumes FC gives desired level of coverage
- Confidence Intervals still with frequentist interpretation
- Bayesian interpretation for "classification probability" of points in parameter space for desired confidence regions
- A good summary would be "Accelerating Frequentist CI search by estimating CI edges through Bayesian ML"

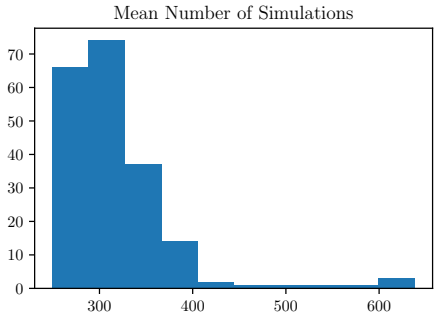
Pseudo-code

Algorithm 1 $\mathcal{GP}$ iterative confidence contour finding

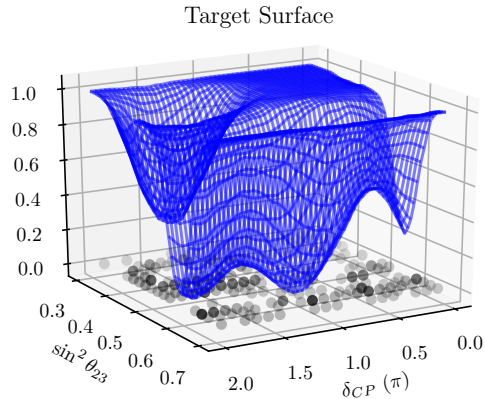
for each iteration $t = 1, 2, ...$ **do**
  Propose new points in parameter space $\arg\max_\theta a(\theta)$
  for each point θ' **do**
    Simulate likelihood ratio distribution
    for $k = 1, 2, ...$ **do**
      Perform a pseudo experiment
      Maximize the likelihood with respect to (θ, δ)
      Maximize the likelihood with constraint $\theta = \theta'$
    end for
    Obtain critical value $c(\theta')$
  end for
  Update $\mathcal{GP}$ approximation $\hat{c}(\theta)$
  Update confidence contours
end for

Results : NH, $sin^2\theta_{23} - \delta_{CP}$

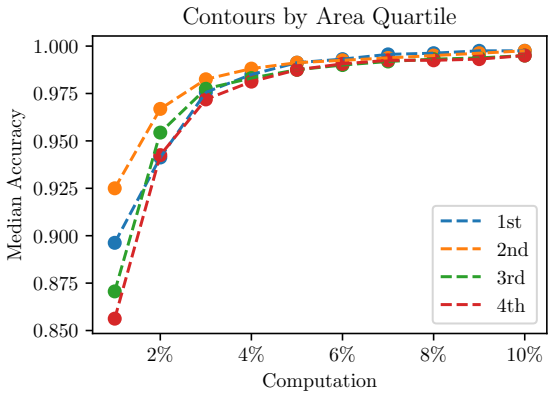




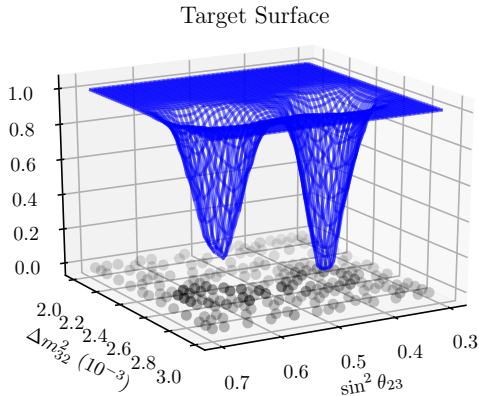
NH, $sin^2\theta_{23} - \delta_{CP}$



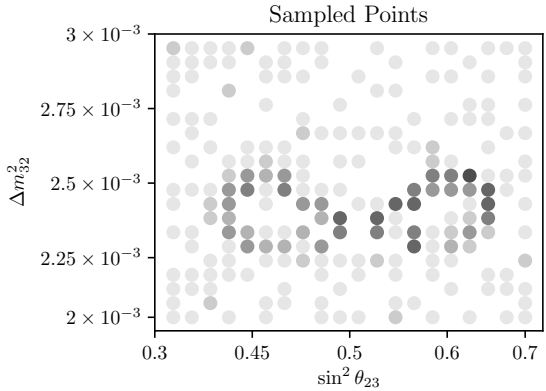
NH, $sin^2\theta_{23} - \delta_{CP}$



NH, $sin^2\theta_{23} - \Delta m^2_{32}$



NH, $sin^2\theta_{23} - \Delta m^2_{32}$



NH, $sin^2\theta_{23} - \Delta m^2_{32}$