Robot Physicist

Team = [Manoj Kumar, Michele Ceru, Phil Yeres]
Advisor = Kyle Cranmer

Theoretical Documentation

- Drafted a theoretical explanation of the model we're developing
- Received feedback from our advisors on how to better frame our analysis in the language of Bayesian inference

1 Information gain calculation:

We indicate with x the data that is the output of the experiment. This output depends on θ (that represent a constant of nature) and Φ (that represent the setting to the experiment). We call $P(x|\theta,\Phi)$ the probability distribution of the data given θ and Φ.
The information gain for this distribution is defined as:

$$EIG(\Phi) = \int P(x|\Phi)\Big[H[P(\theta)] - H[P(\theta|x,\Phi)]\Big]dx \tag{1}$$

where we indicated with H the entropy, that for a discrete distribution is defined as:

$$H[P] = -\sum_{k\geq 1} p_k \log p_k \tag{2}$$

To compute that we need to sample from $P(x|\Phi)$ and know $P(\theta|x,\Phi)$.

- We don't have $P(x|\Phi)$ because the data of the experiment is conditioned on θ as well, consequently we can only sample from the distribution $P(x|\theta,\Phi)$. But we can calculate it using the following:

$$P(x|\Phi) = \int P(x,\theta|\Phi)d\theta = \int P(x|\theta,\Phi)P(\theta|\Phi)d\theta = E[P(x|\theta,\Phi)] \tag{3}$$

Where the last equality is the expected value of $P(x|\theta,\Phi)$ under the distribution $P(\theta|\Phi)$. Discretising the integral:

$$P(x|\Phi) = \sum_{i=1}^{n} P(x|\theta_i,\Phi)P(\theta_i|\Phi) = \frac{1}{n}\sum_{i=1}^{n} P(x|\theta_i,\Phi) \tag{4}$$

Where in the last equality we are assuming to have an uniform distribution $P(\theta_i|\Phi) = 1/n$. We use the black box to generate the distributions $P(x|\theta_i,\Phi)$ for n values of θ:

$$\begin{aligned}
&\theta_1 \rightarrow P(x|\theta_1,\Phi)\\
&\theta_2 \rightarrow P(x|\theta_2,\Phi)\\
&\dots\\
&\theta_n \rightarrow P(x|\theta_n,\Phi)
\end{aligned} \tag{5}$$

each of these distribution is obtained from a histogram of the sampled data. Using these we can compute $P(x|\Phi)$.

Since $P(x|\Phi)$ is now discretized because of the histogram, we sample from a multinomial with the parameters, x_i's given by the bin centers and the probability computed from the normalized histogram.

This is done from lines 92 to 107 in the code.

1

Docker Deployment on AWS

- Received physics experiment simulator from our advisor
- Deployed physics simulation to AWS, and profiled performance

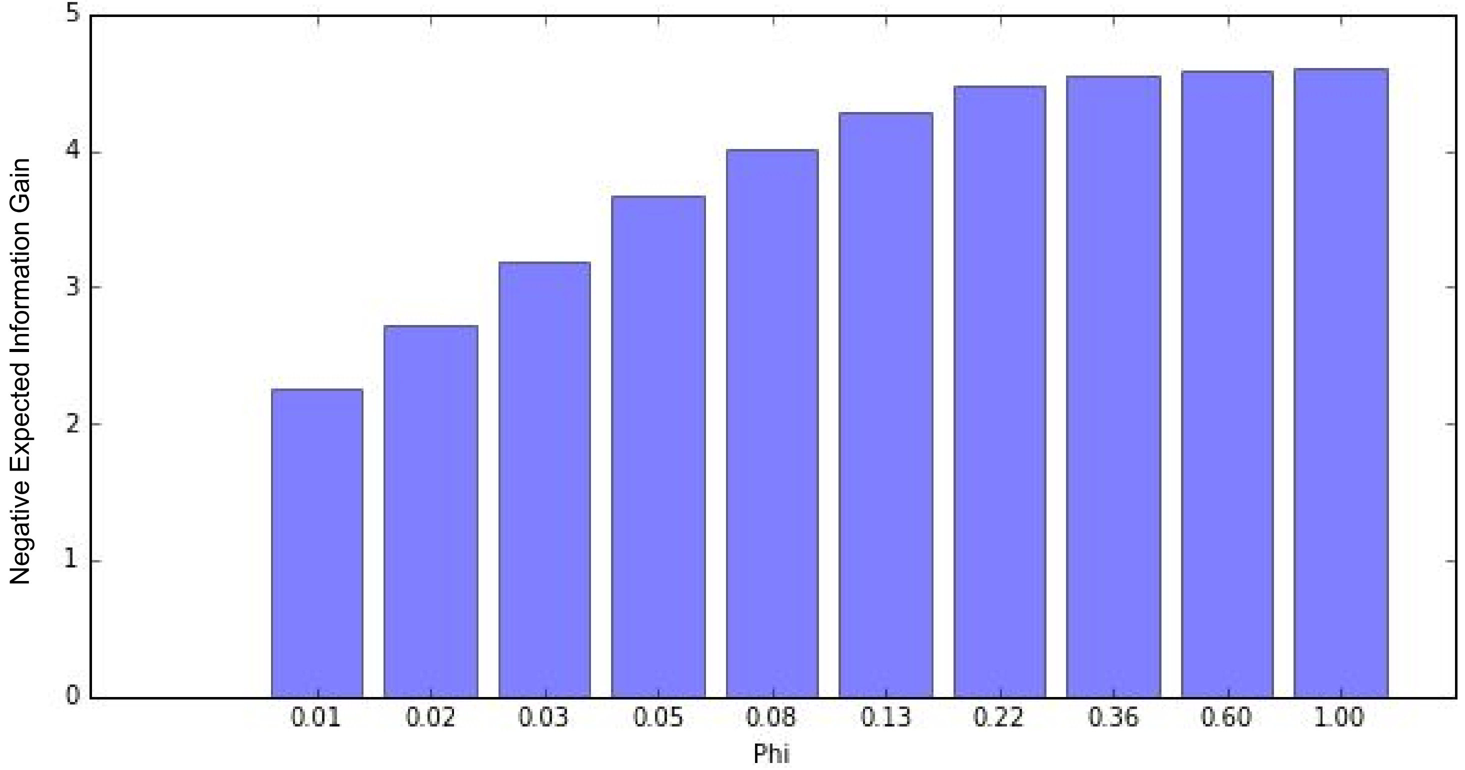
```
       __|  __|_  )
       _|  (     /   Amazon Linux AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-ami/2016.09-release-notes/
4 package(s) needed for security, out of 6 available
Run "sudo yum update" to apply all updates.
[ec2-user@ip-172-31-15-94 ~]$ export TOP=https://raw.githubusercontent.com/lukasheinrich/weinberg-exp/master/example_yadage
[ec2-user@ip-172-31-15-94 ~]$ eval "$(curl https://raw.githubusercontent.com/diana-hep/yadage/master/yadagedocker.sh)"
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   182  100   182    0     0    789      0 --:--:-- --:--:-- --:--:--   787
[ec2-user@ip-172-31-15-94 ~]$ yadage-run -t $TOP workdir rootflow.yml -p nevents=25000 -p seeds=[1,2,3,4] -a $TOP/input.zip \
>            -p runcardtempl=run_card.templ -p proccardtempl=sm_proc_card.templ \
>            -p sqrtshalf=45 -p polbeam1=0 -p polbeam2=0
```

… a single run (25k events) takes between 5 and 10 minutes.

Robot Physicist Version 3

- Revised calculation of expected information gain
- Profiled code to identify performance/scaling bottlenecks
- Documented code to tie it more closely to the theoretical writeup

Visualizing Beta Results



Potential Next Steps

- Expand system to operate on multidimensional inputs
- Create visualizations that give insight into the optimization process
- Do more learning around distributing Docker images for parallel execution

Thanks!