OASIS: "Better" simulated events to allow for fewer simulated events

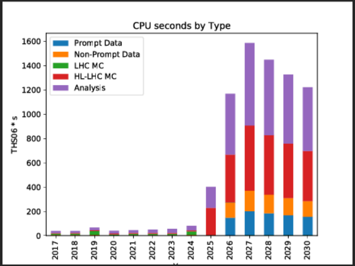
Prasanth Shyamsundar
University of Florida

based on [arXiv:2006.16972]
"OASIS: Optimal Analysis-Specific Importance Sampling for event generation"
Konstantin T. Matchev, Prasanth Shyamsundar

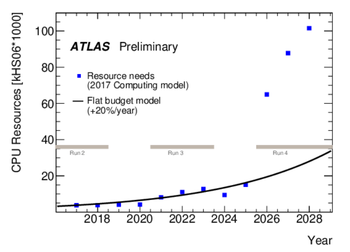
LPC Physics Forum, Fermilab
July 30, 2020

Motivation

- Simulations in HEP are computationally expensive.
  - Detector simulation is the most resource intensive part of the pipeline.
  - Projected HL-LHC computational requirements may not be met.
    "Billion dollar problem"
  - Need to speed up the simulation pipeline.



CMS

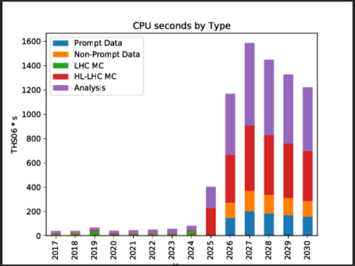


ATLAS

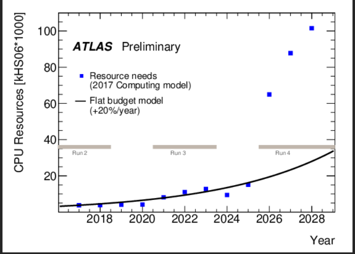
J. Albrecht *et al.* [HEP Software Foundation], "A Roadmap for HEP Software and Computing R&D for the 2020s," Comput. Softw. Big Sci. **3**, no.1, 7 (2019) [arXiv:1712.06982 [physics.comp-ph]].

Motivation

- Simulations in HEP are computationally expensive.
  - Detector simulation is the most resource intensive part of the pipeline.
  - Projected HL-LHC computational requirements may not be met.
    "Billion dollar problem"
  - Need to speed up the simulation pipeline. **Require fewer simulated events?**



CMS



ATLAS

J. Albrecht *et al.* [HEP Software Foundation], "A Roadmap for HEP Software and Computing R&D for the 2020s," Comput. Softw. Big Sci. **3**, no.1, 7 (2019) [arXiv:1712.06982 [physics.comp-ph]].

Importance Sampling

- The simulation pipeline starts with the parton level hard scattering.
- At the parton level, we can compute the probability density of a given event.

  (under a given theory/set of param values)
- Ingredients:
  - Matrix element
  - Parton distribution functions
- Given an oracle for a distribution, how do we sample events as per the distribution?

  Answer: Importance Sampling

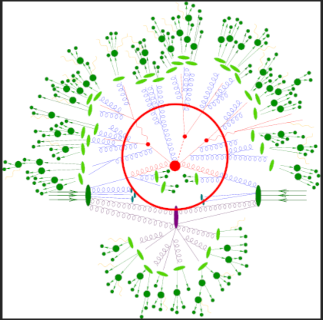
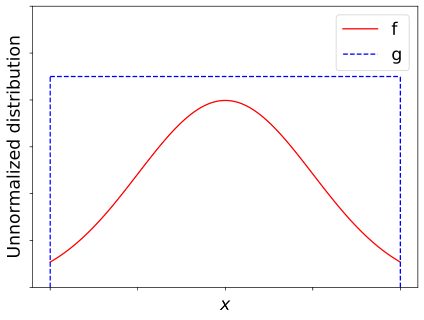


Image from the Sherpa Team

Importance Sampling

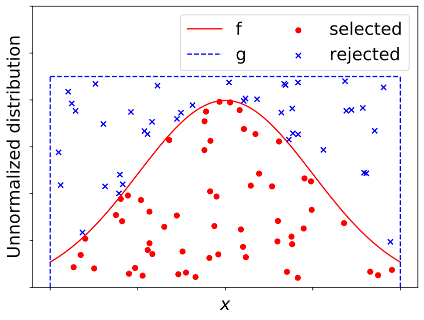


- f = distribution to sample from
  g = distribution we can sample from
  (both unnormalized)
- Throw darts uniformly at random into the "box".
  Or sample events according to g.
- Option 1: Unweighting
  - Accept the events that fall under f.
    Or accept event i with probability $f(x_i)/g(x_i)$.
- Option 2: Weighted events
  - Accept all events, but weight them

$$w_i = f(x_i)/g(x_i)$$

- The "box" g doesn't have to be a rectangle. Just needs to be something we can sample from.

Importance Sampling



- f = distribution to sample from
  g = distribution we can sample from
  (both unnormalized)
- Throw darts uniformly at random into the "box".
  Or sample events according to g.
- Option 1: Unweighting
  - Accept the events that fall under f.
    Or accept event i with probability $f(x_i)/g(x_i)$.
- Option 2: Weighted events
  - Accept all events, but weight them

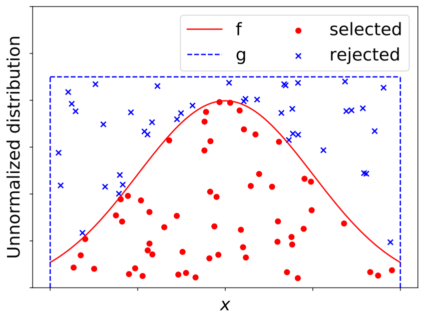
$$w_i = f(x_i)/g(x_i)$$

- The "box" g doesn't have to be a rectangle. Just needs to be something we can sample from.

Importance Sampling

Current philosophy: Try to make g close to f

Rationale 1:



Unweighting efficiency... circular argument

We want unweighted events

$\Downarrow$

$g \rightarrow f/F$ reduces wastage (lesser fraction of events thrown out)

$g \rightarrow f/F$ is ideal

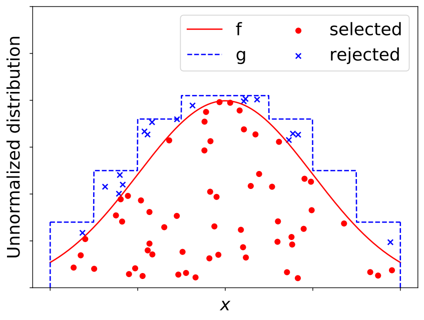
$\Downarrow$

We should unweight events at the parton level before moving onto the rest of the (computationally expensive) simulation pipeline

Importance Sampling

Current philosophy: Try to make g close to f

Rationale 1:



Unweighting efficiency... circular argument

We want unweighted events

$\Downarrow$

$g \to f/F$ reduces wastage (lesser fraction of events thrown out)

$g \to f/F$ is ideal

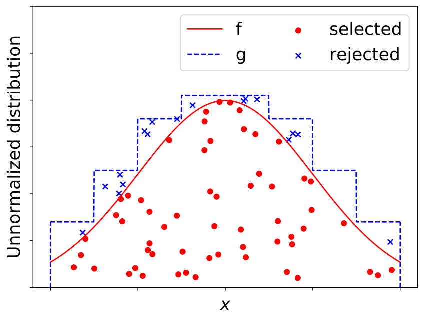
$\Downarrow$

We should unweight events at the parton level before moving onto the rest of the (computationally expensive) simulation pipeline

Importance Sampling

Current philosophy: Try to make g close to f

Rationale 2:



Cross-section estimation

$$F \equiv \int dx\, f(x) = \int dx\, g(x)\, \frac{f(x)}{g(x)}$$

$$= E_g[w] \qquad (\, g \text{ is normalized} \,)$$

$$\Rightarrow \hat{F} = \frac{1}{N_s} \sum_{i=1}^{N_s} w_i$$

$$\text{var}\left[\hat{F}\right] = \frac{\text{var}\,[w]}{N_s} \qquad (\, g \to f\,/\,F \text{ reduces variance} \,)$$

Estimation of F is related to counting experiments

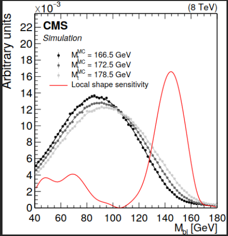
But... HEP analyses have come a long way
from counting experiments!

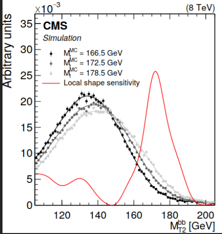
Weighted events = Yet unexplored degree of freedom

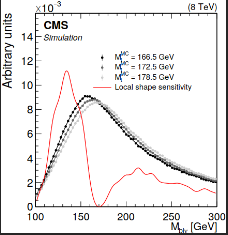
OASIS abondons the notion that $g \to f/F$ is the best strategy

- Nature:
  - Produces unweighted events
  - Constrained to be distributed as per f/F
- Weighted simulations:
  - Not constrained... Sampling distribution g can be whatever we want!
  - OASIS exploits this freedom to an unprecedented degree
- Current usage examples of weighted events:
  - Oversampling tails:
    Extract the sensitivity from the tails without wasting resources on the bulk
  - (Also reweighting events, combining different processes)
- **Why would we want to deviate from f/F on purpose?**
  - Focus on the regions of phase space important to the analysis.

An example: Top mass measurement







A. M. Sirunyan *et al.* [CMS], "Measurement of the top quark mass in the dileptonic $t\bar{t}$ decay channel using the mass observables $M_{b\ell}$, M_{T2}, and $M_{b\ell\nu}$ in pp collisions at $\sqrt{s} = 8$ TeV," Phys. Rev. D **96**, no.3, 032002 (2017) [arXiv:1704.06142 [hep-ex]].

- Different regions of the phase-space are sensitive to the value of a parameter (or presence of a signal) to different extents.
- More simulated events $\rightarrow$ smaller theory error bars
- Reducing the theory error bars everywhere (maintaining the same ratios between error bars) is not the optimal strategy!

OASIS elevator pitch

Optimal Analysis-Specific Importance Sampling

- Choose the sampling distribution optimally to maximize the sensitivity of the analysis at hand, for a given computational budget.
- Reach the target sensitivity with fewer simulated events.
- Piggyback on existing importance sampling techniques.
  (FOAM, VEGAS, machine-learning-based, etc)
- Save, in computational budget,

Hundreds of

OASIS for parton level analysis

- To pick a good sampling distribution g, we need to understand the relationship between the sampling distribution and the sensitivity of the analysis.
- Let θ be a parameter we want to measure by analyzing the parton level events $\{x_i\}$. Let L be the integrated luminosity.
- Fisher Information:

$$\mathcal{I}(\theta) = L \int dx \, \frac{1}{f(x\,;\,\theta)} \left[\frac{\partial f(x\,;\,\theta)}{\partial \theta}\right]^2$$

$$\mathrm{var}\left[\hat{\theta}(\textsf{Data})\,;\,\theta_0\right] \geq \frac{1}{\mathcal{I}(\theta_0)}$$

- The lower bound is achievable in the asymptotic limit by the maximum likelihood fit or minimum-χ^2 fit (fine binning).

Fisher Information for simulation based analyses

$$\mathcal{I}(\theta) = L \int dx \, \frac{1}{f(x;\theta)} \left[\frac{\partial f(x;\theta)}{\partial \theta} \right]^2$$

- Note that there's no g in the expression. This is for analyses based on the functional form of $f(x;\theta)$.
- What about analyses based on simulations?
  (N_s events distributed as per g)

$$\mathcal{I}(\theta) = \int dx \, \frac{1}{L f(x;\theta)} \left[L \frac{\partial f(x;\theta)}{\partial \theta} \right]^2$$

compare to $\sum_{i \in x \text{ bins}} \frac{s_i^2}{n_i}$ or $\sum_{i \in x \text{ bins}} \frac{s_i^2}{\sigma^2_{i,real\ stat}}$

$$\mathcal{I}_{\text{MC}}(\theta) = \int dx \, \frac{\left[L \frac{\partial f(x;\theta)}{\partial \theta} \right]^2}{L f(x;\theta) + N_s g(x) \left[\frac{L}{N_s} w(x;\theta) \right]^2}$$

$$\sigma^2_{i,real\ stat} \to \sigma^2_{i,real\ stat} + \sigma^2_{i,sim\ stat}$$

"s" $\sim$ difference between expected counts for θ and $\theta + \delta\theta$

Fisher Information for simulation based analyses

$$\mathcal{I}_{\mathsf{MC}}(\theta) = \int dx \, \frac{\left[L \, \dfrac{\partial f(x\,;\,\theta)}{\partial \theta}\right]^2}{Lf(x\,;\,\theta) + N_s g(x) \, \left[\dfrac{L}{N_s} \, w(x)\right]^2}$$

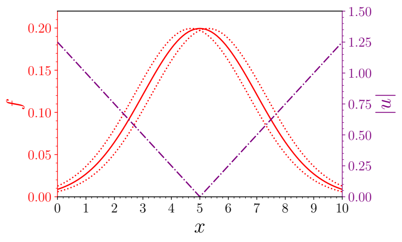
$$\Rightarrow \frac{\mathcal{I}_{\mathsf{MC}}(\theta)}{L} = \int dx \, \frac{f(x\,;\,\theta) \, \left[\partial_\theta [\ln f(x\,;\,\theta)]\right]^2}{1 + \dfrac{L}{N_s} w(x\,;\,\theta)}$$

$$\equiv \int dx \, \frac{f(x) \, u^2(x)}{1 + \dfrac{L}{N_s} w(x)} \qquad \text{where } \; u(x) \equiv \partial_\theta [\ln f(x\,;\,\theta)] = \frac{1}{f} \frac{\partial f}{\partial \theta}$$

$u(x)$ is a per-event score that captures the sensitivity of event to θ.
Can be computed using the matrix element oracle.

Some intuition + toy example

Measuring the mean of a normal dist



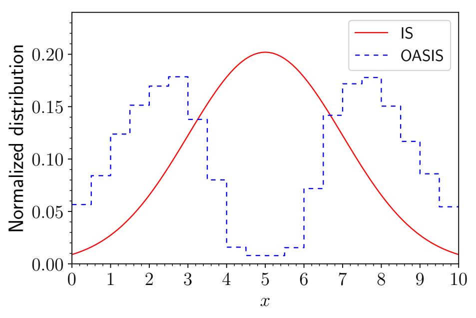
$$\theta_0 = 5$$

$$u = \frac{1}{f}\frac{\partial f}{\partial \theta}$$

$$\frac{\mathcal{I}_{\text{MC}}}{L} = \int dx \, \frac{f(x)\, u^2(x)}{1 + \frac{L}{N_s} w(x)}$$

- LHS: to maximize by picking a good sampling dist g.
- L/N_s is a heuristic parameter specifying our computational budget $\frac{L}{N_s} = F^{-1}\frac{N_r}{N_s}$
- g enters through w. Low w is good, but...
  $E_g[w] = \int dx g(x)\, f(x)/g(x) = F$ (fixed)
- **Assign low weights w where u is high (makes sense).**
- $\frac{L}{N_s}w(x)$ captures improvement from increasing sim.
- 1 captures the diminishing of returns.
  (real data is finite)

Training the sampling distribution



Ideal case Importance Sampling (IS) & Trained OASIS

- Parameterize g using $\vec{\varphi}$ as a piece-wise constant distribution given by

$$g(x) = \frac{p_{cell(x)}}{\text{Volume}_{cell(x)}}$$

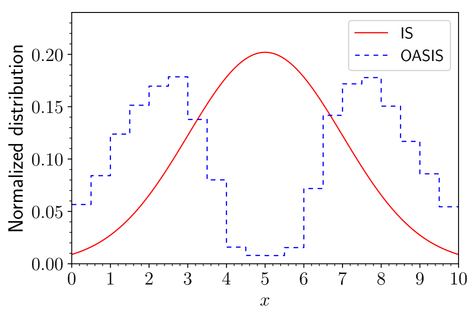
$$p_{cell\ i} = \frac{e^{\varphi_i}}{\sum_j e^{\varphi_j}} \qquad \text{(softmax)}$$

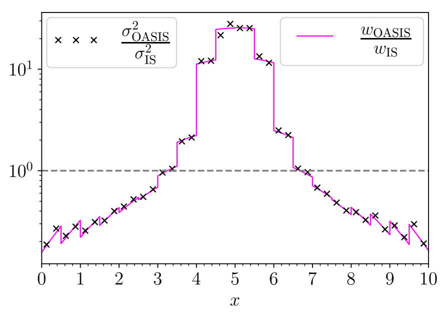
- Set $L/N_s = 1$ ($N_s \approx N_r$)
- Use gradient ascent to maximize $\mathcal{I}_{\text{MC}}$ (using preliminary/preexisting simulations as training data).

Weights

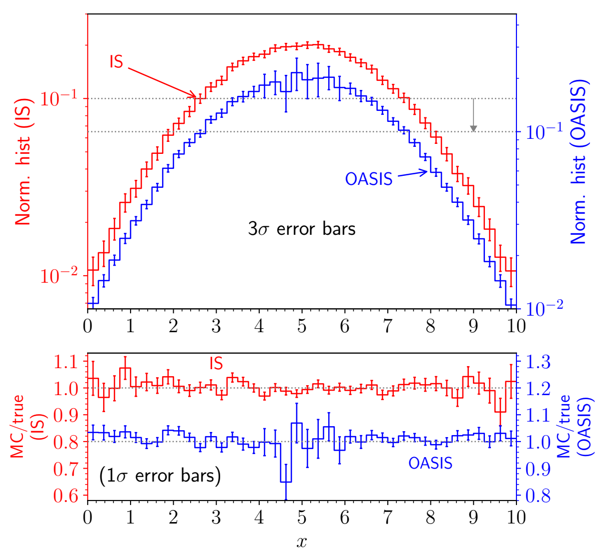
The weights compensate for the difference between g and $f\,/\,F$

$$w(x) = \frac{f(x)}{g(x)}$$





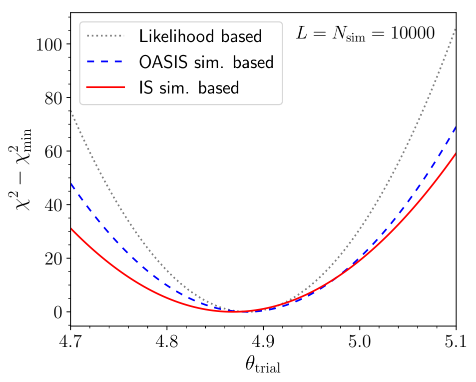
Effect on histograms



- Appropriately weighted histograms under OASIS and IS (100,000 events).
- Plotted on a log scale (with a shift).
- Both are consistent with the true distribution — importance sampling is a robust technique.
- IS has smaller error bars near the center.
- OASIS has smaller error bars away from the center.
- OASIS prioritizes based on utility to θ measurement.

  (Error bar ratios shown in previous slide)

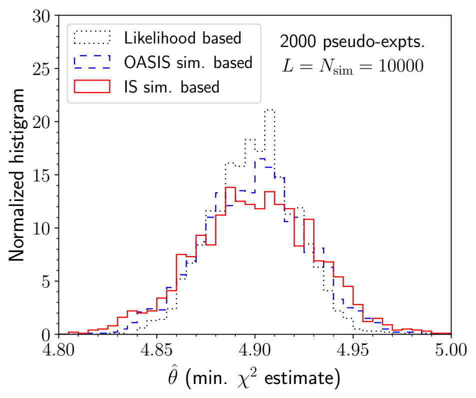
Effect on the measurement of θ



More concave $\sim$ smaller error bar

- Set $\theta_{\text{true}} = 4.9$
  - Simulate "real events", setting $L = 10,000$. $F(\theta_{\text{true}}) \approx 0.9875$
  - 9887 events produced in this pseudo-expt.
- Set simulation $\theta_0 = 5.0$
  (value at which OASIS is optimized)
  - Simulate 10,000 "simulated events" each under IS and OASIS.
  - Reweight them for different values of θ_{trial}.
- Perform simulation-based minimum-χ^2 estimation (40 bins).
- Gray dotted line is the likelihood based estimation (infinite simulation limit).

Effect on the measurement of θ



2000 such pseudo experiments

- Set $\theta_{\text{true}} = 4.9$
  - Simulate "real events", setting $L = 10,000$. $F(\theta_{\text{true}}) \approx 0.9875$
  - 9887 events produced in this pseudo-expt.
- Set simulation $\theta_0 = 5.0$
  (value at which OASIS is optimized)
  - Simulate 10,000 "simulated events" each under IS and OASIS.
  - Reweight them for different values of θ_{trial}.
- Perform simulation-based minimum-χ^2 estimation (40 bins).
- Gray dotted line is the likelihood based estimation (infinite simulation limit).

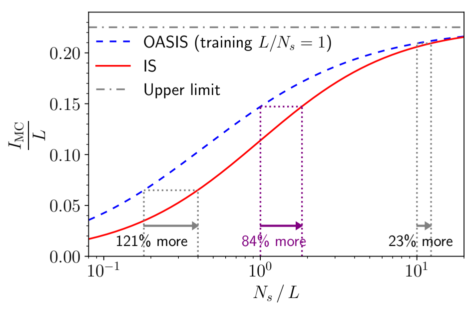
Effect on the measurement of θ

L	10,000			100,000		
N_s	10,000			100,000		
θ_{true}	4.9			4.9		
Training L/N_s	1			1		
Simulation θ_0	5			5		
Pseudo-expts.	2000			500		
	ave. $\hat{\theta}$	stdev $\hat{\theta}$	$[\mathcal{I}_{\text{MC}}(\theta_{\text{true}})]^{-1/2}$	ave. $\hat{\theta}$	stdev $\hat{\theta}$	$[\mathcal{I}_{\text{MC}}(\theta_{\text{true}})]^{-1/2}$
Likelihood-based	4.8997(5)	2.15(3)E−2	2.108(1)E−2	4.9001(3)	6.9(2)E−3	6.667(3)E−3
OASIS-based	4.9000(6)	2.64(4)E−2	2.611(2)E−2	4.8998(4)	8.5(3)E−3	8.258(5)E−3
IS-based	4.8999(7)	3.03(5)E−2	2.957(19)E−2	4.9004(4)	9.6(3)E−3	9.390(19)E−3

Simulation parameters and summary statistics of the results from the simulated pseudo-experiments to measure θ_{true}.

Note: $\mathcal{I}_{\text{MC}}$ is a good measure of sensitivity.

Resource conservation

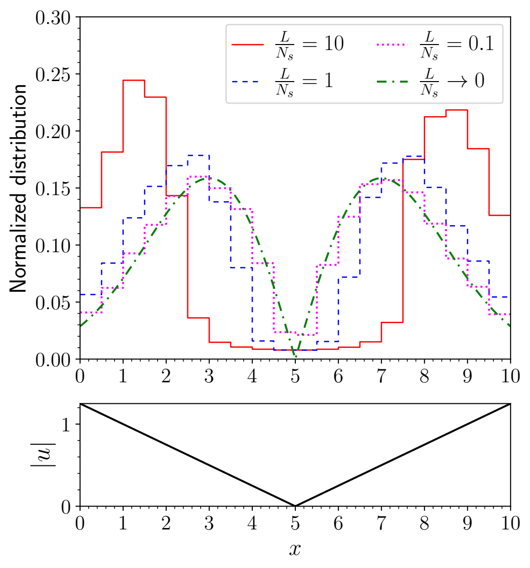


Upper-limit achieved in infinite statistics limit

- The $L\,/\,N_s$ set at training is just a heuristic parameter.
- The sampling distribution can be used to produce any number of events.
- OASIS achieves target sensitivities with fewer events than the ideal case IS.
- For a given number of simulated events, OASIS offers better sensitivity than IS.
- We're on a log scale...

  These numbers are impressive!
- We can do better than 23% at $N_s\,/\,L = 10$ if we train our sampling distribution there... Let's do that!

Varying the training L/N_s and special cases



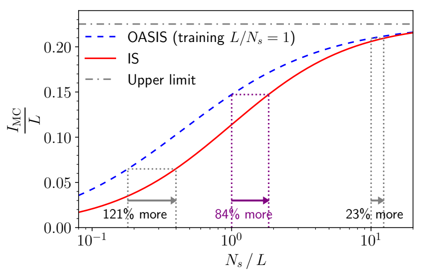
- All OASIS distributions prioritize regions of higher $|u|$.
- As training L/N_s decreases, the sampling distribution is more lenient towards low $|u|$ regions.
- Rationale: In the small N_s limit, focus on the regions of the *highest* $|u|$.

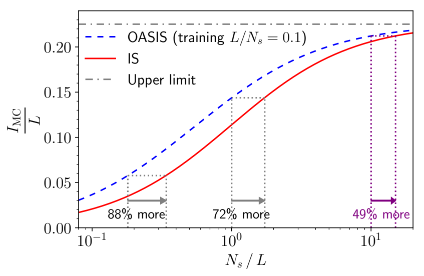
  (like a delta function)

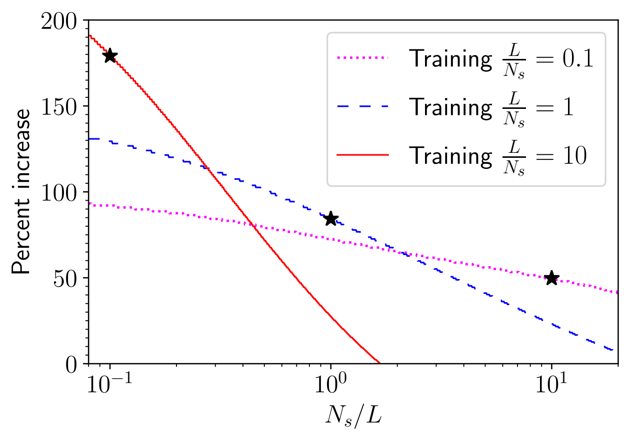
$$\frac{\mathcal{I}_{\mathrm{MC}}}{L} = \int dx \, \frac{f(x)\, u^2(x)}{\cancel{X} + \frac{L}{N_s} w(x)}$$

- As N_s increases, the utility of high $|u|$ regions saturates, so move towards lower $|u|$ regions.
- In the $N_s \to \infty$ limit, $g_{\mathrm{optimal}} \propto f|u|$.

More money plots







Resource conservation offered by OASIS distributions trained for different values of $L\,/\,N_s$.

OASIS at the analysis level

- Parton level events get mapped to analysis variables in a **probabilistic many-to-many** manner, via
  - Parton showers and Initial State Radiation
  - Hadronization
  - Detector simulation
  - Event reconstruction (+ some particles are invisible)
  - Event selection/categorization
  - High level variable calculation
- Also, analysis level datasets are composed of several subsamples.
- There are model uncertainties unrelated to simulation statistics

Q1) How is the sampling distribution related to sensitivity at the analysis level? (How do our equations change?)

Q2) How do we implement OASIS at the parton level when the quantity we are optimizing lives in the analysis realm?

How do the equations change?

- Let v be the possibly-multi-dimensional analysis level variable.
  (including categorization/event selection information)
- x is mapped to v via some transfer function.
- $\mathcal{F}(v\,;\,\theta)$ corresponds to $f(x\,;\,\theta)$
  $\mathcal{U}(v\,;\,\theta) = \partial_\theta\big[\ln[\mathcal{F}(v\,;\,\theta)]\big]$
- Events with the same v value can have different weights. $\mathcal{I}_{\text{MC}}$ becomes...

$$\frac{\mathcal{I}_{\text{MC}}}{L} = \int_{\text{selected events}} dv\, \frac{\mathcal{F}(v)\,\mathcal{U}^2(v)}{1 + \dfrac{L}{N_s}\,\dfrac{E_g[w^2\,|\,v]}{E_g[w\,|\,v]}}$$

- Multiple subsamples and systematics unrelated to simulation statistics...

$$\frac{\mathcal{I}_{\text{MC}}}{L} = \int_{\text{selected events}} dv\, \frac{\mathcal{F}(v)\,\mathcal{U}^2(v)}{1 + \dfrac{\sigma^2_{syst}(v)}{\sigma^2_{real\ \ stat}(v)} + \sum_k \dfrac{\mathcal{F}^{(k)}(v)}{\mathcal{F}(v)}\,\dfrac{L}{N_s^{(k)}}\,\dfrac{E_{g^{(k)}}[w^2\,|\,v]}{E_{g^{(k)}}[w\,|\,v]}}$$

Implementing OASIS at the analysis level

$$\frac{\mathcal{I}_{\text{MC}}}{L} = \int_{\text{selected events}} dv \, \frac{\mathcal{F}(v)\,\mathcal{U}^2(v)}{1 + \frac{\sigma^2_{syst}(v)}{\sigma^2_{real\ stat}(v)} + \sum_k \frac{\mathcal{F}^{(k)}(v)}{\mathcal{F}(v)} \frac{L}{N_s^{(k)}} \frac{E_{g^{(k)}}[w^2 \mid v]}{E_{g^{(k)}}[w \mid v]}}$$

- This expression lives at the analysis level. Importance sampling happens at the parton level...
- Simplifying observation: It is always better to minimize the variance of w in a given v bin. $E_g[w^2] = \text{var}_g\,[w] + (E_g[w])^2$.
- Limit attention to sampling distributions under which the weights (roughly) only depend on v.

$$\frac{\mathcal{I}_{\text{MC}}}{L} = \int_{\text{selected events}} dv \, \frac{\mathcal{F}(v)\,\mathcal{U}^2(v)}{1 + \frac{\sigma^2_{syst}(v)}{\sigma^2_{real\ stat}(v)} + \sum_k \frac{\mathcal{F}^{(k)}(v)}{\mathcal{F}(v)} \frac{L}{N_s^{(k)}} w^{(k)}(v)}$$

Stage 1: Taking stock at the analysis level

$$\frac{\mathcal{I}_{\text{MC}}}{L} = \int_{\text{selected events}} dv \, \frac{\mathcal{F}(v)\,\mathcal{U}^2(v)}{1 + \frac{\sigma^2_{syst}(v)}{\sigma^2_{real\ \ stat}(v)} + \sum_k \frac{\mathcal{F}^{(k)}(v)}{\mathcal{F}(v)} \frac{L}{N_s^{(k)}} w^{(k)}(v)}$$

Learn the "target distribution" or "target weights" $w^{(k)}_{\text{target}}(v)$ (up to a mult. constant)

- In this stage, the analysis groups decide how they want their simulated data to be distributed in the phase space of the analysis variable.
- This expression can be maximized using the same technique we saw earlier.

- *Trained OASIS distribution optimizing too aggressively?*
  Make it less aggressive by hand.
- *Signal search analysis?* Replace $\mathcal{U}$ with $s(v)/b(v)$.
- *Want simulations in control regions that aren't sensitive to θ?*
  Fix $\mathcal{U}$ in those regions (or the $w^{(k)}_{\text{target}}$) by hand.
- *Multiple analyses using the same dataset?* Find a middle ground

Try it out!

"How would the sensitivity change if we had more events here and less events there?"

Stage 1: Taking stock at the analysis level

$$\frac{\mathcal{I}_{\text{MC}}}{L} = \int_{\text{selected events}} dv \, \frac{\mathcal{F}(v)\, \mathcal{U}^2(v)}{1 + \frac{\sigma^2_{syst}(v)}{\sigma^2_{real\ stat}(v)} + \sum_k \frac{\mathcal{F}^{(k)}(v)}{\mathcal{F}(v)} \frac{L}{N_s^{(k)}} w^{(k)}(v)}$$

Learn the "target distribution" or "target weights" $w^{(k)}_{\text{target}}(v)$ (up to a mult. constant)

- In this stage, the analysis groups decide how they want their simulated data to be distributed in the phase space of the analysis variable.
- This expression can be maximized using the same technique we saw earlier.

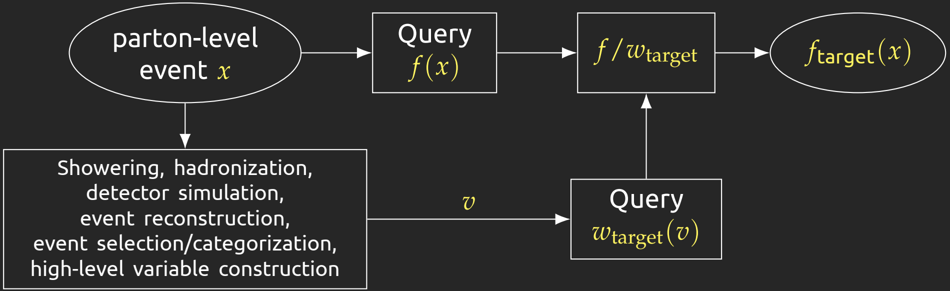
- *Trained OASIS distribution optimizing too aggressively?*
  Make it less aggressive by hand.
- *Signal search analysis?* Replace $\mathcal{U}$ with $s(v)/b(v)$.
- *Want simulations in control regions that aren't sensitive to θ?*
  Fix $\mathcal{U}$ in those regions (or the $w^{(k)}_{\text{target}}$) by hand.
- *Multiple analyses using the same dataset?* Find a middle ground :^)

Try it out!

"How would the sensitivity change if we had more events here and less events there?"

Stage 2: Translating the target weights to parton-level

- Importance sampling algorithms (FOAM, VEGAS, machine-learning-based) need an oracle which can be queried for $f(x)$ (unnormalized).
- They can train a sampling distribution g that mimics the oracle.
- Replace the oracle for f with the oracle for $f_{\text{target}}(x)$:



- Key idea: The map from x to v is approximately many-to-one.
  Non-determinism in $f_{\text{target}}(x)$ is low.
- f_{target} will have the same singularity structure as f... Fast sims are good enough for training... If v is rejected, return an appropriate low f_{target} value...

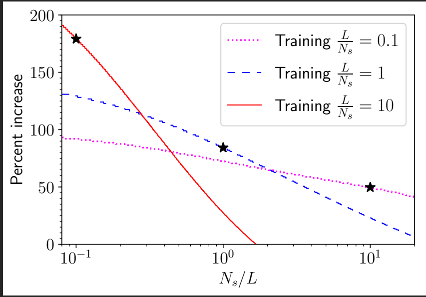
Outlook

Untapped and unexplored optimization

- The performance boost we see here is significant.
- This should not be surprising...
  We're not tweaking an existing approach to eke out some more sensitivity.
- We're opening an avenue of optimization that hasn't been explored yet.
- When working on the paper, a bug in the code led to a sampling distribution far from optimal — not avoiding the middle of the histogram as aggressively. Even that led to significant improvements.
  (See bonus slide)

Complementary to approaches that seek to speed-up the simulation pipeline

- Speed up using GPUs? GANs?
  OASIS can play along.



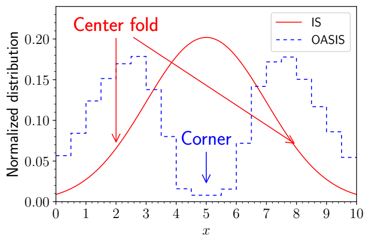
% increase in comp. requirements, when using IS instead of OASIS, to reach the same sensitivity

Outlook

Is OASIS just introducing a compromise, because we cannot generate the amount of data we need?

- OASIS *improves the compromise.
- By not simulating infinite statistics, we are already cutting corners.
- OASIS makes sure that what we are cutting are, in fact, corners.

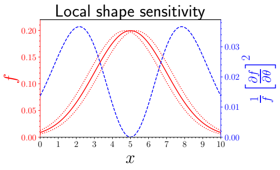
- It makes sense to use OASIS even if we have “enough” computational resources.

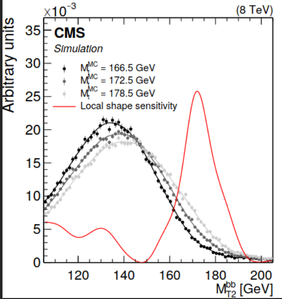


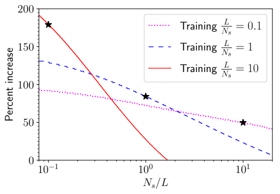
Outlook

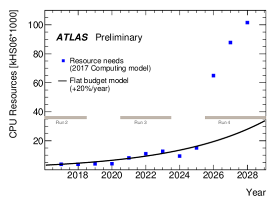
- Lookin at these plots...
  (notes on next slide)
- We are probably looking at savings of the order of **hundreds of millions of dollars** for HL-LHC alone.
- Implementation will likely be “simple”.
- Will require unprecedented level of cooperation between
  - MC theorists
  - MC groups within experiments
  - Physics analysis groups

Thank You! Questions?









Jump to
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26

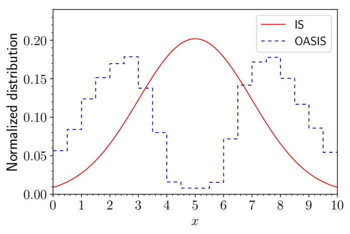
Notes for previous slide

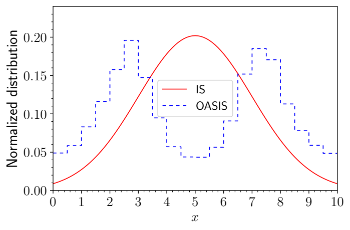
Things to consider:

- The similarity of the “local shape sensitivity” plots in the top row...
- The improvements seen in the bottom-left panel...
- The improvements needed in the bottom-right panel...
- “Billion dollar problem”...
- One the one hand, OASIS may not be appropriate or possible for some analyses...
- On the other hand, for events that don’t make it past the selection cuts, OASIS will lead to much greater resource conservation, by aggressively undersampling them...

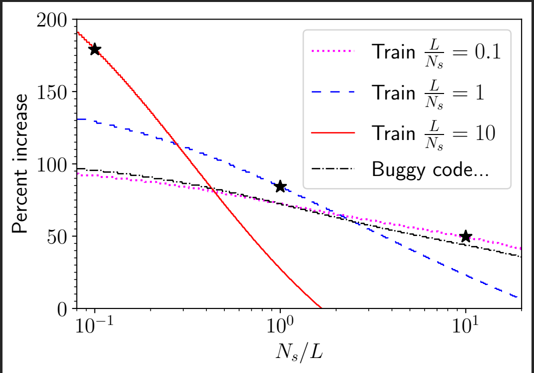
Bonus 1: Buggy code

Properly trained





Buggy code



OASIS doesn't have to be perfect to make a difference

Bonus 2: Special use cases...

- OASIS might be particularly useful for targeted analysis-specific QCD background simulation.
- I mentioned that nature is constrained to produce unweighted events.
  But maybe not...
  We have binary (in/out) triggers and we have unbiased prescale triggers.
  If there's place for a hybrid, OASIS-like ideas can help optimize it.