Efficient Simulated Data in Particle Physics: Event Generation, Reweighting, and Fits

Conceptual Summary for Electron Scattering Experiments (CLAS12 at Jefferson Lab)

Key Frameworks: `elSpectro` · `brufit` · ML Fast Simulation (arXiv:2207.11254)

Overview of the Simulation and Analysis Pipeline

1. **Statistical Motivation:** Minimizing variance and maximizing Kish's effective sample size (N_{eff}).
2. **Event Generation (`elSpectro`):** Optimizing Accept/Reject efficiency via factorized grid envelopes.
3. **Pre-Simulation Filtering (`clas12-elSpectro`):** Generator-level fiducial cuts and acceptance normalization.
4. **Post-Simulation Reweighting:** Density ratios in reconstructed space via histograms and ML classifiers.
5. **Maximum Likelihood Fits (`brufit`):** Dynamic physics model updates via weighted normalization integrals.
6. **Electron Scattering Kinematics (Q^2, t):** Virtual photon flux sampling and t-reweighting vs. t-binning.
7. **Machine-Learned Fast Simulation (arXiv:2207.11254):** Factorized per-particle acceptance and resolution.

Importance Sampling of Physical Observables

- **Foundational Goal:** Minimize the statistical uncertainty of extracted observables from a finite number of Monte Carlo (N_{MC}) events:

$$\langle \mathcal{O} \rangle = \int \mathcal{O}(x)\, p_{\text{data}}(x)\, dx$$

- When generating MC from a baseline distribution $p_{\text{gen}}(x)$, the integral is evaluated by weighting each event by the density ratio:

$$\langle \mathcal{O} \rangle \approx \frac{1}{N_{\text{MC}}} \sum_{i=1}^{N_{\text{MC}}} \mathcal{O}(x_i)\, w_i, \quad w_i = \frac{p_{\text{data}}(x_i)}{p_{\text{gen}}(x_i)}$$

Effective Sample Size (N_{eff})

Statistical power depends directly on the variance of the weights (Var_w) via **Kish's Effective Sample Size**:

$$N_{\text{eff}} = \frac{\left(\sum w_i\right)^2}{\sum w_i^2} \approx \frac{N_{\text{MC}}}{1 + \text{Var}_w/\bar{w}^2}$$

Key Conceptual Takeaway

- **Large Mismatch ($p_{\text{gen}} \neq p_{\text{data}}$):** Produces large weight variance ($\text{Var}_w \gg 0$), drastically reducing N_{eff} and wasting CPU cycles in low-probability phase space.
- **Theoretical Optimum ($p_{\text{gen}}(x) \equiv p_{\text{data}}(x)$):** Gives $w_i = 1$, $\text{Var}_w = 0$, and $N_{\text{eff}} = N_{\text{MC}}$. Matching simulation to data is a strict mathematical requirement for optimal statistical efficiency.

2. Event Generation and the Accept/Reject Efficiency

The Accept/Reject Method

- **Objective:** Generate unweighted events ($w = 1$) distributed according to a complex theoretical target distribution $f(x)$.
- **Envelope Requirement:** Introduce an inexpensive proposal distribution $g(x)$ scaled by a bounding constant M such that $f(x) \leq M \cdot g(x)$ across all kinematics x:

$$M = \max_x \left(\frac{f(x)}{g(x)} \right)$$

- **Generation Efficiency:** The global acceptance efficiency is exactly $\epsilon_{\text{gen}} = 1/M$.
- **Bottleneck:** If $g(x)$ poorly approximates $f(x)$ (e.g., flat generation under sharp resonance peaks), the local maximum of $f(x)/g(x)$ becomes large $\implies$ high rejection rate.

Adaptive Grid Envelopes in `elSpectro`

- Event generators such as `elSpectro` overcome high rejection rates by **discretizing phase space into grids**.
- **Factorized Histograms:** The proposal envelope $g(x)$ is constructed using factorized 1D and 2D histograms (e.g., separating virtual photon flux, momentum transfer, and decay kinematics).
- **Local Shape Wrapping:** Because the step-function histogram envelope is tailored to tightly wrap the shape of $f(x)$ locally, the ratio $f(x)/g(x)$ is minimized across the entire phase space.
- **Result:** Keeps $M \approx 1$, optimizing computational generation speed while producing exact, unweighted events.

3. Pre-Simulation Filtering and Acceptance Tracking

Generator-Level Fiducial Filtering Before GEANT4 (GEMC)

- **Computational Bottleneck:** Event generation in `elSpectro` is fast, whereas GEANT4 detector tracking (GEMC) and event reconstruction are computationally expensive.
- **Standard Practice:** Filter out generated events that fall entirely outside the physical geometric acceptance of the detector prior to running GEANT4.
- **CLAS12 Example (`clas12-elSpectro`):** In electron scattering, an electron generated at an angle too far forward to strike the active volume of the forward spectrometer (or outside the Forward Tagger / Forward Detector angular windows) is rejected immediately at the generator level.

Crucial Requirement: Downstream Cross-Section Bookkeeping

- While fiducial pre-filtering greatly increases simulation efficiency, **the removed phase space must be explicitly accounted for in the final cross-section normalization**.
- When evaluating detector acceptance:

$$\epsilon = \frac{N_{\text{reco}}}{N_{\text{gen}}}$$

the baseline generated luminosity (N_{gen} in the denominator and the associated generator cross section) **must correspond to the exact fiducial volume defined by the pre-cuts**, rather than the unconstrained 4π solid angle.

4. Post-Simulation Event Reweighting and Density Ratios

Reweighting in Reconstructed Space (Without Unfolding)

- When physics models are updated or generator parameters are tuned to data, baseline MC $p_{\text{base}}(x)$ is reweighted to match the target distribution $p_{\text{target}}(x)$ via $w(x) = p_{\text{target}}(x)/p_{\text{base}}(x)$ (keeping $w(x) \approx 1$ to preserve N_{eff}).
- **Experimental Reality:** True kinematics are unobservable in real data; without full multidimensional unfolding, the density ratio **must be evaluated between reconstructed real data and reconstructed MC** (x_{reco}).

1. Histograms (Low Dimensionality: 1D / 2D)

- Evaluate weights by directly dividing the normalized target (data) histogram by the base MC histogram:

$$w_{\text{bin}} = \frac{H_{\text{target}}(x_{\text{reco}})}{H_{\text{base}}(x_{\text{reco}})}$$

- **When to use:** Simple, transparent, and robust for well-populated 1D or 2D kinematic projections.

2. Probabilistic Classifiers (High-D: 4D / 5D)

- In correlated multi-dimensional phase spaces where histograms suffer from empty bins, train a classifier (e.g., DNN or BDT) to distinguish target ($y = 1$) from base ($y = 0$) events.
- A calibrated classifier probability $s(x)$ gives the exact same density ratio:

$$w(x) = \frac{p_{\text{target}}(x)}{p_{\text{base}}(x)} = \frac{s(x)}{1 - s(x)}$$

Weighted Normalization in Extended Likelihoods

- In unbinned extended maximum likelihood parameter extractions (such as amplitude and moment fits performed in `brufit`), minimizing $-\ln \mathcal{L}(\theta)$ requires evaluating the expected number of accepted events $N_{\text{exp}}(\theta)$ for each trial parameter hypothesis θ.
- The normalization integral over the detector acceptance is computed by summing over the weighted Monte Carlo events:

$$N_{\text{exp}}(\theta) \propto \sum_{j=1}^{N_{\text{MC}}} w_j \, \mathcal{P}(x_j \mid \theta)$$

where $\mathcal{P}(x_j \mid \theta)$ is the theoretical intensity and w_j incorporates generator/reweighting corrections.

Key Practical Advantages

- **Dynamic Physics Updates:** Modifying the expected yield via w_j allows the fit function in `brufit` to dynamically incorporate updated physics models or background tunings **without generating and processing a new GEANT4 MC sample** for every theoretical variation.
- **Controlling Fit Variance:** Because $N_{\text{exp}}(\theta)$ is evaluated from a finite MC sum, keeping w_j close to 1 ($\text{Var}_w \approx 0 \implies N_{\text{eff}} \approx N_{\text{MC}}$) is essential to prevent MC statistical fluctuations in the normalization integral from inflating parameter uncertainties on θ.

6. Kinematic Dependencies in Electron Scattering: Q^2 and t

Generator Baseline Requirements for Electron Scattering

- **Virtual Photon Flux (Q^2):** Generating scattered electrons with flat phase space is computationally non-viable due to the steep $1/Q^4$ behavior of the virtual photon propagator. Baseline MC (p_{base}) must be generated using a nominal virtual photon flux (as implemented in `elSpectro`).
- **Momentum Transfer (t):** Hadronic production is sharply peaked at low $|t|$ (e.g., $\propto e^{-b|t|}$), requiring a nominal t-slope in the generator to concentrate events in physically populated regions.

Method 1: Integrating over t via Reweighting

- Generate MC with a nominal t-slope, reweight so the simulated t-distribution matches data, and integrate over the full t-range.
- **Purpose:** Corrects the integrated cross section for strong variations of detector acceptance with t.
- **Vulnerability:** Relies heavily on the global reweighting model. If the generated nominal slope deviates too far from true data, large weight variance at high $|t|$ **drastically reduces** N_{eff}.

Method 2: Binning in t and Independent Fits

- Divide data and MC into discrete, narrow t-bins and perform the likelihood extraction (`brufit`) independently within each bin.
- **Physics Advantage:** Directly isolates the physical t-evolution of the cross section.
- **Statistical Robustness:** Because both acceptance and cross section vary little across a narrow bin, intra-bin reweighting is minimal ($w \approx 1$), preserving N_{eff} and making the extraction robust against global generator mismodeling.

7. Machine-Learned Fast Simulation (arXiv:2207.11254)

1. Motivation & Factorized Architecture

- Bypasses GEANT4 tracking and reconstruction overheads ($\geq$ 10–100 kHz rate).
- **Per-Particle Factorization:** Rather than simulating whole multi-particle events at once, treats each generated particle species individually in two sequential stages: **Acceptance** and **Reconstruction**.

2. Fast Acceptance via Density Ratios

- Binary classifier separates generated truth from accepted MC across (p, θ, ϕ):

$$w \approx \frac{c(x)}{1 - c(x)}$$

- **DNN + GBDT Cascade:** High-capacity DNNs with a **Gaussian quantile transform** capture 2D angular correlations and zero-acceptance gaps, followed by a **GBDT reweighting pass** to correct residual 1D deviations.

3. Fast Resolution via Nearest-Neighbour DTs

- Uses **over-fitted Decision Tree (DT) regressors** as a fast nearest-neighbour search mapping (p, θ, ϕ) + uniform random inputs to $(\delta p, \delta\theta, \delta\phi)$.
- **Ensemble Smoothing:** Picking randomly from an **ensemble of multiple DTs** (e.g., 50 trees with distinct random inputs) yields smooth continuous resolutions while preserving non-Gaussian tails and multi-dimensional correlations.

4. Applications & Statistical Caveat Mitigation

- **Uses:** Toy-MC parameter extraction tests, high-statistics background templates, and fast pre-filtering before GEANT4.
- **Statistical Power:** While generic ML generative models are capped by training statistics, **factorization mitigates this caveat**: independent per-particle models randomly mix in multi-body final states, creating novel combinatorial kinematics and vastly increasing effective statistical power.