高能物理实验中的蒙特卡洛模拟与探测器接受度

本文大部分使用AI生成。本人思考了很多问题，但是奈何这期作业数学方面我真的心有余而力不足，很多问题我只在脑子里思考，实际推导是真差得多。特别是线性代数方面的知识，上学期还能游刃有余，这学期我还得花时间补补，这期作业还有很多地方要学习。

1. 蒙特卡洛方法简史

布丰投针（1777）
法国博物学家布丰提出：在间距为 a 的平行线上随机投掷长度为 l（$l < a$）的针，针与线相交的概率 $p = 2l/(\pi a)$。通过大量投针实验统计相交次数，即可估算 π 值。这是人类首次用随机试验来求解确定性数学问题，其内核——利用统计频率逼近概率，再由概率反推物理量——正是蒙特卡洛思想的最早萌芽。

费米的中子扩散（1930年代）
恩里科·费米在罗马研究慢中子核反应时，面对中子输运和慢化这一极其复杂的随机过程，独立发明了一种手工随机抽样方法。他使用机械计算器产生伪随机数，跟踪单个中子的运动轨迹（散射、吸收、泄漏），通过对大量中子命运的统计，成功估算了临界质量等关键参数。费米的工作证明：即使在无电子计算机的时代，蒙特卡洛思想也能解决最尖端的核物理问题。

曼哈顿计划与正式诞生（1940年代）
二战期间，洛斯阿拉莫斯实验室为设计原子弹须精确计算中子链式反应。问题涉及高维积分和复杂边界，解析方法彻底失效。数学家斯坦尼斯瓦夫·乌拉姆在养病时联想到扑克牌的随机性，提出用随机抽样模拟核反应的设想。随后他与冯·诺依曼深入讨论，冯·诺依曼系统化了该方法，并以摩纳哥赌城“蒙特卡洛”命名。1949年，冯·诺依曼发表《与随机数字相关的各种技术》，结合第一台电子计算机ENIAC，首次完成了大规模中子输运的蒙特卡洛模拟。该方法从此成为严谨的数值计算框架。

历史意义
蒙特卡洛方法的发展实现了三次关键跨越：

- 方法论革命：开创了“用随机模拟求解确定性问题”的新范式；
- 计算物理融合：直接推动了计算机编程和随机数生成技术的发展；
- 跨学科普适性：从核物理扩展到粒子物理、统计力学、金融工程等领域，成为复杂系统建模的基石。

---

2. 现代高能物理中的蒙特卡洛模拟

2.1 什么是蒙特卡洛模拟

在高能物理中，蒙特卡洛（MC）模拟是指从粒子对撞的硬过程出发，经过量子场论计算、强子化、探测器响应、信号数字化到物理对象重建的全链条虚拟实验。标准流程为：

1. **事例产生**（Event Generation）
   基于量子色动力学（QCD）和电弱理论的事例产生器（如 Pythia, Herwig, Sherpa）：
   - 硬散射：按微扰理论计算反应截面，产生夸克、胶子、轻子等基本粒子；
   - 部分子簇射：模拟初末态辐射，形成喷注内部结构；
   - 强子化：用唯象模型将带色部分子组合成强子；
   - 强子衰变：产生寿命足够长、能飞入探测器的稳定粒子（$e^{\pm}, \mu^{\pm}, \gamma, \pi^{\pm}, K^{\pm}, p, n$ 等）。
2. **探测器模拟**（Detector Simulation）
   使用 Geant4 工具包，精确构建探测器的三维几何模型（包括磁场、支撑结构等），跟踪每个粒子在材料中的输运过程，模拟电磁簇射、强子簇射、电离能损、切伦科夫辐射等，记录在灵敏体积内沉积的能量和时间。
3. **数字化与重建**（Digitization & Reconstruction）
   - 将能量沉积转换为探测器实际输出的电信号（加入噪声、死道、饱和效应）；
   - 模拟触发系统的事例选择逻辑；
   - 用与真实数据完全相同的算法重建径迹、能量团、粒子种类。

最终得到的 MC 样本在格式上与真实实验数据完全一致，并附带“MC 真值”（MC truth），即每个重建对象对应的原始粒子信息。这一样本是连接理论与观测的桥梁。

2.2 为什么需要蒙特卡洛模拟

- **连接理论与观测**：理论计算的对象是夸克、胶子等基本粒子，而探测器看到的是强子化后的喷注、轻子、光子。MC 是唯一能从第一性原理建立两者定量关系的工具。
- **测量效率与接受度**：要得到绝对截面或分支比，必须修正探测器的选择效应（效率 ϵ 和接受度 A），这些修正因子全部依赖于 MC 模拟。
- **理解本底**：海量标准模型过程会模拟真实信号。通过 MC 样本可以精确估计信号区内的本底成分与产额，并验证本底估计方法的可靠性。
- **信号显著性判别**：通过模拟纯本底实验，构造本底涨落的期望分布，量化观测到的超出的统计显著性（如 5σ）。
- **探测器设计与优化**：在实验建造前，通过大规模 MC 研究比较不同设计方案，优化子探测器布局、磁场强度、触发逻辑等。
- **统计推断与系统误差**：通过修改模拟参数评估理论模型依赖、探测器几何不确定性等带来的系统误差。

现代高能物理的任何一项定量结果，其可信度都直接建立在蒙特卡洛模拟的精确性之上。

---

3. 核心方法与数学原理

蒙特卡洛模拟的根基在于“生成服从任意目标概率密度 $f(x)$ 的样本”，因为其核心思想是**将物理过程或数学问题转化为随机抽样实验**，而所有后续的统计推断都建立在样本分布的正确性之上。这可以从三个层面理解：

（1）数学本质：用随机抽样计算积分

蒙特卡洛方法最根本的用途是计算高维积分。关键点在于：有了分布就有了一切。任何物理量（截面、效率、接受度等）最终都可表达为某一概率密度下的期望值：

$$I = \int g(x) f(x)\, dx = \mathbb{E}_f[g(x)].$$

这个积分的常规处理方法是直接求解，但是在大部分情况下是很难求解的，因此需要通过积分估计。唯一需要的就是能够从 $f(x)$ 中独立抽取样本 $\{x_i\}$，然后用样本均值 $\frac{1}{N}\sum_i g(x_i)$ 逼近 I。**没有从 $f(x)$ 抽样的能力，这一框架就彻底崩塌。** 大数定律保证该估计收敛，中心极限定理提供误差估计，而这一切的前提都是样本服从正确的 $f(x)$。

（2）物理模拟的本质：将自然规律翻译为概率密度

在高能物理中，粒子反应的末态粒子配置（动量、方向、种类）完全由量子场论的**微分散射截面**决定。微分截面本身就是一种概率密度：$\frac{d\sigma}{d\Omega} \propto$ 粒子飞入某方向的事例权重。
完整的 MC 模拟链（事例产生→探测器模拟→重建）本质上是反复执行：

- 从微分截面 $f_{\text{physics}}(x)$ 抽样产生一个末态粒子配置；
- 在探测器模拟中，再从粒子与物质相互作用的概率分布抽样其命运；
- 在数字化中，再从电子学噪声等分布抽样得到电信号。
  整条链上的每一个环节都是“按某个分布随机抽样”。因此，能够从任意给定的 $f(x)$ 生成样本，就是让模拟能够“忠实复现”物理规律的最底层能力。

（3）通用性与可解耦性

实际物理分布 $f(x)$ 往往极为复杂（高维、无解析形式、仅能计算函数值）。逆变换法和接受-拒绝法的存在，保证了**哪怕 $f(x)$ 无法积分、无法求逆，只要我们能计算它的值，就能生成其样本**。这一保证赋予了蒙特卡洛方法处理任意复杂系统的灵活性，使其成为“不可计算”问题面前的通用工具。
同时，它将**分布生成**与**物理内容**解耦：算法只关心“如何从给定的 $f(x)$ 抽样”，而不需要知道 $f(x)$ 的物理来源。这样，理论物理学家可以任意修改模型、更换微分截面，模拟程序只需更新 $f(x)$ 的计算模块，而抽样引擎保持不变。

一句话概括：**蒙特卡洛模拟是用随机样本的统计行为替代解析推导，而“生成服从目标分布的样本”就是构造这个随机统计实验的唯一入口。** 没有这一根基，整个方法体系便无从建立。这正是冯·诺依曼、乌拉姆等人当年在曼哈顿计划中首先要解决“如何生成任意分布随机数”这一问题的根本原因。

既然蒙特卡洛模拟的根基在于：从均匀随机数 $U \sim \text{Uniform}(0,1)$ 生成服从任意目标概率密度 $f(x)$ 的样本。高能物理中的微分截面常无法直接写出累积分布的逆函数，因此发展了两类基础算法。

3.1 逆变换法（Inverse Transform Method）

历史：该方法是概率积分变换的直接推论，很早就被统计学家熟知，20世纪40年代随蒙特卡洛模拟兴起被正式确立为通用技术。
原理推导：

设连续随机变量 X 有严格递增的累积分布函数（CDF）$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt$，其逆函数为 $F^{-1}(u)$。则有以下定理：

> 若 $U \sim \text{Uniform}(0,1)$，则 $X = F^{-1}(U)$ 服从目标分布 $F(x)$。

证明：对任意实数 x，

$$P(X \le x) = P(F^{-1}(U) \le x) = P(U \le F(x)) = F(x).$$

第一个等号代入 X 的定义；第二个等号是因为 F 严格递增，事件 $F^{-1}(U) \le x$ 等价于 $U \le F(x)$；第三个等号直接利用均匀分布的性质 $P(U \le a) = a$（$0 \le a \le 1$）。因此 X 的 CDF 正好是 F，其概率密度即为 $f(x)$。

适用条件：要求 F^{-1} 能解析表达或高效数值计算。例如指数分布 $f(x) = \lambda e^{-\lambda x}$，$F^{-1}(u) = -\frac{1}{\lambda}\ln(1-u)$，可直接生成。

3.2 接受-拒绝法（Acceptance-Rejection Method）

历史：由冯·诺依曼于1947年左右提出，用于在早期计算机上生成任意分布的随机数，1951年系统发表。这是蒙特卡洛方法走向实用的关键突破。
问题设定：目标密度 $f(x)$ 无法快速求逆，但能容易地计算函数值。我们寻找一个“提议分布” $g(x)$，它易于采样，且存在常数 $C \ge 1$ 使得对定义域内所有 x 有

$$f(x) \le C g(x).$$

C 可取 $C = \sup_x \frac{f(x)}{g(x)}$，以保证封套函数 $Cg(x)$ 完全覆盖 $f(x)$。

算法步骤：

1. 从提议分布 $g(x)$ 中抽取候选点 x。
2. 独立地生成均匀随机数 $u \sim U(0,1)$。
3. 如果 $u \le \frac{f(x)}{Cg(x)}$，则**接受** x，输出该样本；否则**拒绝** x，返回步骤 1。

正确性证明（两种视角）：

(a) 条件概率视角
在单次试验中，候选点 x 被抽中的密度为 $g(x)$，且它被接受的概率为 $\alpha(x) = \frac{f(x)}{Cg(x)}$。因此，输出样本（被接受条件下的 x）的联合非归一化密度为

$$p_{\text{acc}}(x) = g(x) \cdot \alpha(x) = g(x) \cdot \frac{f(x)}{Cg(x)} = \frac{f(x)}{C}.$$

整体接受概率为

$$P_{\text{accept}} = \int p_{\text{acc}}(x)\,dx = \int \frac{f(x)}{C}\,dx = \frac{1}{C}.$$

于是，在“被接受”这一条件下的条件概率密度为

$$\frac{p_{\text{acc}}(x)}{P_{\text{accept}}} = \frac{f(x)/C}{1/C} = f(x).$$

因此，算法最终输出的每一个 x 都精确服从目标分布 $f(x)$。

(b) 几何-均匀视角（冯·诺依曼原始理解）
在二维平面中考虑曲线 $y_{\max}(x) = Cg(x)$ 下方的区域 $\mathcal{R} = \{(x, y) : 0 \le y \le Cg(x)\}$。

- 从 $g(x)$ 抽取 x，再在 $[0, Cg(x)]$ 上均匀抽取 y，等价于在区域 $\mathcal{R}$ 内产生均匀分布的点（联合密度为常数 $1/C$）。
- 接受条件 $u \le \frac{f(x)}{Cg(x)}$ 可改写为 $y \le f(x)$（因为 $y = u \cdot Cg(x)$）。因此，只有落在目标曲线 $y = f(x)$ 下方的点才被保留。
- 保留点的 x 坐标边缘密度正比于该垂直线段内曲线的“高度” $f(x)$，即 $\int_0^{f(x)} dy = f(x)$。归一化后即为 $f(x)$。

这一构造表明：**无论 $g(x)$ 是否为常数，只要实现了在 $Cg(x)$ 下方的均匀采样，最终输出的 x 就一定服从 $f(x)$。** 提议分布 $g(x)$ 仅影响抽样效率（接受概率 $1/C$），不影响结果分布的正确性。

常数 C 的选取与存在性：
要求 C 是 $\frac{f(x)}{g(x)}$ 的上确界。只要 g 的支撑集覆盖 f 的支撑集且二者之商有界，C 一定存在。实践中为了高效，常选择形状接近 f 的 g，使 C 尽可能小。

非常重要的一点：“归一化”如何自动完成
算法输出的 x 是直接从被接受的事例中取得的，其数值就是原先的候选值，未经任何缩放。被拒事件的丢弃这一动作，改变了样本中不同 x 的相对比例，使得整体分布恰好重塑为 $f(x)$。数学上，接受率 $1/C$ 在条件概率的分母中被约掉，因此最终样本无需任何额外的归一化操作。

3.3 微分散射截面与概率密度

在高能物理中，产生粒子的方向、能量分布由理论给出的**微分散射截面** $\frac{d\sigma}{d\Omega}$ 决定。

- 总截面 σ 具有面积量纲（$1\ \text{barn} = 10^{-28}\ \text{m}^2$），它源于经典散射图像：$\sigma$ 相当于靶粒子对入射粒子提供的“有效面积”，通过 $R = \sigma L$（L 为亮度）将通量换算为事例率。量子力学中该图像得以保留，但 σ 由相互作用动力学决定，不再对应几何尺寸。
- $\frac{d\sigma}{d\Omega}$ 是单位立体角内的截面，它给出了末态粒子飞入某方向的概率权重。例如在质心系中，反应 $e^+e^- \to \mu^+\mu^-$ 的微分散射截面（无极化、忽略电子质量）为

  $$\frac{d\sigma}{d\Omega} \propto 1 + \cos^2\theta,$$

  其中 θ 是 μ^- 相对于电子束流方向的极角。
- 将这一角分布归一化，就得到了 $\cos\theta$ 的概率密度函数：

$$f(\cos\theta) = \frac{1+\cos^2\theta}{\int_{-1}^{1}(1+t^2)dt} = \frac{3}{8}(1+\cos^2\theta), \quad \cos\theta \in [-1,1].$$

在蒙特卡洛事例产生器中，正是依据此类 $f(x)$ 抽样末态粒子的四动量。

---

4. 探测器的接受度

4.1 物理意义

接受度（Acceptance，符号 A）定义为：**在没有任何探测效率损失（理想重建、无死道）的条件下，仅因探测器的几何覆盖范围和基本运动学阈值，能被“看见”的事例占理论上产生的事例的比例。**

- 真实探测器不可能覆盖全立体角，比如径迹室的赝快度覆盖通常为 $|\eta| < 2.5$，对应 $|\cos\theta| \lesssim 0.9$；强子量能器可能延伸到更大快度，但沿束流方向仍有缝隙。
- 此外，低横动量粒子因弯曲半径太小或能量太低，即使击中探测器也可能无法触发或重建。
  因此，接受度本质上是**探测器几何边界与最低运动学要求所施加的必然相空间截断**。它纯粹由粒子层面的选择（particle-level selection）决定，与探测器具体响应的微观过程无关。将接受度 A 与重建、识别、触发效率 ϵ 相乘，便得到总有效选择效率 $A \times \epsilon$，用于将观测事例数修正回理论截面。

4.2 如何计算接受度

接受度**仅用产生器层面的粒子信息计算，不涉及探测器的详细模拟**，从而干净分离几何效应与效率。
标准步骤（以双缪子末态为例）：

1. 用 MC 产生器生成大量信号事例，记录每个末态粒子的四动量。生成时尽可能不加任何与探测器有关的截断。
2. 定义**基准选择（fiducial selection）**，直接映射探测器的几何接受和触发/运动学条件。例如：
   - 两个 μ 子的 $|\cos\theta| < 0.8$；
   - 横动量 $p_T > 20\ \mathrm{GeV}$；
   - 双缪子不变质量窗口等。
3. 对每个生成事例，检查末态粒子是否满足基准选择。
4. 接受度 $A = \dfrac{N_{\text{passed}}}{N_{\text{generated}}}$，其中分母是总生成事例数，分子是满足基准选择的事例数。这本质上是在生成相空间内，对微分截面做加权积分：

$$A = \frac{1}{\sigma_{\text{generated}}} \int_{\text{fiducial}} \frac{d\sigma}{d\Omega} d\Omega.$$

---

5. 完整习题示例：$e^+e^- \to \mu^+\mu^-$ 的蒙特卡洛模拟与接受度

5.1 物理图景与目标分布

质心系能量 $\sqrt{s} = 10\ \mathrm{GeV}$，正负电子对撞湮灭为虚光子，产生一对 $\mu^+\mu^-$。量子电动力学预言微分散射截面正比于 $1 + \cos^2\theta$，θ 为 μ^- 的极角。由于动量守恒，在质心系中 μ^+ 与 μ^- 背对背飞行，故 $\cos\theta_{\mu^+} = -\cos\theta_{\mu^-}$，两者的 $|\cos\theta|$ 相等。
探测器径迹室仅覆盖 $|\cos\theta| < 0.8$。一个事例能被“看见”的条件是两个 μ 子都击中径迹室，即 $|\cos\theta| < 0.8$。

令 $x = \cos\theta$，其概率密度归一化为

$$f(x) = \frac{1 + x^2}{\int_{-1}^{1}(1 + t^2)dt} = \frac{3}{8}(1 + x^2), \quad x \in [-1, 1].$$

$f(x)$ 最大值为 0.75（在 $x = \pm 1$ 处）。

5.2 接受度解析计算

接受度 $A = P(|x| < 0.8) = \int_{-0.8}^{0.8} f(x)\,dx$。
计算：

$$\begin{aligned} A &= \frac{3}{8}\int_{-0.8}^{0.8}(1 + x^2)\,dx = \frac{3}{4}\int_{0}^{0.8}(1 + x^2)\,dx \\ &= \frac{3}{4}\left[x + \frac{x^3}{3}\right]_0^{0.8} = \frac{3}{4}\left(0.8 + \frac{0.512}{3}\right) \\ &\approx \frac{3}{4} \times 0.9707 \approx 0.728. \end{aligned}$$

约 72.8% 的事例满足几何接受条件。

5.3 蒙特卡洛模拟设计：接受-拒绝法生成 $\cos\theta$

我们采用最简单的封套：$g(x)$ 为 $[-1, 1]$ 上的均匀分布，密度 $g(x) = 0.5$。取 $C = \sup_x f(x)/g(x) = 0.75/0.5 = 1.5$，则封套曲线为常数 $Cg(x) = 0.75$，形成一个高度 0.75、宽度 2 的矩形。

算法步骤：

1. 生成 $x \sim U[-1, 1]$：`x = 2*rand.Rndm() - 1`；
2. 独立生成 $y \sim U[0, 0.75]$：`y = 0.75*rand.Rndm()`；
3. 若 $y \le f(x) = \frac{3}{8}(1 + x^2)$，则接受 x，否则重复；
4. 对接受的 x，若 $|x| < 0.8$ 则计数器 `nAccepted` 加一；
5. 同时可模拟探测器效率：对每个几何接受的 μ 子，独立地以概率 $\epsilon = 0.9$ 判断是否重建成功（再生成均匀随机数，小于0.9即成功），两个 μ 都成功则计数器 `nDetected` 加一。

MC估计量：

$$\hat{A} = \frac{N_{\text{acc}}}{N}, \quad \sigma_{\hat{A}} = \sqrt{\frac{\hat{A}(1-\hat{A})}{N}}.$$

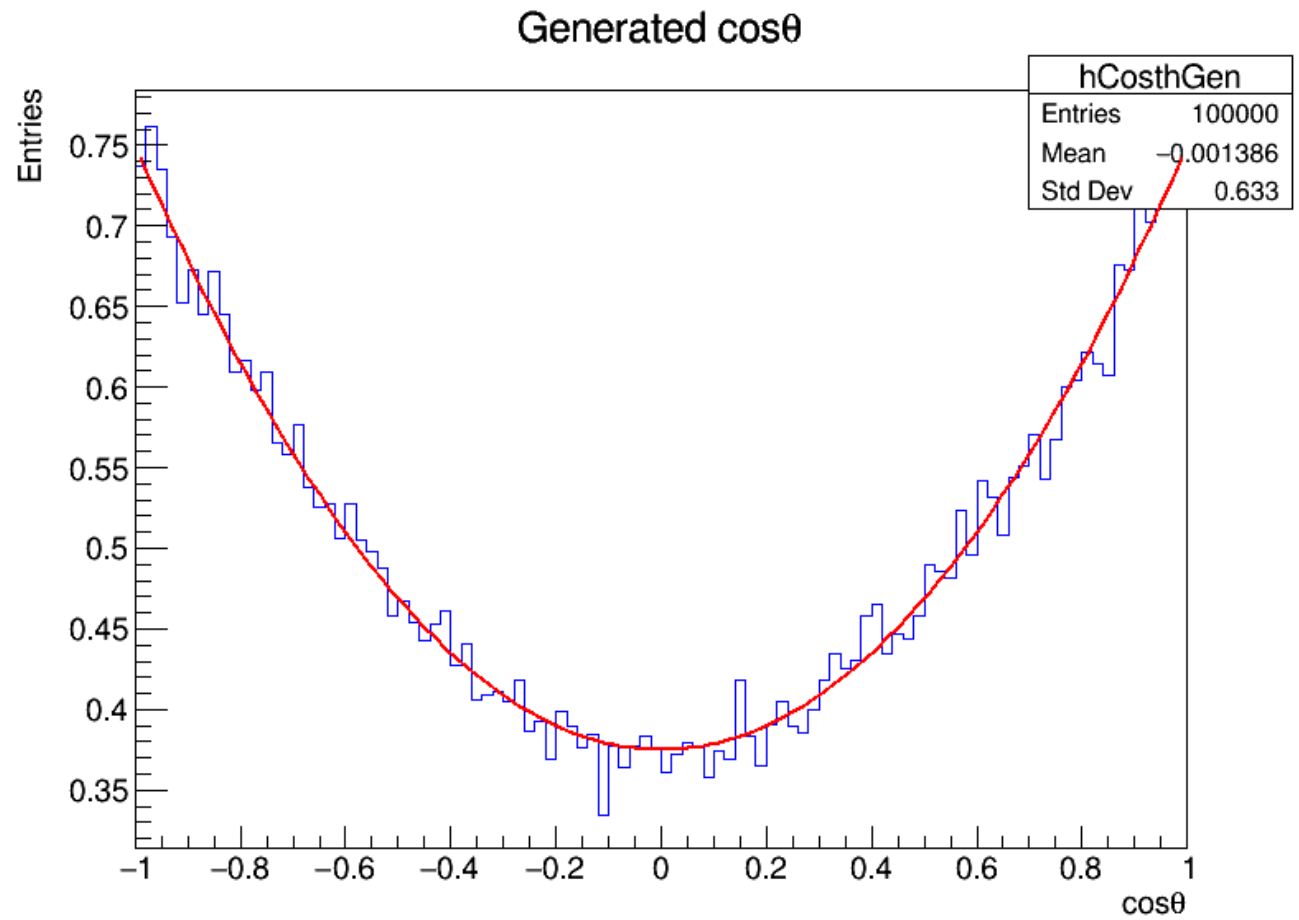
总有效效率 $\hat{\epsilon}_{\text{total}} = \frac{N_{\text{detected}}}{N}$，预期值 $A \cdot \epsilon^2 = 0.728 \times 0.81 = 0.589$。

5.4 ROOT 完整代码实现（C++）

以下宏可直接在 ROOT 环境运行。代码展示了接受-拒绝法抽样、几何接受判断、重建效率模拟，并与解析理论曲线对比验证。

```
// MC_exercise.C
#include "TRandom3.h"
#include "TH1D.h"
#include "TCanvas.h"
#include "TF1.h"
#include "TMath.h"
#include <iostream>

void MC_exercise(int N = 100000) {
    // 初始化随机数生成器 (Mersenne Twister)
    TRandom3 rand(0); // seed=0 使用机器时间种子

    // 目标概率密度 f(x) = 3/8*(1+x^2)
    auto f = [](double x) { return 3.0/8.0 * (1 + x*x); };

    // 直方图用于检验分布
    TH1D *hCosthGen = new TH1D("hCosthGen", "Generated
cos#theta;cos#theta;Entries/bin", 100, -1, 1);
    TH1D *hCosthAcc = new TH1D("hCosthAcc", "Geometrically accepted
cos#theta;cos#theta;Entries/bin", 100, -1, 1);

    int nGenerated = 0;
    int nAccepted  = 0;   // 几何接受
    int nDetected  = 0;   // 几何接受 + 双缪子重建成功

    double eff_muon = 0.9; // 单缪子重建效率

    while (nGenerated < N) {
        // ---- 步骤1：接受-拒绝法生成服从 f(x) 的 x ----
        double x, y;
        do {
            x = 2.0 * rand.Rndm() - 1.0;    // U[-1,1]
            y = 0.75 * rand.Rndm();          // U[0,0.75]
        } while (y > f(x));                  // 若在曲线上方则重试

        hCosthGen->Fill(x);
        nGenerated++;

        // ---- 步骤2：几何接受度判断 ----
```

```
        if (TMath::Abs(x) < 0.8) {
            nAccepted++;
            hCosthAcc->Fill(x);

            // ---- 步骤3：模拟重建效率 ----
            bool reco_mu_minus = (rand.Rndm() < eff_muon);
            bool reco_mu_plus  = (rand.Rndm() < eff_muon);
            if (reco_mu_minus && reco_mu_plus) {
                nDetected++;
            }
        }
    }

    // ---- 结果输出 ----
    double A_est = double(nAccepted) / N;
    double A_err = TMath::Sqrt(A_est * (1 - A_est) / N);
    double eff_total_est = double(nDetected) / N;
    double eff_total_err = TMath::Sqrt(eff_total_est * (1 - eff_total_est) /
N);
    double A_analytic = 0.728;

    std::cout << "========== 蒙特卡洛模拟结果 ==========" << std::endl;
    std::cout << "总生成事例数: " << N << std::endl;
    std::cout << "解析接受度 A: " << A_analytic << std::endl;
    std::cout << "MC 估计接受度 A: " << A_est << " ± " << A_err << std::endl;
    std::cout << "MC 总有效效率 (A·ε²): " << eff_total_est << " ± " <<
eff_total_err << std::endl;
    std::cout << "解析总效率: " << A_analytic * eff_muon * eff_muon <<
std::endl;

    // ---- 绘制分布验证 ----
    TCanvas *c1 = new TCanvas("c1", "cos#theta distribution", 800, 600);
    hCosthGen->SetLineColor(kBlue);
    hCosthGen->Scale(1.0 / hCosthGen->Integral("width")); // 归一化为概率密度
    hCosthGen->Draw("hist");

    TF1 *f_theory = new TF1("f_theory", "3.0/8.0*(1+x*x)", -1, 1);
    f_theory->SetLineColor(kRed);
    f_theory->SetLineWidth(2);
    f_theory->Draw("same");

    c1->SaveAs("cos_theta_distribution.png");
}
```




```
hyp@fedora:~/Download/workarea — root -l MC_exercise.C
~/Download/workarea

bash: /home/hyp/root_install/bin/thisroot.sh: No such file or directory
hyp@fedora:~/Download/workarea$ root -l MC_exercise.C
root [0]
Processing MC_exercise.C...
========== 蒙特卡洛模拟结果 ==========
总生成事例数: 100000
解析接受度 A: 0.728
MC 估计接受度 A: 0.72777 ± 0.00140755
MC 总有效效率 (A·ε²): 0.5895 ± 0.0015556
解析总效率: 0.58968
Info in <TCanvas::Print>: file cos_theta_distribution.png has been created
root [1]
```

更简化的演示——直接使用 `TF1::GetRandom()`：

```
TF1 *fpdf = new TF1("fpdf", "3.0/8.0*(1+x*x)", -1, 1);
double x = fpdf->GetRandom(); // 内部自动采用数值方法生成
```

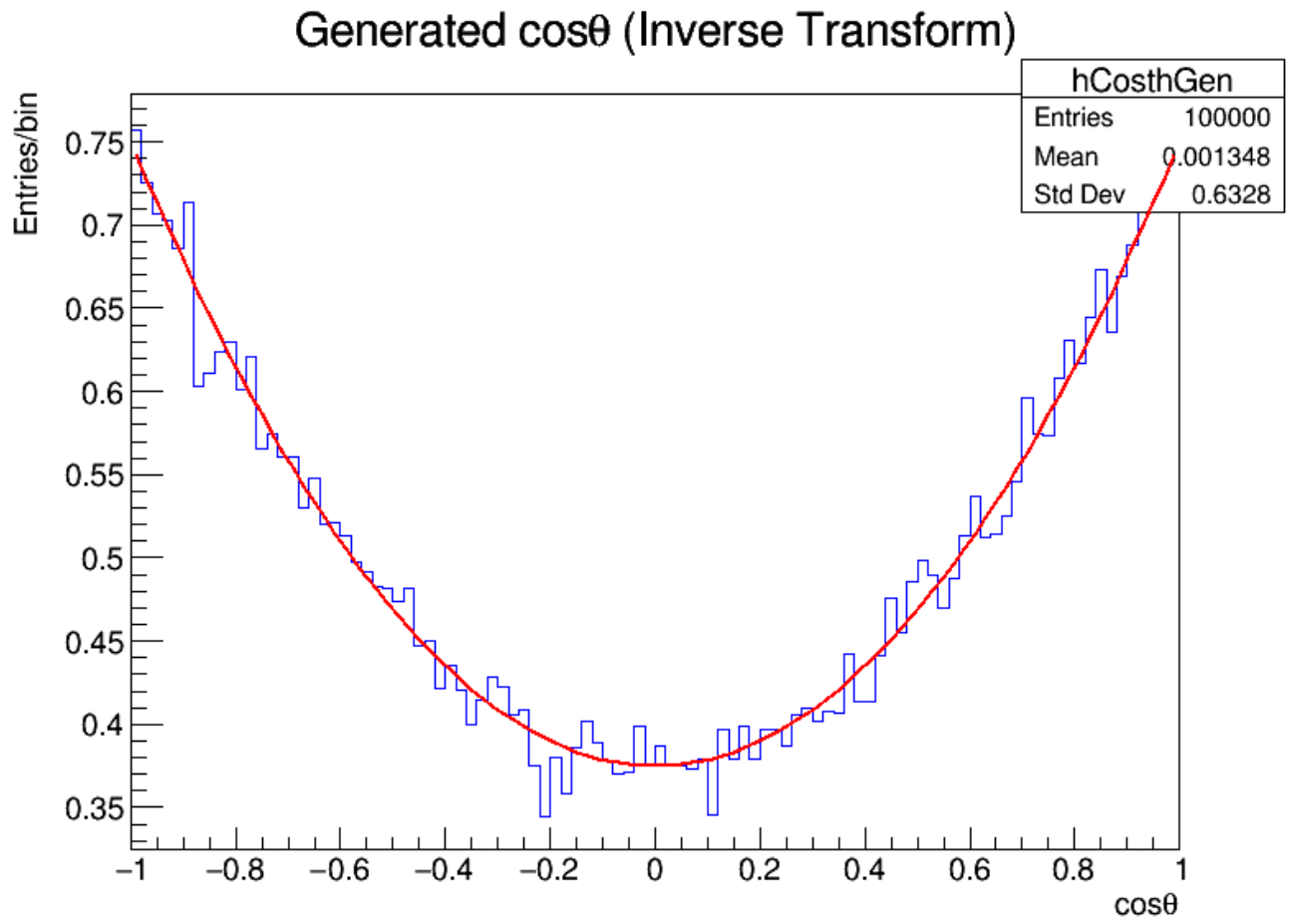
这种方法在 ROOT 中极为方便，其内部原理往往是数值化的逆变换或自适应接受-拒绝法。

5.5 两种生成方法的对比：逆变换法 vs 接受-拒绝法

为了加深理解，我们还可通过数值求解逆函数来直接生成，并与接受-拒绝法比较。以下代码用牛顿法解 $F(x) = u$，展示两者结果完全一致。

```
// MC_inverse.C
// 使用逆变换法（牛顿迭代求根）生成 e+e- -> mu+mu- 事例，并计算接受度
#include "TRandom3.h"
#include "TH1D.h"
#include "TCanvas.h"
#include "TF1.h"
#include "TMath.h"
#include <iostream>

void MC_inverse(int N = 100000) {
    // 初始化随机数生成器
    TRandom3 rand(0); // 0 表示使用机器时间种子

    // 目标概率密度 f(x) 及其累积分布函数 F(x)
    auto f = [](double x) { return 3.0/8.0 * (1 + x*x); };
    auto F = [](double x) { return 3.0/8.0 * (x + x*x*x/3.0 + 4.0/3.0); };
    // 注意：F(-1)=0，F(1)=1

    // 直方图
    TH1D *hCosthGen = new TH1D("hCosthGen", "Generated cos#theta (Inverse
Transform);cos#theta;Entries/bin", 100, -1, 1);
    TH1D *hCosthAcc = new TH1D("hCosthAcc", "Geometrically accepted
cos#theta;cos#theta;Entries/bin", 100, -1, 1);

    int nGenerated = 0;
    int nAccepted  = 0;
    int nDetected  = 0;

    double eff_muon = 0.9; // 单缪子重建效率

    // 牛顿迭代参数
    const int maxIter = 20;
    const double tol = 1e-10;

    while (nGenerated < N) {
        // ---- 步骤1: 逆变换法生成 cosθ ----
        double u = rand.Rndm();          // U[0,1)
        double x = 2.0 * u - 1.0;        // 初始猜测（线性近似）
        // 牛顿法解方程 F(x) - u = 0
        for (int iter = 0; iter < maxIter; ++iter) {
            double fx = F(x) - u;
            if (TMath::Abs(fx) < tol) break;
            double fpx = f(x);           // F'(x) = f(x)
            x -= fx / fpx;
        }
        // 保证 x 在 [-1,1] 内（极小数值误差修正）
        if (x < -1.0) x = -1.0;
        if (x >  1.0) x =  1.0;
```

```
            hCosthGen->Fill(x);
            nGenerated++;

            // ---- 步骤2: 几何接受度判断 ----
            if (TMath::Abs(x) < 0.8) {
                nAccepted++;
                hCosthAcc->Fill(x);

                // ---- 步骤3: 模拟重建效率 ----
                bool reco_mu_minus = (rand.Rndm() < eff_muon);
                bool reco_mu_plus  = (rand.Rndm() < eff_muon);
                if (reco_mu_minus && reco_mu_plus) {
                    nDetected++;
                }
            }
        }

        // ---- 结果输出 ----
        double A_est = double(nAccepted) / N;
        double A_err = TMath::Sqrt(A_est * (1 - A_est) / N);
        double eff_total_est = double(nDetected) / N;
        double eff_total_err = TMath::Sqrt(eff_total_est * (1 - eff_total_est) /
N);
        double A_analytic = 0.728;

        std::cout << "========== 逆变换法蒙特卡洛模拟结果 ==========" << std::endl;
        std::cout << "总生成事例数: " << N << std::endl;
        std::cout << "解析接受度 A: " << A_analytic << std::endl;
        std::cout << "MC 估计接受度 A: " << A_est << " ± " << A_err << std::endl;
        std::cout << "MC 总有效效率 (A·ε²): " << eff_total_est << " ± " <<
eff_total_err << std::endl;
        std::cout << "解析总效率: " << A_analytic * eff_muon * eff_muon <<
std::endl;

        // ---- 绘制分布验证 ----
        TCanvas *c1 = new TCanvas("c1", "Inverse Transform Method", 800, 600);
        hCosthGen->SetLineColor(kBlue);
        hCosthGen->Scale(1.0 / hCosthGen->Integral("width")); // 归一化到概率密度
        hCosthGen->Draw("hist");

        TF1 *f_theory = new TF1("f_theory", "3.0/8.0*(1+x*x)", -1, 1);
        f_theory->SetLineColor(kRed);
        f_theory->SetLineWidth(2);
        f_theory->Draw("same");

        c1->SaveAs("inverse_transform_distribution.png");
}
```




```
hyp@fedora:~/Download/workarea — root -l MC_inverse.C
~/Download/workarea

bash: /home/hyp/root_install/bin/thisroot.sh: No such file or directory
hyp@fedora:~/Download/file-transfer$ cd /home/hyp/Download/workarea/
hyp@fedora:~/Download/workarea$ root -l MC_inverse.C
root [0]
Processing MC_inverse.C...
========== 逆变换法蒙特卡洛模拟结果 ==========
总生成事例数: 100000
解析接受度 A: 0.728
MC 估计接受度 A: 0.72593 ± 0.00141052
MC 总有效效率 (A·ε²): 0.58787 ± 0.00155653
解析总效率: 0.58968
Info in <TCanvas::Print>: file inverse_transform_distribution.png has been created
root [1]
```

两个直方图形状完全一致，且与理论曲线 $f(x)$ 吻合。这说明：无论是通过确定性数值求逆，还是通过随机筛选，都能正确生成目标分布。在高能物理的真实复杂环境中，由于截面函数常为多维且无解析逆，接受-拒绝法及其变体成为通用标准。

---

6. 总结

蒙特卡洛模拟是高能物理实验的“虚拟实验室”。它始于从均匀随机数生成任意分布的数学方法（逆变换法、接受-拒绝法），进而依据量子场论给出的微分散射截面生成物理事例，经过探测器的全模拟，最终产出可与真实数据比对的分析样本。探测器的接受度作为其中最基础的概念，通过粒子层面的基准选择将几何覆盖损失量化为一个可精确计算的比例 A，与探测器效率 ϵ 一起构成连接观测计数与理论截面的桥梁。上述习题从 $1+\cos^2\theta$ 分布出发，完整演示了接

受-拒绝法的数学构造、接受度的解析与模拟计算，以及效率因子的组合，浓缩了现代高能物理中蒙特卡洛方法的核心思想与工作流程。

AMS 重核分析筛选条件详解

> 对 AMS 实验中重宇宙线核素（$Z \geq 9$）分析的径迹筛选和电荷筛选条件进行逐条详细解释，阐明每条筛选的物理动机、压制的背景类型以及筛选阈值取值的物理依据。

---

0. 总览：筛选的层级逻辑

AMS 在 13.5 年运行中采集了约 2.4×10^{11} 个宇宙线事例，而最终用于重核通量分析的事例数仅为 10^5~10^6 量级。筛选链条的逻辑是从粗到细、逐级收紧：

```
第0级：触发（Trigger）
  → 必须满足基本触发条件（TOF 符合 + 径迹探测器击中）
第1级：径迹存在性与质量（Track Quality）
  → 必须有一根在探测器内被足够多探测层确认的"好"径迹
第2级：核素身份验证（Charge Identification）
  → 径迹的电荷必须由多个独立子探测器一致确认为目标 Z 值
第3级：残余背景扣除（Residual Background Subtraction）
  → 即使通过上述筛选，仍有核相互作用和电荷混淆残余——用独立方法估计并扣除
```

下面逐条解释第 1 级和第 2 级中每条筛选条件的物理含义。

第一部分：径迹质量筛选 (Track Quality Selection)

> 算法版本：V6 alignment + GBL (General Broken Lines) track fitting

1.1 `InnerNHit >= 5`

含义：粒子在内层径迹探测器（Inner Tracker，共 9 层硅微条，编号 L1~L9）中至少留下 5 个击中点（hit）。

物理动机——为什么是 5?

GBL 径迹拟合算法需要估计 5 个螺旋线参数（曲率 κ、方位角 ϕ_0、极角 θ、横向冲击参数 d_0、纵向冲击参数 z_0）。每个击中点提供 1~2 个测量坐标（取决于该层是单面条还是双面条），因此：

- **至少 5 个测量坐标**才能完全约束 5 个参数（自由度 ≥ 0）
- 若 InnerNHit = 4，则自由度 = -1，参数完全无法被独立约束——解退化
- 若 InnerNHit = 5，则恰好饱和（零自由度），但没有任何冗余来检验拟合质量，也无法识别野值

实际中更倾向于 > 5 个击中点，因为：

- 额外自由度允许计算 χ^2 来评估拟合质量
- 冗余测量可以检测并剔除某些坏击中点（如大残差 → 可能来自 δ 射线或噪声）
- 粒子的多次库仑散射使实际径迹偏离螺旋线——更多测量点可以平均掉散射效应

压制的背景：

- 假径迹（随机噪声击中被误关联为径迹）：如果只有 3-4 个击中点，随机噪声点恰好排成近似螺旋线的概率虽然小，但乘上 2.4×10^{11} 的总事例数后会产生可观的假径迹数量。要求 ≥ 5 个击中点使假径迹概率急剧下降。
- 低能粒子过早停止：低能（< 1 GeV/n）的核素可能无法穿透所有层，仅在内层沉积电荷——这类事件径迹太短，刚度测量不可靠。

1.2 跨层逻辑 `L2 && (L3|L4) && (L5|L6) && (L7|L8)`

含义：

- `L2`： 径迹**必须**在 L2 层有击中
- `(L3|L4)`： L3 或 L4 中至少一层有击中
- `(L5|L6)`： L5 或 L6 中至少一层有击中
- `(L7|L8)`： L7 或 L8 中至少一层有击中

即径迹必须跨越从 L2 到 L7/L8 的完整探测空间。

物理动机——为什么必须跨越多层?

回顾第七次作业中推导的动量分辨率公式：

$$\frac{\sigma(p_T)}{p_T} \propto \frac{\sigma\, p_T}{BL^2}$$

其中 L 是径迹在横向平面的有效测量弧长。对于 AMS 的永磁体谱仪，径迹测量的有效臂长取决于第一层和最后一层击中之间的跨度。相邻层的击中提供的臂长太短，对曲率的约束极弱。

为什么是成对 (L3|L4)、(L5|L6)、(L7|L8)?

AMS 的内层径迹探测器结构是：部分层可能因为读出电子学通道失效、传感器死区、或粒子恰好穿过层间缝隙而缺失击中。采用成对逻辑（同一径向区域的两层中任一层有击中即可）提供了必要的**冗余度**——如果单层要求的效率是 95%，则成对冗余下的效率是 $1-(1-0.95)^2 = 99.75\%$。这在不牺牲径迹质量的前提下显著提高了事例存活率。

为什么 L1 不在成对逻辑中?

L1 的几何位置特殊——它是探测器的最内层，在 L1 与 L2 之间有 TRD 和上层 TOF 等相当厚度的材料。许多粒子在穿过这些材料时发生核相互作用，在 L1 被测量后就在 L2 之前碎裂

了。因此：

- 若要 L1 有击中（做全跨度分析）→ 使用单独的 `L1XY` 条件
- 若不需要 L1（仅用内层 L2-L8）→ 仍可通过 `InnerNHit>=5 && L2 && ...` 的基本要求

压制的背景：

- 保证了动量测量的刚度 L。臂长不足的径迹刚度误差极大，重建出的动量可能严重偏离真实值——这种径迹如果被保留，会在 Unfolding 的迁移矩阵中产生巨大的非对角元，使反演无法进行。

1.3 `L1XY`

含义：粒子必须在 L1 层同时有 X 和 Y 两个投影方向的击中（L1 是双面硅微条，分别测量 x 和 y 坐标）。

物理动机：

L1 是 AMS 径迹探测器的最内层，位于所有其他探测器材料之前。经过 L1 的粒子尚未与 TRD、上层 TOF 等大量材料发生相互作用。因此 L1 具有独特的物理价值：

1. **最干净的电荷测量**：粒子在击中 L1 时尚未被核相互作用改变其核素身份。L1 的电荷测量 (q_{l1}) 最接近粒子的真实电荷。
2. **相互作用标记**：如果 L1 测到的电荷 q_{l1} 与内层测到的 q_{inn} 不一致，基本可以断定粒子在 L1 与 L2 之间发生了核相互作用——这正是最重要的背景来源之一。
3. **完整径迹重建**：L1 + 内层 (L2-L8) 的组合提供了从最内层到外层的完整径迹，使得冲击参数（impact parameter）和入射角的测量精度最高。

为何要求 XY 双方向?

AMS 的硅微条层分 X 面和 Y 面（正反两面微条方向互相垂直），分别测量粒子在弯曲平面（x-z）和非弯曲平面（y-z）的坐标。刚度测量主要依赖弯曲平面的信息，但 y 向信息是必需的：

- 确认击中确实来自同一个粒子（XY 坐标在 3D 空间中确定一个点的位置）
- 径迹的空间角度 θ 和 ϕ 需要两个方向的投影

1.4 `L9XY`（全跨度 FS 分析专用）

含义：粒子必须在最外层 L9 也有 XY 双方向击中。

物理动机：

第七次作业中我们已经知道：动量分辨率 $\sigma(p_T)/p_T \propto 1/L^2$。加入 L9 击中使径迹臂长从 L2~L8（约 0.5 m）扩展到 L1~L9（约 1 m），臂长翻倍意味着**动量分辨率改善为原来的** $1/4$。

全跨度 (Full Span, FS) 分析用于需要最高刚度精度的场合，如：

- 高刚度区间（> 100 GV），此时内层径迹在探测器尺度内弯曲极小，必须有最大臂长才能提取可靠的 sagitta
- 对刚度分辨率要求最严苛的核素（如 Ca，因其通量测量精度是论文的核心结果之一）

代价：`L9XY` 要求显著降低了事例数（L9 覆盖的立体角比内层小），因此仅在统计量充裕且分辨率要求高的分析中使用。对 P、Cl、K 的标准分析与对 Ar、Ca 的 FS 分析会采用不同的径迹选择策略。

1.5 χ^2 相关筛选

1.5.1 `innerNormChisqY < 10`

含义：Y 方向（弯曲平面，r-ϕ 平面）的归一化 χ^2（即 χ^2/ndf，每自由度的 χ^2）必须小于 10。

物理动机：

在 GBL 径迹拟合中，χ^2 衡量的是各测量点的实际位置与拟合螺旋线之间的偏离程度：

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i^{\mathrm{meas}} - y_i^{\mathrm{fit}})^2}{\sigma_i^2}$$

其中 σ_i 包含了探测器位置分辨率和多次库仑散射的贡献。

归一化 χ^2 很大的物理原因：

1. **核相互作用**：粒子在探测器内与材料中的原子核碰撞，方向发生突变，后续击中点系统性地偏离原螺旋线。这是最重要的——大 χ^2 是核相互作用的标志。
2. **δ 射线**：粒子电离产生的二次电子（δ 电子）在相邻微条中沉积额外电荷，使击中点重建位置偏离真实位置。
3. **多次大角度库仑散射**：虽然 GBL 算法已经对多次散射做了处理（将其纳入过程噪声），但单次大角度散射（Mott 散射尾部）仍然可能导致个别测量点出现大残差。

为什么阈值取 10？

对于典型的 6~8 个自由度的拟合，$\chi^2/\mathrm{ndf} \sim 1$ 代表好的拟合。阈值 10 不是严格的理论界限，而是通过数据驱动的优化确定的：

- 太低（如 < 3）→ 剔除过多真实径迹（多次散射本身就会增大 χ^2）
- 太高（如 > 50）→ 保留太多核相互作用事件作为背景

该值通常通过检查 χ^2 分布的长尾部，在信号效率与背景纯净度之间取得平衡。

1.5.2 L1Inner 一致性 χ^2（L1Inner 分析专用）

`L1InnerNormChisqY < 10 && L1InnerChisqY < 10`

含义：L1 的击中点与内层 (L2-L8) 单独拟合得到的径迹在 L1 层预测位置之间的偏离不能太大。

物理动机：

单独用 L2-L8 做一次径迹拟合（内层迹），然后外推到 L1 的位置，看 L1 的实际击中点落在外推预测的什么位置。差值 Δy 对应的 χ^2 如果很大（> 10），说明：

- L1 的击中点和内层径迹不"兼容"
- 可能原因：粒子在 L1 和 L2 之间发生了核相互作用，导致径迹方向的突变

虽然内层径迹仍是碎裂产物（与目标核素相同），但径迹方向在核反应后发生了偏转——L1 之前的径迹和 L1 之后的径迹方向不完全一致。这虽然不是必须剔除的理由（碎片本身仍是我们想测量的），但较大的不一致提示该事件可能带有额外的系统不确定性，在要求高纯净度的分析（L1Inner 分析）中会被剔除。

1.5.3 L1InnerL9 全跨度一致性 χ^2（FS 分析专用）

`L1InnerL9NormChisqY < 10 && (L1InnerL9ChisqY - InnerChisqY)/2 < 10`

含义：将 L1、L9 两层加入与 L2-L8 的联合拟合后，χ^2 的增量不能过大：

$$\frac{\chi^2_{\text{L1+L9+Inner}} - \chi^2_{\text{Inner}}}{2} < 10$$

物理动机：

内层 (L2-L8) 的拟合提供了径迹的基础 χ^2。加入 L1 和 L9 两个额外击中点后，"新增的坏拟合程度"（$\Delta\chi^2/\Delta\text{ndf}$）应该合理。

分母用 2 是因为 L1 和 L9 各贡献一个弯曲平面 (y) 的测量值，合起来增加了 2 个自由度。

$\Delta\chi^2$ 过大的可能原因：

1. **L1 上方核相互作用**：粒子在到达 L1 之前（碳纤维支撑结构）就发生了核反应。L1 测到的是原始重核，而 L2-L9 测到的是碎裂产物。L1 的击中如果加入拟合（假设仍是同一个粒子），χ^2 会急剧增大。
2. **L9 处于探测器边界**：L9 在探测器边缘，击中位置重建误差大，或粒子在 L8 和 L9 之间的支撑材料中发生了散射。

这个筛选条件在全跨度分析中尤其重要——因为 FS 分析对动量精度的要求高，任何包含可疑测量点（L1 或 L9 的残差异常大）的径迹如果被保留，其刚度重建值可能严重偏离真实值。

第二部分：电荷筛选 (Charge Selection)

> 电荷标定基于 YJ 的校准结果；所有子探测器的电荷重建值已在数据/MC 对比中经过验证。

电荷筛选是整个分析链条中**最核心的背景压制手段**。AMS 通过多个独立的子探测器测量粒子的电荷 Z：

- **内层径迹探测器** (Inner Tracker, L2-L8)：`q_inn`
- **L1 层** (最内层硅微条)：`q_l1` （含 x 和 y 两个投影分量）
- **L9 层** (最外层硅微条，FS 分析)：`q_l9`
- **上层 TOF** (Upper Time-of-Flight)：`q_utof`
- **下层 TOF** (Lower TOF)：`q_ltof`

每个子探测器各自独立依据粒子的电离能损 $\mathrm{d}E/\mathrm{d}x \propto Z^2$ 重建电荷值。

2.1 `Z >= 9`

含义：仅分析电荷数 ≥ 9（氟 F 及以上）的重核。

物理动机：

- AMS 的重核分析系列（Na/Al → Ne/Mg/S → P/Cl/Ar/K/Ca）聚焦于 $Z \geq 9$ 的核素，原因是：轻核（He、C、N、O）丰度极高但物理上已有充分的测量，而中重核（$Z \geq 9$）的通量低、测量难度大，但包含了宇宙线起源和传播的关键信息（原初分量 vs 次级碎片分量的分离）
- 技术原因：Z 越大，$\mathrm{d}E/\mathrm{d}x \propto Z^2$ 越大，电荷峰之间的间距（以电荷单位计）越宽（因为 $\Delta(\mathrm{d}E/\mathrm{d}x) \propto 2Z\Delta Z$），电荷分辨率越好。对 $Z < 9$ 的轻核，相邻 Z 的电荷峰重叠严重，需要不同的分析策略（如利用 TOF + 刚度测量质量来区分同位素）

2.2 内层电荷 `q_inn` 筛选

```
fabs(q_inn - Z) < (Z >= 14) ? 0.5 : 0.0075 * pow(Z, 1.414) + 0.198
```

含义：内层径迹探测器重建的电荷 q_{inn} 与目标核素电荷 Z 之差的绝对值必须小于一个 Z 依赖的阈值 $\Delta(Z)$：

$$\Delta(Z) = \begin{cases} 0.5, & Z \geq 14 \\ 0.0075 \times Z^{1.414} + 0.198, & Z < 14 \end{cases}$$

阈值函数的物理依据：

1. $Z \geq 14$ **时固定 0.5**：对较重的核素（Si 及以上），内层径迹探测器的电荷分辨率 $\sigma_Z \approx 0.2$–0.3 电荷单位。0.5 的窗口对应于约 2σ，能在保留 > 95% 信号效率的同时有效压制相邻 Z 核素的污染。

2. $Z < 14$ **时随 Z 变化的阈值**：低 Z 核素电离能损较小，统计涨落相对更大，电荷分辨率较差。阈值按 $0.0075Z^{1.414} + 0.198$ 缩放：

   - $Z = 9$ (F)：$\Delta \approx 0.0075 \times 9^{1.414} + 0.198 \approx 0.37$
   - $Z = 10$ (Ne)：$\Delta \approx 0.0075 \times 10^{1.414} + 0.198 \approx 0.42$
   - $Z = 11$ (Na)：$\Delta \approx 0.0075 \times 11^{1.414} + 0.198 \approx 0.48$

   可以看到，阈值随 Z 增长逐渐接近 0.5，在 Z=14 处平滑衔接。
3. **为什么用 $Z^{1.414}$ 而不是 Z？** $1.414 \approx \sqrt{2}$。电荷分辨率由电离能损的涨落决定，满足 $\sigma_Z \propto 1/\sqrt{N_{\text{clusters}}}$，而簇团数与 $\mathrm{d}E/\mathrm{d}x \propto Z^2$ 成正比，故 $\sigma_Z \propto 1/Z$。但实际中还有电子学噪声等与 Z 无关的贡献，使得总 σ_Z 不完全按 $1/Z$ 变化。$Z^{1.414} = Z^{\sqrt{2}}$ 这个经验指数是 YJ 通过对不同核素的电荷分布峰做高斯拟合后提取的经验参数化。

为什么不同核素的阈值不同?

因为背景的风险不对称。以测量 P (Z=15) 为例：

- S (Z=16) 的通量远高于 P（S 是原初成分为主的稳定核素，P 主要是次级碎片）
- S 电荷峰的右侧尾部如果落入 P 的电荷窗口 → 污染 P 样本

如果对 P 和对 Si 使用相同的阈值 0.5，对 P 来说 S 的污染会比对 Si 来说 Al 的污染更严重（因为 S 相对 P 的丰度比远高于 Al 相对 Si 的丰度比）。但这里对所有 $Z \geq 14$ 使用了统一的 0.5 阈值，说明对于重核（$Z \geq 14$），内层电荷分辨率已经好到 0.5 窗口足够窄，相邻 Z 的污染可忽略。额外的压制由后续 L1 电荷筛选提供。

2.3 L1 电荷筛选

2.3.1 电荷范围

```
Z - 0.0585*pow(Z, 1.15) - 0.35 < q_l1 < Z + 0.0334*pow(Z, 1.15) + 0.20
```

含义：L1 层测得的电荷 q_{l1} 必须落在目标 Z 附近的一个**不对称窗口**内。

为什么上下限不对称?

充电荷 Z 附近电荷分布的形状不是对称的高斯分布——在低端有来自以下效应的长尾：

1. **核相互作用**：如果粒子在击中 L1 之前（支撑结构）发生了部分碎裂，L1 可能测到介于原始核素和碎片之间的中间电荷
2. δ **射线逃逸**：高能 δ 电子可能逃离 L1 灵敏体积，使测得的电离能损偏低 → q_{l1} 偏低
3. **电荷共享效应**：如果粒子正好击中微条边缘，电荷可能被相邻两条微条分走，每条收集的电荷都偏低

因此**下限窗口更宽**——需要更多的余量来容纳向下涨落的电荷测量。

高端尾部主要来自**朗道涨落**（少数极大能量转移的碰撞），但其向高端的涨落幅度远小于向低端的涨落（朗道分布本身是不对称的，高能尾部以 $1/\Delta E^2$ 衰减）。因此**上限窗口更窄**。

为什么使用 $Z^{1.15}$ 的标度？

电荷分布的宽度随 Z 变化。电离能损的相对涨落（FWHM/mean）随 Z 增大而减小（因为电离能损均值和绝对涨落的标度不同）。$Z^{1.15}$ 是经验拟合的结果，近似描述了绝对窗口宽度随 Z 的变化。

对各核素的具体数值（近似）：

- $Z = 15$ (P)：窗口约为 [14.1, 15.6]，即下限约 -0.9，上限约 +0.6
- $Z = 20$ (Ca)：窗口约为 [19.0, 20.7]，即下限约 -1.0，上限约 +0.7

2.3.2 X-Y 电荷一致性

```
fabs(q_l1x - q_l1y) / (q_l1x + q_l1y) < 0.2
```

含义：L1 层 X 面和 Y 面分别重建的电荷值 q_{l1x} 和 q_{l1y} 必须一致——二者的相对差异不能超过 20%。

物理动机：

L1 的双面硅微条结构意味着同一个粒子穿过 L1 时在两个投影面上各自沉积能量。对于一个正常入射的粒子：

- q_{l1x} 和 q_{l1y} 应该测量到近似相同的电荷（都正比于 $\mathrm{d}E/\mathrm{d}x$）
- 二者的差异来自：条间增益涨落、δ 射线在各方向的各向异性、电荷共享的偶然不对称

什么情况下 X-Y 电荷不一致？

1. **大角度入射粒子**：粒子以极大的倾角穿过 L1，在 X 面和 Y 面穿过的硅厚度不同 → 沉积能量的路径长度不同。不过对 AMS 垂直方向接收的宇宙线（主要在 ±40° 内），该效应通常不会造成 > 20% 的差异。
2. **δ 射线**：如果一个高能 δ 电子恰好沿 X 方向（微条长度方向）传播，可能在 X 面的多根微条上沉积电荷，导致 X 面电荷重建偏高，而 Y 面不受影响。
3. **噪声或电子学异常**：单面读出电子学的偶然噪声脉冲叠加在信号上。

20% 的阈值是一个经验选择——对于正常入射的重核，L1 X-Y 电荷一致性通常在 5%~10% 以内。超过 20% 的事例几乎可以肯定是反常事件。

2.3.3 电荷状态标志位

```
(qstatus_l1 & 0x10013D) == 0
```

含义：L1 电荷重建的状态标志位（status flag）必须指示重建质量良好。`0x10013D` 是位掩码——各比特位对应不同的异常状态：

- 如果 `qstatus_l1` 中 `0x10013D` 的任一比特被置 1，说明电荷重建过程中出现了某种问题

典型异常状态包括：

- **饱和**（saturation）：沉积能量超出 ADC 量程，测得的电荷被截断——实际电荷可能远大于测量值
- **死道**（dead channel）：粒子击中了已知为死道的微条，电荷信息不可靠
- **边界击中**（edge hit）：粒子击中了微条的最边缘（< 1 pitch 范围），电荷收集效率低
- **多粒子击中**：同一层有两个或以上的粒子同时穿过（在宇宙线中较少见，但在高 rate 条件下重要）
- **重建算法异常**：如聚类形状异常（cluster shape）、拟合失败等

这一筛选是整个链条中**最基础的质量控制**——在讨论任何物理结果之前，必须保证每层探测器的基本读出和重建是可靠的。

2.4 L9 电荷筛选（FS 全跨度分析专用）

```
Z - 0.0284*pow(Z, 1.15) - 0.17 < q_l9 < Z + 0.0585*pow(Z, 1.15) + 0.35
```

含义：与 L1 电荷筛选类似，但窗口参数有所不同。

与 L1 电荷筛选的差异：

1. **窗口更宽**：L9 的下限系数为 0.0284（vs L1 的 0.0585），上限系数为 0.0585（vs L1 的 0.0334）。总体而言 L9 的窗口比 L1 宽松——因为 L9 位于探测器最外层，粒子到达 L9 之前已经穿过了 L1-L8 的所有硅层 + TRD + TOF 等材料，多次库仑散射和可能的轻微核相互作用使电荷测量精度变差。
2. **不对称性反转**：L1 的窗口下限比上限宽（0.0585 vs 0.0334），而 L9 的上限比下限宽（0.0585 vs 0.0284）。这是因为：
   - **L1**（探测器入口）：向下涨落是主要问题（散射后 $\mathrm{d}E/\mathrm{d}x$ 偏低、核相互作用的电荷损失），下限更宽
   - **L9**（探测器出口）：粒子经过了全部材料，部分粒子可能经历了轻微的电荷改变（如从核反应中获得少量电荷？不，实际上是电离涨落在长路径累积后更对称了）；同时 L9 处的径迹角度更大（因为粒子经过了整个磁场的偏转），角度效应导致的测量偏差在两个方向上都有
3. **FS 分析的价值**：L9 电荷筛选为全跨度分析提供了额外的核素身份验证——如果一个重核在 L1 处被确认为 P (Z=15)，在 L2-L8 也被确认为 P，但在 L9 处的电荷明显偏低（比如接近 Si），则怀疑在 L8 和 L9 之间发生了电荷改变相互作用。

2.5 TOF 电荷筛选

2.5.1 上层 TOF 电荷

```
Z - 0.625 - 0.0225*(Z-9) < q_utof < Z + 1.5
```

含义：上层 TOF 测得的电荷必须在 $[Z - 0.625 - 0.0225(Z-9),\ Z + 1.5]$ 的范围内。

TOF 电荷分辨率为什么差?

TOF 用的是塑料闪烁体 + 光电倍增管 (PMT)，其电荷测量原理是闪烁光产额 → PMT 光电子数 → 脉冲高度。与硅微条的精密电离测量相比，闪烁体的光产额涨落大得多（Fano 因子 ~ 1 vs 硅的 ~ 0.1），因此 TOF 的电荷分辨率（$\sigma_Z \approx 0.5$–1.0 电荷单位）远差于径迹探测器。

为什么对 TOF 使用如此宽的窗口?

TOF 的物理角色不是提供精密电荷测量，而是提供**冗余的身份一致性检验**和**速度测量**（通过飞行时间测 $\beta = v/c$）。即使其电荷分辨率差，如果 q_{utof} 与目标 Z 差了 3 个电荷单位（比如目标是 P (Z=15) 但 TOF 给出 Z≈10），那显然不是核素电荷涨落的正常范围——应该剔除。

下限的线性递增：$-0.625 - 0.0225 \times (Z-9)$ 意味着对更重的核素，下限绝对值更大（更宽松）。这是因为：

- 重核的 TOF 信号幅度大，但相对涨落不会按同样比例缩小
- 重核在 TOF 闪烁体中更可能产生核相互作用（截面随 Z 增大），导致信号偏离

上限固定 +1.5：向上涨落主要来自朗道涨落尾部和偶然的 PMT 后脉冲叠加，与 Z 的关系较弱。

2.5.2 下层 TOF 电荷（FS 分析专用）

```
q_ltof > Z - 0.625 - 0.0225*(Z-9)
```

含义：下层 TOF 的电荷只需满足下限，没有上限约束。

物理动机：

下层 TOF 位于探测器的底部（粒子从上向下穿过 AMS），粒子到达下层 TOF 时已经经过了整个探测器，可能已经历了核相互作用。如果粒子发生了电荷改变反应（丢失了一些质子和中子），q_{ltof} 会比原始 Z 偏小。因此只设置下限可以：

- 剔除那些电荷测量明显不合理的低值（如 PMT 噪声触发的虚假 TOF 信号）
- 保留真实但电荷可能轻微偏低的核相互作用幸存者

上限不设限制，因为一个 Z=15 的粒子在下层 TOF 显示 Z=20 的唯一可能是 PMT 异常或另一个粒子偶然同时穿过（极低概率），这些情况已被其他筛选条件覆盖。

第三部分：背景类型总结

通过上述筛选条件后，仍未完全消除的背景可分为三大类：

3.1 电荷混淆背景（Charge Confusion）

机制：相邻 Z 核素电荷测量的统计涨落尾部互相重叠。

典型量级（经过上述筛选后）：
核素	电荷混淆残余
P (Z=15)	< 2%
Cl, Ar, K, Ca	< 1%

为什么 Cl/Ar/K/Ca 比 P 更好？ 因为这些核素质量更大，$\mathrm{d}E/\mathrm{d}x$ 更大，电荷峰间距更宽，相邻 Z 的电荷混淆自然更小。此外 P (Z=15) 有 Z=14 (Si) 和 Z=16 (S) 两个高丰度邻居，而某些核素如 Ar (Z=18) 的相邻核素 Cl (Z=17) 和 K (Z=19) 丰度相对较低。

3.2 核相互作用背景（Nuclear Interactions）

机制：一个更重的原初宇宙线核素（如 Fe (Z=26)、Cr (Z=24) 等）在探测器材料中与原子核碰撞，碎裂产生一个较轻的碎片（如 P、Ca 等），该碎片通过全部径迹和电荷筛选，表现为"目标核素"。

为什么这是最危险的背景？

- 碎片与真正的目标核素在物理上**无法区分**：电荷相同、质量近似、运动学也相似
- 碎片穿过的探测器层数足够多，能产生符合所有径迹质量筛选条件的"好"径迹
- 电荷筛选也不能辨别——因为碎片本身的电荷就是目标 Z

仅有的辨别手段：

1. 如果核反应发生在 L1 之上（支撑结构）→ L1 测的是原始重核的电荷，内层测的是碎片电荷 → **L1 电荷与内层电荷不一致**（但 L1 上方材料少，此类事件占比小）
2. 如果核反应发生在 L1 与 L2 之间（TRD + 上层 TOF）→ L1 电荷 = 原始重核，L2-L8 电荷 = 碎片 → **需用 L1 电荷模板拟合来估计**
3. 如果核反应发生在 L2 之后 → 无法通过单个事件的任何直接测量辨别 → **只能通过 MC 模拟统计估计**

扣除方法（对应 PRL 论文中的讨论）：

反应位置	估计方法
L1 上方 (支撑结构)	MC 模拟（Geant4），用 AMS 实测截面数据验证
L1 与 L2 之间 (TRD+TOF)	**L1 电荷模板拟合法**——用纯净的 Si~Ti 样本在 L2 层提取电荷响应函数，然后在 L1 层拟合数据中的电荷分布，残差即为本底
L2 以下各层之间	MC 模拟

3.3 假径迹背景（Fake Tracks）

机制：噪声击中 + 不相关的真实击中 + 低能粒子击中被径迹寻找算法（track finder）错误地组合成一根径迹。

为什么经过上述筛选后基本可以忽略？

- `InnerNHit >= 5` + 多层要求 + $\chi^2 < 10$ 这一组合对假径迹的压制效率极高
- 假径迹几乎不可能同时满足"5 层以上有击中"和"拟合 χ^2 合理"两个条件
- 即使在 2.4×10^{11} 的事例总数下，假径迹的绝对数也极小

可能的残余：极少数情况下，两根不相关的真实径迹的部分击中点可能被混淆——但这在 AMS 的低事例率环境中概率极低（AMS 在 ISS 上的平均触发率约数百 Hz，同时有两个粒子穿过的概率 < 0.1%）。

附录：筛选条件的物理解释速查表

筛选条件	一句话解释	主要压制的背景
`InnerNHit >= 5`	至少 5 层击中才能可靠拟合 5 个螺旋线参数	假径迹、径迹过短
`L2 && (L3|L4) && (L5|L6) && (L7|L8)`	径迹跨越整个内层，保证刚度量程	臂长不足 → 动量误差巨大
`L1XY`	最内层双方向击中，最干净的电荷和方向测量	核相互作用标记数据缺失
`L9XY` (FS)	最外层击中，最大臂长 = 最优动量分辨率	高刚度区间分辨率不足
`innerNormChisqY < 10`	拟合质量检验——偏离螺旋线太远 = 坏事件	核相互作用、δ 射线、假径迹
L1Inner χ^2 限制	L1 和内层径迹的相容性	L1-L2 间核相互作用
L1InnerL9 $\Delta\chi^2$ 限制	加入 L1/L9 后新增的 χ^2 必须合理	L1 上方或L8-L9 间核相互作用、边界效应
`Z >= 9`	仅分析重核	—（分析范围定义）
`|q_inn - Z| < Δ(Z)`	内层电荷匹配目标 Z	相邻 Z 核素的电荷混淆
q_{l1} 窗口 (不对称)	L1 电荷在下宽上窄的窗口内	电荷向下涨落（核反应、δ 逃逸）和向上涨落（朗道尾部）
L1 X-Y 一致性 < 20%	q_{l1x} 和 q_{l1y} 必须一致	大角度入射、δ 射线偏斜、电子学异常
L1 质量标志位	排除饱和、死道、边界击中	电荷重建算法失败

筛选条件	一句话解释	主要压制的背景
q_{l9} 窗口 (FS)	L9 电荷验证——到达出口时仍是同一核素	中途核相互作用改变电荷
TOF 电荷窗口	冗余电荷验证（分辨率差但仍有辨别力）	显著电荷不一致的大错误

Unfolding——探测器刚度迁移修正完整理论

> 从物理问题出发，系统阐述刚度（动量）迁移修正的数学框架：响应矩阵的构建、反演的病态性、正则化策略、主流 Unfolding 方法及其在高能物理中的应用。

---

1. 问题起源：为什么需要 Unfolding?

1.1 物理图景

在径迹探测器中，带电粒子的**真实刚度**（rigidity $R_{\text{true}} = p/Z$，即动量除以电荷数）是我们想要知道的物理量。然而探测器测量给出的只能是**重建刚度** R_{meas}，二者之间存在差异，原因是：

- **有限位置分辨率**：硅微条 pitch 决定了击中点位置精度，位置误差经 sagitta 关系 $\sigma(p_T)/p_T \propto \sigma p_T/(BL^2)$ 转化为刚度测量误差
- **多次库仑散射**：粒子穿过探测器物质时方向发生随机偏转，污染了弯曲曲率的测量
- **重建算法近似**：径迹拟合（如卡尔曼滤波）的线性化近似、磁场图的精度有限

结果就是：一个真实刚度在第 j 个 bin 的粒子，有可能被测量到第 i 个 bin——发生了**bin 间迁移**（bin migration）。

1.2 一个直观例子

假设真实刚度谱在某个 bin 有一个尖峰（如共振态质量峰），由于探测器分辨率，重建后的峰会变宽，且可能偏离真实位置。若直接对测量谱除以 bin 宽度并声称这就是"真实谱"，就犯了两个错误：

1. **分辨率展宽**：峰被人为加宽
2. **效率损失**：总有部分粒子因几何接受度或重建失败而未被记录

Unfolding 的任务就是：**从测量谱 $g(R_{\text{meas}})$ 反推出真实谱 $f(R_{\text{true}})$，同时正确传递统计误差和系统偏差。**

---

2. 数学表述

2.1 离散化与响应矩阵

将真实刚度范围分为 m 个 bin，测量刚度范围分为 n 个 bin（通常 $n \geq m$）。定义：

- $\boldsymbol{f} = (f_1, f_2, \ldots, f_m)^\top$：真实能谱，$f_j$ 是真实刚度落在第 j 个 bin 的事例数
- $\boldsymbol{g} = (g_1, g_2, \ldots, g_n)^\top$：测量能谱，$g_i$ 是重建刚度落在第 i 个 bin 的事例数

二者通过**响应矩阵**（response matrix）A_{ij} 联系：

$$\boxed{g_i = \sum_{j=1}^{m} A_{ij}\, f_j + \varepsilon_i}$$

或用矩阵形式：

$$\boxed{\boldsymbol{g} = \mathbf{A}\,\boldsymbol{f} + \boldsymbol{\varepsilon}}$$

其中 $\boldsymbol{\varepsilon}$ 是测量统计噪声（各 bin 事例数的泊松涨落或高斯近似）。

2.2 响应矩阵的物理含义

A_{ij} 的物理解释：**一个真实刚度在 bin j 的粒子，被探测器测量后落入 bin i 的概率**。

由定义可知响应矩阵的列归一化性质：

$$\sum_{i=1}^{n} A_{ij} = \varepsilon_j \leq 1$$

其中 ε_j 是真实 bin j 的总探测效率（包括几何接受度、重建效率、筛选效率等）。若没有任何事例丢失，则列和为 1。

响应矩阵的构建完全依赖**蒙特卡洛模拟**：

1. 用事例产生器（Pythia/Herwig）生成已知真实刚度的事例
2. 经过 Geant4 全模拟 + 重建算法，得到重建刚度
3. 填充二维直方图 $(R_{\text{true}}, R_{\text{meas}})$，归一化后得到 A_{ij}

2.3 "Folding"与"Unfolding"

- **Folding（正演）**：已知 $\boldsymbol{f}$ 和 $\mathbf{A}$，预测 $\boldsymbol{g} = \mathbf{A}\boldsymbol{f}$。是正问题，数学上适定（well-posed）。
- **Unfolding（反演）**：已知 $\boldsymbol{g}$ 和 $\mathbf{A}$，估计 $\boldsymbol{f}$。是反问题，数学上**病态**（ill-posed）。

前向过程是信息丢失的（多对一映射），反演过程必须从有限信息中恢复丢失的细节，这就是病态性的来源。

---

3. 直接反演为何失败

3.1 普通最小二乘解

若忽略正则化，直接用最小二乘最小化：

$$\chi^2 = (\boldsymbol{g} - \mathbf{A}\boldsymbol{f})^\top \mathbf{V}^{-1} (\boldsymbol{g} - \mathbf{A}\boldsymbol{f})$$

其中 $\mathbf{V} = \mathrm{Cov}(\boldsymbol{g})$ 是测量协方差矩阵（对角元为 g_i 的方差，若用泊松统计则为 g_i；若 bins 之间有关联则含非对角元）。

令 $\partial\chi^2/\partial\boldsymbol{f} = 0$，得正规方程：

$$\mathbf{A}^\top \mathbf{V}^{-1} \mathbf{A}\, \boldsymbol{f} = \mathbf{A}^\top \mathbf{V}^{-1} \boldsymbol{g}$$

形式解：

$$\hat{\boldsymbol{f}} = (\mathbf{A}^\top \mathbf{V}^{-1} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{V}^{-1} \boldsymbol{g}$$

3.2 病态的根源——小特征值放大噪声

对信息矩阵 $\mathbf{A}^\top \mathbf{V}^{-1} \mathbf{A}$ 做特征分解。响应矩阵相邻 bin 之间的迁移跃迁高度相关，使得矩阵具有许多**接近于零的特征值**。这些特征值对应的模式（特征向量）在数据中几乎无法被约束——它们表现为 $\boldsymbol{g}$ 中微小的统计涨落，经过 λ_k^{-1} 的反演放大后，在 $\hat{\boldsymbol{f}}$ 中产生**剧烈、高频的振荡**。

物理上：

- 若探测器分辨率 σ_R 将谱的精细结构抹平了，那么要恢复比 σ_R 更细的 bin 结构，本质上是从噪声中提取信息
- 响应矩阵的每一列在相邻 bin 之间高度相关，说明这些 bin 之间不是独立可测的——有效自由度远小于 bin 数

3.3 数值示例

假设真实谱是平滑的指数衰减 $f(R) \propto e^{-R/R_0}$，经过高斯分辨率 σ_R 的探测器测量。直接求逆会产生如下现象：

- 某些 bin 的事例数变为**负值**（非物理）
- 相邻 bin 出现剧烈的锯齿振荡（χ^2 虽小但解不合理）
- 解的统计误差远大于 bin 内的事例数

这被称为"**反演问题的病态性**"（ill-posedness of inverse problems）。1902年 Hadamard 定义了适定问题的三个条件：解存在、解唯一、解连续依赖于数据——Unfolding 问题违反了第三条。

---

4. 正则化：核心思想

4.1 偏差-方差权衡

所有 Unfolding 方法的核心都是在**偏差**和**方差**之间做权衡：

极端	偏差	方差	解的表现
无正则化（直接求逆）	最小（无偏）	巨大	剧烈振荡、负值
过强正则化	大（过平滑）	小	丢失真实结构
适中正则化	可接受	可控	**物理上合理的谱**

正则化本质上引入了对解的**先验假设**：真实谱应当是平滑的、非负的、或与某个参考谱（如 MC 生成器给出的理论谱）不应差得太远。

4.2 正则化的一般形式

引入惩罚项 $\mathcal{R}(\boldsymbol{f})$ 修改目标函数：

$$\mathcal{L}(\boldsymbol{f}) = \underbrace{(\boldsymbol{g}-\mathbf{A}\boldsymbol{f})^\top \mathbf{V}^{-1}(\boldsymbol{g}-\mathbf{A}\boldsymbol{f})}_{\text{数据拟合项（}\chi^2\text{）}} + \tau \cdot \underbrace{\mathcal{R}(\boldsymbol{f})}_{\text{正则化项}}$$

其中 τ 是**正则化强度参数**，控制平滑程度：

- $\tau \to 0$：退化为普通最小二乘（无正则化）
- $\tau \to \infty$：完全由先验信息决定（忽略数据）

正则化项的选择决定了方法的名称。

---

5. 四类主流 Unfolding 方法

5.1 一、Bin-by-bin 修正（最简单但最粗糙）

方法：对每个测量 bin 独立地乘以一个校正因子：

$$\hat{f}_i = \frac{g_i}{\varepsilon_i} \quad \text{或} \quad \hat{f}_i = C_i \cdot g_i$$

其中 ε_i（效率）和 C_i（校正因子）均由 MC 模拟确定。

优点：实现极为简单。

缺点（致命的）：

- 完全**忽略了 bin 间迁移**——假设测量 bin i 中的事例全部来自真实 bin i
- 当迁移显著时（如分辨率 σ_R 大于 bin 宽），会导致严重偏差
- 在高能物理中仅适用于分辨率远优于 bin 宽的情况（如量能器能量重建）

适用条件：$\sigma_R \ll \Delta R_{\text{bin}}$，即探测器分辨率远好于 bin 宽度。

5.2 二、SVD 展开法（Höcker–Kartvelishvili Unfolding）

> 已有独立文件详细讨论：Unfolding奇异值分解（SVD）展开。此处重点放在该方法的物理直觉和与刚度迁移的关联上。

核心思路：

1. 用 $\mathbf{V}^{-1/2}$ 将加权最小二乘化为普通最小二乘问题：$\mathbf{C} = \mathbf{V}^{-1/2}\mathbf{A}$，$\boldsymbol{y} = \mathbf{V}^{-1/2}\boldsymbol{g}$
2. 对 $\mathbf{C}$ 做奇异值分解：$\mathbf{C} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{W}^{\top}$
3. 解可写为奇异向量的线性组合：

$$\boldsymbol{f} = \sum_{k=1}^{m} \frac{\boldsymbol{u}_k^{\top}\boldsymbol{y}}{\sigma_k}\boldsymbol{w}_k$$

4. **小奇异值 σ_k 对应高频振荡模式**——这些模式携带的信息被噪声主导，应在求和时截断：

$$\boldsymbol{f}_{\text{reg}} = \sum_{k=1}^{k_{\max}} \frac{\boldsymbol{u}_k^{\top}\boldsymbol{y}}{\sigma_k}\boldsymbol{w}_k, \quad k_{\max} < m$$

正则化参数：截断阶数 $k_{\max}$。

物理意义：

- 低阶奇异向量（大 σ_k）描述谱的平滑整体形状，对噪声不敏感
- 高阶奇异向量（小 σ_k）描述 bin 级精细结构，完全被噪声淹没
- 截断本质上是**将有效 bin 数从 m 降为 $k_{\max}$**，用降低 bin 间独立性换取稳定性

在刚度迁移中的应用：

- 探测器分辨率 $\sigma_R \propto p_T^2$（随动量增大而变差），迁移矩阵在不同动量区域呈现不同的"宽度"
- SVD 能自然适应这种变化——在分辨率好的低动量区保留更多奇异值，差的高动量区自然被截断

5.3 三、贝叶斯迭代展开（D'Agostini 方法）

5.3.1 贝叶斯公式框架

由 G. D'Agostini 于 1995 年提出，基于贝叶斯定理将 Unfolding 视为一个概率推断问题。

令 C_j 表示"真实 bin 为 j"这一事件，E_i 表示"测量 bin 为 i"这一事件。

贝叶斯定理：

$$P(C_j|E_i) = \frac{P(E_i|C_j)\cdot P_0(C_j)}{\sum_k P(E_i|C_k)\cdot P_0(C_k)}$$

其中：

- $P(E_i|C_j) = A_{ij}$ — 响应矩阵（smeared 到 bin i 的概率）
- $P_0(C_j)$ — **先验分布**（初始猜测的真实谱形状）
- $P(C_j|E_i)$ — 后验概率（给定测量 i，真实值为 j 的概率）

5.3.2 展开公式

给定先验 $P_0(C_j)$，测量 g_i 个事例在 bin i，则真实 bin j 的估计为：

$$\boxed{\hat{f}_j = \frac{1}{\varepsilon_j} \sum_{i=1}^{n} g_i \cdot P(C_j|E_i)}$$

其中 $\varepsilon_j = \sum_i P(E_i|C_j)$ 是效率（可能不是 1，因为某些事例未被重建）。

5.3.3 迭代改进

关键创新：若对结果不满意（如残余偏差），可将当前估计 $\hat{f}_j$ 作为新的先验 $P_0(C_j)$，重复展开：

1. **第 0 次**：从 MC 生成器谱或平坦分布作为先验
2. **第 1 次迭代**：用贝叶斯公式得到第一次估计
3. **第 2 次迭代**：将第一次估计作为先验，再做一次展开
4. ...重复直到收敛或满足终止条件

正则化效果：迭代次数 N_{iter} 就是正则化参数。

- 迭代 0 次 → 返回先验（过强正则化）
- 迭代 1-2 次 → 通常给出合理的折中
- 迭代过多次 → 噪声被逐步放大，趋向于直接反演的病态解

5.3.4 为什么迭代次数需要控制

每次迭代，算法将残差 $g_i - \sum_j A_{ij}\hat{f}_j^{(k)}$ 重新分配。经过多轮迭代后，统计涨落被逐步放大为 bin 间振荡——这与 SVD 方法中保留过多奇异值导致振荡的机制类似。

实践中的经验法则：

- 迭代次数的选择通过**MC 闭合测试**确定：用已知真实谱的模拟数据运行 unfolding，比较不同迭代次数下的偏差和方差
- 通常 $N_{\text{iter}} \leq 4$ 能给出稳定结果

5.3.5 协方差矩阵

D'Agostini 给出了解析的协方差传播公式。对于第 k 次迭代：

$$\text{Cov}(\hat{f}_r, \hat{f}_s) = \sum_i \left(\frac{\partial \hat{f}_r}{\partial g_i}\right)\left(\frac{\partial \hat{f}_s}{\partial g_i}\right)\sigma_{g_i}^2$$

实际使用中常通过**玩具蒙特卡洛**（toy MC）来估计协方差：对测量谱做多次泊松涨落，每次重新运行 unfolding，从结果分布中直接计算协方差。这更加稳健，尤其在迭代次数较多时。

5.4 四、TUnfold（Tikhonov 正则化 + 多种约束）

5.4.1 Tikhonov 正则化

由 S. Schmitt 开发的 ROOT 工具包 `TUnfold`，将正则化项选为解的**曲率**（二阶差分）惩罚：

$$\mathcal{R}(\boldsymbol{f}) = \sum_{j=2}^{m-1} \left[(\Delta^2 f)_j\right]^2 = \sum_{j=2}^{m-1} \left(f_{j-1} - 2f_j + f_{j+1}\right)^2$$

写成矩阵形式：$\mathcal{R}(\boldsymbol{f}) = \boldsymbol{f}^\top \mathbf{L}^\top \mathbf{L} \boldsymbol{f}$，其中 $\mathbf{L}$ 是二阶差分矩阵：

$$\mathbf{L} = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix}$$

目标函数：

$$\mathcal{L}(\boldsymbol{f}) = (\boldsymbol{g} - \mathbf{A}\boldsymbol{f})^\top \mathbf{V}^{-1} (\boldsymbol{g} - \mathbf{A}\boldsymbol{f}) + \tau \cdot \boldsymbol{f}^\top \mathbf{L}^\top \mathbf{L} \boldsymbol{f}$$

5.4.2 解析解

对 $\mathcal{L}(\boldsymbol{f})$ 求导并令其为 0：

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}} = -2\mathbf{A}^\top \mathbf{V}^{-1}(\boldsymbol{g} - \mathbf{A}\boldsymbol{f}) + 2\tau \mathbf{L}^\top \mathbf{L}\boldsymbol{f} = 0$$

得到 Tikhonov 正则化解：

$$\boxed{\hat{\boldsymbol{f}}_\tau = \left(\mathbf{A}^\top \mathbf{V}^{-1} \mathbf{A} + \tau \mathbf{L}^\top \mathbf{L}\right)^{-1} \mathbf{A}^\top \mathbf{V}^{-1} \boldsymbol{g}}$$

对比无正则化解 $(\mathbf{A}^\top \mathbf{V}^{-1}\mathbf{A})^{-1}$，增加了一项 $\tau \mathbf{L}^\top \mathbf{L}$。这一项将 $\mathbf{A}^\top \mathbf{V}^{-1}\mathbf{A}$ 的零（或极小）特征值"抬升"，使得矩阵可稳定求逆。

5.4.3 与 SVD 的等价性

将 $\mathbf{C} = \mathbf{V}^{-1/2}\mathbf{A}$ 和 $\mathbf{L}$ 代入，Tikhonov 正则化等价于在 SVD 展开中采用平滑衰减因子：

$$\phi_k = \frac{\sigma_k^2}{\sigma_k^2 + \tau^2}$$

当 $\tau \to 0$，$\phi_k \to 1$（无正则化）；当 $\tau \to \infty$，$\phi_k \to 0$（返回先验）。

5.4.4 TUnfold 的额外功能

- **多维度 unfolding**：可同时展开多个变量（如刚度和角度）
- **不等式约束**：可强制 $f_j \geq 0$（非负约束），通过二次规划求解
- **面积守恒约束**：$\sum_j f_j$ 可固定为某一物理预期值
- **与真实分布形状约束**：可要求解不偏离理论模型的形状太远

5.5 方法对比总结

方法	正则化方式	正则化参数	优点	缺点
Bin-by-bin	无（忽略迁移）	—	极简单	仅适用于 $\sigma_R \ll \Delta$bin
SVD	奇异值截断	$k_{\max}$	数学清晰，直观	硬截断产生 bin 间相关性
贝叶斯迭代	迭代次数	N_{iter}	物理直觉好，非负	迭代次数需仔细调谐
TUnfold (Tikhonov)	曲率惩罚	τ	连续平滑，约束灵活	需选择惩罚矩阵形式

---

6. 正则化参数的选择策略

6.1 L-曲线法

绘制两条曲线的双对数图：

- **横轴**：残差范数 $||\boldsymbol{g} - \mathbf{A}\hat{\boldsymbol{f}}||^2$（越小越好）
- **纵轴**：解的范数 $||\mathbf{L}\hat{\boldsymbol{f}}||^2$（越小越平滑）

随正则化参数变化，曲线呈现特征性的"L"形拐角——拐点处偏差和方差达到最优平衡。

6.2 MC 闭合测试

这是高能物理中最可信的方法：

1. 用 MC 产生已知真实谱 $\boldsymbol{f}^{\text{true}}$ 的样本，分成两半
2. 一半用于构建响应矩阵 $\mathbf{A}$
3. 另一半作为"伪数据"（pseudo-data），其重建谱为 $\boldsymbol{g}^{\text{pseudo}} = \mathbf{A}\boldsymbol{f}^{\text{MC}}$
4. 对 $\boldsymbol{g}^{\text{pseudo}}$ 做 unfolding，得 $\hat{\boldsymbol{f}}$
5. 比较 $\hat{\boldsymbol{f}}$ 和 $\boldsymbol{f}^{\text{MC}}$，计算偏差平方和 + 方差的综合指标
6. 选择使该指标最小的正则化参数

这保证了所选的参数在**对真实谱无偏的前提下**达到最优统计精度。

6.3 注意事项

- 正则化参数的选择**不得基于真实数据本身**——否则引入确认偏差（confirmation bias），导致结果趋近先验
- 必须通过 MC 模拟、交叉验证或 L-曲线这类客观判据确定
- 最终结果须展示对正则化参数一定范围内变动的敏感性，作为系统误差的一部分

---

7. 误差分析与协方差传播

7.1 统计协方差

由于所有正则化方法的解都是测量值 $\boldsymbol{g}$ 的线性（或可线性化）函数，统计误差的传播可用标准公式：

$$\boxed{\mathrm{Cov}(\hat{\boldsymbol{f}}) = \mathbf{J} \cdot \mathrm{Cov}(\boldsymbol{g}) \cdot \mathbf{J}^{\top}}$$

其中 $\mathbf{J}_{jk} = \partial \hat{f}_j / \partial g_k$ 是雅可比矩阵（unfolding 矩阵）。

以 Tikhonov 正则化为例，解析形式为：

$$\mathbf{J} = \left(\mathbf{A}^{\top}\mathbf{V}^{-1}\mathbf{A} + \tau \mathbf{L}^{\top}\mathbf{L}\right)^{-1} \mathbf{A}^{\top}\mathbf{V}^{-1}$$

对贝叶斯迭代方法，$\mathbf{J}$ 无封闭解析式，需通过**玩具 MC** 估计：对 $\boldsymbol{g}$ 做 N_{toy} 次独立泊松涨落后运行 unfolding，从结果分布直接计算协方差矩阵。

7.2 正则化偏差

正则化在降低方差的同时**引入了偏差**。偏差的估算：

$$\boldsymbol{b} = \mathbb{E}[\hat{\boldsymbol{f}}] - \boldsymbol{f}^{\mathrm{true}}$$

对于线性正则化方法（Tikhonov, SVD），偏差可解析计算：

$$\boldsymbol{b} = \left[\left(\mathbf{A}^{\top}\mathbf{V}^{-1}\mathbf{A} + \tau \mathbf{L}^{\top}\mathbf{L}\right)^{-1}\mathbf{A}^{\top}\mathbf{V}^{-1}\mathbf{A} - \mathbf{I}\right]\boldsymbol{f}^{\mathrm{true}}$$

方括号中的项衡量了"正则化对真实解的扭曲"。它依赖于未知的 $\boldsymbol{f}^{\mathrm{true}}$，因此在实践中通过 MC 模拟各种可能的真实谱形状来评估偏差的范围。

7.3 总误差

总均方误差的分解：

$$\mathrm{MSE} = \underbrace{\mathrm{tr}[\mathrm{Cov}(\hat{\boldsymbol{f}})]}_{\text{统计方差}} + \underbrace{\|\boldsymbol{b}\|^2}_{\text{正则化偏差平方}}$$

这清楚地展示了偏差-方差权衡：增大正则化 → 方差 ↓ 但偏差 ↑，减小正则化 → 偏差 ↓ 但方差 ↑。最优正则化参数使 MSE 最小。

---

8. 高能物理中的完整 Unfolding 工作流

8.1 标准流程

以 CMS/ATLAS 实验中的微分截面测量为例：

```
┌──────────────────────────────────────────────────────────────┐
│ 第1步：选择基准相空间（fiducial phase space）                  │
│   在生成器层面定义"能被探测器看到的"相空间范围                  │
│   这定义了 unfolding 的目标——修正到粒子级（particle level）     │
├──────────────────────────────────────────────────────────────┤
│ 第2步：构建响应矩阵                                           │
│   MC 信号样本 → 粒子级信息 + 重建级信息                        │
│   填充 2D 直方图 (R_true, R_meas) → 归一化得 A_ij              │
├──────────────────────────────────────────────────────────────┤
│ 第3步：本底扣除                                               │
│   测量谱 = 信号 + 本底                                        │
│   从数据中减去估计的本底（来自 MC 或控制区）                    │
│   得到 g_i（仅含信号的测量谱）                                 │
├──────────────────────────────────────────────────────────────┤
│ 第4步：Unfolding                                              │
│   选取方法（SVD/贝叶斯迭代/TUnfold）                           │
│   用 MC 闭合测试确定正则化参数                                 │
│   对 g 运行 unfolding → 得到 f_hat                            │
├──────────────────────────────────────────────────────────────┤
│ 第5步：效率修正                                               │
│   除以 bin-by-bin 效率 ε_j（如果响应矩阵未包含效率）           │
│   得到最终的微分截面 dσ/dR 或 dN/dR                           │
├──────────────────────────────────────────────────────────────┤
│ 第6步：不确定性评估                                           │
│   统计误差：从协方差矩阵传播或 toy MC                          │
│   系统误差：改变正则化参数、响应矩阵、本底估计等 → 重新 unfolding │
│   展示各项误差的贡献（通常画为分级误差带）                      │
└──────────────────────────────────────────────────────────────┘
```

8.2 关键检查点

在公布任何 Unfolding 结果之前，必须展示以下验证：

1. **闭合测试**：用 MC 伪数据做 unfolding，验证结果与 MC 真实谱一致（在误差范围内）
2. **应力测试**：在 MC 中注入人为的谱形畸变（如尖峰、突变），验证 unfolding 能恢复
3. **独立性测试**：用一半 MC 样本构建响应矩阵，对另一半做 unfolding，避免自洽偏差

4. **正则化敏感性**：展示改变正则化参数（如 $k_{\max}$ 变动 ±1，或 τ 变动 ±20%）对结果的影响

---

9. 在刚度（动量）测量中的特殊考量

9.1 非高斯迁移尾部

动量重建中，多次库仑散射和硬韧致辐射会导致迁移矩阵出现远离对角线的**非高斯尾部**（fat tails）。例如：

- 电子在探测器材料中发生韧致辐射 → 能量损失，刚度被低估
- 强子发生核相互作用 → 径迹断裂或方向突变

对于这些情况：

- 响应矩阵必须在 MC 中充分建模这些效应（Geant4 全模拟至关重要）
- Unfolding 方法需要能处理远离对角线的迁移——SVD 和 TUnfold 在此表现通常优于贝叶斯方法

9.2 动量依赖的分辨率

探测器刚度分辨率随动量的变化：

- **低动量**（$p_T \lesssim 10$ GeV）：多次散射主导，$\sigma(p_T)/p_T$ 近似常数
- **中动量**（$10 \lesssim p_T \lesssim 100$ GeV）：位置误差主导，$\sigma(p_T)/p_T \propto p_T$
- **高动量**（$p_T \gtrsim 100$ GeV）：分辨率极差，迁移矩阵几乎退化

这意味着：

- 在低 bin 使用较窄的 bin 宽度是可接受的
- 在高 bin 必须增大 bin 宽度以保持每个 bin 内的对角迁移比例
- 响应矩阵各列的有效宽度在不同区域差异显著

9.3 Bin 宽度的选择

Bin 宽度的黄金法则：

$$\boxed{\Delta R_{\text{bin}} \gtrsim \sigma_R(R)}$$

即 bin 宽度不应小于该刚度处的探测器分辨率。若 bin 太窄：

- 迁移矩阵严重非对角（对角元 < 50%）
- Unfolding 所需的正则化极强，引入巨大模型依赖
- 结果对正则化参数极其敏感

实践中，bin 宽度通常在谱的低端和高端有所不同（**变宽度 binning**），以保持近似恒定的对角纯度。

10. 常用工具

工具	语言/框架	方法	特点
RooUnfold	C++/ROOT	SVD, 贝叶斯, Bin-by-bin	最广泛使用，简单易用
TUnfold	C++/ROOT	Tikhonov 正则化	支持约束和正则化扫描
PyUnfold	Python	贝叶斯迭代	适合小型分析
OmniFold	Python/TensorFlow	迭代神经 Unfolding	无 bin 离散化，处理高维

---

11. 总结

Unfolding 的本质是将一个**数学上病态的反问题**通过引入合理的**先验假设**（平滑性、非负性）转化为可求解的适定问题。核心张力在于：

- **统计噪声**要求平滑（减小方差）
- **物理真实**要求保留细节（减小偏差）

任何 Unfolding 结果的可靠性都取决于：

1. 响应矩阵的精度（MC 模拟对探测器响应的忠实程度）
2. 正则化参数选择的客观性（闭合测试而非人工调参）
3. 对正则化偏差的诚实估计

刚度迁移修正是高能物理数据分析中最关键也最容易被滥用的步骤之一。理解其数学基础，才不至于在"修数据"的过程中无意间创造了物理信号。

> **参考文献**
>

> - G. D'Agostini, *NIM A* 362 (1995) 487–498 — 贝叶斯迭代 Unfolding
> - A. Höcker & V. Kartvelishvili, *NIM A* 372 (1996) 469–481 — SVD Unfolding
> - S. Schmitt, *JINST* 7 (2012) T10003 — TUnfold: Tikhonov regularization
> - V. Blobel, *Unfolding methods in high-energy physics experiments* (DESY 84-118)
> - G. Cowan, *Statistical Data Analysis*, Oxford (1998), Chapter 11