

The event of interest

Let $X \sim p(x)$ and let $g(x)$ be a real-valued *score* or *performance function*.

Rare-event probability

$$A_\gamma = \{x : g(x) \geq \gamma\}, \quad \mathcal{I} = P_\gamma = \mathbb{P}_p(g(X) \geq \gamma).$$

$$I_\gamma(x) = \mathbf{1}\{g(x) \geq \gamma\}, \quad \mathcal{I} = \mathbb{E}_p[I_\gamma(X)] = \int I_\gamma(x)p(x) dx.$$

- γ is a prescribed, fixed failure or exceedance threshold.
- We care about the probability under **the original density** p , even when we simulate from another density.

Source: Section 4.5, Eq. (4.22), with the probability written unconditionally for clarity.

Why ordinary Monte Carlo struggles

Draw $X_1, \dots, X_N \stackrel{\text{iid}}{\sim} p$ and estimate

$$\hat{\mathcal{I}}_{\text{MC}} = \frac{1}{N} \sum_{i=1}^N I_{\gamma}(X_i).$$

Since $I_{\gamma}(X)$ is Bernoulli with probability $\mathcal{I}$,

$$\text{Var}(\hat{\mathcal{I}}_{\text{MC}}) = \frac{\mathcal{I}(1 - \mathcal{I})}{N}, \quad \frac{\text{SE}(\hat{\mathcal{I}}_{\text{MC}})}{\mathcal{I}} = \sqrt{\frac{1 - \mathcal{I}}{N\mathcal{I}}}.$$

Interpretation

When $\mathcal{I} \ll 1$, the relative error is approximately $(N\mathcal{I})^{-1/2}$: it is the **number of rare-event hits** that matters.

No hits does not mean zero probability

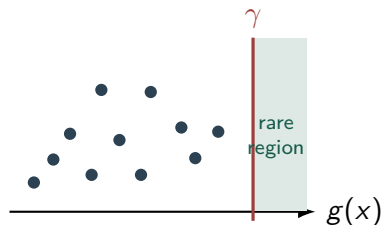
If the true probability is $\mathcal{I}$, the number of hits H satisfies

$$H \sim \text{Binomial}(N, \mathcal{I}), \quad \mathbb{P}(H = 0) = (1 - \mathcal{I})^N.$$

With $\mathcal{I} \ll 1$, this is approximately $e^{-N\mathcal{I}}$.

Key question

What if even a large simulation produces **no observations** in A_γ ?



Review: cross-entropy learns a proposal

For nonnegative f , the (formal) zero-variance IS proposal is

$$q^*(x) = \frac{f(x)p(x)}{\mathcal{I}}, \quad \mathcal{I} = \mathbb{E}_p[f(X)] > 0.$$

In the rare-event setting, $f = l_\gamma$, so

$$q_\gamma^*(x) = \frac{l_\gamma(x)p(x)}{P_\gamma} = p(x \mid g(X) \geq \gamma).$$

CE objective (from Sections 4.1–4.3)

Fit $q(x \mid \theta)$ to q_γ^* by minimising $D_{\text{KL}}(q_\gamma^* \parallel q_\theta)$, equivalently

$$\theta^* = \arg \max_{\theta \in \Omega} \mathbb{E}_p[l_\gamma(X) \log q(X \mid \theta)].$$

Why one-stage CE may fail

Suppose we draw $X_i \sim p$ and approximate the CE objective by

$$\hat{F}(\theta) = \frac{1}{n} \sum_{i=1}^n l_{\gamma}(X_i) \log q(X_i \mid \theta).$$

The rare-event difficulty

If none of the X_i satisfies $g(X_i) \geq \gamma$, then

$$l_{\gamma}(X_i) = 0 \text{ for all } i \quad \implies \quad \hat{F}(\theta) = 0 \text{ for every } \theta.$$

The objective tells us *nothing* about which proposal to choose.

Replace the single hard target by a sequence of easier targets.

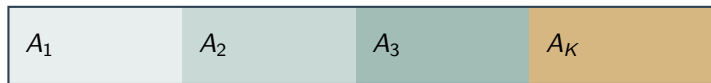
The multi-stage idea: nested events

Choose thresholds that increase towards the final level:

$$\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_K = \gamma.$$

At stage k define

$$A_k = \{x : g(x) \geq \gamma_k\}, \quad I_{\gamma_k}(x) = \mathbf{1}\{g(x) \geq \gamma_k\}.$$



Each event is contained in the preceding event; all use the same original p .

The events become rarer under p , while the proposal adapts to keep the currently selected event observable.

How do we choose the intermediate threshold?

At iteration k , sample $X_1, \dots, X_n \sim q(\cdot \mid \theta_{k-1})$ and compute

$$S_i = g(X_i), \quad i = 1, \dots, n.$$

Choose an **elite fraction** $\rho \in (0, 1)$, for example $\rho = 0.05$ or 0.10 .

Adaptive level rule

Let $\hat{Q}_{1-\rho}(S_1, \dots, S_n)$ be the sample $(1 - \rho)$ -quantile of the *scores*. Set

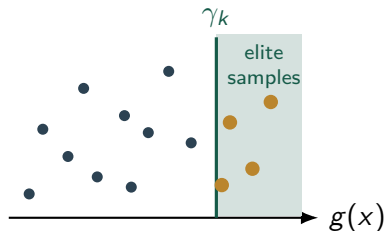
$$\gamma_k = \min\left\{\gamma, \hat{Q}_{1-\rho}(S_1, \dots, S_n)\right\}.$$

Before the target is reached, approximately ρn samples satisfy $S_i \geq \gamma_k$. The uploaded notes say “quantile of x_i ”; the quantile must be taken of $g(X_i)$.

What does the elite fraction accomplish?

For $n = 2,500$ and $\rho = 0.10$:

- About 250 sampled scores are above the temporary threshold.
- These are the **elite samples** used to refit the proposal.
- The updated proposal places more mass near the target region.
- The next threshold is computed from *new* samples.



Why the CE update needs importance weights

Samples in stage k come from q_{k-1} , not p . The ideal objective is

$$F_k(\theta) = \mathbb{E}_p[l_{\gamma_k}(X) \log q(X | \theta)].$$

By changing measure,

$$F_k(\theta) = \mathbb{E}_{q_{k-1}} \left[l_{\gamma_k}(X) \underbrace{\frac{p(X)}{q(X | \theta_{k-1})}}_{w_{k-1}(X)} \log q(X | \theta) \right].$$

Stage- k weighted CE fit

$$\theta_k = \arg \max_{\theta \in \Omega} \frac{1}{n} \sum_{i=1}^n l_{\gamma_k}(X_i) w_{k-1}(X_i) \log q(X_i | \theta).$$

This is the multi-stage version of the adaptive CE objective in the course notes, Sections 4.3 and 4.5.

Understanding the weight in the update

$$\underbrace{I_{\gamma_k}(X_i)}_{\text{keeps elite samples}} \times \underbrace{w_{k-1}(X_i)}_{\text{corrects for sampling from } q_{k-1}} \times \underbrace{\log q(X_i | \theta)}_{\text{fits the new proposal}} .$$

- The **indicator** determines which samples contribute at the current threshold.
- The **importance weight** accounts for the difference between p and the current proposal.
- The **optimisation variable** is the *new* parameter θ ; weights are computed using the *old* parameter θ_{k-1} .

Important distinction

The elite quantile is selected using the sampled $g(X_i)$ values, while the *refitting* uses weighted elite samples.

Gaussian proposals: closed-form update

Suppose q_θ is an independent d -dimensional Gaussian with parameters $\theta = (\mu, \sigma^2)$. Define the nonnegative fitting weights

$$a_i = I_{\gamma_k}(X_i) \frac{p(X_i)}{q(X_i | \theta_{k-1})}, \quad A = \sum_{i=1}^n a_i.$$

Coordinate-wise weighted maximum likelihood

$$\mu_{k,j} = \frac{\sum_{i=1}^n a_i X_{i,j}}{A}, \quad \sigma_{k,j}^2 = \frac{\sum_{i=1}^n a_i (X_{i,j} - \mu_{k,j})^2}{A}.$$

These are Section 4.4's Gaussian CE formulas with $f = I_{\gamma_k}$. In practice, use a variance floor or smoothing when the elite sample is too concentrated; the worked example below fixes $\sigma^2 = 1$.

Deriving the Gaussian mean update

For fixed variance σ^2 , write $q(x \mid \mu) = \mathcal{N}(\mu, \sigma^2)$. The terms depending on μ are

$$\hat{F}_k(\mu) = -\frac{1}{2n\sigma^2} \sum_{i=1}^n a_i (X_i - \mu)^2 + \text{constant}.$$

Differentiate:

$$\frac{d\hat{F}_k}{d\mu} = \frac{1}{n\sigma^2} \sum_{i=1}^n a_i (X_i - \mu).$$

Set the derivative to zero:

$$\sum_{i=1}^n a_i X_i - \mu \sum_{i=1}^n a_i = 0 \quad \Longrightarrow \quad \boxed{\mu_k = \frac{\sum_i a_i X_i}{\sum_i a_i}.$$

The objective is a concave quadratic whenever at least one $a_i > 0$.

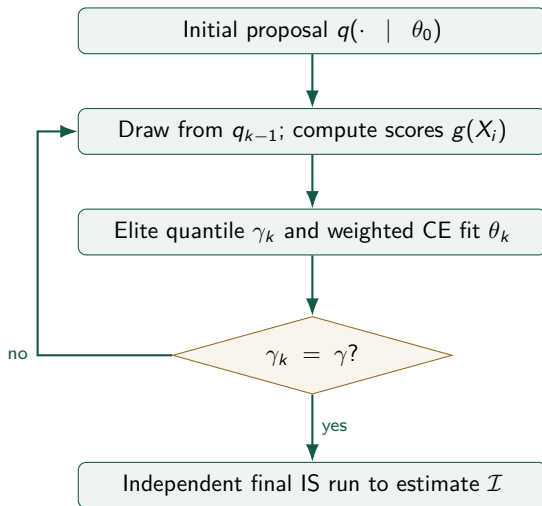
Full multi-stage CE algorithm

- ① **Initialise:** choose θ_0 (often $q_{\theta_0} = p$), elite fraction ρ , stage sample size n , and target γ .
- ② For $k = 1, 2, \dots$:
 - ① Draw $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} q(\cdot \mid \theta_{k-1})$.
 - ② Compute $S_i = g(X_i)$ and $\gamma_k = \min\{\gamma, \hat{Q}_{1-\rho}(S)\}$.
 - ③ Compute $a_i = \mathbf{1}\{S_i \geq \gamma_k\} p(X_i)/q(X_i \mid \theta_{k-1})$.
 - ④ Fit $\theta_k = \arg \max_{\theta} \sum_i a_i \log q(X_i \mid \theta)$.
- ③ **Stop adapting** when $\gamma_k = \gamma$ (or report failure if the iteration limit is reached).
- ④ **Estimate $\mathcal{I}$:** draw a fresh, independent sample from $q(\cdot \mid \theta_k)$ and apply importance sampling.

Why γ_0 does not appear

An initial level is unnecessary: the algorithm starts directly from θ_0 and constructs γ_1 .

An algorithm flow diagram



The final probability estimate

Once CE has produced the final proposal $q_K(x) = q(x \mid \theta_K)$, draw **fresh, independent** samples $Y_1, \dots, Y_M \sim q_K$.

Ordinary importance sampling, now with a learned proposal

$$\hat{\mathcal{I}}_{\text{IS}} = \frac{1}{M} \sum_{i=1}^M l_{\gamma}(Y_i) \frac{p(Y_i)}{q_K(Y_i)}.$$

Conditional on the trained proposal (assuming proper support),

$$\mathbb{E}[\hat{\mathcal{I}}_{\text{IS}} \mid q_K] = \int l_{\gamma}(y) \frac{p(y)}{q_K(y)} q_K(y) dy = \mathcal{I}.$$

CE is used to *design* q_K ; the likelihood ratio gives the probability under the original p .

Estimating uncertainty and checking the weights

For $Y_i \sim q_K$, define the contributions

$$Z_i = l_\gamma(Y_i) \frac{p(Y_i)}{q_K(Y_i)}, \quad \hat{\mathcal{I}}_{\text{IS}} = \bar{Z}.$$

The usual estimated Monte Carlo standard error is

$$\widehat{\text{SE}}(\hat{\mathcal{I}}_{\text{IS}}) = \frac{s_Z}{\sqrt{M}}, \quad s_Z^2 = \frac{1}{M-1} \sum_{i=1}^M (Z_i - \bar{Z})^2.$$

Useful diagnostics

- relative MCSE: $\widehat{\text{SE}}/\hat{\mathcal{I}}$;
- fraction of final samples in A_γ ;
- concentration of nonzero weights.

Do not confuse

A high proportion of sampled rare events *does not by itself* imply a good estimator: a few huge weights may dominate.

Worked example: a five-standard-deviation event

Original model and rare event

$$X \sim \mathcal{N}(0, 1), \quad g(x) = x, \quad \gamma = 5, \quad \mathcal{I} = \mathbb{P}(X \geq 5) = 1 - \Phi(5).$$

$$\mathcal{I} \approx 2.8665 \times 10^{-7}.$$

The event occurs on average just **once in roughly 3.49 million** independent draws from p . This Gaussian example is a teaching extension, not a numerical example from the uploaded notes.

For this example, crude Monte Carlo almost always returns zero

With $N = 20,000$ samples from $\mathcal{N}(0, 1)$:

$$\mathbb{E}[H] = N\mathcal{I} = 20,000(2.8665 \times 10^{-7}) \approx 0.00573.$$

$$\mathbb{P}(H = 0) = (1 - \mathcal{I})^{20,000} \approx 0.99428.$$

Consequence

In about 99.4% of such experiments, the crude Monte Carlo estimate is precisely zero.

We need to sample near $x = 5$ rather than repeatedly near $x = 0$.

Choose a Gaussian IS family

We restrict the proposal to a normal distribution with shiftable mean:

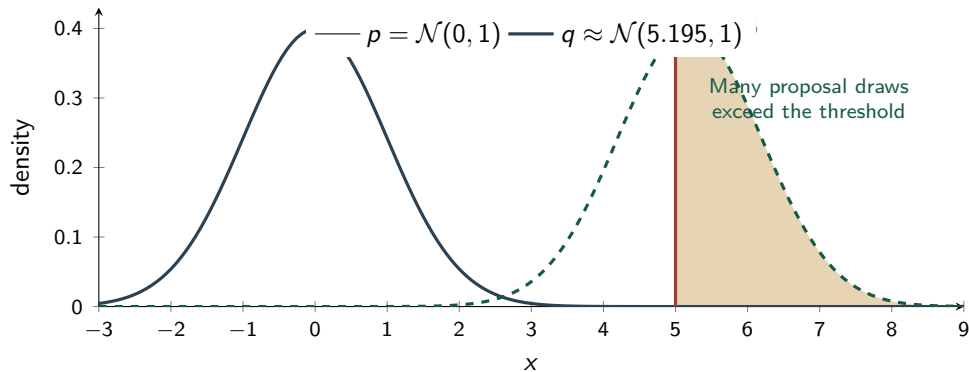
$$q(x \mid \mu) = \mathcal{N}(\mu, 1), \quad \mu_0 = 0.$$

Its importance weight is especially simple:

$$\begin{aligned} w_\mu(x) &= \frac{p(x)}{q(x \mid \mu)} = \frac{\exp(-x^2/2)}{\exp[-(x - \mu)^2/2]} \\ &= \exp\left(-\mu x + \frac{\mu^2}{2}\right). \end{aligned}$$

Here $g(x) = x$, so each stage uses the $(1 - \rho)$ -quantile of the simulated X values.

Visualising the change of measure



The area under q past 5 is large; each of those events must still be reweighted to recover the tiny probability under p .

Stage-by-stage CE: a reproducible simulation

Settings: $n = 2,500$ per stage, $\rho = 0.10$, $\mu_0 = 0$, target $\gamma = 5$, unit proposal variance.

Stage k	Current μ_{k-1}	Selected γ_k	Elite count	Updated μ_k
1	0.000	1.260	250	1.724
2	1.724	3.006	250	3.289
3	3.289	4.595	250	4.785
4	4.785	5.000	1,028	5.195

The fitted mean reached the rare-event region in four stages without an initial target-event hit. Illustrative simulation: NumPy seed 20261030; interpolation quantiles, no smoothing. Results vary by random seed.

Understand the first stage

At the start, $q_0 = p = \mathcal{N}(0, 1)$ and the temporary event is $X \geq \gamma_1$.

$$\gamma_1 \approx Q_{0.90}(\mathcal{N}(0, 1)) \approx 1.282 \quad (\text{sample value: } 1.260).$$

Since $w_0(x) = p(x)/q_0(x) = 1$, the update is the ordinary mean of the elite sample:

$$\mu_1 = \frac{\sum_{i=1}^{2500} \mathbf{1}\{X_i \geq \gamma_1\} X_i}{\sum_{i=1}^{2500} \mathbf{1}\{X_i \geq \gamma_1\}} \approx 1.724.$$

Important idea

Stage 1 learns from a *moderate tail event*; there is no need to observe the final event $X \geq 5$ yet.

Why weighted refitting matters from stage 2 onwards

At stage 2 the samples come from $q_1 = \mathcal{N}(\mu_1, 1)$, and not from p .

$$\gamma_2 \approx 3.006, \quad w_1(x) = \exp\left(-\mu_1 x + \frac{\mu_1^2}{2}\right).$$

Therefore the CE mean update is

$$\mu_2 = \frac{\sum_{i=1}^n \mathbf{1}\{X_i \geq \gamma_2\} w_1(X_i) X_i}{\sum_{i=1}^n \mathbf{1}\{X_i \geq \gamma_2\} w_1(X_i)} \approx 3.289.$$

- Elite scores determine which observations qualify.
- Weights ensure the *fitted target remains tied to the original model* p .

The final estimator in the Gaussian example

A *fresh* set of $M = 20,000$ samples is drawn from $Y_i \sim \mathcal{N}(5.1947, 1)$.

$$\hat{\mathcal{I}}_{\text{IS}} = \frac{1}{M} \sum_{i=1}^M \mathbf{1}\{Y_i \geq 5\} \exp\left(-5.1947 Y_i + \frac{5.1947^2}{2}\right).$$

Illustrative simulation

$$\hat{\mathcal{I}}_{\text{IS}} \approx 2.8350 \times 10^{-7}$$

$$\widehat{\text{SE}} \approx 4.80 \times 10^{-9}$$

Reference and precision

$$\mathcal{I} = 2.8665 \times 10^{-7}$$

$$\widehat{\text{RSE}} \approx 1.69\%.$$

Simulation in this lecture: independent final sample, seed 20261030. Actual results vary with the draw.

Why the final Gaussian shift is sensible

Recall the optimal zero-variance density for this event:

$$q^*(x) = p(x \mid X \geq 5).$$

Within the family $q(x \mid \mu) = \mathcal{N}(\mu, 1)$, the **population CE optimum** is

$$\mu^* = \mathbb{E}_p[X \mid X \geq 5].$$

For a standard Gaussian, using φ and Φ for its pdf and cdf,

$$\mathbb{E}[X \mid X \geq a] = \frac{\int_a^\infty x \varphi(x) dx}{1 - \Phi(a)} = \frac{\varphi(a)}{1 - \Phi(a)}.$$

At $a = 5$,

$$\mu^* \approx 5.1865, \quad \mu_4 \approx 5.1947.$$

The best Gaussian in this fixed-variance family is not the truncated Gaussian itself; CE fits an approximation.

What does the example show about efficiency?

	Crude MC	Multi-stage CE + IS
Original law	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$
Sampling law for final estimate	$\mathcal{N}(0, 1)$	$\mathcal{N}(5.195, 1)$
Final draws	20,000	20,000
Observed estimate	0	2.8350×10^{-7}
Useful target samples	0	11,381
Relative MCSE	not informative at 0	about 1.69%

CE also used $4 \times 2,500 = 10,000$ pilot draws to train the proposal. Counts and estimates are for a single seeded run, not guaranteed results.

Choosing ρ and the stage sample size

Elite fraction ρ

- Smaller ρ : more aggressive threshold jumps, fewer elite samples.
- Larger ρ : easier fitting, but possibly more stages.
- Typical teaching choices: 0.05 or 0.10.

Stage sample size n

- Larger n improves quantile and CE parameter estimates.
- Small ρn makes fitting noisy.
- Monitor effective weighted elite count as well as raw elite count.

No universal best settings

Choose n and ρ based on tail complexity, model dimension, weight stability and available simulation budget.

Practical difficulties and possible remedies

Problem	Sensible response
Too few effective weighted elites	Increase stage sample size or use a less aggressive ρ .
Gaussian variance shrinks nearly to zero	Add variance floors, shrinkage or a fixed-variance family.
Event has several disconnected modes	Try mixture proposals or multiple initialisations.
Proposal misses important target regions	Check support/coverage; consider defensive mixture IS.
Reusing pilot draws for final estimate	Use an independent final IS sample for the simple unbiasedness argument.
These are implementation extensions beyond the short Section 4.5 description.	

CE versus exponential tilting

Both approaches try to build a better importance-sampling law.

	Exponential tilting	Cross-entropy
Main design principle	Choose a tilt parameter (e.g. from a large-deviation or variance criterion).	Fit a parameterised proposal close to the ideal conditional event density.
What is adapted?	The tilt parameter.	Proposal parameters and intermediate event levels.
For extremely rare events	May be effective when analytic tilting structure is available.	Multi-stage levels avoid the initial zero-hit failure.

The techniques are related but not identical; both must use valid likelihood-ratio weights to estimate $\mathcal{I}$.