Empirical Distributions and Importance Sampling Foundations

Particle Filtering in Dynamical Systems

Sai Sampath Kedari
sampath@umich.edu

Contents

1 Empirical Distribution

1.1 Monte Carlo Estimation

Let X be a random variable with probability density function $f(x)$, and consider the expectation

$$\mathbb{E}[h(X)] = \int h(x)\, f(x)\, dx.$$

When this integral cannot be evaluated analytically, it is approximated using samples drawn from the distribution $f(x)$. Assume that we can generate independent samples

$$X_1, X_2, \ldots, X_N \sim \text{i.i.d. } f.$$

Monte Carlo estimation approximates the expectation by the sample mean

$$\hat{\mu}_N^{\mathrm{MC}} = \frac{1}{N} \sum_{i=1}^{N} h(X_i).$$

After the samples are realized as numerical values $x_1, \ldots, x_N$, this becomes

$$\hat{\mu}_N^{\mathrm{MC}} = \frac{1}{N} \sum_{i=1}^{N} h(x_i),$$

which is a deterministic quantity.

By the Law of Large Numbers, $\hat{\mu}_N^{\mathrm{MC}}$ converges to $\mathbb{E}[h(X)]$ as N increases.

1.2 A Shift in Perspective

The Monte Carlo estimator can be written as

$$\mathbb{E}_f[h(X)] \;\approx\; \hat{\mu}_N^{\mathrm{MC}} = \frac{1}{N} \sum_{i=1}^{N} h(x_i) = \sum_{i=1}^{N} \left(\frac{1}{N} \right) h(x_i).$$

This expression is simply a weighted sum of the values $h(x_i)$, where each sample contributes equally with weight $1/N$.

Now consider the following viewpoint. We can treat the finite set of samples

$$\{x_1, x_2, \ldots, x_N\}$$

as the support of a discrete probability distribution that assigns probability mass $1/N$ to each sample.

Under this discrete uniform distribution, the expectation of $h(X)$ is

$$\mathbb{E}[h(X)] = \sum_{i=1}^{N} \left(\frac{1}{N} \right) h(x_i),$$

which is exactly the Monte Carlo average.

Thus, the Monte Carlo estimator is not merely an average. It is the expectation of $h(X)$ under a discrete probability distribution supported on the observed samples, with equal probability mass assigned to each.

This change in viewpoint, from viewing the estimator as an average to viewing it as an expectation under a data-driven distribution, leads naturally to the concept of the empirical distribution.

1.3 The Empirical Distribution

After drawing N samples from the distribution $f(X)$ and observing the realizations

$$x_1, x_2, \ldots, x_N,$$

we now forget how these samples were generated. At this stage, the values $\{x_i\}_{i=1}^N$ are known, fixed points in the state space.

Using only these observed values, we define a probability distribution directly from the data.

1.3.1 Discrete Uniform Empirical Distribution

The simplest choice is to assign equal probability mass to each observed sample. This defines a discrete *uniform* distribution on the finite set

$$\mathcal{X}_N = \{x_1, x_2, \ldots, x_N\}.$$

We denote this distribution by $\hat{P}_N$ and define its probability mass function as

$$\hat{P}_N(X = x_i) = \frac{1}{N}, \qquad i = 1, \ldots, N.$$

All probability mass is concentrated on the observed samples, and no probability mass is assigned anywhere else. This data-driven distribution is called the *empirical distribution.*

1.3.2 Continuous Representation Using Dirac Delta Functions

The same discrete uniform distribution can be written in a continuous form using Dirac delta functions. We define

$$\hat{P}_N(X = x) = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}(x), \qquad x \in \operatorname{supp} f(X).$$

Here, the Dirac delta is defined simply as

$$\delta_{x_i}(x) = \begin{cases} 1, & x = x_i, \\ 0, & x \neq x_i. \end{cases}$$

This representation states that:

- the probability mass is $\frac{1}{N}$ at each observed point x_i,
- the probability is zero at all other points in the support of $f(X)$.

Writing the empirical distribution in this form allows us to treat both discrete and continuous distributions in a unified way. In particular, expectations can always be written as integrals, without switching between sums and integrals depending on the type of distribution.

1.3.3 Expectation Under the Empirical Distribution

Using the continuous representation, the expectation of a function $h(X)$ under the empirical distribution is

$$\mathbb{E}_{\hat{P}_N}[h(X)] = \int h(x)\, \hat{P}_N(x)\, dx.$$

Substituting the definition of $\hat{P}_N$,

$$\mathbb{E}_{\hat{P}_N}[h(X)] = \int h(x)\, \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}(x)\, dx.$$

Pulling the sum outside the integral,

$$= \frac{1}{N} \sum_{i=1}^{N} \int h(x)\, \delta_{x_i}(x)\, dx.$$

By the definition of the Dirac delta,

$$\int h(x)\, \delta_{x_i}(x)\, dx = h(x_i),$$

and therefore

$$\boxed{\mathbb{E}_{\hat{P}_N}[h(X)] = \frac{1}{N} \sum_{i=1}^{N} h(x_i).}$$

This shows that the Monte Carlo estimator is exactly the expectation of $h(X)$ under the empirical distribution.

1.3.4 Empirical Distribution as a Distributional Approximation

The empirical distribution should be viewed as a discrete approximation of the true, unknown continuous distribution $f(X)$.

Instead of working directly with $f(X)$, which is typically unavailable, we replace it with $\hat{P}_N$, a distribution supported only on sampled points. As the number of samples increases, the empirical distribution becomes a finer and finer approximation of $f(X)$.

By the Law of Large Numbers,

$$\mathbb{E}_{\hat{P}_N}[h(X)] = \frac{1}{N} \sum_{i=1}^{N} h(x_i) \;\longrightarrow\; \mathbb{E}_f[h(X)] \quad \text{as } N \to \infty.$$

Thus, although $\hat{P}_N$ is discrete, expectations computed under it converge to expectations under the true continuous distribution. In this sense, the empirical distribution converges to $f(X)$ in an expectation sense as the number of samples grows.

Remark

The key idea is simple:

> A continuous distribution can be approximated by a discrete uniform distribution supported on samples drawn from it, and expectations under the true distribution can be approximated by exact expectations under this empirical distribution.

In later sections, this same viewpoint will be extended by changing how probability mass is assigned to the sampled points.

Why This Matters (and Why We Are Doing This)

The idea of the empirical distribution is not a side detail. It is the foundation of everything that follows.

Once we accept that a distribution can be approximated by a discrete distribution supported on samples, a unifying picture emerges.

- **Monte Carlo** uses a *uniform empirical distribution*.
- **Importance Sampling** uses a *non-uniform empirical distribution*.
- **Self-Normalized Importance Sampling** uses a *normalized weighted empirical distribution*.
- **Particle Filters** repeatedly construct, propagate, and update empirical distributions over time.

Across all these methods, the structure is the same:

- the samples form the *support* of the distribution,
- the weights represent *probability mass*,
- estimation is done by computing expectations under an empirical distribution.

What changes from one method to another is not the idea of an empirical distribution, but how probability mass is assigned to the sampled points and how this distribution is updated.

In the sections that follow, we will move beyond the uniform empirical distribution and show how the target distribution $f(X)$ can be approximated using *non-uniform* empirical distributions. This shift leads directly to importance sampling, self-normalized importance sampling, and ultimately to particle filtering.

2 Importance Sampling and Self-Normalized Importance Sampling

2.1 Scope and Context

This section provides a brief recap of importance sampling (IS) and self-normalized importance sampling (SNIS). Detailed and intuition-driven explanations of these Monte Carlo methods are available in my `MonteCarlo-Statistical-Methods` repository and are assumed as background. This section is intended as a concise consolidation of those results.

Repository: `https://github.com/SaiSampathKedari/MonteCarlo-Statistical-Methods`

Our goal here is to motivate the use of importance weights, derive the importance sampling identity, identify its limitations when normalizing constants are unknown, and arrive naturally at the self-normalized importance sampling estimator. This will serve as the direct precursor to interpreting SNIS in terms of weighted empirical distributions.

2.2 Problem Setup

Let $f(x)$ denote a target probability density on a space $\mathcal{X}$. We are interested in computing expectations of the form

$$\mu := \mathbb{E}_f[h(X)] = \int h(x)\, f(x)\, dx,$$

for a measurable function $h : \mathcal{X} \to \mathbb{R}$.

In many settings, direct sampling from f is either infeasible or inefficient. Instead, we introduce a proposal density $g(x)$ from which sampling is possible.

Throughout, we assume the standard support condition:

$$f(x) > 0 \;\Rightarrow\; g(x) > 0,$$

ensuring that all regions relevant under f are accessible under g.

2.3 Importance Sampling: Weight Intuition

When sampling from g instead of f, different regions of the space are sampled with frequencies that generally differ from those under f.

- If $g(x) < f(x)$, the proposal undersamples x relative to the target.
- If $g(x) > f(x)$, the proposal oversamples x relative to the target.

Importance sampling corrects for this mismatch by assigning a weight to each sample that adjusts its contribution according to how representative it is under the target distribution. This correction factor is the *importance weight*

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

Weights do not alter the sampled values themselves; they only modulate how much each sample contributes to the final estimate.

2.4 Derivation of the Importance Sampling Identity

Starting from the target expectation,

$$\mu = \int h(x)\, f(x)\, dx,$$

we multiply and divide the integrand by $g(x)$:

$$\mu = \int h(x)\, \frac{f(x)}{g(x)}\, g(x)\, dx = \mathbb{E}_g[h(X)\, w(X)]\,, \quad w(x) = \frac{f(x)}{g(x)}.$$

This identity allows expectations under f to be expressed as expectations under g, provided importance weights are applied.

2.5 Importance Sampling Estimator

Given i.i.d. samples $X_1, \ldots, X_N \sim g$, the importance sampling estimator is

$$\hat{\mu}_N^{\mathrm{IS}} = \frac{1}{N} \sum_{i=1}^{N} h(X_i)\, w(X_i) = \frac{1}{N} \sum_{i=1}^{N} h(X_i)\, \frac{f(X_i)}{g(X_i)}.$$

This estimator is unbiased:

$$\mathbb{E}_g[\hat{\mu}_N^{\mathrm{IS}}] = \mu.$$

The variance of the estimator depends critically on how well the proposal g aligns with the regions of $\mathcal{X}$ where $h(x)f(x)$ is large. Importance sampling is particularly effective in rare-event and tail-probability estimation, where naive Monte Carlo methods suffer from extreme variance.

2.6 Limitation: Unknown Normalizing Constants

In many practical applications, the densities f and g are known only up to normalizing constants:

$$f(x) = \frac{\hat{f}(x)}{c}, \qquad g(x) = \frac{\hat{g}(x)}{b},$$

where $\hat{f}(x)$ and $\hat{g}(x)$ are computable pointwise, but the constants c and b are unknown.

The importance weight then becomes

$$w(x) = \frac{f(x)}{g(x)} = \frac{b}{c}\, \frac{\hat{f}(x)}{\hat{g}(x)},$$

which cannot be evaluated exactly due to the unknown ratio b/c. This renders the standard importance sampling estimator infeasible.

2.7 Self-Normalized Importance Sampling

We start from the target expectation

$$\mu = \mathbb{E}_f[h(X)] = \int h(x)\, f(x)\, dx.$$

When direct sampling from $f(x)$ is not possible, we introduce a proposal distribution $g(x)$ with the same support and rewrite the expectation as

$$\mu = \int h(x)\, \frac{f(x)}{g(x)}\, g(x)\, dx.$$

Assume that both densities are known only up to normalizing constants:

$$f(x) = \frac{\hat{f}(x)}{c}, \qquad g(x) = \frac{\hat{g}(x)}{b},$$

where $\hat{f}(x)$ and $\hat{g}(x)$ are computable pointwise, but c and b are unknown.

Substituting these expressions gives

$$\mu = \int h(x)\,\frac{\hat{f}(x)/c}{\hat{g}(x)/b}\,g(x)\,dx = \frac{b}{c}\int h(x)\,\frac{\hat{f}(x)}{\hat{g}(x)}\,g(x)\,dx.$$

This can be written compactly as an expectation under the proposal distribution:

$$\mu = \frac{b}{c}\,\mathbb{E}_g\left[h(X)\,\frac{\hat{f}(X)}{\hat{g}(X)}\right].$$

The ratio b/c is unknown. To eliminate it, we use the normalization condition of $f(x)$. Since $f(x)$ is a probability density,

$$1 = \int f(x)\,dx = \int \frac{\hat{f}(x)}{c}\,dx = \frac{b}{c}\int \frac{\hat{f}(x)}{\hat{g}(x)}\,g(x)\,dx = \frac{b}{c}\,\mathbb{E}_g\left[\frac{\hat{f}(X)}{\hat{g}(X)}\right].$$

Solving for the unknown ratio gives

$$\frac{b}{c} = \frac{1}{\mathbb{E}_g\left[\frac{\hat{f}(X)}{\hat{g}(X)}\right]}.$$

Substituting this back into the expression for μ yields the *self-normalized importance sampling identity*:

$$\boxed{\mu = \frac{\mathbb{E}_g\left[h(X)\,\frac{\hat{f}(X)}{\hat{g}(X)}\right]}{\mathbb{E}_g\left[\frac{\hat{f}(X)}{\hat{g}(X)}\right]}.}$$

2.8 Self-Normalized Importance Sampling Estimator

Assume that we draw N independent samples from the proposal distribution:

$$x_1, x_2, \ldots, x_N \sim g.$$

Once sampled, these are realized values and are therefore written in lowercase.

Define the unnormalized importance weights as

$$\hat{w}_i := \hat{w}(x_i) = \frac{\hat{f}(x_i)}{\hat{g}(x_i)}, \qquad i = 1, \ldots, N.$$

The self-normalized importance sampling (SNIS) estimator of $\mu = \mathbb{E}_f[h(X)]$ is then given by

$$\hat{\mu}_N^{\mathrm{SNIS}} = \frac{\sum_{i=1}^N \hat{w}_i\,h(x_i)}{\sum_{i=1}^N \hat{w}_i}.$$

It is convenient to define the normalized weights

$$\tilde{w}_i := \frac{\hat{w}_i}{\sum_{j=1}^{N} \hat{w}_j}, \qquad \sum_{i=1}^{N} \tilde{w}_i = 1.$$

With this normalization, the estimator can be written compactly as

$$\boxed{\hat{\mu}_N^{\mathrm{SNIS}} = \sum_{i=1}^{N} \tilde{w}_i \, h(x_i).}$$

Unlike standard importance sampling, the SNIS estimator is generally biased due to the ratio of random sums. However, it is consistent under mild regularity conditions.

Transition

At this point, the estimator is expressed as a weighted sum of function evaluations, with nonnegative weights summing to one. In the next section, these normalized weights will be reinterpreted as probability mass, allowing the estimator to be viewed as an expectation under a weighted empirical distribution.

3 Self-Normalized Importance Sampling as a Discrete Approximation of the Target Distribution

3.1 Starting Point: The SNIS Estimator

From the previous section, we obtained the self-normalized importance sampling (SNIS) estimator

$$\hat{\mu}_N^{\mathrm{SNIS}} = \sum_{i=1}^{N} \tilde{w}_i \, h(x_i), \qquad \sum_{i=1}^{N} \tilde{w}_i = 1,$$

where

$$x_1, \ldots, x_N \sim g$$

are samples drawn from the proposal distribution $g(x)$, and $\tilde{w}_i$ are the normalized importance weights.

At this stage, the estimator is typically viewed as a weighted sum used to approximate $\mathbb{E}_f[h(X)]$. In this section, we take a different perspective.

> *Rather than viewing SNIS as an estimator, we ask what probability distribution it implicitly defines.*

This shift is essential for understanding particle filtering.

3.2 Working at the Level of Realized Samples

Once the samples $x_1, \ldots, x_N$ have been drawn from $g(x)$, they are no longer random variables. They are fixed, realized points in the state space.

At this point, we are given:

- a finite set of support points $\{x_1, \ldots, x_N\}$,
- a set of nonnegative weights $\{\tilde{w}_1, \ldots, \tilde{w}_N\}$ that sum to one.

No additional randomness remains. Everything that follows is a deterministic construction based on these realized values.

3.3 SNIS Induces a Non-Uniform Empirical Distribution

In Section 1, we introduced the empirical distribution as a discrete distribution supported on sampled points. Here, SNIS induces a *non-uniform* version of that same object.

We define the empirical distribution induced by SNIS as

$$\hat{P}_N^{\mathrm{SNIS}}(X = x_i) = \tilde{w}_i, \qquad i = 1, \ldots, N.$$

Equivalently, using the Dirac delta representation introduced earlier, this distribution can be written as

$$\boxed{\hat{P}_N^{\mathrm{SNIS}}(dx) = \sum_{i=1}^{N} \tilde{w}_i \, \delta_{x_i}(dx), \qquad x \in \operatorname{supp} f.}$$

This expression should be read literally:

- the support of the distribution is the set of sampled points $\{x_i\}$,
- the probability mass at each point x_i is $\tilde{w}_i$,
- the distribution assigns zero probability everywhere else.

Thus, SNIS constructs a discrete probability distribution whose support is determined by the proposal distribution and whose probability mass is determined by importance weights.

3.4 The SNIS Estimator as an Exact Expectation

Once the empirical distribution $\hat{P}_N^{\mathrm{SNIS}}$ is defined, the SNIS estimator admits an exact probabilistic interpretation.

For any function h,

$$\mathbb{E}_{\hat{P}_N^{\mathrm{SNIS}}}[h(X)] = \sum_{i=1}^{N} \tilde{w}_i \, h(x_i).$$

Comparing with the definition of $\hat{\mu}_N^{\mathrm{SNIS}}$, we obtain the exact identity

$$\boxed{\hat{\mu}_N^{\mathrm{SNIS}} = \mathbb{E}_{\hat{P}_N^{\mathrm{SNIS}}}[h(X)].}$$

There is no approximation in this step. The SNIS estimator is precisely the expectation of $h(X)$ under the empirical distribution $\hat{P}_N^{\mathrm{SNIS}}$.

3.5 What Is Being Approximated

The approximation in SNIS occurs at the distributional level.

The true target distribution $f(x)$ is unknown and typically intractable. SNIS replaces it with the discrete empirical distribution

$$f \;\approx\; \hat{P}_N^{\mathrm{SNIS}}.$$

Once this approximation is constructed, *all* expectations under f are approximated by exact expectations under $\hat{P}_N^{\mathrm{SNIS}}$:

$$\mathbb{E}_f[h(X)] \;\approx\; \mathbb{E}_{\hat{P}_N^{\mathrm{SNIS}}}[h(X)].$$

Thus, SNIS should be understood as a method for constructing a probability distribution that approximates f, not merely as a tool for computing a single expectation.

3.6 Support and the Role of the Proposal Distribution

The empirical distribution $\hat{P}_N^{\mathrm{SNIS}}$ assigns probability mass only to points that were sampled from the proposal distribution g:

$$\mathrm{supp}(\hat{P}_N^{\mathrm{SNIS}}) = \{x_1, \ldots, x_N\}.$$

As a result:

- regions not visited by the proposal receive zero probability mass,
- importance weights can only redistribute mass among sampled points,
- no weighting scheme can recover regions where no samples exist.

This limitation is fundamental and highlights the critical role of proposal design.

3.7 Particles and the Link to Particle Filtering

With this viewpoint, the meaning of a particle becomes precise:

- a particle is a realized sample location x_i,
- the particle weight $\tilde{w}_i$ is the probability mass assigned to that location,
- the set $\{(x_i, \tilde{w}_i)\}_{i=1}^N$ defines a discrete probability distribution.

Particle filtering builds directly on this construction. At each time step, a particle filter maintains, propagates, and updates an empirical distribution of exactly this form.

Conclusion

Self-normalized importance sampling constructs the empirical distribution

$$\hat{P}_N^{\mathrm{SNIS}}(dx) = \sum_{i=1}^{N} \tilde{w}_i \, \delta_{x_i}(dx),$$

which serves as a discrete approximation of the target distribution $f(x)$.

The SNIS estimator is simply the exact expectation of a function under this empirical distribution. Particle filtering extends this idea by evolving such empirical distributions sequentially over time.