Introduction to Monte Carlo and importance sampling

Sara Pérez Vieites

Aalto University
Finnish Center for Artificial Intelligence (FCAI)

CPS group meeting

8 October 2024

Index

Simple Monte Carlo or *crude* Monte Carlo

Problem:: Compute the expected value of a variable $Y = f(X)$, where

- the r.v. $X \in D \subseteq \mathbb{R}^d$ has a probability density function $p(x)$
- f is a real-valued function defined over D

$$\mu = \mathbb{E}_{p(x)}[f(X)] = \int_D f(x)p(x)dx. \tag{1}$$

If there is no analytical solution: Monte Carlo estimator

$$\widehat{\mu}_{MC} = \frac{1}{N}\sum_{i=1}^{N} f(X_i) \quad \text{where} \quad X_i \sim p(x) \text{ are i.i.d..} \tag{2}$$

Simple Monte Carlo or *crude* Monte Carlo

Problem:: Compute the expected value of a variable $Y = f(X)$, where

- the r.v. $X \in D \subseteq \mathbb{R}^d$ has a probability density function $p(x)$
- f is a real-valued function defined over D

$$\mu = \mathbb{E}_{p(x)}[f(X)] = \int_D f(x)p(x)dx. \tag{1}$$

If there is no analytical solution: Monte Carlo estimator

$$\widehat{\mu}_{MC} = \frac{1}{N}\sum_{i=1}^{N} f(X_i) \quad \text{where} \quad X_i \sim p(x) \text{ are i.i.d..} \tag{2}$$

Why does Monte Carlo work?

1. **Consistency:**

$$\lim_{N\to\infty} \widehat{\mu}_{MC} = \mathbb{E}_{p(x)}[f(X)] = \mu \quad \text{(almost surely)}.$$

- The **Law of Large Numbers (LLN)** ensures that with enough samples, the estimate becomes increasingly accurate.

2. **Unbiasedness:**

$$\mathbb{E}_{p(x)}[\widehat{\mu}_{MC}] = \mu \quad \text{for any } N.$$

- The estimator $\widehat{\mu}_{MC}$ does not systematically overestimate or underestimate μ.

Why does Monte Carlo work?

1. Consistency:

$$\lim_{N\to\infty} \widehat{\mu}_{MC} = \mathbb{E}_{p(x)}[f(X)] = \mu \quad \text{(almost surely)}.$$

- The **Law of Large Numbers (LLN)** ensures that with enough samples, the estimate becomes increasingly accurate.

2. Unbiasedness:

$$\mathbb{E}_{p(x)}[\widehat{\mu}_{MC}] = \mu \quad \text{for any } N.$$

- The estimator $\widehat{\mu}_{MC}$ does not systematically overestimate or underestimate μ.

Why does Monte Carlo work?

3. Error/Variance of the Estimator:

$$\mathrm{MSE}(\widehat{\mu}_{MC}) = \mathbb{E}_{p(x)}[(\widehat{\mu}_{MC} - \mu)^2] = \frac{\mathrm{Var}_{p(x)}[f(X)]}{N}$$

- The accuracy (reduction in variance) improves as N increases: the larger the number of samples N, the lower the variance.
- The variance depends on $\mathrm{Var}_{p(x)}[f(X)]$: a larger variance in $f(X)$ results in a slower reduction in error.

  Warning: $\mathrm{Var}_{p(x)}[f(X)]$ depends on dimension in general!

Why does Monte Carlo work?

3. Error/Variance of the Estimator:

$$\mathrm{MSE}(\widehat{\mu}_{MC}) = \mathbb{E}_{p(x)}[(\widehat{\mu}_{MC} - \mu)^2] = \frac{\mathrm{Var}_{p(x)}[f(X)]}{N}$$

- The accuracy (reduction in variance) improves as N increases: the larger the number of samples N, the lower the variance.
- The variance depends on $\mathrm{Var}_{p(x)}[f(X)]$: a larger variance in $f(X)$ results in a slower reduction in error.

  Warning: $\mathrm{Var}_{p(x)}[f(X)]$ depends on dimension in general!

Why does Monte Carlo work?

3. Error/Variance of the Estimator:

$$\mathrm{MSE}(\widehat{\mu}_{MC}) = \mathbb{E}_{p(x)}[(\widehat{\mu}_{MC} - \mu)^2] = \frac{\mathrm{Var}_{p(x)}[f(X)]}{N}$$

- The accuracy (reduction in variance) improves as N increases: the larger the number of samples N, the lower the variance.
- The variance depends on $\mathrm{Var}_{p(x)}[f(X)]$: a larger variance in $f(X)$ results in a slower reduction in error.

  Warning: $\mathrm{Var}_{p(x)}[f(X)]$ depends on dimension in general!

Advantages and Disadvantages of Monte Carlo

Advantages:

- **General applicability.** Can be applied to a wide range of problems.
- **Unbiased estimator** with convergence guarantees, based on the Law of Large Numbers (LLN).
- **Easy implementation.**

Disadvantages:

- **Computational cost:** Requires a large number of samples for accurate results.
- **Sampling issues:** Inefficient or problematic when sampling from $p(x)$ is either impossible or just inefficient.

Advantages and Disadvantages of Monte Carlo

Advantages:

- **General applicability.** Can be applied to a wide range of problems.
- **Unbiased estimator** with convergence guarantees, based on the Law of Large Numbers (LLN).
- **Easy implementation.**

Disadvantages:

- **Computational cost:** Requires a large number of samples for accurate results.
- **Sampling issues:** Inefficient or problematic when sampling from $p(x)$ is either impossible or just inefficient.

Addressing Sampling Issues: Importance Sampling

Sampling from $p(x)$:

- **Impossible:** $p(x)$ is too complex or unknown (e.g., complex high-dimensional distributions).
- **Inefficient:** Rare event estimation or distributions with high variance can lead to very few relevant samples, slowing convergence.

Solution: Importance Sampling (IS)

- IS solves this problem by sampling from a **simpler, more convenient distribution** $q(x)$, and reweighting the samples to reflect the target distribution $p(x)$.
- This allows us to focus the sampling in regions where the contributions to the integral are more significant.

Addressing Sampling Issues: Importance Sampling

Sampling from $p(x)$:

- **Impossible:** $p(x)$ is too complex or unknown (e.g., complex high-dimensional distributions).
- **Inefficient:** Rare event estimation or distributions with high variance can lead to very few relevant samples, slowing convergence.

Solution: Importance Sampling (IS)

- IS solves this problem by sampling from a **simpler, more convenient distribution** $q(x)$, and reweighting the samples to reflect the target distribution $p(x)$.
- This allows us to focus the sampling in regions where the contributions to the integral are more significant.

Importance Sampling

$$\mu = \int_D f(x)p(x)dx = \int_D f(x)\frac{p(x)}{q(x)}q(x)dx \quad (3)$$

- $q(x)$ is the *importance distribution* (helps to obtain more samples from region D),
- $p(x)$ is the *nominal distribution*,
- $\frac{p(x)}{q(x)}$ is the *importance weight*, to adjust our estimate to account for having oversampled in this region. While for simple MC we do not need to evaluate the PDF of $p(x)$, with IS we do need!

$$\widehat{\mu}_{IS} = \frac{1}{N}\sum_{i=1}^{N} f(X_i)\frac{p(X_i)}{q(X_i)} = \frac{1}{N}\sum_{i=1}^{N} f(X_i)w_i \quad (4)$$

where $w_i = \frac{p(X_i)}{q(X_i)}$ are the **importance weights**, and $X_i \sim q(x)$.

Importance Sampling

$$\mu = \int_D f(x)p(x)dx = \int_D f(x)\frac{p(x)}{q(x)}q(x)dx \tag{3}$$

- $q(x)$ is the *importance distribution* (helps to obtain more samples from region D),
- $p(x)$ is the *nominal distribution*,
- $\frac{p(x)}{q(x)}$ is the *importance weight*, to adjust our estimate to account for having oversampled in this region. While for simple MC we do not need to evaluate the PDF of $p(x)$, with IS we do need!

$$\widehat{\mu}_{IS} = \frac{1}{N}\sum_{i=1}^{N} f(X_i)\frac{p(X_i)}{q(X_i)} = \frac{1}{N}\sum_{i=1}^{N} f(X_i)w_i \tag{4}$$

where $w_i = \frac{p(X_i)}{q(X_i)}$ are the **importance weights**, and $X_i \sim q(x)$.

Importance Sampling

$$\mu = \int_D f(x)p(x)dx = \int_D f(x)\frac{p(x)}{q(x)}q(x)dx \tag{3}$$

- $q(x)$ is the *importance distribution* (helps to obtain more samples from region D),
- $p(x)$ is the *nominal distribution*,
- $\frac{p(x)}{q(x)}$ is the *importance weight*, to adjust our estimate to account for having oversampled in this region. While for simple MC we do not need to evaluate the PDF of $p(x)$, with IS we do need!

$$\widehat{\mu}_{IS} = \frac{1}{N}\sum_{i=1}^{N} f(X_i)\frac{p(X_i)}{q(X_i)} = \frac{1}{N}\sum_{i=1}^{N} f(X_i)w_i \tag{4}$$

where $w_i = \frac{p(X_i)}{q(X_i)}$ are the **importance weights**, and $X_i \sim q(x)$.

Variance of IS Estimator

$$\text{MSE}(\widehat{\mu}_{IS}) = \mathbb{E}_{q(x)}\left[(\widehat{\mu}_{IS} - \mu)^2\right] = \frac{\text{Var}_{q(x)}\left[w(X)f(X)\right]}{N}$$

where $w(X) = \frac{p(X)}{q(X)}$.

Key Points:

- The variance of IS depends on the choice of $q(x)$: a good proposal distribution $q(x)$ should closely resemble $p(x)$ in regions where $f(x)$ contributes significantly to the integral.
- If $q(x)$ is poorly chosen, the weights $w(x)$ can become large, leading to high variance.
- **Goal of IS:** To choose $q(x)$ such that

$$\text{Var}_{q(x)}\left[w(X)f(X)\right] < \text{Var}_{p(x)}[f(X)]$$

Variance of IS Estimator

$$\mathrm{MSE}(\widehat{\mu}_{IS}) = \mathbb{E}_{q(x)}\left[(\widehat{\mu}_{IS} - \mu)^2\right] = \frac{\mathrm{Var}_{q(x)}\left[w(X)f(X)\right]}{N}$$

where $w(X) = \frac{p(X)}{q(X)}$.

Key Points:

- The variance of IS depends on the choice of $q(x)$: a good proposal distribution $q(x)$ should closely resemble $p(x)$ in regions where $f(x)$ contributes significantly to the integral.
- If $q(x)$ is poorly chosen, the weights $w(x)$ can become large, leading to high variance.
- **Goal of IS:** To choose $q(x)$ such that

$$\mathrm{Var}_{q(x)}\left[w(X)f(X)\right] < \mathrm{Var}_{p(x)}[f(X)]$$

Variance of IS Estimator

$$\mathrm{MSE}(\widehat{\mu}_{IS}) = \mathbb{E}_{q(x)}\left[\left(\widehat{\mu}_{IS} - \mu\right)^2\right] = \frac{\mathrm{Var}_{q(x)}\left[w(X)f(X)\right]}{N}$$

where $w(X) = \frac{p(X)}{q(X)}$.

Key Points:

- The variance of IS depends on the choice of $q(x)$: a good proposal distribution $q(x)$ should closely resemble $p(x)$ in regions where $f(x)$ contributes significantly to the integral.
- If $q(x)$ is poorly chosen, the weights $w(x)$ can become large, leading to high variance.
- **Goal of IS:** To choose $q(x)$ such that

$$\mathrm{Var}_{q(x)}\left[w(X)f(X)\right] < \mathrm{Var}_{p(x)}\left[f(X)\right]$$

Variance of IS Estimator

$$\text{MSE}(\widehat{\mu}_{IS}) = \mathbb{E}_{q(x)}\left[(\widehat{\mu}_{IS} - \mu)^2\right] = \frac{\text{Var}_{q(x)}\left[w(X)f(X)\right]}{N}$$

where $w(X) = \frac{p(X)}{q(X)}$.

Key Points:

- The variance of IS depends on the choice of $q(x)$: a good proposal distribution $q(x)$ should closely resemble $p(x)$ in regions where $f(x)$ contributes significantly to the integral.
- If $q(x)$ is poorly chosen, the weights $w(x)$ can become large, leading to high variance.
- **Goal of IS:** To choose $q(x)$ such that

$$\text{Var}_{q(x)}\left[w(X)f(X)\right] < \text{Var}_{p(x)}[f(X)]$$

Example: Rare Event

Goal: For $p(x) = \mathcal{N}(0,1)$, compute

$$\mathbb{P}[X > 3] = \mathbb{E}_{p(x)}[\mathbb{I}(X > 3)] = \int \underbrace{\mathbb{I}(X > 3)}_{\text{Indicator function}} p(x)dx \tag{5}$$

Monte Carlo (MC):
$X_i \sim p(x) = \mathcal{N}(0,1)$

$$\widehat{\mu}_{MC} = \frac{\sum_{i=1}^{N} \mathbb{I}(X_i > 3)}{N}$$

Importance Sampling (IS):
$X_i \sim q(x) = \mathcal{N}(3,1)$

$$\widehat{\mu}_{IS} = \frac{\sum_{i=1}^{N} \mathbb{I}(X_i > 3) w_i}{N}$$

$$w_i = \frac{p(X_i)}{q(X_i)} = \frac{\mathcal{N}(X_i \mid 0,1)}{\mathcal{N}(X_i \mid 3,1)}$$

Example: Rare Event

Goal: For $p(x) = \mathcal{N}(0,1)$, compute

$$\mathbb{P}[X > 3] = \mathbb{E}_{p(x)}[\mathbb{I}(X > 3)] = \int \underbrace{\mathbb{I}(X > 3)}_{\text{Indicator function}} \; p(x)dx \tag{5}$$

Monte Carlo (MC):
$X_i \sim p(x) = \mathcal{N}(0,1)$

$$\widehat{\mu}_{MC} = \frac{\sum_{i=1}^{N} \mathbb{I}(X_i > 3)}{N}$$

Importance Sampling (IS):
$X_i \sim q(x) = \mathcal{N}(3,1)$

$$\widehat{\mu}_{IS} = \frac{\sum_{i=1}^{N} \mathbb{I}(X_i > 3) w_i}{N}$$

$$w_i = \frac{p(X_i)}{q(X_i)} = \frac{\mathcal{N}(X_i \mid 0,1)}{\mathcal{N}(X_i \mid 3,1)}$$

Example: Rare Event

Goal: For $p(x) = \mathcal{N}(0,1)$, compute

$$\mathbb{P}[X > 3] = \mathbb{E}_{p(x)}[\mathbb{I}(X > 3)] = \int \underbrace{\mathbb{I}(X > 3)}_{\text{Indicator function}} p(x)dx \qquad (5)$$

Monte Carlo (MC):
$X_i \sim p(x) = \mathcal{N}(0,1)$

$$\widehat{\mu}_{MC} = \frac{\sum_{i=1}^{N} \mathbb{I}(X_i > 3)}{N}$$

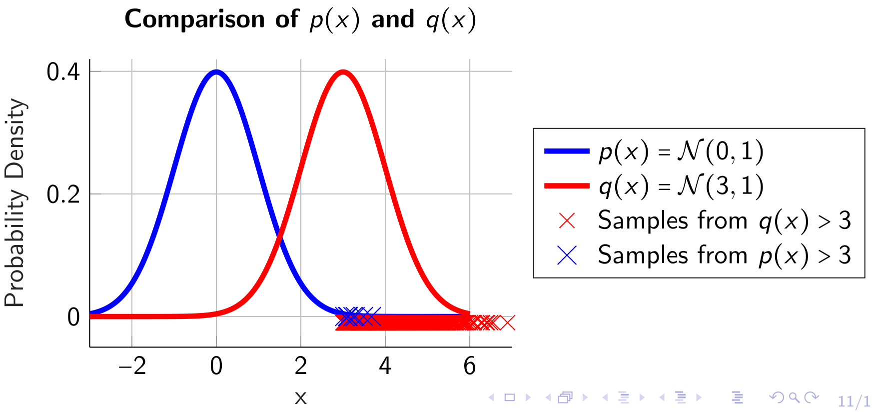
Importance Sampling (IS):
$X_i \sim q(x) = \mathcal{N}(3,1)$

$$\widehat{\mu}_{IS} = \frac{\sum_{i=1}^{N} \mathbb{I}(X_i > 3) w_i}{N}$$

$$w_i = \frac{p(X_i)}{q(X_i)} = \frac{\mathcal{N}(X_i \mid 0,1)}{\mathcal{N}(X_i \mid 3,1)}$$

Example: Rare Event

- **MC**: Samples are mostly centered around 0, making it inefficient for rare events like $X > 3$.
- **IS**: Focuses sampling in the tail region $X > 3$, reducing variance and improving efficiency.



Examples of IS in Control and RL

1. MPPI (Model Predictive Path Integral Control) [Asmar et al. 2023]:

- We sample actions $u_t^i \sim q(u)$ (e.g., Gaussian noise around nominal controls).
- We weight the sampled actions based on the cost function $e^{-\frac{1}{\lambda}S(u_t^i)}$.

2. Off-policy Evaluation (Reinforcement Learning):

- We use samples (s^i, u^i, r^i, s^{i+1}) from a different behavior policy $u^i \sim q(u) = \pi_a(u \mid s)$.
- We weight each sample with the new target policy $\pi_b(u \mid s)$, using the importance weight:

$$w^i = \frac{\pi_b(u^i \mid s^i)}{\pi_a(u^i \mid s^i)}$$

Examples of IS in Control and RL

1. **MPPI (Model Predictive Path Integral Control)** [Asmar et al. 2023]:
   - We sample actions $u_t^i \sim q(u)$ (e.g., Gaussian noise around nominal controls).
   - We weight the sampled actions based on the cost function $e^{-\frac{1}{\lambda}S(u_t^i)}$.

2. **Off-policy Evaluation** (Reinforcement Learning):
   - We use samples (s^i, u^i, r^i, s^{i+1}) from a different behavior policy $u^i \sim q(u) = \pi_a(u \mid s)$.
   - We weight each sample with the new target policy $\pi_b(u \mid s)$, using the importance weight:

$$w^i = \frac{\pi_b(u^i \mid s^i)}{\pi_a(u^i \mid s^i)}$$

Thank you!

- Book about **Monte Carlo, importance sampling**, MCMC, and QMC.

  Owen, A. B. (2013). Monte Carlo theory, methods and examples.
  Link to the FREE book (pdf).

- Asmar, D. M., Senanayake, R., Manuel, S., & Kochenderfer, M. J. (2023, May). Model predictive optimized path integral strategies. In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 3182-3188). IEEE.
- Metelli, A. M., Papini, M., Montali, N., & Restelli, M. (2020). Importance sampling techniques for policy optimization. Journal of Machine Learning Research, 21(141), 1-75.

- Next step: Bayesian inference, Kalman filters, and sequential Monte Carlo (particle filters).

  Särkkä, S., & Svensson, L. (2023). Bayesian filtering and smoothing (Vol. 17). Cambridge university press.
  Link to the FREE book (pdf).

Thank you!

- Book about **Monte Carlo, importance sampling**, MCMC, and QMC.

  Owen, A. B. (2013). Monte Carlo theory, methods and examples.
  Link to the FREE book (pdf).

- Asmar, D. M., Senanayake, R., Manuel, S., & Kochenderfer, M. J. (2023, May). Model predictive optimized path integral strategies. In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 3182-3188). IEEE.
- Metelli, A. M., Papini, M., Montali, N., & Restelli, M. (2020). Importance sampling techniques for policy optimization. Journal of Machine Learning Research, 21(141), 1-75.

- Next step: Bayesian inference, Kalman filters, and sequential Monte Carlo (particle filters).

  Särkkä, S., & Svensson, L. (2023). Bayesian filtering and smoothing (Vol. 17). Cambridge university press.
  Link to the FREE book (pdf).