STATS 406 F15: Lab 06
Importance sampling

Disclaimer: This note uses its own notation. Do not make direct notational match with lecture notes.

1 Scenarios where naive Monte Carlo integration is a poor choice

- Example 1* **(Locally concentrated integrand)**
  Consider the following function:

$$f(x,y) := 0.5e^{-90(x-0.5)^2-45(y+0.1)^4} + e^{-45(x+0.4)^2-60(y-0.5)^2}$$

  Goal: compute

$$I_1 := \int_{-1}^{1}\int_{-1}^{1} f(x,y)\mathrm{d}x\mathrm{d}y$$

  (* Illustration : see Lab_6.nb (in Mathematica) *)

- Example 2 **(High-dimensional integration)**
  For a p-dimensional variable $\boldsymbol{x} := (x_1,\ldots,x_p)^T \in \mathbb{R}^p$ and a funciton $f : \mathbb{R}^p \to \mathbb{R}$, consider the integral:

$$I_2 := \int_{[-1,1]^p} f(\boldsymbol{x})\mathbb{1}[\|\boldsymbol{x}\|_2 \le 1]\mathrm{d}\boldsymbol{x}$$

  Here, the integrand $f(\boldsymbol{x})\mathbb{1}[\|\boldsymbol{x}\|_2 \le 1]$ is only non-zero inside the p-dimensional unit ball: $\mathcal{B}^p := \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \le 1\}$.

  * If we follow the naive Monte Carlo integration, in which we sample $\boldsymbol{x}$ by uniformly sampling each of its element x_i from Unif$[-1,1]$, that is, sampling uniformly from the p-dimensional unit cube $\mathcal{C}^p := \{\boldsymbol{x} : \|\boldsymbol{x}\|_\infty \le 1\}$, what will happen?
  * **Volume of the p-dimensional unit cube $\mathcal{C}^p$:**

$$\text{Volume}(\mathcal{C}^p) = 2^p$$

*Section 2.5.1, "*Monte Carlo Strategies in Scientific Computing*" by Jun Liu

* **Volume of the p-dimensional unit ball $\mathcal{B}^p$:**

$$\text{Volume}(\mathcal{B}^p) = \frac{(\sqrt{\pi}\times 1)^p}{\Gamma\left(\frac{p}{2}+1\right)} \overset{\text{large } p}{\sim} \frac{1}{\sqrt{p\pi}}\left(\frac{2\pi e}{p}\right)^{\frac{p}{2}} 1^p = \text{Constant}\times p^{-\frac{p+1}{2}}$$

where the first approximation is by Stirling's formula.

* **The proportion of "wasted" sample points goes lightning fast to 100% as p increases!**

* The "curse of dimensionality": the‘ sample size needed increses exponentially with the dimension.

2 Importance sampling(primary form)

Importance sampling is a representative technique for carrying out Monte Carlo integration.

- **Goal:** compute $I = \int f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$[†]. In examples we saw last week, $\boldsymbol{x} = x$ is a scalar(1-dimensional); in Example 1, $\boldsymbol{x} = (x, y)$ is 2-dimensional; in Example 2, $\boldsymbol{x} \in \mathbb{R}^p$.

- A general approach is to choose a distribution $\pi(\boldsymbol{x})$ and rewrite

$$I = \int f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int \frac{f(\boldsymbol{x})}{\pi(\boldsymbol{x})}\cdot\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mathbb{E}\left[\frac{f(\boldsymbol{X})}{\pi(\boldsymbol{X})}\right]$$

  Thus a natural estimator for I is therefore

$$\hat{I} := \hat{\mathbb{E}}\left[\frac{f(\boldsymbol{X})}{\pi(\boldsymbol{X})}\right] = \frac{1}{n}\left\{\frac{f(\boldsymbol{X_1})}{\pi(\boldsymbol{X_1})},\ldots,\frac{f(\boldsymbol{X_n})}{\pi(\boldsymbol{X_n})}\right\} \tag{1}$$

  where $\boldsymbol{X}_i \overset{\text{PDF}}{\sim} \pi(\boldsymbol{x})$.

- $\hat{I}$ is always an unbiased estimator of I regardless of the choice of π, in the sense that

$$\mathbb{E}[\hat{I}] = \mathbb{E}\left[\frac{f(\boldsymbol{X})}{\pi(\boldsymbol{X})}\right]$$

  Good and bad choices of π are differentiated by the "stability" of the resulting estimator $\hat{I}$, mathematically described by $\text{Var}\left[\hat{\mathbb{E}}\left[\frac{f(\boldsymbol{X})}{\pi(\boldsymbol{X})}\right]\right]$.

- **How to choose a good π?**
  Recall $I := \mathbb{E}\left[\frac{f(\boldsymbol{X})}{\pi(\boldsymbol{X})}\right] = \int f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$. Now we show how to choose π to minimize the variance. Notice

$$\begin{aligned}\text{Var}\left(\frac{f(\boldsymbol{X})}{\pi(\boldsymbol{X})}\right) &= \int\left\{\frac{f(\boldsymbol{x})}{\pi(\boldsymbol{x})} - I\right\}^2\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int\left\{\frac{f^2(\boldsymbol{x})}{\pi(\boldsymbol{x})} - 2f(\boldsymbol{x})I + I^2\pi(\boldsymbol{x})\right\}\mathrm{d}\boldsymbol{x}\\ &= \int\frac{f^2(\boldsymbol{x})}{\pi(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} - I^2\end{aligned}$$

[†]Within this course, only consider non-negative $f(\boldsymbol{x})$'s.

Since π is a density,

$$\int \pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$$

By Cauchy-Schwartz inequality, we have

$$\left(\int \frac{f^2(\boldsymbol{x})}{\pi(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}\right)\left(\int \pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right) \geq \left(\int f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)^2$$

where equality holds only when

$$\frac{f^2(\boldsymbol{x})}{\pi(\boldsymbol{x})} \Big/ \pi(\boldsymbol{x}) = \text{Constant}$$

Therefore, the best choice of π is

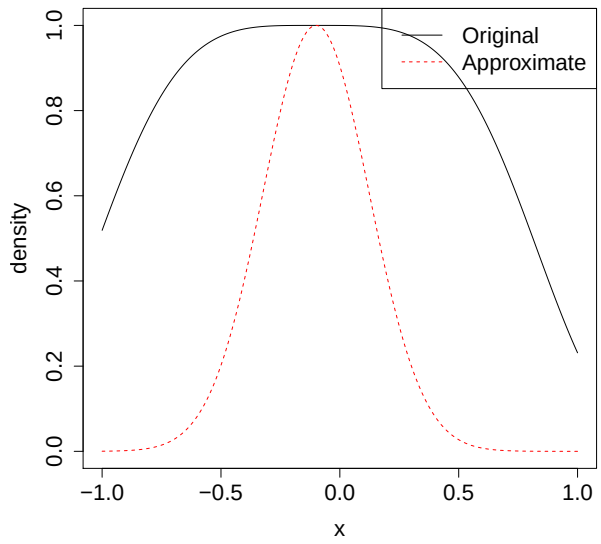
$$\pi(\boldsymbol{x}) \propto f(\boldsymbol{x})$$

- In practice, we do not always know how to sample from a $\pi(\boldsymbol{x})$ that follows the shape of $f(\boldsymbol{x})$, but we usually can find a π whose shape is similar to f, and such π is usually a decent choice.

3 Revisit Examples 1 and 2 with importance sampling

Example 1

- Telling from visual, we should choose $\pi(x, y)$ to be a (truncated) mixture of 2-dimensional normal distributions, with a shape similar to $f(x, y)$.
- We approximately express f to be proportional to a mixture of normal distributions. In order to do this we first approximate $e^{-(y+0.1)^4}$ with $e^{-10(y+0.1)^2}$:



rewrite

$$
\begin{aligned}
f(x,y) &\approx 0.5e^{-90(x-0.5)^2-10(y+0.1)^2} + e^{-45(x+0.4)^2-60(y-0.5)^2} \\
&= 0.5 \times \exp\left[-\left\{\frac{(x-0.5)^2}{2\times\frac{1}{180}} + \frac{(y+0.1)^2}{2\times\frac{1}{20}}\right\}\right] \\
&+ 1 \times \exp\left[-\left\{\frac{(x+0.4)^2}{2\times\frac{1}{90}} + \frac{(y-0.5)^2}{2\times\frac{1}{120}}\right\}\right] \\
&\propto 0.464 \times \mathcal{N}\left[\begin{pmatrix}0.5\\-0.1\end{pmatrix}, \begin{pmatrix}\frac{1}{180} & 0\\ 0 & \frac{1}{20}\end{pmatrix}\right] + 0.536 \times \mathcal{N}\left[\begin{pmatrix}-0.4\\0.5\end{pmatrix}, \begin{pmatrix}\frac{1}{90} & 0\\ 0 & \frac{1}{120}\end{pmatrix}\right] \quad (2)
\end{aligned}
$$

Therefore, in the importance sampling, we should choose $\pi(\boldsymbol{x})$ as follows:

$$
\pi(\boldsymbol{x}) = \pi(x,y)\begin{cases}\propto \text{PDF of RHS of (2)} & \text{if } (x,y)\in[-1,1]^2\\ = 0 & \text{otherwise}\end{cases}
$$

- We have the following algorithm to sample from $\pi(\boldsymbol{x})$:
  1. Toss a coin with probability 0.464 to sample from the first normal distribution and with probability 0.536 to sample from the second normal distribution.
  2. Sample $\boldsymbol{X} = (X, Y)$ from the chosen multivariate normal distribution
  3. Accept $\boldsymbol{X} = (X, Y)$ if it falls within $[-1, 1]^2$, otherwise reject it and return to Step 1.
- Then estimate I by $\hat{I}$ defined in (1).

```
# Numerical experiments for Example 1: see Lab_6.r
```

Example 2

- The choice of $\pi(\boldsymbol{x})$ depends on the shape of $f(\boldsymbol{x})$. For example, choosing $\pi(\boldsymbol{x})$ to be a multivariate normal distribution usually alleviates (but may not completely remedy) the curse of dimensionality.

```
# A numerical example in 4−dimensional space: see Lab_6.r
```

4 A useful variant estimator(reviewing lecture content)

- **Beginning of the story – a slightly different goal:** Now we want to evalute $\mathbb{E}[g(\boldsymbol{X})]$, where the random variable $\boldsymbol{X} \overset{\text{PDF}}{\sim} p(\boldsymbol{x})$. Written in integration form:

$$
J := \int g(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \quad (3)
$$

- If the shape of $g(\boldsymbol{x})$ suggests that an importance sampling is needed and we know the exact $p(\boldsymbol{x})$, we can choose a proper density $q(\boldsymbol{x})$, and rewrite J:

$$J := \int g(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int \frac{g(\boldsymbol{x})p(\boldsymbol{x})}{q(\boldsymbol{x})} \times q(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mathbb{E}\left[\frac{g(\boldsymbol{X})p(\boldsymbol{X})}{q(\boldsymbol{X})}\right]$$

  Then compute the estimator (similar to (1))

$$\hat{J} := \frac{1}{n}\left\{\frac{g(\boldsymbol{X_1})p(\boldsymbol{X_1})}{q(\boldsymbol{X_1})} + \ldots + \frac{g(\boldsymbol{X_n})p(\boldsymbol{X_n})}{q(\boldsymbol{X_n})}\right\} \tag{4}$$

  where $\boldsymbol{X}_i \overset{\text{PDF}}{\sim} q(\boldsymbol{x})$.

- But sometimes, we only know $p(\boldsymbol{x})$ up to a normalizing constant, i.e. $p(\boldsymbol{x}) = C \times p_0(\boldsymbol{x})$.

  * A familiar example: $p(\boldsymbol{x})$ is a truncated normal distribution.
  * A more realistic-like (though still artificial) example:
    1. Define random variable $\boldsymbol{U} := (U_1, \ldots, U_p) \in \{0,1\}^p$.
    2. U_k's are independent, $U_k \sim$Bernoulli(v_k), $k = 1, \ldots, p$.
    3. Given $\boldsymbol{U}$, $\boldsymbol{x}$ follows a conditional distribution $p(\boldsymbol{x}|\boldsymbol{U})$.
    4. Obviously,

$$p(\boldsymbol{x}) \propto \sum_{u_1,\ldots,u_p \in \{0,1\}^p} \left[p(\boldsymbol{x}|\boldsymbol{u}) \times \prod_{k=1}^{p} \left\{u_k^{v_k}(1-u_k)^{1-v_k}\right\}\right]$$

    5. When p is large, it is computationally infeasible to enumerate all possibilities for $\boldsymbol{u} \in \{0,1\}^p$, which is necessary for computing the normalizing constant in $p(\boldsymbol{x})$.
  * Now we don't know the normalizing constant in $p(\boldsymbol{x})$, the problem (3) becomes different from the problem "compute an integral by Monte Carlo methods", because the integrand is not fully specified.

  Notice that (4) requires the knowledge of the normalizing constant (must know exact $p(\boldsymbol{x})$) and is thus inapplicable here.

- **A useful variant of** (4)
  Consider $\tilde{J}$, in which we are only allowed to use p_0, not p, as follows:

$$\tilde{J} := \frac{\dfrac{g(\boldsymbol{X_1})p_0(\boldsymbol{X_1})}{q(\boldsymbol{X_1})} + \ldots + \dfrac{g(\boldsymbol{X_n})p_0(\boldsymbol{X_n})}{q(\boldsymbol{X_n})}}{\dfrac{p_0(\boldsymbol{X_1})}{q(\boldsymbol{X_1})} + \ldots + \dfrac{p_0(\boldsymbol{X_n})}{q(\boldsymbol{X_n})}} \tag{5}$$

  where $\boldsymbol{X}_i \overset{\text{PDF}}{\sim} q(\boldsymbol{x})$. Obviously, apart from the difference in p and p_0, $\hat{J}$ just amounts to replace the denominator of $\tilde{J}$ by n.

* **Advantage of** (5): it suffices to know $p_0(\boldsymbol{x})$, since C in the numerator and the denominator cancel out.

* **Asymptotic unbiasedness of $\tilde{J}$:**

  1. **The numerator:** by Strong Law of Large Numbers,

$$\left\{\frac{g(\boldsymbol{X_1})p_0(\boldsymbol{X_1})}{q(\boldsymbol{X_1})}+\ldots+\frac{g(\boldsymbol{X_n})p_0(\boldsymbol{X_n})}{q(\boldsymbol{X_n})}\right\}\Big/n \overset{\text{a.s.}}{\to} \mathbb{E}\left[\frac{g(\boldsymbol{X})p(\boldsymbol{X})}{q(\boldsymbol{X})}\right]\Big/C =: J/C \tag{6}$$

  2. **The denominator:** by Strong Law of Large Numbers,

$$\left\{\frac{p_0(\boldsymbol{X_1})}{q(\boldsymbol{X_1})}+\ldots+\frac{p_0(\boldsymbol{X_n})}{q(\boldsymbol{X_n})}\right\}\Big/n \overset{\text{a.s.}}{\to} \mathbb{E}\left[\frac{p(\boldsymbol{X})}{q(\boldsymbol{X})}\right]\Big/C \tag{7}$$

$$=\int\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\Big/C=\int p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\Big/C=1\Big/C \tag{8}$$

  3. Divide the numerator and denominator of the RHS of (5) by n and apply (6) and (8). By Slutsky's theorem: `https://en.wikipedia.org/wiki/Slutsky%27s_theorem`[‡], we have

$$\tilde{J}\overset{\text{a.s.}}{\to} J$$

  as $n\to\infty$.

[‡]We used the a.s. convergence version Slutsky's theorem, which is not shown in the Wiki page but can be proved similarly with the continuous mapping theorem.