MAP2212 - EP3 Métodos de integração de Monte Carlo (Quase Aleatório

Vinícius da Costa Collaço - 11811012

Maio de 2022

1 Introdução

Esse relatório visa apresentar uma solução para o segundo exercício programa (EP3) proposto na matéria MAP2212/2022 (Laboratório de Computação e Simulação) do curso de bacharelado de matemática aplicada e computacional (BMAC) do instituito IME-USP.

O objetivo é implementar quatro variantes do método estocástico de integração de Monte Carlo, com um gerador de número quase aleatórios, para integrar a função $f(x) = \exp(-ax)\cos(bx)$ no intervalo $[0, 1]$, onde $a = 0, RG, b = 0, CPF$. RG e CPF são os números de identificação do autor. Para proteção dos dados do autor, serão utilizados somente 6 dígitos significativos do RG e CPF, essa alteração não traz mudança significativa na função ou nos métodos utilizados.

Portanto, a integral cujo o valor será estimado será:

$$\gamma = \int_0^1 e^{-0.460279x}\cos(0,382023x)dx \tag{1}$$

A estimativa da integral deverá ser calculada com um erro mínimo de:

$$\frac{|\hat{\gamma} - \gamma|}{\gamma} \leq 0.0005 \tag{2}$$

Utilizando a linguagem *Python* e bibliotecas adequadas[3, 8, 4, 7, 2], a integral em 1 será estimada utilizando os seguintes métodos de Monte Carlo:

1. *Crude*
2. *Hit or Miss*
3. *Importance Sampling*
4. *Control Variates*

2 Implementação do gerador de números quase aleatórios

Para a geração de números quase aleatórios desse EP, será utilizada a biblioteca chaospy[2], com a geração de uma sequência de baixa discrepância Sobol, tendo o algoritmo implementado na própria biblioteca.

No EP será necessária a execução de duas distribuições distintas; a distribuição uniforme e a distribuição exponencial truncada. Para comparação visual foi gerada na figura1 pontos com as distribuições uniforme e exponencial truncada em [0,1] e com os geradores pseudo aleatório e quase aleatório.

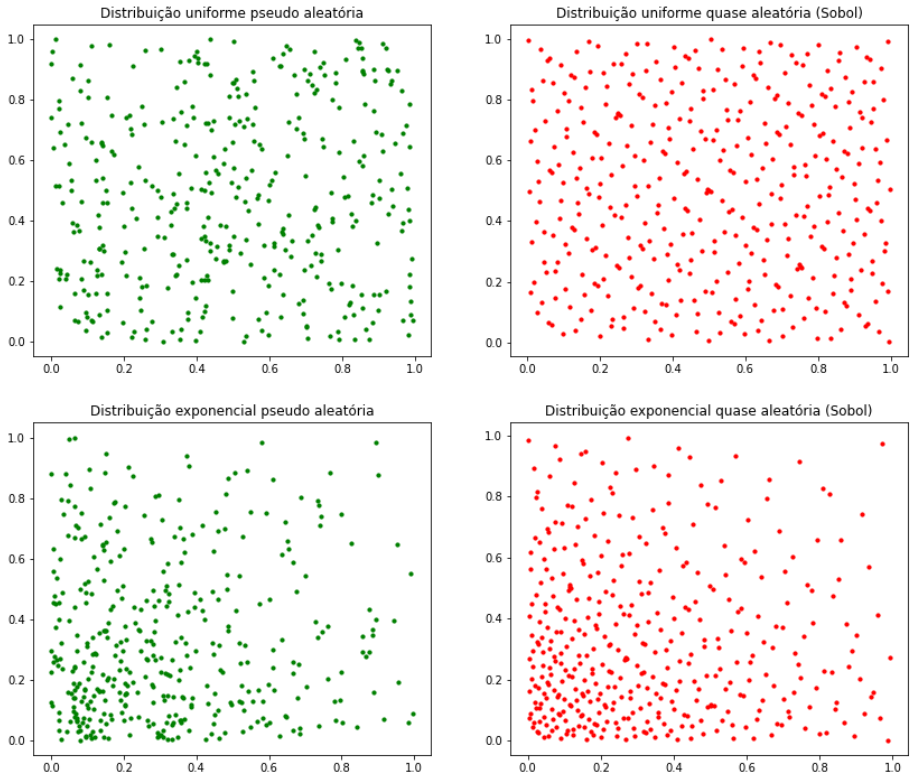

Figura 1: Comparação dos geradores pseudo aleatório e quase aleatório

Mesmo com distribuições distintas pôde-se observar pontos "mais bem comportados"com o gerador de números quase aleatórios.

3 Definindo o tamanho da amostra (n)

Para a definição do valor do n, iremos utilizar a aproximação assintótica de uma distribuição Bernoulli.

Supondo que o tamanho da amostra seja relativamente grande, pelo teorema do limite central, podemos aproximar a Bernoulli por uma normal, tendo:

$$P(|\hat{p} - p| \leq \varepsilon) \geq \gamma$$

[6]

$$P(-\varepsilon \leq \hat{p} - p \leq \varepsilon) = P\left(\frac{-\sqrt{n}\varepsilon}{\sigma} \leq Z \leq \frac{\sqrt{n}\varepsilon}{\sigma}\right) \approx \gamma$$

obtendo finalmente

$$n = \frac{\sigma^2 {Z_\gamma}^2}{\varepsilon^2} \tag{3}$$

esse resultado obtido em 3 será utilizado nos quatro métodos

3.1 Intervalo de confiança

Para o problema será utilizado um intervalo de confiança $\gamma = 95\%$, escolhido arbitrariamente obtendo assim o Z_γ da $N(0, 1)$, portanto $Z_\gamma = 1,96$ e será utilizado nos quatro métodos.

3.2 Erro amostral (ε)

O erro amostral é dado por $|\hat{\gamma} - \gamma|$ com o erro exigido pelo problema em 2, temos que o erro amostral será dado por:

$$\varepsilon = 0.0005 \cdot \gamma$$

porém como não sabemos o valor real de gamma, será utilizado o estimador $\hat{\gamma}$ para cálculo do erro amostral, portanto o erro amostral será dado por:

$$\varepsilon = 0.0005 \cdot \hat{\gamma} \tag{4}$$

o valor do estimador será obtido em cada método através de uma amostra piloto. Para meios de comparação será utilizado o mesmo tamanho de amostra para o piloto, que será descrito nas próximas seções.

3.3 Variância

Para obtenção da variância, em cada método será utilizada a variância amostral ($\hat{\sigma}^2$) obtida a partir da amostra piloto

3.4 Amostra Piloto

Para o tamanho da amostra piloto foi escolhido arbitrariamente $n = 100$

4 Método *Crude*

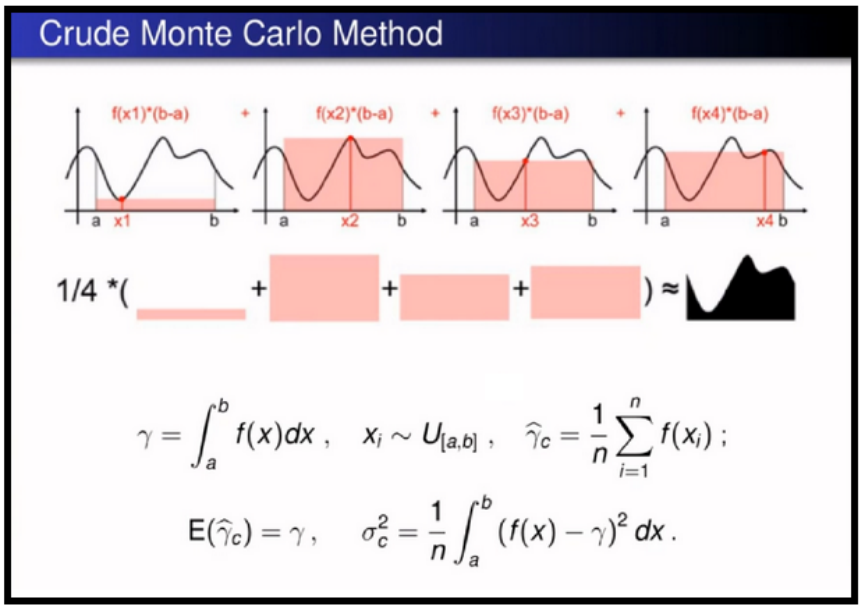

Figura 2: Método Crude

Um resumo do método *Crude* pode ser observado na figura 2

4.1 Estimador $\hat{\gamma}_c$ inicial

Para o cálculo do $\hat{\gamma}_c$ será utilizado um método empírico, uma amostra piloto na função crude(Seed,n), com $n = 100$ e $Seed = 38$, para ser possível a replicação. Resultando em

$$\hat{\gamma}_c = 0.784909$$

4.2 Variância amostral no Método *Crude*

Para o cálculo das variâncias empíricas foi criada a função variancias (Seed, n), retornando a variância de cada método. As variâncias poderiam ser retornadas nas funções principais, mas optou-se por criar outra função para que as funções que foram exigidas no exercício retornassem somente os valores como foram descritos.

Para o método *Crude*, a variância pode ser calculada por:

$$\hat{\sigma_c}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(f(x) - \hat{\gamma}_c)^2$$

Com $Seed = 38$ e $n = 100$, a variância amostral calculada na função variancia, resultou em:

$$\hat{\sigma_c}^2 = 0,0141315$$

Com o resultado obtido em 2, temos:

4.3 Erro amostral *Crude* (ε_c)

$$\varepsilon_c = 0,0005 \cdot \hat{\gamma} = 0,0005 \cdot 0,784909 = 0,000392455$$

4.4 Cálculo do n para o método *Crude* (n_c)

O cálculo do n pode ser obtido através da equação em 3, com os valores amostrais obtidos anteriormente, temos:

$$n_c = \frac{0,0141315 \cdot 1,96^2}{0,000392455^2} = 352468,75$$

portanto:

$$n_c = 352469$$

5 Método *Hit or Miss*

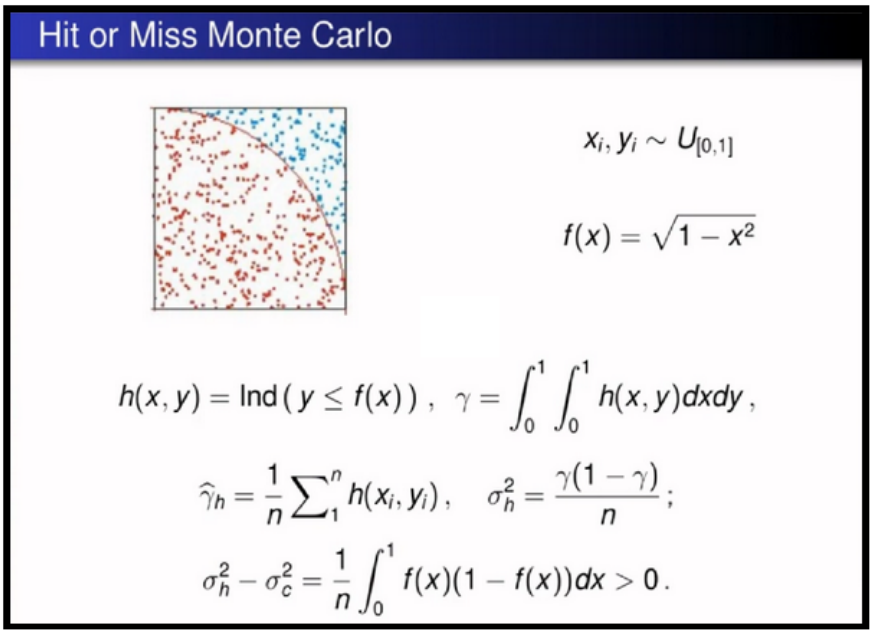

Figura 3: Método Hit or Miss

Um resumo do método *Hit or Miss* pode ser observado na figura 3

5.1 Estimador $\hat{\gamma}_{hom}$ inicial

Para o cálculo do $\hat{\gamma}_{hom}$ será utilizado um método empírico, uma amostra piloto na função hit_or_miss(Seed,n), com $n = 100$ e $Seed = 38$, para ser possível a replicação, resultando em:

$$\hat{\gamma}_{hom} = 0,78$$

5.2 Variância amostral no Método *Hit or Miss*

Para o método *Hit or Miss*, a variância pode ser calculada por:

$$\hat{\sigma}^2_{hom} = \hat{\gamma}_{hom}(1 - \hat{\gamma}_{hom})$$

Com $Seed = 38$ e $n = 100$, a variância amostral calculada na função variancia, resultou em:

$$\hat{\sigma}^2_{hom} = 0,171599$$

5.3 Erro amostral *Hit or Miss* (ε_{hom})

$$\varepsilon_{hom} = 0,0005 \cdot \hat{\gamma} = 0,0005 \cdot 0,78 = 0,00039$$

5.4 Cálculo do n para o método *Hit or Miss* (n_{hom})

O cálculo do n pode ser obtido através da equação em 3, com os valores amostrais obtidos anteriormente, temos:

$$n_{hom} = \frac{0,171599 \cdot 1,96^2}{0,00039^2} = 4334112,82$$

portanto:

$$n_{hom} = 4334113$$

6 Método *Importance Sampling*

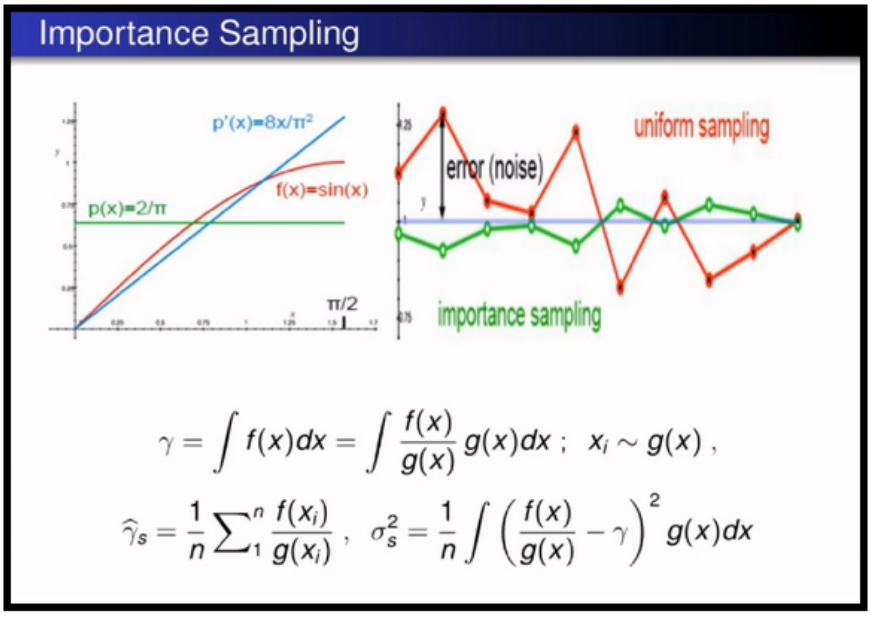

Figura 4: Método Importance Sampling

Um resumo do método *Importance Sampling* pode ser observado na figura 4.

Para esse método utilizou-se a função de probabilidade exponencial truncada da biblioteca chaospy[2], a função chaospy.TruncExponential, com parâmetros $upper = 1$ e $scale = 1/0.460279$, para resultar em uma $g(x)$, no intervalo $[0, 1]$, da seguinte forma:

$$g(x) = \frac{0,460279 \cdot \exp(-0,460279x)}{1 - \exp(-0,460279)}$$

Dessa forma,

$$\frac{f(x)}{g(x)} = \frac{\exp(-0,460279x) \cdot \cos(0,382023x)}{\frac{0,460279 \cdot \exp(-0,460279x)}{1-\exp(-0,460279)}}$$

Resultando em:

$$\frac{f(x)}{g(x)} = \frac{\cos(0,382023x) \cdot (1 - \exp(-0,460279))}{0,460279} \tag{5}$$

Essa a escolha da $g(x)$, simplificou algebricamente o resultado em 5, além de ser de fácil implementação da sua distribuição. Podemos observar a forte correlação no intervalo $[0, 1]$ através da imagem 5, gerada pela plataforma Desmos[1].

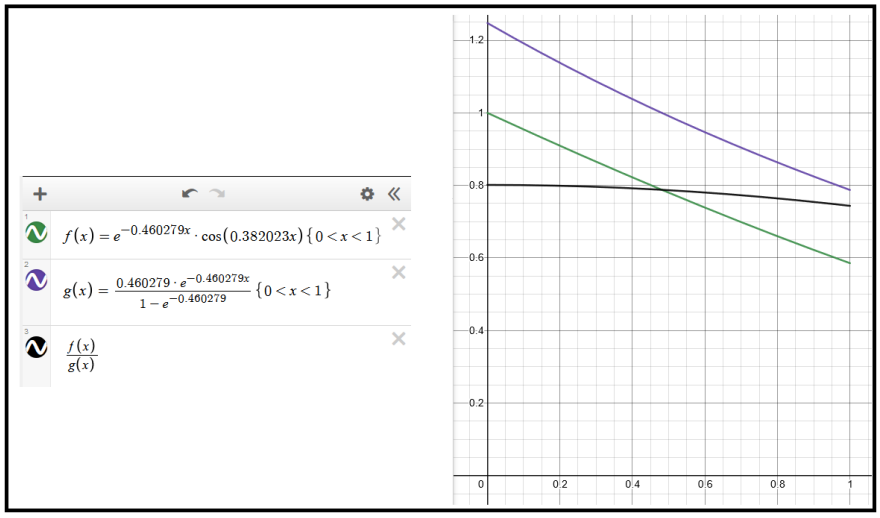

Figura 5: Curvas do Importance Sampling gerada pela plataforma Desmos

Foi plotado também a equação 5, podendo ser observada um valor próximo da integral preterida.

6.1 Estimador $\hat{\gamma}_{is}$ inicial

Para o cálculo do $\hat{\gamma}_{is}$ será utilizado um método empírico, uma amostra piloto na função importance_sampling(Seed,n), com $n = 100$ e $Seed = 38$, para ser possível a replicação, resultando em:

$$\hat{\gamma}_{is} = 0,784540$$

6.2 Variância amostral no Método *Importance Sampling*

Para o método *Importance Sampling*, a variância foi calculada empiricamente com auxílio da função numpy.var() da biblioteca numpy [3], com grau de liberdade = 1:

Com $Seed = 38$ e $n = 100$, a variância amostral calculada na função variancia(), resultou em:

$$\hat{\sigma}^2_{is} = 0,000271609$$

6.3 Erro amostral *Importance Sampling* (ε_{is})

$$\varepsilon_{is} = 0,0005 \cdot \hat{\gamma} = 0,0005 \cdot 0,784540 = 0,000392270$$

6.4 Cálculo do n para o método *Importance Sampling* (n_{is})

O cálculo do n pode ser obtido através da equação em 3, com os valores amostrais obtidos anteriormente, temos:

$$n_{is} = \frac{0,000271609 \cdot 1,96^2}{0,000392270^2} = 6780,88$$

portanto:

$$n_{is} = 6781$$

7 Método *Control Variates*

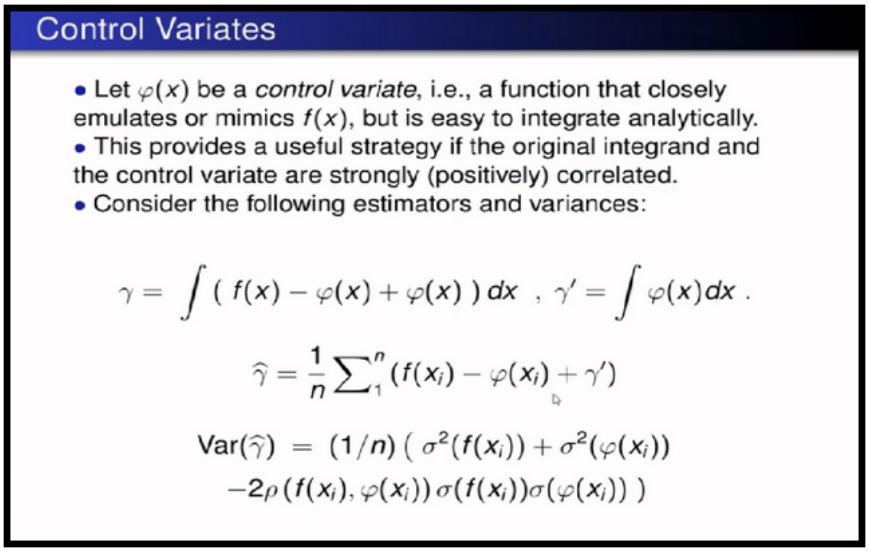

Figura 6: Método Control Variates

Um resumo do método *Control Variates* pode ser observado na figura 6.

Para a escolha da função $\varphi(x)$ de controle, foi utilizado a ferramenta gráfica Desmos[1], para achar uma

função com alta correlação e fácil integração no intervalo $[0, 1]$.

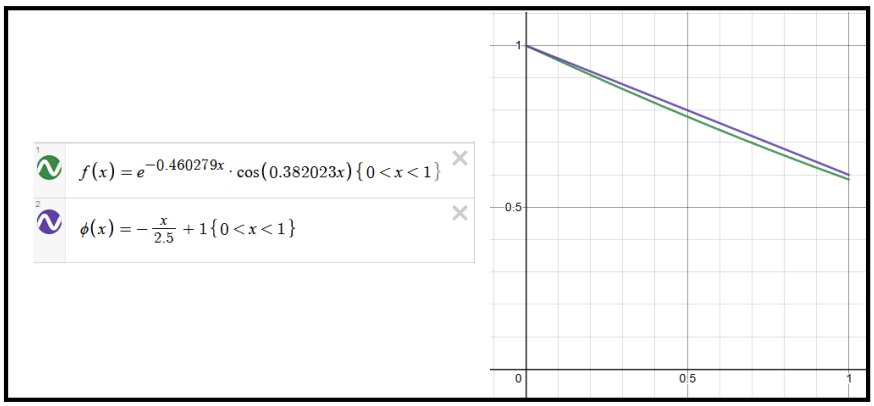

Figura 7: Curvas da função $f(x)$ e $\varphi(x)$ pela plataforma Desmos

Optou-se por:

$$\varphi(x) = -\frac{x}{2,5} + 1$$

Na figura 7 pode-se observar uma alta correlação entre as duas figuras.

7.1 Integral de $\varphi(x)$

$$\int_0^1 \varphi(x)dx = \int_0^1 -\frac{x}{2,5}dx + \int_0^1 1dx = -0,2 + 1 = 0,8$$

7.2 Estimador $\hat{\gamma}_{cv}$ inicial

Para o cálculo do $\hat{\gamma}_{cv}$ será utilizado um método empírico, uma amostra piloto na função control_variate(Seed,n), com $n = 100$ e $Seed = 38$, para ser possível a replicação, resultando em:

$$\hat{\gamma}_{cv} = 0,784190$$

7.3 Variância amostral no Método *Control Variate*

No método *Control Variate* a variância pode ser calculada por:

$$var(\hat{\gamma}_{cv}) = (1/n)var(f(x)) + var(\varphi(x) - 2 \cdot cov(f(x), \varphi(x))$$

Como a variância de $f(x)$ e $\varphi(x)$ é desconhecida, foi calculada empiricamente a covariância amostral e as variâncias amostrais com auxílio da função numpy.var(), e numpy.cov() da biblioteca numpy [3], com grau de liberdade = 1:

Com $Seed = 38$ e $n = 100$, a variância amostral calculada na função variancia(), resultou em:

$$\hat{\sigma}^2_{cv} = 0,0000336645$$

7.4 Erro amostral *Control Variate* (ε_{cv})

$$\varepsilon_{cv} = 0,0005 \cdot \hat{\gamma} = 0,0005 \cdot 0,784190 = 0,000392095$$

7.5 Cálculo do n para o método *Control Variate* (n_{cv})

O cálculo do n pode ser obtido através da equação em 3, com os valores amostrais obtidos anteriormente, temos:

$$n_{cv} = \frac{0,0000352536 \cdot 1,96^2}{0,0003920953^2} = 841,20$$

portanto:

$$n_{cv} = 842$$

8 Resultados e discussões

Para comparação dos métodos será utilizado o seguinte método de cálculo de eficiência: [5]

$$Eff(\hat{\gamma}) = [MSE(\hat{\gamma}) \times C(\hat{\gamma})]^{-1}$$

Como os estimadores são não-viesados a eficiência pode ser dada por:

$$Eff(\hat{\gamma}) = [\hat{\sigma}^2(\hat{\gamma}) \times C(\hat{\gamma})]^{-1}$$

onde $\hat{\sigma}^2(\hat{\gamma})$ é a variância do estimador e $C(\hat{\gamma})$ o tempo estimado para calcular $\hat{\gamma}$

Portanto quanto maior a eficiência melhor o método, contabilizando o tempo de execução, em consequência o algorítimo utilizado

	$\hat{\gamma}$	Erro $\left(\frac{\lvert\hat{\gamma}-\gamma\rvert}{\gamma}\right)$	$\hat{\sigma}^2(\hat{\gamma})$	n	$C(\hat{\gamma})$(s)	Eficiência ($\hat{\gamma}$)
Crude (pseudo)	0,784203	$1,651 \times 10^{-4}$	0,01154	296596	0,02039	4246,2
Crude (sobol)	0,784280	$1,547 \times 10^{-6}$	0,0141315	352469	1,7292	40,9
Hit or Miss (pseudo)	0,784141	$3,469 \times 10^{-4}$	0,1824	4852548	0,7539	7,3
Hit or Miss (sobol)	0,784286	$7,901 \times 10^{-6}$	0,171599	4334113	21,9554	0,26
Import. Samp. (pseudo)	0,784316	$6,765 \times 10^{-5}$	0,0002875	7187	0,0007908	4398001,2
Import. Samp. (sobol)	0,784180	$4,459 \times 10^{-8}$	0,000271609	6781	0,03504	105083,3
Control Variate (pseudo)	0,784143	$7,975 \times 10^{-5}$	0,00003554	884	0,004347	6525271,3
Control Variate (sobol)	0,784274	$7,767 \times 10^{-6}$	0,00003366	842	0,009617	3088758,1

Tabela 1: Tabela Comparativa

No geral tivemos uma redução de variância com o gerador de números pseudo aleatório sobol em comparação com o gerador pseudo aleatório e consequentemente uma redução da amostra para atingir o erro

proposto, com exceção do método crude, isso pode ter ocorrido pela decisão inicial de se usar uma amostra inicial de apenas 100 amostras.

Como a integral proposta é de fácil cálculo, podemos obter o erro real, e no geral com o gerador quase aleatório tivemos erro menor, e bem pequeno. Tendo um indício que com uma amostra inicial maior o n poderia ter uma redução no geral.

Porém mesmo nos casos com menor variância e consequentemente n menor tivemos uma eficiência pior, pois o gerador de números quase aleatórios é mais lento que o gerador de números pseudo aleatórios.

9 Conclusão

A tarefa foi eficiente para mostrar o método de Monte Carlo nativo, e algumas variantes com redução de variância, além de aplicação de bibliotecas para geração de números pseudo aleatórios. Na tabela 1, podemos observar que o método mais eficiente foi o método *Control Variate*, seguido do *Importance Sampling*, em ambos foi possível encontrar facilmente funções apropriadas e tal fato se mostrou efetivo nos resultados. O método *Crude* foi o método de mais fácil implementação, porém precisou de uma amostra grande para atingir o intervalo de confiança desejado. O método *Hit or Miss* foi o com cálculo de variância mais fácil de se calcular, porém por ter um n grande e necessitar de dois arrays de pontos foi o método menos eficiente.

Referências

[1] Inc. Desmos. Plataforma desmos. `https://www.desmos.com/`, 2022. [Online; acessado 19-Abril-2022].

[2] Jonathan Feinberg and Hans Petter Langtangen. Chaospy: An open source tool for designing methods of uncertainty quantification. *Journal of Computational Science*, 11:46–57, 2015.

[3] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[4] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[5] C. Lemieux. *Monte Carlo and Quasi-Monte Carlo Sampling*. Springer Series in Statistics. Springer, 2008.

[6] Wilton de O. Bussab Pedro A. Morettin. *Estatística Básica,* 6^a *Edição*. 2010.

[7] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[8] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.