Adaptive Importance Sampling in Monte Carlo Integration

Diogo C. Netto*

Abstract

In the following paper, we analyze two of the Monte Carlo with Adaptive Importance Sampling algorithms for numerical integration. We start our survey with a brief analysis of the standard Monte Carlo method, showing how Importance Sampling can be used as an alternative to reduce variance and motivating the search for adaptive methods. We then turn our focus to an analysis of the adaptive scheme proposed by Ryu and Boyd, in which we present a convexity result which allows us to reduce the problem of accelerating the convergence of the Monte Carlo estimator to a convex optimization problem. We also analyze an adaptive scheme proposed by Oh and Berger and highlight one of its asymptotic properties. We conclude by reproducing the implementations of Ryu and Boyd and of Oh and Berger and comparing their numerical convergence with deterministic schemes of quadrature.

1 Introduction

A significant number of statistical problems can be reduced to the calculation of integrals of the form

$$I = \int_{\mathcal{X}} \phi(x)\pi(x)dx, \tag{1}$$

where $x \in \mathcal{X} \subset \mathbb{R}^p$, $\phi(\cdot) : \mathcal{X} \to \mathbb{R}$ and $\pi(\cdot)$ is a probability density function over the alphabet $\mathcal{X}$. For instance, the Bayesian least squares estimator $\mathbf{x}_{BLS}(y)$ can be calculated through the equation (1) with the choice of functions $\phi(\mathbf{x}) = \mathbf{x}$ and $\pi(\cdot) = p_{X|Y}(\cdot|y)$.

In most applications, the integrals in (1) are analytically intractable and one must use methods of numerical integration. In high-dimensional problems, the stochastic approach is often preferred.

Within the plethora of Stochastic methods, the standard Monte Carlo approach rises as a simple and easy to implement method, which we describe in the following lines. Given the distribution $\pi(\cdot) : \mathcal{X} \to \mathbb{R}$, the standard Monte Carlo method consists of generating *iid* samples $X_1, ..., X_n \sim \pi(\cdot)$ and estimating I through

$$\hat{I}_n^{MC} = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i). \tag{2}$$

One may easily see that $\hat{I}_n^{MC}$ is unbiased:

$$\mathbb{E}[\hat{I}_n^{MC}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\phi(X_i)]. = \mathbb{E}[\phi(\cdot)] = \int_{\mathcal{X}} \phi(x)\pi(x)dx, \tag{3}$$

*Massachusetts Institute of Technology, MA.

and has variance given by

$$Var[\hat{I}_n^{MC}] = \frac{1}{n^2}\sum_{i=1}^{n} Var[\phi(X_i)] = \frac{1}{n}\left(\int_{\mathcal{X}} \phi^2(x)\pi(x)dx - I^2\right). \tag{4}$$

As an attempt to reduce the variance of the estimator (and accelerate convergence as a consequence), one may use importance sampling, which consists of taking *i.i.d* samples $X_1, ..., X_n \sim \tilde{\pi}(\cdot)$, where $\tilde{\pi}(\cdot)$ is an auxiliary distribution, and estimating I through

$$\hat{I}_n^{IS} = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)\frac{\pi(X_i)}{\tilde{\pi}(X_i)}. \tag{5}$$

The estimator above is clearly unbiased and has variance given by

$$Var[\hat{I}_n^{IS}] = \frac{1}{n^2} Var_{\tilde{\pi}}\left[\frac{\phi(X_i)\pi(X_i)}{\tilde{\pi}(X_i)}\right] = \frac{1}{n}\left(\int_{\mathcal{X}} \frac{\phi^2(x)\pi^2(x)}{\tilde{\pi}(x)dx} - I^2\right). \tag{6}$$

A rearrangement of the terms allows us to write

$$Var[\hat{I}_n^{IS}] = \frac{1}{n}\left(\int_{\mathcal{X}} \frac{\phi^2(x)\pi^2(x)dx}{\tilde{\pi}(x)} - I^2\right) = \frac{1}{n}\left(\int_{\mathcal{X}} \frac{(\phi(x)\pi(x) - I\tilde{\pi}(x))^2 dx}{\tilde{\pi}(x)}\right), \tag{7}$$

from which we conclude that the variance is minimized for

$$\tilde{\pi}(\cdot) \propto \phi(\cdot)\pi(\cdot). \tag{8}$$

One may note that choosing $\tilde{\pi}(\cdot)$ wisely can reduce the variance of the estimator, but it's often difficult to perform in practice. A simple heuristic may be to choose the sampling distribution from a parameterized family and adjust the natural parameter such that the condition given in (8) holds approximately.

More sophisticated approaches consist of choosing the sampling distribution from a parameterized family, but gradually improving the sampler by adjusting the natural parameter in each iteration. Within this framework, we may, for example, choose the natural parameter as the solution for a convex optimization problem involving the family of sampling distributions, $\pi(\cdot)$ and $\phi(\cdot)$. We may also adjust the natural parameter directly through a suitable empirical average with respect to the past samples. These are the approaches followed, respectively, by Ryu and Boyd and Oh and Berger, which we analyze in the next sections.

2 Adaptive Importance Sampling via Stochastic Convex Programming

The ansatz of Ryu and Boyd's adaptive scheme is to choose the sampling distribution from a parameterized linear exponential family, reducing, as we will show, the update of the sampling distribution to a problem of Convex Optimization, for which Gradient Descent methods may be used and for which theoretical guarantees about convergence can be made.

In order to define the parameterized family of distributions, let us first choose a parameter space $\Theta \subset \mathbb{R}^k$, a natural statistc $T(\cdot) : \mathbb{R}^p \to \mathbb{R}^k$ and a base distribution $h(\cdot) : \mathcal{X} \to \mathbb{R}_+$. A parameterized distribution $\tilde{\pi}_\theta(x)$ can be written as

$$\tilde{\pi}_\theta(x) = \exp(\theta^T T(x) - A(\theta))h(x), \tag{9}$$

where $A(\theta)$ is a normalizing factor given by

$$A(\theta) = \log \int \exp(\theta^T T(x))h(x)dx. \tag{10}$$

Finally, $\mathcal{P} = \{\tilde{\pi}_\theta(\cdot)|\theta \in \Theta\}$ defines the exponential family under consideration.

Next, we define $V(\cdot) : \mathbb{R}^p \to \mathbb{R}$ to be the per-sample variance of the importance sampling estimator for distributions in $\mathcal{P}$

$$V(\theta) = Var_{X\sim\tilde{\pi}_\theta}\left[\frac{\phi(X)\pi(X)}{\tilde{\pi}_\theta(X)}\right] = \int \frac{\phi^2(x)\pi^2(x)}{\tilde{\pi}_\theta(x)}dx - I^2. \tag{11}$$

That is it, the importance sampling estimator obtained through n *i.i.d* samples from $\tilde{\pi}_\theta(\cdot)$ has variance $\frac{V(\theta)}{n}$. Let us now prove a convexity result for $V(\theta)$.

THEOREM 2.1. *The per-sample variance $V(\theta)$ of the importance sampling estimator is a convex function in θ.*

Proof. We begin by proving that $A(\cdot)$ is convex in θ:

$$\exp(A(\eta\theta_1 + (1-\eta)\theta_2)) = \int \exp((\eta\theta_1 + (1-\eta)\theta_2)^T T(x))h(x)dx. \tag{12}$$

Define the norm $\|f\|_p = (\int |f|^p dx)^{\frac{1}{p}}$. Hölder's inequality states that if $p, q \in \mathbb{R}_+$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$ then

$$\|fg\|_1 \leq \|f\|_p \|g\|_q. \tag{13}$$

Choosing $f(\theta) = h(x)\exp(\eta\theta_1^T T(x)), g(\theta) = h(x)\exp((1-\eta)\theta_2^T T(x))$ yields

$$\exp(A(\eta\theta_1 + (1-\eta)\theta_2)) \leq \left(\int \exp(\theta_1^T T(x))h(x)dx\right)^\eta \left(\int \exp(\theta_2^T T(x))h(x)dx\right)^{1-\eta}, \tag{14}$$

that is it,

$$\exp(A(\eta\theta_1 + (1-\eta)\theta_2)) \leq \exp(A(\theta_1))^\eta \exp(A(\theta_2))^{1-\eta}. \tag{15}$$

Taking the log on both sides gives the desired result of the convexity of $A(\theta)$

$$A(\eta\theta_1 + (1-\eta)\theta_2) \leq \eta A(\theta_1) + (1-\eta)A(\theta_2). \tag{16}$$

Since $A(\theta)$ and $\theta^T T(x)$ are convex in θ, then $\exp(A(\theta) - \theta^T T(x))$ is also convex in θ. From this result, we conclude that

$$V(\theta) = \int \phi^2(x)\pi^2(x)\exp(A(\theta) - \theta^T T(x))dx - I^2 \tag{17}$$

is convex in θ.

□

The convexity of $V(\theta)$ motivates us to look for Gradient Descent methods. We can compute the gradient $\nabla_\theta V(\theta)$ as

$$\nabla_\theta V(\theta) = \nabla_\theta \left[\int \frac{\phi^2(x)\pi^2(x)}{\tilde{\pi}_\theta} dx - I^2 \right], \tag{18}$$

from which we obtain

$$\nabla_\theta V(\theta) = \int \nabla_\theta \left[\phi^2(x) \frac{\pi^2(x)}{\tilde{\pi}_\theta} \right] dx = \int (\nabla_\theta A(\theta) - T(x)) \frac{\phi^2(x)\pi^2(x)}{\tilde{\pi}_\theta^2(x)} \tilde{\pi}_\theta(x) dx. \tag{19}$$

That is it,

$$\nabla_\theta V(\theta) = \mathbb{E}_{\tilde{\pi}_\theta} \left[(\nabla_\theta A(\theta) - T(x)) \frac{\phi^2(x)\pi^2(x)}{\tilde{\pi}_\theta^2(x)} \right] \tag{20}$$

In short, by taking a sample $X \sim \tilde{\pi}_\theta$ and computing $\gamma = (\nabla_\theta A(\theta) - T(X)) \frac{\phi^2(x)\pi^2(X)}{\tilde{\pi}_\theta^2(X)}$, we obtain a vector γ satisfying $\mathbb{E}[\gamma] = \nabla_\theta V(\theta)$, which is one of the fundamental ideas of Ryu and Boyd's approach.

Let us now present the Adaptive Importance Sampling via Stochastic Convex Programming algorithm.

Algorithm 1 Adaptive Importance Sampling via Stochastic Convex Programming (AIS-SCP)

Input: initial estimate of θ_1, number of iterations N and step-size C.
$n \leftarrow 1, \hat{I}_1 \leftarrow 0$.
for $n < N$ **do**
  Sample $X_n \sim \tilde{\pi}_\theta$.
  Set $\hat{I}_{n+1} \leftarrow \frac{\hat{I}_n \cdot n + \frac{\phi(X_n)\pi(X_n)}{\tilde{\pi}_\theta(X_n)}}{n+1}$.
  Set $\gamma_n \leftarrow (\nabla_\theta A(\theta) - T(X_n)) \frac{\phi^2(X_n)\pi^2(X_n)}{\tilde{\pi}_\theta^2(X_n)}$.
  Set $\theta_{n+1} \leftarrow \theta_n - \frac{C\gamma_n}{\sqrt{n}}$.
end for
return $\hat{I}_N$.

Finally, we highlight a convergence result for the Stochastic Convex Programming approach. The proof is presented in Ryu and Boyd, but it is out of scope for the purposes of this paper.

THEOREM 2.2. *Let* $K(\theta) = \int \frac{\phi^4(x)\pi^4(x)}{\tilde{\pi}_\theta{}^3(x)} dx$. *Assuming that* $\Theta \subset \{\theta | K(\theta) < \infty\}$ *is nonempty, convex and compact, then*

$$\sqrt{n}(\hat{I}_n - I) \xrightarrow{d} \mathcal{N}(0, V^*), \tag{21}$$

where $V^* = \inf_{\theta \in \Theta} V(\theta)$.

3 Adaptive Importance Sampling via Cumulative Average Parameter Update

In this section, we discuss the adaptive algorithm proposed by Oh and Berger. Similarly to Ryu and Boyd's approach, in this scheme we also choose the sampling distribution from a parameterized family and update the corresponding natural parameter in each iteration.

The fundamental differences, as we will see, are that the family of distributions doesn't need to be an exponential family and also, that the parameter is adapted according to a cumulative average with respect to the past samples, rather than chosen in order to minimize an objective function.

Let $\Theta \subset \mathbb{R}^k$ be a parameter space and $\mathcal{P} = \{\tilde{\pi}_\theta | \theta \in \Theta\}$ be the parameterized family of distributions. Let us also assume that in the optimal scenario, $\theta = \mathbb{E}[\zeta(\theta)]$, where $\zeta(\theta) = [\zeta_1(\theta), ..., \zeta_k(\theta)]^T$.

We can motivate this optimality condition from a statistical analysis for a particular example. Indeed, suppose that we are trying to perform the computations for a distribution $\pi(\cdot) : \mathcal{X} \subset \mathbb{R}^p \to \mathbb{R}_+$ and we choose the family $\mathcal{P}$ to be the set of p-dimensional multivariate normal distributions. The true (μ, Σ) parameters of the normal distribution take the form of expectations of suitable functions of the random variable, and its maximum likelihood versions $(\hat{\mu}_{ML}, \hat{\Sigma}_{ML})$ take the form of empirical averages over the dataset, from which we motivate the ideal parameterization in terms of expectations of appropriate functions of the random variable.

The ansatz of Oh and Berger's approach is to update the parameter θ according to a cumulative empirical average of $\zeta(\cdot)$ over the dataset, as we describe below. For purposes of notation, let $\phi(\cdot) = [\phi_1(\cdot), ..., \phi_p(\cdot)]^T$.

Algorithm 2 Adaptive Importance Sampling via Cumulative Average Parameter Update (AIS-CAPU)

Input: Stopping criterion, initial estimate of θ, and allocation policy $n_k|_{k\in\mathbb{N}}$.
$j \leftarrow 1$
while Stopping criterion is not satisfied **do**
  Draw n_j *i.i.d* samples $X_1^{(j)}, ..., X_{n_j}^{(j)} \sim \tilde{\pi}_{\theta_j}$.
  Let $w^{(j)}(\cdot) = \frac{\pi(\cdot)}{\tilde{\pi}_{\theta_j}(\cdot)}$. Define the functional $W_j(h) = \sum_{i=1}^{n_j} h(X_i^{(j)}) w^{(j)}(X_i^{(j)})$.
  Compute $W_j(\zeta_i)$, for $i = 1, ..., k$ and $W_k(\phi_i)$ for $i = 1, ..., p$.
  Set $\theta_{j+1} = \left[\frac{\sum_{i=1}^j W_i(\zeta_1)}{\sum_{i=1}^j W_i(\mathcal{I})}, ..., \frac{\sum_{i=1}^j W_i(\zeta_p)}{\sum_{i=1}^j W_i(\mathcal{I})} \right]^T$, where $\mathcal{I}$ denotes the identity function.
  Set $j \leftarrow j + 1$.
end while
return $\hat{I}_j = \left[\frac{\sum_{i=1}^j W_i(\phi_1)}{\sum_{i=1}^j W_i(\mathcal{I})}, ..., \frac{\sum_{i=1}^j W_i(\phi_p)}{\sum_{i=1}^j W_i(\mathcal{I})} \right]^T$.

Again, we highlight a convergence result from Oh and Berger, whose proof is out of scope for the purposes of this paper.

THEOREM 3.1. *Suppose that $\tilde{\pi}_{\theta_j}(\cdot)$ has the same support as $\pi(\cdot)$ for all $j \in \mathbb{N}$, $\mathbb{E}[\phi(\cdot)]$ exists and that $w^{(j)}$ is bounded by a constant for all $j \in \mathbb{N}$, then*

$$\hat{I}_j \xrightarrow{a.s.} I, \text{ as } \sum_{i=1}^{j} n^{(i)} \to \infty. \tag{22}$$

4 Numerical Experiments

In this section, we analyze the numerical behavior of the two algorithms described in the previous sections, comparing their convergence with deterministic schemes of quadrature.

We decided to adopt the number of samples drawn as the common unit of cost for the comparison between the stochastic methods. Indeed, generating a random sample from the distribution was the most expensive step (in terms of runtime) in each iteration, which justifies our choice of unit of cost. Runtime was only used to compare the stochastic methods with deterministic schemes.

For the purposes of testing, we decided to estimate the expected value of a multivariate student's t-distribution with 2 degrees of freedom. We chose as our parameterized family of sampling distributions the set of multivariate normals $\mathcal{P} = \{\mathcal{N}_d(\theta, I_d) | \theta \in \mathbb{R}^d\}$, where d was a parameter that we vary in our tests.

4.1 Comparison between Stochastic Methods

For the purposes of a comparison between the stochastic approaches, we analyzed the accuracy of their estimates for the expected value of a 2-dimensional student's t-distribution with 2 degrees of freedom as a function of the number of samples drawn. The relative error was measured according to $\epsilon_{rel} = \frac{\|\hat{I} - I\|}{\|I\|}$, where $\hat{I}$ is the estimated expected value and I is the true expected value. Furthermore, each algorithm was run 20 times for the same input in order to reduce noise.

After drawing 5×10^5 samples for this particular problem, we were able to observe a relative error of approximately 3×10^{-2} in both methods. The results were consistent with the ones found in Oh and Berger and Ryu and Boyd, respectively.

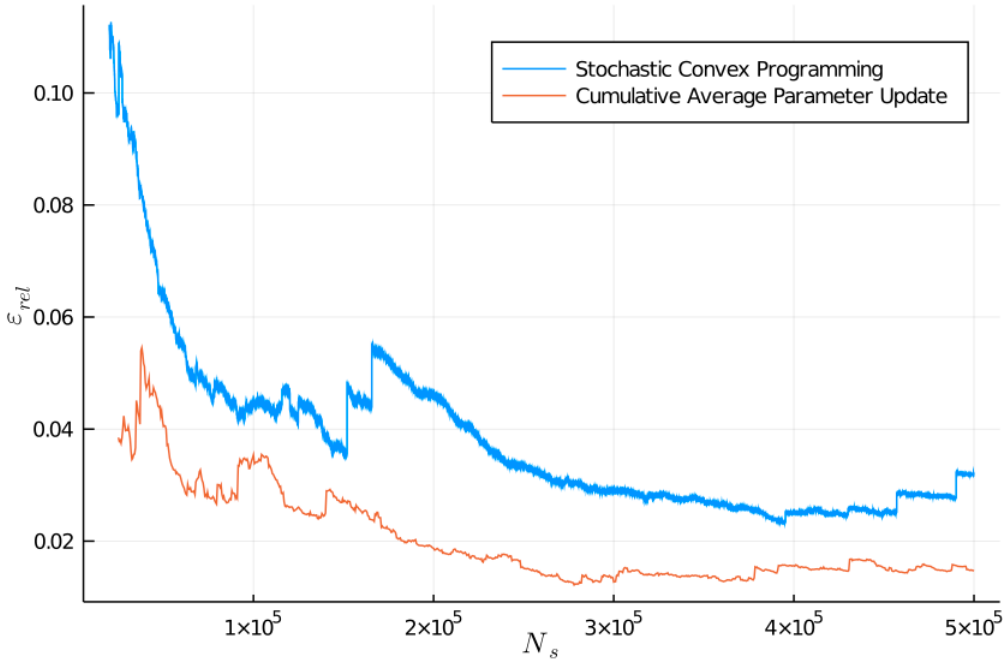

FIG. 1. *Relative error for both methods as a function of the number of samples drawn.*

4.2 Comparison between Stochastic Methods and Quadrature Schemes

In order to present a more extensive analysis, we compared the performance and accuracy of the stochastic methods to the deterministic scheme of quadrature used by Cubature.jl.

To perform these tests, we estimated the expected value of the first coordinate of a random variable distributed according to a d-dimensional student's t-distribuion. Also, we kept the accuracy fixed for problems in the same dimension.

Finally, we performed the change of variables $x_j = \frac{t_j}{1-t_j^2}, \forall 1 \leq j \leq d$ when running Cubature.jl in order to transform the improper integral into an integral over the unit cube $[-1,1]^d$ and ease the computation.

We summarize the runtime results in the table below.

d	$\epsilon_{rel} = \frac{\lvert\hat{I}^{(1)} - I^{(1)}\rvert}{\lvert I^{(1)}\rvert}$	Curbature.jl (s)	AIS-SCP (s)	AIS-CAPU (s)
2	0.1	0.68	4.51	5.28
3	0.1	0.81	5.50	8.12
4	0.3	1.35	5.63	9.98
5	0.3	2.55	7.05	12.25
6	0.3	$> 10^3$	7.31	15.64
7	0.3	$> 10^3$	7.34	20.10

TABLE 1

Comparison of runtimes between stochastic and quadrature methods. Accuracy was fixed for tests with data in the same dimension.

We were able to verify the better performance of stochastic methods for high-dimensional problems.

4.3 Experimental Convergence Rates

According to the Central Limit Theorem for the Stochastic Convex Programming approach,

$$\sqrt{n}(\hat{I}_n - I) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V^*), \tag{23}$$

from which we conclude that for one-dimensional problems,

$$\mathbb{P}\left[|\hat{I}_n - I| \geq \alpha\sqrt{\frac{V*}{n}}\right] \doteq 2(1 - \Phi(\alpha)), \tag{24}$$

so that for any $\epsilon > 0$ there exists $C(\epsilon)$ such that

$$\frac{|\hat{I}_n - I|}{|I|} \dot{\leq} C(\epsilon)\sqrt{\frac{V^*}{n}} \tag{25}$$

with probability $1 - \epsilon$. Because of this, we expect that with high-probability, we should be able to bound the error by $Cn^{-\frac{1}{2}}$, for $C \in \mathbb{R}_+$.

As shown in Oh and Berger, a Central Limit Theorem also holds for the Cumulative Average Parameter Update estimator. Because of this, we also expect that with high-probability the $\sim n^{-\frac{1}{2}}$ error upper-bound holds.

To confirm these bounds, we repeated each algorithm for the same input 10 times and obtained the average errors.We were able to observe the the $\sim n^{-\frac{1}{2}}$ error upper decay given by the Central Limit Theorem experimentally, as shown in the plot.

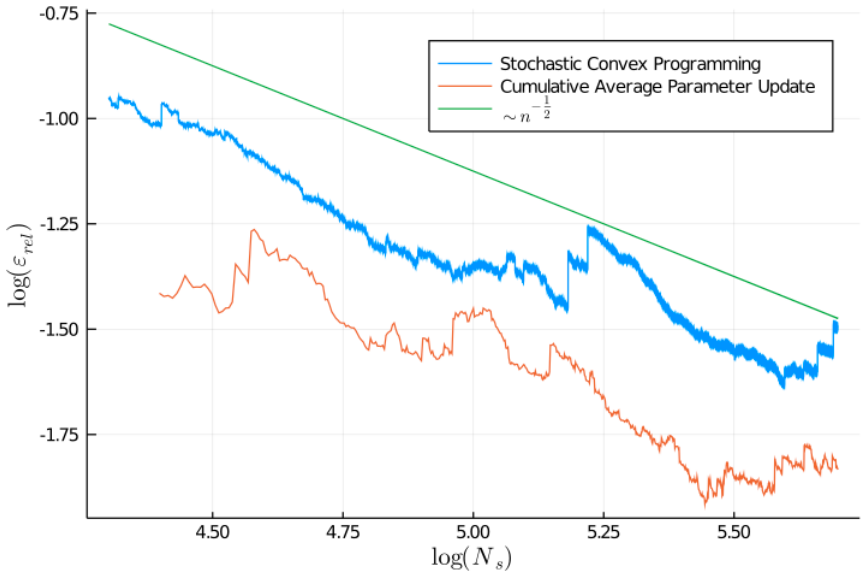

FIG. 2. *Experimental confirmation of the* $\sim n^{-\frac{1}{2}}$ *error upper-bound for the stochastic methods.*

References

[1] Ryu, E. and Boyd, S., Adaptive Importance Sampling via Stochastic Convex Programming, arXiv preprint, 2015.

[2] Oh, M-S and Berger, J., Adaptive importance sampling in Monte Carlo integration, Journal of Statistical Computation and Simulation, 1992.

[3] Evans, M. and Swartz, T., Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems, Statistical Science, 1995.

[4] Owen, A. and Zhou, Y., Safe and Effective Importance Sampling. Journal of the American Statistical Association. 2000.

[5] Johnson, S., Cubature.jl, `https://github.com/stevengj/cubature/`