Adaptive Stratified Sampling for Monte-Carlo integration of Differentiable functions

Alexandra Carpentier
Statistical Laboratory, CMS
Wilberforce Road, Cambridge
CB3 0WB UK
a.carpentier@statslab.cam.ac.uk

Rémi Munos
INRIA Lille - Nord Europe
40, avenue Halley
59000 Villeneuve d'ascq, France
remi.munos@inria.fr

Abstract

We consider the problem of adaptive stratified sampling for Monte Carlo integration of a differentiable function given a finite number of evaluations to the function. We construct a sampling scheme that samples more often in regions where the function oscillates more, while allocating the samples such that they are well spread on the domain (this notion shares similitude with low discrepancy). We prove that the estimate returned by the algorithm is almost similarly accurate as the estimate that an optimal oracle strategy (that would know the variations of the function *everywhere*) would return, and provide a finite-sample analysis.

1 Introduction

In this paper we consider the problem of numerical integration of a differentiable function $f : [0,1]^d \to \mathbb{R}$ given a finite budget n of evaluations to the function that can be allocated sequentially.

A usual technique for reducing the mean squared error (w.r.t. the integral of f) of a Monte-Carlo estimate is the so-called stratified Monte Carlo sampling, which considers sampling into a set of strata, or regions of the domain, that form a partition, i.e. a stratification, of the domain (see [10][Subsection 5.5] or [6]). It is efficient (up to rounding issues) to stratify the domain, since when allocating to each stratum a number of samples proportional to its measure, the mean squared error of the resulting estimate is always smaller or equal to the one of the crude Monte-Carlo estimate (that samples uniformly the domain).

Since the considered functions are differentiable, if the domain is stratified in K hyper-cubic strata of same measure and if one assigns uniformly at random n/K samples per stratum, the mean squared error of the resulting stratified estimate is in $O(n^{-1}K^{-2/d})$. We deduce that if the stratification is built *independently* of the samples (before collecting the samples), and if n is known from the beginning (which is assumed here), the minimax-optimal choice for the stratification is to build n strata of same measure and minimal diameter, and to assign only one sample per stratum uniformly at random. We refer to this sampling technique as Uniform stratified Monte-Carlo. The resulting estimate has a mean squared error of order $O(n^{-(1+2/d)})$. The arguments that advocate for stratifying in strata of same measure and minimal diameter are closely linked to the reasons why quasi Monte-Carlo methods, or low discrepancy sampling schemes are efficient techniques for integrating smooth functions. See [9] for a survey on these techniques.

It is minimax-optimal to stratify the domain in n strata and sample one point per stratum, but it would also be interesting to adapt the stratification of the space with respect to the function f. For example, if the function has larger variations in a region of the domain, we would like to discretize the domain in smaller strata in this region, so that more samples are assigned to this region. Since f is initially unknown, it is not possible to design a good stratification before sampling. However an efficient algorithm should allocate the samples in order to estimate online the variations of the

function in each region of the domain while, *at the same time*, allocating more samples in regions where f has larger local variations.
The papers [5, 7, 3] provide algorithms for solving a similar trade-off when the stratification is fixed: these algorithms allocate more samples to strata in which the function has larger variations. It is, however, clear that the larger the number of strata, the more difficult it is to allocate the samples almost optimally in the strata.

Contributions: We propose a new algorithm, Lipschitz Monte-Carlo Upper Confidence Bound (LMC-UCB), for tackling this problem. It is a two-layered algorithm. It first stratifies the domain in $K \ll n$ strata, and then allocates uniformly to each stratum an initial small amount of samples in order to estimate roughly the variations of the function per stratum. Then our algorithm sub-stratifies each of the K strata according to the estimated local variations, so that there are in total approximately n sub-strata, and allocates one point per sub-stratum. In that way, our algorithm discretizes the domain into more refined strata in regions where the function has higher variations. It cumulates the advantages of quasi Monte-Carlo and adaptive strategies.

More precisely, our contributions are the following:

- We prove an asymptotic lower bound on the mean squared error of the estimate returned by an optimal oracle strategy that has access to the variations of the function f everywhere and would use the best stratification of the domain with hyper-cubes (possibly of heterogeneous sizes). This quantity, since this is a lower-bound on any oracle strategies, is smaller than the mean squared error of the estimate provided by Uniform stratified Monte-Carlo (which is the non-adaptive minimax-optimal strategy on the class of differentiable functions), and also smaller than crude Monte-Carlo.
- We introduce the algorithm LMC-UCB, that sub-stratifies the K strata in hyper-cubic sub-strata, and samples one point per sub-stratum. The number of sub-strata per stratum is linked to the variations of the function in the stratum. We prove that algorithm LMC-UCB is asymptotically as efficient as the optimal oracle strategy. We also provide finite-time results when f admits a Taylor expansion of order 2 in every point. By tuning the number of strata K wisely, it is possible to build an algorithm that is almost as efficient as the optimal oracle strategy.

The paper is organized as follows. Section 2 defines the notations used throughout the paper. Section 3 states the asymptotic lower bound on the mean squared error of the optimal oracle strategy. In this Section, we also provide an intuition on how the number of samples into each stratum should be linked to the variation of the function in the stratum in order for the mean squared error of the estimate to be small. Section 4 presents the algorithm LMC-UCB and the first Lemma on how many sub-strata are built in the initial strata. Section 5 finally states that the algorithm LMC-UCB is almost as efficient as the optimal oracle strategy. We finally conclude the paper. Due to the lack of space, we also provide experiments and proofs in the Supplementary Material (see also [2]).

2 Setting

We consider a function $f : [0,1]^d \to \mathbb{R}$. We want to estimate as accurately as possible its integral according to the Lebesgue measure, i.e. $\int_{[0,1]^d} f(x)dx$. In order to do that, we consider algorithms that stratify the domain in two layers of strata, one more refined than the other. The strata of the refined layer are referred to as sub-strata, and we sample in the sub-strata. We will compare the performances of the algorithms we construct, with the performances of the optimal oracle algorithm that has access to the variations $||\nabla f(x)||_2$ of the function f everywhere in the domain, and is allowed to sample the domain where it wishes.

The first step is to partition the domain $[0,1]^d$ in K measurable *strata*. In this paper, we assume that $K^{1/d}$ is an integer[1]. This enables us to partition, in a natural way, the domain in K *hyper-cubic* strata $(\Omega_k)_{k\leq K}$ of same measure $w_k = \frac{1}{K}$. Each of these strata is a region of the domain $[0,1]^d$, and the K strata form a partition of the domain. We write $\mu_k = \frac{1}{w_k}\int_{\Omega_k} f(x)dx$ the mean and $\sigma_k^2 = \frac{1}{w_k}\int_{\Omega_k}\big(f(x)-\mu_k\big)^2 dx$ the variance of a sample of the function f when sampling f at a point chosen at random according to the Lebesgue measure conditioned to stratum Ω_k.

[1]This is not restrictive in small dimension, but it may become more constraining for large d.

We possess a budget of n samples (which is assumed to be known in advance), which means that we can sample n times the function at any point of $[0,1]^d$. We denote by $\mathcal{A}$ an algorithm that sequentially allocates the budget by sampling at round t in the stratum indexed by $k_t \in \{1,\dots,K\}$, and returns after all n samples have been used an estimate $\hat{\mu}_n$ of the integral of the function f.

We consider strategies that sub-partition each stratum Ω_k in hyper-cubes of same measure in Ω_k, but of heterogeneous measure among the Ω_k. In this way, the number of sub-strata in each stratum Ω_k can adapt to the variations f within Ω_k. The algorithms that we consider return a sub-partition of each stratum Ω_k in S_k sub-strata. We call $\mathcal{N}_k = (\Omega_{k,i})_{i\leq S_k}$ the sub-partition of stratum Ω_k. In each of these sub-strata, the algorithm allocates at least one point[2]. We write $X_{k,i}$ the first point sampled uniformly at random in sub-stratum $\Omega_{k,i}$. We write $w_{k,i}$ the measure of the sub-stratum $\Omega_{k,i}$. Let us write $\mu_{k,i} = \frac{1}{w_{k,i}}\int_{\Omega_{k,i}} f(x)dx$ the mean and $\sigma^2_{k,i} = \frac{1}{w_{k,i}}\int_{\Omega_{k,i}} \big(f(x)-\mu_{k,i}\big)^2 dx$ the variance of a sample of f in sub-stratum $\Omega_{k,i}$ (e.g. of $X_{k,i} = f(U_{k,i})$ where $U_{k,i} \sim \mathcal{U}_{\Omega_{k,i}}$).
This class of $2-$layered sampling strategies is rather large. In fact it contains strategies that are similar to low discrepancy strategies, and also to any stratified Monte-Carlo strategy. For example, consider that all K strata are hyper-cubes of same measure $\frac{1}{K}$ and that each stratum Ω_k is partitioned into S_k hyper-rectangles $\Omega_{k,i}$ of minimal diameter and same measure $\frac{1}{KS_k}$. If the algorithm allocates one point per sub-stratum, its sampling scheme shares similarities with quasi Monte-Carlo sampling schemes, since the points at which the function is sampled are well spread.

Let us now consider an algorithm that first chooses the sub-partition $(\mathcal{N}_k)_k$ and then allocates deterministically 1 sample uniformly at random in each sub-stratum $\Omega_{k,i}$. We consider the stratified estimate $\hat{\mu}_n = \sum_{k=1}^K \sum_{i=1}^{S_k} \frac{w_{k,i}}{S_k} X_{k,i}$ of μ. We have

$$\mathbb{E}(\hat{\mu}_n) = \sum_{k=1}^{K}\sum_{i=1}^{S_k} \frac{w_{k,i}}{S_k}\mu_{k,i} = \sum_{k\leq K}\sum_{i=1}^{S_k}\int_{\Omega_{k,i}} f(x)dx = \int_{[0,1]^d} f(x)dx = \mu,$$

and also

$$\mathbb{V}(\hat{\mu}_n) = \sum_{k\leq K}\sum_{i=1}^{S_k}(\frac{w_{k,i}}{S_k})^2\mathbb{E}(X_{k,i}-\mu_{k,i})^2 = \sum_{k\leq K}\sum_{i=1}^{S_k}\frac{w_{k,i}^2}{S_k^2}\sigma^2_{k,i}.$$

For a given algorithm $\mathcal{A}$ that builds for each stratum k a sub-partition $\mathcal{N}_k = (\Omega_{k,i})_{i\leq S_k}$, we call *pseudo-risk* the quantity

$$L_n(\mathcal{A}) = \sum_{k\leq K}\sum_{i=1}^{S_k}\frac{w_{k,i}^2}{S_k^2}\sigma^2_{k,i}. \tag{1}$$

Some further insight on this quantity is provided in the paper [4].

Consider now the uniform strategy, i.e. a strategy that divides the domain in $K = n$ hyper-cubic strata. This strategy is a fairly natural, minimax-optimal *static* strategy, on the class of differentiable function defined on $[0,1]^d$, when no information on f is available. We will prove in the next Section that its asymptotic mean squared error is equal to

$$\frac{1}{12}\Big(\int_{[0,1]^d} ||\nabla f(x)||_2^2 dx\Big)\frac{1}{n^{1+\frac{2}{d}}}.$$

This quantity is of order $n^{-1-2/d}$, which is smaller, as expected, than $1/n$: this strategy is more efficient than crude Monte-Carlo.

We will also prove in the next Section that the minimum asymptotic mean squared error of an optimal *oracle* strategy (we call it "oracle" because it builds the stratification using the information about the variations $||\nabla f(x)||_2$ of f in every point x), is larger than

$$\frac{1}{12}\Big(\int_{[0,1]^d} (||\nabla f(x)||_2)^{\frac{d}{d+1}} dx\Big)^{2\frac{(d+1)}{d}}\frac{1}{n^{1+\frac{2}{d}}}$$

This quantity is always smaller than the asymptotic mean squared error of the Uniform stratified Monte-Carlo strategy, which makes sense since this strategy assumes the knowledge of the variations of f everywhere, and can thus adapt accordingly the number of samples in each region. We define

$$\Sigma = \frac{1}{12}\Big(\int_{[0,1]^d} (||\nabla f(x)||_2)^{\frac{d}{d+1}} dx\Big)^{2\frac{(d+1)}{d}}. \tag{2}$$

[2]This implies that $\sum_k S_k \leq n$.

Given this minimum asymptotic mean squared error of an optimal oracle strategy, we define the pseudo-regret of an algorithm $\mathcal{A}$ as

$$R_n(\mathcal{A}) = L_n(\mathcal{A}) - \Sigma \frac{1}{n^{1+\frac{2}{d}}}. \tag{3}$$

This pseudo-regret is the difference between the pseudo-risk of the estimate provided by algorithm $\mathcal{A}$, and the lower-bound on the optimal oracle mean squared error. In other words, this pseudo-regret is the price an adaptive strategy pays for not knowing in advance the function f, and thus not having access to its variations. An efficient adaptive strategy should aim at minimizing this gap coming from the lack of informations.

3 Discussion on the optimal asymptotic mean squared error

3.1 Asymptotic lower bound on the mean squared error, and comparison with the Uniform stratified Monte-Carlo

A first part of the analysis of the exposed problem consists in finding a good point of comparison for the pseudo-risk. The following Lemma states an asymptotic lower bound on the mean squared error of the optimal oracle sampling strategy.

Lemma 1 *Assume that f is such that ∇f is continuous and $\int ||\nabla f(x)||_2^2 dx < \infty$. Let $\left((\Omega_k^n)_{k\leq n}\right)_n$ be an arbitrary sequence of partitions of $[0,1]^d$ in n strata such that all the strata are hyper-cubes, and such that the maximum diameter of each stratum goes to 0 as $n \to +\infty$ (but the strata are allowed to have heterogeneous measures).Let $\hat{\mu}_n$ be the stratified estimate of the function for the partition $(\Omega_k^n)_{k\leq n}$ when there is one point pulled at random per stratum. Then*

$$\lim \inf_{n\to\infty} n^{1+2/d}\mathbb{V}(\hat{\mu}_n) \geq \Sigma.$$

The full proof of this Lemma is in the Supplementary Material, Appendix B (see also [2]).

We have also the following equality for the asymptotic mean squared error of the uniform strategy.

Lemma 2 *Assume that f is such that ∇f is continuous and $\int ||\nabla f(x)||_2^2 dx < \infty$. For any $n = l^d$ such that l is an integer (and thus such that it is possible to partition the domain in n hyper-cubic strata of same measure), define $\left((\Omega_k^n)_{k\leq n}\right)_n$ as the sequence of partitions in hyper-cubic strata of same measure $1/n$. Let $\hat{\mu}_n$ be the stratified estimate of the function for the partition $(\Omega_k^n)_{k\leq n}$ when there is one point pulled at random per stratum. Then*

$$\lim \inf_{n\to\infty} n^{1+2/d}\mathbb{V}(\hat{\mu}_n) = \frac{1}{12}\Big(\int_{[0,1]^d} ||\nabla f(x)||_2^2 dx\Big).$$

The proof of this Lemma is substantially similar to the proof of Lemma 1 in the Supplementary Material, Appendix B (see also [2]). The only difference is that the measure of each stratum Ω_k^n is $1/n$ and that in Step 2, instead of Fatou's Lemma, the Theorem of dominated convergence is required.

The optimal rate for the mean squared error, which is also the rate of the Uniform stratified Monte-Carlo in Lemma 2, is $n^{-1-2/d}$ and is attained with ideas of low discrepancy sampling. The constant can however be improved (with respect to the constant in Lemma 2), by adapting to the specific shape of each function. In Lemma 1, we exhibit a lower bound for this constant (and without surprises, $\frac{1}{12}\Big(\int_{[0,1]^d} ||\nabla f(x)||_2^2 dx\Big) \geq \Sigma$). Our aim is to build an adaptive sampling scheme, also sharing ideas with low discrepancy sampling, that attains this lower-bound.

There is one main restriction in both Lemma: we impose that the sequence of partitions $\left((\Omega_k^n)_{k\leq n}\right)_n$ is composed only with strata that have the shape of an hyper-cube. This assumption is in fact reasonable: indeed, if the shape of the strata could be arbitrary, one could take the level sets (or approximate level sets as the number of strata is limited by n) as strata, and this would lead to $\lim_{n\to\infty} \inf_\Omega n^{1+2/d}\mathbb{V}(\hat{\mu}_{n,\Omega}) = 0$. But this is not a fair competition, as the function is unknown, and determining these level sets is actually a much harder problem than integrating the function.
The fact that the strata are hyper-cubes appears, in fact, in the bound. If we had chosen other shapes, e.g. l_2 balls, the constant $\frac{1}{12}$ in front of the bounds in both Lemma would change[3]. It is however not

[3]The $\frac{1}{12}$ comes from computing the variance of an uniform random variable on $[0,1]$.

possible to make a finite partition in l_2 balls of $[0,1]^d$, and we chose hyper-cubes since it is quite easy to stratify $[0,1]^d$ in hyper-cubic strata.

The proof of Lemma 1 makes the quantity $s^*(x) = \frac{(||\nabla f(x)||_2)^{\frac{d}{d+1}}}{\int_{[0,1]^d}(||\nabla f(u)||_2)^{\frac{d}{d+1}}du}$ appear. This quantity is proposed as "asymptotic optimal allocation", i.e. the asymptotically optimal number of sub-strata one would ideally create in any small sub-stratum centered in x. This is however not very useful for building an algorithm. The next Subsection provides an intuition on this matter.

3.2 An intuition of a good allocation: Piecewise linear functions

In this Subsection, we (i) provide an example where the asymptotic optimal mean squared error is also the optimal mean squared error at finite distance and (ii) provide explicitly what is, in that case, a good allocation. We do that in order to give an intuition for the algorithm that we introduce in the next Section.

We consider a partition in K hyper-cubic strata Ω_k. Let us assume that the function f is affine on all strata Ω_k, i.e. on stratum Ω_k, we have $f(x) = \Big(\langle \theta_k, x\rangle + \rho_k\Big)\mathbb{I}\{x \in \Omega_k\}$. In that case $\mu_k = f(a_k)$ where a_k is the center of the stratum Ω_k. We then have:

$$\sigma_k^2 = \frac{1}{w_k}\int_{\Omega_k}(f(x)-f(a_k))^2dx = \frac{1}{w_k}\int_{\Omega_k}\Big(\langle\theta_k,(x-a_k)\rangle\Big)^2dx = \frac{1}{w_k}\Big(\frac{||\theta_k||_2^2}{12}w_k^{1+2/d}\Big) = \frac{||\theta_k||_2^2}{12}w_k^{2/d}.$$

We consider also a sub-partition of Ω_k in S_k hyper-cubes of same size (we assume that $S_k^{1/d}$ is an integer), and we assume that in each sub-stratum $\Omega_{k,i}$, we sample one point. We also have $\sigma_{k,i}^2 = \frac{||\theta_k||_2^2}{12}\big(\frac{w_k}{S_k}\big)^{2/d}$ for sub-stratum $\Omega_{k,i}$.

For a given k and a given S_k, all the $\sigma_{k,i}$ are equals. The pseudo-risk of an algorithm $\mathcal{A}$ that divides each stratum Ω_k in S_k sub-strata is thus

$$L_n(\mathcal{A}) = \sum_{k\le K}\sum_{i\le S_k}\frac{w_k^2}{S_k^2}\frac{||\theta_k||_2^2}{12}\big(\frac{w_k}{S_k}\big)^{2/d} = \sum_{k\le K}\frac{w_k^{2+2/d}}{S_k^{1+2/d}}\frac{||\theta_k||_2^2}{12} = \sum_{k\le K}\frac{w_k^2}{S_k^{1+2/d}}\sigma_k^2.$$

If an unadaptive algorithm $\mathcal{A}^*$ has access to the variances σ_k^2 in the strata, it can choose to allocate the budget in order to minimize the pseudo-risk. After solving the simple optimization problem of minimizing $L_n(\mathcal{A})$ with respect to $(S_k)_k$, we deduce that an optimal oracle strategy on this stratification would divide each stratum k in $S_k^* = \frac{(w_k\sigma_k)^{\frac{d}{d+1}}}{\sum_{i\le K}(w_i\sigma_i)^{\frac{d}{d+1}}}n$ sub-strata[4]. The pseudo-risk for this strategy is then

$$L_{n,K}(\mathcal{A}^*) = \frac{\Big(\sum_{k\le K}(w_k\sigma_k)^{\frac{d}{d+1}}\Big)^{2\frac{(d+1)}{d}}}{n^{1+2/d}} = \frac{\Sigma_K^{2\frac{(d+1)}{d}}}{n^{1+2/d}}, \tag{4}$$

where we write $\Sigma_K = \sum_{i\le K}(w_i\sigma_i)^{\frac{d}{d+1}}$. We will call in the paper *optimal proportions* the quantities

$$\lambda_{K,k} = \frac{(w_k\sigma_k)^{\frac{d}{d+1}}}{\sum_{i\le K}(w_i\sigma_i)^{\frac{d}{d+1}}}. \tag{5}$$

In the specific case of functions that are piecewise linear, we have $\Sigma_K = \sum_{k\le K}(w_k\sigma_k)^{\frac{d}{d+1}} = \sum_{k\le K}(w_k\frac{||\theta_k||_2}{2\sqrt{3}}w_k^{1/d})^{\frac{d}{d+1}} = \int_{[0,1]^d}\frac{(||\nabla f(x)||_2)^{\frac{d}{d+1}}}{12^{\frac{d}{2(d+1)}}}dx$. We thus have

$$L_{n,K}(\mathcal{A}^*) = \Sigma\frac{1}{n^{1+\frac{2}{d}}}. \tag{6}$$

This optimal oracle strategy attains the lower bound in Lemma 1. We will thus construct, in the next Section, an algorithm that learns and adapts to the optimal proportions defined in Equation 5.

[4]We deliberately forget about rounding issues in this Subsection. The allocation we provide might not be realizable (e.g. if S_k^* is not an integer), but plugging it in the bound provides a lower bound on any realizable performance.

4 The Algorithm LMC-UCB

4.1 Algorithm LMC-UCB

We present the algorithm Lipschitz Monte Carlo Upper Confidence Bound ($LMC-UCB$). It takes as parameter a partition $(\Omega_k)_{k\leq K}$ in $K \leq n$ hyper-cubic strata of same measure $1/K$ (it is possible since we assume that $\exists l \in \mathbb{N}/l^d = K$). It also takes as parameter an uniform upper bound L on $||\nabla f(x)||_2^2$, and δ, a (small) probability. The aim of algorithm $LMC-UCB$ is to sub-stratify each stratum Ω_k in $\lambda_{K,k} = \frac{(w_k\sigma_k)^{\frac{d}{d+1}}}{\sum_{i=1}^K (w_i\sigma_i)^{\frac{d}{d+1}}} n$ hyper-cubic sub-strata of same measure and sample one point per sub-stratum. An intuition on why this target is relevant was provided in Section 3.

Algorithm LMC-UCB starts by sub-stratifying each stratum Ω_k in $\bar{S} = \left\lfloor \left(\left(\frac{n}{K}\right)^{\frac{d}{d+1}} \right)^{1/d} \right\rfloor^d$ hyper-cubic strata of same measure. It is possible to do that since by definition, $\bar{S}^{1/d}$ is an integer. We write this first sub-stratification $\mathcal{N}'_k = (\Omega'_{k,i})_{i\leq \bar{S}}$. It then pulls one sample per sub-stratum in $\mathcal{N}'_k$ for each Ω_k.

It then sub-stratifies again each stratum Ω_k using the informations collected. It sub-stratifies each stratum Ω_k in

$$S_k = \max\left\{ \left\lfloor \frac{w_k^{\frac{d}{d+1}} \left(\hat{\sigma}_{k,K\bar{S}} + A(\frac{w_k}{\bar{S}})^{1/d} \sqrt{\frac{1}{\bar{S}}} \right)^{\frac{d}{d+1}}}{\sum_{i=1}^K w_i^{\frac{d}{d+1}} \left(\hat{\sigma}_{i,K\bar{S}} + A(\frac{w_i}{\bar{S}})^{1/d} \sqrt{\frac{1}{\bar{S}}} \right)^{\frac{d}{d+1}}} (n - K\bar{S}) \right\rfloor^{1/d} \right\rfloor^d, \bar{S} \right\} \tag{7}$$

hyper-cubic strata of same measure (see Figure 1 for a definition of A). It is possible to do that because by definition, $S_k^{1/d}$ is an integer. We call this sub-stratification of stratum Ω_k stratification $\mathcal{N}_k = (\Omega_{k,i})_{i\leq S_k}$. In the last Equation, we compute the empirical standard deviation in stratum Ω_k at time $K\bar{S}$ as

$$\hat{\sigma}_{k,K\bar{S}} = \sqrt{\frac{1}{\bar{S}-1} \sum_{i=1}^{\bar{S}} \left(X_{k,i} - \frac{1}{\bar{S}} \sum_{j=1}^{\bar{S}} X_{k,j} \right)^2}. \tag{8}$$

Algorithm LMC-UCB then samples in each sub-stratum $\Omega_{k,i}$ one point. It is possible to do that since, by definition of S_k, $\sum_k S_k + K\bar{S} \leq n$

The algorithm outputs an estimate $\hat{\mu}_n$ of the integral of f, computed with the first point in each sub-stratum of partition $\mathcal{N}_k$. We present in Figure 1 the pseudo-code of algorithm LMC-UCB.

Input: Partition $(\Omega_k)_{k\leq K}$, L, δ, set $A = 2L\sqrt{d}\sqrt{\log(2K/\delta)}$
Initialize: $\forall k \leq K$, sample 1 point in each stratum of partition $\mathcal{N}'_k$
Main algorithm:
Compute S_k for each $k \leq K$
Create partition $\mathcal{N}_k$ for each $k \leq K$
Sample a point in $\Omega_{k,i} \in \mathcal{N}_k$ for $i \leq S_k$
Output: Return the estimate $\hat{\mu}_n$ computed when taking the first point $X_{k,i}$ in each sub-stratum $\Omega_{k,i}$ of $\mathcal{N}_k$, that is to say $\hat{\mu}_n = \sum_{k=1}^K w_k \sum_{i=1}^{S_k} \frac{X_{k,i}}{S_k}$

Figure 1: Pseudo-code of LMC-UCB. The definition of $\mathcal{N}'_k$, $\bar{S}$, $\mathcal{N}_k$, $\Omega_{k,i}$ and S_k are in the main text.

4.2 High probability lower bound on the number of sub-strata of stratum Ω_k

We first state an assumption on the function f.

Assumption 1 *The function f is such that ∇f exists and $\forall x \in [0,1]^d$, $||\nabla f(x)||_2^2 \leq L$.*

The next Lemma states that with high probability, the number S_k of sub-strata of stratum Ω_k, in which there is at least one point, adjusts "almost" to the unknown optimal proportions.

Lemma 3 *Let Assumption 1 be satisfied and $(\Omega_k)_{k\leq K}$ be a partition in K hyper-cubic strata of same measure. If $n \geq 4K$, then with probability at least $1-\delta$, $\forall k$, the number of sub-strata satisfies*

$$S_k \geq \max\left[\lambda_{K,k} \left[n - 7(L+1)d^{3/2}\sqrt{\log(K/\delta)}(1 + \frac{1}{\Sigma_K}) K^{\frac{1}{d+1}} n^{\frac{d}{d+1}} \right], \bar{S} \right].$$

The proof of this result is in the Supplementary Material (Appendix C) (see also [2]).

4.3 Remarks

A sampling scheme that shares ideas with quasi Monte-Carlo methods: Algorithm $LMC-UCB$ *almost* manages to divide each stratum Ω_k in $\lambda_{K,k}n$ hyper-cubic strata of same measure, each one of them containing at least one sample. It is thus possible to build a learning procedure that, at the same time, estimates the empirical proportions $\lambda_{K,k}$, and allocates the samples proportionally to them.

The error terms: There are two reasons why we are not able to divide *exactly* each stratum Ω_k in $\lambda_{K,k}n$ hyper-cubic strata of same measure. The first reason is that the true proportions $\lambda_{K,k}$ are unknown, and that it is thus necessary to estimate them. The second reason is that we want to build strata that are hyper-cubes of same measure. The number of strata S_k needs thus to be such that $S_k^{1/d}$ is an integer. We thus also loose efficiency because of rounding issues.

5 Main results

5.1 Asymptotic convergence of algorithm LMC-UCB

By just combining the result of Lemma 1 with the result of Lemma 3, it is possible to show that algorithm LMC-UCB is asymptotically (when K goes to $+\infty$ and $n \geq K$) as efficient as the optimal oracle strategy of Lemma 1.

Theorem 1 *Assume that ∇f is continuous, and that Assumption 1 is satisfied. Let $(\Omega_k^n)_{n,k\leq K_n}$ be an arbitrary sequence of partitions such that all the strata are hyper-cubes, such that $4K_n \leq n$, such that the diameter of each strata goes to 0, and such that $\lim_{n\to+\infty} \frac{1}{n}\Big(K_n\big(\log(K_n n^2)\big)^{\frac{d+1}{2}}\Big) = 0$. The regret of LMC-UCB with parameter $\delta_n = \frac{1}{n^2}$ on this sequence of partition, where for sequence $(\Omega_k^n)_{n,k\leq K_n}$ it disposes of n points, is such that*

$$\lim_{n\to\infty} n^{1+2/d} R_n(\mathcal{A}_{LMC-UCB}) = 0.$$

The proof of this result is in the Supplementary Material (Appendix D) (see also [2]).

5.2 Under a slightly stronger Assumption

We introduce the following Assumption, that is to say that f admits a Taylor expansion of order 2.

Assumption 2 *f admits a Taylor expansion at the second order in any point $a \in [0,1]^d$ and this expansion is such that $\forall x, |f(x) - f(a) - \langle \nabla f, (x-a)\rangle| \leq M||x-a||_2^2$ where M is a constant.*

This is a slightly stronger assumption than Assumption 1, since it imposes, additional to Assumption 1, that the variations of $\nabla f(x)$ are uniformly bounded for any $x \in [0,1]^d$. Assumption 2 implies Assumption 1 since $\big|||\nabla f(x)||_2 - ||\nabla f(0)||_2\big| \leq M||x-0||_2$, which implies that $||\nabla f(x)||_2 \leq ||\nabla f(0)||_2 + M\sqrt{d}$. This implies in particular that we can consider $L = ||\nabla f(0)||_2 + M\sqrt{d}$. We however do not need M to tune the algorithm LMC-UCB, as long as we have access to L (although M appears in the bound of next Theorem).

We can now prove a bound on the pseudo-regret.

Theorem 2 *Under Assumptions 1 and 2, if $n \geq 4K$, the estimate returned by algorithm $LMC-UCB$ is such that, with probability $1-\delta$, we have*

$$R_n(\mathcal{A}_{LMC-UCB}) \leq \frac{1}{n^{\frac{d+2}{d}}}\Big[M(L+1)^4\Big(1+\frac{3Md}{\Sigma}\Big)^4\Big(650d^{3/2}\sqrt{\log(K/\delta)}K^{\frac{1}{d+1}}n^{-\frac{1}{d+1}} + 25d\big(\frac{1}{K}\big)^{\frac{1}{d+1}}\Big)\Big].$$

A proof of this result is in the Supplementary Material (Appendix E) (see also [2]).

Now we can choose optimally the number of strata so that we minimize the regret.

Theorem 3 *Under Assumptions 1 and 2, the algorithm $LMC-UCB$ launched on $K_n = \Big\lfloor(\sqrt{n})^{1/d}\Big\rfloor^d$ hyper-cubic strata is such that, with probability $1-\delta$, we have*

$$R_n(\mathcal{A}_{LMC-UCB}) \leq \frac{1}{n^{1+\frac{2}{d}+\frac{1}{2(d+1)}}}\Big[700M(L+1)^4 d^{3/2}\Big(1+\frac{3Md}{\Sigma}\Big)^4\sqrt{\log(n/\delta)}\Big].$$

5.3 Discussion

Convergence of the algorithm LMC-UCB to the optimal oracle strategy: When the number of strata K_n grows to infinity, but such that $\lim_{n\to+\infty} \frac{1}{n}\Big(K_n\big(\log(K_n n^2)\big)^{\frac{d+1}{2}}\Big) = 0$, the pseudo-regret of algorithm LMC-UCB converges to 0. It means that this strategy is asymptotically as efficient as (the lower bound on) the optimal oracle strategy. When f admits a Taylor expansion at the first order in every point, it is also possible to obtain a finite-time bound on the pseudo-regret.

A new sampling scheme: The algorithm $LMC-UCB$ samples the points in a way that takes advantage of both stratified sampling and quasi Monte-Carlo. Indeed, LMC-UCB is designed to cumulate (i) the advantages of quasi Monte-Carlo by spreading the samples in the domain and (ii) the advantages of stratified, adaptive sampling by allocating more samples where the function has larger variations. For these reasons, this technique is efficient on differentiable functions. We illustrate this assertion by numerical experiments in the Supplementary Material (Appendix A) (see also [2]).

In high dimension: The bound on the pseudo-regret in Theorem 3 is of order $n^{-1-\frac{2}{d}} \times poly(d)n^{-\frac{1}{2(d+1)}}$. In order for the pseudo-regret to be negligible when compared to the optimal oracle mean squared error of the estimate (which is of order $n^{-1-\frac{2}{d}}$) it is necessary that $poly(d)n^{-\frac{1}{2(d+1)}}$ is negligible compared to 1. In particular, this says that n should scale exponentially with the dimension d. This is unavoidable, since stratified sampling shrinks the approximation error to the asymptotic oracle only if the diameter of each stratum is small, i.e. if the space is stratified in every direction (and thus if n is exponential with d). However Uniform stratified Monte-Carlo, also for the same reasons, shares this problem[5].

We emphasize however the fact that a (slightly modified) version of our algorithm is more efficient than crude Monte-Carlo, up to a negligible term *that depends only of* $poly(\log(d))$. The bound in Lemma 3 depends of $poly(d)$ only because of rounding issues, coming from the fact that we aim at dividing each stratum Ω_k in hyper-cubic sub-strata. The whole budget is thus not completely used, and only $\sum_k S_k + K\bar{S}$ samples are collected. By modifying LMC-UCB so that it allocates the remaining budget uniformly at random on the domain, it is possible to prove that the (modified) algorithm is always at least as efficient as crude Monte-Carlo.

Conclusion

This work provides an adaptive method for estimating the integral of a differentiable function f. We first proposed a benchmark for measuring efficiency: we proved that the asymptotic mean squared error of the estimate outputted by the optimal oracle strategy is lower bounded by $\Sigma\frac{1}{n^{1+2/d}}$. We then proposed an algorithm called LMC-UCB, which manages to learn the amplitude of the variations of f, to sample more points where theses variations are larger, and to spread these points in a way that is related to quasi Monte-Carlo sampling schemes. We proved that algorithm LMC-UCB is asymptotically as efficient as the optimal, oracle strategy. Under the assumption that f admits a Taylor expansion in each point, we provide also a finite time bound for the pseudo-regret of algorithm LMC-UCB. We summarize in Table 1 the rates and finite-time bounds for crude Monte-Carlo, Uniform stratified Monte-Carlo and LMC-UCB. An interesting extension of this work would be to

	Pseudo-Risk:		
Sampling schemes	Rate	Asymptotic constant	+ Finite-time bound
Crude MC	$\frac{1}{n}$	$\int_{[0,1]^d}\big(f(x)-\int_{[0,1]^d} f(u)du\big)^2 dx$	$+0$
Uniform stratified MC	$\frac{1}{n^{1+\frac{2}{d}}}$	$\frac{1}{12}\Big(\int_{[0,1]^d} \|\|\nabla f(x)\|\|_2^2 dx\Big)$	$+O(\frac{d}{n^{1+\frac{2}{d}+\frac{1}{2d}}})$
LMC-UCB	$\frac{1}{n^{1+\frac{2}{d}}}$	$\frac{1}{12}\Big(\int_{[0,1]^d}(\|\|\nabla f(x)\|\|_2)^{\frac{d}{d+1}} dx\Big)^{2\frac{(d+1)}{d}}$	$+O(\frac{d^{\frac{11}{2}}}{n^{1+\frac{2}{d}+\frac{1}{2(d+1)}}})$

Table 1: Rate of convergence plus finite time bounds for Crude Monte-Carlo, Uniform stratified Monte Carlo (see Lemma 2) and LMC-UCB (see Theorems 1 and 3).

adapt it to $\alpha-$Hölder functions that admit a Riemann-Liouville derivative of order α. We believe that similar results could be obtained, with an optimal constant and a rate of order $n^{1+2\alpha/d}$.

Acknowledgements This research was partially supported by Nord-Pas-de-Calais Regional Council, French ANR EXPLO-RA (ANR-08-COSI-004), the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement 270327 (project CompLACS), and by Pascal-2.

[5]When d is very large and n is not exponential in d, then second order terms, depending on the dimension, take over the bound in Lemma 2 (which is an asymptotic bound) and $poly(d)$ appears in these negligible terms.

References

[1] J.Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[2] A. Carpentier and R. Munos. Adaptive Stratified Sampling for Monte-Carlo integration of Differentiable functions. Technical report, *arXiv:0575985*, 2012.

[3] A. Carpentier and R. Munos. Finite-time analysis of stratified sampling for monte carlo. In *In Neural Information Processing Systems (NIPS)*, 2011a.

[4] A. Carpentier and R. Munos. Finite-time analysis of stratified sampling for monte carlo. Technical report, INRIA-00636924, 2011b.

[5] Pierre Etoré and Benjamin Jourdain. Adaptive optimal allocation in stratified sampling methods. *Methodol. Comput. Appl. Probab.*, 12(3):335–360, September 2010.

[6] P. Glasserman. *Monte Carlo methods in financial engineering*. Springer Verlag, 2004. ISBN 0387004513.

[7] V. Grover. Active learning and its application to heteroscedastic problems. *Department of Computing Science, Univ. of Alberta, MSc thesis*, 2009.

[8] A. Maurer and M. Pontil. Empirical bernstein bounds and sample-variance penalization. In *Proceedings of the Twenty-Second Annual Conference on Learning Theory*, pages 115–124, 2009.

[9] H. Niederreiter. Quasi-monte carlo methods and pseudo-random numbers. *Bull. Amer. Math. Soc*, 84(6):957–1041, 1978.

[10] R.Y. Rubinstein and D.P. Kroese. *Simulation and the Monte Carlo method*. Wiley-interscience, 2008. ISBN 0470177942.