Comparison of independent and stratified sampling schemes in problems of global optimization

Marina Kondratovich Anatoly A. Zhigljavsky

University of St.Petersburg, Russia

Abstract

Let $\mathcal{X}$ be a compact subset of R^d and $\{x_1, \ldots, x_N\}$ be a sample of random points in $\mathcal{X}$. It is well–known that any properly organized stratified sampling procedure is superior to the independent sampling with respect to the variance of Monte Carlo estimates of integrals of functions in $L_2(\mathcal{X})$. We prove similar results for some perfomance characteristics important in global optimization. We also demonstrate that the stratified sample with the maximum stratification is optimum, in a suitable sense.

1 Introduction

Let $(\mathcal{X}, \mathcal{B}, P)$ be a measure space, $\mathcal{X}$ be a compact subset of $\mathbf{R}^d$, $\mathcal{B}$ be the σ-algebra of Borel subsets of X and P be a probability measure on $(\mathcal{X}, \mathcal{B})$, we call it *uniform distribution.* Let also $\mathcal{F} \subseteq C(\mathcal{X})$ be a functional class, N be a fixed number, $\Xi = (x_1, x_2, ..., x_N) \in \mathcal{X}^N$, $M[f, \Xi] = \max_{x_i \in \Xi} f(x_i)$.

If Ξ is a random vector in $\mathcal{X}^N$ with certain distribution $Q(\mathrm{d}\Xi)$, then the ordered pair $\Pi = (M[f, \Xi], Q)$ is called *random search procedure* for the global maximum of $f \in F$.

Consider a partition $\mathcal{P}_m$ of $\mathcal{X}$ into m disjoint connected subsets of positive measure:

$$\mathcal{P}_m : \mathcal{X} = \bigcup_{i=1}^{m} \mathcal{X}_i,\ \mathcal{X}_i \in \mathcal{B},\ q_i = P(X_i) > 0 \text{ for } i = 1, \ldots, m,\ \ \mathcal{X}_i \cap \mathcal{X}_j = \emptyset \text{ for } i \neq j.$$

Since P is a probability measure, $\sum_{i=1}^{m} q_i = 1$. Define the uniform probability measure P_i on $\mathcal{X}_i$ by $P_i(A) = P(A \cap \mathcal{X}_i)/q_i$ for every $A \in \mathcal{B}$.

Given a partition $\mathcal{P}_m$ and a collection of integers $L = \{l_1, \ldots, l_m\}$ such that $\sum_{i=1}^{m} l_i = N$, the *stratified sample* $\Xi_{m,L}$ can be defined as

$$\Xi_{m,L} = (x_{1,1}, \ldots, x_{1,l_1}, \ldots, x_{m,1}, \ldots, x_{m,l_m}) \quad (1)$$

where for every $i=1,\ldots,m,\ \ x_{i,1},\ldots,x_{i,l_i}$ are independent random variables with the uniform distribution P_i on $\mathcal{X}_i$. (In practice, an additional randomization on the order of generation of $x_{i,j}$ is sometimes useful as well.)

We shall call the stratified sample (1) *proper stratified sample* if the number of points in $\mathcal{X}_i$ is proportional to $q_i = P(\mathcal{X}_i)$, that is,

$$l_i = Nq_i \quad \text{for all } \ i = 1,\ldots,m\,. \tag{2}$$

The joint distribution of the random vector (1) is

$$Q_{m,L}(d\,\Xi_{m,L}) = \prod_{i_1=1}^{l_1} P_1(dx_{1,i_1}) \times \ldots \times \prod_{i_m=1}^{l_m} P_m(dx_{m,i_m})\,.$$

The random search procedure $\Pi_{m,L} = (M[f,\Xi_{m,L}], Q_{m,L})$ with $m>1$ corresponds to the stratified sampling on $\mathcal{X}$, and $\Pi_1 = \Pi_{1,N} = (M[f,\Xi_{1,N}], Q_{1,N})$ corresponds to the independent sampling from the distribution P.

For a fixed $f \in \mathcal{F}$, let $\Psi_f(\Pi)$ be a criterion for comparison of procedures Π. In line with the general concept of domination, we say that Π dominates Π' in $\mathcal{F}$ if $\Psi_f(\Pi) \le \Psi_f(\Pi')$ for every $f \in \mathcal{F}$ and there exists a function $f_* \in \mathcal{F}$ such that $\Psi_{f_*}(\Pi) < \Psi_{f_*}(\Pi')$.

Below we consider two related dominance criteria: (i) the cumulative distribution function (c.d.f.) of the record value $M[f,\Xi] = \max_{x_i \in \Xi} f(x_i)$ achieved at the sample points, and (ii) k-th moment of the difference $\max_{\mathcal{X}} f - M[f,\Xi]$, for every $k>0$. Some other criteria are studied in [1]. The results of the paper confirm the assertion of [1] that the "reduction of randomness" and "increase of uniformity" improves the performance of Monte Carlo procedures for global optimization.

2 Stochastic dominance with respect to the record value

Let us consider the stochastic dominance when the criterion $\Psi_f(\Pi)$ is the c.d.f. of the record value $M[f,\Xi] = \max_{x_i \in \Xi} f(x_i)$:

$$F_{f,\Pi}(t) = P(M[f,\Xi] \le t), \ \ t \in (\min f, \max f). \tag{3}$$

In this case, the dominance of a procedure Π over Π' in $\mathcal{F}$ means that $F_{f,\Pi}(t) \le F_{f,\Pi'}(t)$ for all real t and $f \in \mathcal{F}$ and there exists $f_* \in \mathcal{F}$ such that $F_{f_*,\Pi}(t) < F_{f_*,\Pi'}(t)$ for all $t \in (\min f_*, \max f_*)$.

Theorem 1. Let $\mathcal{P}_m$ be a fixed partition of $\mathcal{X}$ into $m \le N$ subsets, $\mathcal{F} = C^p(\mathcal{X})$ for some $0 \le p \le \infty$ and $\Pi_{m,L} = (M[f,\Xi_{m,L}], Q_{m,L})$ be a stratified sampling random search procedure such that $L=\{l_1,\ldots,l_m\}$, $l_i \ge 0$, $\sum_{i=1}^m l_i = N$. Then

(i) if the stratified sample $\Xi_{m,L}$ is proper, that is (2) holds, then the stratified sampling random search procedure $\Pi_{m,L}$ stochastically dominates the independent random sampling procedure Π_1 in $\mathcal{F}$, with respect to the criterion (3);

(ii) if (2) does not hold for at least one i, then $\Pi_{m,L}$ does not stochastically dominate Π_1: there exists $f^* \in \mathcal{F}$ such that for some t $F_{m,L}(f_*, t) > F_1(f_*, t)$ where $F_{m,L}(f, t) = F_{f,\Pi_{m,L}}(t)$ and $F_1(f, t) = F_{f,\Pi_1}(t)$ are the c.d.f. (3) for the stratified and independent sampling procedures, respectively.

Proof. Let f be an arbitrary function in $\mathcal{F}$. Then the c.d.f. $F_1(f, t)$ for the independent sampling procedure $\Pi_1 = (M[f, \Xi_{1,N}], Q_{1,N})$ can be rewritten as

$$F_1(f,t) = P(f(x_{1,j}) \le t, j = 1, \ldots, N) = P^N(f(x_{1,j}) \le t) = P^N(A_t)$$

where $A_t = f^{-1}((-\infty, t])$ is the inverse image of the set $(-\infty, t]$. Since $\{\mathcal{X}_i\}_{i=1}^m$ is a complete system of events, we have

$$P(A_t) = \sum_{i=1}^{m} P(A_t \cap \mathcal{X}_i) = \sum_{i=1}^{m} \beta_i$$

where

$$\beta_i = P(A_t \cap \mathcal{X}_i), \;\; i = 1, \ldots, m, \quad \sum_{i=1}^{m} \beta_i = P(A_t) \le 1.$$

We thus have

$$F_1(f,t) = \left(\sum_{i=1}^{m} \beta_i \right)^N .$$

For the stratified sampling procedure $\Pi_{m,L} = (M[f, \Xi_{m,L}], Q_{m,L})$ the c.d.f. $F_{m,L}(f, t)$ can be analogously rewritten as

$$F_{m,L}(f,t) = P(f(x_{1,1}) \le t, \ldots, f(x_{1,l_1}) \le t, \ldots, f(x_{m,l_m}) \le t) =$$

$$\prod_{i=1}^{m} P_i^{l_i}(f(x_{i,j}) \le t) = \prod_{i=1}^{m} (P(\{f(x_{i,j}) \le t\} \cap \mathcal{X}_i)/q_i)^{l_i} = \prod_{i=1}^{m} \left(\frac{\beta_i}{q_i} \right)^{l_i} .$$

For every $i = 1, \ldots, m$, set

$$\gamma_i = \frac{l_i}{N}, \quad \alpha_i = \frac{\beta_i}{q_i} = P(A_t \cap X_i)/P(X_i).$$

Then

$$0 < \gamma_i < 1, \quad 0 \le \alpha_i \le 1 \quad \text{for} \;\; i = 1, \dots, m, \;\; \sum_{i=1}^{m} \gamma_i = 1$$

and the vector $\alpha = (\alpha_1, \dots, \alpha_m)$ may get any value in the interior of the cube $[0, 1]^m$ depending on f and t.

The representations for the c.d.f. $F_{m,L}(f, t)$ and $F_1(f, t)$ and (4) imply that the following two inequalities are equivalent:

$$F_{m,L}(f, t) \le F_1(f, t) \quad \Longleftrightarrow \quad \prod_{i=1}^{m} \alpha_i^{\gamma_i} \le \sum_{i=1}^{m} q_i \alpha_i$$

which we rewrite in a more convenient form

$$F_{m,L}(f, t) \le F_1(f, t) \quad \Longleftrightarrow \quad \sum_{i=1}^{m} \gamma_i \log \alpha_i \le \log \left(\sum_{i=1}^{m} q_i \alpha_i \right) \tag{4}$$

Analogous equivalence takes place when the sign $\le$ in (4) is substituted for the strict inequality sign.

Let us now prove (i). If (2) holds then $\gamma_i = q_i$ for all $i = 1, \dots, m$ and the validity of the second inequality in (4), for every $\alpha \in [0, 1]^m$ and thus for every $f \in \mathcal{F}$, follows from the concavity of the logarithm. Consider a function $f_* \in \mathcal{F}$ such that $0 \le f \le 1$, $f_*(x) = 0$ for all $x \in \mathcal{X}_1$ and $\max_{x \in \mathcal{X}_2} = 1$. Then

$$\alpha_1 = P(A_t \cap X_1)/P(X_1) = 1 \;\; \text{and} \;\; \alpha_2 = P(A_t \cap X_2)/P(X_2) < 1$$

for all $t \in (0, 1) = (\min f, \max f)$. Therefore the values α_i are not all equal each other and the strict concavity of the logarithm implies the strict inequality in (4).

Let us now turn to (ii). Assume that (2) does not hold. Then there exists $i_0 \le m$ such that $\gamma_{i_0} < q_{i_0}$. Consider a function $f^* \in \mathcal{F}$ such that $f^*(x) = 0$ for all $x \in \mathcal{X} \setminus \mathcal{X}_{i_0}$ and $\max_{x \in \mathcal{X}_{i_0}} = 1$. Then $\alpha_j = 1$ for all $j \ne i_0$ and α_{i_0} gets all values in $(0, 1)$ depending on t.

Let us show that for the function f^* the inequality

$$\sum_{i=1}^{m} \gamma_i \log \alpha_i > \log \left(\sum_{i=1}^{m} q_i \alpha_i \right) \tag{5}$$

holds for all sufficiently large $\alpha_{i_0} < 1$. Denote $\varepsilon = 1 - \alpha_{i_0} > 0$ and rewrite the inequality (5) as $h(\varepsilon) > 0$ where

$$h(\varepsilon) = \gamma_{i_0} \log(1 - \varepsilon) - \log(1 - q_{i_0} \varepsilon) \,.$$

At $\varepsilon = 0$ we have $h(0) = 0$ and $h'(0) = -\gamma_{i_0} + q_{i_0} > 0$, where we have used the fact that $\gamma_{i_0} < q_{i_0}$. This implies $h(\varepsilon) > 0$ for all sufficiently small

$\varepsilon > 0$ and therefore the validity of (5). In its turn, it yields that the inequality $F_{m,L}(f^*, t) > F_1(f^*, t)$ holds for all t sufficiently close to 1. ∎

Corollary. Analogously to (i) in Theorem 1 we can easily get that if $m' < m$, $\mathcal{P}_m$ is a subpartition of a partition $\mathcal{P}_{m'}$, $\Xi_{m,L}$ is a proper stratified sample and $\Pi_{m,L}$ and $\Pi_{m',L'}$ are the random search procedures, corresponding to the stratified samples $\Xi_{m,L}$ and $\Xi_{m',L'}$, then $\Pi_{m,L}$ stochastically dominates $\Pi_{m',L'}$ in $\mathcal{F} = C^p(\mathcal{X})$ for every $0 \le p \le \infty$. This particularly implies that the stratified sample $\Xi_{m,L}$ with the maximum stratification, that is, when $P(\mathcal{X}_i) = 1/m$ and $L = (1, \ldots, 1)$, generates the best possible random search procedure $\Pi_{m,L}$, with respect to the stochastic dominance based on the c.d.f. (3).

3 Asymptotic criteria

In the present section we only consider proper stratified sampling procedures $\Pi_{m,l} = \Pi_{m,L}$ where $P(\mathcal{X}_i) = 1/m$ and $l_i = l$ for all $i = 1, \ldots, m$. We also assume that $N = ml$, l=const, $m \to +\infty$, that is, the number of subsets in the partition $\mathcal{P}_m$ tends to infinity but the number of points in each subset stays constant.

As the criteria for comparison of procedures, consider now k-th moment of the random variable $(M[f] - M[f, \Xi_{m,l}])$ where $M[f] = \max_{x \in \mathcal{X}} f(x)$ and $\Xi_{m,l} = \Xi_{m,L}$ for $L = (l, \ldots, l)$:

$$\Psi_f(m, l) = \mathbf{E}(M[f] - M[f, \Xi_{m,l}])^k, \quad k > 0\,. \tag{6}$$

Theorem 1 implies that the stratified sampling procedure $\Pi_{m,l}$ is superior to the independent sampling procedure $\Pi_{1,N}$ with respect to the criteria (6), for every $k > 0$. Theorem 2 below establishes the qualitative result concerning this superiority.

Let the class of distributions $\{P\}$ and the functional class $\mathcal{F}_*$ of continuous functions $f(x) = \varphi(x - x_*)$ with a unique point of the global maximum $x_*(f)$, at this point $f(x_*) = M[f]$, satisfy the following two conditions.

Condition A. For the c.d.f. $F_f(t) = \int_{f(x) \le t} P(dx)$, the function $V_f(v) = 1 - F_f(M[f] - 1/v), v > 0$, regularly varies at infinity with some index $-\alpha < 0$, that is, $\lim_{v \to \infty} V_f(uv)/V_f(v) = u^{-\alpha}$ for all $u > 0$.

Condition B. The point x_* has a certain distribution $R(dx)$ on $(\mathcal{X}, \mathcal{B})$ which is equivalent to the Lebesgue measure on $(\mathcal{X}, \mathcal{B})$.

The above conditions have been introduced and studied in [1,2]. It is shown in these works that conditions A and B hold in a rather general setup. Condition A is satisfied, for example, when $\mathcal{X}$ is a compact subset of R^d, the measure P is equivalent to the Lebesgue measure in some neighbourhood of x_*, $\nabla\varphi(0) = 0$, and the matrix $\nabla^2\varphi(0)$ is non-singular, in this case $\alpha = d/2$.

The following theorem generalizes to arbitrary $k > 0$ the result of [1] for $k = 1, 2$.

Theorem 2. Assume that the conditions A and B are satisfied, $N = ml$, l=const, $m \to \infty$. Then for every $k > 0$ with R-probability 1

$$\frac{\Psi_f(m,l)}{\Psi_f(1,N)} = \frac{\mathbf{E}(M[f] - M[f, \Xi_{m,l}])^k}{\mathbf{E}(M[f] - M[f, \Xi_{1,N}])^k} = r(l,k,\alpha) + o(1), \quad N \to +\infty, \tag{7}$$

where

$$r(l,k,\alpha) = \frac{l^{k/\alpha}\Gamma(l+1)}{\Gamma(k/\alpha + l + 1)}, \tag{8}$$

and $\Gamma(\cdot)$ is the gamma function. Moreover, $r(l,k,\alpha) < 1$ for every $l, k, \alpha > 0$, function $r(l,k,\alpha)$ is strictly increasing as a function of l and $\lim_{l\to\infty} r(l,k,\alpha) = 1$.

Proof. If the condition A holds, then

$$\lim_{N\to+\infty} F^N(M + (M - \theta_N)t) = \exp\{-(-t)^\alpha\}, \quad \forall t \le 0,$$

where $F = F_f$, $M = M[f] = \max f$ and θ_N is the $(1-1/N)$-quantile of the c.d.f. F $F(\theta_N) = 1 - 1/N$. This implies

$$M - F^{-1}(y) \sim (M - \theta_N)(-N \log y)^{1/\alpha}, \quad N \to \infty, \quad \forall y \in (0,1]$$

For the independent sample

$$\Psi_f(1,N) = \mathbf{E}(M[f] - M[f, \Xi_{1,N}])^k = N \int_{-\infty}^{+\infty} (M - x)^k F^{N-1}(x) dF(x) =$$

$$N \int_0^1 (M - F^{-1}(y))^k y^{N-1} dy \sim N \int_0^1 \left((M - \theta_N)(-N\log y)^{1/\alpha}\right)^k y^{N-1} dy =$$

$$N^{1+k/\alpha}(M - \theta_N)^k \int_0^1 (\log \frac{1}{y})^{k/\alpha} y^{N-1} dy =$$

$$N^{1+k/\alpha}(M - \theta_N)^k \int_0^{+\infty} z^{k/\alpha} e^{-Nz} dz = (M - \theta_N)^k \Gamma(k/\alpha + 1)$$

Therefore,

$$\mathbf{E}(M[f] - M[f, \Xi_{1,N}])^k \sim (M - \theta_N)^k \Gamma(k/\alpha + 1), \quad N \to \infty \tag{9}$$

Consider the stratified sample. By the condition B, the probability that the point of the global maximum of f is on the boundary of one of the sets $\mathcal{X}_i$ is 0. Consider l largest values from the collection $\{f(x_{i,j});\ i = 1,\ldots,m;\ j = 1,\ldots,l\}$. Since f is a continuous function and the global maximum is attained at a single point x_*, then as $m \to +\infty$ all l points with the largest function f values are located in one set, denote it $X_{i_\diamond}$. Therefore

$$F_\diamond(t) = P_{i_\diamond}(f(x) \le t) = m\left(P(f(x) \le t) - \frac{m-1}{m}\right) = mF(t) - m + 1.$$

Let $\theta_{\diamond l}$ be the $(1-1/l)$-quantile of the c.d.f. $F_\diamond(t)$. Since

$$1 - F_\diamond(\theta_{\diamond l}) = 1/l,\ 1 - mF(\theta_{\diamond l}) + m - 1 = 1/l,\ 1 - F(\theta_{\diamond l}) = \frac{1}{ml} = \frac{1}{n}\,,$$

we get $\theta_N = \theta_{\diamond l}$.

Since $F(t)$ satisfies condition A

$$F_\diamond(M + (M - \theta_{\diamond l})t) \sim m\exp\{-(-t)^{-\alpha}/(lm)\} - m + 1 \sim 1 - (-t)^\alpha/l\,.$$

Hence

$$M - F_\diamond^{-1}(y) \sim (M - \theta_N)\left(l(1-y)\right)^{1/\alpha},\quad m \to +\infty\,.$$

When $m \to \infty$, we thus get for every $k > 0$

$$\Psi_f(m,l) = \mathbf{E}(M[f] - M[f,\Xi_{m,l}])^k = l\int_{-\infty}^{+\infty}(M-x)^k F_\diamond^{l-1}(x)dF_\diamond(x) =$$

$$l\int_0^1 (M - F_\diamond^{-1}(y))^k y^{l-1}dy \sim l\int_0^1 \left((M-\theta_n)l^{1/\alpha}(1-y)^{1/\alpha}\right)^k y^{l-1}dy =$$

$$(M-\theta_N)^k l^{1+k/\alpha}\int_0^1 y^{l-1}(1-y)^{k/\alpha}dy = (M-\theta_N)^k\frac{\Gamma(k/\alpha+1)l^{k/\alpha}\Gamma(l+1)}{\Gamma(k/\alpha+l+1)}\,.$$

This and (9) yields (7).

Let us now fix k,α and study the function $r(l) = r(l,k,\alpha)$. Since $r(1) = 1/\Gamma(2+k/\alpha) < 1$ and the application of the Stirling formula yields

$$\lim_{l\to\infty} r(l) = \lim_{l\to+\infty}\frac{l^{k/\alpha}(2\pi l)^{1/2}\exp(-l)l^l}{(2\pi(l+k/\alpha))^{1/2}\exp\{-(l+k/\alpha)\}(l+k/\alpha)^{l+k/\alpha}} = 1\,,$$

to complete the proof we only need to show that the function $r(l)$ is strictly increasing. Indeed,

$$\log r(l+1) - \log r(l) = \log\left(1+\frac{1}{l}\right)^{k/\alpha} - \log\left(1+\frac{k/\alpha}{l+1}\right)$$

This expression is positive for every $l, k, \alpha > 0$ since

$$\left(1+\frac{1}{l}\right)^{k/\alpha} - \left(1+\frac{k/\alpha}{l+1}\right) = \left(\frac{k}{\alpha}\frac{1}{l}+\frac{\frac{k}{\alpha}(\frac{k}{\alpha}-1)}{2!}\frac{1}{l^2}+\dots\right) - \left(1+\frac{k}{l}\frac{1}{l}-\frac{k}{\alpha}\frac{1}{l^2}+\dots\right) =$$

$$\frac{k}{\alpha}\left(\frac{k/\alpha-1+2!}{2!l^2}+\frac{(k/\alpha-1)(k/\alpha-2)-3!}{3!l^3}+\dots\right)$$

and the positivity of the last expression follows from the fact that this series is alternating with rapid convergence and positive first term. ■

References

[1] A.A. Zhigljavsky *Theory of Global Random Search*, Kluwer Academic Publishers, Dordrecht, 1991.

Department of Mathematics
University of St.Petersburg
Bibliotechnaya sq. 2
198904 St.Petersburg
Russia

e-mail zh@stat.math.lgu.spb.su