About an Unbiased Estimate of the Gradient with Minimum Variance in the Planning of the Experiment

Sidnyaev N.I.*, Razgulyaev S.V., Shafikova S.E.

Abstract This paper considers algorithms of search of an extremum, to solve the problem of planning the experiment using the gradient method. A feature of the algorithm is that when searching for the motion (when the maximum) does not occur in the direction of the gradient, which is unknown to us, but its estimate. Estimate of the gradient at the point when this factor space is based on the results of measurements carried out in the neighborhood. Researcher’s task is to build a sensible plan with center to determine the estimate of the gradient in it.

Keywords Theory of Experiment Planning, Unbiased Estimate of the Gradient, The Response Function, The Method of Least Squares

1. Introduction

The task of finding the extremum of the mean characteristics, usually the task of finding the extremum of the response function \( \eta = f(x_1, x_2, ..., x_k) \). Search extremum response function is performed by the response surface study. This study is carried out by measuring the response surface at different points in the factor space.

Use for this purpose immediately known methods of finding the extremum function of many variables cannot be as "measurement" of the response function at each point in the factor space, which put the experience happens to fail.

1. One of the most famous classes of gradient methods for searching the extremes of the response function, which are based on gradient method is used not the gradient, and its evaluation. We present a general formulation of the gradient estimation problem and its solution.

Let the response function

\[
\eta = f(x_1, x_2, ..., x_k)
\]

defined in \( G \subseteq \mathbb{R}^k \). Consider an arbitrary point \( \bar{X} \in G \). Using a point \( \bar{X} = (X_1^0, X_2^0, ..., X_k^0) \) as the center of the plan [4], see full or fractional factorial experiment [1,2]. Denoted by \( X_i^0 \) taken by the variable \( X_i (i = 1, 2, ..., k) \) in the \( l \)-th level \((l = 1, 2)\). Upper \( X_i^0 \) and lower \( X_i^0 \) levels chosen symmetrically relative to the center plan to satisfy the condition \( X_i = (X_i^0 + X_i^0)/2 \) \((i = 1, 2, ..., k)\) [3]. It is obvious that the encoded variable

\[
x_i = (x_i - X_i^0) / S_i^0, i = 1, 2, ..., k,
\]

where \( S_i^0 = (X_i^0 - X_i^0)/2 \) - the interval of variation.

Express the response function (1) through coded variables:

\[
\eta = f(x_i, x_i, ..., x_i).
\]
2. The Task of Estimating the Gradient

The task of estimating the gradient will understand some estimate of the components of the response function $f'(x_1, x_2, \ldots, x_k)$ in point $x^0 = (x_1^0, x_2^0, \ldots, x_k^0)'$, where $x_i^0 = 0$ $(i = 1, \ldots, k)$. Suppose that in a neighborhood of $(x_1^0, x_2^0, \ldots, x_k^0)'$ function (3) can be decomposed by the Taylor series expansion of the form:

$$\eta = f(x_1, x_2, \ldots, x_k) = f(x_1^0, x_2^0, \ldots, x_k^0) + \sum_{i=1}^{k} \frac{\partial f}{\partial x_i} x_i + \frac{1}{2} \sum_{j=1}^{k} \sum_{i=1}^{k} \frac{\partial^2 f}{\partial x_i \partial x_j} x_i x_j + o(\|x - x^0\|^2).$$

Using the values

$$\beta_0^0 = f_1(x_1^0, x_2^0, \ldots, x_k^0); \quad \beta_j^0 = \frac{\partial f}{\partial x_j} x_j^0; \quad \beta_{ij}^0 = \frac{\partial^2 f}{\partial x_i \partial x_j} x_j^0,$$

we can write the response function

$$\eta = \beta_0^0 + \sum_{i=1}^{k} \beta_i^0 x_i + \sum_{j=1}^{k} \beta_j^0 x_j x_j + \sum_{i=1}^{k} \beta_{ij}^0 x_i x_j + o(\|x - x^0\|^2).$$

(4)

It should be noted that $\text{grad } f(x^0) = (\beta_0^0, \beta_1^0, \ldots, \beta_k^0)'$. Therefore, the problem of estimating the gradient is reduced to finding the OLS estimates of the unknown parameters $\beta_0^0, \beta_1^0, \ldots, \beta_k^0$. Let $\mathcal{D}(x_{iu})$ $(i = 1, 2, \ldots, k; u = 1, 2, \ldots, N)$ matrix of full or fractional factorial plan with center $x^0 = (x_1^0, x_2^0, \ldots, x_k^0)'$. For example, a full factorial plan matrix type $2^k$ [6] has the form:

$$\mathcal{D} = \begin{pmatrix}
-1 & -1 & \to y_1 \\
1 & -1 & \to y_2 \\
-1 & 1 & \to y_3 \\
1 & 1 & \to y_4
\end{pmatrix}$$

for the case when $\eta = f_i(x_1, x_2)$. For simplicity we assume that in the region $\mathcal{T} = \{(x_1, x_2, \ldots, x_k)' : -1 \leq x_i \leq 1\}$ response surface quite accurately approximated by a hyperplane, i.e.

$$\eta \approx \beta_0^0 + \sum_{i=1}^{k} \beta_i^0 x_i.$$  

(5)

Then, if $X = (x_{iu})$ $(j = 0, 1, \ldots, k; u = 1, 2, \ldots, N)$ - matrix of independent variables corresponding to the matrix plan $\mathcal{B} = (x_{iu})$ and the response function (5), The OLS estimator $\hat{\beta}_i^0$ is

$$\hat{\beta}_j = \frac{1}{N} \sum_{u=1}^{N} x_{ju} y_u,$$

where $y_1, y_2, \ldots, y_N$ - supervision in plan points.

As $\hat{\beta}_1^0, \hat{\beta}_2^0, \ldots, \hat{\beta}_k^0$ are estimates of the components of the gradient $\beta_0^0, \beta_1^0, \ldots, \beta_k^0$ or partial $\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_k}$ then the OLS estimate of the gradient of the response function at the point $(x_1^0, x_2^0, \ldots, x_k^0)'$ equal

$$\text{grad } f_1(x_1^0, x_2^0, \ldots, x_k^0) = (\hat{\beta}_1^0, \hat{\beta}_2^0, \ldots, \hat{\beta}_k^0)'.$$  

(6)

Evaluation of the gradient of the response (6) coincides with the point of the $(x_1^0, x_2^0, \ldots, x_k^0)'$ estimates its gradient at this point. Evaluating the gradient in the problem of finding an extremum is assumed that the response function near of the center plan factorial experiment accurately enough approximated by a hyperplane. In real situations, this condition is often not fulfilled. In this case evaluation of the gradient component may be biased. There is a relationship between the components of the gradient estimates and the estimates of the model parameters used to describe the response surface. Knowing bias of the estimates of the model parameters, we can find the bias of the gradient estimation. This implies an extremely important practical research conclusion: even the use of crude models to describe the response surface can not lead to a shift in the center of the gradient evaluation plan. This fact partly explains the reason for the relatively successful application of the method of Box and Wilson in applied research with a relatively crude approximation of the response surface.

Problem of unbiased estimation of the gradient $\text{grad } f(x_1^0, x_2^0, \ldots, x_k^0)$ during the decomposition of the response function $\eta = f(x_1^0, x_2^0, \ldots, x_k^0)'$ of the Taylor series (4) near of the center plan [6] $x^0 = (x_1^0, x_2^0, \ldots, x_k^0)'$ is to find unbiased estimates of the components of the gradient

$$\beta_j = \frac{\partial f}{\partial x_j} x_j^0,$$

where $x_j^0 = 0$ $(i = 1, 2, \ldots, k)$. Consider the case where

$$\eta = f(x_1, x_2, \ldots, x_k) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ij} x_i^2.$$

(7)

Obviously, $\text{grad } f(x_1^0, x_2^0, \ldots, x_k^0)' = (\beta_1, \beta_2, \ldots, \beta_k)'$. The following lemma is fair:

**Lemma 1.** For the response function of the form (7) using a fractional factorial design type $2^k$ OLS-estimate of the gradient $\text{grad } f(x_1^0, x_2^0, \ldots, x_k^0)'$ will be unbiased if and only if the variables $x_1, x_2, \ldots, x_k$ satisfy
\[ \pm \tilde{x}_i \neq \tilde{x}_j, \quad i, j = 1,2,\ldots, k; \quad i \neq j, \quad (8) \]
\[ \pm \tilde{x}_i \neq \tilde{x}_j \otimes \tilde{x}_s, \quad i, l, \quad s = 1,2,\ldots, k; \quad i \neq s, \quad (9) \]

**Proof:** Let \( \overline{D} = (\tilde{x}_i) \) (i = 1,2,..., k) - matrix factor type \( 2^{k-q} \) and also conditions (8), (9) are satisfied. Write response function (7) as
\[
\eta = \sum_{i \in \tilde{D}_k} \beta_i x_i + \beta_0 + \sum_{i \in \tilde{D}_k, j \in \tilde{D}} \beta_{ij} x_i x_j + \sum_{i \in \tilde{D}_k} \beta_{ij}^2 x_i^2 \]

(10)

Then not hard to believe that the columns of the matrix of independent variables \( X = (\tilde{x}_j) \) (j = 1,2,..., p+1) corresponding response function (10), satisfy the conditions of
\[
\| \tilde{x}_j \|^2 = N, \quad j = 1,2,\ldots, p+1; \quad (11) \]
\[ \tilde{x}_i^T \tilde{x}_j = 0, \quad i = 1,2,\ldots, k; \quad j = 1,2,\ldots, p+1; \quad i \neq j \quad (12) \]

It follows that the matrix \( X \) can be represented as a matrix block [1]
\[ X = (X^x, X^*) \quad \text{where} \quad X^x = (\tilde{x}_j)_{(j = 1,2,\ldots,k)}, \quad X^* = X^x R \]
For two arbitrary plans \( \xi_1(N_u) \) and \( \xi_2(N_u) \) with matrix \( \tilde{D}_1 = (x_{u}^{(1)}) \) and \( \tilde{D}_2 = (x_{u}^{(2)}) \)

(i = 1,2,...;u = 1,...,N_u) there is orthogonal matrix R: \( R = R(D_1) \), whence \( R = R(D_1)^{-1} \) [1]. According (12)
\[ X^x x^* = 0 \quad \text{and hence [1,2], OLS - evaluation} \]
\[ \hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2,\ldots, \hat{\beta}_k)^T \quad \text{gradient vector} \quad (\hat{\beta}_1, \hat{\beta}_2,\ldots, \hat{\beta}_k)^T \]

gradient vector and equal \( \hat{\beta}_j = (X^o x)^T X^o y \) where \( Y = (y_1, y_2,\ldots, y_N)^T \) - vector of supervision.

Because the columns of the matrix \( X^0 \) are pair wise orthogonal [1], then
\[ \hat{\beta}_j = (N)^{-1} \tilde{x}_j Y, \quad j = 1,\ldots, k. \]

(13)

It is easy to see that if the OLS - estimate of the gradient
\[ \hat{\nabla} d f(x^0_1, x^0_2,\ldots, x^0_k) \]
is unbiased, then (11), (12).

**Consequence 1.** If the matrix of the plan \( \tilde{D} = (x_{u})_{(u = 1,\ldots, N)} \) is a matrix of the full plan \( 2^k \), then conditions (11), (12) is carried out for any k and therefore unbiased OLS - estimate of the gradient exists and is defined by (13).

**Consequence 2.** If the conditions (11) and (12) of Lemma 1 are satisfied, then for the approximation of the response surface can be used dependence of the form
\[ \eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \]

In this case, OLS - estimate of the gradient models (7) and (13) is the same.

Note that condition (11), (12) is performed only in the case where the fractional resolutio replicates \( 2^{k-q} \) greater than three [2]. Therefore reformulate Lemma 1 with that in mind.

**Lemma 2.** Application of fractional replicate type \( 2^{k-q} \) allows to obtain an unbiased OLS - estimate of the gradient \( \hat{\nabla} d f(x^0_1, x^0_2,\ldots, x^0_k) \) of the response function (7) at the point \( (x^0_1, x^0_2,\ldots, x^0_k)^T \) if and only if its resolution is greater than three.

3. Examples of Specific Implementations

**Example 1.** Response function \( \eta = f(x_1, x_2, x_3, x_4) \) has the form (7). It is easy to see that half-replicate \( 2^{4-1} \) with defining contrast \( l = x_1 x_2 x_3 x_4 \) as a resolution equal to four, and by Lemma 2 allows to obtain an unbiased estimate of the gradient \( \hat{\nabla} d f(0,0,0,0) \) of the response function in the center of the plan. Obvious then
\[ \hat{\nabla} d f(0,0,0,0) = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)^T \]

where
\[ \hat{\beta}_1 = \frac{1}{4} (-y_1 + y_2 - y_3 + y_4 - y_5 + y_6 - y_7 + y_8); \]
\[ \hat{\beta}_2 = \frac{1}{4} (-y_1 - y_2 + y_3 + y_4 - y_5 + y_6 + y_7 + y_8); \]
\[ \hat{\beta}_3 = \frac{1}{4} (-y_1 - y_2 - y_3 - y_4 + y_5 + y_6 + y_7 + y_8); \]
\[ \hat{\beta}_4 = \frac{1}{4} (-y_1 + y_2 + y_3 - y_4 + y_5 - y_6 - y_7 + y_8). \]

If for this purpose use half-replicate \( 2^{4-1} \) with defining contrast \( l = x_1 x_2 x_4 \), then the OLS-estimate of the gradient will be biased, and [2]
\[ M \hat{\nabla} d f(0,0,0,0) = (\beta_1, \beta_2, \beta_3, \beta_4)^T = (\beta_1, \beta_2, \beta_3, \beta_4)^T. \]

In this case, only an estimate of the gradient component \( \beta_1 \) is biased; estimates other components are biased.

**Example 2.** Response function \( \eta = f(x_1, x_2,\ldots, x_5) \) has the form (3). Let's make sure that the 1/4 - replicate \( 2^{4-1} \), defined
by generating relations \( x_6 = x_1 x_2 x_3, x_7 = x_4 x_5 x_6 \), as a resolution equal to four [1], and therefore allows to obtain an unbiased OLS-estimate of the gradient in the center of the plan. Indeed, determining contrasts defining fractional replicate [2,5] are equal 1 = \( x_1 x_2 x_3 x_6 \); 1 = \( x_4 x_5 x_7 \).

Out of here defining the generalized contrast [6] \( 1 = x_1 x_2 x_3 x_6 = x_3 x_4 x_5 x_7 = x_1 x_2 x_3 x_6 x_7 \) has a resolution equal to four, i.e., fractional replicate is a replicate of \( 2^{7-2} \) [1]. It is easy to check that the class of fractional replicate \( 2^{7-3} \) fractional replicate does not exist, allowing for unbiased estimation of the gradient [1, 2, and 6].

These examples illustrate a general approach to solving the problem of estimating the gradient in situations where the model approximately describes the response surface [3-5]. From this discussion it follows that in some important practice [3, 4] cases is possible to match gradient estimates for models of varying complexity. Is a very complex practice [3, 4] cases is possible to match gradient estimates for models of varying complexity. Is a very complex practice [3, 4] cases is possible to match gradient estimates for models of varying complexity.

Example 2. Assume that the distribution of the directions grad \( f(\tau) = (\beta_\tau^0, \beta_\tau^2, \ldots, \beta_\tau^0) \)' know that it belongs to some parameter family of distributions. In case of continuous random variable it means that known views density \( p(\text{grad} f(\tau)/\theta) \), but not know the value of the parameter \( K \) determining the concrete density [5,7]. Parameter \( \theta \) can be a vector. For example, for a normal distribution \( \theta = (\mu, \sigma) \)

\[
p(\text{grad} f(\tau)/\theta) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(\text{grad} f(\tau) - \mu)^2}{2\sigma^2} \right).
\]

In the case of a discrete random variable \( p(\text{grad} f(\tau)/\theta) \), will denote the probability \( p(X=\text{grad} f(x)) \). For example, for Poisson distribution

\[
p(\text{grad} f(\tau)/\theta) = \frac{\theta^x e^{-\theta}}{x!}, \quad x=0,1,2,3,\ldots
\]

Let \( \tau(\theta) \) be a numeric parameter of interest to us. For example, \( \tau(\theta) = \theta, \tau(\mu, \sigma) = \mu, \tau(\mu, \sigma) = \mu/\sigma \).

Consider the problem of estimation of the gradient, consisting in the construction of a function \( \text{grad} f(\tau) = (\tau_1, \ldots, \tau_n) \), to the substitution of the instead of the arguments \( x_1, \ldots, x_n \) sampling data of the directions of the gradient we got the numbers, “close” to \( \tau(\theta) \). This proximity can only be done in average. Therefore, the requirements to the quality of the estimates of the gradient, is formulated in probabilistic terms relating to the distribution of the assessments covered as random values [7, 8]. This is the requirement to estimate values in most of the experiments were close to the estimated value of the parameter can be formulated as the following definition.

**Definition 1.** Estimate \( \text{grad} f(\tau) = (\tau_1, \ldots, \tau_n) \) is called for the unbiased \( \tau(\theta) \), if \( M_\theta \text{grad} f(\tau) = \tau(\theta) \) for all \( \theta \), where \( M_\theta \) symbol mathematical expectation, provided that the random vector \( \text{grad} f(\tau) = (x_1, \ldots, x_n) \) have the distribution

\[
L(\text{grad} f(\tau)/\theta) = p(\text{grad} f(x_1)/\theta) \ldots p(\text{grad} f(x_n)/\theta).
\]

But, as a rule, only one unbiasedness requirement does not emit estimation for unambiguously. Therefore, the following desirable requirement is the requirement of minimum of variance of this estimation [7]. To statistics might serve as a good estimate of this parameter \( \tau(\theta) \), it is necessary that the distribution of this statistics was concentrated in sufficiently close to the unknown value \( \tau(\theta) \), so that the probability of large deviations of the statistics from \( \tau(\theta) \) was quite small [7]. Then the systematic application of repeated this statistics as an estimation of the characteristics on the average will be obtained sufficient accuracy. The probability of large deviations will be small, and they will be rare.

Thus, among all unbiased estimates \( \text{grad} f(\tau) \) for \( \tau(\theta) \) more desirable is the estimation of the gradient, which has a minimum variance

\[
D_\theta(\text{grad} f(\tau)) = M(\text{grad} f(\tau) - \tau(\theta))^2 \quad \text{for all} \, \theta.
\]

**Definition** The unbiased estimate of the gradient with minimum variance is estimated \( \text{grad} f(\tau) \) that

\[
D(\text{grad} f(\tau)) \leq D(\text{grad} f(\tau)) \quad \text{for all} \, \theta \quad \text{for any unbiased estimate the gradient parameter} \, \tau(\theta).
\]

These requirements generally provide the estimate clearly, if such an estimate gradient exists at all. The existence of the unbiased estimate of the gradient with minimum variance (UEGMV) takes place, not always, as far as the variance for these estimates should be minimal evenly \( \theta \). This fact is perhaps the most serious argument against such strong requirements such as unlimited increasing the number of observations, i.e.

\[
\lim_{n \to \infty} P \left( \left| \text{grad} f(\tau) - \tau(\theta) \right| < \varepsilon \right) = 1 \quad \text{for} \quad \forall \varepsilon > 0
\]

So, the viability of the gradient estimation means that if a sufficiently large number of observations \( n \) with arbitrarily high accuracy deviation estimate the gradient from the true parameter value less any pre-specified value.

To unbiased estimate of the gradient was valid enough to variance estimation tends to zero if \( n \to \infty \) (this follows from the Chebyshev inequality[7]).

**Example 3.** Let's show that the estimates of the gradient

a. \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} \text{grad} f(x_i) \) - the sample mean,

b. \( \bar{S}^2 = \frac{1}{n} \sum_{i=1}^{n} (\text{grad} f(x_i) - \bar{x})^2 \) - sample variance,

Are valid estimates.

**Decision.** According to the law of large numbers
\[ \left\lfloor \frac{1}{n} \sum_{1}^{n} \nabla f(x) - M(\nabla f(x)) \right\rfloor < \varepsilon \rightarrow 1, n \rightarrow \infty \]

i.e. estimates a) valid, because
\[ \sum \xi = \frac{1}{n} \sum_{1}^{n} \nabla f(x) \]

Example 4. (The existence of several unbiased estimates of the gradient.)
\[ \nabla f(x_1), \ldots, \nabla f(x_n) \] are independent identically distributed random variables having a Poisson distribution
\[ p_\theta(\nabla f(x) = k) = \frac{\theta^k}{k!} e^{-\theta}, \quad k = 0, 1, \ldots \]

Let’s show that there is not unbiased estimate the gradient.
\[ \tau(\theta) \text{ i.e. } M_\theta \nabla f(x) = \tau(\theta) \] whence it follows that
\[ \sum_{k=0}^{\infty} \frac{\theta^k}{k!} e^{-\theta} = \frac{1}{\theta} \]

Whence it follows that
\[ \sum_{k=0}^{\infty} \frac{\theta^k}{k!} \theta^k = \frac{1}{\theta} \sum_{k=0}^{\infty} \frac{\theta^k}{k!} \theta^k. \]

Finally
\[ \sum_{k=0}^{\infty} \frac{\theta^k}{k!} \theta^k = \sum_{k=0}^{\infty} \frac{1}{(k+1)!} \theta^k. \]

And as easy to understand, there is no function \( \nabla f(x) \) [7], for which the last equality it would be fair for all \( \theta > 0 \).

Example 6. (when unbiased estimate of the gradient is absurd)
A random variable \( \nabla f(x) \) has a geometric distribution [8]:
\[ p\{\nabla f(x) = k\} = q^k p, \quad k \geq 0, \quad p + q = 1 \]

(the number of failures preceding the first success in a sequence of Bernoulli trials).

The unbiased estimate of the parameter \( p \) is the estimate of
\[ \nabla f(x) = \begin{cases} 1, & x = 0, \\ 0, & x > 0. \end{cases} \]

Indeed,
\[ \sum_{k=0}^{\infty} \frac{\theta^k}{k!} \theta^k = \sum_{k=0}^{\infty} \frac{\theta^k}{k!} \theta^k \Rightarrow \nabla f(x) = \begin{cases} 1, & k = 0, \\ 0, & k > 0. \end{cases} \]

However, estimate \( \nabla f(x) \) has the lowest variance.

Example 5. (When there are no unbiased assessment of the gradient.)

One observation \( \nabla f(x) \), \( \nabla f(x) \) is a random variable which has a Poisson distribution with an unknown parameter \( \theta \)
\[ p_\theta(\nabla f(x) = k) = \frac{\theta^k}{k!} e^{-\theta}, \quad k = 0, 1, \ldots \]

Decision. Suppose that \( \nabla f(x) \) is the unbiased estimate \( \tau(\theta) \) i.e. \( M_\theta \nabla f(x) = \tau(\theta) \)
whence it follows that
\[ \sum_{k=0}^{\infty} \frac{\theta^k}{k!} e^{-\theta} \theta^k = \frac{1}{\theta} \]

Finally
\[ \sum_{k=0}^{\infty} \frac{\theta^k}{k!} \theta^k = \frac{1}{\theta} \sum_{k=0}^{\infty} \frac{\theta^k}{k!} \theta^k. \]

And as easy to understand, there is no function \( \nabla f(x) \) [7], for which the last equality it would be fair for all \( \theta > 0 \).

Example 6. (when unbiased estimate of the gradient is absurd)
A random variable \( \nabla f(x) \) has a geometric distribution [8]:
\[ p\{\nabla f(x) = k\} = q^k p, \quad k \geq 0, \quad p + q = 1 \]

(the number of failures preceding the first success in a sequence of Bernoulli trials).

The unbiased estimate of the parameter \( p \) is the estimate of
\[ \nabla f(x) = \begin{cases} 1, & x = 0, \\ 0, & x > 0. \end{cases} \]

Indeed,
\[ \sum_{k=0}^{\infty} \frac{\theta^k}{k!} \theta^k = \sum_{k=0}^{\infty} \frac{\theta^k}{k!} \theta^k \Rightarrow \nabla f(x) = \begin{cases} 1, & k = 0, \\ 0, & k > 0. \end{cases} \]

However, estimate \( \nabla f(x) \) has the lowest variance.
For that, make \( \hat{\text{grad}} f_1(x) = t_1 \), \( \hat{\text{grad}} f_2(x) = t_2 \) and \( \hat{\text{grad}} f_3(x) = t_3 \).

In fact, consider a new estimate \( t_3 = \frac{1}{2}(t_1 + t_2) \), for which

\[
M_{\phi 3} = \frac{1}{2}(M_{t_1} + M_{t_2}) = \tau(\theta),
\]
\[
Dt_3 = \frac{1}{4}(Dt_1 + Dt_2 + 2 \text{cov}(t_1, t_2)).
\]

It is known that \( \text{cov}(t_1, t_2) \leq \sqrt{Dt_1 Dt_2} \) [7,9,10], and the sign of equality occurs only when almost everywhere

\[
(t_1 - \tau) = k(\theta)(t_2 - \tau),
\]

Where \( k(\theta) \) any function from \( \theta \). If we mark \( Dt_1 = Dt_2 = V \), we have shown that \( \text{cov}(t_1, t_2) \leq V \), and therefore \( Dt_3 \leq V \), with the sign of equality occurs only when the condition (14).

Find \( k(\theta) \). Know \( V = \text{cov}(t_1, t_2) = k(\theta)V \) whence it follows that \( k(\theta) = 1 \) and therefore \( t_1 = t_2 \) and correspondingly \( \hat{\text{grad}} f_1(x) = \hat{\text{grad}} f_2(x) \) almost everywhere.

These examples allow you to move on to more complex tasks [11], namely to the problems of comparison of the effectiveness of plans at the estimation of the gradient.

4. Conclusions

Thus, the existence of the unbiased estimate of the gradient with minimum variance occurs not always, as far as the variance for these estimates should be minimal uniformly in the parameter. The viability of the gradient estimation means that if a sufficiently large number of observations \( n \) with arbitrarily high accuracy deviation estimate the gradient from the true parameter value less any pre-specified value. To unbiased estimate of the gradient was wealthy enough to variance estimation tends to zero \( n \to \infty \).

**REFERENCES**


