Bayesian Sample Size Estimation

N. S. Kiselev[1, *] **and A. V. Grabovoy**[1, **]

[1] *Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russian Federation*

Abstract—The paper investigates the problem of estimating a sufficient sample size. The issue of determining a sufficient sample size without specifying a statistical hypothesis about the distribution of model parameters is considered. Two approaches to determining a sufficient sample size based on likelihood function values on resampled subsets are proposed. These approaches are based on heuristics about the behavior of the likelihood function with a large number of objects in the sample. Two approaches to determining a sufficient sample size based on the proximity of posterior distributions of model parameters on similar subsets are suggested. The correctness of the presented approaches is proven under certain restrictions on the model used. A theorem about moments of the limit posterior distribution of parameters in a linear regression model is proven. A method for forecasting the likelihood function in case of an insufficient sample size is proposed. A computational experiment is conducted to analyze the properties of the proposed methods.

Keywords and phrases: *Sufficient sample size, Bayesian inference, Bootstrapping, Posterior distributions similarity, Linear regression*

1. INTRODUCTION

The task of supervised machine learning involves selecting a predictive model from a parametric family. This choice is usually based on certain statistical hypotheses, such as maximizing a quality functional.

Definition 1. *A model that satisfies these statistical hypotheses is called an* ***adequate*** *model.*

When planning a computational experiment, it is necessary to estimate the minimum sample size — the number of objects required to build an adequate model.

Definition 2. *The sample size required to build an adequate predictive model is called* ***sufficient****.*

This work addresses the issue of determining the sufficient sample size. There are numerous studies dedicated to this topic, with approaches classified into statistical, Bayesian, and heuristic methods.

Some of the early articles on this topic [1, 2] formulate a specific statistical criterion, where the sample size estimation method associated with this criterion guarantees achieving a fixed statistical power with a Type I error not exceeding a specified value. Statistical methods have certain limitations associated with their practical application. They allow for estimating the sample size based on assumptions about the data distribution and information about the agreement of observed values with the assumptions of the null hypothesis.

The class of Bayesian methods for sample size estimation is quite wide. In the work [3] the sufficient sample size is determined based on maximizing the expected utility function. This may explicitly include parameter distribution functions and penalties for increasing the sample size. This work also considers alternative approaches based on constraining a certain quality criterion for estimating model parameters. Among these criteria, the Average Posterior Variance Criterion (APVC), Average Coverage Criterion (ACC), Average Length Criterion (ALC), and Effective Sample

[*] E-mail: kiselev.ns@phystech.edu

[**] E-mail: grabovoy.av@phystech.edu

Size Criterion (ESC) stand out. These criteria have been further developed in other works, for example, [4] and [5]. Over time, the authors of [6] conducted a theoretical and practical comparison of methods from [1–3].

Authors like [7], as well as [8], discuss the differences between Bayesian and frequentist approaches in determining sample size. They also propose robust methods for the Bayesian approach and provide illustrative examples for some probabilistic models.

In the work [9], various methods for estimating sample size in generalized linear models are considered, including statistical, heuristic, and Bayesian methods. Methods such as Lagrange Multiplier Test, Likelihood Ratio Test, Wald Test, Cross-Validation, Bootstrap, Kullback-Leibler Criterion, Average Posterior Variance Criterion, Average Coverage Criterion, Average Length Criterion, and Utility Maximization are analyzed. The authors point out the potential development of combining Bayesian and statistical approaches to estimate sample size for insufficient available sample sizes.

In [10] a method for determining sample size in logistic regression is discussed, using cross-validation and Kullback-Leibler divergence between posterior distributions of model parameters on similar subsamples. Similar subsamples are those that can be obtained from each other by adding, removing, or replacing one object.

This work utilizes a genetic algorithm [11] to approximate a given set of functions. A genetic algorithm is an optimization process of a population of candidates (referred to as individuals) evolving towards better solutions [12]. Each individual has a set of characteristics (genes or phenotypes) that can change during evolution. Changes occur through crossover or mutation operations. Evolution starts with a random population, and each generation is considered as a basis for generating the next one. The fitness of individuals is measured in each generation, and individuals with high fitness are selected to create a new generation [13]. The algorithm terminates after reaching the maximum number of generations or achieving satisfactory results. Thus, each new generation becomes more adapted to the environment.

2. PROBLEM STATEMENT

An object is defined as a pair $(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}^n$ is the feature vector, and $y \in \mathbb{Y}$ is the target variable. In regression problems $\mathbb{Y} = \mathbb{R}$, and in K-class classification problems $\mathbb{Y} = \{1, \ldots, K\}$.

The feature-object matrix for a sample $\mathfrak{D}_m = \left\{(\mathbf{x}_i, y_i)\right\}, i \in \mathcal{I} = \{1, \ldots, m\}$ of size m is called the matrix $\mathbf{X}_m = \left[\mathbf{x}_1, \ldots, \mathbf{x}_m\right]^\top \in \mathbb{R}^{m \times n}$.

The target variable vector for a sample $\mathfrak{D}_m = \left\{(\mathbf{x}_i, y_i)\right\}, i \in \mathcal{I} = \{1, \ldots, m\}$ of size m is denoted by $\mathbf{y}_m = \left[y_1, \ldots, y_m\right]^\top \in \mathbb{Y}^m$.

A model is a parametric family of functions f, mapping the Cartesian product of the set of feature vector values $\mathbb{X}$ and the set of parameter values $\mathbb{W}$ to the set of target variable values $\mathbb{Y}$:

$$f : \mathbb{X} \times \mathbb{W} \to \mathbb{Y}.$$

A probabilistic model is a joint distribution

$$p(y, \mathbf{w}|\mathbf{x}) = p(y|\mathbf{x}, \mathbf{w})p(\mathbf{w}) : \mathbb{Y} \times \mathbb{W} \times \mathbb{X} \to \mathbb{R}^+,$$

where $\mathbf{w} \in \mathbb{W}$ is the set of model parameters, $p(y|\mathbf{x}, \mathbf{w})$ specifies the likelihood of an object, and $p(\mathbf{w})$ represents the prior distribution of parameters.

The likelihood function of a sample $\mathfrak{D}_m = \left\{(\mathbf{x}_i, y_i)\right\}, i \in \mathcal{I} = \{1, \ldots, m\}$ of size m, where $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are i.i.d. together, is defined as

$$L(\mathfrak{D}_m, \mathbf{w}) = p(\mathbf{y}_m|\mathbf{X}_m, \mathbf{w}) = \prod_{i=1}^{m} p(y_i|\mathbf{x}_i, \mathbf{w}).$$

Its logarithm

$$l(\mathfrak{D}_m, \mathbf{w}) = \sum_{i=1}^{m} \log p(y_i|\mathbf{x}_i, \mathbf{w})$$

is called the logarithmic likelihood function. Unless stated otherwise, we consider samples to be i.i.d.

The maximum likelihood estimate (MSE) of a set of parameters $\mathbf{w} \in \mathbb{W}$ based on the sample $\mathfrak{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$ of size m is given by

$$\hat{\mathbf{w}}_m = \arg\max_{\mathbf{w} \in \mathbb{W}} L(\mathfrak{D}_m, \mathbf{w}).$$

The task is to determine the sufficient sample size m^*. Let a criterion T be given. E.g. it can be constructed based on heuristics regarding the behaviour of model parameters.

Definition 3. *The sample size m^* is called* ***sufficient*** *according to the criterion T, if T holds for all $k \geqslant m^*$.*

It should be noted that it is possible for $m^* \leqslant m$ or $m^* > m$. These two cases will be considered separately later on.

3. SUFFICIENT SAMPLE SIZE DOES NOT EXCEED THE AVAILABLE ONE

In this section, we will assume that $m^* \leqslant m$ is valid. This means that we just need to formalize which sample size can be considered sufficient.

3.1. Analysis of the behavior of the likelihood function

To determine sufficiency, we will use the likelihood function. When there are enough objects available, it is quite natural to expect that the resulting parameter estimate will not change much from one sample realization to another [2, 14]. The same can be said about the likelihood function. Thus, we formulate which sample size can be considered sufficient.

Definition 4. *Let's fix some positive number $\varepsilon > 0$. The sample size m^* is called* ***D-sufficient*** *if for all $k \geqslant m^*$*

$$D(k) = \mathbb{D}_{\mathfrak{D}_k} L(\mathfrak{D}_m, \hat{\mathbf{w}}_k) \leqslant \varepsilon.$$

On the other hand, when there are enough objects available, it is also quite natural that when adding another object to consideration, the resulting parameter estimate will not change much. Let's formulate another definition.

Definition 5. *Let's fix some positive number $\varepsilon > 0$. The sample size m^* is called* ***M-sufficient*** *if for all $k \geqslant m^*$*

$$M(k) = \left| \mathbb{E}_{\mathfrak{D}_{k+1}} L(\mathfrak{D}_m, \hat{\mathbf{w}}_{k+1}) - \mathbb{E}_{\mathfrak{D}_k} L(\mathfrak{D}_m, \hat{\mathbf{w}}_k) \right| \leqslant \varepsilon.$$

In the definitions above instead of the likelihood function $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$ we can consider its logarithm $l(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$.

Suppose that $\mathbb{W} = \mathbb{R}^n$. Recall that the Fisher information is called the matrix

$$\left[\mathcal{I}(\mathbf{w})\right]_{ij} = -\mathbb{E}\left[\frac{\partial^2 \log p(\mathbf{y}|\mathbf{x}, \mathbf{w})}{\partial w_i \partial w_j}\right].$$

A known result is the asymptotic normality of the maximum likelihood estimate, that is, $\sqrt{k}\left(\hat{\mathbf{w}}_k - \mathbf{w}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}^{-1}(\mathbf{w})\right)$. Convergence in the distribution generally does not imply convergence of the moments of a random vector. Nevertheless, if we assume the latter, then in some models it is possible to prove the correctness of our proposed definition of M-sufficient sample size.

For convenience, we denote the distribution parameters $\hat{\mathbf{w}}_k$ as follows: mathematical expectation $\mathbb{E}_{\mathfrak{D}_k} \hat{\mathbf{w}}_k = \mathbf{m}_k$ and the covariance matrix $\mathbb{D}_{\mathfrak{D}_k} \hat{\mathbf{w}}_k = \mathbf{\Sigma}_k$. Then the following theorem holds, the proof of which is given in the section A.

Theorem 1 (Kiselev, 2023). *Let $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ and $\|\boldsymbol{\Sigma}_{k+1} - \boldsymbol{\Sigma}_k\|_F \to 0$ as $k \to \infty$. Then, in the linear regression model, the definition of M-sufficient sample size is correct. Namely, for any $\varepsilon > 0$, there is such a m^* that for all $k \geqslant m^*$ $M(k) \leqslant \varepsilon$ is satisfied.*

Corollary 1. *Let $\|\mathbf{m}_k - \mathbf{w}\|_2 \to 0$ and $\|\boldsymbol{\Sigma}_k - [k\mathcal{I}(\mathbf{w})]^{-1}\|_F \to 0$ for $k \to \infty$. Then, in the linear regression model, the definition of an M-sufficient sample size is correct.*

By condition, one sample is given. Therefore, in the experiment it is not possible to calculate the mathematical expectation and variance specified in the definitions. To evaluate them, we will use the bootstrap technique. Namely, we will generate from the given $\mathfrak{D}_m$ a number of B subsamples of size k with a return. For each of them, we get an estimate of the parameters $\hat{\mathbf{w}}_k$ and calculate the value of $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$. For the estimation, we will use a sample mean and an unbiased sample variance (for bootstrap samples).

The definitions proposed above can also be applied in those problems where an arbitrary loss function is minimized rather than the likelihood function is maximized. We do not provide any theoretical justification for this, but in practice such a heuristic turns out to be quite successful.

3.2. *Analysis of the posterior distribution of model parameters*

In [10], it is proposed to use the Kullback-Leibler divergence to estimate a sufficient sample size in a binary classification problem. The idea is based on the fact that if two subsamples differ from each other by one object, then the posterior distributions obtained from them should be close. This proximity is determined by the Kullback-Leibler divergence.

In this paper, it is proposed to develop this approach, to investigate it not only in the classification problem, but also in the regression problem. As a measure of proximity, it is proposed to use not only the Kullback-Leibler divergence, but also the s-score similarity function from [15].

Consider two subsamples $\mathfrak{D}^1 \subseteq \mathfrak{D}_m$ and $\mathfrak{D}^2 \subseteq \mathfrak{D}_m$. Let $\mathcal{I}_1 \subseteq \mathcal{I} = \{1, \dots, m\}$ and $\mathcal{I}_2 \subseteq \mathcal{I} = \{1, \dots, m\}$ — corresponding to them subsets of indexes.

Definition 6. *Subsamples $\mathfrak{D}^1$ and $\mathfrak{D}^2$ are called **similar** if $\mathcal{I}_2$ can be obtained from $\mathcal{I}_1$ by deleting, replacing or adding one element, that is*

$$|\mathcal{I}_1 \triangle \mathcal{I}_2| = |(\mathcal{I}_1 \setminus \mathcal{I}_2) \cup (\mathcal{I}_2 \setminus \mathcal{I}_1)| = 1.$$

Consider two similar subsamples $\mathfrak{D}_k = (\mathbf{X}_k, \mathbf{y}_k)$ and $\mathfrak{D}_{k+1} = (\mathbf{X}_{k+1}, \mathbf{y}_{k+1})$ of sizes k and $k+1$, respectively. This means that the larger one is obtained by adding one element to the smaller one. Let's find the posterior distribution of the model parameters over these subsamples:

$$p_k(\mathbf{w}) = p(\mathbf{w}|\mathfrak{D}_k) = \frac{p(\mathfrak{D}_k|\mathbf{w})p(\mathbf{w})}{p(\mathfrak{D}_k)} \propto p(\mathfrak{D}_k|\mathbf{w})p(\mathbf{w}),$$

$$p_{k+1}(\mathbf{w}) = p(\mathbf{w}|\mathfrak{D}_{k+1}) = \frac{p(\mathfrak{D}_{k+1}|\mathbf{w})p(\mathbf{w})}{p(\mathfrak{D}_{k+1})} \propto p(\mathfrak{D}_{k+1}|\mathbf{w})p(\mathbf{w}).$$

Definition 7. *Let's fix some positive number $\varepsilon > 0$. The sample size m^* is called **KL-sufficient** if for all $k \geqslant m^*$*

$$KL(k) = D_{KL}(p_k \| p_{k+1}) = \int p_k(\mathbf{w}) \log \frac{p_k(\mathbf{w})}{p_{k+1}(\mathbf{w})} d\mathbf{w} \leqslant \varepsilon.$$

For a pair of normal distributions, the Kullback-Leibler divergence has a fairly simple form. Assume that the posterior distribution is normal, that is, $p_k(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_k, \boldsymbol{\Sigma}_k)$. Guided by the heuristic that the convergence of the moments of such a distribution should entail the proximity of posterior distributions on similar subsamples, the following statement can be formulated.

Theorem 2 (Kiselev, 2024)**.** *Let $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ and $\|\boldsymbol{\Sigma}_{k+1} - \boldsymbol{\Sigma}_k\|_F \to 0$ as $k \to \infty$. Then, in a model with a normal posterior distribution of parameters, the definition of a KL-sufficient sample size is correct. Namely, for any $\varepsilon > 0$, there is such a m^* that for all $k \geqslant m^*$ $KL(k) \leqslant \varepsilon$ is satisfied.*

In this paper, it is proposed to use the s-score similarity function from [15] as a measure of proximity of distributions:

$$\text{s-score}(g_1, g_2) = \frac{\int_{\mathbf{w}} g_1(\mathbf{w}) g_2(\mathbf{w}) d\mathbf{w}}{\max_{\mathbf{b}} \int_{\mathbf{w}} g_1(\mathbf{w} - \mathbf{b}) g_2(\mathbf{w}) d\mathbf{w}}.$$

Definition 8. *Let's fix some positive number $\varepsilon > 0$. The sample size m^* is called **S-sufficient** if for all $k \geqslant m^*$*

$$S(k) = \textit{s-score}(p_k, p_{k+1}) \geqslant 1 - \varepsilon.$$

As in the case of a KL-sufficient sample size, in a model with a normal posterior distribution, it is possible to write an expression for the criterion used. Thus, one more statement can be formulated.

Theorem 3 (Kiselev, 2024)**.** *Let $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ as $k \to \infty$. Then, in a model with a normal posterior distribution of parameters, the definition of an S-sufficient sample size is correct. Namely, for any $\varepsilon > 0$, there is such a m^* that for all $k \geqslant m^*$ $S(k) \geqslant 1 - \varepsilon$ is satisfied.*

Let the linear regression model have a normal prior distribution of parameters. By the conjugacy property of the prior distribution and likelihood, the posterior distribution is also normal. Thus, we come to one of the simplest examples of a model for which the theorems presented above are valid. In fact, simpler statements can be formulated for linear regression.

Theorem 4 (Kiselyov, 2024)**.** *Let the sets of values of the features and the target variable be bounded, that is, $\exists M \in \mathbb{R} : \|\mathbf{x}\|_2 \leqslant M$ and $|y| \leqslant M$. If $\lambda_{\min}\left(\mathbf{X}_k^\top \mathbf{X}_k\right) = \omega(\sqrt{k})$ for $k \to \infty$, then in a linear regression model with a normal prior distribution of parameters $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ and $\|\boldsymbol{\Sigma}_{k+1} - \boldsymbol{\Sigma}_k\|_F \to 0$ as $k \to \infty$.*

4. SUFFICIENT SAMPLE SIZE IS LARGER THAN AVAILABLE

In this section, we will assume that $m^* > m$ is valid.

The problem arises of predicting the mathematical expectation of the likelihood function / loss function at $k > m$. In general, this is quite a difficult task. In this paper, it is proposed to analyze a large number of open datasets from [16] in order to find a parametric family of functions that should approximate the dependence of the loss function on the sample size used. It is proposed to study datasets with regression and classification tasks separately.

4.1. A genetic algorithm in the task of approximating a set of functions

One of the simplest search algorithms in terms of implementation and logic is the genetic algorithm. Using it, we will build a method for finding the desired family of functions.

Suppose that for N different datasets, a graph is plotted of the dependence of the average value of the loss function (or likelihood function with a minus sign) on the sample size used. Let's bring these N dependencies to the same scale on both axes. To do this, subtract the minimum value, and then divide by the maximum value. In this case, the graph of each dependence is squared $[0; 1]^2$, It starts at the point $(0; 1)$ and ends at the point $(1; 0)$.

A population in a genetic algorithm is a set of parametric families of functions. For example, one individual can be a family of $w_0 + w_1 \cdot \log(w_2 \cdot x) + w_3 \cdot x^2$, where x is a variable, and $\mathbf{w}$ is a vector of parameters. The initial population is initialized randomly. The simplest unary functions are used: $1, x, \sin x, \cos x, \exp x, \log x, \operatorname{ctg} x$ and $\operatorname{cth} x$, as well as the simplest binary functions: $+, -, *$ and $/$. Each individual is represented using a binary tree, the nodes of which contain the above-mentioned functions, and the leaves are necessarily 1 or x. At the same time, each node is assigned its own component of the parameter vector.

The fitness of an individual is measured as follows. For each of the N approximated dependencies, the problem of selecting a vector of parameters is solved. The mean squared error is minimized. The resulting MSE value is averaged over all N dependencies. The final value determines the fitness of the individual.

The crossover is implemented in such a way that a random subtree of one of the parent individuals is replaced by a random subtree of the other. A mutation replaces a function in a random node of the tree with another random function.

The algorithm terminates after a given number of generations. An individual from the last generation with the best fitness is selected. The solution is an appropriate parametric family of functions.

5. COMPUTATIONAL EXPERIMENT

An experiment has been conducted to analyze the properties of the proposed methods for estimating a sufficient sample size. The experiment consists of several parts. In the first part, estimates of a sufficient sample size are considered in the case when a sufficient sample size does not exceed the available one. The second part examines the results obtained under the conditions that a sufficient sample size is larger than the available one.

5.1. Sufficient sample size does not exceed the available one

5.1.1. Convergence of the proposed functions Synthetic data is generated from a linear regression model. The number of objects is 1000, the number of features is 20. One object is sequentially removed from the given sample until the number of objects in the subsample is equal to the number of features. This process is repeated $B = 1000$ times. As a result, for each sample size k, the value of each of the functions $D(k)$, $M(k)$, $KL(k)$ and $S(k)$ defined in Chapter 3 is obtained (the logarithm of the likelihood function is used here). The following is a Fig. 1, which shows the resulting dependencies.

The graphs obtained confirm the results obtained in the Theorems 1, 2 and 3. The values of the functions $D(k)$, $M(k)$ and $KL(k)$ tend to zero as the size of the subsample increases. The values of $S(k)$ tend to one as the size of the subsample increases.

5.1.2. Determining a sufficient sample size Synthetic data is generated from a linear regression model. The number of objects is 1000, the number of features is 20. The following are graphs of the logarithm of the likelihood function of the sample, as well as the functions $D(k)$ and $M(k)$ for the logarithm of the likelihood function. $B = 1000$ bootstrap samples were used. The definition of D-sufficient and M-sufficient sample sizes has been performed. For D-sufficiency, $\varepsilon = 3 \cdot 10^1$ is selected, for M-sufficiency $\varepsilon = 4 \cdot 10^{-1}$. The results are shown in Fig. 2.

The second synthetic sample is generated from a logistic regression model. The number of objects is 1000, the number of features is 20. Similar graphs are shown in Fig. 3. For D-sufficiency, $\varepsilon = 3 \cdot 10^1$ was used, for M-sufficiency $\varepsilon = 6 \cdot 10^{-1}$.

The following is an example of determining a sufficient sample size based on real data. The Abalone dataset from [16] is used with the regression task. The number of objects is 4177, the number of features is 8. The results are shown in Fig. 4. The definition of D-sufficiency uses $\varepsilon = 2.5 \cdot 10^{-3}$, for M-sufficiency, $\varepsilon = 8 \cdot 10^{-3}$ is taken.

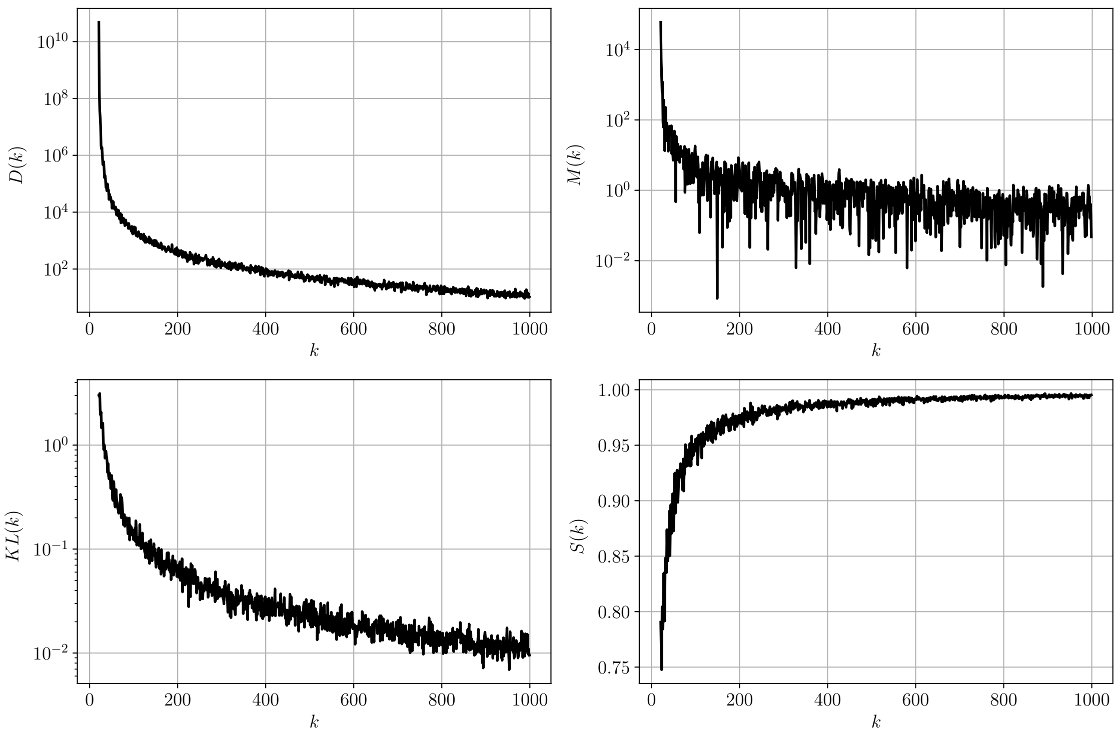

Figure 1. Convergence of functions for synthetic sampling (linear regression)

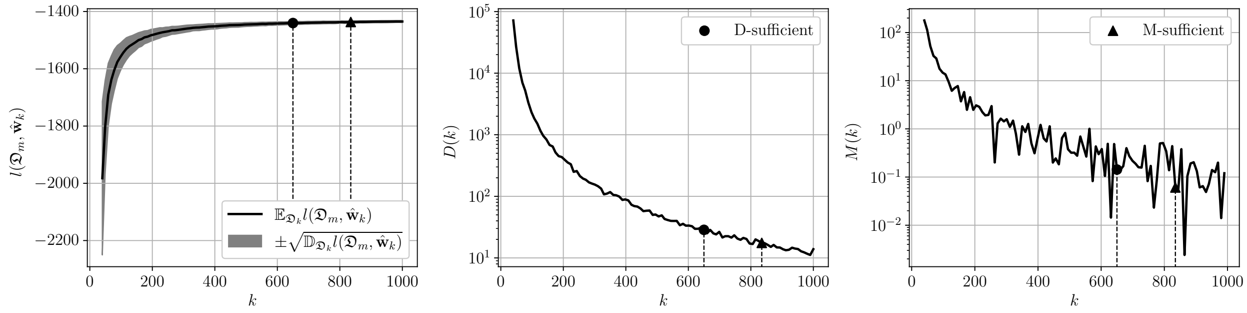

Figure 2. Synthetic sample (linear regression) at $m^* \leqslant m$

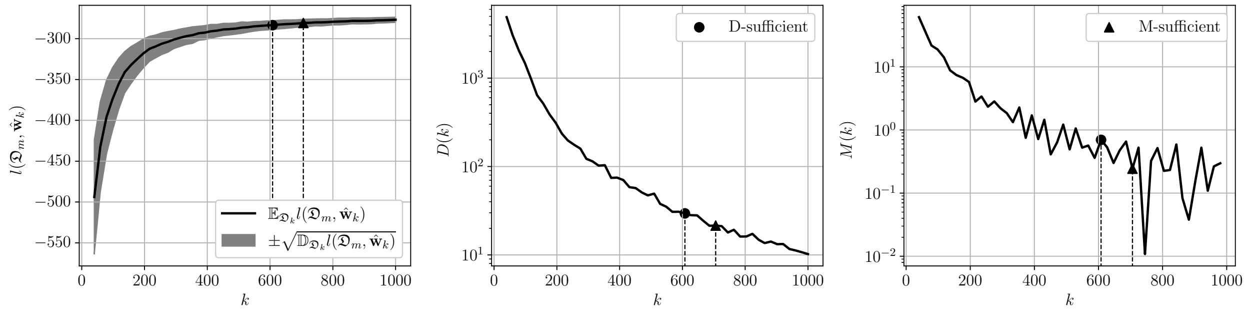

Figure 3. Synthetic sample (logistic regression) at $m^* \leqslant m$

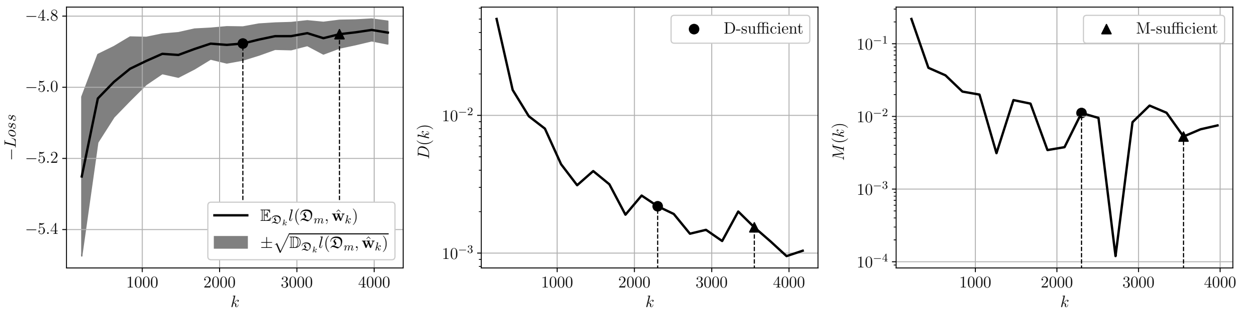

Figure 4. Abalone sample (regression) at $m^* \leqslant m$

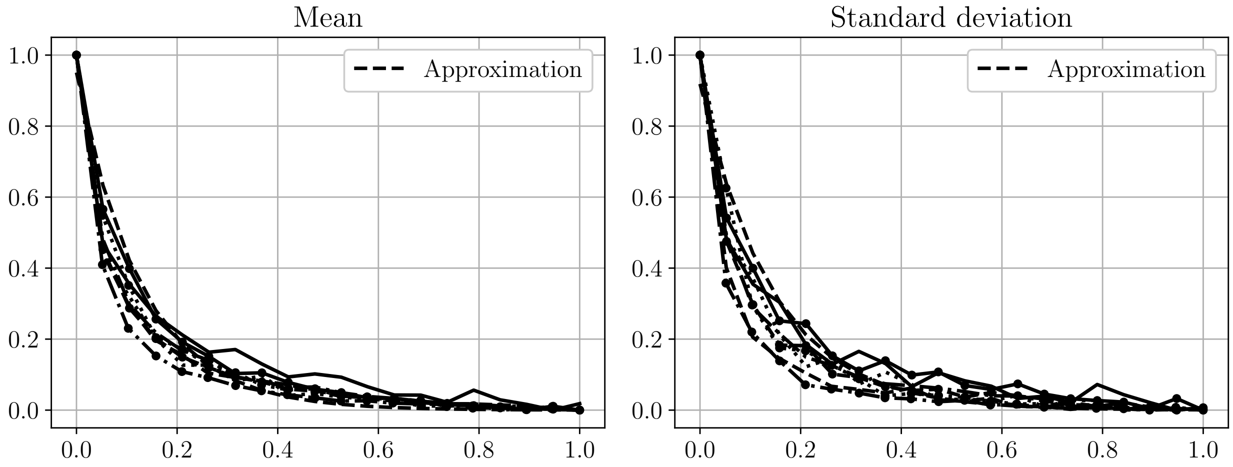

Figure 5. Behavior of the loss function in the regression problem

5.2. Sufficient sample size is larger than the available one

5.2.1. Determination of a parametric family of functions using a genetic algorithm The implementation of the genetic algorithm given in the section 4.4.1 can be found in the repository `https://github.com/kisnikser/Bayesian-Sample-Size-Estimation`. To analyze the dependence of the loss function on the sample size used in the regression task, the following datasets from [16] were used: Abalone, Auto MPG, Liver Disorders, Wine Quality, Parkinsons Telemonitoring, Bike Sharing Dataset, Real estate valuation and Heart failure clinical records. The quadratic loss function MSE was chosen. The regression problem for each of them was solved using linear regression from [17]. Averaging was performed on $B = 100$ bootstrap samples. As mentioned earlier, all dependencies are reduced to the same scale on both axes. The resulting graphs are shown in Fig. 5. On the left is a graph for the sample average. On the right is a graph for the sample standard deviation.

The application of the genetic algorithm leads to the same family of functions for approximating the mean and standard deviation in the regression problem:

$$w_0 + w_1 \cdot \exp(w_2 \cdot x).$$

The classification task used 12 datasets from [16]: Automobile, Breast Cancer Wisconsin (Diagnostic), Car Evaluation, Credit Approval, Glass Identification, Ionosphere, Iris, Tic-Tac-Toe Endgame, Congressional Voting Records, Wine, Zoo and Heart failure clinical records. The classification problem for each of them was solved using logistic regression from [17]. Averaging was performed on $B = 100$ bootstrap samples. All the curves are also reduced to the same scale on both axes. The resulting graphs are shown in Fig. 6. As before, there is a graph for the sample mean on the left, and a graph for the sample standard deviation on the right.

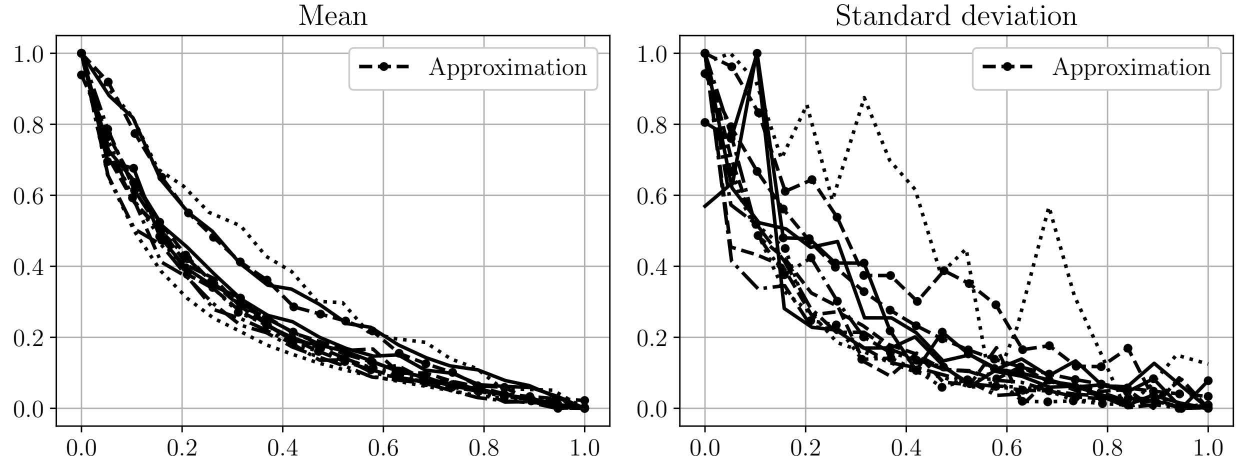

Figure 6. Behavior of the loss function in the classification problem

The application of a genetic algorithm for the mean value leads to the same family of functions as in the regression problem:

$$w_0 + w_1 \cdot \exp(w_2 \cdot x).$$

The standard deviation in the case of a classification problem for each sample has its own dependence on the sample size. Thus, predicting variance for classification turns out to be quite a difficult task.

5.2.2. Prediction of the likelihood function For synthetic samples, the approximation of likelihood functions is carried out. The mean and variance are approximated by the parametric family of functions given in the previous paragraph.

The division into training and test samples was carried out in the ratio of 70:30. The approximation was performed only on the training part. A sufficient sample size was in the test part. In Fig. 7 and Fig. 8 the true and approximated dependencies for synthetic data are presented. It also indicates the sample sizes determined by the true dependence D-sufficient and M-sufficient.

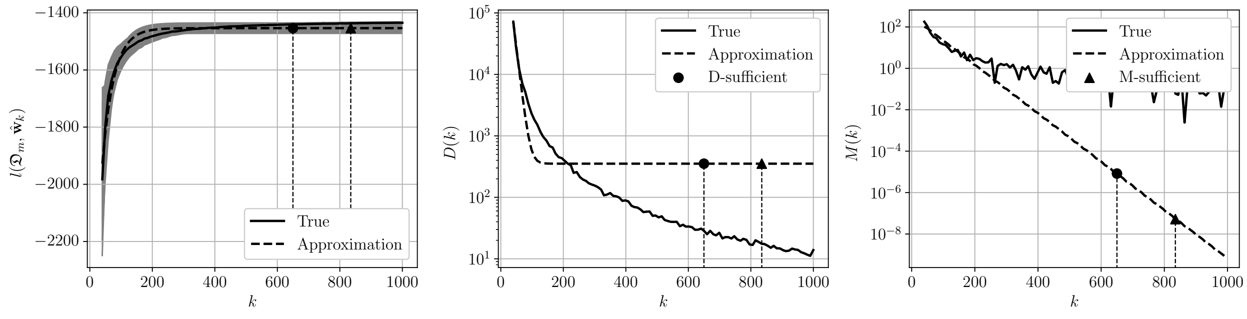

Figure 7. Synthetic sample (linear regression) at $m^* > m$

Next, similar results were obtained for the Abalone sample from [16]. They are shown in Fig. 9.

6. CONCLUSION

Approaches to determining a sufficient sample size based on the values of the likelihood function on the bootstrapped subsamples and the proximity of posterior distributions of model parameters on similar subsamples are proposed. The first two allow you to determine a sufficient sample size on any dataset with a regression or classification task. The correctness of the proposed approaches

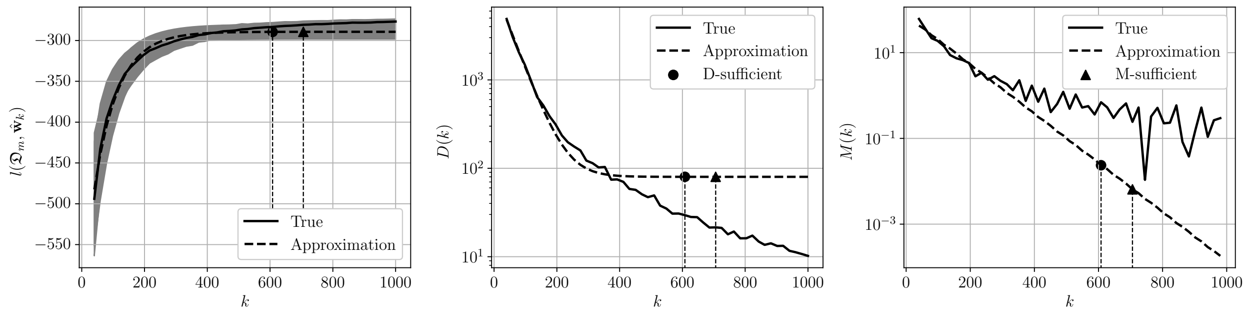

Figure 8. Synthetic sample (logistic regression) at $m^* > m$

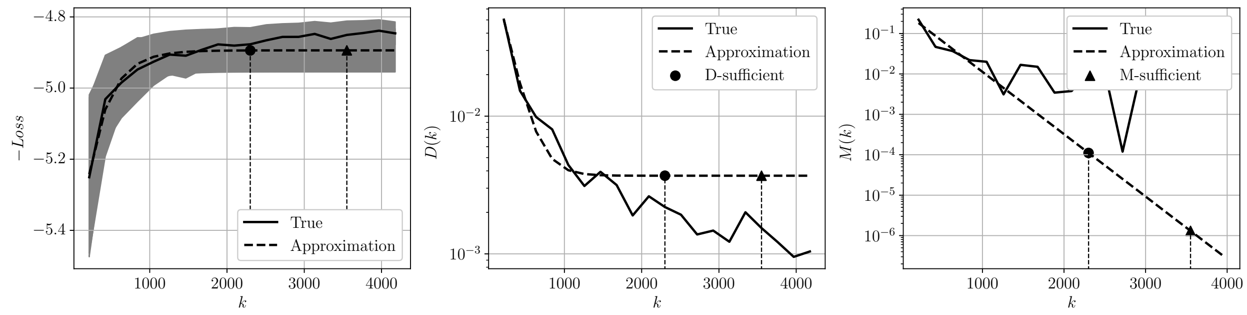

Figure 9. Abalone sample (regression) at $m^* > m$

is proved under certain restrictions on the model used, and a method for predicting the likelihood function in the case of insufficient sample size is proposed. The theorem on the moments of the limit posterior distribution of parameters in a linear regression model is proved. The conducted computational experiment makes it possible to analyze the properties of the proposed methods and their effectiveness. A parametric family of functions approximating the error function for a set of datasets is defined.

Appendix A: Proofs of theorems

Proof (Theorem 1). Consider the definition of an M-sufficient sample size in terms of the logarithm of the likelihood function. In a linear regression model

$$L\left(\mathfrak{D}_m, \hat{\mathbf{w}}_k\right) = p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{w}}_k) = \prod_{i=1}^{m} p(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_k) = \prod_{i=1}^{m} \mathcal{N}\left(y_i|\hat{\mathbf{w}}_k^\top \mathbf{x}_i, \sigma^2\right) =$$

$$= \left(2\pi\sigma^2\right)^{-m/2} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_k\|_2^2\right).$$

Take a logarithm:

$$l\left(\mathfrak{D}_m, \hat{\mathbf{w}}_k\right) = \log p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{w}}_k) = -\frac{m}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_k\|_2^2.$$

Let's take the mathematical expectation of $\mathfrak{D}_k$, given that $\mathbb{E}_{\mathfrak{D}_k}\hat{\mathbf{w}}_k = \mathbf{m}_k$ and $\text{cov}(\hat{\mathbf{w}}_k) = \mathbf{\Sigma}_k$:

$$\mathbb{E}_{\mathfrak{D}_k} l\left(\mathfrak{D}_m, \hat{\mathbf{w}}_k\right) = -\frac{m}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\Big(\|\mathbf{y} - \mathbf{X}\mathbf{m}_k\|_2^2 + \text{tr}\left(\mathbf{X}^\top\mathbf{X}\mathbf{\Sigma}_k\right)\Big).$$

Let's write down an expression for the difference in mathematical expectations:

$$\mathbb{E}_{\mathfrak{D}_{k+1}} l(\mathfrak{D}_m, \hat{\mathbf{w}}_{k+1}) - \mathbb{E}_{\mathfrak{D}_k} l(\mathfrak{D}_m, \hat{\mathbf{w}}_k) =$$

$$= \frac{1}{2\sigma^2}\Big(\|\mathbf{y} - \mathbf{X}\mathbf{m}_k\|_2^2 - \|\mathbf{y} - \mathbf{X}\mathbf{m}_{k+1}\|_2^2\Big) + \frac{1}{2\sigma^2}\mathrm{tr}\Big(\mathbf{X}^\top\mathbf{X}\left(\mathbf{\Sigma}_k - \mathbf{\Sigma}_{k+1}\right)\Big) =$$

$$= \frac{1}{2\sigma^2}\Big(2\mathbf{y}^\top\mathbf{X}(\mathbf{m}_{k+1} - \mathbf{m}_k) + (\mathbf{m}_k - \mathbf{m}_{k+1})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{m}_k + \mathbf{m}_{k+1})\Big) +$$

$$+ \frac{1}{2\sigma^2}\mathrm{tr}\Big(\mathbf{X}^\top\mathbf{X}\left(\mathbf{\Sigma}_k - \mathbf{\Sigma}_{k+1}\right)\Big).$$

The value of the function $M(k)$ is a module from the above expression. Let's apply the triangle inequality for the module, and then evaluate each term.
We estimate the first term using the Cauchy-Schwarz inequality:

$$\left|\mathbf{y}^\top\mathbf{X}(\mathbf{m}_{k+1} - \mathbf{m}_k)\right| \leqslant \|\mathbf{X}^\top\mathbf{y}\|_2\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2.$$

The second term is estimated using the Cauchy-Schwarzy inequality, the consistency property of the spectral matrix norm, as well as the limitation of the sequence of vectors $\mathbf{m}_k$, which follows from the presented convergence condition:

$$\left|(\mathbf{m}_k - \mathbf{m}_{k+1})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{m}_k + \mathbf{m}_{k+1})\right| \leqslant \|\mathbf{X}(\mathbf{m}_k - \mathbf{m}_{k+1})\|_2\|\mathbf{X}(\mathbf{m}_k + \mathbf{m}_{k+1})\|_2 \leqslant$$

$$\leqslant \|\mathbf{X}\|_2^2\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2\|\mathbf{m}_k + \mathbf{m}_{k+1}\|_2 \leqslant C\|\mathbf{X}\|_2^2\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2.$$

We estimate the last term using the Holder's inequality for the Frobenius norm:

$$\left|\mathrm{tr}\Big(\mathbf{X}^\top\mathbf{X}\left(\mathbf{\Sigma}_k - \mathbf{\Sigma}_{k+1}\right)\Big)\right| \leqslant \|\mathbf{X}^\top\mathbf{X}\|_F\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_{k+1}\|_F.$$

Finally, since $\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2 \to 0$ and $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_{k+1}\|_F \to 0$ as $k \to \infty$, then $M(k) \to 0$ as $k \to \infty$, which proves the theorem. □

Proof (Corollary). From the convergence conditions given, it follows that $\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2 \to 0$ and $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_{k+1}\|_F \to 0$ for $k \to \infty$. The application of the 1 theorem completes the proof. □

Proof (Theorem 2). The Kullback-Leibler divergence for a pair of normal posterior distributions has the form

$$D_{\mathrm{KL}}\left(p_k\|p_{k+1}\right) = \frac{1}{2}\left(\mathrm{tr}\left(\mathbf{\Sigma}_{k+1}^{-1}\mathbf{\Sigma}_k\right) + (\mathbf{m}_{k+1} - \mathbf{m}_k)^\top\mathbf{\Sigma}_{k+1}^{-1}(\mathbf{m}_{k+1} - \mathbf{m}_k) - n + \log\left(\frac{\det\mathbf{\Sigma}_{k+1}}{\det\mathbf{\Sigma}_k}\right)\right).$$

Let's express $\mathbf{\Sigma}_{k+1}$ as $\mathbf{\Sigma}_{k+1} = \mathbf{\Sigma}_k + \Delta\mathbf{\Sigma}$. Let's consider each term separately.

$$\mathrm{tr}\left(\mathbf{\Sigma}_{k+1}^{-1}\mathbf{\Sigma}_k\right) = \mathrm{tr}\left(\left(\mathbf{\Sigma}_k + \Delta\mathbf{\Sigma}\right)^{-1}\mathbf{\Sigma}_k\right) \to \mathrm{tr}\mathbf{I}_n = n \text{ as } \|\Delta\mathbf{\Sigma}\|_F \to 0,$$

$$\left|(\mathbf{m}_{k+1} - \mathbf{m}_k)^\top\mathbf{\Sigma}_{k+1}^{-1}(\mathbf{m}_{k+1} - \mathbf{m}_k)\right| \leqslant \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2^2\|\mathbf{\Sigma}_{k+1}^{-1}\|_2 \; to 0 \text{ as } \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0,$$

$$\log\left(\frac{\det\mathbf{\Sigma}_{k+1}}{\det\mathbf{\Sigma}_k}\right) = \log\left(\frac{\det\left(\mathbf{\Sigma}_k + \Delta\mathbf{\Sigma}\right)}{\det\mathbf{\Sigma}_k}\right) \to \log\det\mathbf{I}_n = \log 1 = 0 \text{ as } \|\Delta\mathbf{\Sigma}\|_F \to 0,$$

from where we have the required. □

Proof (Theorem 3). Let's use the s-score expression for a pair of normal posterior distributions from [15]:

$$\text{s-score}(p_k, p_{k+1}) = \exp\left(-\frac{1}{2}(\mathbf{m}_{k+1} - \mathbf{m}_k)^\top\left(\mathbf{\Sigma}_k + \mathbf{\Sigma}_{k+1}\right)^{-1}(\mathbf{m}_{k+1} - \mathbf{m}_k)\right).$$

Because

$$\left|(\mathbf{m}_{k+1} - \mathbf{m}_k)^\top\left(\mathbf{\Sigma}_k + \mathbf{\Sigma}_{k+1}\right)^{-1}(\mathbf{m}_{k+1} - \mathbf{m}_k)\right| \leqslant \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2^2\|\left(\mathbf{\Sigma}_k + \mathbf{\Sigma}_{k+1}\right)^{-1}\|_2 \to 0$$

if $\|\mathbf{m}_{k+1}-\mathbf{m}_k\|_2 \to 0$, then the value of the quadratic form inside the exponent tends to zero. Therefore, s-score$(p_k, p_{k+1}) \to 1$ as $\|\mathbf{m}_{k+1}-\mathbf{m}_k\|_2 \to 0$. □

Proof (Theorem 4). Let be a normal prior distribution of parameters $p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right)$. In a linear regression model, likelihood is normal, namely

$$p(\mathbf{y}|\mathbf{X},\mathbf{w}) = \mathcal{N}\left(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}\right) = \left(2\pi\sigma^2\right)^{-m/2}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2\right).$$

Using the conjugacy of the prior distribution and likelihood, it is easy to find the parameters of the posterior distribution:

$$p(\mathbf{w}|\mathbf{X},\mathbf{y}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m},\boldsymbol{\Sigma}\right),$$

where

$$\boldsymbol{\Sigma} = \left(\alpha\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}\right)^{-1}, \qquad \mathbf{m} = \left(\mathbf{X}^\top\mathbf{X} + \alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}.$$

Consider the expression $\|\boldsymbol{\Sigma}_{k+1}-\boldsymbol{\Sigma}_k\|_2$ norms of difference of covariance matrices for subsamples of size k and $k+1$. Let's introduce the notation $\mathbf{A}_k = \frac{1}{\sigma^2}\mathbf{X}_k^\top\mathbf{X}_k$. Given the formulas above, we have

$$\|\boldsymbol{\Sigma}_{k+1}-\boldsymbol{\Sigma}_k\|_2 = \left\|\left(\alpha\mathbf{I}+\mathbf{A}_{k+1}\right)^{-1} - \left(\alpha\mathbf{I}+\mathbf{A}_k\right)^{-1}\right\|_2 =$$

$$= \left\|\left(\alpha\mathbf{I}+\mathbf{A}_{k+1}\right)^{-1}\left(\mathbf{A}_{k+1}-\mathbf{A}_k\right)\left(\alpha\mathbf{I}+\mathbf{A}_k\right)^{-1}\right\|_2 \leqslant$$

Let's use the submultiplicativity of the spectral matrix norm.

$$\leqslant \left\|\left(\alpha\mathbf{I}+\mathbf{A}_{k+1}\right)^{-1}\right\|_2 \left\|\left(\alpha\mathbf{I}+\mathbf{A}_k\right)^{-1}\right\|_2 \|\mathbf{A}_{k+1}-\mathbf{A}_k\|_2 =$$

Now let's use the expression of the spectral matrix norm in terms of the maximum eigenvalue.

$$= \frac{1}{\lambda_{\min}\left(\alpha\mathbf{I}+\mathbf{A}_{k+1}\right)}\frac{1}{\lambda_{\min}\left(\alpha\mathbf{I}+\mathbf{A}_k\right)}\|\mathbf{A}_{k+1}-\mathbf{A}_k\|_2 \leqslant$$

$$\leqslant \frac{1}{\lambda_{\min}\left(\mathbf{A}_{k+1}\right)}\frac{1}{\lambda_{\min}\left(\mathbf{A}_k\right)}\|\mathbf{A}_{k+1}-\mathbf{A}_k\|_2 =$$

$$= \sigma^2\frac{1}{\lambda_{\min}\left(\mathbf{X}_{k+1}^\top\mathbf{X}_{k+1}\right)}\frac{1}{\lambda_{\min}\left(\mathbf{X}_k^\top\mathbf{X}_k\right)}\left\|\mathbf{X}_{k+1}^\top\mathbf{X}_{k+1}-\mathbf{X}_k^\top\mathbf{X}_k\right\|_2.$$

Further, since by the condition $\|\mathbf{x}\|_2 \leqslant M$, then

$$\left\|\mathbf{X}_{k+1}^\top\mathbf{X}_{k+1}-\mathbf{X}_k^\top\mathbf{X}_k\right\|_2 = \left\|\sum_{i=1}^{k+1}\mathbf{x}_i\mathbf{x}_i^\top - \sum_{i=1}^{k}\mathbf{x}_i\mathbf{x}_i^\top\right\|_2 = \left\|\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top\right\|_2 = \lambda_{\max}\left(\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top\right) =$$

A matrix of unit rank has a single nonzero eigenvalue.

$$= \mathbf{x}_{k+1}^\top\mathbf{x}_{k+1} = \|\mathbf{x}_{k+1}\|_2^2 \leqslant M^2.$$

By condition $\lambda_{\min}\left(\mathbf{X}_k^\top\mathbf{X}_k\right) = \omega(\sqrt{k})$, then $\|\boldsymbol{\Sigma}_{k+1}-\boldsymbol{\Sigma}_k\|_2 = o(k^{-1})$ as $k\to\infty$. Next, we will use the equivalence of matrix norms, namely

$$\|\boldsymbol{\Sigma}_{k+1}-\boldsymbol{\Sigma}_k\|_F \leqslant \sqrt{k}\|\boldsymbol{\Sigma}_{k+1}-\boldsymbol{\Sigma}_k\|_2 = o(k^{-1/2}) \text{ as } k\to\infty,$$

which was exactly what needed to be proved. Now let's estimate the norm of the difference in mathematical expectations.

$$\|\mathbf{m}_{k+1}-\mathbf{m}_k\|_2 = \left\|\left(\mathbf{X}_{k+1}^\top\mathbf{X}_{k+1}+\alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{X}_{k+1}^\top\mathbf{y}_{k+1}-\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{X}_k^\top\mathbf{y}_k\right\|_2 =$$

Consider that $\mathbf{X}_{k+1}^\top = [\mathbf{X}_k^\top, \mathbf{x}_{k+1}]$ and $\mathbf{y}_{k+1} = [\mathbf{y}_k, y_{k+1}]^\top$, then $\mathbf{X}_{k+1}^\top\mathbf{X}_{k+1} = \mathbf{X}_k^\top\mathbf{X}_k + \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top$ and $\mathbf{X}_{k+1}^\top\mathbf{y}_{k+1} = \mathbf{X}_k^\top\mathbf{y}_k + \mathbf{x}_{k+1}y_{k+1}$.

$$= \left\|\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}+\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top\right)^{-1}\left(\mathbf{X}_k^\top\mathbf{y}_k+\mathbf{x}_{k+1}y_{k+1}\right)-\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{X}_k^\top\mathbf{y}_k\right\|_2 =$$

Let's take out the multiplier in the first term:

$$\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}+\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top\right)^{-1} = \left(\mathbf{I}+\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top\right)^{-1}\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)^{-1}.$$

Next, we will take out the common multiplier for both terms.

$$= \left\|\left[\left(\mathbf{I}+\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top\right)^{-1}-\mathbf{I}\right]\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{X}_k^\top\mathbf{y}_k+\right.$$

$$\left.+\left(\mathbf{X}_{k+1}^\top\mathbf{X}_{k+1}+\alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{x}_{k+1}y_{k+1}\right\|_2 =$$

Let's use the triangle inequality, as well as the consistency and submultiplicativity property of the spectral norm.

$$\leqslant \left\|\left(\mathbf{I}+\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top\right)^{-1}-\mathbf{I}\right\|_2\left\|\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)^{-1}\right\|_2\left\|\mathbf{X}_k^\top\mathbf{y}_k\right\|_2+$$

$$+\left\|\left(\mathbf{X}_{k+1}^\top\mathbf{X}_{k+1}+\alpha\sigma^2\mathbf{I}\right)^{-1}\right\|_2\|\mathbf{x}_{k+1}y_{k+1}\|_2$$

Let's evaluate each term separately. In the first multiplier of the first term, we apply the formula for the difference of inverse matrices, as we did with covariance matrices.

$$\left\|\left(\mathbf{I}+\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top\right)^{-1}-\mathbf{I}\right\|_2 \leqslant$$

$$\leqslant \left\|\left(\mathbf{I}+\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top\right)^{-1}\right\|_2\cdot\|\mathbf{I}\|_2\cdot\left\|\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top\right\|_2 \leqslant$$

Again, we use submultiplicativity, as well as an expression for the norm of a matrix of unit rank.

$$\leqslant \frac{1}{\lambda_{\min}\left(\mathbf{I}+\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top\right)}\frac{\|\mathbf{x}_{k+1}\|_2^2}{\lambda_{\min}\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)} \leqslant$$

$$\leqslant \frac{1}{1+\lambda_{\min}\left(\left(\mathbf{X}_k^\top\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top\right)}\frac{M^2}{\lambda_{\min}\left(\mathbf{X}_k^\top\mathbf{X}_k\right)} \leqslant$$

The minimum eigenvalue of the product of matrices is estimated by the product of their minimum eigenvalues. In addition, the minimum eigenvalue of the matrix of unit rank $\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top$ is zero.

$$\leqslant \frac{1}{1+\lambda_{\max}\left(\mathbf{X}_k^\top\mathbf{X}_k + \alpha\sigma^2\mathbf{I}\right)\lambda_{\min}\left(\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top\right)} \frac{M^2}{\lambda_{\min}\left(\mathbf{X}_k^\top\mathbf{X}_k\right)} = \frac{M^2}{\lambda_{\min}\left(\mathbf{X}_k^\top\mathbf{X}_k\right)}.$$

The second and third multipliers of the first term are evaluated as follows.

$$\left\|\left(\mathbf{X}_k^\top\mathbf{X}_k + \alpha\sigma^2\mathbf{I}\right)^{-1}\right\|_2 \left\|\mathbf{X}_k^\top\mathbf{y}_k\right\|_2 \leqslant \frac{\left\|\mathbf{X}_k^\top\mathbf{y}_k\right\|_2}{\lambda_{\min}\left(\mathbf{X}_k^\top\mathbf{X}_k\right)} = \frac{\left\|\sum\limits_{i=1}^{k}\mathbf{x}_i y_i\right\|_2}{\lambda_{\min}\left(\mathbf{X}_k^\top\mathbf{X}_k\right)} \leqslant \frac{kM^2}{\lambda_{\min}\left(\mathbf{X}_k^\top\mathbf{X}_k\right)}$$

Finally, let's evaluate the second term.

$$\left\|\left(\mathbf{X}_{k+1}^\top\mathbf{X}_{k+1} + \alpha\sigma^2\mathbf{I}\right)^{-1}\right\|_2 \|\mathbf{x}_{k+1}y_{k+1}\|_2 \leqslant \frac{M^2}{\lambda_{\min}\left(\mathbf{X}_{k+1}^\top\mathbf{X}_{k+1}\right)}$$

In total, we have the following estimate.

$$\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \leqslant \frac{kM^3}{\lambda_{\min}^2\left(\mathbf{X}_k^\top\mathbf{X}_k\right)} + \frac{M^2}{\lambda_{\min}\left(\mathbf{X}_{k+1}^\top\mathbf{X}_{k+1}\right)} = k \cdot o(k^{-1}) + o(k^{-1/2}) = o(1) \text{ as } k \to \infty$$

Thus, we obtained the required convergence. □

REFERENCES

[1] C. J. Adcock. “A Bayesian approach to calculating sample sizes,”Journal of the Royal Statistical Society: Series D (The Statistician) **37** (4-5), 433–439 (1988).

[2] L. Joseph, D. B. Wolfson, and R. D. Berger. “Sample size calculations for binomial proportions via highest posterior density intervals,”Journal of the Royal Statistical Society. Series D (The Statistician), **44**(2), 143–154 (1995).

[3] D. V. Lindley. “The choice of sample size,”Journal of the Royal Statistical Society: Series D (The Statistician), **46**(2), 129–138 (1997).

[4] T. Pham-Gia. “On bayesian analysis, bayesian decision theory and the sample size problem,”Journal of the Royal Statistical Society: Series D (The Statistician), **46**(2), 139–144 (1997).

[5] A. E. Gelfand and F. Wang. “A simulation-based approach to bayesian sample size determination for performance under a given model and for separating models,”Statistical Science, **17** (2), (2002).

[6] J. Cao, J. J. Lee, and S. Alber. “Comparison of bayesian sample size criteria: Acc, alc, and woc,”Journal of Statistical Planning and Inference, **139**(12), 4111–4122 (2009).

[7] P. Brutti, F. De Santis, and S. Gubbiotti. “Bayesian-frequentist sample size determination: a game of two priors,”METRON, **72**(2), 133–151 (2014).

[8] H. Pezeshk, N. Nematollahi, V. Maroufy, and J. Gittins. “The choice of sample size: a mixed bayesian / frequentist approach,”Statistical Methods in Medical Research, **18**(2), 183–194 (2008).

[9] A. V. Grabovoy, T. T. Gadaev, A. P. Motrenko, and V. V. Strijov. “Numerical methods of sufficient sample size estimation for generalised linear models,”Lobachevskii Journal of Mathematics, **43**(9), 2453–2462 (2022).

[10] A. Motrenko, V. Strijov, and G.-W. Weber. “Sample size determination for logistic regression,”Journal of Computational and Applied Mathematics, **255**, 743–752 (2014).

[11] D. E. Goldberg and J. H. Holland. “Genetic algorithms and machine learning,”Machine Learning, **3**(2), 95–99 (1988).

[12] S. Mirjalili. “Genetic algorithm,”Evolutionary Algorithms and Neural Networks: Theory and Applications, 43–55 (2019).

[13] O. Kramer. “Genetic Algorithms,”Springer International Publishing, 11–19 (2017).

[14] L. Joseph, R. D. Berger, and P. B MΓ́elisle. “Bayesian and mixed bayesian/likelihood criteria for sample size determination,”Statistics in Medicine, **16**(7), 769–781 (1997).

[15] A. A. Aduenko. “Selection of multimodels in classification tasks,”Ph.d. diss., Moscow (2017).

[16] K. Markelle, L. Rachel, and N. Kolby. “The UCI machine learning repository. ”

[17] F. Pedregosa et. al. “Scikit-learn: Machine learning in Python,”Journal of Machine Learning Research, **12**, 2825–2830 (2011).