Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Киселев Никита Сергеевич

БАЙЕСОВСКИЙ ПОДХОД К ВЫБОРУ ОПТИМАЛЬНОГО РАЗМЕРА ВЫБОРКИ

03.03.01 — Прикладные математика и физика

Выпускная квалификационная работа бакалавра

Научный руководитель:
Грабовой Андрей Валериевич,
канд. физ.-мат. наук

Москва — 2024

Аннотация

При построении модели машинного обучения неизбежно возникает проблема сбора данных для ее обучения. Зачастую, по той или иной причине, естественное требование при этом — минимизировать количество таких данных. В настоящей работе исследуется задача определения достаточного размера выборки. Рассматривается проблема определения достаточного размера выборки без постановки статистической гипотезы о распределении параметров модели.

Предлагаются два подхода к определению достаточного размера выборки по значениям функции правдоподобия на подвыборках с возвращением. Эти подходы основываются на эвристиках о поведении функции правдоподобия при большом количестве объектов в выборке. Предлагаются два подхода к определению достаточного размера выборки на основании близости апостериорных распределений параметров модели на схожих подвыборках. Доказывается корректность представленных подходов при определенных ограничениях на используемую модель. Доказываются теоремы о моментах и полном байесовском прогнозе предельного апостериорного распределения параметров в модели линейной регрессии. Предлагается метод прогнозирования функции правдоподобия в случае недостаточного размера выборки. Проводится вычислительный эксперимент для анализа свойств предложенных методов.

Содержание

Введение

Задача машинного обучения с учителем предполагает выбор предсказательной модели из некоторого параметрического семейства. Обычно такой выбор связан с некоторыми статистическими гипотезами, например, максимизацией некоторого функционала качества. Модель, которая соответствует этим статистическим гипотезам, называется *адекватной* моделью [1, 2, 3].

При планировании вычислительного эксперимента требуется оценить минимальный размер выборки — количество объектов, необходимое для построения адекватной модели. Размер выборки, необходимый для построения адекватной модели прогнозирования, называется *достаточным* [4, 5, 6].

В работе рассматривается проблема определения достаточного размера выборки. Этой теме посвящено большое число работ. Используемые в них подходы можно разделить на статистические, байесовские и эвристические.

Одни из первых исследований по данной теме [7, 8] формулируют определенный статистический критерий, где связанный с данным критерием метод оценки размера выборки гарантирует достижение фиксированной статистической мощности с величиной ошибки первого рода, не превышающей заданного значения. К статистическим методам относятся метод множителей Лагранжа [9], метод проверки отношения правдоподобия [10], метод Вальда [11]. Статистические методы имеют ряд ограничений, которые связаны с их применением на практике. Они позволяют оценить размер выборки, исходя из предположений о распределении данных и информации о соответствии наблюдаемых величин предположениям нулевой гипотезы.

Байесовский подход тоже имеет место в данной проблеме. В работе [12] достаточный размер выборки определяется исходя из максимизации ожидаемой функции полезности. Она может включать в себя в явном виде функции распределения параметров и штрафы за увеличение размера выборки. Также в этой работе рассматриваются альтернативные подходы, основанные на ограничении некоторого критерия качества оценки параметров модели. Среди таких критериев можно выделить критерий средней апостериорной дисперсии APVC, критерий среднего покрытия ACC, критерий средней длины ALC и критерий эффективного объема выборки ESC. Эти критерии получили свое развитие в других работах, например, [13] и [14]. Спустя время, авторы [15] провели теоретическое и практическое сравнение методов из [7, 8, 12].

Авторы [16], как и [17], рассматривают различия между байесовским и частотным подходами при определении размера выборки. Также они предлагают робастные методы для байесовского подхода и приводят наглядные примеры для некоторых вероятностных моделей.

В работе [18] рассматриваются различные методы оценки размера выборки в обобщенных линейных моделях, включая статистические, эвристические и байесовские методы. Анализируются такие методы, как тест на множители Лагранжа, тест на отношение правдоподобия, статистика Вальда, кросс-валидация, бутстрап, критерий Кульбака-Лейблера, критерий средней апостериорной дисперсии, критерий среднего охвата, критерий средней длины и максимизация полезности. Авторы указывают на возможное развитие темы, которое заключается в поиске метода, сочетающего байесовский и статистический подходы для оценки размера выборки для недостаточного доступного размера выборки.

В [19] рассматривается метод определения размера выборки в логистической регрессии, использующий кросс-валидацию и дивергенцию Кульбака-Лейблера между апостериорными распределениями параметров модели на схожих подвыборках. Под схожими подвыборками понимают такие подвыборки, которые могут быть получены друг из друга добавлением, удалением или заменой одного объекта.

Генетический алгоритм [20] используется с целью аппроксимации заданного набора функций. Он представляет собой процесс оптимизации популяции кандидатов (называемых особями), который эволюционирует в сторону лучших решений [21]. Каждая особь имеет набор характеристик (генов или фенотипов), которые могут изменяться в процессе эволюции. Изменение происходит с помощью операции кроссинговера или мутации. Эволюция начинается со случайной популяции, и каждое поколение рассматривается как основа для генерации следующего. Приспособленность особей измеряется в каждом поколении, и особи с высокой приспособленностью выбираются для создания нового поколения [22]. Алгоритм завершается после достижения максимального числа поколений или достижения удовлетворительных результатов. Таким образом, каждое новое поколение становится более приспособленным к окружающей среде.

В настоящей работе рассматриваются несколько подходов к определению достаточного размера выборки. Предлагается оценивать математическое ожидание и дисперсию функции правдоподобия на бутстрапированных подвыборках.

Малое изменение этих величин при добавлении очередного объекта свидетельствует о достижении достаточного числа объектов в выборке. Доказывается корректность определения в модели линейной регрессии. Представленный метод легко использовать и на практике. Для этого предлагается подсчитывать значение функции ошибки, а не правдоподобия. Также в работе предлагается метод, который позволяет оценить достаточный размер выборки в случае, если данных объектов недостаточно. Используется генетический алгоритм для аппроксимации большого числа зависимостей функции ошибки от размера выборки на открытых датасетах с задачами регрессии и классификации.

Также в настоящей работе приводятся два подхода, основанных на расстоянии между апостериорными распределениями. Предлагается рассмотреть две подвыборки, вторая из которых получена добавлением одного объекта к первой. Апостериорные распределения параметров модели по этим подвыборкам оказываются близки, если размер выборки достаточен. Предлагается в качестве меры близости распределений использовать дивергенцию Кульбака-Лейблера [19], а также функцию сравнения моделей s-score [23]. Новизна данной работы заключается в доказательстве корректности предложенных методов. Корректность доказывается в вероятностной модели с нормальным апостериорным распределением параметров. Для модели линейной регрессии доказывается теорема о моментах предельного апостериорного распределения параметров.

1. Постановка задачи

Объектом называется пара $(\mathbf{x}, y)$, где $\mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}^n$ есть вектор признакового описания объекта, а $y \in \mathbb{Y}$ есть значение целевой переменной. В задаче регрессии $\mathbb{Y} = \mathbb{R}$, а в задаче K-классовой классификации $\mathbb{Y} = \{1, \ldots, K\}$.

Матрицей объекты-признаки для выборки $\mathfrak{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$ размера m называется матрица $\mathbf{X}_m = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^\intercal \in \mathbb{R}^{m \times n}$.

Вектором ответов (вектором значений целевой переменной) для выборки $\mathfrak{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$ размера m называется вектор $\mathbf{y}_m = [y_1, \ldots, y_m]^\intercal \in \mathbb{Y}^m$.

Моделью называется параметрическое семейство функций f, отображающих декартово произведение множества значений признакового описания объектов $\mathbb{X}$ и множества значений параметров $\mathbb{W}$ во множество значений целевой переменной $\mathbb{Y}$:

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

Вероятностной моделью называется совместное распределение вида

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

где $\mathbf{w} \in \mathbb{W}$ есть набор параметров модели, $p(y|\mathbf{x}, \mathbf{w})$ задает правдоподобие объекта, а $p(\mathbf{w})$ задает априорное распределение параметров.

Функцией правдоподобия простой выборки $\mathfrak{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$ размера m называется функция

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

Ее логарифм

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

называется логарифмической функцией правдоподобия. Далее, если не оговорено противное, будем считать выборку простой.

Оценкой максимума правдоподобия набора параметров $\mathbf{w} \in \mathbb{W}$ по подвы-

борке $\mathfrak{D}_k$ размера k называется

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

Ставится задача определения достаточного размера выборки m^*. Пусть задан некоторый критерий T. Он может быть построен, например, на основе эвристик о поведении параметров модели.

Определение 1. *Размер выборки m^* называется **достаточным** согласно критерию T, если T выполняется для всех $k \geqslant m^*$.*

Стоит учесть, что возможно $m^* \leqslant m$ или $m^* > m$. Эти два случая будут отдельно рассмотрены далее.

2. Достаточный размер выборки не превосходит доступный

В этом разделе будем считать, что достоверно $m^* \leqslant m$. Это означает, что нам нужно просто формализовать, какой размер выборки можно считать достаточным.

2.1. Анализ поведения функции правдоподобия

Для определения достаточности будем использовать функцию правдоподобия. Когда в наличии имеется достаточно объектов, вполне естественно ожидать, что от одной реализации выборки к другой полученная оценка параметров не будет сильно меняться [8, 24]. То же можно сказать и про функцию правдоподобия. Таким образом, сформулируем, какой размер выборки можно считать достаточным.

Определение 2. *Зафиксируем некоторое положительное число $\varepsilon > 0$. Размер выборки m^* называется* ***D-достаточным****, если для всех $k \geqslant m^*$*

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

С другой стороны, когда в наличии имеется достаточно объектов, также вполне естественно, что при добавлении очередного объекта в рассмотрение полученная оценка параметров не будет сильно меняться. Сформулируем еще одно определение.

Определение 3. *Зафиксируем некоторое положительное число $\varepsilon > 0$. Размер выборки m^* называется* ***M-достаточным****, если для всех $k \geqslant m^*$*

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

Замечание. *В определениях 2 и 3 вместо функции правдоподобия $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$ можно рассматривать ее логарифм $l(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$.*

Предположим, что $\mathbb{W} = \mathbb{R}^n$. Напомним, что информацией Фишера называется матрица

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

Известным результатом является асимптотическая нормальность оценки максимума правдоподобия, то есть $\sqrt{k}\,(\hat{\mathbf{w}}_k - \mathbf{w}) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}^{-1}(\mathbf{w})\right)$. Из сходимости

по распределению в общем случае не следует сходимость моментов случайного вектора. Тем не менее, если предположить последнее, то в некоторых моделях можно доказать корректность предложенного нами определения M-достаточного размера выборки.

Для удобства обозначим параметры распределения $\hat{\mathbf{w}}_k$ следующим образом: математическое ожидание $\mathbb{E}\hat{\mathbf{w}}_k = \mathbf{m}_k$ и матрица ковариации $\mathbb{D}\hat{\mathbf{w}}_k = \mathbf{\Sigma}_k$. Тогда имеет место следующая теорема, доказательство которой приведено в разделе 4.2.1..

Теорема 1 (Киселев, 2023). *Пусть* $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ *и* $\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_F \to 0$ *при* $k \to \infty$. *Тогда в модели линейной регрессии определение M-достаточного размера выборки является корректным. А именно, для любого* $\varepsilon > 0$ *найдется такой* m^*, *что для всех* $k \geqslant m^*$ *выполнено* $M(k) \leqslant \varepsilon$.

Следствие. *Пусть* $\|\mathbf{m}_k - \mathbf{w}\|_2 \to 0$ *и* $\|\mathbf{\Sigma}_k - [k\mathcal{I}(\mathbf{w})]^{-1}\|_F \to 0$ *при* $k \to \infty$. *Тогда в модели линейной регрессии определение M-достаточного размера выборки является корректным.*

По условию задана одна выборка. Поэтому в эксперименте нет возможности посчитать указанные в определениях математическое ожидание и дисперсию. Для их оценки воспользуемся техникой бутстрап. А именно, сгенерируем из заданной $\mathfrak{D}_m$ некоторое число B подвыборок размера k с возвращением. Для каждой из них получим оценку параметров $\hat{\mathbf{w}}_k$ и посчитаем значение $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$. Для оценки будем использовать выборочное среднее и несмещенную выборочную дисперсию (по бутстрап-выборкам).

Предложенные выше определения можно применять и в тех задачах, когда минимизируется произвольная функция потерь, а не максимизируется функция правдоподобия. Мы не приводим никаких теоретических обоснований этого, однако на практике такая эвристика оказывается достаточно удачной.

2.2. Анализ апостериорного распределения параметров модели

2.2.1. Сходимость апостериорных распределений

В работе [19] предлагается использовать дивергенцию Кульбака-Лейблера для оценки достаточного размера выборки в задаче бинарной классификации. Идея основывается на том, что если две подвыборки отличаются друг от друга

на один объект, то полученные по ним апостериорные распределения должны быть близки. Эта близость определяется дивергенцией Кульбака-Лейблера.

В настоящей работе предлагается развить этот подход, исследовать его не только в задаче классификации, но и в задаче регрессии. В качестве меры близости предлагается использовать не только дивергенцию Кульбака-Лейблера, но и функцию сходства s-score из [23].

Рассмотрим две подвыборки $\mathfrak{D}^1 \subseteq \mathfrak{D}_m$ и $\mathfrak{D}^2 \subseteq \mathfrak{D}_m$. Пусть $\mathcal{I}_1 \subseteq \mathcal{I} = \{1, \ldots, m\}$ и $\mathcal{I}_2 \subseteq \mathcal{I} = \{1, \ldots, m\}$ — соответствующие им подмножества индексов.

Определение 4. *Подвыборки $\mathfrak{D}^1$ и $\mathfrak{D}^2$ называются **схожими**, если $\mathcal{I}_2$ может быть получено из $\mathcal{I}_1$ удалением, заменой или добавлением одного элемента, то есть*

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

Рассмотрим две схожие подвыборки $\mathfrak{D}_k = (\mathbf{X}_k, \mathbf{y}_k)$ и $\mathfrak{D}_{k+1} = (\mathbf{X}_{k+1}, \mathbf{y}_{k+1})$ размеров k и $k+1$ соответственно. Это означает, что большая из них получена добавлением одного элемента к меньшей. Найдем апостериорное распределение параметров модели по этим подвыборкам:

$$p_j(\mathbf{w}) = p(\mathbf{w}|\mathfrak{D}_j) = \frac{p(\mathfrak{D}_j|\mathbf{w})p(\mathbf{w})}{p(\mathfrak{D}_j)} \propto p(\mathfrak{D}_j|\mathbf{w})p(\mathbf{w}), \quad j = k, k+1.$$

Определение 5. *Зафиксируем некоторое положительное число $\varepsilon > 0$. Размер выборки m^* называется **KL-достаточным**, если для всех $k \geqslant m^*$*

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

Для пары нормальных распределений дивергенция Кульбака-Лейблера имеет достаточно простой вид. Предположим, что апостериорное распределение является нормальным, то есть $p_k(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_k, \mathbf{\Sigma}_k)$. Руководствуясь эвристикой, что сходимость моментов такого распределения должна влечь за собой близость апостериорных распределений на схожих подвыборках, можно сформулировать следующее утверждение.

Теорема 2 (Киселев, 2024)**.** *Пусть $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ и $\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_F \to 0$ при $k \to \infty$. Тогда в модели с нормальным апостериорным распределением*

параметров определение KL-достаточного размера выборки является корректным. А именно, для любого $\varepsilon > 0$ найдется такой m^*, что для всех $k \geqslant m^*$ выполнено $KL(k) \leqslant \varepsilon$.

В настоящей работе предлагается в качестве меры сходства распределений использовать меру сходства s-score из [23]:

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

Определение 6. *Зафиксируем некоторое положительное число $\varepsilon > 0$. Размер выборки m^* называется* ***S-достаточным****, если для всех $k \geqslant m^*$*

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

Как и в случае KL-достаточного размера выборки, в модели с нормальным апостериорным распределением есть возможность записать выражение для используемого критерия. Таким образом, можно сформулировать еще одно утверждение.

Теорема 3 (Киселев, 2024)**.** *Пусть $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ при $k \to \infty$. Тогда в модели с нормальным апостериорным распределением параметров определение S-достаточного размера выборки является корректным. А именно, для любого $\varepsilon > 0$ найдется такой m^*, что для всех $k \geqslant m^*$ выполнено $S(k) \geqslant 1 - \varepsilon$.*

Пусть в модели линейной регрессии задано нормальное априорное распределение параметров. По свойству сопряженности априорного распределения и правдоподобия апостериорное распределение также является нормальным. Таким образом, мы приходим к одному из простейших примеров модели, для которой справедливы теоремы, представленные выше. На самом деле для линейной регрессии можно сформулировать более простые утверждения.

Теорема 4 (Киселев, 2024)**.** *Пусть множества значений признаков и целевой переменной ограничены, то есть $\exists M \in \mathbb{R} : \|\mathbf{x}\|_2 \leqslant M$ и $|y| \leqslant M$. Если $\lambda_{\min}(\mathbf{X}_k^{\intercal}\mathbf{X}_k) = \omega(\sqrt{k})$ при $k \to \infty$, то в модели линейной регрессии с нормальным априорным распределением параметров $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ и $\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_F \to 0$ при $k \to \infty$.*

2.2.2. Сходимость полного байесовского прогноза

В предыдущем разделе было показано, что при увеличении используемого размера выборки в модели линейной регрессии наблюдается сходимость апостериорных распределений параметров на схожих подвыборках. Однако на практике нас не сильно интересует вопрос устойчивости именно параметров используемой модели. Больший интерес представляет прогноз, который можно сделать на тестовой выборке, предварительно настроив модель на обучающей.

Пусть предварительно было произведено разбиение выборки $\mathfrak{D}_m$ на обучающую и тестовую, то есть

$$\mathfrak{D}_m = \mathfrak{D}_{m_1}^{\text{train}} \sqcup \mathfrak{D}_{m_2}^{\text{test}},$$

где $\mathfrak{D}_{m_1}^{\text{train}} = (\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$ и $\mathfrak{D}_{m_2}^{\text{test}} = (\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}})$.

Рассмотрим подвыборку $(\mathbf{X}_k, \mathbf{y}_k) \subset \mathfrak{D}_{m_1}^{\text{train}}$ и сформулируем теорему о близости полных байесовских прогнозов, сделанных на схожих подвыборках обучающей выборки.

Теорема 5 (Киселев, 2024)**.** *Пусть множества значений признаков и целевой переменной ограничены, то есть* $\exists M \in \mathbb{R} : \|\mathbf{x}\|_2 \leqslant M$ *и* $|y| \leqslant M$*. Если* $\lambda_{\min}\left(\mathbf{X}_k^{\intercal}\mathbf{X}_k\right) = \omega(\sqrt{k})$ *при* $k \to \infty$*, то в модели линейной регрессии с нормальным априорным распределением параметров*

$$\left\|\mathbb{E}\left[\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_{k+1}, \mathbf{y}_{k+1}\right] - \mathbb{E}\left[\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_k, \mathbf{y}_k\right]\right\|_2 \to 0,$$

$$\left\|\mathbb{D}\left[\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_{k+1}, \mathbf{y}_{k+1}\right] - \mathbb{D}\left[\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_k, \mathbf{y}_k\right]\right\|_F \to 0,$$

а потому и

$$D_{\text{KL}}\left(p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_k, \mathbf{y}_k) \| p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_{k+1}, \mathbf{y}_{k+1})\right) \to 0$$

при $k \to \infty$*.*

3. Достаточный размер выборки больше доступного

В этом разделе будем считать, что достоверно $m^* > m$.

Возникает задача прогнозирования математического ожидания функции правдоподобия / функции ошибки при $k > m$. В общем виде это достаточно трудная задача. В настоящей работе предлагается проанализировать большое число открытых датасетов из [25], чтобы найти параметрическое семейство функций, которыми стоит аппроксимировать зависимость функции ошибки от используемого размера выборки. Предлагается отдельно исследовать датасеты с задачами регрессии и классификации.

3.1. Генетический алгоритм в задаче аппроксимации набора функций

Одним из наиболее простых с точки зрения реализации и логики алгоритмов перебора является генетический алгоритм. С помощью него построим метод нахождения искомого семейства функций.

Пусть для N различных датасетов построен график зависимости среднего значения функции ошибки (или функции правдоподобия со знаком минус) от используемого размера выборки. Приведем эти N зависимостей к одинаковому масштабу по обеим осям. Для этого вычтем минимальное значение, а затем поделим на максимальное значение. В таком случае график каждой зависимости лежит в квадрате $[0;1]^2$, начинается в точке $(0;1)$ и заканчивается в точке $(1;0)$.

Популяцией в генетическом алгоритме является набор параметрических семейств функций. Например, одной особью может быть семейство $w_0 + w_1 \cdot \log(w_2 \cdot x) + w_3 \cdot x^2$, где x есть переменная, а $\mathbf{w}$ есть вектор параметров. Начальная популяция инициализируется случайным образом. Используются простейшие унарные функции: $1, x, \sin x, \cos x, \exp x, \log x, \operatorname{ctg} x$ и $\operatorname{cth} x$, а также простейшие бинарные функции: $+, -, *$ и $/$. Каждая особь представляется с помощью бинарного дерева, в узлах которого стоят вышеупомянутые функции, а листьями являются обязательно 1 или x. При этом за каждым узлом закрепляется своя компонента вектора параметров.

Приспособленность особи измеряется следующим образом. Для каждой из N аппроксимируемых зависимостей решается задача подбора вектора параметров. Минимизируется среднеквадратичное отклонение. Полученное значение MSE усредняется по всем N зависимостям. Итоговое значение определяет приспособленность особи.

Кроссинговер реализуется так, что случайное поддерево одного из особей-родителей заменяется случайным поддеревом другого. Мутация заменяет функцию в случайном узле дерева на другую случайную функцию.

Алгоритм завершается по прошествии заданного числа поколений. Выбирается особь из последнего поколения с наилучшей приспособленностью. Решением является соответствующее параметрическое семейство функций.

4. Вычислительный эксперимент

Проводится эксперимент для анализа свойств предложенных методов оценки достаточного размера выборки. Эксперимент состоит из нескольких частей. В первой части рассматриваются оценки достаточного размера выборки в случае, когда достаточный размер выборки не превосходит доступный. Во второй части исследуются результаты, полученные в условиях того, что достаточный размер выборки больше доступного.

4.1. Достаточный размер выборки не превосходит доступный

4.1.1. Бутстрапирование функции правдоподобия

Сходимость функций $D(k)$ и $M(k)$. Синтетические данные сгенерированы из моделей линейной и логистической регрессий. Число объектов 1000, число признаков 20. Используется $B = 1000$ бустрапированных подвыборок. Подсчитываются значения функций $D(k)$ и $M(k)$. Датасет с задачей регрессии Liver Disorders из [25] содержит 345 объектов и 5 признаков. Мы также используем $B = 1000$ бутстрапированных подвыборок для оценки математического ожидания и дисперсии функции ошибки.

На Рис. 1 можно видеть полученные зависимости между используемым размером подвыборки k и рассматриваемыми функциями $D(k)$ и $M(k)$ для синтетической выборки с задачей регрессии. Результаты для синтетической выборки с задачей классификации представлены на Рис. 2. В то же время, на Рис. 3 мы видим аналогичные графики для датасета Liver Disorders. Видно, что во всех случаях значения функций $D(k)$ и $M(k)$ стремятся к нулю при увеличении размера выборки. Эти эмпирические результаты подтверждают теоретические, полученные ранее.

Варьирование гиперпараметра ε для достаточного размера выборки. В определениях D-достаточности и M-достаточности участвует гиперпараметр ε, который отвечает за порог для достаточного размера выборки m^*. С целью изучения зависимости между ними, мы представляем Рис. 4, который демонстрирует, какой размер выборки следует выбрать, чтобы обеспечить определенный уровень уверенности.

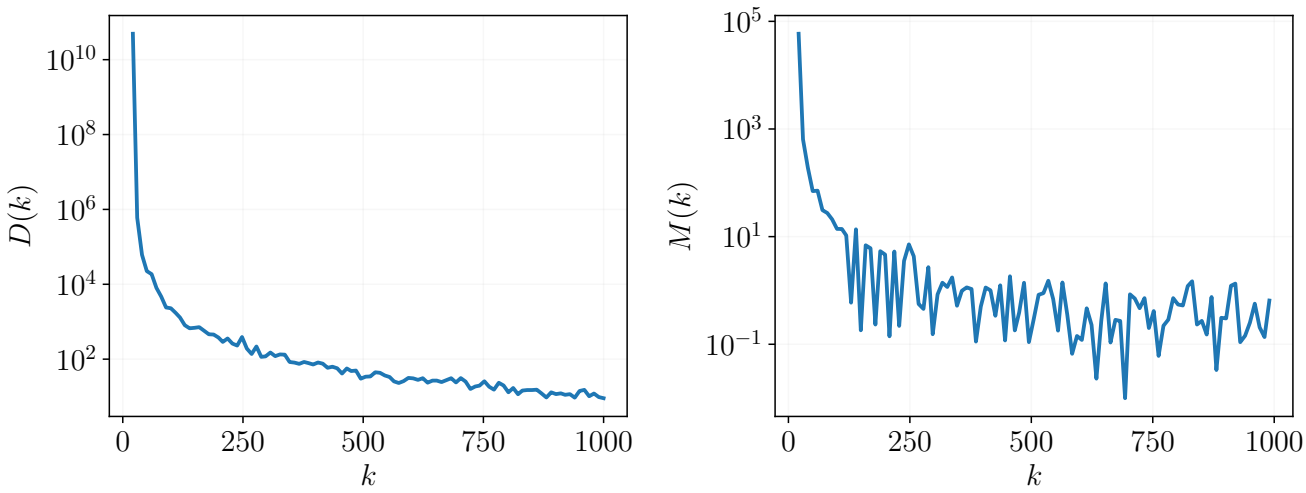

Рис. 1: Синтетическая выборка (линейная регрессия)

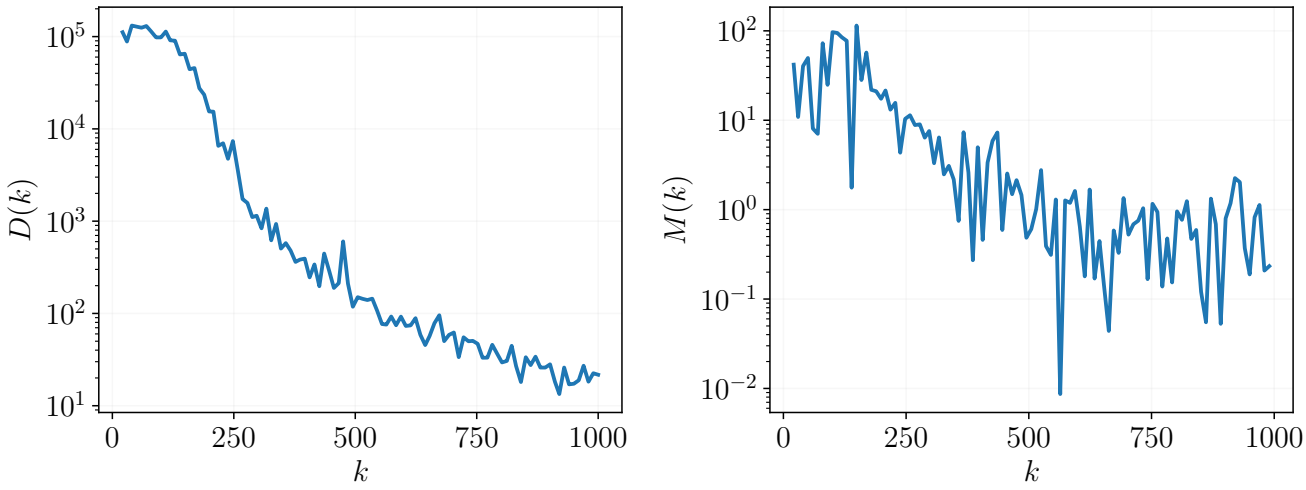

Рис. 2: Синтетическая выборка (логистическая регрессия)

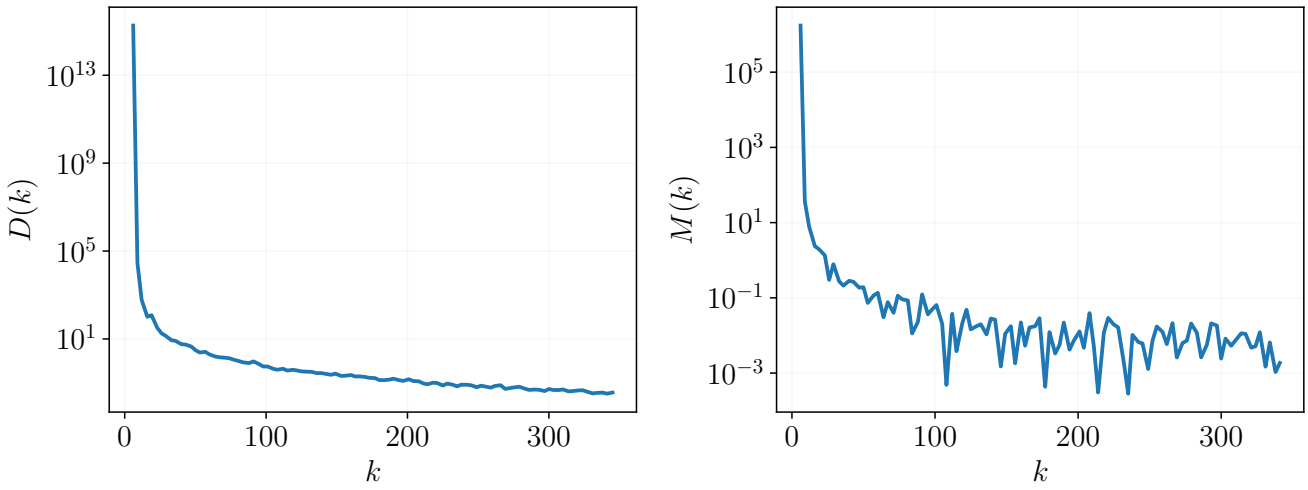

Рис. 3: Выборка Liver Disorders (регрессия)

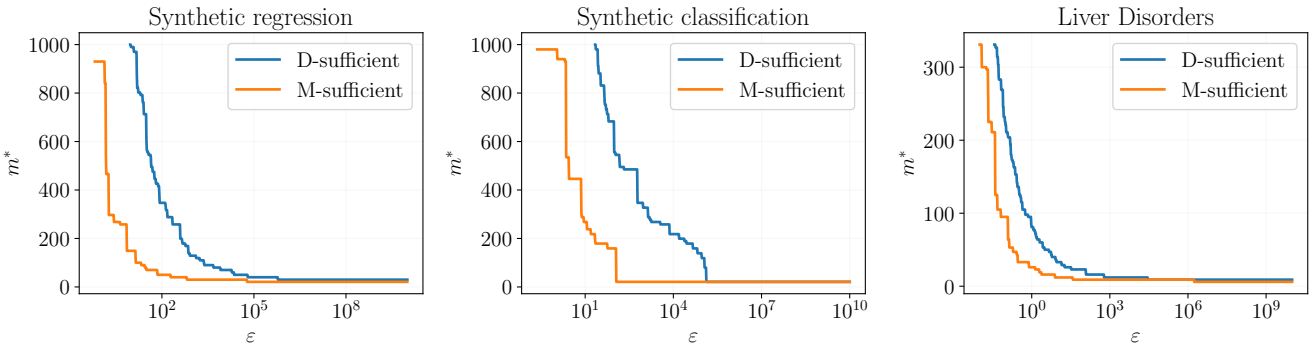

Рис. 4: Достаточный размер выборки в зависимости от гиперпараметра ε

4.1.2. Близость апостериорных распределений

Сходимость апостериорных распределений. Используется синтетическая выборка, сгенерированная из модели линейной регрессии. С целью упрощения визуализации рассматриваются одномерный и двумерный случаи. Число объектов 100, среднеквадратичное отклонение шума 1, априорное распределение параметров $\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{I})$. Далее представлены Рис. 5 и Рис. 6, на которых изображено сходство апостериорных распределений, определенных на схожих подвыборках размера k и $k+1$.

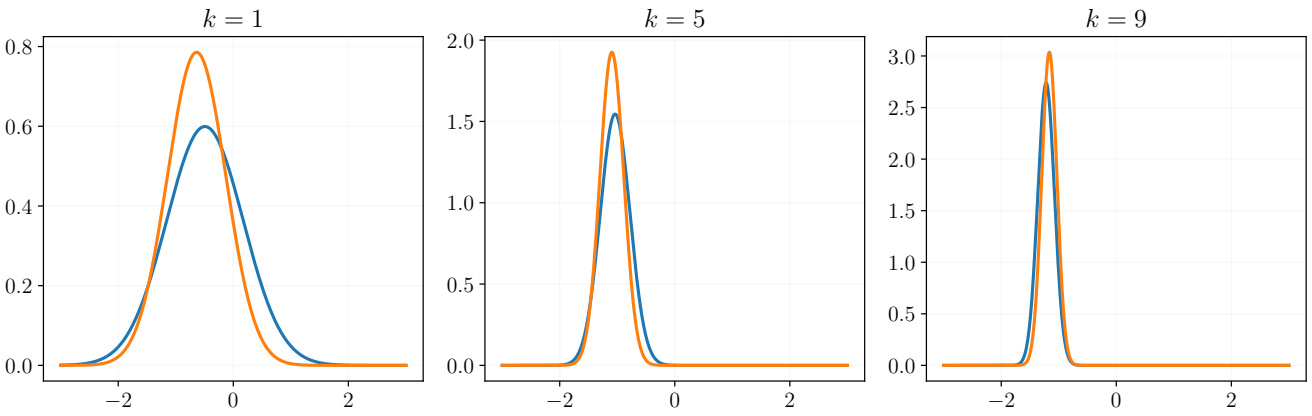

Рис. 5: Апостериорные распределения параметров на схожих подвыборках (одномерный случай — графики плотностей распределения)

Видно, что в обоих случаях распределение параметров становится схожим при росте доступного размера выборки. Эта интуиция и послужила фундаментом для предложенных в настоящей работе методов.

Асимптотическое поведение минимального собственного значения матрицы $\mathbf{X}^\intercal\mathbf{X}$. Синтетические данные сгенерированы из модели линейной регрессии. Число объектов 500, число признаков 10. Один объект последовательно

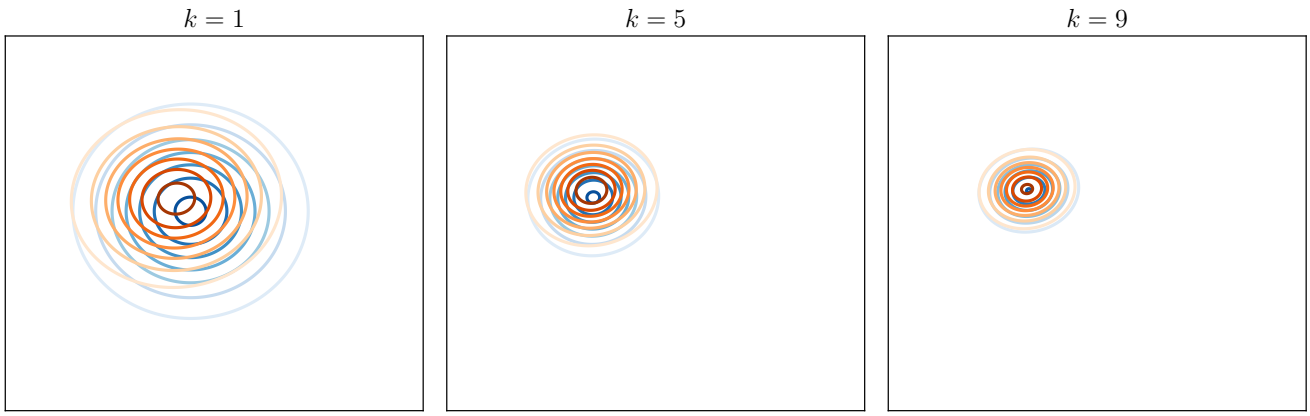

Рис. 6: Апостериорные распределения параметров на схожих подвыборках (двумерный случай — графики линий уровня)

удаляется из данной выборки, пока число объектов в подвыборке не станет равно числу признаков. Для каждого размера выборки k подсчитывается минимальное собственное значение матрицы $\mathbf{X}_k^\intercal \mathbf{X}_k$. Кроме того, подсчитываются значения функций $KL(k)$ и $S(k)$. Такой процесс повторяется $B = 100$ раз.

Датасет с задачей регрессии Liver Disorders из [25] имеет 345 объектов и 5 признаков. Мы аналогично удаляем объекты из выборки один за другим. Подсчитываются минимальное собственное значение и значения функций. Процесс повторяется $B = 1000$ раз.

Рис. 7 показывает асимптотическое поведение минимального собственного значения матрицы $\mathbf{X}_k^\intercal \mathbf{X}_k$. Видно, что при стремлении размера выборки к бесконечности минимальное собственное значение также стремится к бесконечности. Помимо этого, как и требуется в Теореме 4, этот график лежит выше, чем $\sqrt{k}$.

Сходимость функций $KL(k)$ и $S(k)$. На Рис. 8 мы можем наблюдать полученные зависимости между доступным размером выборки k и предложенными функциями $KL(k)$ и $S(k)$ для синтетической выборки с задачей регрессии. В то же время, на Рис. 9 можно видеть аналогичные графики для датасета Liver Disorders. Для обеих выборок значения функции $KL(k)$ стремятся к нулю при увеличении размера выборки, а значения $S(k)$ стремятся к единице. Эти эмпирические результаты подтверждают теоретические, представленные ранее.

Варьирование гиперпараметра ε для достаточного размера выборки. В определениях KL-достаточности и S-достаточности участвует гиперпараметр ε, который отвечает за порог для достаточного размера выборки m^*. С целью изуче-

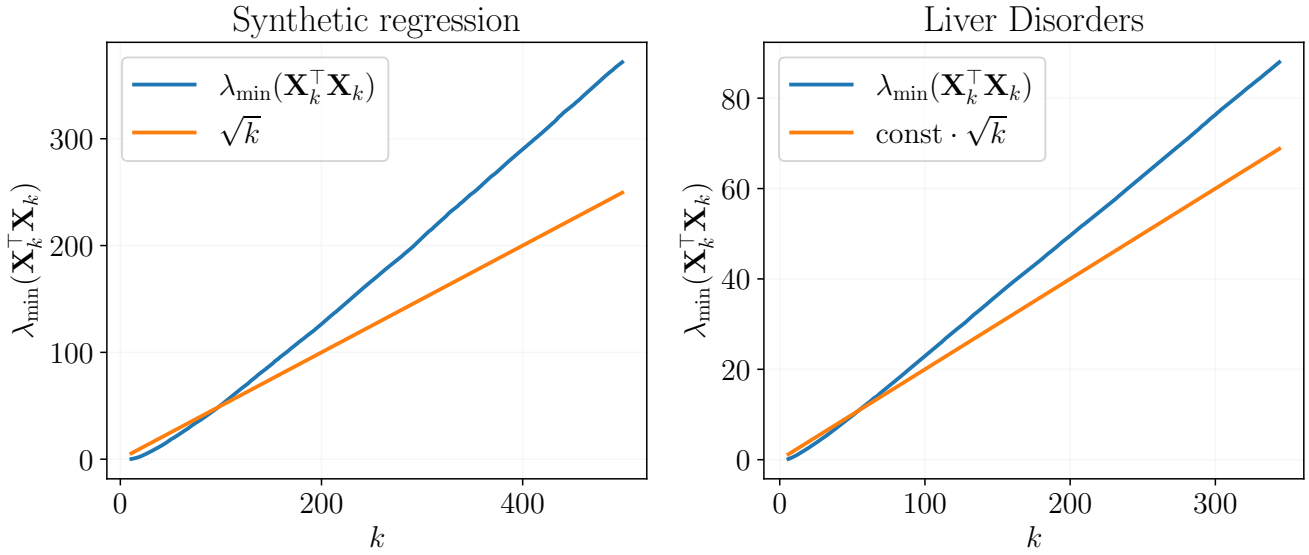

Рис. 7: Минимальное собственное значение матрицы $\mathbf{X}_k^\intercal \mathbf{X}_k$

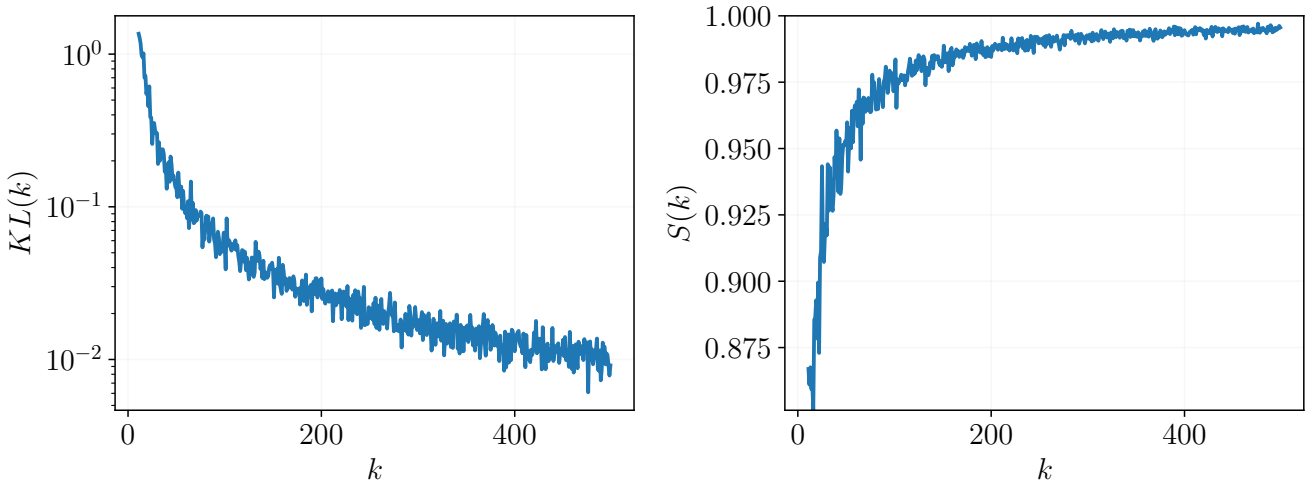

Рис. 8: Синтетическая выборка (линейная регрессия)

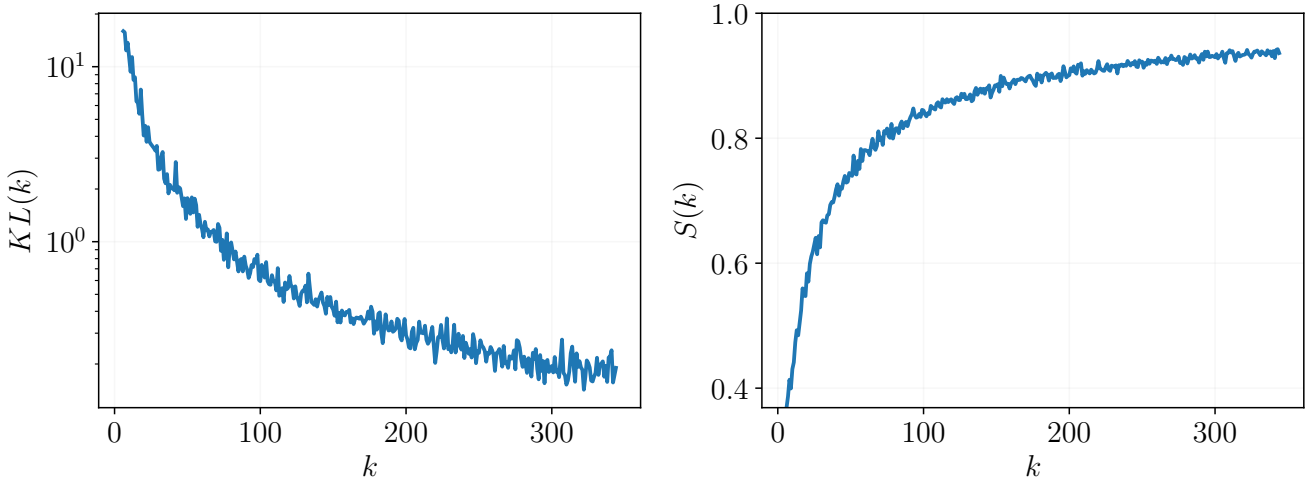

Рис. 9: Выборка Liver Disorders (регрессия)

ния зависимости между ними, мы представляем Рис. 10, который демонстрирует, какой размер выборки следует выбрать, чтобы обеспечить определенный уровень уверенности.

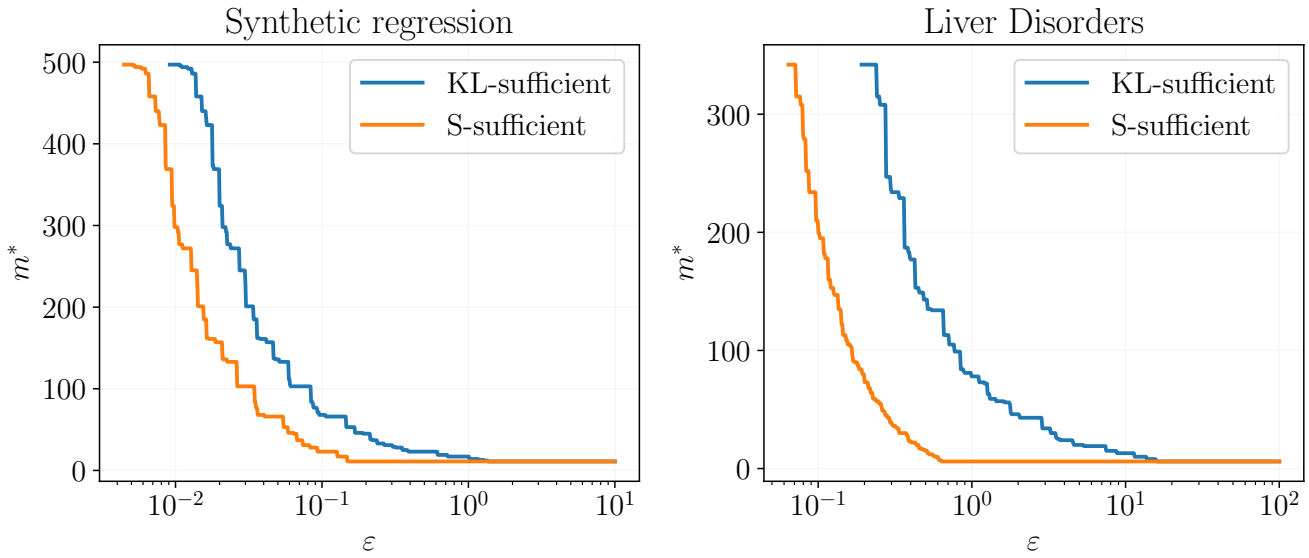

Рис. 10: Достаточный размер выборки в зависимости от гиперпараметра ε

Для $\varepsilon = 10$ достаточно взять число объектов, равное числу признаков. Однако для $\varepsilon = 10^{-2}$ требуется взять всю выборку.

Сравнение подходов на множестве выборок. Чтобы оценить эффективность предложенных методов на разных наборах данных, были выбраны выборки из открытой библиотеки [25]. Подробная информация о каждом наборе данных, количество наблюдений и количество признаков представлены в Таблице 1. Для демонстрационных целей были выбраны такие значения гиперпараметра ε, при которых значения функций $D(k)$ и $M(k)$ уменьшаются в 1000 раз, а значения функций $KL(k)$ и $S(k)$ приближаются к предельным вдвое. Соответствующие результаты приведены в Таблице 1. Пропуски означают, что первоначальный размер выборки недостаточен.

Таблица 1: Выборки с задачей регрессии (пропуски означают, что первоначальный размер выборки недостаточен)

Название выборки	Объектов m	Признаков n	D	M	KL	S
Abalone	4177	8	96	96	3921	4091
Auto MPG	392	8	15	15	62	—
Automobile	159	25	70	156	156	—
Liver Disorders	345	6	12	19	—	—
Servo	167	4	41	—	163	163
Forest fires	517	12	208	—	507	—
Wine Quality	6497	12	144	144	5305	6099
Energy Efficiency	768	9	24	442	—	—
Student Performance	649	32	129	177	636	—
Facebook Metrics	495	18	31	388	475	—
Real Estate Valuation	414	7	15	23	—	—
Heart Failure Clinical Records	299	12	63	224	276	293
Bone marrow transplant: children	142	36	—	—	109	—

4.2. Достаточный размер выборки больше доступного

4.2.1. Определение параметрического семейства функций с помощью генетического алгоритма

Реализацию генетического алгоритма, приведенного в разделе 3.1., можно найти в репозитории. Для исследования зависимости функции ошибки от используемого размера выборки в задаче регрессии использовались следующие датасеты из [25]: Abalone, Auto MPG, Liver Disorders, Wine Quality, Parkinsons Telemonitoring, Bike Sharing Dataset, Real estate valuation и Heart failure clinical records. Была выбрана квадратичная функция потерь MSE. Задача регрессии для каждого из них решалась с помощью линейной регрессии из [26]. Усреднение производилось по $B = 100$ бутстрап-выборкам. Как было сказано ранее, все зависимости приводятся к одинаковому масштабу по обеим осям. Полученные графики представлены на Рис. 11. Слева находится график для выборочного среднего. Справа находится график для выборочного среднеквадратичного отклонения.

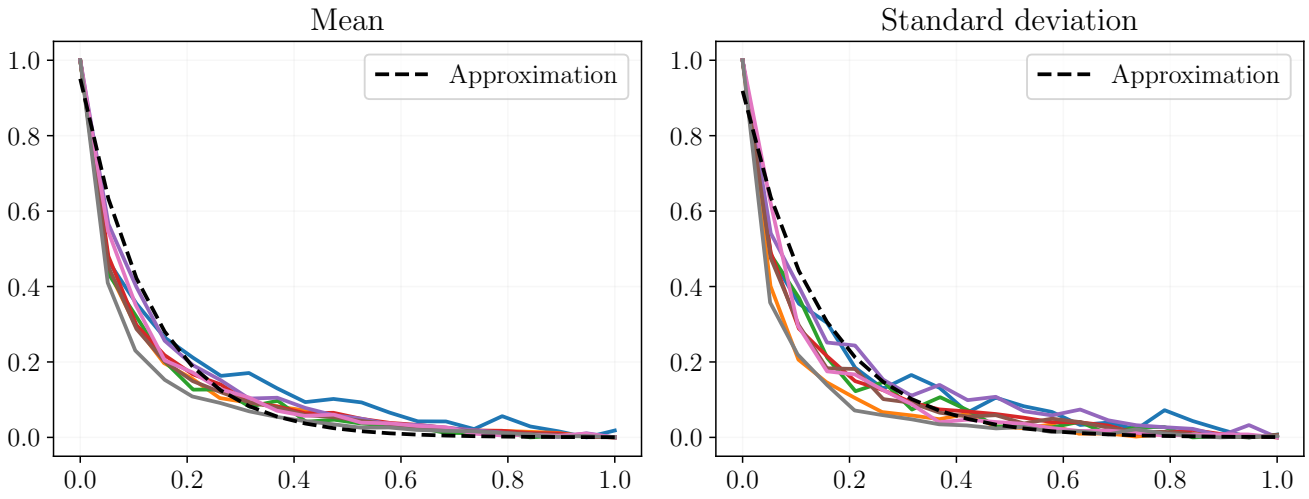

Рис. 11: Поведение функции ошибки в задаче регрессии

Применение генетического алгоритма приводит к одинаковому семейству функций для аппроксимации среднего и среднеквадратичного отклонения в задаче регрессии:

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

В задаче классификации использовалось 12 датасетов из [25]: Automobile, Breast Cancer Wisconsin (Diagnostic), Car Evaluation, Credit Approval, Glass Identification, Ionosphere, Iris, Tic-Tac-Toe Endgame, Congressional Voting Records,

Wine, Zoo и Heart failure clinical records. Задача классификации для каждого из них решалась с помощью логистической регрессии из [26]. Усреднение производилось по $B = 100$ бутстрап-выборкам. Все завимисости также приводятся к одинаковому масштабу по обеим осям. Полученные графики представлены на Рис. 12. Как и ранее, слева находится график для выборочного среднего, справа находится график для выборочного среднеквадратичного отклонения.

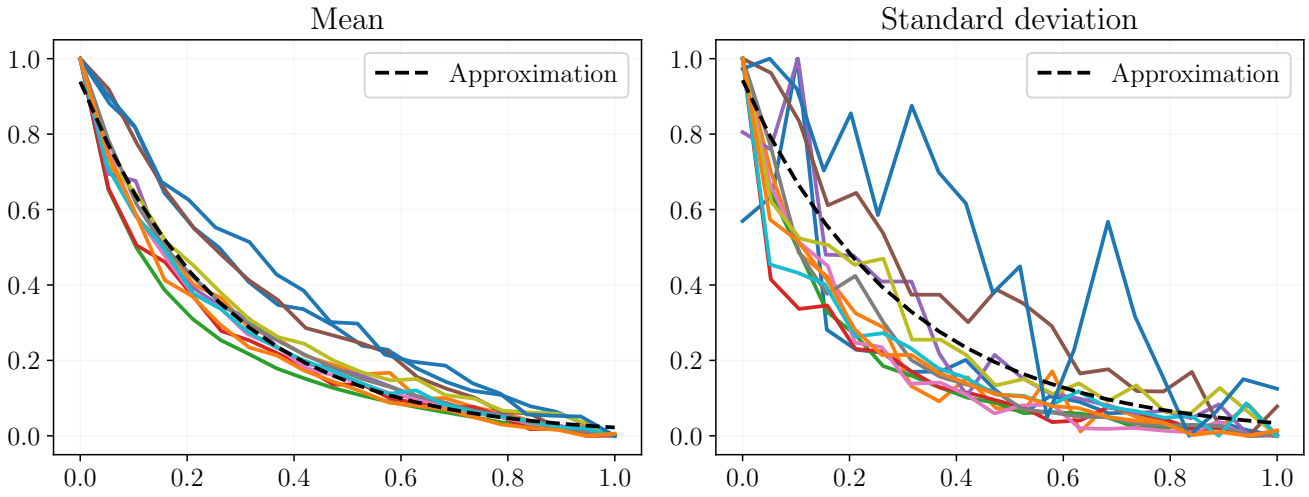

Рис. 12: Поведение функции ошибки в задаче классификации

Применение генетического алгоритма для среднего значения приводит к такому же семейству функций, как и в задаче регрессии:

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

Среднеквадратичное отклонение в случае задачи классификации для каждой выборки имеет свою зависимость от размера выборки. Таким образом, прогнозировать дисперсию для классификации оказывается достаточно сложной задачей.

Заключение

В данной работе рассматривалась проблема выбора достаточного размера выборки. Заключается она в том, чтобы определить необходимое число объектов для построения адекватной модели машинного обучения. Несмотря на то, что понятие достаточности зачастую является неформальным и эвристическим, существующие методы позволяют определять достаточное число объектов. Однако большинство из них либо не имеют строгих математических обоснований, либо применимы только для проверки некоторой статистической гипотезы о распределении параметров модели.

Для построения универсального и применимого на практике критерия достаточности в работе было проведено исследование:

1. Значений функции правдоподобия на бутстрапированных подвыборках;
2. Расстояния между апостериорными распределениями параметров модели на схожих подвыборках.

Первая часть была посвящена выборочной оценке математического ожидания и дисперсии функции правдоподобия выборки. Значение полного правдоподобия в точке, полученной методом максимума правдоподобия на некоторой подвыборке, является случайной величиной, зависящей от использованной подвыборки. В настоящей работе был проведен анализ моментов этой случайной величины.

Второй подход, изученный в работе, отсылает к близости апостериорных распределений параметров модели, полученных на подвыборках, отличающихся на один объект. Эти распределения были сравнены между собой при помощи дивергенции Кульбака-Лейблера и функции близости s-score.

В результате проведенного анализа были предложены четыре новых подхода к определению достаточного размера выборки: D-, M-, KL- и S-достаточность. Согласно каждому из них, достаточность определяется как близость к нулю или единице соответствующей функции. При этом, если задана некая модель, подход можно считать корректным, если в ней при стремлении размера выборки к бесконечности и соответствующая функция устремляется к своему предельному значению.

Была показана корректность предложенных определений при определенных условиях на используемую вероятностную модель. Кроме того, были доказа-

ны теоремы о близости моментов и полных байесовских прогнозов в модели линейной регрессии с нормальным априорным распределением параметров. Для эмпирического подтверждения полученных теоретических результатов был проведен вычислительный эксперимент. Целью эксперимента стала проверка соответствующих сходимостей и сравнение предложенных подходов между собой.

Таким образом, результаты настоящей работы могут быть применимы и в практических задачах. Это возможно, в частности, потому, что подходы, использующие функцию правдоподобия, допускают ее замену на функцию ошибки на обучающей выборке, и это подтверждается численными экспериментами. Более того, в других двух подходах расстояние между распределениями может быть вычислено и сэмплированием, что позволит приспособить их к случаю произвольной модели.

Тем не менее существенным ограничением рассматриваемых подходов является требуемая в доказательствах линейность модели. Учитывая популярность и постоянное развитие нейронных сетей, появляется естественное желание получить критерий достаточности и для таких моделей, существенно нелинейных. Дальнейшие планы развития настоящей работы направлены именно в сторону нейронных сетей.

Список литературы

1. A genetic algorithm-based, hybrid machine learning approach to model selection / Robert R Bies, Matthew F Muldoon, Bruce G Pollock et al. // *Journal of pharmacokinetics and pharmacodynamics*. — 2006. — Vol. 33, no. 2. — P. 195.

2. *Cawley Gavin C, Talbot Nicola LC*. On over-fitting in model selection and subsequent selection bias in performance evaluation // *The Journal of Machine Learning Research*. — 2010. — Vol. 11. — Pp. 2079–2107.

3. *Raschka Sebastian*. Model evaluation, model selection, and algorithm selection in machine learning // *arXiv preprint arXiv:1811.12808*. — 2018.

4. Sample size selection in optimization methods for machine learning / Richard H Byrd, Gillian M Chin, Jorge Nocedal, Yuchen Wu // *Mathematical programming*. — 2012. — Vol. 134, no. 1. — Pp. 127–155.

5. Predicting sample size required for classification performance / Rosa L Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, Long H Ngo // *BMC medical informatics and decision making*. — 2012. — Vol. 12. — Pp. 1–10.

6. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review / Indranil Balki, Afsaneh Amirabadi, Jacob Levman et al. // *Canadian Association of Radiologists Journal*. — 2019. — Vol. 70, no. 4. — Pp. 344–353.

7. *Adcock C. J.* A Bayesian Approach to Calculating Sample Sizes // *The Statistician*. — 1988. — Vol. 37, no. 4/5. — P. 433.

8. *Joseph Lawrence, Wolfson David B., Berger Roxane Du*. Sample Size Calculations for Binomial Proportions via Highest Posterior Density Intervals // *Journal of the Royal Statistical Society. Series D (The Statistician)*. — 1995. — Vol. 44, no. 2. — Pp. 143–154.

9. *Self Steven G, Mauritsen Robert H*. Power/sample size calculations for generalized linear models // *Biometrics*. — 1988. — Pp. 79–86.

10. *Shieh Gwowen*. On power and sample size calculations for likelihood ratio tests

in generalized linear models // *Biometrics*. — 2000. — Vol. 56, no. 4. — Pp. 1192–1196.

11. *Shieh Gwowen*. On power and sample size calculations for Wald tests in generalized linear models // *Journal of Statistical Planning and Inference*. — 2005. — Vol. 128, no. 1. — Pp. 43–59.

12. *Lindley Dennis V.* The choice of sample size // *Journal of the Royal Statistical Society: Series D (The Statistician)*. — 1997. — jul. — Vol. 46, no. 2. — P. 129–138.

13. *Pham T.* On Bayesian analysis, Bayesian decision theory and the sample size problem // *Journal of the Royal Statistical Society: Series D (The Statistician)*. — 1997. — jul. — Vol. 46, no. 2. — P. 139–144.

14. *Gelfand Alan E., Wang Fei*. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models // *Statistical Science*. — 2002. — may. — Vol. 17, no. 2.

15. *Cao Jing, Lee J. Jack, Alber Susan*. Comparison of Bayesian sample size criteria: ACC, ALC, and WOC // *Journal of Statistical Planning and Inference*. — 2009. — dec. — Vol. 139, no. 12. — P. 4111–4122.

16. *Brutti Pierpaolo, De Santis Fulvio, Gubbiotti Stefania*. Bayesian-frequentist sample size determination: a game of two priors // *METRON*. — 2014. — may. — Vol. 72, no. 2. — P. 133–151.

17. The choice of sample size: a mixed Bayesian / frequentist approach / Hamid Pezeshk, Nader Nematollahi, Vahed Maroufy, John Gittins // *Statistical Methods in Medical Research*. — 2008. — apr. — Vol. 18, no. 2. — P. 183–194.

18. Numerical Methods of Sufficient Sample Size Estimation for Generalised Linear Models / A. V. Grabovoy, T. T. Gadaev, A. P. Motrenko, V. V. Strijov // *Lobachevskii Journal of Mathematics*. — 2022. — sep. — Vol. 43, no. 9. — P. 2453–2462.

19. *Motrenko Anastasiya, Strijov Vadim, Weber Gerhard-Wilhelm*. Sample size determination for logistic regression // *Journal of Computational and Applied Mathematics*. — 2014. — Vol. 255. — Pp. 743–752.

20. *Goldberg David E., Holland John H.* Genetic Algorithms and Machine Learning // *Machine Learning*. — 1988. — Vol. 3, no. 2. — Pp. 95–99.

21. *Mirjalili Seyedali.* Genetic Algorithm // Evolutionary Algorithms and Neural Networks: Theory and Applications. — Cham: Springer International Publishing, 2019. — Pp. 43–55.

22. *Kramer Oliver.* Genetic Algorithms // Genetic Algorithm Essentials. — Cham: Springer International Publishing, 2017. — Pp. 11–19.

23. *Адуенко Александр Александрович.* Выбор мультимоделей в задачах классификации.

24. *Joseph Lawrence, Berger Roxane Du, Bélisle Patrick.* Bayesian and mixed Bayesian/likelihood criteria for sample size determination // *Statistics in Medicine*. — 1997. — Vol. 16, no. 7. — Pp. 769–781.

25. *Markelle Kelly, Rachel Longjohn, Kolby Nottingham.* The UCI Machine Learning Repository. — URL: https://archive.ics.uci.edu.

26. Scikit-learn: Machine Learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort et al. // *Journal of Machine Learning Research*. — 2011. — Vol. 12. — Pp. 2825–2830. — URL: https://scikit-learn.org.

27. *Murphy Kevin P.* Probabilistic Machine Learning: An introduction. — MIT Press, 2022.

Приложение

Доказательство (Теорема 1). Рассмотрим определение M-достаточного размера выборки в терминах логарифма функции правдоподобия. В модели линейной регрессии

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

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

Прологарифмируем:

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

Возьмем математическое ожидание по $\mathfrak{D}_k$, учитывая, что $\mathbb{E}\hat{\mathbf{w}}_k = \mathbf{m}_k$ и $\mathbb{D}\hat{\mathbf{w}}_k = \mathbf{\Sigma}_k$:

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

Запишем выражение для разности математических ожиданий:

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

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

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

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

Значение функции $M(k)$ есть модуль от вышеприведенного выражения. Применим неравенство треугольника для модуля, а затем оценим каждое слагаемое. Первое слагаемое оценим, используя неравенство Коши-Буняковского:

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

Второе слагаемое оценим, используя неравенство Коши-Буняковского, свойство согласованности спектральной матричной нормы, а также ограниченность последовательности векторов $\mathbf{m}_k$, которая следует из предъявленной в условии сходимости:

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

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

Последнее слагаемое оценим, используя неравенство Гельдера для нормы Фробениуса:

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

Наконец, поскольку $\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2 \to 0$ и $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_{k+1}\|_F \to 0$ при $k \to \infty$, то $M(k) \to 0$ при $k \to \infty$, что доказывает теорему. □

Доказательство (Следствие). Из приведенных в условии сходимостей следует, что $\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2 \to 0$ и $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_{k+1}\|_F \to 0$ при $k \to \infty$. Применение Теоремы 1 заканчивает доказательство. □

Доказательство (Теорема 2). Дивергенция Кульбака-Лейблера для пары нормальных апостериорных распределений имеет вид

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

$$-n + \log\left(\frac{\det \mathbf{\Sigma}_{k+1}}{\det \mathbf{\Sigma}_k}\right)\Big).$$

Представим $\mathbf{\Sigma}_{k+1}$ как $\mathbf{\Sigma}_{k+1} = \mathbf{\Sigma}_k + \Delta\mathbf{\Sigma}$. Рассмотрим в отдельности каждое слагаемое.

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

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

при $\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$,

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

при $\Delta\boldsymbol{\Sigma}\|_F \to 0$, откуда и имеем требуемое. □

Доказательство (Теорема 3). Воспользуемся выражением s-score для пары нормальных априорных распределений из [23]:

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

Поскольку

$$\left|(\mathbf{m}_{k+1} - \mathbf{m}_k)^\intercal \left(\boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_{k+1}\right)^{-1} (\mathbf{m}_{k+1} - \mathbf{m}_k)\right| \leqslant$$

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

при $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$, то значение квадратичной формы внутри экспоненты стремится к нулю. Следовательно, $\text{s-score}(p_k, p_{k+1}) \to 1$ при устремлении $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$. □

Доказательство (Теорема 4). Пусть задано нормальное априорное распределение параметров $p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right)$. В модели линейной регрессии правдоподобие является нормальным, а именно

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

Используя сопряженность априорного распределения и правдоподобия, легко найти параметры апостериорного распределения:

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

где

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

Рассмотрим выражение $\|\boldsymbol{\Sigma}_{k+1} - \boldsymbol{\Sigma}_k\|_2$ нормы разности матриц ковариции для подвыборок размера k и $k+1$. Введем обозначение $\mathbf{A}_k = \frac{1}{\sigma^2}\mathbf{X}_k^\intercal\mathbf{X}_k$. Учитывая формулы выше, имеем

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

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

Воспользуемся субмультипликативностью спектральной матричной нормы.

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

Теперь воспользуемся выражением спектральной матричной нормы через максимальное собственное значение.

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

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

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

Далее, поскольку по условию $\|\mathbf{x}\|_2 \leqslant M$, то

$$\|\mathbf{X}_{k+1}^{\intercal}\mathbf{X}_{k+1}-\mathbf{X}_{k}^{\intercal}\mathbf{X}_{k}\|_2 = \left\|\sum_{i=1}^{k+1}\mathbf{x}_i\mathbf{x}_i^{\intercal}-\sum_{i=1}^{k}\mathbf{x}_i\mathbf{x}_i^{\intercal}\right\|_2 = \|\mathbf{x}_{k+1}\mathbf{x}_{k+1}^{\intercal}\|_2 =$$

Матрица единичного ранга имеет единственное ненулевое собственное значение.

$$= \lambda_{\max}(\mathbf{x}_{k+1}\mathbf{x}_{k+1}^{\intercal}) = \mathbf{x}_{k+1}^{\intercal}\mathbf{x}_{k+1} = \|\mathbf{x}_{k+1}\|_2^2 \leqslant M^2.$$

По условию $\lambda_{\min}(\mathbf{X}_k^{\intercal}\mathbf{X}_k) = \omega(\sqrt{k})$, тогда $\|\mathbf{\Sigma}_{k+1}-\mathbf{\Sigma}_k\|_2 = o(k^{-1})$ при $k\to\infty$. Далее воспользуемся эквивалентностью матричных норм, а именно

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

что и требовалось доказать. Теперь оценим норму разности математических ожиданий.

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

Учтем, что $\mathbf{X}_{k+1}^{\intercal} = [\mathbf{X}_k^{\intercal}, \mathbf{x}_{k+1}]$ и $\mathbf{y}_{k+1} = [\mathbf{y}_k, y_{k+1}]^{\intercal}$, тогда $\mathbf{X}_{k+1}^{\intercal}\mathbf{X}_{k+1} = \mathbf{X}_k^{\intercal}\mathbf{X}_k + \mathbf{x}_{k+1}\mathbf{x}_{k+1}^{\intercal}$ и $\mathbf{X}_{k+1}^{\intercal}\mathbf{y}_{k+1} = \mathbf{X}_k^{\intercal}\mathbf{y}_k + \mathbf{x}_{k+1}y_{k+1}$.

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

Вынесем множитель в первом слагаемом:

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

$$\cdot \left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2\mathbf{I}\right)^{-1}.$$

Далее вынесем общий множитель у обоих слагаемых.

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

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

Воспользуемся неравенством треугольника, а также свойством согласованности и субмультипликативности спектральной нормы.

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

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

Оценим по отдельности каждое слагаемое. В первом множителе первого слагаемого применим формулу для разности обратных матриц, как мы делали с ковариационными матрицами.

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

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

Снова используем субмультипликативность, а также выражение для нормы матрицы единичного ранга.

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

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

Минимальное собственное значение произведения матриц оценивается произведением их минимальных собственных значений. Кроме того, минимальное собственное значение матрицы единичного ранга $\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\intercal$ равно нулю.

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

Второй и третий множители первого слагаемого оцениваются следующим образом.

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

Наконец, оценим второе слагаемое.

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

Итого, имеем следующую оценку.

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

при $k\to\infty$. Таким образом, получили требуемую сходимость. □

Доказательство (Теорема 5). Задано нормальное априорное распределение параметров $p(\mathbf{w})=\mathcal{N}\left(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I}\right)$. Для модели линейной регрессии

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

Апостериорное распределение параметров:

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

где

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

Полный байесовский прогноз для одиночной модели выражается по формуле

$$p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_k, \mathbf{y}_k) = \int p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{w}) p(\mathbf{w}|\mathbf{X}_k, \mathbf{y}_k) d\mathbf{w} =$$

$$= \int \mathcal{N}\left(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}\mathbf{w}, \sigma^2\mathbf{I}\right) \mathcal{N}\left(\mathbf{w}|\mathbf{m}_k, \mathbf{\Sigma}_k\right) d\mathbf{w}.$$

Полученный интеграл несложно вычислить, например, как в [27]:

$$= \mathcal{N}\left(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}\mathbf{m}_k, \sigma^2\mathbf{I} + \mathbf{X}_{\text{test}}\mathbf{\Sigma}_k\mathbf{X}_{\text{test}}^{\intercal}\right).$$

В условиях настоящей теоремы справедливы результаты Теоремы 4. Следовательно, $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ и $\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_2 \to 0$ при $k \to \infty$. Далее, остается показать сходимость матожидания и ковариационной матрицы нормального распределения прогноза модели. Для математического ожидания при $k \to \infty$ выполнено

$$\|\mathbf{X}_{\text{test}}\mathbf{m}_{k+1} - \mathbf{X}_{\text{test}}\mathbf{m}_k\|_2 \leqslant \|\mathbf{X}_{\text{test}}\|_2 \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0.$$

Для ковариационной матрицы при $k \to \infty$ справедливо следующее:

$$\|\sigma^2\mathbf{I} + \mathbf{X}_{\text{test}}\mathbf{\Sigma}_{k+1}\mathbf{X}_{\text{test}}^{\intercal} - \sigma^2\mathbf{I} - \mathbf{X}_{\text{test}}\mathbf{\Sigma}_k\mathbf{X}_{\text{test}}^{\intercal}\|_F \leqslant$$

Пользуясь субмультипликативностью нормы Фробениуса,

$$\leqslant \|\mathbf{X}_{\text{test}}\|_F^2 \|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_F \to 0.$$

Далее, имея сходимость матожидания и ковариационной матрицы, действуем аналогично доказательству Теоремы 2 и получаем, что

$$D_{\text{KL}}\left(p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_k, \mathbf{y}_k) \| p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_{k+1}, \mathbf{y}_{k+1})\right) \to 0$$

при $k \to \infty$. □