Восстановление сигналов методами стохастической оптимизации

Василевский Алексей, Григорянц Сергей, Потапов Георгий, Федорец Никита

Октябрь 2019

Аннотация

Мы реализуем методы, описанные в [JN19]

Введение:

Пусть $\omega^K = (\omega_1, ..., \omega_K)$, $\omega_k = (\eta_k, y_k)$, где $\eta \in \mathbf{R}^{n\times m}$, $y_k \in \mathbf{R}^m$ Предположим, что наблюдения описываются Generalized Linear Model, то есть условное математическое ожидание при условии η по y есть $f(\eta^T x)$, $f : R^m \to R^m$, где $x \in R^m$ - неизвестный вектор параметров. Цель - восстановить x по наблюдениям ω^K. Стандартный подход - выбрать функцию потерь и $l(y, \Theta) : \mathbf{R}^m \times \mathbf{R}^m \to \mathbf{R}$ и восстановить x, как оптимальное решение оптимизационной задачи

$$\min_{u\in\mathcal{X}} \mathbf{E}_{\omega\sim P_x}\{l(y, f(\eta^T u))\}, \tag{1}$$

где P_x - распределение наблюдений ω, а $\mathcal{X}$ заранее известное множество параметров. Другими словами, это сводится к задаче стохастической оптимизации (1), у которой неточное решение ищется через наблюдения ω^K. Это так же может быть сделано через усреднение

$$\frac{1}{K}\sum_{k=1}^{K} l(y_k, f(\eta_k^T u)) \tag{2}$$

математического ожидания (1) по $x \in \mathcal{X}$ методом *Sample Average Approximation* (SAA), или применив итеративный алгоритм стохастической оптимизации *Stochastic Approximation* (SA).
Предполагая, что условная вероятность при условии η y индуцировано P_x принадлежит известному параметрическому семейству $\mathcal{P} = \{P^\theta : \theta \in \Theta \subset \mathbf{R}^m\}$, в частности $P^x_{|\eta} = P^{f(\eta^T x)}$, стандартный выбор функции потерь определяется с помощью максимального правдоподобия: при условии, что распределение P имеет плотность p_θ, тогда

$$\ell(y, \theta) = -\ln(p_\theta(y)).$$

Например, в классической линейной регрессии $m = 1$, $f(s) = (1 + e^{-s})^{-1}$, $\Theta = (0, 1)$, и P^θ, $\theta \in \Theta$, это распределение Бернулли, то есть y принимает значение 1 с вероятностью $(1 + \exp\{-\eta^T x\})^{-1}$ и 0 с соответствующей, тогда

$$\ell(y, f(\eta^T u)) = \ln(1 + \exp\{\eta^T u\}) - y\eta^T u.$$

В этом случае, задача (1) становится задачей оптимизации

$$\min_{u\in\mathcal{X}} \mathbf{E}_{(\eta,y)\sim P_x} \left\{\ln(1+\exp\{\eta^T u\}) - y\eta^T u\right\}, \tag{3}$$

и ее SAA становится

$$\min_{u\in\mathcal{X}} \frac{1}{K}\sum_{k=1}^{K} \left[\ln(1+\exp\{\eta_k^T u\}) - y_k\eta_k^T u\right]; \tag{4}$$

Предполагая, что $\mathcal{X}$ выпуклая, обе эти задачи становятся выпуклыми, что подразумевает возможность глобального решения SAA достаточно эффективно, аналогично использованию хороших свойств сходимости SA.
Более обще, распределение наблюдений из экспоненциального семейства, отрицательная логарифмическая функция правдоподобия имеет вид

$$\{\ell(y,\eta^T u) = F(\eta^T u) - y\eta^T u,$$

с выпуклой функцией распределения F, и соответствующей минимизацией функции потерь

$$\min_{u\in\mathcal{X}} \mathbf{E}_{(\eta,y)\sim P_x} \left\{F(\eta^T u) - y\eta^T u\right\}.$$

В этом случае, так же как и в случае логистической регрессии, SAA или SA могут применяться для вычисления Оценки максимального правдоподобия параметра модели. Однако, что предположение на экспоненциальная семейство довольно сильное. Если

$$y = f(\eta^T x) + \xi,\ \xi \sim \mathcal{N}(0,\sigma^2 I_m).$$

$\ell(\cdot)$ становится

$$\min_{u\in\mathcal{X}} \mathbf{E}_{\eta\sim Q}\{\|f(\eta^T x) - f(\eta^T u)\|_2^2\}, \min_{u\in\mathcal{X}} \left\{\frac{1}{K}\sum_{k=1}^{k} \|y_k - f(\eta_k^T u)\|_2^2\right\},$$

где Q распределение η. Если f нелинейна, тогда обе задачи обычно не выпуклы. Цель следующего заключается в том, чтобы предложить альтернативу подгонке модели через (1) на основе ML подхода с использованием функции потерь для оценки параметров.

1.1 Необходимые определения и факты:

Непрерывное векторное поле $f : \mathbf{R}^m \to \mathbf{R}^m$ называется *строго монотонным* с модулем монотонности $\varkappa$ на множестве $Z \subset \mathbf{R}^m$, если для всех $z, z' \in Z$ выполняется:

$$\left[g(z) - g(z')\right]^T \left[z - z'\right] \geqslant \varkappa\|z - z'\|_2^2 \tag{5}$$

Будем говорить, что поле f — *строго монотонное* на Z, если его модуль монотонности положителен.

Для дальнейшей работы потребуются следующие свойства монотонных векторных полей:

I. Для монотонного векторного поля f в пространстве $\mathbf{R}^m$ и $n\times m$-матрицы A векторное поле

$$g(x) = Af(A^T x + a)$$

является монотонным полем в пространстве $\mathbf{R}^n$. Кроме того, если f монотонно с модулем монотонности $\varkappa \geqslant 0$ на замкнутом выпуклом множестве $Z \subset \mathbf{R}^m$, а замкнутое выпуклое множество $X \subset \mathbf{R}^n$ таково, что $A^T x + a \in Z$ при $x \in X$, то на множестве X поле g монотонно с модулем монотонности $\sigma^2\varkappa$, где σ — минимальное сингулярное число матрицы A.

II. Если S — польское пространство и $f(x,s) : \mathbf{R}^m \times S \to \mathbf{R}^m$ — борелевская функция, являющаяся монотонным полем при любом фиксированном $s \in S$, а $\mu(ds)$ — борелевская вероятностная мера на S, а векторное поле

$$F(x) = \int_S f(x,s)\mu(ds) \tag{6}$$

корректно определено при всех x, то поле F монотонно. Кроме того, если X — замкнутое выпуклое подмножество $\mathbf{R}^m$, а борелевская функция $\varkappa(s)$ такова, что при всех $s \in S$ поле $f(\cdot, s)$ монотонно на X с модулем монотонности $\varkappa(s)$, то поле F также монотонна на X, с модулем монотонности $\int_S \varkappa(s)\mu(ds)$.

Далее, пусть $\mathcal{X} \subset \mathbf{R}^m$ - непустой выпуклый компакт, и $\mathcal{H}$ -определённое на нём монотонное векторное поле. Вектор z_* называется *слабым решением* вариационного неравенства $\mathrm{VI}(\mathcal{X}, \mathcal{H})$, ассоциированного с если выполнено:

$$\mathcal{H}(z)^T(z - z_*) \geq 0, \ \forall x \in \mathcal{X} \tag{7}$$

Хорошо известно, что слабое решение всегда существует и при этом корень векторного поля им будет. Также в условиях непрерывности $\mathcal{H}$ будет верно, что слабое решение будет также и сильным решением, то есть:

$$\mathcal{H}(z_*)^T(z - z_*) \geq 0, \ \forall x \in \mathcal{X} \tag{8}$$

Также используется следующий факт:

Лемма 1.1 *Если в условиях определения VI($\mathcal{X}, \mathcal{H}$) поле $\mathcal{H}$ строго монотонно с модулем монотонности $\varkappa \geq 0$, то слабое решение z_* единственно и выполнено:*

$$\mathcal{H}(z)^T(z - z_*) \geq \varkappa||z - z_*||_2^2 \tag{9}$$

1.2 Допускаемые предположения:

Авторы решают задачу в следующих предположениях:

- **A.1.** Векторное поле f непрерывно и монотонно, а векторное поле

$$F(z) = \mathbf{E}_{\eta\sim Q}\{\eta f(\eta^T z)\} \tag{10}$$

корректно определено.

- **A.2.** $\mathcal{X}$ — непустое выпуклое компактное множество, а векторное поле F монотонно и его модуль монотонности $\varkappa$ положителен.

- **A.3.** Для некоторого $M < \infty$ и любого $x \in \mathcal{X}$ выполнено, что

$$\mathbf{E}_{(\eta,y)\sim P_x}\{\|\eta y\|_2^2\} < M^2 \tag{11}$$

Эти предположения выполнены, в частности, если все моменты случайной величины η конечны, а $\mathbf{E}_{\eta\sim Q}\{\eta\eta^T\} \succ \mathbf{0}$, а на поле f наложены следующие условия:

1. f непрерывно дифференцируема
2. $d^T f' d > 0$ ля всех z и всех $d \neq 0$
3. $\|f(z)\|_2 \leqslant C(1 + \|z\|_2^p)$ для некоторых $C \geqslant 0,\ p \geqslant 0$ и всех z

Основная идея:

Основная идея происходящего процесса заключается в том, что искомое $x \in \mathcal{X}$ является корнем следующего векторного поля:

$$G(z) = F(z) - F(x), \tag{12}$$

где F определено выше. Тогда из предположений **A.1-3** будет следовать, что G строго монотонно на $\mathcal{X}$, из чего будет следовать что искомый корень единственен. Далее можно заметить,что корень является слабым решением $\mathrm{VI}(G)$.
То есть если бы мы знали G, то мы могли бы найти решение $\mathrm{VI}(G)$ градиентными методами. То есть следующими операциями:

$$z_k = \mathrm{Proj}_{\mathcal{X}}[z_{k-1} - \gamma_k G(z_{k-1})],\ k = 1, 2, ..., K,$$

где

- $\mathrm{Proj}_{\mathcal{X}}[z] = \mathrm{argmin}_{u \in \mathcal{X}} \|z - u\|_2$;
- $\gamma_k > 0$ данный размер шага;
- начальная z_0 выбирается произвольно из $\mathcal{X}$.

Известно, что в предположениях **A.1-3**, этот метод дает сколь угодно хорошее приближение решения для достаточно большого K. Трудность заключается в том, что значения G неизвестны, а есть лишь реализации случайных величин. Поэтому необходимо ввести аналогичные векторные поля для наших наблюдений:

$$G_{\eta,y}(z) = \eta f(\eta^T z) - \eta y : \tag{13}$$

Лемма 2.1 $\forall x \in \mathcal{X}$ *верно:*

1. $\mathbf{E}_{(\eta,y)\sim P_x}\{G_{\eta,y}(z)\} = G(z)\ \forall z \in \mathbf{R}^n$
2. $||F(x)||_2 \leq M$
3. $\mathbf{E}_{(\eta,y)\sim P_x}\{||G_{\eta,y}(z)||_2^2\} \leq 4M^2\ \forall z \in \mathcal{X}$

Данная теория даёт возможность пользоваться двумя стандартными подходами.

2.1 Аппроксимация средними

Лемма (2.1) говорит о том, что $G_{\eta,y}(z)$ есть несмещённая оценка $G(z)$ с равномерно ограниченными по y, z конечными матожиданиями и дисперсией, поэтому поле

$$G_{\omega^K}(z) = \frac{1}{K}\sum_{k=1}^{K}[\eta_k f(\eta_k^T z) - \eta_k y_k] \tag{14}$$

равномерно сходится по распределению к $G(z)$ вследствие закона больших чисел. Тогда из строгой монотонности f и леммы (1.1) будет следовать строгая монотонность $G_{\omega^K}(z)$, откуда в свою очередь следует то, что асимптотически почти наверное слабое решение $\mathrm{VI}(G_{\omega^K}, \mathcal{X})$ сходится к слабому решению $\mathrm{VI}(G, \mathcal{X})$, а предпоследнее уже можно вычислить эффективно.

2.2 Стохастическая аппроксимация

Также можно воспользоваться стохастическим приближением, имея в виду несмещенность оценки $G_{(\eta_k,y_k)}(z)$ для $G(z)$. Тогда получим следующую реккуренту:

$$z_k = \mathrm{Proj}_{\mathcal{X}}[z_{k-1} - \gamma_k G_{(\eta_k,y_k)}(z_{k-1})],\ 1 \le k \le K, \tag{15}$$

Для нее верно следующее

Утверждение 2.1 *В предположении выполнения* ***A.1-3****, с шагами:*

$$\gamma_k = [\varkappa(k+1)]^{-1},\ k = 1, 2, ... \tag{16}$$

Для любого сигнала $x \in \mathcal{X}$ последовательность оценок $\widehat{x}_k(\omega^k) = z_k$, полученная из рекурренты (15) при $\omega_k = (\eta_k, y_k)$ для каждого k подчиняется закону:

$$\mathbf{E}_{\omega^k \sim P_x^k}\left\{\|\widehat{x}_k(\omega^k) - x\|_2^2\right\} \le \frac{4M^2}{\varkappa^2(k+1)},\ k = 0, 1, ... \tag{17}$$

Из этого утверждения мы можем сделать вывод, что просто реализовав данную рекурренту мы получаем хорошее приближение для решения.

Реализация и результаты

Мы рассмотрели методы следующую задачу:

- $\mathcal{X} = \{x \in \mathbf{R}^n : \|x\|_2 \le 2\}$;
- Распределение η – $\mathcal{N}(0, I_n)$;
- $f(x) = x^3 + x$;
- Распределение y при условии η – $\mathcal{N}(f(\eta^T x), I_n)$
- $x^* = \mathbf{1}$

Нижней оценкой на модуль непрерывности была выбрана 1, т.к. $f'(x) = 3x^2 + 1 \ge 1$. Заметим также, что $f''(x) = 6x$, и поэтому эта задача не является выпуклой на $\mathcal{X}$. Также рассмотрим еще задачу логистической регрессии:

- $\mathcal{X} = \{x \in \mathbf{R}^n : \|x\|_2 \le 2\}$;
- Распределение η – $\mathcal{N}(0, I_n)$;
- $f(x) = (1 + e^{-x})^{-1}$;
- Распределение y про условии η – $Bern(f(\eta^T x))$
- $x^* = \mathbf{1}$

Для логистической регрессии реализованные методы работают не так хорошо. Дело в том, что в f есть экспонента, то наблюдения могут давать для нее довольно большой разброс, также наблюдения будут равны 0 или 1 лишь с правильной частотой. Поэтому наблюдаемое поле может иметь корень, далёкий от корня искомого поля. Из-за этого мы увеличивали размер выборок в 10 раз относительно первой тестовой постановки.

3.1 Аппроксимация средними

Ниже приведены графики сходимости при использовании аналогов обычных градиентных методов нахождения корней векторного поля при аппроксимации средними, для размерности 2 в первой постановке и 5 во второй.

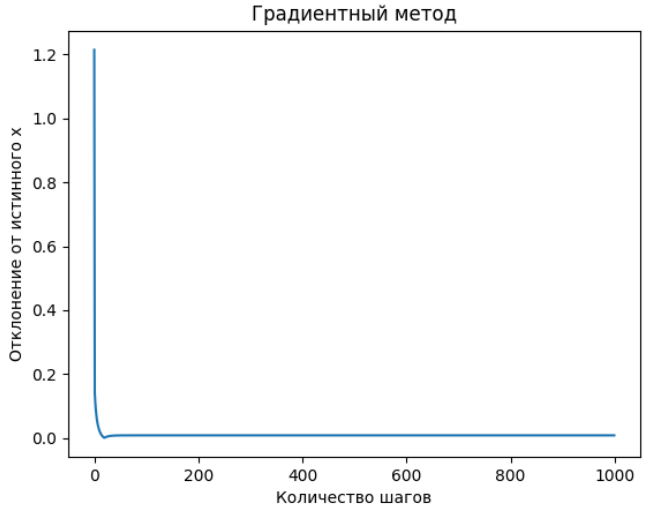


Рис. 1: Метод градиентного спуска для функции $x + x^3$ с x размерности 2

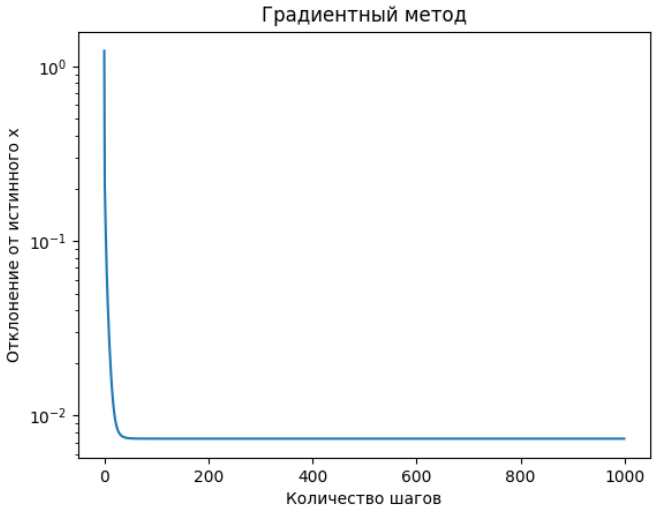


Рис. 2: Метод градиентного спуска для функции $x+x^3$ с x размерности 2 (логарифмическая шкала)

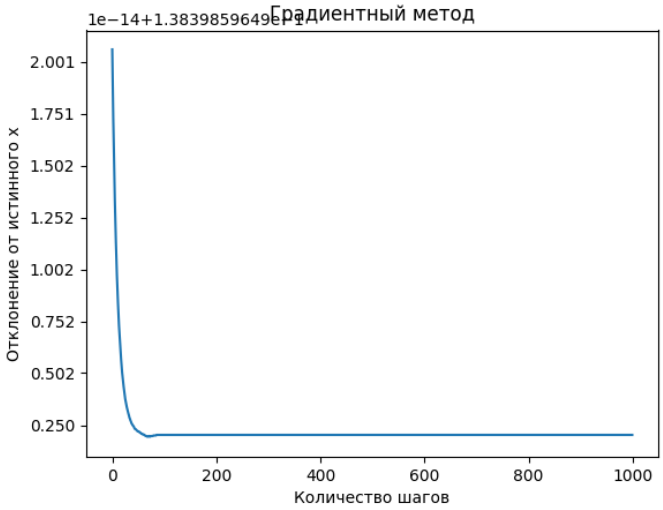


Рис. 3: Метод градиентного спуска для логистической регрессии с x размерности 5

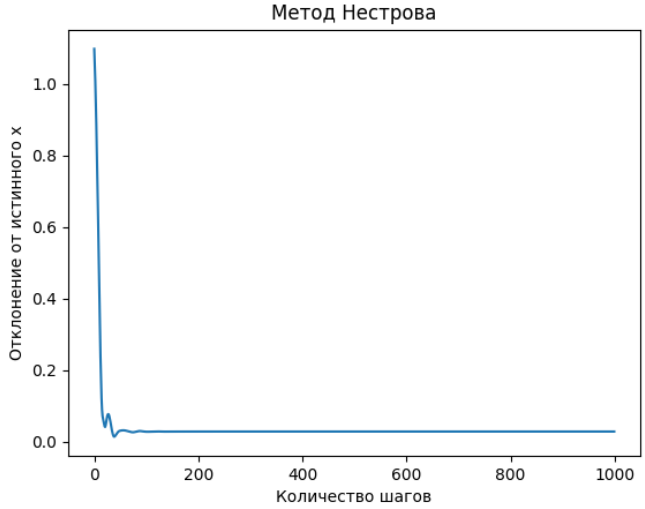


Рис. 4: Метод Нестерова для функции $x + x^3$ с x размерности 2

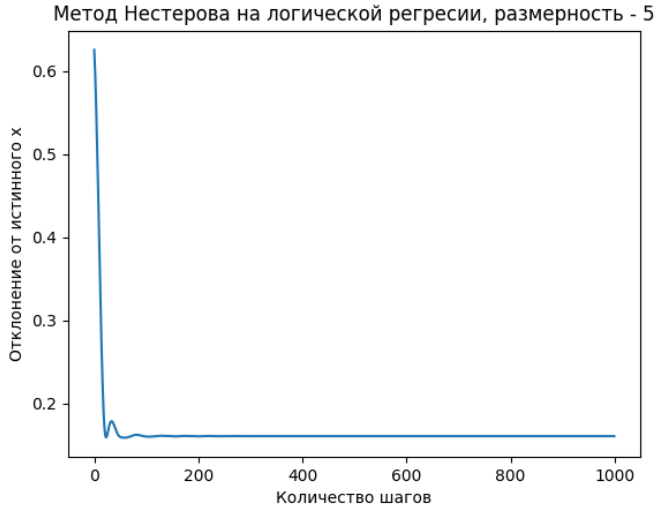


Рис. 5: Метод Нестерова для логистической регрессии с x размерности 5

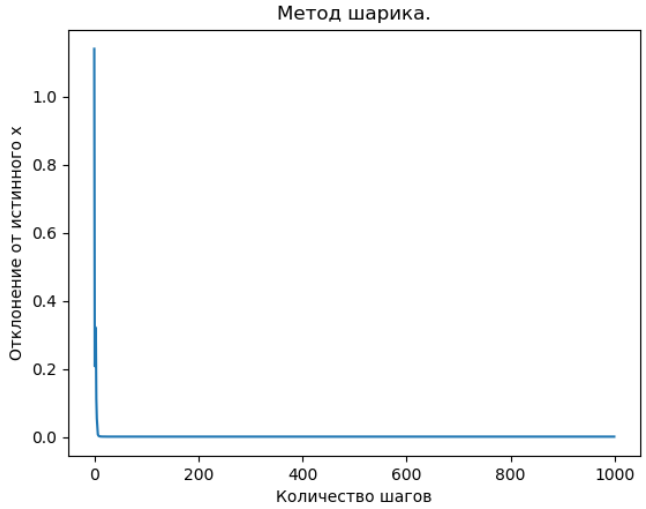


Рис. 6: Метод тяжелого шарика для функции $x + x^3$ с x размерности 2

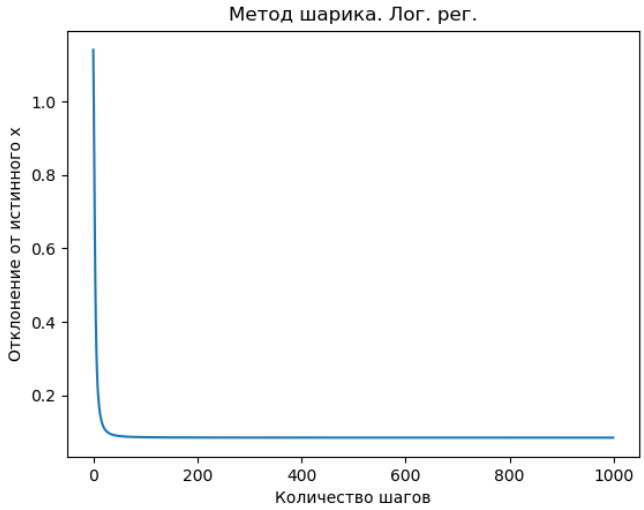


Рис. 7: Метод тяжелого шарика для логистической регрессии

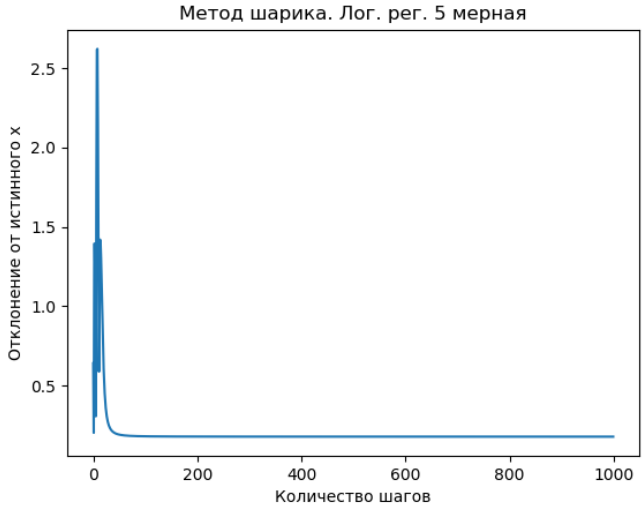


Рис. 8: Метод тяжелого шарика для логистической регрессии с x размерности 5

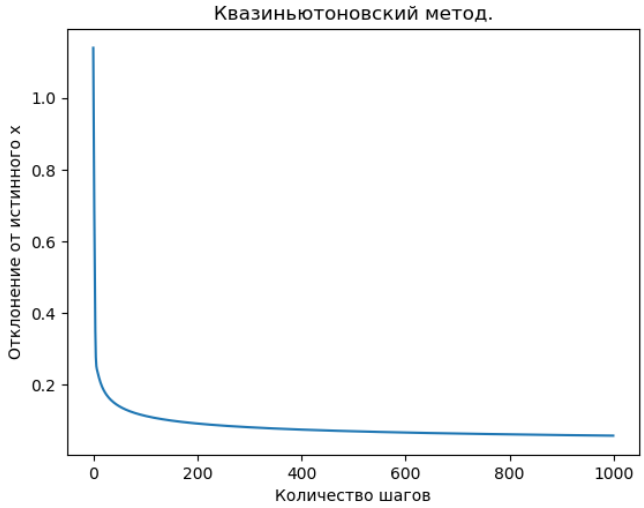


Рис. 9: Метод тяжелого шарика для функции $x + x^3$ с x размерности 2

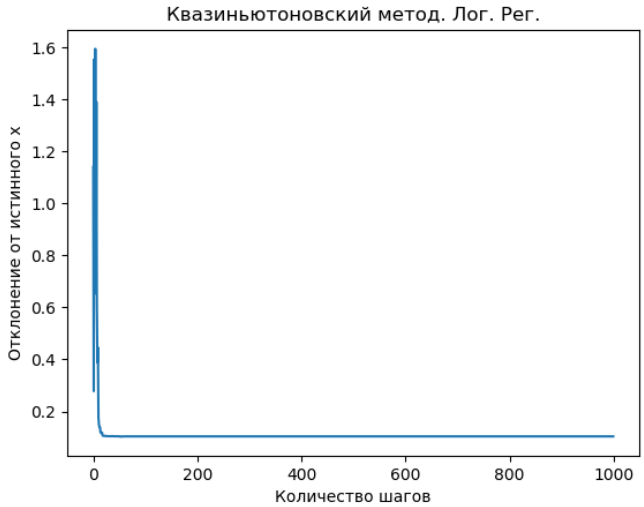


Рис. 10: Метод тяжелого шарика для логистической регрессии

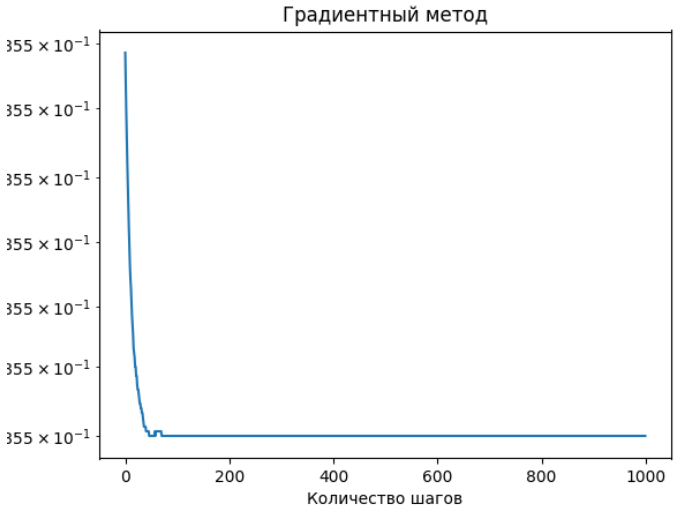


Рис. 11: Метод градиентного спуска для логистической регрессии с x размерности 5 (логарифмическая шкала)

3.2 Методы с усреднением в логарифмической шкале

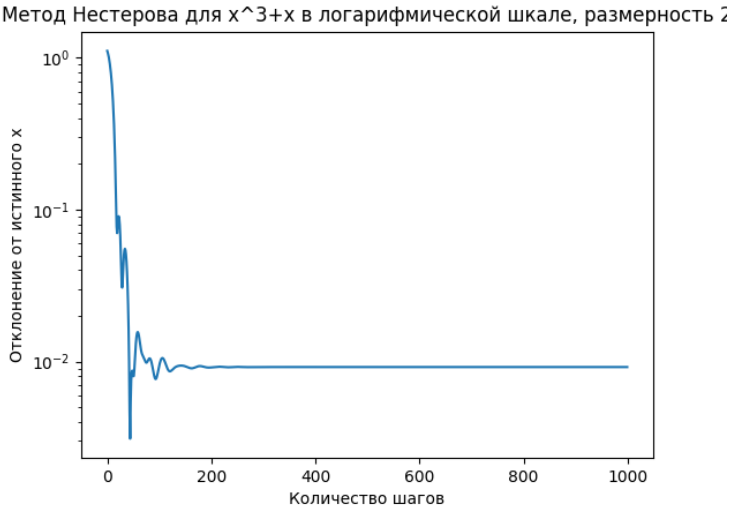


Рис. 12: Метод Нестерова для функции $x + x^3$ с x размерности 2, логарифмическая шкала

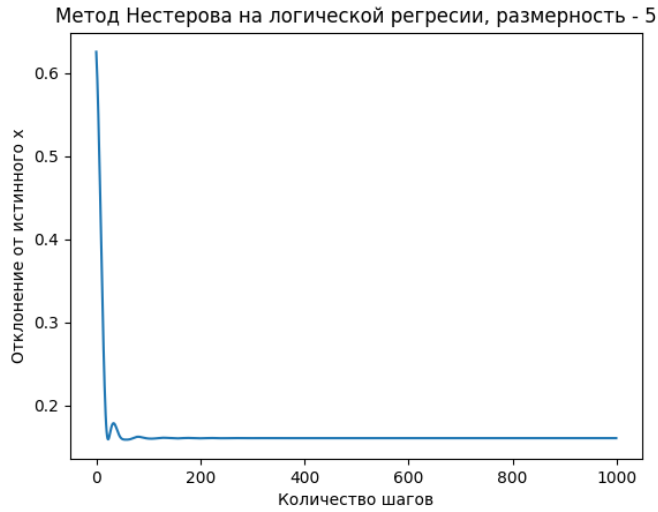


Рис. 13: Метод Нестерова для логистической регрессии с x размерности 5, логарифмическая шкала

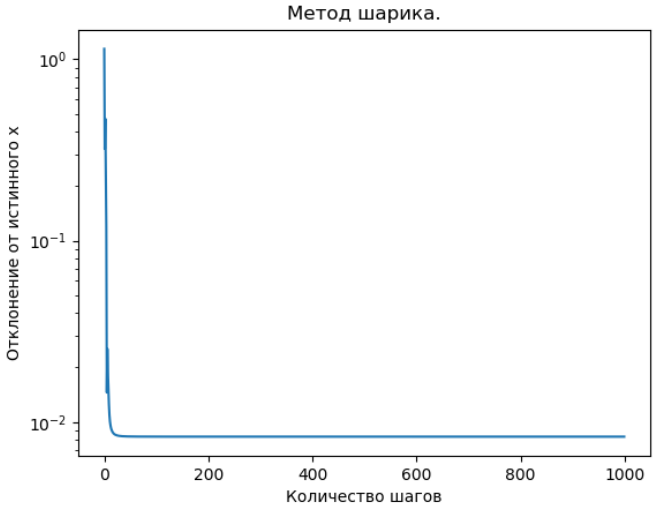


Рис. 14: Метод тяжелого шарика для функции $x + x^3$ с x размерности 2, логарифмическая шкала

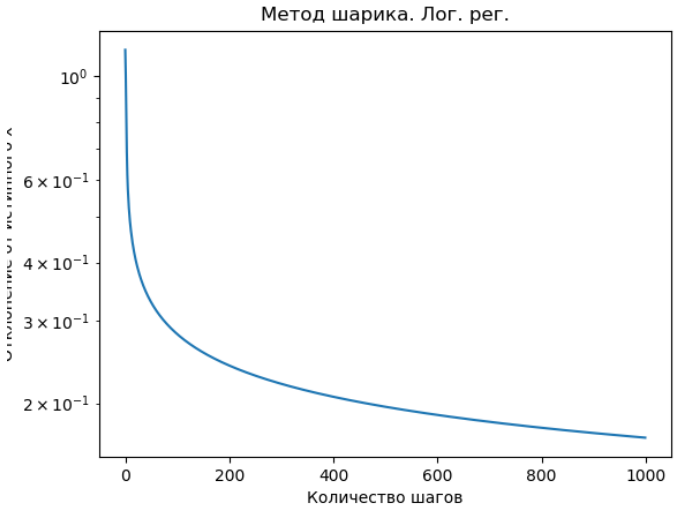


Рис. 15: Метод тяжелого шарика для логистической регрессии, логарифмическая шкала

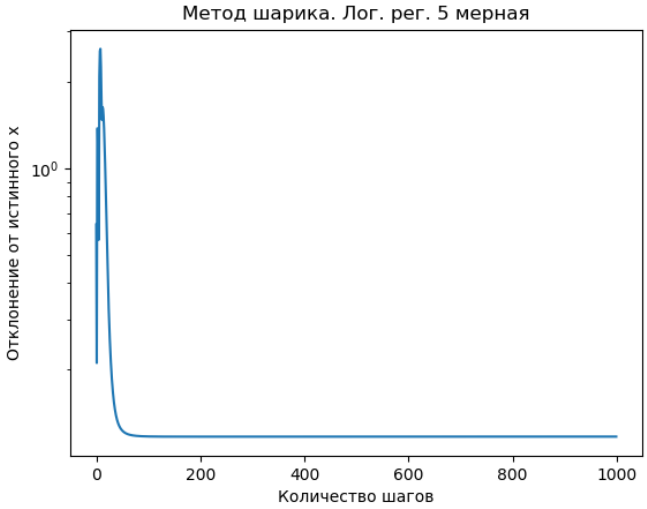


Рис. 16: Метод тяжелого шарика для логистической регрессии с x размерности 5, логарифмическая шкала

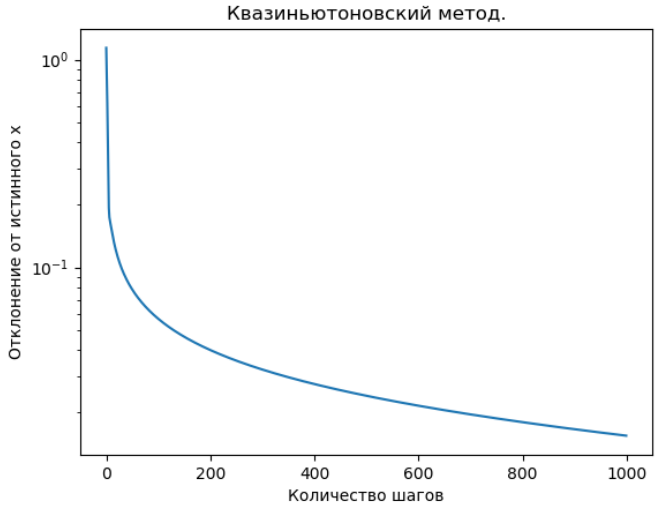


Рис. 17: Метод тяжелого шарика для функции $x + x^3$ с x размерности 2, логарифмическая шкала

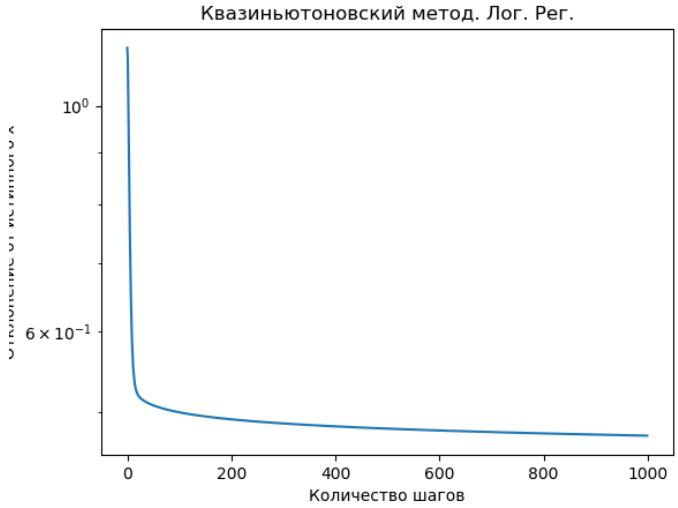
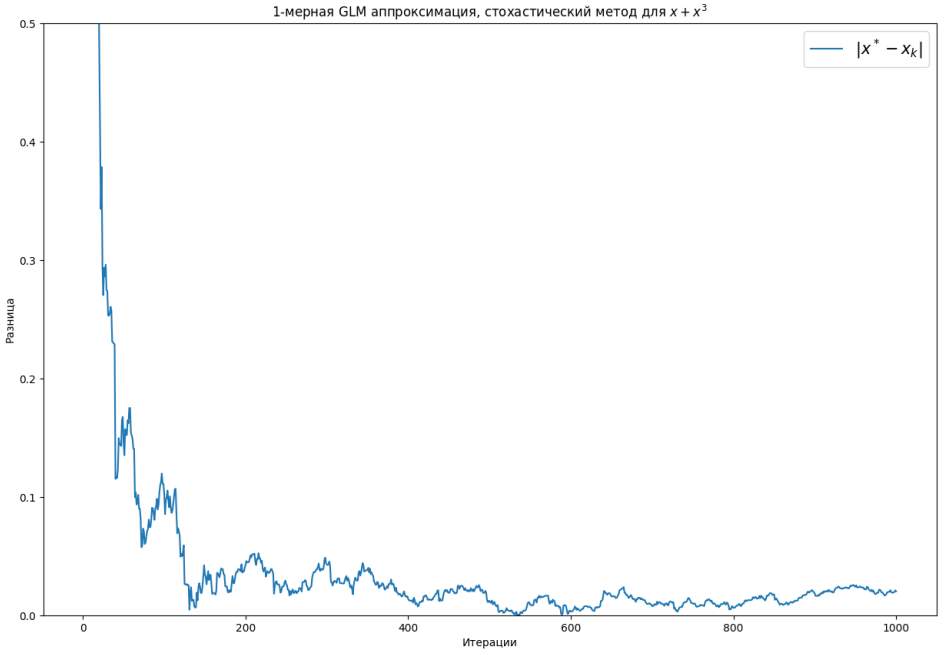
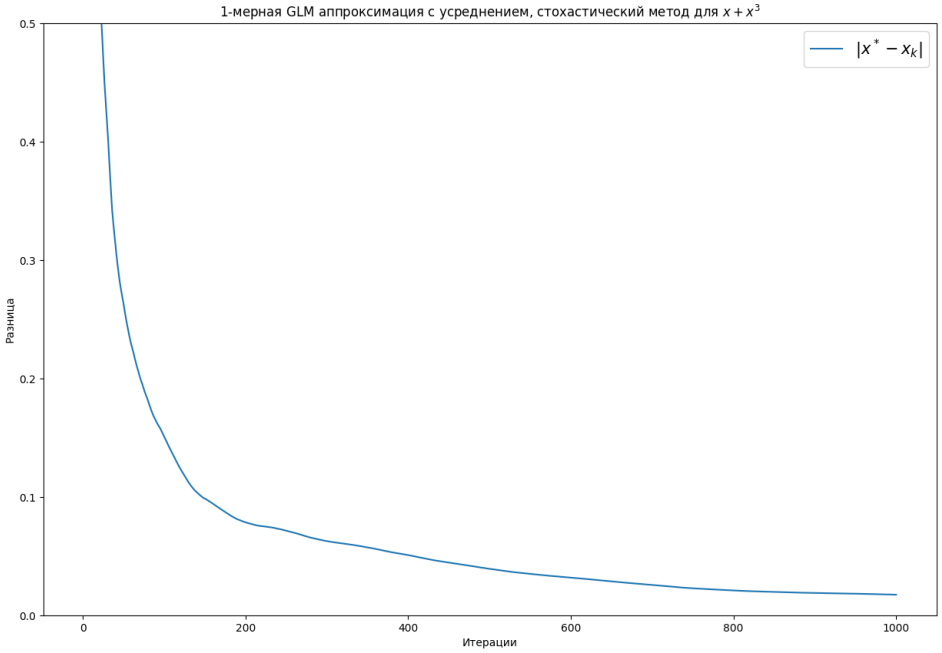


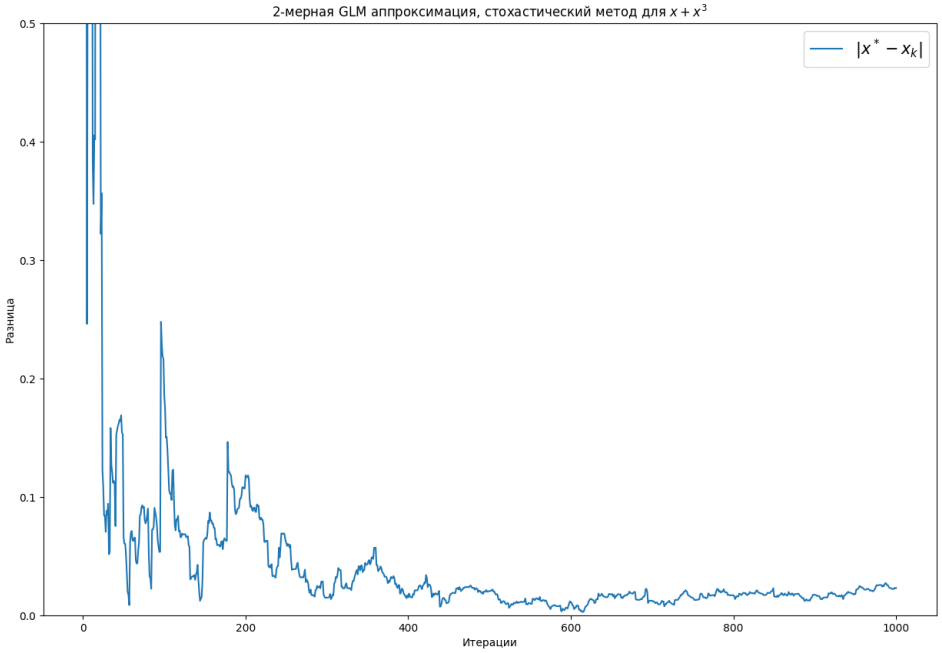
Рис. 18: Метод тяжелого шарика для логистической регрессии, логарифмическая шкала

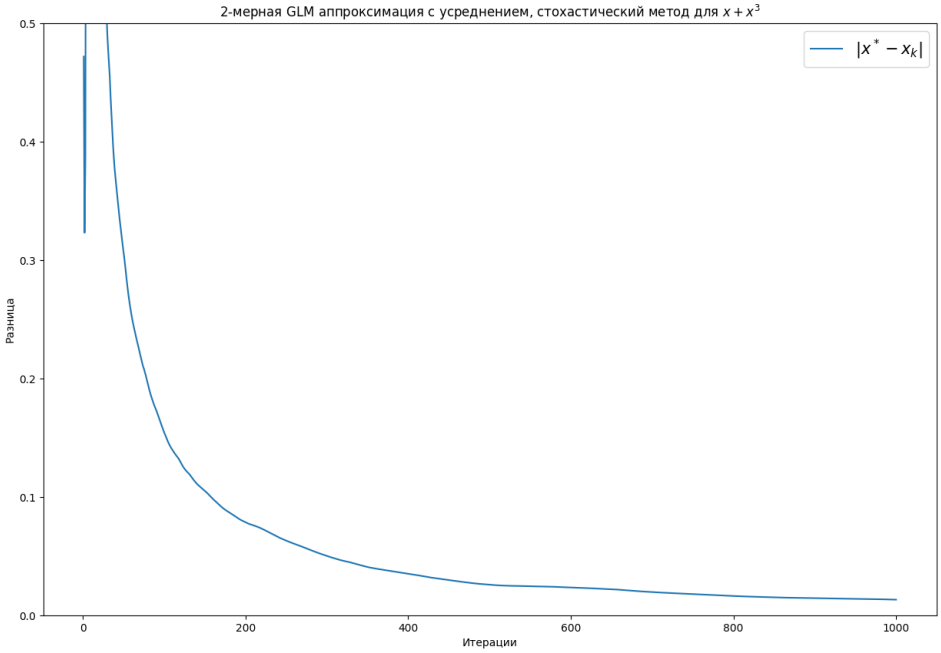
3.3 Стохастическая аппроксимация

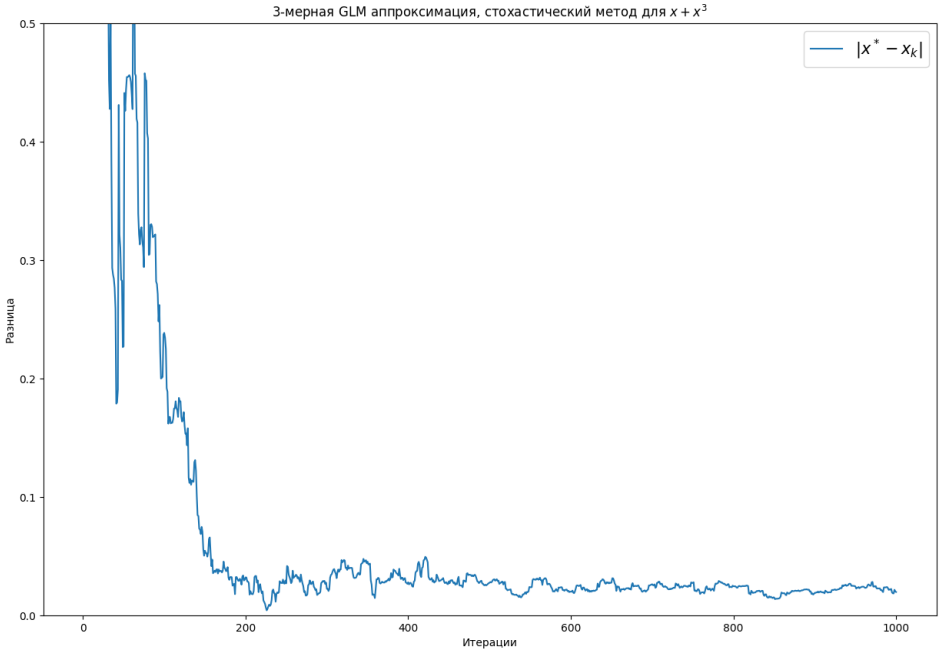
Ниже приведены графики сходимости к x^* при $n = 1, 2, 3$ при использовании SAA.

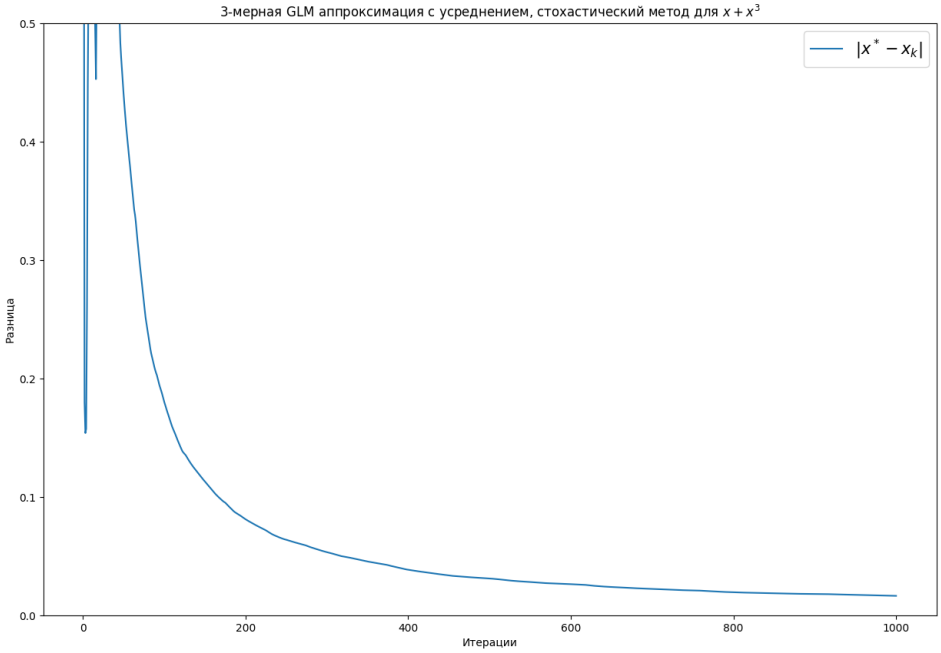


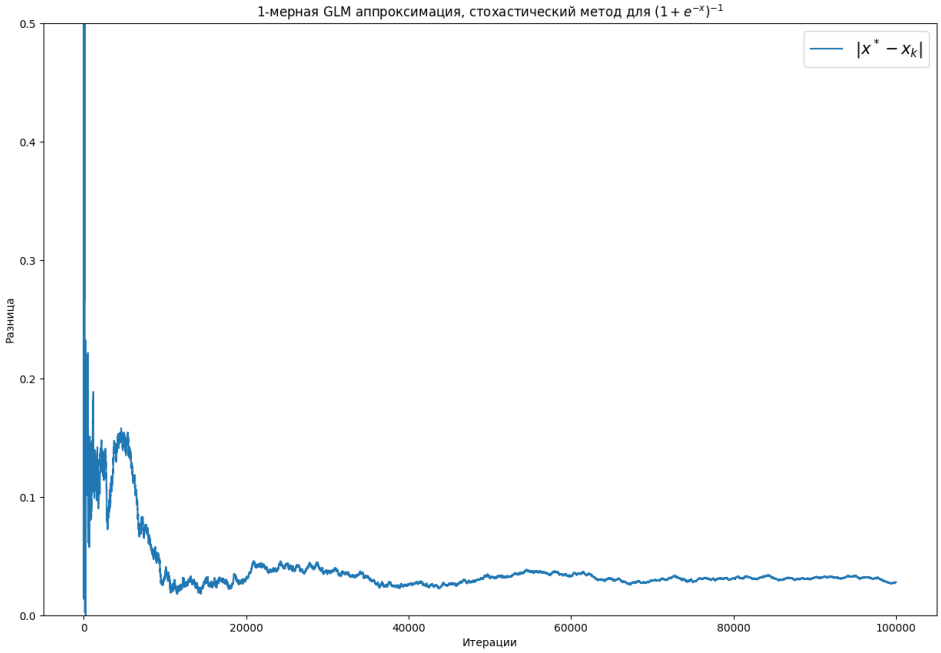

1-мерная GLM аппроксимация с усреднением, стохастический метод для $x + x^3$
$|x^* - x_k|$
Разница
Итерации
0.0
0.1
0.2
0.3
0.4
0.5
0
200
400
600
800
1000

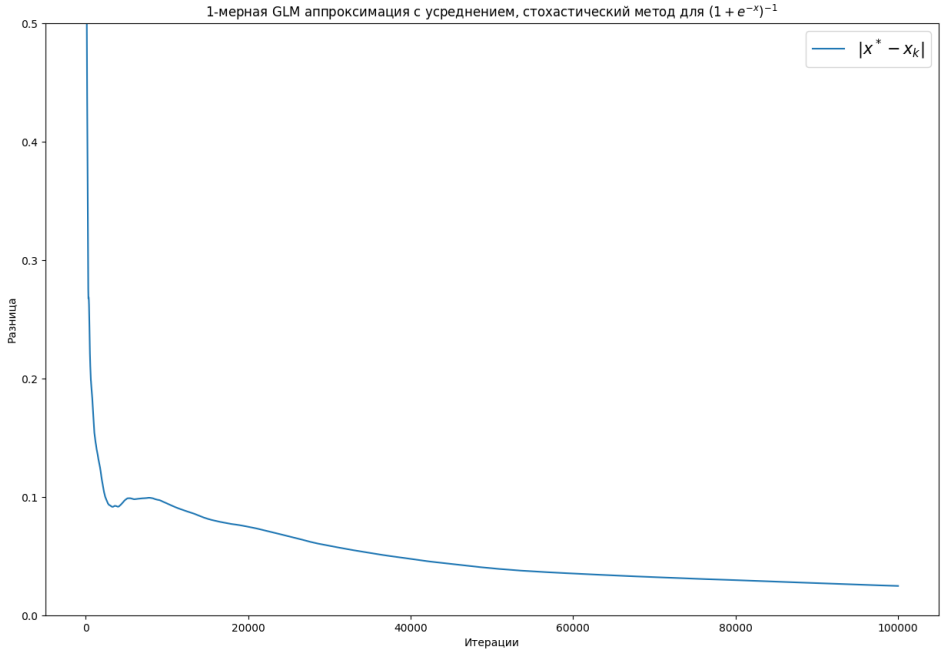

2-мерная GLM аппроксимация, стохастический метод для $x + x^3$
$|x^* - x_k|$
Разница
Итерации
0.0
0.1
0.2
0.3
0.4
0.5
0
200
400
600
800
1000

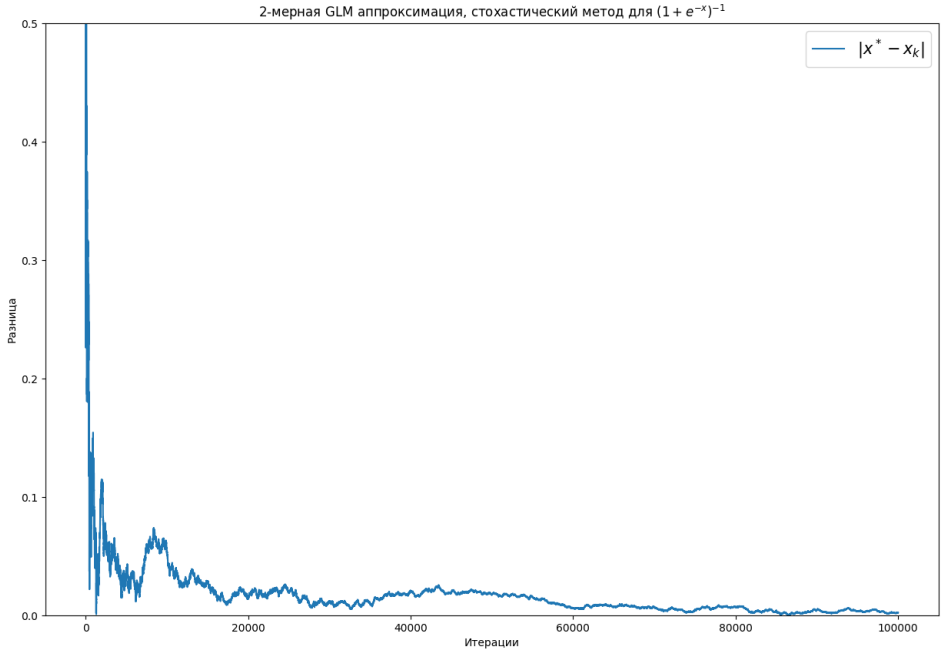

2-мерная GLM аппроксимация с усреднением, стохастический метод для $x + x^3$
$|x^* - x_k|$
Разница
Итерации
0.0
0.1
0.2
0.3
0.4
0.5
0
200
400
600
800
1000

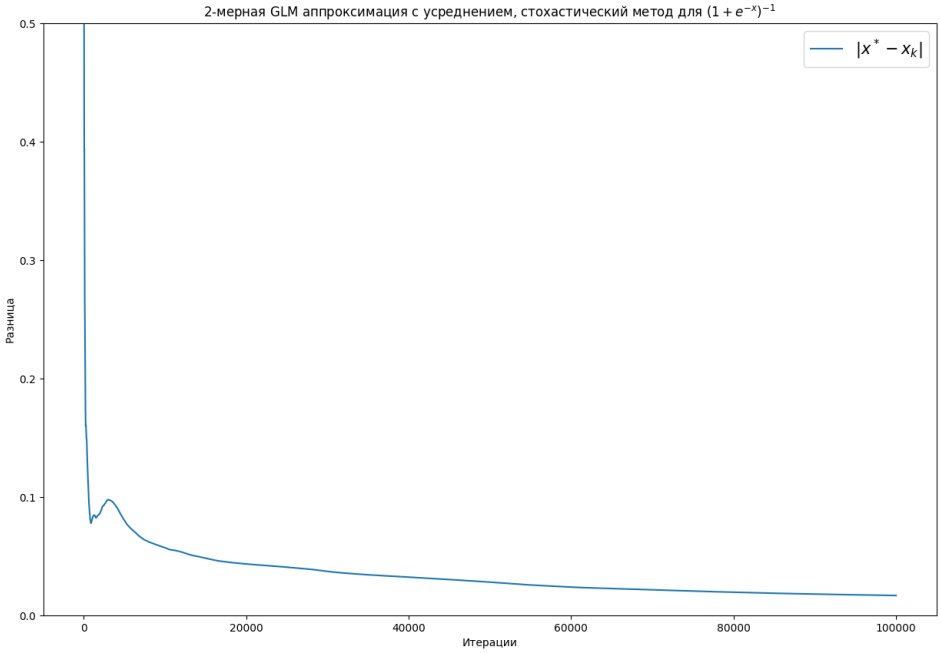

3-мерная GLM аппроксимация, стохастический метод для $x + x^3$
$|x^* - x_k|$
Разница
Итерации
0.0
0.1
0.2
0.3
0.4
0.5
0
200
400
600
800
1000

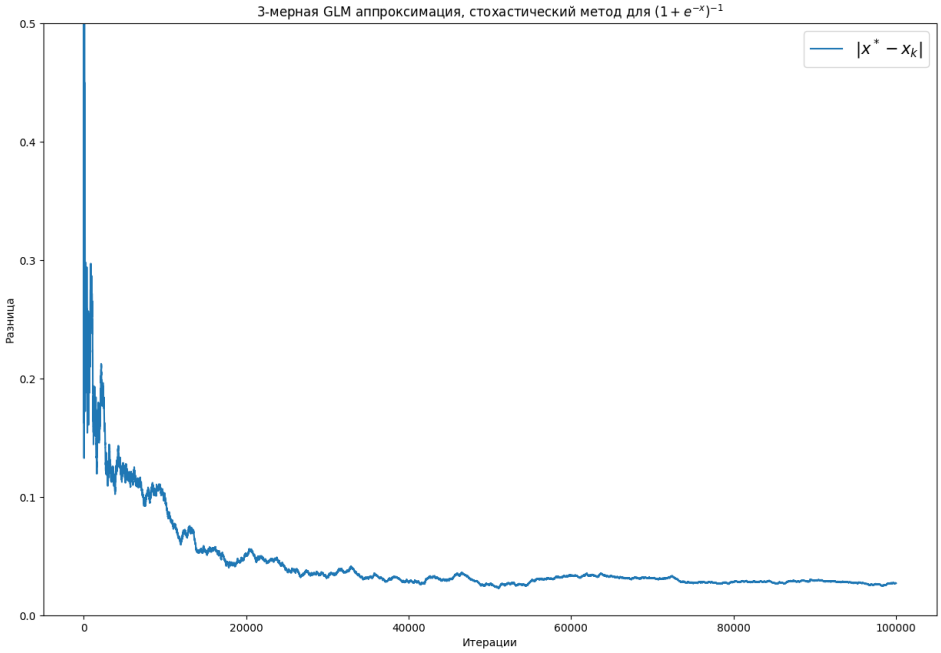

3-мерная GLM аппроксимация с усреднением, стохастический метод для $x + x^3$
$|x^* - x_k|$
Разница
Итерации
0.0
0.1
0.2
0.3
0.4
0.5
0
200
400
600
800
1000

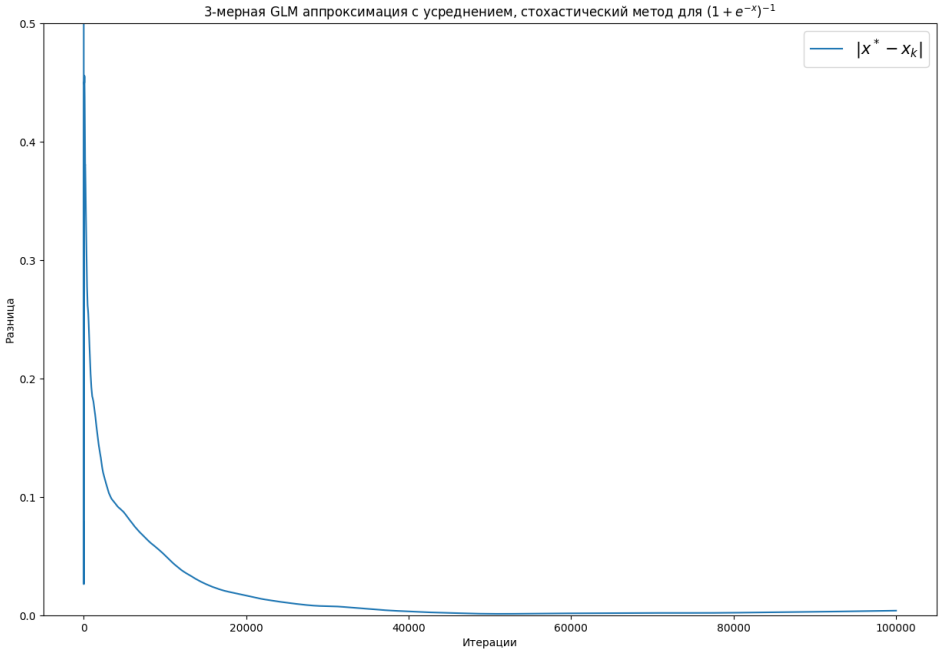

1-мерная GLM аппроксимация, стохастический метод для $(1 + e^{-x})^{-1}$
$|x^* - x_k|$
Разница
Итерации
0.0
0.1
0.2
0.3
0.4
0.5
0
20000
40000
60000
80000
100000


1-мерная GLM аппроксимация с усреднением, стохастический метод для $(1+e^{-x})^{-1}$
$|x^* - x_k|$
Разница
Итерации
0.0
0.1
0.2
0.3
0.4
0.5
0
20000
40000
60000
80000
100000


2-мерная GLM аппроксимация, стохастический метод для $(1+e^{-x})^{-1}$
$|x^* - x_k|$
Разница
Итерации
0.0
0.1
0.2
0.3
0.4
0.5
0
20000
40000
60000
80000
100000


2-мерная GLM аппроксимация с усреднением, стохастический метод для $(1+e^{-x})^{-1}$
$|x^* - x_k|$
Разница
Итерации
0.0
0.1
0.2
0.3
0.4
0.5
0
20000
40000
60000
80000
100000


3-мерная GLM аппроксимация, стохастический метод для $(1+e^{-x})^{-1}$
$|x^* - x_k|$
Разница
Итерации
0.0
0.1
0.2
0.3
0.4
0.5
0
20000
40000
60000
80000
100000


3-мерная GLM аппроксимация с усреднением, стохастический метод для $(1+e^{-x})^{-1}$
$|x^* - x_k|$
Разница
Итерации
0.0
0.1
0.2
0.3
0.4
0.5
0
20000
40000
60000
80000
100000

Результаты

Исходя из полученных результатов, можно считать, что предложенные методы действительно могут использоваться для восстановления сигналов, замещая работу по сведению данной задачи к выпуклой. Однако необходимое число наблюдений может быть довольно большим в случае не непрерывных распределений y, поэтому возможно иногда будет более просто находить функции потерь.

Вклады

Все участники внесли по 0.25 вклада.

Список литературы

[JN19] Anatoli Juditsky and Arkadi Nemirovski. Signal recovery by stochastic optimization, 2019.