Неявный стохастический градиент

Владислав Чабаненко

Факультет ВМК
МГУ им. М. В. Ломоносова

14 ноября 2014 г.
Спецсеминар "Байесовские методы машинного обучения"

Содержание

Интуиция неявных методов

В численных методах для решения дифференциальных уравнений используются *явные* и *неявные* методы.

- Следующее приближение для *явного* метода вычисляется явно через предыдущее:

$$y_{t+1} = f(y_t)$$

- Для нахождения следующего приближения в *неявном* методе требуется решить уравнение:

$$F(y_t, y_{t+1}) = 0$$

Motivating example: метод Эйлера

- Рассмотрим задачу Коши:

$$y'(t) = f(t, y), t \geq 0, y(0) = y_0.$$

- Итерация прямого (явного) метода Эйлера:

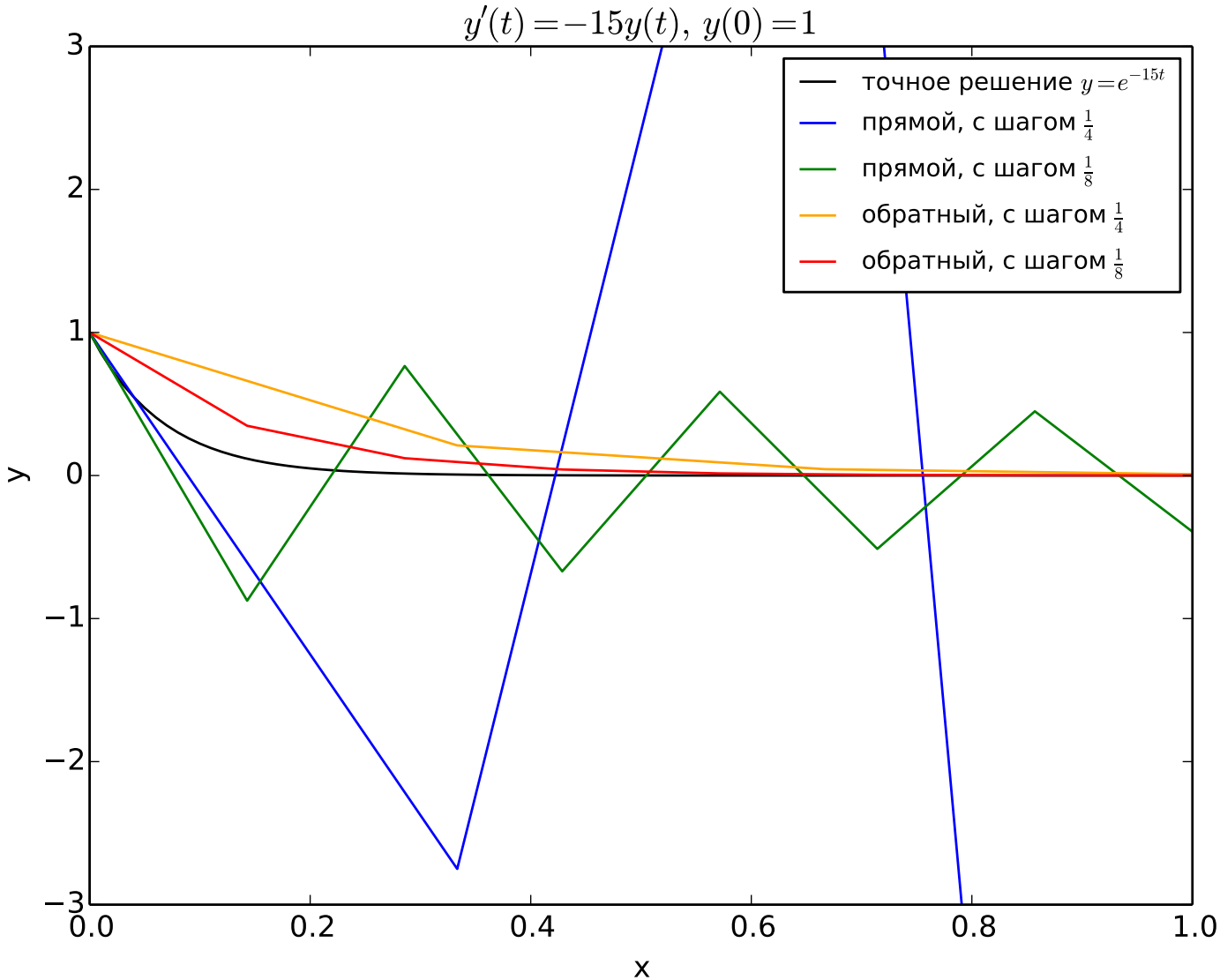
$$y_{k+1} = y_k + h \cdot f(t_k, y_k)$$

- Итерация обратного (неявного) метода Эйлера:

$$y_{k+1} = y_k + h \cdot f(t_{k+1}, \mathbf{y_{k+1}})$$

- h называется *шагом* метода

Motivating example: метод Эйлера



Градиентный спуск

- Пусть X — вектор реализаций случайной величины, $X_i \sim Pr(x|\theta)$.
- Наша задача — по данной выборке X оценить параметр распределения θ.
- Один из способов решения задачи — **градиентный метод**, где в качестве максимизируемого функционала выступает логарифм правдоподобия

$$\log \mathcal{L}(\theta|X) = \sum_i \log Pr(X_i|\theta) \to \max_\theta$$

- Итерационный процесс:

$$\theta_0 - \texttt{начальное приближение}$$
$$\theta_t = \theta_{t-1} + \gamma_t \nabla \mathcal{L}(\theta_{t-1}|X),$$

где γ_t — шаг градиента для каждой итерации.

Стохастический градиентный спуск

- **Метод стохастического градиента** вычисляет градиент только на одном сэмпле и сразу изменяет оценку θ.
- Итерационная схема:

$$\theta_0 - \texttt{начальное приближение}$$
$$\texttt{получаем сэмпл} \quad x_t$$
$$\theta_t = \theta_{t-1} + \gamma_t \nabla_{x_t} \mathcal{L}(\theta_{t-1}|x_t),$$

- Плюсы стохастического градиента:
  - Если выборка большая, то мы можем выбирать на каждом шаге случайный элемент из выборки и считать градиент только по нему
  - Можем онлайн решать задачу оценки параметра модели, если нам подается по одному сэмплу

Неявный метод стохастического градиента

- Явный метод стохастического градиента очень чувствителен к шагу градиента
- Итерационная схема для неявного метода:

$$\theta_0 - \texttt{начальное приближение}$$
$$\texttt{получаем сэмпл} \quad x_t$$
$$\theta_t = \theta_{t-1} + \gamma_t \nabla_{x_t} \mathcal{L}(\theta_t | x_t).$$

- Далее будем сравнивать явный и неявный методы стохастического градиента

Generalized linear model

- Авторы статьи исследовали методы стохастического градиентного спуска на примере GLM
- GLM (generalized linear model) — обобщение обычной линейной регрессии на случай, если шум имеет не нормальное распределение
- Обычная линейная регрессия

$$y \sim \mathcal{N}(\theta^T x, \sigma^2)$$

- Каноническое представление распределений GLM

$$p_{\text{glm}}(y|\theta, x) = c(y, \phi) \exp\left(\frac{\theta^T x y - b(\theta^T x)}{\phi}\right),$$

  ϕ — фиксированный дисперсионный параметр.
- Отметим, что распределения семейства GLM являются членами экспоненциального семейства

$$p_{\text{exp}}(y|\theta) = c(y) \exp\left(\theta^T u(y) - b(\theta)\right)$$

Примеры распределений семейства GLM 1

$$p_{\text{glm}}(y|\theta, x) = c(y, \phi) \exp\left(\frac{\theta^T x y - b(\theta^T x)}{\phi}\right)$$

Семейство включает такие распределения как

- Нормальное

$$y \sim \mathcal{N}(\mu, \sigma^2)$$

Сведем его к каноническому виду:

$$\begin{aligned} p(y|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{y^2}{2\sigma^2}\right) \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(\frac{\mu y}{\sigma^2}\right) \\ &= \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{y^2}{2\sigma^2}\right)\right] \exp\left(\frac{\mu y - \mu^2/2}{\sigma^2}\right) \end{aligned}$$

Таким образом,
$\mathbb{E}(y|\mu, \sigma^2) = \mu = \theta^T x, \phi = \sigma^2, b(\theta^T x) = (\theta^T x)^2/2.$

Примеры распределений семейства GLM 2

$$p_{\text{glm}}(y|\theta, x) = c(y, \phi) \exp\left(\frac{\theta^T x y - b(\theta^T x)}{\phi}\right)$$

- Распределение Пуассона

$$y \sim Poi(\lambda)$$

Сведем его к каноническому виду:

$$p(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} = \frac{e^{-\lambda}e^{y\log\lambda}}{y!}$$
$$= \left[\frac{1}{y!}\right] \exp\left(\frac{\log\lambda^y - e^{\log\lambda}}{1}\right)$$

Таким образом, $\mathbb{E}(y|\lambda) = \lambda = e^{\theta^T x}, \phi = 1, b(\theta^T x) = e^{\theta^T x}$.

Примеры распределений семейства GLM 3

$$p_{\text{glm}}(y|\theta, x) = c(y, \phi) \exp\left(\frac{\theta^T x y - b(\theta^T x)}{\phi}\right)$$

- Гамма-распределение

$$y \sim Ga(\alpha, \beta),$$

$$\mathbb{E}(y|\alpha, \beta) = \frac{\alpha}{\beta} = \frac{1}{\theta^T x}, \phi = \frac{1}{\alpha}, b(\theta^T x) = -\frac{1}{(\theta^T x)^2}$$

Постановка задачи для GLM

- Нам онлайн подаются признаки $x \in \mathbb{R}^p$ и отклик $y \in \mathbb{R}$ модели. Мы хотим оценить параметр $\theta^* \in \mathbb{R}^p$ для модели.
- Будем максимизировать логарифм правдоподобия для GLM методом стохастического градиентного спуска

$$\log \mathcal{L}(\theta | Y, X) = \sum_i \left(\frac{1}{\phi} \left(\theta^T x_i y_i - b(\theta^T x_i) \right) \right) + \text{const}$$

Итерационная схема для GLM

$$\log \mathcal{L}(\theta|Y,X) = \sum_i \left(\frac{1}{\phi} \left(\theta^T x_i y_i - b(\theta^T x_i) \right) \right) + \text{const}$$

- Можно показать, что для GLM итерационная схема будет записываться в следующем виде
- Для явного метода:

$$\theta_n = \theta_{n-1} + \gamma_n (y_n - h(\theta_{n-1}^T x_n)) x_n,$$

- для неявного метода:

$$\theta_n = \theta_{n-1} + \gamma_n (y_n - h(\theta_n^T x_n)) x_n,$$

- где $h(\theta^T x) = \mathbb{E}(y|\theta, x)$, а $1/\phi$ мы внесли в шаг γ_n.
- Для канонической формы распределений семейства GLM функция $h(\cdot)$ гладкая и **монотонная**

Случай неявного пересчета 1

Требуется решить уравнение относительно θ_n

$$\boldsymbol{\theta_n} = \theta_{n-1} + \gamma_n(y_n - \mathbf{h}(\boldsymbol{\theta_n^T}\mathbf{x_n}))x_n,$$

- Отметим, что сложно найти θ_n в общем случае из-за того, что функция h произвольная
- К тому же θ_n — вектор, а $h(\theta_n^T x_n)$ — число, и нам нужно решать уравнение одновременно для всех измерений
- На самом деле, это облегчает задачу

Случай неявного пересчета 2

Требуется решить уравнение относительно θ_n

$$\boldsymbol{\theta_n} = \theta_{n-1} + \gamma_n(y_n - \mathbf{h}(\boldsymbol{\theta}_\mathbf{n}^\mathbf{T}\mathbf{x_n}))x_n,$$

но для вычисления $\boldsymbol{\theta_n}$ достаточно знать только $\mathbf{h}(\boldsymbol{\theta}_\mathbf{n}^\mathbf{T}\mathbf{x_n})$.

- Применим операцию транспонирования к обеим частям нашего уравнения и домножим скалярно справа на x_n:

$$\boldsymbol{\theta}_\mathbf{n}^\mathbf{T}\mathbf{x_n} = \theta_{n-1}^T x_n + \gamma_n(y_n - \mathbf{h}(\boldsymbol{\theta}_\mathbf{n}^\mathbf{T}\mathbf{x_n}))||x_n||^2$$

- Теперь к обеим частям уравнения применим функцию $h(\cdot)$ как к числу:

$$\mathbf{h}(\boldsymbol{\theta}_\mathbf{n}^\mathbf{T}\mathbf{x_n}) = h(\theta_{n-1}^T x_n + \gamma_n(y_n - \mathbf{h}(\boldsymbol{\theta}_\mathbf{n}^\mathbf{T}\mathbf{x_n}))||x_n||^2)$$

Случай неявного пересчета 3

$$\mathbf{h(\theta_n^T x_n)} = h(\theta_{n-1}^T x_n + \gamma_n(y_n - \mathbf{h(\theta_n^T x_n)})||x_n||^2)$$

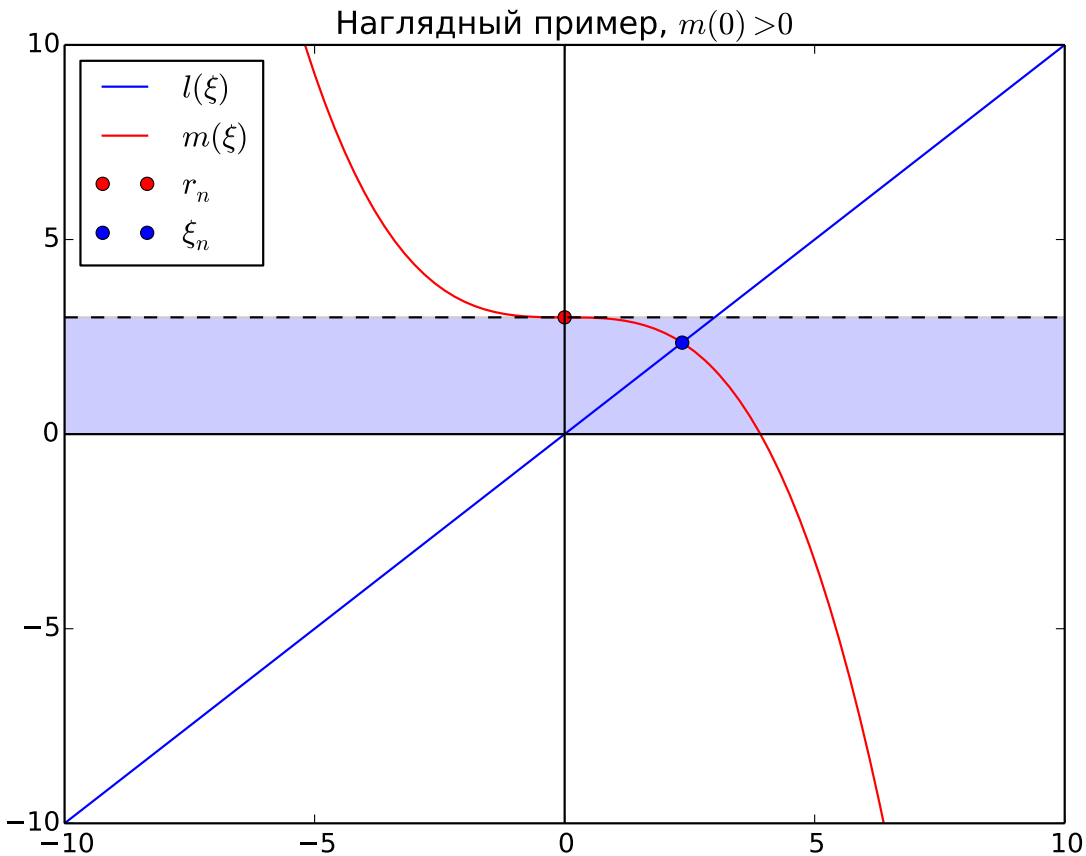
- Мы получили **одномерное** уравнение относительно $\xi_n = \mathbf{h(\theta_n^T x_n)}$:

$$\xi_n = h(\theta_{n-1}^T x_n + \gamma_n(y_n - \xi_n)||x_n||^2)$$

- Если $h(\cdot)$ — монотонно возрастающая (как в случае нормального и пуассоновского распределений), то выражение справа будет монотонно убывать. Тогда полученное уравнение мы можем решать с помощью методов оптимизаций для одномерных уравнений (метод секущей, метод дихотомии) на некотором отрезке.
- Если $h(\cdot)$ — монотонно убывающая (например, для гамма распределения), то в общем случае не гарантируется даже существования решения. Оставим этот случай для дальнейших исследований.
- Далее будем работать с монотонно возрастающими $h(\cdot)$

Границы поиска 1

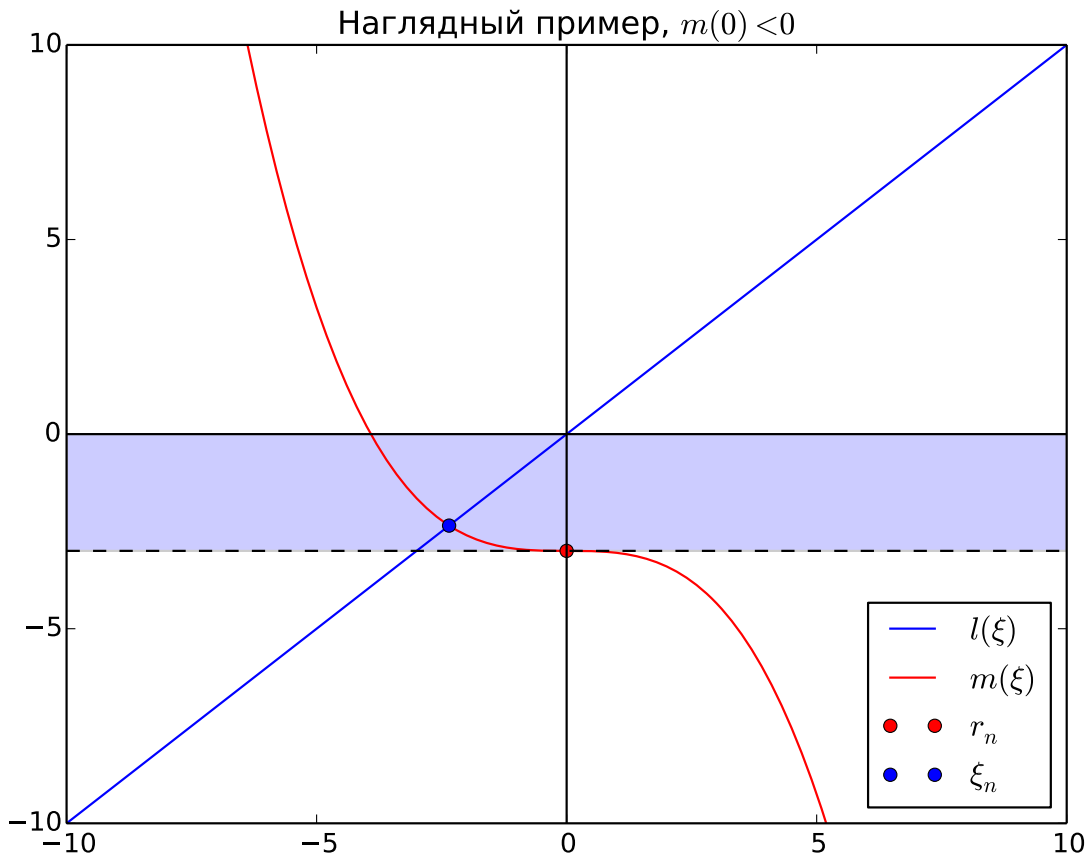
$$\xi_n = h(\theta_{n-1}^T x_n + \gamma_n(y_n - \xi_n)||x_n||^2)$$



$l(\xi) = \xi,\ m(\xi) = h(\theta_{n-1}^T x_n + \gamma_n(y_n - \xi)||x_n||^2)$
$m(0) = h(\theta_{n-1}^T x_n + \gamma_n y_n||x_n||^2) = r_n$

Границы поиска 2

$$\xi_n = h(\theta_{n-1}^T x_n + \gamma_n(y_n - \xi_n)||x_n||^2)$$



$l(\xi) = \xi,\ m(\xi) = h(\theta_{n-1}^T x_n + \gamma_n(y_n - \xi)||x_n||^2)$
$m(0) = h(\theta_{n-1}^T x_n + \gamma_n y_n||x_n||^2) = r_n$

Алгоритм для неявного метода

Получаем алгоритм для нахождения θ_n для неявной итерации:

for $n = 1, 2, \ldots$ **do**

$r_n \leftarrow h(\theta_{n-1}^T x_n + \gamma_n y_n ||x_n||^2)$

Решаем численно уравнение на отрезке $[0, r_n]$ относительно ξ_n:

$$\xi_n = h(\theta_{n-1}^T x_n + \gamma_n (y_n - \xi_n) ||x_n||^2)$$

$\boldsymbol{\theta}_n \leftarrow \boldsymbol{\theta}_{n-1} + \gamma_n (y_n - \xi_n) x_n$

end

Теорема 1

Пусть $\gamma_n > 0$ — убывающая последовательность, такая что $\sum_{i=1}^{\infty} \gamma_i = \infty$ и $\sum_{i=1}^{\infty} \gamma_i^2 < \infty$, тогда

- смещенность явного метода стохастического градиента удовлетворяет:

$$\mathbb{E}\left(\theta_n^{\mathrm{sgd}} - \theta^*\right) = (I - \gamma_n \phi \mathcal{I}(\theta^*))\mathbb{E}\left(\theta_{n-1}^{\mathrm{sgd}} - \theta^*\right) + \mathbb{E}\left(\mathcal{O}||\theta_{n-1}^{\mathrm{sgd}} - \theta^*||^2\right) x_n,$$

- смещенность неявного метода стохастического градиента удовлетворяет:

$$\mathbb{E}\left(\theta_n^{\mathrm{im}} - \theta^*\right) = (I + \gamma_n \phi \mathcal{I}(\theta^*))^{-1}\left(\mathbb{E}\left(\theta_{n-1}^{\mathrm{im}} - \theta^*\right) + \mathbb{E}\left(\mathcal{O}||\theta_{n-1}^{\mathrm{im}} - \theta^*||^2\right) x_n\right),$$

- где $\mathcal{I}(\theta^*) = \mathrm{Cov}(\nabla_\theta \log \mathcal{L}(\theta^*|y, x))$ — матрица информации Фишера, **симметричная положительно определенная** матрица

Доказательство

Докажем для явного пересчета.

- Запишем итерационную формулу:

$$\theta_n = \theta_{n-1} + \gamma_n(y_n - h(\theta_{n-1}^T x_n))x_n$$

- Вычтем θ^* из обеих частей и возьмем матожидание:

$$\mathbb{E}(\theta_n - \theta^*) = \mathbb{E}(\theta_{n-1} - \theta^*) + \mathbb{E}\gamma_n y_n x_n - \mathbb{E}\gamma_n h(\theta_{n-1}^T x_n)x_n$$

$$\mathbb{E}(\theta_n - \theta^*) = \mathbb{E}(\theta_{n-1} - \theta^*) + \mathbb{E}\gamma_n h(\theta^{*T} x_n)x_n - \mathbb{E}\gamma_n h(\theta_{n-1}^T x_n)x_n$$

- Разложим $h(\theta_{n-1}^T x_n)$ по Тейлору в точке θ^* следующим образом:

$$h(\theta_{n-1}^T x_n) = h(\theta^{*T} x_n) + h'(\theta^{*T} x_n)(\theta_{n-1} - \theta^*)^T x_n + \mathcal{O}\left(||\theta_{n-1} - \theta^*||^2\right)$$

- Тогда равенство выше преобразуется в

$$\mathbb{E}(\theta_n - \theta^*) = \mathbb{E}(\theta_{n-1} - \theta^*) + \mathbb{E}\gamma_n h(\theta^{*T} x_n)x_n - \mathbb{E}\gamma_n h(\theta^{*T} x_n)x_n - \mathbb{E}\gamma_n h'(\theta^{*T} x_n)x_n(\theta_{n-1} - \theta^*)^T x_n - \mathbb{E}\left(\mathcal{O}||\theta_{n-1} - \theta^*||^2\right)x_n$$

Доказательство

- Упрощая выражение, получим:

$$\mathbb{E}\left(\theta_n - \theta^*\right) = \left(I - \gamma_n h'(\theta^{*T} x_n) x_n x_n^T\right)\mathbb{E}\left(\theta_{n-1} - \theta^*\right) + \mathbb{E}\left(\mathcal{O}||\theta_{n-1} - \theta^*||^2\right) x_n$$

- Можно показать, что

$$\mathcal{I}(\theta^*) = \frac{1}{\phi} h'(\theta^{*T} x_n) x_n x_n^T,$$

- тогда получим окончательное выражение:

$$\mathbb{E}\left(\theta_n - \theta^*\right) = \left(I - \gamma_n \phi \mathcal{I}(\theta^*)\right)\mathbb{E}\left(\theta_{n-1} - \theta^*\right) + \mathbb{E}\left(\mathcal{O}||\theta_{n-1} - \theta^*||^2\right) x_n$$

- Для неявного метода можно проделать аналогичные преобразования над итерационной формулой и получить требуемое равенство.

Стабильность

- Теперь исследуем методы на *стабильность*, то есть на ожидаемую смещенность оценок при малом количестве итераций.
- Упростим формулу для асимптотической смещенности из Теоремы 1, отбросив остаточный член.
- Для явного метода

$$\mathbb{E}\left(\theta_n^{\mathrm{sgd}} - \theta^*\right) = (I - \gamma_n \phi \mathcal{I}(\theta^*))\mathbb{E}\left(\theta_{n-1}^{\mathrm{sgd}} - \theta^*\right) = P_1^n b_0,$$

$P_1^n = \prod_{i=1}^n (I - \gamma_i \phi \mathcal{I}(\theta^*))$

- Для неявного метода

$$\mathbb{E}\left(\theta_n^{\mathrm{im}} - \theta^*\right) = (I + \gamma_n \phi \mathcal{I}(\theta^*))^{-1}\left(\mathbb{E}\left(\theta_{n-1}^{\mathrm{im}} - \theta^*\right)\right) = Q_1^n b_0,$$

$Q_1^n = \prod_{i=1}^n (I + \gamma_i \phi \mathcal{I}(\theta^*))^{-1}$

- b_0 — это смещенность обоих методов для начальной оценки θ_0

Рассуждения о стабильности

$$P_1^n = \prod_{i=1}^{n} (I - \gamma_i \phi \mathcal{I})$$

$$Q_1^n = \prod_{i=1}^{n} (I + \gamma_i \phi \mathcal{I})^{-1}$$

- Оценим векторы $P_1^n b_0$ и $Q_1^n b_0$ по евклидовой норме
- P_1^n и Q_1^n — симметричные матрицы, тогда

$$||P_1^n b_0|| \leq |\lambda|_{\max}^{P_1^n} ||b_0||,$$

$$||Q_1^n b_0|| \leq |\lambda|_{\max}^{Q_1^n} ||b_0||$$

- Скорость сходимости методов зависит от $|\lambda|_{\max}^{P_1^n}$ и $|\lambda|_{\max}^{Q_1^n}$

Случай явного метода

$$P_1^n = \prod_{i=1}^{n} (I - \gamma_i \phi \mathcal{I})$$

- Можно показать, что $|\lambda|_{\max}^{P_1^n} = \prod_{i=1}^{n} |\lambda|_{\max}^{I - \gamma_i \phi \mathcal{I}}$
- Если $\gamma_1 \phi \lambda_{\max}^{\mathcal{I}} < 2$, тогда $|\lambda|_{\max}^{I - \gamma_1 \phi \mathcal{I}} < 1$ и $|\lambda|_{\max}^{P_1^n}$ убывает
- Но может быть $\gamma_1 \phi \lambda_{\max}^{\mathcal{I}} > 2$, тогда максимальным по модулю собственным значением матрицы $I - \gamma_i \phi \mathcal{I}$ на первых итерациях будет $|1 - \gamma_i \phi \lambda_{\max}^{\mathcal{I}}| > 1$, пока будет выполняться условие $\gamma_i \phi \lambda_{\max}^{\mathcal{I}} > 2$
- Тогда на этих первых итерациях отклонение оценки явного метода будет сильно возрастать

Проблемы со стабильностью явного метода

$$P_1^n = \prod_{i=1}^{n} (I - \gamma_i \phi \mathcal{I})$$

- В то же время на достаточно большой итерации будет выполняться $\gamma_i \phi \lambda_{\max}^{\mathcal{I}} < 1$ и $|\lambda|_{\max}^{P_1^n}$ будет соответствовать минимальным собственным значениям матриц $\gamma_i \phi \mathcal{I}$
- Значит скорость сходимости будет лучшей, когда $\gamma_i \phi \lambda_{\min}^{\mathcal{I}}$ близко к 1 $\Rightarrow$ $\lambda_{\max}^{P_1^n}$ будет близко к 0
- В итоге получаем:
  - Для стабильности нужно, чтобы $\gamma_1 \phi \lambda_{\max}^{\mathcal{I}} < 2$
  - Для быстрой сходимости — $\gamma_i \phi \lambda_{\min}^{\mathcal{I}} \approx 1$ на достаточно больших итерациях
- Невозможно удовлетворить сразу эти оба требования
- Напротив, для неявного метода такой проблемы не возникает
- Подтвердим наши наблюдения следующей леммой

Лемма

- Пусть $\lambda_{\max} = \max \operatorname{eig}(\phi \mathcal{I}(\theta^*))$
- Пусть $\gamma_n = \alpha / n$ и $\alpha \lambda_{\max} > 2$, тогда максимально возможное собственное значение
- матрицы P_1^n удовлетворяет:

$$\max_{n>0} \max \operatorname{eig}(P_1^n) = \Theta\left(\frac{2^{\alpha\lambda_{\max}}}{\sqrt{\alpha\lambda_{\max}}}\right),$$

- матрицы Q_1^n неявного метода удовлетворяет:

$$\max_{n>0} \max \operatorname{eig}(Q_1^n) < 1.$$

Докажем вспомогательное утверждение

$$M = \max_{n>0} \left| \prod_{i=1}^{n} \left(1 - \frac{b}{i}\right) \right| \approx \begin{cases} |1-b|, & 0 < b < 2 \\ \frac{2^b}{\sqrt{2\pi b}}, & b > 2 \end{cases}$$

- Для $0 < b < 2$ получаем $|(1-b)(1-\frac{b}{2})\ldots| \leq |1-b|$.
- Теперь пусть $b > 2$. Без ограничения общности будем считать, что b — натуральное четное, тогда:

$$M = (b-1)\left(\frac{b}{2} - 1\right)\ldots\left(\frac{b}{b/2} - 1\right) = \frac{1}{2}\frac{b!}{(b/2)!(b/2)!}$$

Вспомним формулу Стирлинга:

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$$

и применим ее к нашему выражению:

$$M = \frac{1}{2}\frac{b!}{(b/2)!(b/2)!} \approx \frac{1}{2}\frac{\left(\frac{b}{e}\right)^b \sqrt{2\pi b}}{\left(\frac{b}{2e}\right)^b \pi b} = \frac{2^b}{\sqrt{2\pi b}}$$

Доказательство леммы

Вспомним выражения для матриц явного и неявного методов для смещенности.

- Для явного метода $P_1^n = \prod_{i=1}^{n}(I - \gamma_i \phi \mathcal{I}(\theta^*))$, тогда

$$\max_{n>0} \max \operatorname{eig}(P_1^n) = \max_{n>0} \left| \prod_{i=1}^{n} \left(1 - \frac{\alpha \lambda_{\max}}{i}\right) \right| \approx \frac{2^{\alpha\lambda_{\max}}}{\sqrt{2\pi\alpha\lambda_{\max}}},$$

  так как из условия теоремы $\alpha\lambda_{\max} > 2$.

- Для неявного метода $Q_1^n = \prod_{i=1}^{n}(I + \gamma_i \phi \mathcal{I}(\theta^*))^{-1}$, но собственные значения матриц $(I + \gamma_i \phi \mathcal{I}(\theta^*))^{-1}$ меньше единицы, следовательно

$$\max_{n>0} \max \operatorname{eig}(Q_1^n) = \max_{n>0} \left| \prod_{i=1}^{n} \left(1 + \frac{\alpha \lambda_{\max}}{i}\right)^{-1} \right| < 1.$$

Итог

$$||P_1^n b_0|| \leq |\lambda|_{\max}^{P_1^n} ||b_0||, \quad ||Q_1^n b_0|| \leq |\lambda|_{\max}^{Q_1^n} ||b_0||$$

Из теоремы следует, что

- в явной процедуре стохастического градиента эффект начальной смещенности экспоненциально возрастает, прежде чем начать убывать, в случае если шаг градиента $\alpha > 2/\lambda_{\max}$, $\lambda_{\max} = \max \operatorname{eig}(\phi \mathcal{I}(\theta^*))$,
- неявная процедура стохастического градиента не зависит от шага градиента, и завивисомть от начальной смещенности убывает при каждой следующей итерации алгоритма.

В экспериментах мы хотим проверить, что

- Если α большое ($\alpha\lambda_{\max} > 2$), то явный метод дает плохую оценку, а неявный хорошую
- Если α достаточно малое ($\alpha\lambda_{\min} \ll 1$), то оба метода медленно сходятся

Эксперимент №1

Постановка эксперимента: на примере двумерного распределения Пуассона мы

- оценим параметр модели явным и неявным методами
- убедимся в стабильности неявного метода и нестабильности явного при большом шаге градиента
- исследуем, каким должен быть шаг градиента для более надежной работы явного метода

Эксперимент №1

- Имеется модель:
$$y_n \sim Poi(e^{\theta^{*T} x_n}),$$
- Оптимизируемый функционал выглядит:
$$\log \mathcal{L}(\theta|X, Y) = \sum_i \left(\theta^T x_i y_i - e^{\theta^T x_i}\right) + \text{const}$$
- Обучать будем онлайн, получая на каждой итерации x_i, y_i. Генерировать данные будем следующим образом:
  - x_n принимает значения $(0,0)^T, (1,0)^T, (0,1)^T$ с вероятностями $0.6, 0.2$ и 0.2 соответственно,
  - y_n — из модели.
- Положим $\theta^* = (\log 2, \log 4)^T$
- Шаг градиента положим соответствующим случаю $\alpha\lambda_{\max} > 2$ $\gamma_n = 10/3n$, $\alpha = 10/3$
- Будем выполнять $N = 20000$ итераций для $m = 100$ различных начальных оценок
- Начальные оценки будем генерировать из двумерного стандартного нормального распределения $\mathcal{N}_2(0, I)$

Эксперимент №1: Оценка параметра

Запустим оба метода. Всего у нас имеется $m = 100$ итоговых оценок параметра θ^*. Выберем несколько случайных оценок для каждого метода

- Для *явного*: $(0.709, \text{-}\mathbf{24.382})^T$, $(0.71, 1.4)^T$, $(\text{-}\mathbf{9.14}, 1.39)^T$, $(0.694, \text{-}\mathbf{284563.14})^T$, $(0.712, 1.386)^T$
- Для *неявного*: $(0.695, 1.384)^T$, $(0.688, 1.403)^T$, $(0.697, 1.368)^T$, $(0.679, 1.41)^T$, $(0.693, 1.392)^T$
- *Реальное* значение параметра: $(0.693, 1.386)^T$

Заметим сильное отклонение некоторых оценок явного метода.

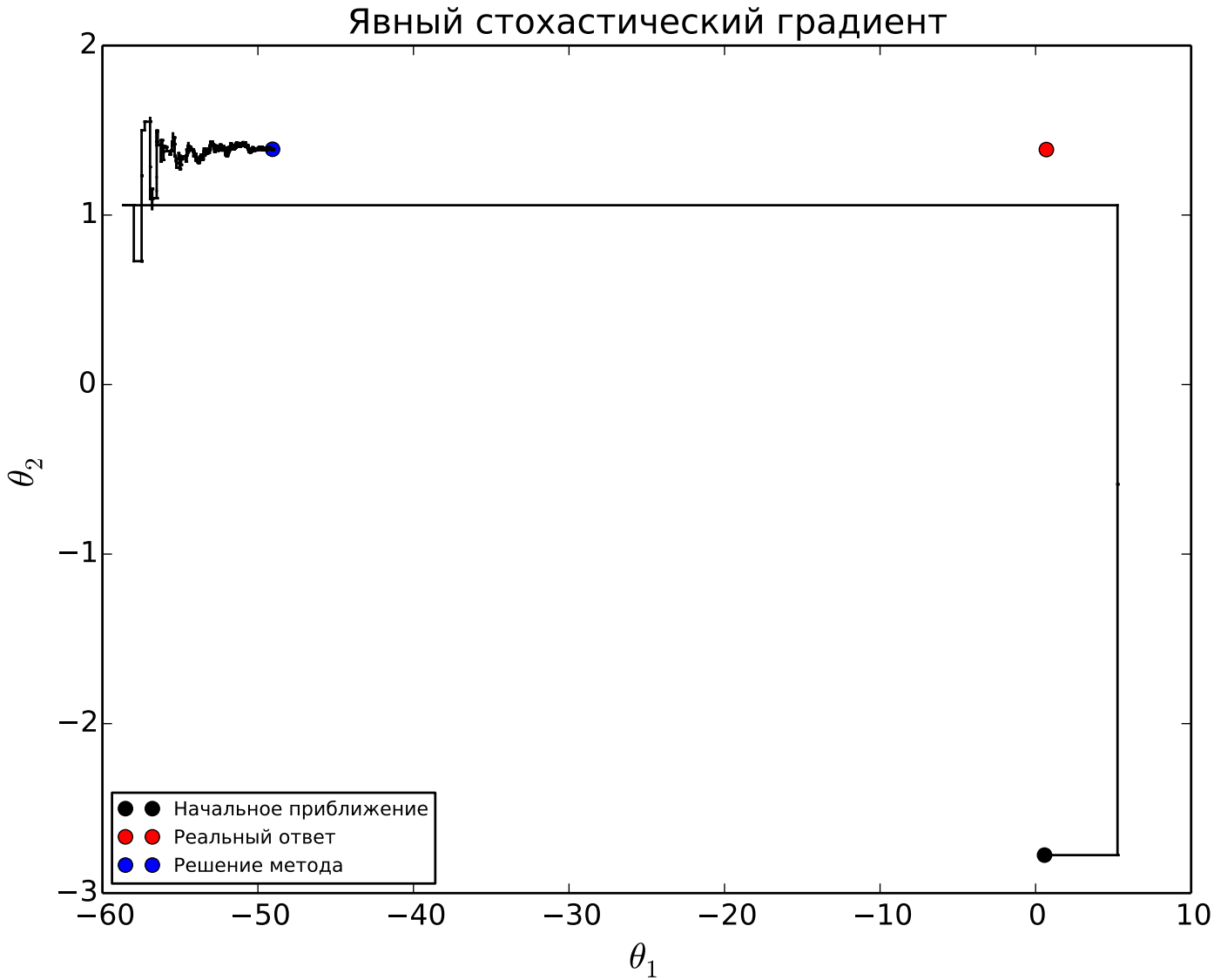


Рис. : Пример плохой сходимости явного метода

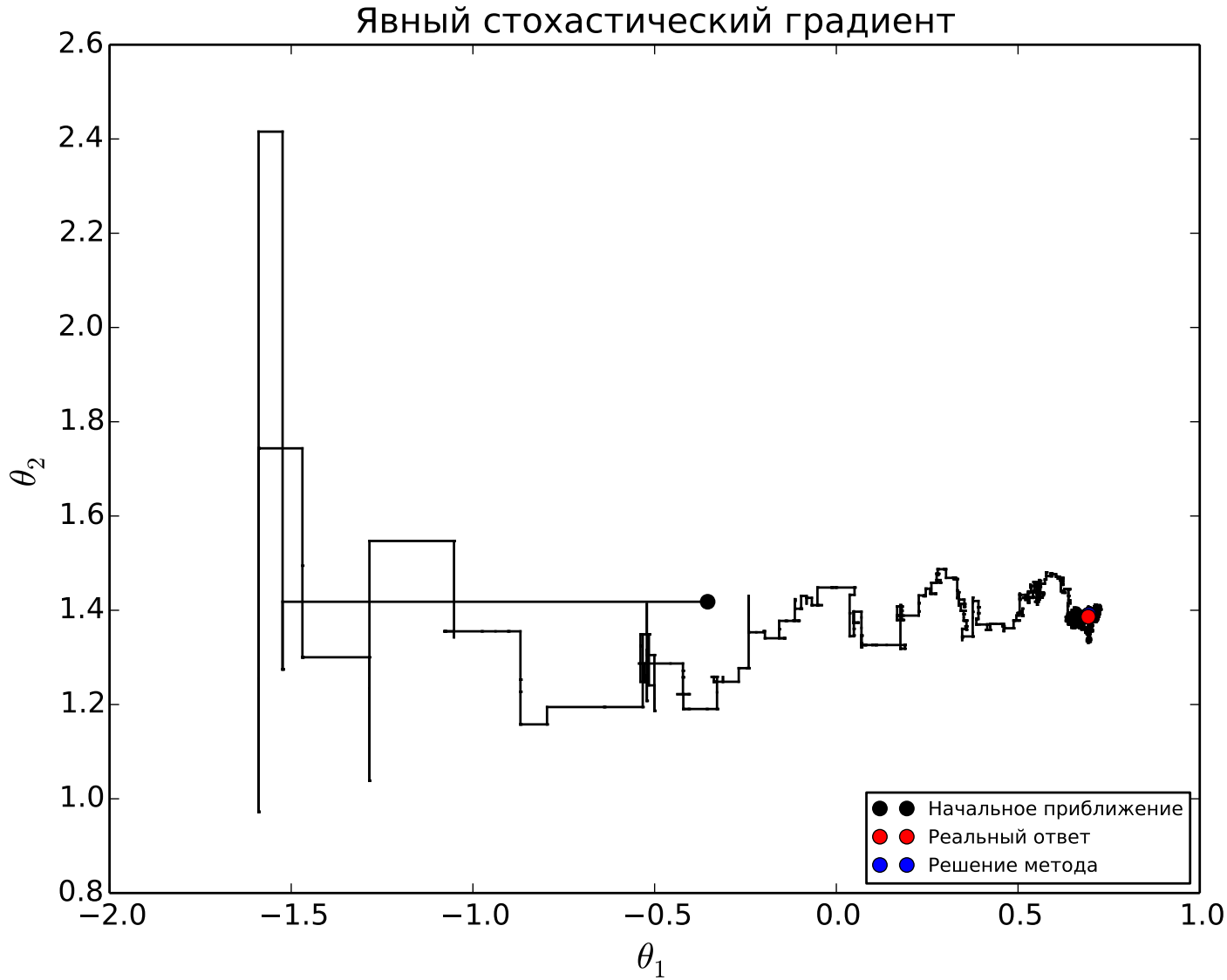


Рис. : Пример хорошей сходимости явного метода

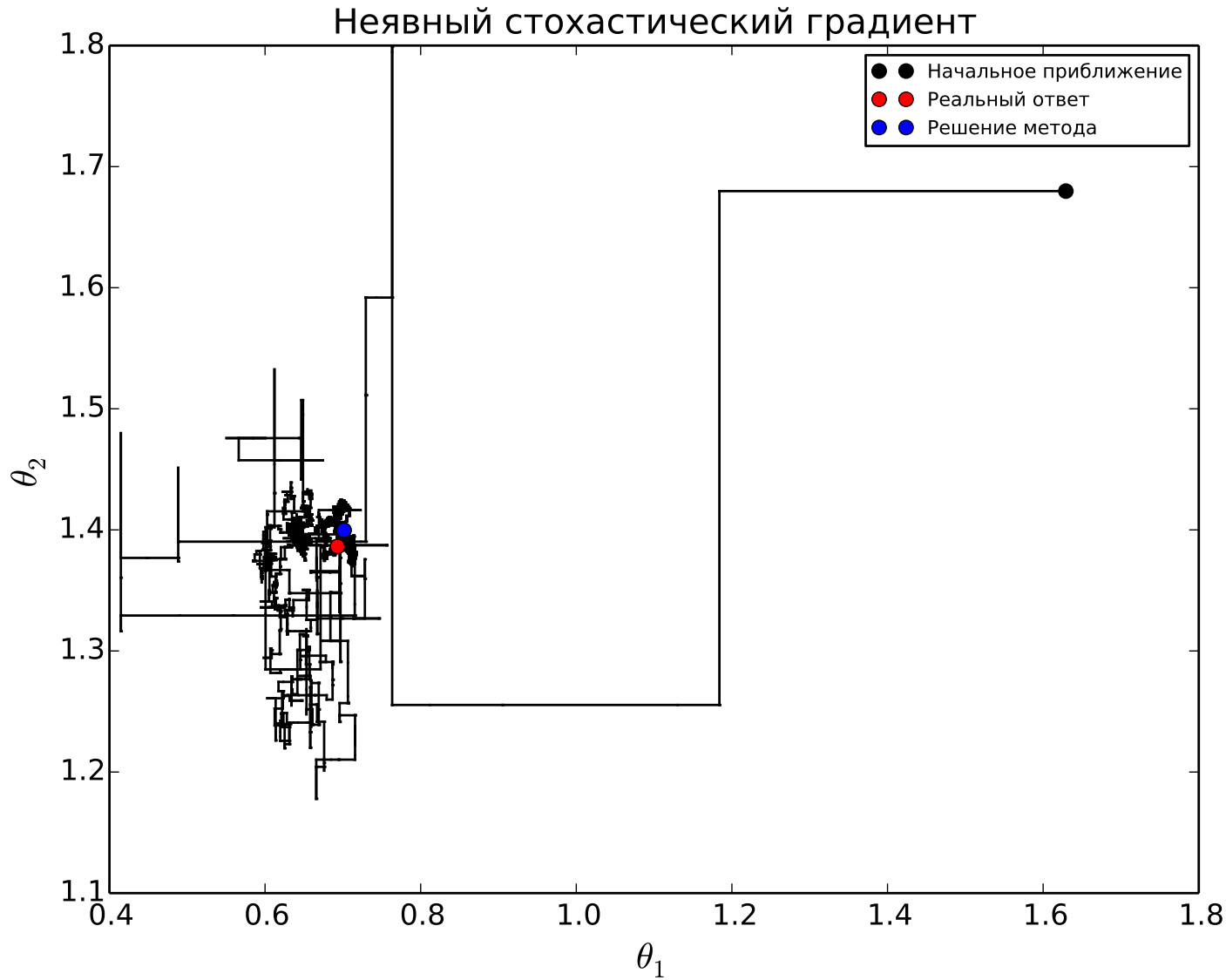


Рис. : Неявный метод работает обычно так

Эксперимент №1: Стабильность

Сравним отклонения оценок явного и неявного методов от реального значения параметра для разных квантилей.

Метод	25%	50%	75%	85%	95%	100%
Явный	0.01	0.03	215.5	1660.3	$\approx 10^5$	$\approx 10^8$
Неявный	0.00	0.01	0.02	0.02	0.03	0.04

Таблица : Квантили для полных 20000 итераций

В таблице взяты квантили отклонений по $m = 100$ оценкам для обоих методов.
Вывод: явный метод дает оценку с погрешностью в 0.03 всего в 50% случаев, неявный метод стабилен и показывает хорошую точность в 100% случаев.

Эксперимент №1: шаг градиента 1

- Вычислим матрицу Фишера

$$\mathcal{I}(\theta^*) = \begin{pmatrix} 0.4 & 0 \\ 0 & 0.8 \end{pmatrix}$$

- Все это время мы работали с плохом шагом для явного метода. Ведь $\alpha\lambda_{max} = 10/3 \times 8/10 = 8/3 > 2$, где $\lambda_{\max} = \max \operatorname{eig} \mathcal{I}(\theta^*)$
- Из доказательства Леммы можно предположить, что если $\alpha\lambda_{max} < 2$, то явный метод будет получать близкие к реальному значению параметра оценки. Проверим, так ли это
- Рассмотрим $\alpha = \frac{10}{3}, \frac{10}{5}, \frac{10}{7}, 1$
- Еще проверим, будет ли явный метод работать лучше при очень малых α

Эксперимент №1: шаг градиента 2

Здесь $10/3 \times \lambda_{\max} = 8/3 > 2$, а $10/5 \times \lambda_{\max} = 8/5 < 2$

Метод	25%	50%	75%	85%	95%	100%
Явный	0.01	0.03	215.5	1660.3	$\approx 10^5$	$\approx 10^8$
Неявный	0.00	0.01	0.02	0.02	0.03	0.04

Таблица : $\alpha = 10/3$

Метод	25%	50%	75%	85%	95%	100%
Явный	0.01	0.02	0.06	36.6	143.9	$\approx 10^5$
Неявный	0.01	0.01	0.02	0.02	0.03	0.04

Таблица : $\alpha = 10/5$

Видим, что для меньшего шага явный метод дает теперь хорошую точность в 75% случаях, но все же не всегда. Неявный метод стабилен.

Эксперимент №1: шаг градиента 3

Здесь $10/5 \times \lambda_{\max} = 8/5 < 2$, $10/7 \times \lambda_{\max} = 8/7 < 2$

Метод	25%	50%	75%	85%	95%	100%
Явный	0.01	0.02	0.06	36.6	143.9	$\approx 10^5$
Неявный	0.01	0.01	0.02	0.02	0.03	0.04

Таблица : $\alpha = 10/5$

Метод	25%	50%	75%	85%	95%	100%
Явный	0.01	0.02	0.07	0.4	130.1	$\approx 10^3$
Неявный	0.01	0.02	0.03	0.04	0.07	0.1

Таблица : $\alpha = 10/7$

При шаге $\alpha = 10/7$ у явного метода неплохая точность получается уже с вероятностью 85%. Заметим, что неявный метод стал давать оценку чуть хуже, чем раньше.

Эксперимент №1: шаг градиента 4

Здесь $10/7 \times \lambda_{\max} = 8/7 > 1$, $1 \times \lambda_{\max} = 8/10 < 1$

Метод	25%	50%	75%	85%	95%	100%
Явный	0.01	0.02	0.07	0.4	130.1	$\approx 10^3$
Неявный	0.01	0.02	0.03	0.04	0.07	0.1

Таблица : $\alpha = 10/7$

Метод	25%	50%	75%	85%	95%	100%
Явный	0.02	0.04	0.1	0.18	2.9	76.8
Неявный	0.02	0.03	0.07	0.1	0.27	0.82

Таблица : $\alpha = 1$

При шаге $\alpha = 1$ в принципе явный метод с большой вероятностью хорошо оценит наш параметр θ^*. И опять заметим, что неявный метода стал давать еще хуже оценку, чем раньше.

Эксперимент №1: шаг градиента 5

Что будет, если взять α еще меньше?

Метод	25%	50%	75%	85%	95%	100%
Явный	0.02	0.04	0.1	0.18	2.9	76.8
Неявный	0.02	0.03	0.07	0.1	0.27	0.82

Таблица : $\alpha = 1$

Метод	25%	50%	75%	85%	95%	100%
Явный	0.80	1.20	1.85	2.10	2.76	3.42
Неявный	0.82	1.22	1.85	2.09	2.76	3.42

Таблица : $\alpha = 1/10$

Видим, что при $\alpha = 1/10$ методы сравнялись по точности. Шаг стал очень маленьким, и теперь оба метода медленно сходятся, им не хватает 20 тысяч итераций.
Отметим, что явный метод теперь в 100% случаев дает хоть и не точную оценку, но более близкую, чем могла быть при больших α.

Эксперимент №2

Постановка эксперимента: на примере многомерного нормального распределения мы

- оценим параметр модели явным и неявным методами
- исследуем отклонения оценок обоих методов при различных параметрах α
- убедимся, что неявный метод работает надежно при больших α, в то время как явный — сильно расходится
- покажем, что неявный метод дает меньшую дисперсию оценки с итерациями, чем явный

Эксперимент №2

- Имеется модель многомерного нормального распределения
- Оптимизируемый функционал имеет следующий вид:

$$\log \mathcal{L}(\theta|X, Y) = \sum_i \left(\theta^T x_i y_i - \theta^T x_i\right) + \text{const}$$

- Оптимизацию будем производить онлайн, на каждой итерации получая x_i, y_i. Генерировать их будем следующим образом:
  - $x_n \sim \mathcal{N}_p(0, V_x)$, где V_x — матрица ковариации такая, что ее собственные значения равномерно распределены на отрезке $[0.2, 1]$, а ее максимальное собственное значение порядка размера пространства: $\lambda_{\max}^{V_x} = 0.1p$
  - $y_n \sim \mathcal{N}(\theta^{*T} x_n, 1)$
- Положим $\theta^* = (1, 1, \ldots, 1)^T$, $p = 20$
- Будем выполнять $N = 1000$ итераций для $m = 2000$ одинаковых начальных оценок
- Сразу отметим, что в нашем случае $\mathcal{I}(\theta^*) = V_x$

Эксперимент №2: оценка параметров

Запустим методы для $\alpha = 4/9$ (отметим, что $\alpha\lambda_{\max} < 1$) и получим $m = 2000$ оценок для каждого метода. Возьмем несколько оценок для первых 10 параметров (из 20)

- Для неявного метода:
  (0.98 1.03 1.01 0.98 0.88 1.02 0.97 0.99 1.12 1.10),
  (1.00 0.96 0.99 1.07 1.00 0.99 1.05 1.03 0.99 1.01),
  (1.01 1.07 0.89 0.91 1.04 0.99 1.01 1.00 0.91 0.94)
- Для явного метода:
  (0.86 **2.63** **2.24** 1.04 0.94 0.72 **2.06** 1.37 1.70 **2.21**),
  (1.01 1.10 1.15 0.97 0.96 1.01 0.93 1.01 0.86 1.00),
  (1.05 0.28 0.80 -**0.72** 0.78 0.85 0.61 0.13 0.64 0.61)
- Реальные значения:
  (1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00)

Видно, что явный метод дает хуже результаты, чем неявный.

Эксперимент №2: визуализируем по итерациям

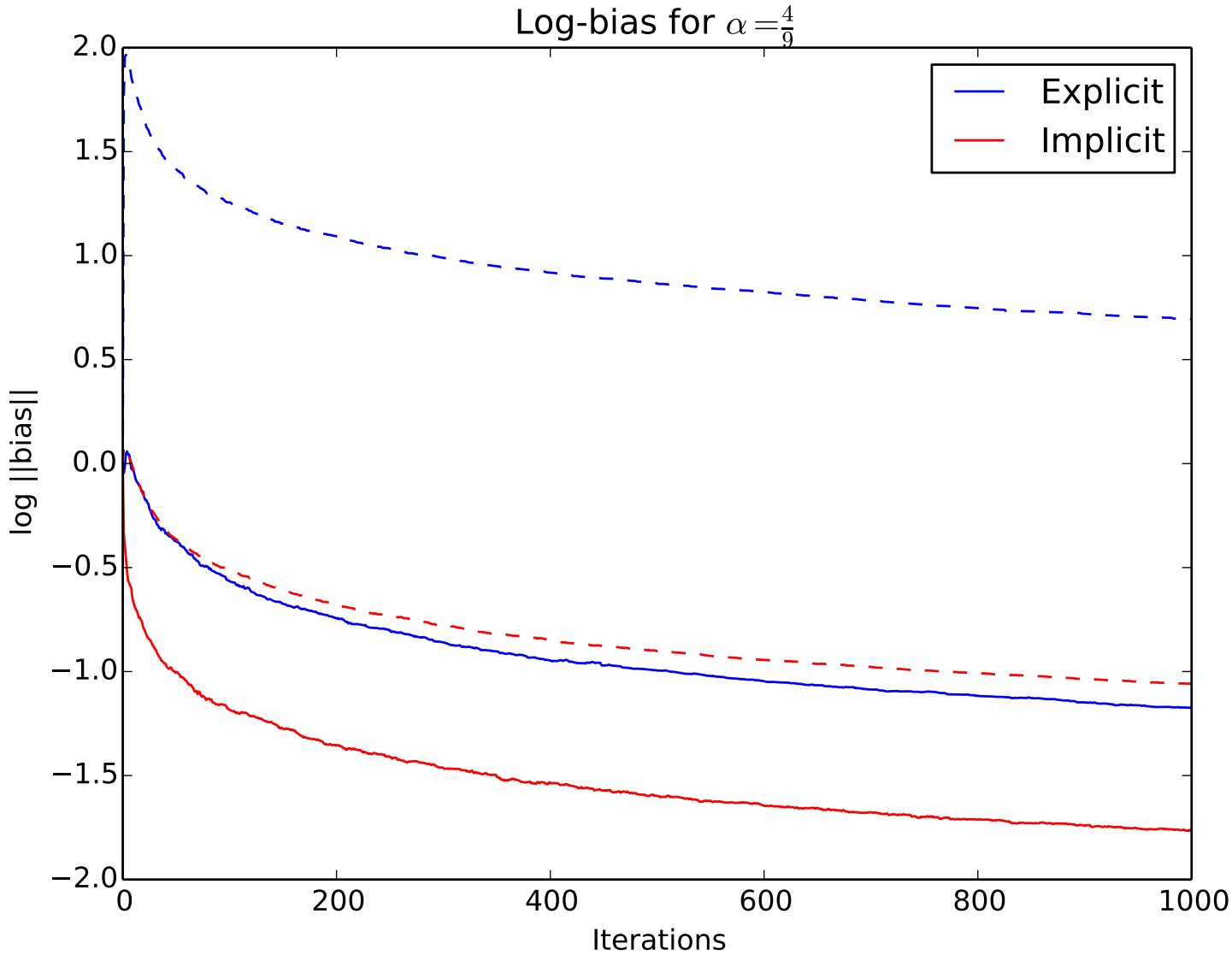


Рис. : Квантили 2% и 95% логарифмов отклонений

Рассмотрим получше предыдущий график

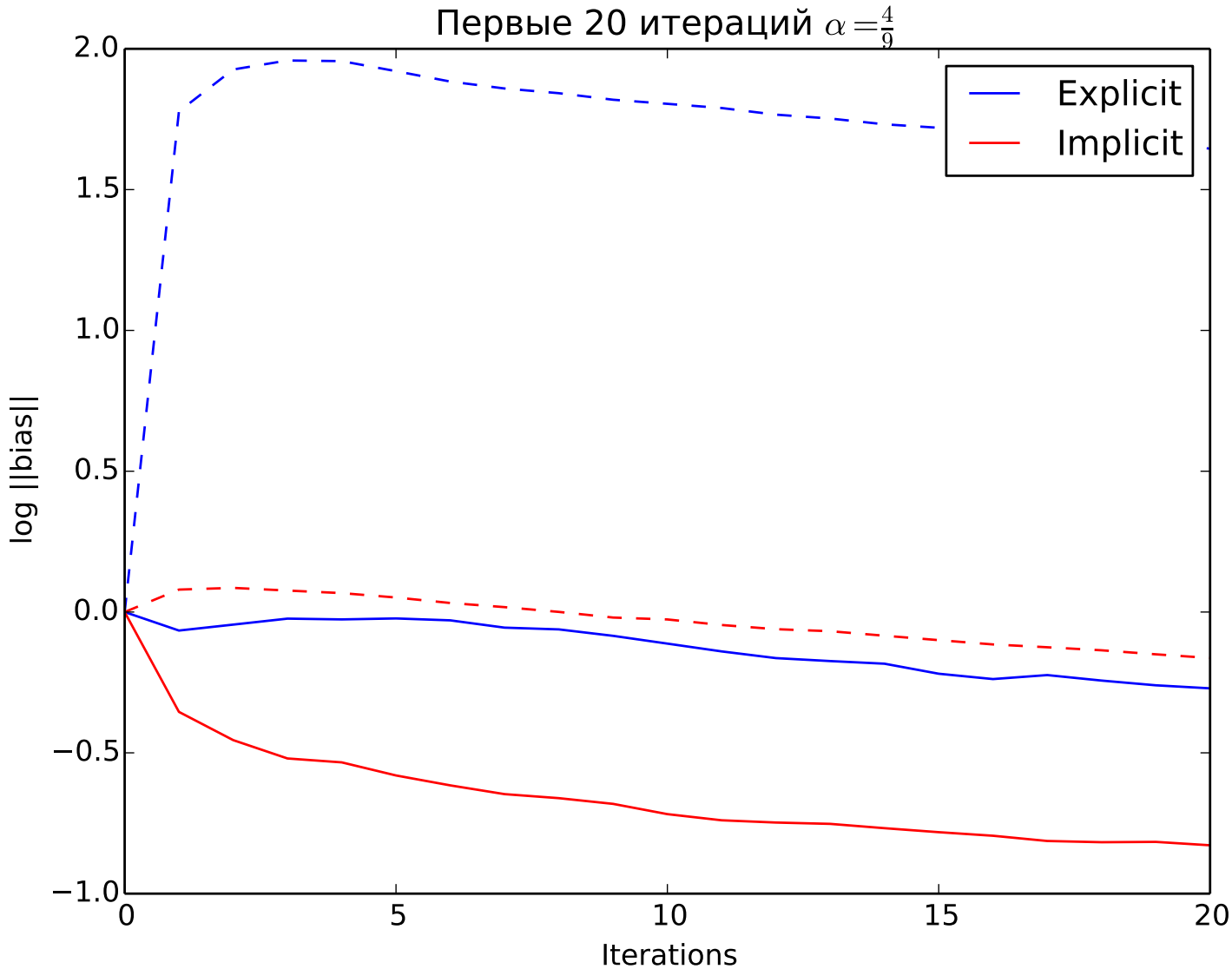


Рис. : Заметим возрастание 95% квантилей на первых итерациях

Попробуем меньшее $\alpha = 1/20$

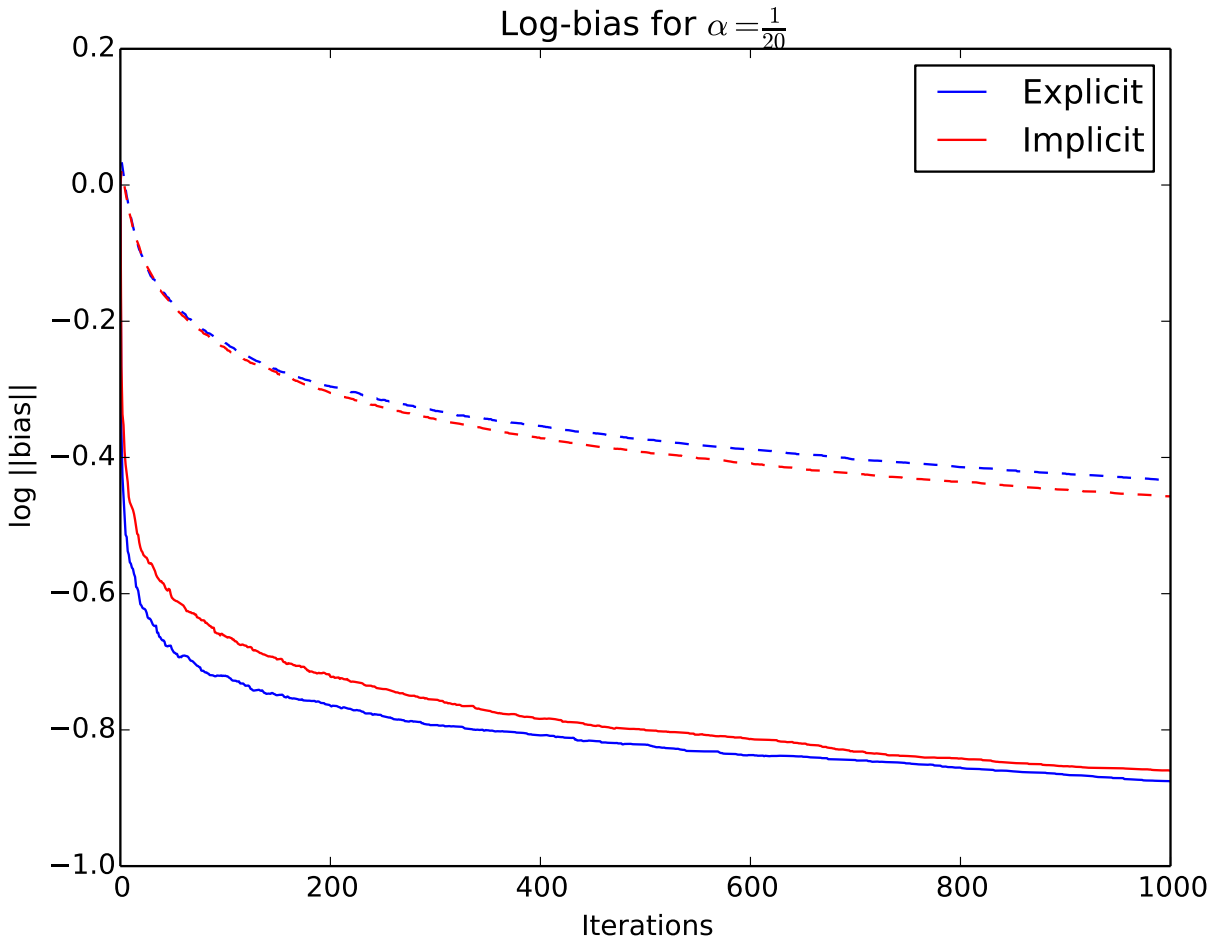


Рис. : Квантили 2% и 95% логарифмов отклонений

Еще раз убеждаемся, что для малых α методы показывают похожие результаты.

Эксперимент №2: различные шаги градиента

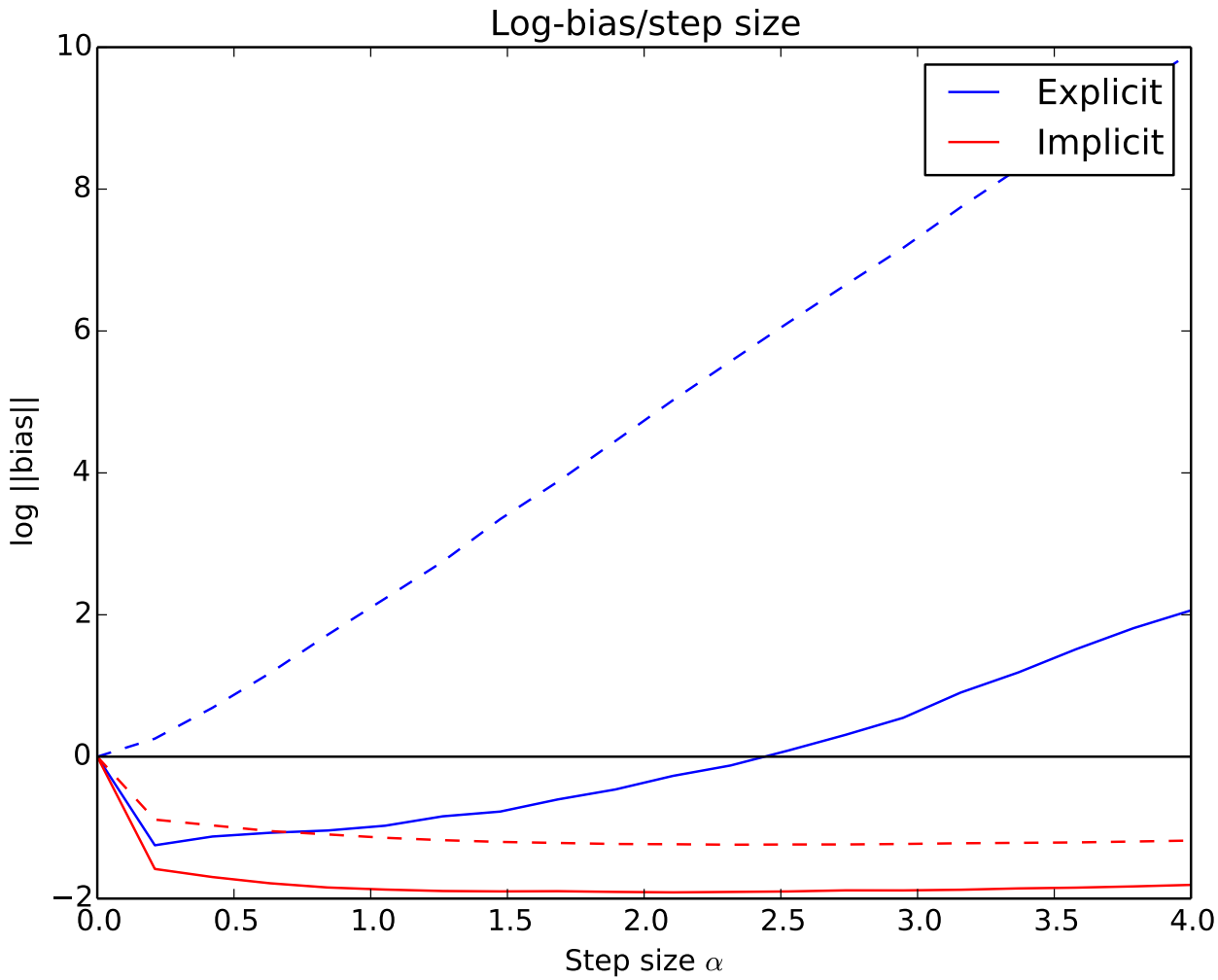


Рис. : Квантили 2% и 95% логарифмов отклонений при $\alpha = 4k/19, k = 0..19$

Эксперимент №2: Надежность

Покажем, что неявный метод дает более надежную оценку. То есть дисперсия оценок неявного метода с итерациями получается меньше, чем для явного.
Для этого будем на каждой итерации вычислять след матрицы ковариаций полученных оценок для обоих методов.

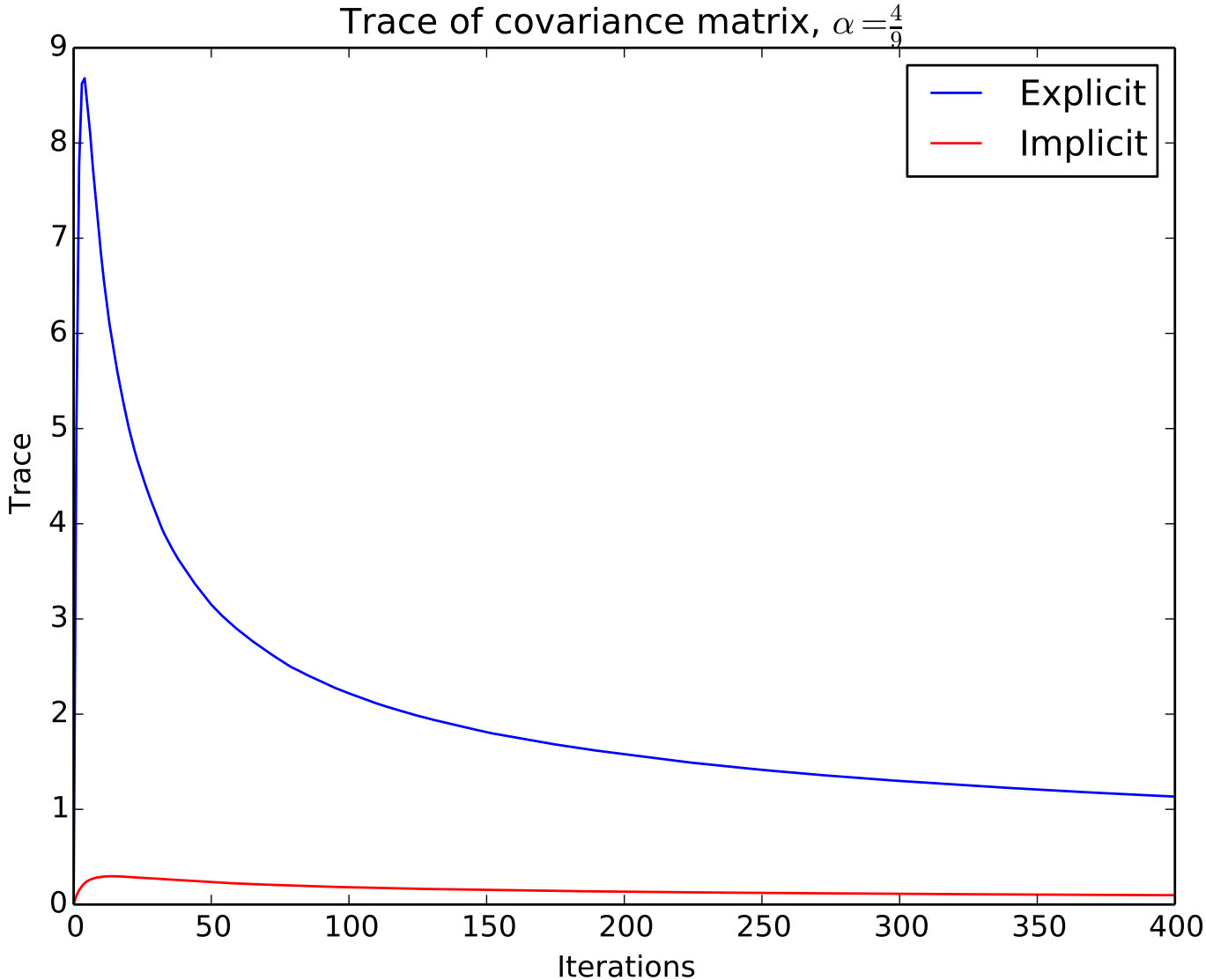


Рис. : Неявный метод показывает меньший разброс оценок

Литература

- Panos Toulis, Edoardo M. Airoldi, "Statistical analysis of stochastic gradient methods for generalized linear models" , ICML, Beijing China, 2014
- Robbins & Monro, "A Stochastic Approximation Method" , 1951
- Nelder and Wedderbur "Generalized linear models" , 1972
- Harold Kushner "Stochastic Approximation: A Survey" , Brown University, 2008