Дополнительная профессиональная программа
"Современные методы теории информации и оптимизации"

2 ноября 2022 г.

Исследование враждебного и случайного шумов в методах безградиентной оптимизации

Авторы:
Востриков Даниил, Конин Георгий,
Малышева Елизавета, Атласова Екатерина

Преамбула

Каждая вычислительная машина имеет конечную точность, которая обусловлена наличием конечной мантиссы. Вследствие этого накапливается ошибка, которая интерпретируется как шум. В данной работе исследуется метод SGD(GD) с использованием безградиентной оптимизации на классе гладких выпуклых задач с различными шумами. Рассматриваются две концепции шума: враждебный и случайный. Мы будем исследовать зависимость невязки по функции ε от размерности пространства d и величины шума Δ, чтобы понять, какой преимущественно характер носит машинный шум: враждебный или случайный.

Безградиентная оптимизация имеет широкое применение в:

1. Reinforcement learning
2. Подборе гиперпараметров
3. Дискретной оптимизации

Постановка задачи

Рассматривается оптимизационная задача:

$$\min_{x \in \mathbb{R}^d} f(x), \text{где } f(x) = \mathbb{E}_\xi f(x, \xi) \qquad (*)$$

"*$\varepsilon-$решением*" задачи $(*)$ назовём такой $\hat{x}^N$, для которого выполнено:

$$\mathbb{E}[f(\hat{x}^N)] - \min_{x \in \mathbb{R}^d} f(x) \leq \varepsilon$$

Предполагаем, что нам недоступны оракулы $1, 2, \ldots$ порядков, то есть мы можем оперировать только *оракулом 0-го порядка* с некоторым шумом δ:

$$x = (x_1, \ldots, x_d)$$

$$\downarrow$$

$$\boxed{\text{black-box}}$$

$$\downarrow$$

$$f_\delta(x) = f(x) + \delta(x)$$

Решаем задачу, пользуясь методом стохастического градиентного спуска (SGD):

$$x^{k+1} = x^k - h \cdot g(x^k), \quad \overline{k = 1, d}$$

где

$$g(x) = (\frac{\partial f_\delta}{\partial x_1}, \ldots, \frac{\partial f_\delta}{\partial x_d})^T$$

— аппроксимация истинного градиента конечными разностями.

Предположения

Предположение 1 (Выпуклость функции)

Функция $f(x)$, выпукла, то есть $\forall x', x'' \in dom f = \{x \in X : f(x) < +\infty\}$ и $\forall \alpha \in [0,1]$ выполнено неравенство Йенсена:

$$f((1-\alpha)x' + \alpha x'') \leq (1-\alpha)f(x') + \alpha f(x'') \qquad (1)$$

Предположение 2 (О липшицевости градиента)

Функция $f(x)$ непрерывно дифференцируема, а $\nabla f(x)$ является L-Липшицевым непрерывным для всех $x \in \mathbb{R}^d$, то есть выполнено:

$$\|\nabla f(x) - \nabla f(y)\|_q \leq L\|x - y\|_p, \tag{2}$$

где $\frac{1}{p} + \frac{1}{q} = 1$ (здесь и далее считаем, что $p = q = 2$.)

Предположение 3 (О липшицевости гессиана)

Функция $f(x)$ дважды непрерывно дифференцируема, а $\nabla^2 f(x)$ является M-Липшицевым непрерывным для всех $x \in \mathbb{R}^d$, то есть выполнено:

$$\|H(x) - H(y)\|_q \leq M\|x - y\|_p, \tag{3}$$

где $\frac{1}{p} + \frac{1}{q} = 1$.

Предположение 4 (Враждебный шум)

Шум зависит от входа x, при этом наиболее сильно зашумляя аппроксимацию градиента; единственное ограничение, которое мы накладываем - ограниченность шума по модулю:

$$\forall x \in X \rightarrow |\delta(x)| \leq \Delta \tag{4}$$

Предположение 5 (Случайный шум)

Для любых выбранных x', x'' шум не зависит от траектории, и второй момент ограничен константой σ^2, т.е.

$$\forall x', x'' \in X \rightarrow \mathbb{E}[\delta(x')^2] \leq \sigma^2, \mathbb{E}[\delta(x'')^2] \leq \sigma^2 \tag{5}$$

Методы аппроксимации градиента

Далее все оценки будем получать при Предположении 4 (считаем шум враждебным, т.к. теория случайного шума прелполагает, что ошибка накапливаться не будет вне зависимости от величины шума или размерности пространства).

1 Конечная прямая разность (FFD)

Первый метод, который мы анализируем - это стандартный метод конечных разностей. Приближение $\nabla f(x)$ прямой конечной разностью (FFD) при $x \in \mathbb{R}^d$ вычисляется с использованием множества $X = \{x + \gamma e_i\}_{i=1}^d \cup \{x\}$, где $\gamma > 0$ - шаг, а $e_i = (0,...1...,0)^T$, $i = \overline{1,d}$ (1 стоит на i-ой позиции), следующим образом:

$$\frac{\partial f_\delta}{\partial x_i} \approx \frac{f_\delta(x + \gamma e_i) - f_\delta(x)}{\gamma} = [g(x)]_i \tag{6}$$

Утверждение 1.1 Если верны предположения (1), (2), (4) и $g(x)$ - аппроксимация $\nabla f(x)$ конечной прямой разностью, то $[g(x)]_i \approx \frac{\partial f}{\partial x_i} + O(\sqrt{\Delta})$

Доказательство. Оценим (6): $f_\delta(x+\gamma e_i) - f_\delta(x) = f(x+\gamma e_i) - f(x) + \delta(x+\gamma e_i) - \delta(x) \leq f(x+\gamma e_i) - f(x) + 2\Delta \approx$ [*Разложим f(x) в ряд Тейлора*] $\approx < f'(x), \gamma e_i > + \frac{<\gamma e_i^T, f''(x)\gamma e_i>}{2} + 2\Delta = \gamma \cdot \frac{\partial f}{\partial x_i} + \frac{\gamma^2 L}{2} + 2\Delta$

Получаем, что $\frac{\partial f_\delta}{\partial x_i} \approx \frac{\gamma \cdot \frac{\partial f}{\partial x_i} + \frac{\gamma^2 L}{2} + 2\Delta}{\gamma} = \frac{\partial f}{\partial x_i} + \frac{\gamma L}{2} + \frac{2\Delta}{\gamma}$

Из приведенных выше оценок видно, что взаимосвязь между шагом γ и шумом Δ играет решающую роль в качестве аппроксимации. В частности, когда Δ равно нулю, то γ может быть выбрано сколь угодно малым и может быть получена близкая аппроксимация $\nabla f(x)$. С другой стороны, когда Δ большое, то малые значения

γ приводят к очень неточным аппроксимациям градиента.

Найдем минимум по γ:
$\frac{\partial(\frac{\partial f_\delta}{\partial x_i})}{\partial \gamma} = \frac{L}{2} - \frac{2\Delta}{\gamma^2} = 0 \to \gamma = 2\sqrt{\frac{\Delta}{L}}$
Значит $[g(x)]_i = \frac{\partial f_\delta}{\partial x_i} = \frac{\partial f}{\partial x_i} + 2\sqrt{\Delta L} = \frac{\partial f}{\partial x_i} + O(\sqrt{\Delta})$

□

Переходим к оценке нормы.

Теорема 1. Если верны предположения (1), (2), (4) и $g(x)$ - аппроксимация $\nabla f(x)$ конечной прямой разностью, то $\forall x \in \mathbb{R}^d$:

$$\|g(x) - \nabla f(x)\| \le \sqrt{d} \cdot \frac{L\gamma}{2} + \sqrt{d} \cdot \frac{2\Delta}{\gamma}$$

При $\gamma = \sqrt{2\frac{\Delta}{L}}$ - оптимальное для FFD, получаем, что $\|g(x) - \nabla f(x)\| \approx \sqrt{d} \cdot O(\sqrt{\Delta})$

Доказательство. Утверждение теоремы напрямую следует из утверждения 1.1. □

Теперь надо понять, как зависит невязка по функции ε от величины шума Δ и размерности пространства d: для этого норму разности $\|g(x) - \nabla f(x)\|$ приравняем к $\frac{\varepsilon}{R}$, где R - это норма точки оптимума.

Утверждение 1.2 Если верны предположения (1), (2), (4) и $g(x)$ - аппроксимация $\nabla f(x)$ конечной прямой разностью, то $\varepsilon = O(R\sqrt{\Delta d})$

Доказательство. $\|g(x) - \nabla f(x)\| \le L \cdot \sqrt{\frac{d\Delta}{L}} + \frac{d\Delta}{\sqrt{\frac{d\Delta}{L}}} = \sqrt{L \cdot d \cdot \Delta} + \sqrt{L \cdot d \cdot \Delta} = 2\sqrt{L \cdot d \cdot \Delta} = \frac{\varepsilon}{R} \to \varepsilon = O(R\sqrt{\Delta d})$ □

2 Конечная центральная разность (FCD)

Второй метод, который мы анализируем - это стандартный метод конечных разностей. Приближение $\nabla f(x)$ центральной конечной разностью (FCD) при $x \in \mathbb{R}^d$ вычисляется с использованием множества $X = \{x+\gamma e_i\}_{i=1}^d \cup \{x - \gamma e_i\}_{i=1}^d$, где $\gamma > 0$ - шаг, а $e_i = (0,...1...,0)^T$, $i = \overline{1,d}$ (1 стоит на i-ой позиции), следующим образом:

$$\frac{\partial f_\delta}{\partial x_i} \approx \frac{f_\delta(x + \gamma e_i) - f_\delta(x - \gamma e_i)}{2\gamma} = [g(x)]_i, \tag{7}$$

Утверждение 2.1 Если верны предположения (1), (3), (4) и $g(x)$ - аппроксимация $\nabla f(x)$ конечной прямой разностью, то $[g(x)]_i \approx \frac{\partial f}{\partial x_i} + O(\Delta^{\frac{2}{3}})$

Доказательство. Оценим (7):

$f_\delta(x + \gamma e_i) - f_\delta(x - \gamma e_i) = f(x + \gamma e_i) - f(x - \gamma e_i) + \delta(x + \gamma e_i) - \delta(x - \gamma e_i) \le f(x + \gamma e_i) - f(x - \gamma e_i) + 2\Delta \approx$ [*Разложим в ряд Тейлора*] $\approx f(x) + < f'(x), \gamma e_i > + \frac{1}{2} < \gamma e_i^T, f''(x)\gamma e_i > + \frac{1}{6}(\text{дописать}) - (f(x) + < f'(x), -\gamma e_i > + \frac{1}{2} < -\gamma e_i^T, f''(x)(-\gamma e_i) > + \frac{1}{6}(\text{дописать})) + 2\Delta = 2 < f'(x), \gamma e_i > + \frac{1}{3}(\text{дописать}) + 2\Delta = 2\gamma \cdot \frac{\partial f}{\partial x_i} + \frac{1}{3}(\text{дописать}) + 2\Delta \le 2\gamma \cdot \frac{\partial f}{\partial x_i} + \frac{1}{3}\gamma^3 \cdot M + 2\Delta$

Получаем, что $\frac{\partial f_\delta}{\partial x_i} \approx \frac{1}{2\gamma} \cdot (2\gamma\frac{\partial f}{\partial x_i} + \frac{1}{3}\gamma^3 M + 2\Delta) = \frac{\partial f}{\partial x_i} + \frac{1}{6}\gamma^2 M + \frac{\Delta}{\gamma}$

Аналогично предыдущему пункту, взаимосвязь между шагом γ и шумом Δ играет решающую роль в качестве аппроксимации.

Найдем минимум по γ:
$\frac{\partial(\frac{\partial f_\delta}{\partial x_i})}{\partial\gamma} = \frac{1}{3}\gamma M - \frac{\Delta}{\gamma^2} = 0 \to \gamma = (\frac{3\Delta}{M})^{\frac{1}{3}}$
Значит, $[g(x)]_i = \frac{\partial f_\delta}{\partial x_i} = \frac{\partial f}{\partial x_i} + \frac{1}{6}(\frac{\Delta}{3M})^{\frac{2}{3}} \cdot M + \frac{2\Delta^{\frac{2}{3}}}{(3M)^{\frac{1}{3}}} = \frac{\partial f}{\partial x_i} + O(\Delta^{\frac{2}{3}})$

□

Переходим к оценке нормы.

Теорема 2. Если верны предположения (1), (3), (4) и $g(x)$ - аппроксимация $\nabla f(x)$ конечной прямой разностью, то $\forall x \in \mathbb{R}^d$:

$$\|g(x) - \nabla f(x)\| \leq \sqrt{d} \cdot \frac{M\gamma^2}{6} + \sqrt{d} \cdot \frac{\Delta}{\gamma}$$

При $\gamma = (\frac{3\Delta}{M})^{\frac{1}{3}}$ - оптимальное для CFD, получаем, что $\|g(x) - \nabla f(x)\| \approx \sqrt{d} \cdot O(\Delta^{\frac{2}{3}})$

Доказательство. Утверждение теоремы напрямую следует из утверждения 2.1. □

Теперь надо понять, как зависит невязка по функции ε от величины шума Δ и размерности пространства d: для этого норму разности $\|g(x) - \nabla f(x)\|$ приравняем к $\frac{\varepsilon}{R}$, где R - это норма точки оптимума.

Утверждение 2.2 Если верны предположения (1), (3), (4) и $g(x)$ - аппроксимация $\nabla f(x)$ конечной прямой разностью, то $\varepsilon = O(R\Delta^{\frac{2}{3}}\sqrt{d})$

Доказательство. $\|g(x) - \nabla f(x)\| \leq \sqrt{d} \cdot \frac{M\gamma^2}{6} + \sqrt{d} \cdot \frac{\Delta}{\gamma} = \sqrt{d}(\frac{M \cdot (\frac{3\Delta}{M})^{\frac{2}{3}}}{6} + \frac{\Delta}{(\frac{3\Delta}{M})^{\frac{1}{3}}}) = \sqrt{d}(M^{\frac{1}{3}} \cdot \frac{3^{\frac{2}{3}}}{6} \cdot \Delta^{\frac{2}{3}} + \Delta^{\frac{2}{3}} \cdot \frac{1}{3^{\frac{1}{3}}} \cdot M^{\frac{1}{3}}) = \frac{\varepsilon}{R} \to \varepsilon = O(R\Delta^{\frac{2}{3}}\sqrt{d})$ □

3 Покомпонентные метод

FWC (forward wise component)
Рассмотрим:

$$g(x) = \frac{d \cdot (f_\delta(x + \gamma e_i) - f_\delta(x)) \cdot e_i}{\gamma}, \tag{8}$$

где e_i - рандомный вектор с 1 на i-ой из d позиций, $\gamma > 0$, d - коэффициент для несмещенности оценки.

Утверждение 3.1 Выражение (8) является несмещенной оценкой реального градиента.

Доказательство. $\mathbb{E}[g(x,e_i)] = \sum\limits_{i=1}^{d} \frac{1}{d} \cdot \frac{d \cdot (f_\delta(x+\gamma e_i) - f_\delta(x)) \cdot e_i}{gamma} = \sum\limits_{i=1}^{d} \frac{\partial f}{\partial x_i} \cdot e_i = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ 0 \\ . \\ . \\ . \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{\partial f}{\partial x_2} \\ . \\ . \\ . \\ 0 \end{pmatrix} + \ldots + \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ . \\ \frac{\partial f}{\partial x_d} \end{pmatrix} = \nabla f(x)$ □

При предположениях (1), (2), (4) оценки для покомпонентного метода FWC, очевидно, будут полностью аналогичны оценкам для конечного метода FFD.

CWC (central wise component)

Рассмотрим:

$$g(x) = \frac{d \cdot (f_\delta(x+\gamma e_i) - f_\delta(x-\gamma e_i)) \cdot e_i}{2\gamma}, \tag{9}$$

где e_i - рандомный вектор с 1 на i-ой из d позиций, $\gamma > 0$, d - коэффициент для несмещенности оценки.
Утверждение 3.1 Выражение (9) является несмещенной оценкой реального градиента.

Доказательство. Аналогично доказательству утверждения 3.1. □

При предположениях (1), (3), (4) оценки для покомпонентного метода CWC, очевидно, будут полностью аналогичны оценкам для конечного метода CFD.

4 l2 рандомизация на сфере

Теперь рассмотрим алгоритм рандомизированной аппроксимации: рандомизация на l_2 сфере $S_2^d = \{a \in \mathbb{R}^d : \|a\|_2 = 1\}$

FSSG2(forward sphere smoothing gradients l2)
а)Пусть $f(x)$ удовлетворяет условиям (1), (2), (4), тогда выберем следующую схему:

$$g(x) = d \cdot \frac{f_\delta(x+\gamma e) - f_\delta(x)}{\gamma} \cdot e \tag{12}$$

где $\gamma > 0$, $e \sim (S_2^d)$.

Опять же, коэффициент d в формуле (12) нужен для несмещенности оценки относительно реального градиента функции $f(x)$.

Теорема 3. $\|\mathbb{E}[g(x)] - \nabla f(x)\| \leq L\gamma + \frac{d\Delta}{\gamma}$
Доказательство: см. в статье Berahas 2021 (`https://arxiv.org/pdf/1905.01332.pdf`)

Найдем оптимальное γ:
$\frac{d}{d\gamma}(L\gamma + \frac{d\Delta}{\gamma}) = L - \frac{d\Delta}{\gamma^2} = 0 \rightarrow \gamma = \sqrt{\frac{d\Delta}{L}}$

Найдем зависимость невязки по функции ε от величины шума Δ и размерности пространства d:
$\|\mathbb{E}[g(x)] - \nabla f(x)\| \leq L \cdot \sqrt{\frac{d\Delta}{L}} + \frac{d\Delta}{\sqrt{\frac{d\Delta}{L}}} = \sqrt{L \cdot d \cdot \Delta} + \sqrt{L \cdot d \cdot \Delta} = 2\sqrt{L \cdot d \cdot \Delta} = \frac{\varepsilon}{R} \rightarrow \varepsilon = 2R\sqrt{L \cdot d \cdot \Delta}$.

CSSG2(central sphere smoothing gradients l2)
б) Пусть $f(x)$ удовлетворяет условиям (1), (3), (4), тогда выберем следующую схему:

$$g(x) = d \cdot \frac{f_\delta(x+\gamma e) - f_\delta(x-\gamma e)}{2\gamma} \cdot e \tag{13}$$

где $\gamma > 0$, $e \sim (S_2^d)$.

Опять же, коэффициент d в формуле (13) нужен для несмещенности оценки относительно реального градиента функции $f(x)$.

Теорема 4. $\|\mathbb{E}[g(x)] - \nabla f(x)\| \leq M\gamma^2 + \frac{d\Delta}{\gamma}$
Доказательство: см. в статье Berahas 2021 (`https://arxiv.org/pdf/1905.01332.pdf`)

Найдем оптимальное γ:
$\frac{d}{d\gamma}(M\gamma^2 + \frac{d\Delta}{\gamma}) = 2M\gamma - \frac{d\Delta}{\gamma^2} = 0 \rightarrow \gamma = (\frac{d\Delta}{2M})^{\frac{1}{3}}$

Найдем зависимость невязки по функции ε от величины шума Δ и размерности пространства d:
$\|\mathbb{E}[g(x)] - \nabla f(x)\| = M \cdot (\frac{d\Delta}{2M})^{\frac{2}{3}} + \frac{d\Delta}{(\frac{d\Delta}{2M})^{\frac{1}{3}}} = \frac{M^{\frac{1}{3}}}{2^{\frac{2}{3}}} \cdot d^{\frac{2}{3}}\Delta^{\frac{2}{3}} + 2^{\frac{1}{3}} \cdot M^{\frac{1}{3}} \cdot d^{\frac{2}{3}} \cdot \Delta^{\frac{2}{3}} = const \cdot M^{\frac{1}{3}} d^{\frac{2}{3}} \Delta^{\frac{2}{3}} = \frac{\varepsilon}{R} \rightarrow \varepsilon \sim R \cdot M^{\frac{1}{3}} \cdot d^{\frac{2}{3}} \cdot \Delta^{\frac{2}{3}}$.