УДК 517.97 + 004.852

ГРАДИЕНТНЫЙ СПУСК И МЕТОДЫ ЕГО ОПТИМИЗАЦИИ

Сороколетов А.А., Блышко Н.В., студенты

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Малышева О.Н. - канд. физ. -мат. наук, доцент

Аннотация. В работе выполнен сравнительный анализ классического градиентного спуска и алгоритма Adam с позиций математического анализа и механической интерпретации. Показано, что градиентный спуск является схемой Эйлера для релаксации системы в поле потенциальных сил, а Adam реализует адаптивное диагональное предобуславливание на основе оценок первого и второго моментов стохастического градиента. Для квадратичной функции и функции Розенброка сопоставлены скорость сходимости, устойчивость к выбору шага.

Ключевые слова. градиентный спуск, Adam, оптимизация градиентного спуска, функция Розенброка, сходимость градиентного спуска, стохастический градиент, поле потенциальных сил.

Введение. Целью работы является сравнительный анализ классического градиентного спуска и алгоритма Adam с позиций математического анализа, методов машинного обучения и физической интерпретации динамики оптимизации.

В математике градиент функции ∇f(x) представляет собой вектор частных производных, направленный в сторону наибольшего роста функции. В задачах машинного обучения градиентный спуск используется для минимизации функции потерь L(w) → min по параметрам модели w. Стохастический градиент является оценкой точного градиента, вычисляемой по мини-выборке данных. Алгоритм Adam развивает этот подход, используя адаптивные оценки первого и второго моментов стохастического градиента для покоординатного выбора эффективного шага обучения.

Минимизация функции потерь $L(w)$ занимает центральное место в задачах машинного обучения и математической физики. Пространство параметров можно интерпретировать как конфигурационное пространство механической системы, а саму функцию потерь - как потенциальную энергию. Такая аналогия позволяет строго описывать методы оптимизации через уравнения движения и свойства диссипативных систем.

Основная часть.

1 Классический градиентный спуск. Для дифференцируемой функции $L(w)$ итерационный процесс имеет вид

Введем функцию потерь L(w) → min, которая интерпретируется как потенциальная энергия системы. Тогда антиградиент -∇L(w) задает поле потенциальной силы, а сам градиентный спуск описывает релаксацию системы в вязкой среде к состоянию минимальной энергии.

$$w_{t+1} = w_t - \eta \nabla L(w_t) \tag{1}$$

где $\eta > 0$ - шаг обучения. При переходе к непрерывному пределу $\tau = \eta t$ получаем уравнение релаксации

$$\frac{dw}{d\tau} = -\nabla L(w). \tag{2}$$

Оно совпадает с движением частицы в вязкой среде без инерционного члена. Для L-липшицева градиента изменение функции Ляпунова оценивается неравенством

$$L(w_{t+1}) - L(w_t) \leq -\eta \left(1 - \frac{\eta L}{2}\right) \| \nabla L(w_t) \|^2 \tag{3}$$

откуда при $0 < \eta < 2/L$ следует монотонное убывание функционала и сходимость метода. На квадратичной функции $L(w) = 1/2\, w^T A w$ скорость сходимости определяется спектральным радиусом матрицы перехода

$$\rho(I - \eta A) = \max_i |1 - \eta \lambda_i|. \tag{4}$$

При большом числе обусловленности минимум расположен в узкой долине, поэтому обычный градиентный спуск требует малого шага и сходится медленно.

2 Алгоритм Adam. В стохастической постановке вместо точного градиента используется оценка $g_t = \nabla \ell(w_t, \xi_t)$. Алгоритм Adam строит экспоненциальные скользящие средние первого и второго моментов:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \tag{5}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2. \tag{6}$$

Поскольку на первых шагах эти оценки смещены к нулю, вводятся поправки

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \qquad \hat{v}_t = \frac{v_t}{1-\beta_2^t}, \tag{7}$$

после чего обновление параметров записывается в форме

$$w_{t+1} = w_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}. \tag{8}$$

Таким образом, Adam автоматически уменьшает эффективный шаг в направлениях с высокой дисперсией градиента и ускоряет движение в более стабильных координатах. С физической точки зрения это эквивалентно адаптивному изменению «трения» вдоль различных направлений движения системы.

3 Численный эксперимент. Сравнение классического градиентного спуска и метода Adam проводилось авторами для квадратичной функции $f(x,y) = x^2 + 10y^2$ и для функции Розенброка

$$f(x,y) = (1-x)^2 + 100(y-x^2)^2, \tag{9}$$

которая моделирует узкий нелинейный овраг и является показательной для методов первого порядка. Численные расчеты показывают, что на хорошо обусловленной квадратичной функции SGD сходится немного быстрее, тогда как на функции Розенброка Adam выигрывает за счет адаптивного масштабирования шага и лучшей устойчивости к выбору параметра η.

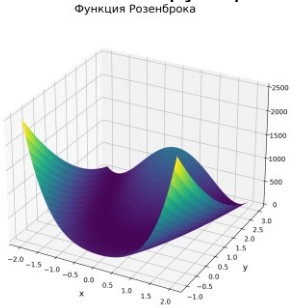

Рисунок 1 - График функции Розенброка

Для квадратичной функции выигрыш от адаптивности невелик, поскольку кривизна поверхности постоянна и одинакова по всем итерациям. Напротив, в узкой долине функции Розенброка эффективный шаг обычного градиентного спуска приходится выбирать слишком малым, что замедляет движение вдоль долины. Adam автоматически уменьшает шаг по жестким координатам и сохраняет более быстрое продвижение к минимуму по мягким направлениям.

Количественные результаты приведены в таблице 1. Они подтверждают, что Adam обеспечивает более высокую устойчивость к выбору шага обучения, хотя на простой квадратичной функции адаптивность не дает преимущества по числу итераций.

На рисунках 2 и 3 приведены кривые сходимости SGD и Adam для квадратичной функции и функции Розенброка. Графики наглядно демонстрируют, что на хорошо обусловленной поверхности различия между методами невелики, тогда как в узкой нелинейной долине адаптивное масштабирование Adam обеспечивает более устойчивое уменьшение функционала.

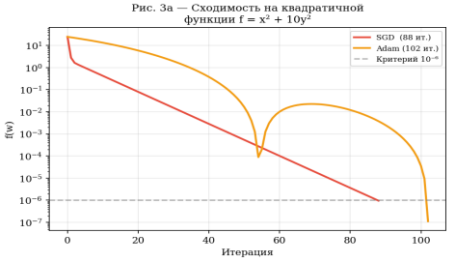

Рисунок 2 - Кривые сходимости SGD и Adam для квадратичной функции

На рисунке 2 обе кривые убывают устойчиво, причем SGD немного быстрее уменьшает ошибку на начальных итерациях. Для квадратичной функции кривизна поверхности постоянна, поэтому адаптивное изменение шага в Adam не дает заметного преимущества.

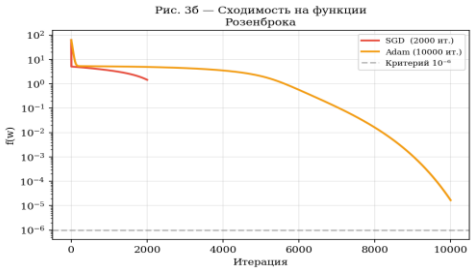

Рисунок 3 - Кривые сходимости SGD и Adam для функции Розенброка

На рисунке 3 SGD испытывает трудности при движении вдоль узкой искривленной долины и требует малого шага. Adam уменьшает шаг по жестким координатам и сохраняет более плавное продвижение к минимуму, поэтому спад функционала остается более устойчивым.

Таблица 1 - Сравнение числа итераций до достижения точности $f(w_t) < 10^{-6}$

Метод	**Квадратичная функция**	**Функция Розенброка**	**Устойчивость к выбору шага**
SGD	88	> 2000	Низкая
Adam	102	10 472	Высокая

Заключение. Градиентный спуск допускает строгую физико-математическую интерпретацию как диссипативное движение к минимуму потенциальной энергии. Алгоритм Adam развивает эту схему за счет адаптивного оценивания моментов и координатного предобусловливания. Проведенный анализ показывает, что классический SGD эффективен на простых и хорошо обусловленных задачах, тогда как Adam особенно полезен на сложных рельефах с резко различающейся кривизной по координатам.

Список использованных источников:

1. Kingma, D. P. Adam: A Method for Stochastic Optimization / D. P. Kingma, J. Ba // ICLR. - 2015.

2. Nesterov, Y. Lectures on Convex Optimization / Y. Nesterov. - 2nd ed. - Springer, 2018. - 589 p.

3. Bottou, L. Optimization Methods for Large-Scale Machine Learning / L. Bottou, F. Curtis, J. Nocedal // SIAM Review. - 2018. - Vol. 60, No. 2. - P. 223-311.

4. Goodfellow, I. Deep Learning / I. Goodfellow, Y. Bengio, A. Courville. - MIT Press, 2016.

UDC 517.97 + 004.852

GRADIENT DESCENT AND THE ADAM ALGORITHM: PHYSICAL-MATHEMATICAL ANALYSIS OF OPTIMIZATION METHODS

Sorokoletov A.A., Blyshko N.V., students

Belarusian State University of Informatics and Radioelectronics[1]
Minsk, Republic of Belarus

Malysheva O.N. - PhD in Physics and Mathematics, Associate Professor

Abstract. The paper compares classical gradient descent and the Adam algorithm from the viewpoints of mathematical analysis and physical interpretation. Gradient descent is treated as the Euler discretisation of a dissipative motion in a potential field, while Adam is interpreted as an adaptive preconditioned method based on first- and second-moment estimates of the stochastic gradient. Numerical experiments on a quadratic function and the Rosenbrock function demonstrate the influence of conditioning on convergence rate and show the practical advantage of Adam on ill-conditioned landscapes.

Keywords. gradient descent, Adam, optimisation, Rosenbrock function, convergence, stochastic gradient, potential energy.