ЭВОЛЮЦИЯ И ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ ГРАДИЕНТНЫХ МЕТОДОВ В ОБУЧЕНИИ АДАПТИВНЫХ СИСТЕМ

Яцук К.М., Пищ Т.Г., студенты

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

Князюк Н.В.– канд. физ.-мат. наук, доцент

Аннотация. В статье исследуется генезис и современное состояние градиентных методов оптимизации для настройки адаптивных систем. Рассматривается переход от классических алгоритмов первого порядка к инерционным и адаптивным процедурам с динамическим шагом. Проведён сравнительный анализ методов Поляка, Нестерова, Немировского и алгоритмов семейства Adam. Установлено, что современные модификации градиентного спуска обеспечивают прирост скорости сходимости устойчивость к стохастическим шумам.

Введение. В современной теории управления и машинного обучения центральное место занимает задача минимизации функционала ошибки, определяющего качество функционирования адаптивной системы. Процесс обучения такой системы сводится к итерационному поиску вектора параметров x^*, доставляющего минимум функции потерь $f(x)$. В условиях экспоненциального роста объемов данных и усложнения архитектур адаптивных моделей, выбор алгоритма оптимизации становится **одним из** определяющих факторов не только качества обучения, но и его технической реализуемости в реальном времени. Градиентные методы оптимизации остаются наиболее востребованным инструментарием благодаря их относительно низкой вычислительной сложности и возможности эффективного распараллеливания. Именно поэтому изучение эволюции градиентных методов и их сравнительный анализ представляют как теоретический, так и прикладной интерес.

1. Теоретические основы и классические подходы. Базовая парадигма градиентного спуска основывается на итерационном процессе движения в направлении антиградиента целевой функции. В общем случае закон обновления параметров имеет вид:

$$x_{k+1} = x_k - \eta_k \nabla f(x_k), \quad (1)$$

где η_k – параметр, определяющий величину шага; x_k – вектор параметров модели на k-й итерации; x_{k+1} – обновлённый вектор параметров на (k+1)-й итерации; $\nabla f(x_k)$ – градиент функции потерь $f(x)$, вычисленный в точке x_k.

Метод наискорейшего спуска, предложенный О. Коши, предполагает аналитическое определение оптимального η_k на каждой итерации. Однако в практических задачах адаптации это часто приводит к «зигзагообразному» движению на овражных поверхностях, так как классический алгоритм игнорирует информацию о локальной кривизне ландшафта функции потерь. На практике это означает, что при вытянутых долинах функции потерь классический градиентный спуск может потребовать в десятки раз больше итераций, чем более современные аналоги. Это ограничение заставляет искать способы учета динамики изменения градиента на предыдущих итерациях.

2. Инерционные методы и ускорение сходимости. Развитие методов в середине XX века было направлено на преодоление ограничений простого градиентного спуска. Метод «тяжелого шарика» Б.Т. Поляка (1964) внедрил концепцию импульса (momentum), позволяющую учитывать предысторию движения [1]:

$$x_{k+1} = x_k - \eta \nabla f(x_k) + \beta(x_k - x_{k-1}), \quad (2)$$

где $\beta \in [0,1)$ – коэффициент инерции. Введение этой составляющей позволяет алгоритму демпфировать высокочастотные колебания и быстрее проходить пологие участки, обеспечивая более **плавное и** стабильное продвижение к глобальному экстремуму в условиях зашумленности.

Качественным скачком стало появление ускоренного метода Ю.Е. Нестерова (1983) [2]. В отличие от метода Поляка, градиент здесь вычисляется в «предсказанной» точке, смещенной по вектору импульса:

$$y_k = x_k + \beta_k(x_k - x_{k-1}), \quad (3)$$

$$x_{k+1} = y_k - \eta \nabla f(y_k). \quad (4)$$

Такой «упреждающий» расчет позволяет алгоритму вовремя замедляться перед минимумом, обеспечивая теоретическую скорость сходимости порядка $O(1/k^2)$, что является оптимальным для гладких выпуклых функционалов.

3. Проксимальные методы и неевклидова геометрия. Проксимальные подходы, фундаментально развитые А.С. Немировским [3], расширили область применения градиентных методов на задачи с негладкими регуляризаторами и сложными ограничениями. Метод зеркального спуска адаптирует геометрию поиска к внутренней структуре пространства параметров, используя функции дивергенции вместо стандартной евклидовой метрики. Это позволяет интегрировать априорные знания о структуре искомого решения (например, разреженность весов при L_1 регуляризации) непосредственно в итерационный процесс, что существенно повышает интерпретируемость и помехоустойчивость адаптивной системы.

4. Адаптивные алгоритмы современности. Для современных адаптивных систем характерна разреженность данных, когда разные параметры требуют существенно разных скоростей обучения. Алгоритмы семейства AdaGrad, RMSprop и Adam решают эту проблему **эффективно** через автоматическую нормировку градиента [4].

В частности, алгоритм Adam (от англ. *Adaptive Moment Estimation* — адаптивная оценка моментов) был предложен Дидериком Кингмой (Diederik P. Kingma) и Джимми Ба (Jimmy Lei Ba) и впервые опубликован в 2014 году в препринте «Adam: A Method for Stochastic Optimization» [5], принятом на конференцию ICLR 2015. Алгоритм объединяет идеи накопления импульса и адаптивной корректировки шага, одновременно отслеживая скользящее среднее градиента (первый момент) и скользящее среднее квадрата градиента (второй момент). Благодаря этому темп обучения настраивается индивидуально для каждого параметра модели, что особенно эффективно при работе с разреженными данными. Статья Кингмы и Ба собрала свыше 200 000 цитирований и стала одной из наиболее цитируемых работ в области машинного обучения за всё время. Вместе с тем в ряде задач выпуклой оптимизации Adam может не сходиться к глобальному минимуму, что послужило основой для последующих модификаций — AMSGrad, AdamW и Nadam, устраняющих данное ограничение.

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

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

где m_t – первый момент (оценка среднего значения градиента) на шаге t; v_t – второй момент (оценка несмещенной дисперсии, т.е. суммы квадратов) на шаге t; m_{t-1} – значения первого момента с предыдущего шага; v_{t-1} – значения второго момента с предыдущего шага; $\beta_1, \beta_2 \in [0,1)$ – коэффициенты затухания (по умолчанию $\beta_1 = 0.9$, $\beta_2 = 0.999$); g_t – градиент функции потерь на итерации t; g_t^2 – поэлементный квадрат градиента .

Механизмы коррекции смещения (bias correction) на начальных итерациях обеспечивают стабильность оценок даже при малом количестве накопленных данных. Это делает Adam универсальным инструментом, минимизирующим необходимость ручной настройки коэффициента обучения и обеспечивающим высокую точность аппроксимации в нелинейных системах.

Заключение. Эволюция градиентных методов прошла путь от простых итерационных процедур до интеллектуальных адаптивных алгоритмов, учитывающих динамику обучения и геометрию параметров. Практическое применение методов типа Adam и NAG позволяет сократить время сходимости на 50–70% по сравнению с классическими подходами. Дальнейшие перспективы исследований лежат в области разработки гибридных архитектур, сочетающих адаптивность современных методов с гарантированной сходимостью второго порядка в условиях высокой размерности. Таким образом, выбор оптимизатора становится неотъемлемой частью проектирования адаптивной системы наравне с выбором её архитектуры.

Список использованных источников:

1. Поляк Б. Т. Введение в оптимизацию. – М.: Наука, Гл. ред. физ.-мат. лит., 1983. – 384 с.

2. Нестеров Ю. Е. Методы выпуклой оптимизации. – М.: МЦНМО, 2010. – 704 с.

3. Немировский А. С., Юдин Д. Б. Сложность задач и эффективность методов оптимизации. – М.: Наука, 1979. – 384 с.

4. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение / пер. с англ. А. А. Слинкина. – 2-е изд. – М.: ДМК Пресс, 2018. – 652 с

5. Kingma D. P., Ba J. Adam: A Method for Stochastic Optimization // International Conference on Learning Representations (ICLR). — 2015. — arXiv:1412.6980.

6. Воронцов К. В. Лекции по методам оптимизации. – М.: МФТИ, 2014. – 164 с.