Известия Гомельского государственного университета
имени Ф. Скорины, № 3 (132), 2022

УДК 004.54

Эффективная адаптация скорости машинного обучения на основе иерархического подхода к оптимизации

С.И. Жогаль, С.П. Жогаль, Н.А. Алёшин, В.В. Орлов

Рассмотрен иерархический подход к адаптации скорости обучения в градиентных методах, называемый оптимизацией скорости обучения (ОСО). ОСО формулирует проблему адаптации скорости обучения как задачу иерархической оптимизации, которая минимизирует функцию потерь по отношению к скорости обучения для текущих параметров и градиентов модели. Затем ОСО оптимизирует скорость обучения на основе метода множителей переменного направления. В процессе этой оптимизации не требуется никакой информации второго порядка и вероятностной модели, поэтому он очень эффективен. Кроме того, ОСО не требует дополнительных гиперпараметров по сравнению с методом градиента с простым экспоненциальным спадом скорости обучения. Если сравнить эффективность оптимизации с современными методами адаптации скорости обучения, а также с наиболее часто используемыми методами адаптивного градиента, то ОСО превосходит другие методы в задачах классификации.
Ключевые слова: глубокое обучение, машинное обучение, математическая оптимизация.

A hierarchical approach to adapting the learning rate in gradient methods, called learning rate optimization (LRO), is considered. LRO formulates the learning rate adaptation problem as a hierarchical optimization problem that minimizes the loss function with respect to the learning rate for current model parameters and gradients. LRO then optimizes the learning rate based on the alternating direction multiplier method. In the process of this optimization it does not require any second order information and a probabilistic model, so it is very efficient. In addition, LRO does not require any additional hyperparameters compared to the gradient method with a simple exponential learning rate decay. If we compare the optimization efficiency with modern learning rate adaptation methods, as well as with the most commonly used SGD adaptive gradient methods, then LRO outperforms all competitors in classification tasks.
Keywords: deep learning, machine learning, mathematical optimization.

Введение. Градиентные методы широко используются для подбора параметров модели в интеллектуальном анализе данных и машинном обучении, благодаря их эффективности. В градиентных методах скорость обучения (или размер шага) является одним из наиболее важных параметров, определяющих общую производительность оптимизации. По этой причине методы адаптации скорости обучения широко изучались с различных точек зрения, таких как информация второго порядка, обучение с подкреплением и статистический тест [1]. Однако эти методы почти не используются на практике из-за сложной реализации, больших трудоемких вычислений.

Для эффективной и действенной адаптации скорости обучения был предложен новый иерархический подход к адаптации скорости обучения, называемый оптимизацией скорости обучения (ОСО). ОСО формулирует проблему адаптации скорости обучения как проблему субоптимизации градиентных методов. В частности, скорость обучения оптимизирована для минимизации линеаризованной функции потерь для текущих параметров и градиентов модели. Затем оптимальная скорость обучения эффективно рассчитывается на основе метода множителей переменных направлений [2]. Поскольку ОСО напрямую оптимизирует скорость обучения, чтобы минимизировать функцию потерь без каких-либо эвристических предположений, мотивация и причина повышения производительности ОСО ясны и хорошо обоснованы.

ОСО не требует каких-либо архитектурных ограничений для моделей прогнозирования, а также трудоемкой информации второго порядка, вероятностных моделей и статистических тестов для адаптации скорости обучения. Следовательно, ОСО эффективен и прост в реализации, дополнительная временная сложность от применения ОСО в градиентных методах незначительна в задачах оптимизации большого масштаба, таких как обучение глубоких

нейронных сетей. Кроме того, ОСО имеет только два гиперпараметра: начальную верхнюю границу скорости обучения и коэффициент затухания, что равно количеству гиперпараметров метода ванильного градиента с экспоненциальным спадом скорости обучения. В ОСО нет дополнительных гиперпараметров, если сравнивать его с методом простого градиента [1].

Метод ОСО значительно улучшил сходимость и оптимизацию метода градиентного спуска. Кроме того, градиентные методы с ОСО значительно превзошли популярные и современные градиентные методы при обучении глубоких нейронных сетей.

Оптимизация скорости обучения. Цель метода ОСО – найти оптимальную скорость обучения для заданных в данный момент параметров модели и градиентов. Таким образом, ОСО является подмодулем градиентных методов для адаптации скорости онлайн-обучения. Вкратце, градиентные методы с ОСО оптимизируют параметры модели в три этапа:

1. Расчет градиентов для целевых и текущих параметров модели.
2. Поиск оптимальной скорости обучения для заданных параметров модели и градиентов.
3. Обновление параметров модели, используя градиенты и оптимизированную скорость обучения.

Цель ОСО – найти параметр η^* – оптимальную скорость обучения, который минимизирует функцию потерь для заданных текущих параметров модели и градиентов как:

$$\eta^* = \arg\min_\eta \left\{ \mathrm{I}(\theta - \eta\nu) + \Omega(\theta - \nu\eta) \right\}, \tag{1}$$

где I – функция потерь обучения, Ω – срок регуляризации θ – параметры модели, ν – адаптивный градиент для обновления θ. Поскольку методы градиента обновляют параметры модели, перемещаясь в направлении, противоположном градиенту $(\theta \leftarrow \theta \leftarrow \eta\nu)$, функция потерь $\mathrm{I}(\theta)$ и срок регуляризации $\Omega(\theta)$ можно переписать как $\mathrm{I}(\theta - \eta\nu)$ и $\Omega(\theta - \eta\nu)$, соответственно.

Однако в реальных задачах непосредственно решать задачу оптимизации в уравнении (1) невозможно из-за сильной нелинейности I. Чтобы справиться с этой трудностью, сначала линеаризуется I вблизи θ в виде:

$$\mathrm{I}(\theta - \eta\nu) \approx \mathrm{I}(\theta) + (\nabla_\theta \mathrm{I})^T (\theta - \eta\nu - \theta) = \mathrm{I}(\theta) - \eta g^T \nu, \tag{2}$$

где $g = \nabla_\theta \mathrm{I}$ является истинным градиентом. Однако для линеаризации уравнения (2) η должен быть достаточно мал. Для этого ограничения мы вводим ограничение неравенства для верхней границы η [3]. Таким образом, задача оптимизации ОСО окончательно определяется следующим образом:

$$\eta^* = \arg\min_{0 \le \eta \le \varepsilon} \left\{ \mathrm{I}(\theta) - \eta g^T \nu + \Omega(\theta - \eta\nu) \right\}, \tag{3}$$

где ε – положительный гиперпараметр, определяющий верхнюю границу скорости обучения. Таким образом, ОСО адаптирует скорость обучения, решая задачу оптимизации с ограничениями в уравнении (3) для текущих параметров модели и градиентов.

Расширенный метод Лагранжа для задачи оптимизации скорости обучения. Расширенный метод Лагранжа – это широко используемый метод оптимизации для решения задачи оптимизации с ограничениями путем преобразования ее в задачу без ограничений [4]. Однако задача оптимизации расширенного метода Лагранжа должна быть определена при ограничениях равенствах. По этой причине необходимо преобразовать задачу оптимизации с ограничениями неравенствами в уравнении (3) к задаче с ограничениями – равенствами. Для этого введем переменную $s = \left[s_1, s_2 \right]^T$, затем преобразуем проблему в проблему с ограничениями равенствами как:

$$\eta^* = \arg\min_{0 \le \eta \le \varepsilon} \left\{ \mathrm{I}(\theta) - \eta g^T \nu + \Omega(\theta - \eta\nu) \right\}, \tag{4}$$

$$\eta - s_1 = 0, \varepsilon - \eta - s_2 = 0, \qquad s_1 \ge 0, s_2 \ge 0,$$

где s_1 и s_2 являются слабыми переменными с неотрицательным ограничением.

Наконец, расширенная функция Лагранжа для задачи определяется как:

$$\mathrm{L}_{\lambda}(\eta, s, \lambda) = \mathrm{I}(\theta) - \eta g^{T} v + \Omega(\theta - \eta v) -$$

$$-\lambda_1(\eta - \mathrm{s}_1) - \lambda_2(\varepsilon + \eta - s_2) + \frac{\mu}{2}(\eta - s_1)^2 + \frac{\mu}{2}(\varepsilon - \eta - s_2)^2, \quad (5)$$

где $\lambda = [\lambda_1, \lambda_2]^T$ является двойственной переменной для расширенной функции Лагранжа и $\mu \geq 0$ является гиперпараметром метода множителей переменных направлений (ММПН) для управления балансом целевой функции и штрафного члена для ограничений равенств. В целом, μ просто настроен на постепенное увеличение в процессе оптимизации, чтобы гарантировать выполнимость решения для ограничений равенств [5].

Оптимизация. Оптимизация ОСО реализуется на основе ММПН, чтобы найти оптимальную скорость обучения, которая минимизирует расширенную функцию Лагранжа в уравнении (5). Основанное на ММПН решение для ОСО имеет два преимущества:

1. Оно эффективно, поскольку правила обновления получаются на основе простых скалярных вычислений.

2. Легко расширяется до нескольких ограничений и скоростей обучения, адаптированных для каждого параметра модели.

В математической оптимизации ММПН широко используется для решения задачи оптимизации [6], содержащей различные типы первичных переменных x и z с ограничениями равенствами:

$$x, z = \arg\min_{x,z} f(x) + g(z) \quad (6)$$

$$Ax + Bz = c,$$

где A и B матрицы коэффициентов ограничений равенств, c является константой. ММПН итеративно находит оптимальные значения основных переменных путем минимизации расширенной функции Лагранжа $L_{\mu}(x, z, \lambda)$ для задачи (6). В частности, ММПН оптимизирует простые и двойственные переменные по методу Гаусса-Зейделя [6], за один проход $L_{\mu}(x, z, \lambda)$ минимизируется по отношению к x и z альтернативно для фиксированной переменной λ. Затем, $L_{\mu}(x, z, \lambda)$ сводится к минимуму по λ для фиксированных x и z.

Задача оптимизации скорости обучения уравнения (4) представляет собой проблему с ограничениями равенствами, а также имеет два типа основных переменных: скорость обучения и переменная резерва [7]. То есть, для основных переменных η и s задача оптимизации скорости обучения имеет ту же структуру, что и задача (6), $f(\eta) = I(\theta) - \eta g^{T} v + \Omega(\theta - \eta v), g(s) = 0$ для $s = [s_1, s_2]^T$, и ограничения равенства $\eta - s_1 = 0$ и $\varepsilon - \eta - s_2 = 0$. Таким образом, основные и двойные переменные задачи оптимизации скорости обучения могут быть оптимизированы с помощью ММПН [8] следующим образом:

$$\eta^{(t+1)} \leftarrow \arg\min_{\eta} \mathrm{L}_{\mu}(\eta, s^{(t)}, \lambda^{(t)}), \quad (7)$$

$$s^{(t+1)} \leftarrow [\arg\min_{s} L_{\mu}(\eta^{(t+1)}, s, \lambda^{(t)})]_{+}, \quad (8)$$

$$\lambda_1^{(t+1)} \leftarrow \lambda_1^{(t)} - \mu(\eta^{(t+1)} - s_1^{(t+1)}), \quad (9)$$

$$\lambda_2^{(t+1)} \leftarrow \lambda_2^{(t)} - \mu(\varepsilon - \eta^{(t+1)} - s_2^{(t+1)}), \quad (10)$$

где $[\ ]_{+}$ является операцией поэлементной максимизации, где максимизация применяется к уравнению (8) для удовлетворения неотрицательного ограничения переменной резерва [9]. В этом процессе оптимизации, поскольку переменная резерва не зависит от функции потерь и члена регуляризации в задаче оптимизации скорости обучения, правила обновления переменной резерва могут быть указаны как [9]:

$$s_1^{(t+1)} = [\eta^{(t+1)} - \frac{\lambda_1^{(t)}}{\mu}]_+, \quad (11)$$

$$s_2^{(t+1)} = [\varepsilon - \eta^{(t+1)} - \frac{\lambda_2^{(t)}}{\mu}]_+. \quad (12)$$

Сходимость ОСО зависит от члена регуляризации $\Omega(\theta - \eta\nu)$. В частности, поскольку условия потерь уже выпуклы относительно η, сходимость ОСО гарантируется, когда $\Omega(\theta - \eta\nu)$ выпукла. Однако, в целом член регуляризации определяется как выпуклая функция или намеренно сведен к выпуклой функции в машинном обучении [10]. Таким образом, сходимость ОСО гарантируется в общих случаях машинного обучения.

Литература

1. Byrd, R. H. A stochastic quasi-Newton method for large-scale optimization / R. H. Byrd, S. L. Hansen, J. Nosedal, Y. A. Singer // SIAM Journal on Optimization. – 2016. – № 26 (2). – P. 1008–1031.

2. Agarwal, A. Noisy matrix decomposition via convex relaxation : Optimal rates in high dimensions / A. Agarwal, S. N. Negahban, M. J. Wainwright // The Annals of Statistics. – 2012. – № 40.2. – P. 1171–1197.

3. Chen, C. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent / C. Chen, B. He, Y. Ye, X. Yuan // Mathematical Programming's. – 2016. – Vol. 155, № 1–2. – P. 57–79.

4. Duchi, J. Adaptive subgradient methods for online learning and stochastic optimization / J. Duchi, E. Hazan, Y. Singer // Journal of Machine Learning Research. – 2011. – № 12 (61). – P. 2121–2159.

5. Fadil, S. Linear inversion of band-limited reflection seismograms / S. Fadil // SIAM Journal on Scientific and Statistical Computing. – 1986. – Vol. 7, № 4. – P. 1307–1330.

6. Gabay, D. A dual algorithm for the solution of nonlinear variational problems via finite element approximation / D. Gabay, B. Mercier // Computes and Mathematics with Applications. – 1976. – № 2. – P. 17–40.

7. Hestenes, M. R. Multiplier and gradient methods / M. R. Hestenes // Journal of Optimization Theory and Applications. – 1969. – № 4. – P. 303–320.

8. LeCun ,Y. Gradient-based learning applied to document recognition / Y. LeCun, L. Bottou, Y. Bengio, P. Haffner // Proceedings of the IEEE. –1998. – № 86 (11). – P. 2278–2324.

9. Lin, T. Global convergence of unmodified 3-block ADMM for a class of convex minimization problems / T. Lin, S. Ma, S. Zhang // Journal of Scientific Computing. – 2017. – № 76 (1). – P. 69–88.

10. Schaul, T. No more learning rates / T. Schaul, S. Zhang, Y. LeCun // Proceedings of machine learning research. – 2013. – Vol. 28. – P. 343–351.

Гомельский государственный
университет имени Франциска Скорины

Поступила в редакцию 08.04.2022