Методы с предобуславливанием и затуханием весов

Матвей Вадимович Крейнин

Московский Физико-Технический Институт

Кафедра: Интеллектуальный анализ данных

Научный руководитель: кандидат ф.-м. наук А. Н. Безносиков

2024

Постановка задачи

Минимизация функции:

$$\min_{w \in \mathbb{R}^d} f(w) \tag{1}$$

Классическое решение проблемы минимизации функции в машинном обучении:

$$w_t = w_{t-1} - \eta \nabla f(w_t).$$

Алгоритмы с предобуславливанием:

$$w_{t+1} = w_t - \eta D_t^{-1} g_t$$

где g_t это несмещённый стохастический градиент, D_t это матрица предобуславливания.

Различные способы задания матрицы

AdaGrad:

$$D_t = diag\left\{\sqrt{\sum_{i=0}^{t} g_i \odot g_i}\right\}$$

RMSProp and Adam:

$$D_t^2 = \beta D_{t-1}^2 + (1-\beta) diag\{g_t \odot g_t\}$$

OASIS:

$$D_t = diag\{z \odot \nabla^2 f(w_t) z\}$$

где z это случайный вектор из распределения Рандемахера.

Новая минимизируемая функция

$$\min_{w \in \mathbb{R}^d} F(w) := f(w) + r(w)$$

где $r(w)$ это функция регуляризации.

Algorithm 1 Различные способы использования предобуславливания с регуляризацией

Require: η – шаг обучения, f – оптимзируемая функция

while w не сойдется **do**

$t = t + 1$

$g_t \leftarrow$ стохастический градиент f

$g_t \leftarrow g_t + \nabla r(w_t)$ обычная регуляризация

$D_t \leftarrow$ матрица предобуславливания с помощью g_t

$w_t \leftarrow w_{t-1} - \eta \cdot D_t^{-1} g_t$ обычная регуляризация,

$w_t \leftarrow w_{t-1} - \eta \cdot D_t^{-1} (g_t + \nabla r(w_t))$ масштабированное затухание весов,

$w_t \leftarrow w_{t-1} - \eta \cdot D_t^{-1} g_t - \eta \cdot \nabla r(w_t)$ затухание весов,

end while

Альтернативный взгляд

Вынесем D_t^{-1} за скобки и получим новый градиент:

$$w_{t+1} = w_t - \eta D_t^{-1}(\nabla f(w_t) + D_t \nabla r(w_t))$$

Новая функция регуляризации $\nabla \tilde{r}(w) = D_t \nabla r(w)$.
Задача минимизации, которая решается на самом деле:

$$\min_{w \in \mathbb{R}^d} \tilde{F}(w) := f(w) + \tilde{r}(w)$$

где $\tilde{F}(w)$ изменяется на каждом шаге.

Предположения на функции

Предположение

(Структура регулязитора) Регуляризатор r сепарабелен, то есть он может быть представлен в следующем виде: $r(w) = \sum_{i=1}^{d} r_i(w^i)$, где $r_i(x) \geq 0$ для $i \in \overline{1, d}$ и $x \in \mathbb{R}$.

Предположение

(Структура матрицы предобуславливания) Матрица предобуславливания D_t может быть представлена в следующем виде: $D_t = diag\left\{d_t^1 \ldots, d_t^d\right\}$.

Предположение

(Сильная выпуклость) Сушествует μ_f такая, что $\forall x, y \in \mathbb{R}^d$ выполняется:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu_f}{2} ||x - y||_2^2$$

Предположения на функции

Предположение

(L-гладкость)

- *Градиент функции f является L_f-гладким, то есть существует такая константа $L_f > 0$ такая, что $\forall x, y \in \mathbb{R}^d$,*

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L_f}{2} \|x - y\|^2.$$

- *Градиент функции r является L_r-гладким, то есть существует такая константа $L_r > 0$ такая, что $\forall x, y \in \mathbb{R}^d$,*

$$r(x) \leq r(y) + \langle \nabla r(y), x - y \rangle + \frac{L_r}{2} ||x - y||^2.$$

Предположения на функции

Предположение

(Ограниченность регуляризатора) Регуляризатор ограничен, то есть существует константа $\Omega > 0$ *такая, что* $\forall w \in \mathbb{R}^d$ *выполняется* $|r(w)| \leq \Omega$.

Предположение

(Ограниченность предобуславливателя) Существуют константы $\alpha, \Gamma \in \mathbb{R} : 0 < \alpha < \Gamma$ *такие, что*

$$\alpha I \preccurlyeq D_t \preccurlyeq \Gamma I \Leftrightarrow \frac{I}{\Gamma} \preccurlyeq D_t^{-1} \preccurlyeq \frac{I}{\alpha}.$$

Предположение

(Ожидания) g_t *являются несмещенными и имеют ограниченную вариацию на любом шаге, то есть*

$$\mathbb{E}\left[g_t\right] = \nabla f(w_t), \mathbb{E}\left[||g_t - \nabla f||^2\right] \leq \sigma^2.$$

Леммы

Lemma

(Существование $\widetilde{r}$) Предполагая, что 1, 2 выполняются, функция $\widetilde{r}$ существует и имеет следующую форму:

$$\widetilde{r}_t(w) = \sum_{i=1}^{d} d_t^i r_i(w_i)$$

Lemma

(L-гладкость $\widetilde{r}$) Предполагая, что 1, 2, 4, 6 выполняются, градиент $\widetilde{r}$ является $L_{\tilde{r}}$-непрерывным, то есть существует константа $L_{\tilde{r}} > 0$ такая, что $\forall x, y \in \mathbb{R}^d$,

$$\widetilde{r}_t(x) \leq \widetilde{r}_t(y) + \langle \nabla \widetilde{r}_t(y), x - y \rangle + \frac{L_{\tilde{r}}}{2} ||x - y||^2,$$

и $L_{\tilde{r}} = \Gamma L_r$.

Теоремы

Theorem

Предполагая, что 1, 2, 4, 5, 6 выполняются, положим ошибку $\varepsilon > 0$ *и шаг обучения удовлетворяют условию:* $\eta < \frac{2\alpha}{L_f + \Gamma L_r}$, *где* L_f, L_r *- константа Липшица функций* f *и* r*. Необходимое количество итераций, начиная с начальной точки* $w_0 \in \mathbb{R}^d$ *с* $\Delta_0 = \tilde{F}_0(w_0) - f^*$, *где* $\widetilde{F}_t$ *определено в* (5) *и* f^* *решением задачи* (1), *необходимое для* ε*-приближения*

$$T = \mathcal{O}\left(\frac{\Delta_0 \Gamma}{\left(\eta - \frac{\tilde{L}\eta^2}{2\alpha}\right)\left(\varepsilon - \frac{\delta\Gamma}{\eta - \frac{\tilde{L}\eta^2}{2\alpha}}\right)}\right),$$

где $\widetilde{L} = L_f + \Gamma L_r$ *and* δ *может выбрано сколь угодно малым с помощью выбора гиперпараметров* α, β, Γ

Теоремы

Theorem

Преполагая, что 1, 2, 4, 5, 6, 3 выполняются, положим ошибку $\varepsilon > 0$ *и шаг обучения удовлетворяют условию:* $\eta < \frac{\alpha}{4L_f}$*, гиперпараметры удовлетворяют условиям:* $\lambda < \frac{\alpha\beta^2}{8L_f\Omega_0^2}$*,* $\beta \geq 1 - \frac{\eta(\mu_f+\lambda)\alpha}{2\Gamma^2}$*,* $\Omega_0^2 \geq \frac{\alpha^2\beta^2}{8L_f^2}$*. Получаем оценку на необходимое количество шагов для сходимости алгоритма к заданной точности*

$$T = \max\left\{\log\left(\frac{4\varepsilon(1-\beta)L_f^2}{\alpha\beta^2}\right)\cdot\log\frac{2}{1+\beta};\ \frac{4}{\frac{\alpha}{4L_f}\left(\mu_f + \frac{\alpha\beta^2}{8L_f\Omega_0^2}\right)}\log\left(\frac{2||w_0 - w_*||_2^2}{\varepsilon}\right)\right\}$$

AdamW

Algorithm 2 Adam

Require: $\eta, \beta_1, \beta_2, \epsilon, f, r$

while θ not converged **do**

$t = t + 1$

$g_t = \nabla f(w_{t-1}) + \nabla r(w_{t-1})$ AdamL2

$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$

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

$\hat{m}_t = \frac{m_t}{1-\beta_1^t} + \nabla r(w_{t-1})$ AdamWH

$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$

$w_t = w_{t-1} - \eta \cdot \frac{\hat{m}_t}{\sqrt{v_t} + \epsilon} - \eta \nabla r(w_{t-1})$ AdamW

end while

OASIS

Algorithm 3 OASIS

Require: $w_0, \eta_0, D_0, \theta_0 = +\infty$

$w_1 = w_0 - \eta \hat{D}_0^{-1} \nabla f(w_0)$

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

$g_k = \nabla f(w_k) + \nabla r(w_{t-1})$ OASISL2

$D_k = \beta D_{k-1} + (1 - \beta_2) \cdot diag\left(z_k \odot \nabla^2 \left(f(w_k) + r(w_k)\right) z_k\right)$ OASISWH

$(\hat{D}_k)_{ii} = max\{|D_k|_{i,i}; \alpha\}, \ \forall i = \overline{1, d}$

$\eta_k = min\{\sqrt{1 + \theta_{k-1}} \cdot \eta_{k-1}; \frac{||w_k - w_{k-1}||_{\hat{D}_k}}{2||\nabla f(w_k) - \nabla f(w_{k-1})||^*_{\hat{D}_k}}\}$

$w_{k+1} = w_k - \eta_k g_k D_k^{-1} - \eta \nabla r(w_{t-1})$ OASISW

$\theta_k = \frac{\eta_k}{\eta_{k-1}}$

end for

Experiments

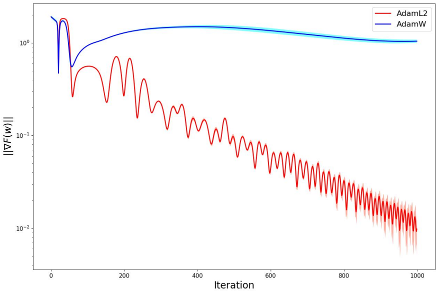

Figure: Adam and AdamW with basic criterion

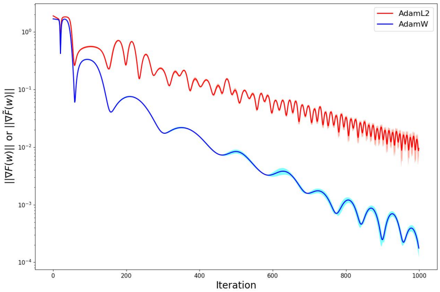

Figure: Adam and AdamW with modified criterion

Список литературы

- Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).
- Jahani, Majid, et al. "Doubly adaptive scaled algorithm for machine learning using second-order information." arXiv preprint arXiv:2109.05198 (2021).
- Sadiev, Abdurakhmon, et al. "Stochastic gradient methods with preconditioned updates." arXiv preprint arXiv:2206.00285 (2022).
- Beznosikov, Aleksandr, et al. "On scaled methods for saddle point problems." arXiv preprint arXiv:2206.08303 (2022).
- Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." arXiv preprint arXiv:1711.05101 (2017).
- Xie, Zeke, Issei Sato, and Masashi Sugiyama. "Stable weight decay regularization." (2020).

На защиту выносится

- Исследована теоретическая сходимость методов.
- Предложен новый метод добавления регуляризатора в методы с предобуславливанием.
- Две доказанные теоремы и леммы при различных предположениях на оптимизируемую функцию.