Методы с предобуславливанием и затуханием весов

Выпускная квалификационная работа бакалавра

Матвей Вадимович Крейнин
Научный руководитель: к.ф.-м.н. А. Н. Безносиков

Кафедра интеллектуальных систем ФПМИ МФТИ
Специализация: Интеллектуальный анализ данных
Направление: 03.03.01 Прикладные математика и физика

2024

Методы с предобуславливанием и затуханием весов

Исследуется сходимость методов оптимизации с предобуславливанием и затуханием весов.

Проблема

Исследование скорости сходимости методов оптимизации с предобуславливанием.

Цель

Оценить скорость сходимости данных методов и предложить альтернативный вариант добавления регуляризации.

Постановка задачи

Классическое решение

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

Метод градиентного спуска

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

Метод Ньютона:

$$w_t = w_{t\_1} - \eta \left(\nabla^2 f(w_{t-1})\right)^{-1} \nabla f(w_{t-1})$$

Метод стохастического градиента

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

g_t - несмещённый стохастический градиент

Методы с предобуславливанием

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

D_t – матрица предобуславливания.

Новая минимизируемая функция

$$\min_{w \in \mathbb{R}^d} F(w) := f(w) + r(w), r(w) \text{ это функция регуляризации.}$$

Algorithm 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 непосредственно:

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

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

Предположения на функции

Предположение (Структура регулязитора)

Регуляризатор r сепарабелен:

$$r(w) = \sum_{i=1}^{d} r_i(w^i),$$

где $r_i(x) \geq 0$ для $i \in \overline{1, d}$ и $x \in \mathbb{R}$.

Предположение (Сильная выпуклость)

$\exists \mu_f : \forall x, y \in \mathbb{R}^d$ *выполняется:*

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

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 (Ограниченность предобуславливателя)

$\exists\ \alpha, \Gamma \in \mathbb{R} : 0 < \alpha < \Gamma$*:*

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

Предположения на функции

Предположение (L-гладкость)

Градиент функции f является L_f-гладким, $\exists\ L_f > 0 : \forall x, y \in \mathbb{R}^d$,

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

Предположение (L-гладкость)

Градиент функции r является L_r-гладким, $\exists\ L_r > 0 : \forall x, y \in \mathbb{R}^d$,

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

Предположение (Ограниченность регуляризатора)

Регуляризатор ограничен, $\exists\ \Omega > 0 : \forall w \in \mathbb{R}^d$ выполняется $|r(w)| \leq \Omega$.

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

Лемма (Существование $\widetilde{r}$)

Предполагая, что 1, 3 выполняются, функция $\widetilde{r}$ *существует и имеет следующую форму:*

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

Лемма

(L-гладкость $\widetilde{r}$*) Предполагая, что 1, 3, 5, 4 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

Теорема

Предполагая, что 1, 3, 5, 7, 4 выполняются, положим ошибку $\varepsilon > 0$ *и шаг обучения удовлетворяют условию:* $\eta < \frac{2\alpha}{L_f + \Gamma L_r}$. *Количество итераций, начиная с начальной точки* $w_0 \in \mathbb{R}^d$, $\Delta_0 = \tilde{F}_0(w_0) - f^*$, *где* $\widetilde{F}_t$ *определено в* (5) *и* f^* *решением задачи* (1), *необходимое для* ε*-приближения нормы градиента к 0, может быть ограничено количеством шагов*

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

$\widetilde{L} = L_f + \Gamma L_r$ *и* $\delta = \frac{(1-\beta)\Gamma^2}{2\alpha}$ *выбирается сколь угодно благодаря* α, β, Γ.

Теоремы

Теорема

Преполагая, что 1, 3, 5, 7, 4, 2 выполняются, положим ошибку $\varepsilon > 0$ *и шаг обучения удовлетворяют условию:*

$$\eta < \eta_{min} = \min\left\{\frac{2L_f\Omega_0^2}{\alpha\beta^2}; \frac{\alpha}{4L_f}; \frac{8\mu_f L_f^2\Omega_0^4}{\alpha^2\beta^4} + \frac{L_f\Omega_0^2}{\alpha\beta^2}\right\},$$

гиперпараметры удовлетворяют условиям: $\lambda < \frac{\alpha\beta^2}{8L_f\Omega_0^2}$, $\beta \geq 1 - \frac{\eta(\mu_f+\lambda)\alpha}{2\Gamma^2}$.
Получаем оценку на необходимое количество шагов для сходимости алгоритма к заданной точности:

$$T = \mathcal{O}\left(\log\left(\frac{R_0^2 + \frac{8\lambda\Omega_0^2\Gamma^2}{\alpha^2(\mu_f+\lambda)}\sigma_0^2}{\varepsilon}\right)\frac{4}{\eta_{min}(\mu_f+\lambda)\cdot\min\left\{1; \frac{2\alpha}{\Gamma^2}\right\}}\right)$$

AdamW

Algorithm 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

Эксперименты

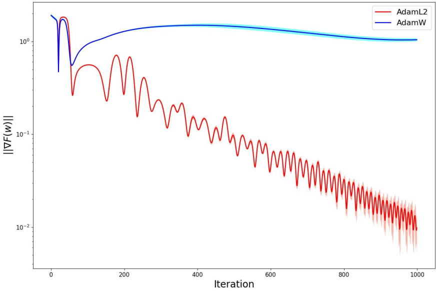

Рис.: Adam and AdamW по критерию, $||\nabla F(w)|| = ||\nabla f(w) + \nabla r(w)||$

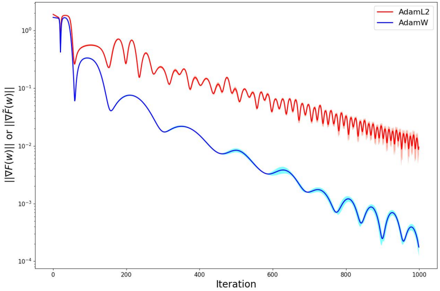

Рис.: Adam and AdamW по критерию, $||\nabla \tilde{F}(w)|| = ||\nabla f(w) + \nabla \tilde{r}(w)||$

На защиту выносится

1. Исследована теоретическая сходимость методов.
2. Новый метод добавления регуляризатора в методы с предобуславливанием.
3. Две леммы о существовании, структуре и гладкости измененного регуляризатора.
4. Теорема о сходимости методов с предобуславливанием и затуханием весов по норме градиента.
5. Теорема о сходимости методов с предобуславливанием и затуханием весов по аргументу.