алгоритмы оптимизации с адаптивным шагом и моментами второго порядка

Методы оптимизации в задаче предискажения сигнала

Масловский Александр

12 августа 2020

Постановка задачи

постановка исходной задачи: $\min_{W \in \mathbb{C}} f(x)$, где f – дифференцируемая функция типа :

$$\sum_{i=0}^{K} \frac{1}{K} (y_i - g(x_i))^2$$

где g(x,W) x, $W \in \mathbb{C}$ - функция предыскажения сигнала

Adagrad

$$g_k = \nabla^r f(x^k, \{\xi_l^{k+1}\}_{l=1}^r)$$

$$x_{k+1} = x_k - \frac{h}{\sqrt{G_k + \epsilon}} \odot g_k$$

Вход: learning rate $h > 0$, starting point $x^0 \in \mathbb{R}^n$, parameter $b_0 > 0$

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

Sample $\nabla f(x^k, \{\xi_i\}_{i=1}^r)$

$b_k^2 = b_{k-1}^2 + \|\nabla f(x^k, \{\xi_i\}_{i=1}^r)\|^2$

$x^{k+1} = x^k - \frac{h}{b_k} \nabla f(x^k, \{\xi_i\}_{i=1}^r)$

end for=0

RMSProp

$$g_k = \nabla^r f(x^k, \{\xi_l^{k+1}\}_{l=1}^r)$$

$$G_{k+1} = \gamma G_k + (1-\gamma) g_k^2 \text{ -exponentioal mean(smoothing)}$$

$$x_{k+1} = x_k - \frac{h}{\sqrt{G_{k+1} + \epsilon}} \odot g_k$$

Вход: learning rate $h, \gamma > 0$, starting point $x^0 \in \mathbb{R}^n$, parameter $b_0 > 0$

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

Sample $\nabla f(x^k, \{\xi_i\}_{i=1}^r)$

$b_k^2 = \gamma b_{k-1}^2 + (1-\gamma)\|\nabla f(x^k, \{\xi_i\}_{i=1}^r)\|^2$

$x^{k+1} = x^k - \frac{h}{\sqrt{b_k + \epsilon}} \nabla f(x^k, \{\xi_i\}_{i=1}^r)$

end for=0

Adadelta

$$g_k = \nabla^r f(x^k, \{\xi_l^{k+1}\}_{l=1}^r)$$

$$G_{k+1} = \gamma_1 G_k + (1-\gamma_1) g_k^2\text{-(smoothing)}$$

$$\delta x_k = x_k - x_{k-1}$$

$$\delta x = \gamma_2 \delta x + (1-\gamma_2)\delta x_k^2 \text{ -(smoothing)}$$

$$x_{k+1} = x_k - \frac{\delta x}{\sqrt{G_{k+1} + \epsilon}} \odot g_k$$

ADAM(Adaptive momdent estimation)

Вход: learning rate $\alpha > 0$, momentum terms $\beta_1 > 0$, $\beta_2 > 0$, $\lambda \in \mathbb{R}$, parameter vector $\theta_0 \in \mathbb{R}^n$, vectors $m_0, v_0 \in^n$.

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

Sample $\nabla f(x^k, \{\xi_i\}_{i=1}^r)$

$m_k = \beta_1 m_{k-1} + (1 - \beta_1) \nabla f(x^k, \{\xi_i\}_{i=1}^r)$

$v_k = \beta_2 v_{k-1} + (1 - \beta_2) \nabla f(x^k, \{\xi_i\}_{i=1}^r)^2$

$\hat{m}_k = \frac{m_k}{1-\beta_1^k}$

$\hat{v}_k = \frac{v_k}{1-\beta_2^k}$

$\theta_k = \theta_{k-1} - (\frac{\alpha}{\sqrt{\hat{v}_k + \epsilon}} \hat{m}_k)$

end for=0

Adam with decoupled weight decay

Вход: learning rate $\alpha > 0$, momentum terms $\beta_1 > 0$, $\beta_2 > 0$, $\lambda \in \mathbb{R}$, parameter vector $\theta_0 \in \mathbb{R}^n$, vectors $m_0, v_0 \in^n$.

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

Sample $\nabla f(x^k, \{\xi_i\}_{i=1}^r)$

$m_k = \beta_1 m_{k-1} + (1 - \beta_1) \nabla f(x^k, \{\xi_i\}_{i=1}^r)$

$v_k = \beta_2 v_{k-1} + (1 - \beta_2) \nabla f(x^k, \{\xi_i\}_{i=1}^r)^2$

$\hat{m}_k = \frac{m_k}{1-\beta_1^k}$

$\hat{v}_k = \frac{v_k}{1-\beta_2^k}$

$\theta_k = \theta_{k-1} - (\frac{\alpha}{\sqrt{\hat{v}_k + \epsilon}} \hat{m}_k + \lambda \theta_{k-1})$

end for=0

AMSGrad

Вход: learning rate $\alpha > 0$, momentum terms $\beta_1 > 0$, $\beta_2 > 0$, $\lambda \in \mathbb{R}$, parameter vector $\theta_0 \in \mathbb{R}^n$, vectors $m_0, v_0 \in^n$.

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

Sample $\nabla f(x^k, \{\xi_i\}_{i=1}^r)$

$m_k = \beta_1 m_{k-1} + (1 - \beta_1) \nabla f(x^k, \{\xi_i\}_{i=1}^r)$

$v_k = \beta_2 v_{k-1} + (1 - \beta_2) \nabla f(x^k, \{\xi_i\}_{i=1}^r)^2$

$\hat{v}_k = \max(\hat{v}_{k-1}, v_k)$

$\theta_k = \theta_{k-1} - \frac{\alpha}{\sqrt{\hat{v}_k + \epsilon}} m_k$

end for=0

QHAdam(QUASI-Hyberbolic Adaptive momdent estimation)

Вход: learning rate $\alpha > 0$, momentum terms $\beta_1, \beta_2 > 0$, $\gamma_1, \gamma_2 > 0$, $\lambda \in \mathbb{R}$, parameter vector $\theta_0 \in \mathbb{R}^n$, vectors $m_0, v_0 \in^n$.

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

$g_k = \nabla f(x^k, \{\xi_i\}_{i=1}^r)$

$m_k = \beta_1 m_{k-1} + (1 - \beta_1)\, g_k$

$v_k = \beta_2 v_{k-1} + (1 - \beta_2)\, g_k^2$

$\theta_k = \theta_{k-1} - \alpha\left(\frac{(1-\gamma_1)g_k + \gamma_1 \cdot m_k}{\sqrt{(1-\gamma_1)g_k^2 + \gamma_2 \cdot v_k} + \epsilon}\right)$

end for=0

алгоритмы с адаптивным шагом

Алгоритм адаптивного градиентного спуска Малитского-Мищенко(2019)

1: $x^0 \in \mathbb{R}^d$, $\lambda_0 > 0$, $\theta_0 = +\infty$, $x^1 = x^0 - \lambda_0 \nabla f(x^0)$
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad \lambda_k = \min\left\{\sqrt{1+\theta_{k-1}}\lambda_{k-1}, \frac{\|x^k - x^{k-1}\|}{2\|\nabla f(x^k) - \nabla f(x^{k-1})\|}\right\}$
4: $\quad x^{k+1} = x^k - \lambda_k \nabla f(x^k)$
5: $\quad \theta_k = \frac{\lambda_k}{\lambda_{k-1}}$
6: **end for**=0

алгоритмы с адаптивным шагом

Алгоритм адаптивного градиентного спуска Малитского-Мищенко(2019) для SGD с моментным членом

$x^0 \in \mathbb{R}^d$, $\lambda_0 > 0$, $\Lambda_0 > 0$, $\theta_0 = \Theta_0 = +\infty$, $y^1 = x^1 = x^0 - \lambda_0 \nabla f(x^0)$

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

$\lambda_k = \min\left\{\sqrt{1+\theta_{k-1}}\lambda_{k-1}, \frac{\|x^k - x^{k-1}\|}{2\|\nabla f(x^k) - \nabla f(x^{k-1})\|}\right\}$

$\Lambda_k = \min\left\{\sqrt{1+\Theta_{k-1}}\Lambda_{k-1}, \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{2\|x^k - x^{k-1}\|}\right\}$

$\beta_k = \frac{\sqrt{1/\lambda_k} - \sqrt{\Lambda_k}}{\sqrt{1/\lambda_k} + \sqrt{\Lambda_k}}$

$y^{k+1} = x^k - \lambda_k \nabla f(x^k)$

$x^{k+1} = y^{k+1} + \beta_k (y^{k+1} - y^k)$

$\theta_k = \frac{\lambda_k}{\lambda_{k-1}}$, $\Theta_k = \frac{\Lambda_k}{\Lambda_{k-1}}$

end for=0