Follow The Moving Leader

Intro to FTML

By henryphzhao(赵鹏昊), 2017.10

“Follow the Moving Leader in Deep Learning”, ICML 2017
Shuai Zheng, James T. Kwok
http://proceedings.mlr.press/v70/zheng17a/zheng17a.pdf

简介

- FTML是FTRL(Follow the regularized leader)的变种
- FTRL 平等地对待所有样本，但是FTML更看重最近的样本
- RMSprop和Adam可以看作是FTML的特例
- FTML集合了RMSprop和Adam的优点，而且克服了它们的缺点

Notation

Notation. For a vector $x \in \mathbb{R}^d$, $\|x\| = \sqrt{\sum_{i=1}^d x_i^2}$, $\mathrm{diag}(x)$ is a diagonal matrix with x on its diagonal, $\sqrt{x}$ is the element-wise square root of x, x^2 denotes the Hadamard (elementwise) product $x \odot x$, and $\|x\|_Q^2 = x^T Q x$, where Q is a symmetric matrix. For any two vectors x and y, x/y, and $\langle x, y \rangle$ denote the elementwise division and dot product, respectively. For a matrix X, $X^2 = XX$, and $\mathrm{diag}(X)$ is a vector with the diagonal of X as its elements. For t vectors $\{x_1, \ldots, x_t\}$, $x_{1:t} = \sum_{i=1}^t x_i$, and $x_{1:t}^2 = \sum_{i=1}^t x_i^2$. For t matrices $\{X_1, \ldots, X_t\}$, $X_{1:t} = \sum_{i=1}^t X_i$.

FTRL简介

- FTRL: the regularization is centered at the origin

At round t, FTRL generates the next iterate θ_t by solving the optimization problem:

α_t 是regularization的系数，也决定了学习率

$$\theta_t = \arg\min_{\theta\in\Theta} \sum_{i=1}^{t} \left(\langle g_i, \theta \rangle + \frac{\alpha_t}{2} \|\theta\|^2 \right),$$

- FTPRL(P for Proximal): centering regularization at each iterate θ_{i-1}

$$\theta_t = \arg\min_{\theta\in\Theta} \sum_{i=1}^{t} \left(\langle g_i, \theta \rangle + \frac{1}{2} \|\theta - \theta_{i-1}\|_{Q_i}^2 \right), \qquad (1)$$

- FTRL有封闭解：

$$\theta_t = \theta_{t-1} - Q_{1:t}^{-1} g_t. \qquad (2)$$

- Per-coordinate learning rate controlled by diagonal matrix Qi

$$Q_{1:t} = \mathrm{diag}\left(\frac{1}{\eta} \left(\sqrt{g_{1:t}^2} + \epsilon \mathbf{1} \right) \right). \qquad (5)$$

Expotential Moving Average

- $v_t = \beta v_{t-1} + (1-\beta)\theta_t$
- $v_t = (1-\beta)\theta_t + (1-\beta)\beta\theta_{t-1} + (1-\beta)\beta^2\theta_{t-2} + \cdots + (1-\beta)\beta^k\theta_{t-k} + \cdots$
- Beta = 0.999 相当于取最近1/(1-0.999)=1000个样本

这是因为 $(1-\varepsilon)^{1/\varepsilon} = \frac{1}{e} \approx 0.35$

v_0= 0， 故需要偏差修正

$$E[\theta_i] = A$$

$$E[(1-\beta)\beta^k\theta_{t-k}] = (1-\beta)\beta^k A$$

$$E[v_t] = \sum_{i=0}^{t}(1-\beta)\beta^i A = (1-\beta^t)A$$

$$E\left[\frac{v_t}{1-\beta^t}\right] = A$$

- 性质 $w_{i,t} = \frac{(1-\beta_1)\beta_1^{t-i}}{1-\beta_1^t}$ **Lemma 1.** $\lim_{\beta_1 \to 1} w_{i,t} = 1/t.$

FTML: 对样本(的loss)做加权平均

- FTRL: each sample's loss (Pi) has the same weight

Recall that at round t, FTRL generates the next iterate θ_t as

$$\theta_t = \arg\min_{\theta \in \Theta} \sum_{i=1}^{t} P_i(\theta), \tag{8}$$

$$P_i(\theta) = \langle g_i, \theta \rangle + \tfrac{1}{2}\|\theta - \theta_{i-1}\|_{Q_i}^2.$$

- FTML: consider only Pi's in a recent window

$$\theta_t = \arg\min_{\theta \in \Theta} \sum_{i=1}^{t} w_{i,t} P_i(\theta), \tag{9}$$

$$w_{i,t} = \frac{(1-\beta_1)\beta_1^{t-i}}{1-\beta_1^t} \tag{10}$$

等价于

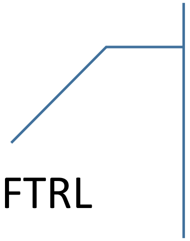
exponential moving average of the P_i's: $S_i = \beta_1 S_{i-1} + (1-\beta_1)P_i$, where $\beta_1 \in [0,1)$ and $S_0 = 0$. This can be easily rewritten as $S_t = (1-\beta_1)\sum_{i=1}^{t} \beta_1^{t-i} P_i$. ~~Instead of~~

~~are normalized to sum to 1.~~ The denominator $1 - \beta_1^t$ plays a similar role as bias correction in Adam. ~~When $\beta_1 = 0$,~~

- 注意：FTML的Pi 和FTRL的Pi并不相同，里面对Qi做了改动，详见下页

FTML: 对学习率做加权平均

- loss函数的二阶导数为学习率的倒数，我们仍然希望这个学习率是各个维度不同的，即为g^2（的加权平均值）



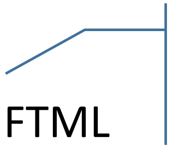
FTRL

$$\theta_t = \arg\min_{\theta \in \Theta} \sum_{i=1}^{t} P_i(\theta), \tag{8}$$

$$Q_{1:t} = \mathrm{diag}\left(\frac{1}{\eta}\left(\sqrt{g_{1:t}^2} + \epsilon \mathbf{1}\right)\right). \tag{5}$$

$$P_i(\theta) = \langle g_i, \theta \rangle + \tfrac{1}{2}\|\theta - \theta_{i-1}\|_{Q_i}^2.$$

Note that the Hessian of the objective in (8) is $Q_{1:t}$. This becomes $\sum_{i=1}^{t} w_{i,t} Q_i$ in (9). ~~Recall that $Q_{1:t}$ depends on~~



FTML

$$\sum_{i=1}^{t} w_{i,t} Q_i = \mathrm{diag}\left(\frac{1}{\eta_t}\left(\sqrt{\frac{v_t}{1-\beta_2^t}} + \epsilon_t \mathbf{1}\right)\right), \tag{11}$$

define $v_i = \beta_2 v_{i-1} + (1-\beta_2) g_i^2$, where $\beta_2 \in [0, 1)$ and $v_0 = 0$, and then correct its bias by dividing by $1 - \beta_2^t$.

- 设dt为loss二阶导数qt的指数加权和，即

$$d_t = \beta_1 d_{t-1} + (1 - \beta_1) Q_t$$

其无偏估计为(11)，可反推得：

Proposition 1. *Define* $d_t = \frac{1-\beta_1^t}{\eta_t}\left(\sqrt{\frac{v_t}{1-\beta_2^t}} + \epsilon_t \mathbf{1}\right)$. *Then,*

$$Q_t = \mathit{diag}\left(\frac{d_t - \beta_1 d_{t-1}}{1 - \beta_1}\right). \tag{12}$$

$\eta_t = \eta/\sqrt{t}$ and $\epsilon_t = \epsilon/\sqrt{t}$,

FTML: 封闭解

- 根据前面讨论，再复述一下FTML问题：

At round t, generate the next iterate θ_t by solving:

$$\theta_t = \arg\min_{\theta\in\Theta} \sum_{i=1}^{t} w_{i,t} P_i(\theta), \qquad (9)$$

$$w_{i,t} = \frac{(1-\beta_1)\beta_1^{t-i}}{1-\beta_1^t} \qquad (10)$$

Define $d_t = \frac{1-\beta_1^t}{\eta_t}\left(\sqrt{\frac{v_t}{1-\beta_2^t}} + \epsilon_t \mathbf{1}\right)$. *Then,*

$$Q_t = diag\left(\frac{d_t - \beta_1 d_{t-1}}{1-\beta_1}\right). \qquad (12)$$

- 其封闭解为：$P_i(\theta) = \langle g_i, \theta\rangle + \frac{1}{2}\|\theta - \theta_{i-1}\|^2_{Q_i}$.

$$\theta_t = \Pi_{\Theta}^{diag(d_t/(1-\beta_1^t))}(-z_t/d_t),$$

where $z_t = \beta_1 z_{t-1} + (1-\beta_1)g_t - \sigma_t\theta_{t-1}$, *and* $\Pi_\Theta^A(x) \equiv \arg\min_{u\in\Theta} \frac{1}{2}\|u-x\|_A^2$ *is the projection onto* Θ *for a given positive semidefinite matrix* A.

$$\sigma_i \equiv d_i - \beta_1 d_{i-1}$$

(9)式是很多小二项式加权求和，这里封闭解就是将其化简为一个二项式，问题就成了如何求解这个最简二项式的一次项和二次项系数。通过导数来求解。

二阶导数由定义，为：$\frac{1}{\eta_t}\left(\sqrt{\frac{v_t}{1-\beta_2^t}} + \epsilon_t \mathbf{1}\right)$

一阶导数推导过程：

$$S_t = \beta_1 S_{t-1} + (1-\beta_1)P_t$$

两边求导，得到一次项的系数Zt

$$z_t = \beta_1 z_{t-1} + (1-\beta_1)\left(g_t - \frac{d_t - \beta_1 d_{t-1}}{1-\beta_1}\theta_{t-1}\right)$$

$z_t = \beta_1 z_{t-1} + (1-\beta_1)g_t - (d_t - \beta_1 d_{t-1})\,\theta_{t-1}$

为得到无偏估计，再除以$(1-\beta_1^t)$

故得到一次项系数$\frac{z_t}{1-\beta_1^t}$，二次项系数为$\frac{1}{\eta_t}\left(\sqrt{\frac{v_t}{1-\beta_2^t}} + \epsilon_t \mathbf{1}\right) = \frac{d_t}{1-\beta_1^t}$，这就是左边的封闭解

FTML和RMSprop的关系

- FTML解的另一种形式：

Theorem 1. *With $\Theta = \mathbb{R}^d$, FTML generates the same updates as:*

$$\theta_t = \theta_{t-1} - diag\left(\frac{1-\beta_1}{1-\beta_1^t}\frac{\eta_t}{\sqrt{v_t/(1-\beta_2^t)}+\epsilon_t\mathbf{1}}\right)g_t. \quad (14)$$

- 特殊情况下退化为RMSprop

When $\beta_1 = 0$ and bias correction for the variance is not used, (14) reduces to RMSprop in (7). ~~However, recall~~

$$\theta_t = \theta_{t-1} - \mathrm{diag}\left(\frac{\eta}{\sqrt{v_t}+\epsilon\mathbf{1}}\right)g_t. \quad (7)$$

$$v_i = \beta v_{i-1} + (1-\beta)g_i^2, \quad (6)$$

FTML和Adam的关系

- FTML的loss函数的Regular项是以各个历史的θ_{i-1}为中心，而Adam都以最近的θ_{t-1}为中心

At iteration t, instead of centering regularization at each θ_{i-1} in (13), consider centering all the proximal regularization terms at the last iterate θ_{t-1}. θ_t then becomes:

$$\arg\min_{\theta\in\Theta} \sum_{i=1}^{t} w_{i,t}\left(\langle g_i, \theta\rangle + \frac{1}{2}\|\theta-\theta_{t-1}\|^2_{\mathrm{diag}\left(\frac{\sigma_i}{1-\beta_1}\right)}\right). \quad (15)$$

Proposition 5. *In (15),*

$$\theta_t = \Pi_\Theta^{A_t}\left(\theta_{t-1} - \boxed{A_t^{-1}}\boxed{\sum_{i=1}^{t} w_{i,t} g_i}\right), \quad (16)$$

where $A_t = diag((\sqrt{v_t/(1-\beta_2^t)} + \epsilon_t \mathbf{1})/\eta_t)$.

As in Adam, $\sum_{i=1}^{t} w_{i,t} g_i$ in (16) can be obtained as $m_t/(1-\beta_1^t)$, where m_t is computed as an exponential moving average of g_t's: $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$.

Adaptive learning rate, like RMSprop

Momentum

Remark: FTML combines all the nice properties

	是否对样本做平均（有关于$\boldsymbol{\beta_1}$）	是否对g^2做无偏估计（有关于$\boldsymbol{\beta_2}$）	**Regularization项的中心**
RMSprop	$\beta_1 = 0$ (and thus relies only on the current sample)	does not correct the bias of the variance estimate	centers the regularization at the current iterates θ_{i-1}
Adam	$\beta_1 > 0$	bias-corrected variance	centers all regularization terms at the last iterate θ_{t-1}
FTML	$\beta_1 > 0$	bias-corrected variance	centers the regularization at the current iterates θ_{i-1}

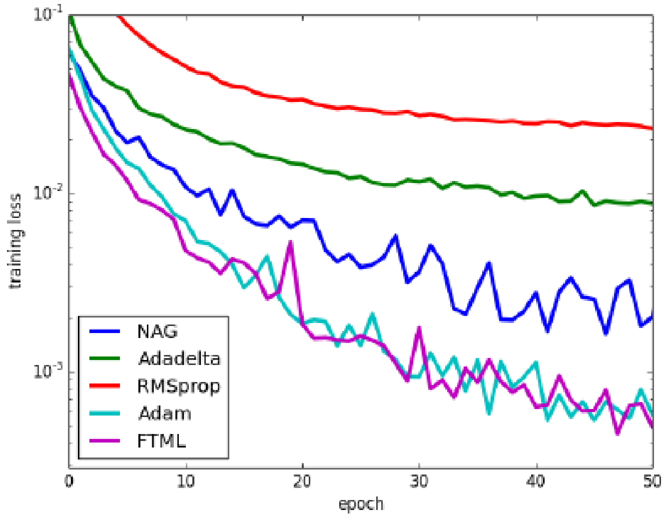
FTML算法实现

Algorithm 1 Follow the Moving Leader (FTML).

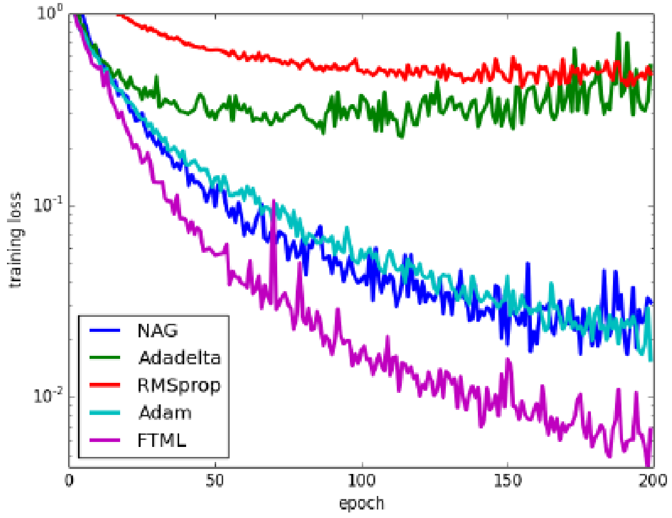
Input: $\eta_t > 0, \beta_1, \beta_2 \in [0, 1), \epsilon_t > 0.$
initialize $\theta_0 \in \Theta; d_0 \leftarrow 0; v_0 \leftarrow 0; z_0 \leftarrow 0;$
for $t = 1, 2, \dots, T$ **do**
  fetch function f_t;
  $g_t \leftarrow \partial_\theta f_t(\theta_{t-1});$
  $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2;$
  $d_t \leftarrow \frac{1-\beta_1^t}{\eta_t}\left(\sqrt{\frac{v_t}{1-\beta_2^t}} + \epsilon_t \mathbf{1}\right);$
  $\sigma_t \leftarrow d_t - \beta_1 d_{t-1};$
  $z_t \leftarrow \beta_1 z_{t-1} + (1 - \beta_1) g_t - \sigma_t \theta_{t-1};$
  $\theta_t \leftarrow \Pi_\Theta^{\mathrm{diag}(d_t/(1-\beta_1^t))}(-z_t/d_t);$
end for
Output: θ_T.

Experiments

- CNN
- Deep Residual Networks
- Memory Networks
- Neural Conversational Model
- Deep Q-Network
- LSTM

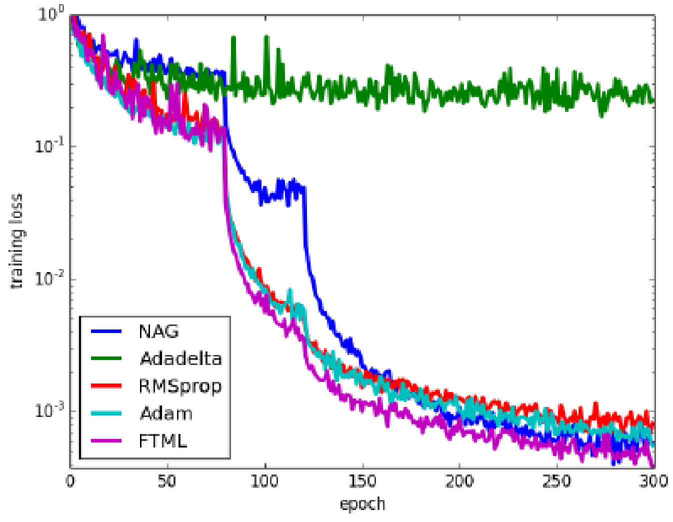


(a) *MNIST.*

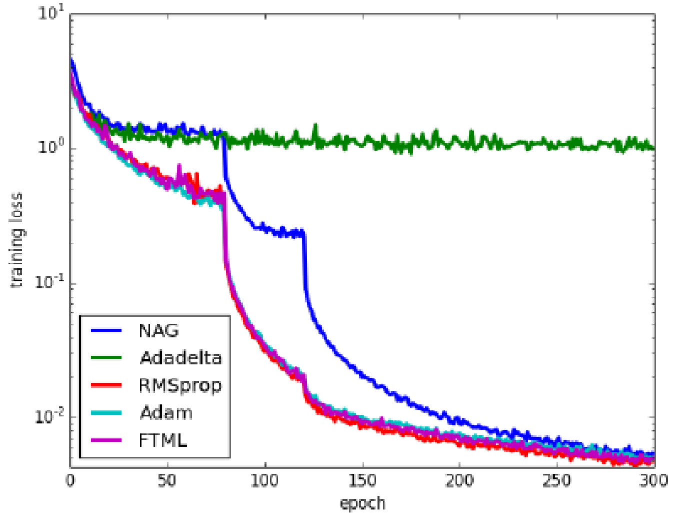


(b) *CIFAR-10.*

Figure 1. Results on convolutional neural network.



(a) *CIFAR-10.*



(b) *CIFAR-100.*

Figure 2. Results on deep residual network.

增加L1和L2正则项

- 一般地，对下列二次函数最优化问题：

$$argmin_w\left(\frac{1}{2}aw^2+bw+l_1|w|+C\right)$$

其中 a>0, l1是L1正则系数, l1>0, C为常数

- 则w有封闭解：

$$w=\begin{cases}0, & |b|\le l_1\\ -\frac{1}{a}(b-sgn(b)l_1), & otherwise\end{cases}$$

- 设FTML有L1正则系数为l1, L2正则系数为l2，则其封闭解对应的a和b为：

$$\theta_t=\Pi_{\Theta}^{diag(d_t/(1-\beta_1^t))}(-z_t/d_t),$$

$$+l_1\|\theta\|_1+\frac{1}{2}l_2\|\theta\|_2$$

$$a=\frac{d_t}{(1-\beta_1^t)}+l_2$$

$$b=\frac{z_t}{1-\beta_1^t}$$

结论

- FTML会对最近的样本的loss加大权重
- 因此，他可以更快的训练到另一个局部最优解
- 而且对数据分布发生变化的情况表现更好
- FTML集合了RMSprop和Adam的优点，避免了他们的缺点
- RMSprop缺点是不够稳定
- Adam缺点是收敛速度没有FTML快。因为它使用了过去所有的梯度值，而随着数据分布变化，其中一部分已经没用了。