Sparse Online Learning via Truncated Gradient [T33]

John Langford1, Lihong Li2, Tong Zhang3
1Yahoo! Research, 2Rutgers University
jl@yahoo-inc.com, lihong@cs.rutgers.edu, tongz@rci.rutgers.edu

Abstract
We introduce Truncated Gradient (TG) that
• suffers small online regret
• approximates L1 regularization
• scales sub-linearly in total number of features
is simple and easy to implement

Stochastic Gradient Descent
For $t=1,2,...$
1. Observe an unlabeled example $x_i$
2. Make prediction based on existing weight $w_t \in R^d$
3. Observe $y_i$ and incur loss $L(w_t, y_i)$ where $z_i = (x_i, y_i)$
4. Perform stochastic gradient descent: $w_{t+1} = w_t - \eta \nabla_t L(w_t, (x_t, y_t))$

Sparse Online Learning
$w_{t+1} = w_t - \eta \nabla_t L(w_t, (x_t, y_t))$

TG Has Small Regret
Assumption #1 $L(w, z)$ is convex in $w$.
Assumption #2 $\exists A > 0$ s.t.
$\sup_{w, \eta} \sum_{t=1}^T \eta L(w_t, z_t) + B \leq A L(w, z) + B$
Theorem: If $\eta_t = 0$, then for all $w \in R^d$:
$\frac{1}{T} \sum_{t=1}^T \eta_t L(w_t, z_t) + B \leq \frac{\eta_t}{T} \sum_{t=1}^T \eta L(w_t, z_t) + B$

TG Approximates L1 Regularization
Theorem: Given training examples $z_i = (x_i, y_i)$ for $i=1,...,n$.
Define $w_{t+1} = (T \eta_t L(w_t, z_t, g_t))$. $\bar{w}_{t+1} = w + (w_{t+1} - \bar{w}_t) / T$
Let $\bar{w}_t$ be Uniform(1, $\ldots$, n), then
$E_{t=1}^\infty R(\bar{w}_t, g) \leq E_{t=1}^\infty \frac{1}{T} R(w_t, g) \leq \inf R(w, g) + o(1)$

TG (w/ Square Loss) is Efficient
Algorithm 1 Truncated Gradient for Square Loss
Inputs:
• threshold $\theta \geq 0$
• gravity sequence $g_t \geq 0$
• learning rate $\eta \in (0, 1)$
• example oracle $O$
initializes weights $w_t \leftarrow 0$ ($j=1, \ldots, d$)
for trial $t = 1, 2, \ldots$
1. Acquire an unlabeled example $x = [x_1, x_2, \ldots, x_d]$ from oracle $O$
2. For $j = 1, \ldots, d$
   (a) if $w_j > \theta$ then $w_j \leftarrow \max\{w_j - g_t y_t, 0\}$
   (b) else if $w_j < -\theta$ then $w_j \leftarrow \min\{|w_j| + g_t y_t, 0\}$
3. Compute prediction: $\hat{y} = \sum_j w_j x_j$
4. Acquire the label $y$ from oracle $O$
5. Update weights for all features $j$: $w_j \leftarrow w_j - 2\eta (y - \hat{y}) x_j$

Computation complexity: $\Theta(k)$ where $k = ||w||_1$
while $\Theta(k \ln d)$ in [Duchi & Singer 2008]

Open Source
Will be available in Vowpal Wabbit (http://www.hunch.net/~vw)