



电子科技大学

University of Electronic Science and Technology of China



# Online Learning Algorithm and Optimization

Reporter: Jiatu(Steven) Shi  
E-mail: jiatu.steven@gmail.com



Data Mining Lab, Big Data Research Center, UESTC  
Email: junmshao@uestc.edu.cn  
<http://staff.uestc.edu.cn/shaojunming>

# Outline

- Introduction
- Truncated Gradient, FOBOS
- RDA (Regularized dual averaging)
- FTRL (Follow-the-regularized-Leader)
- Discussion and Conclusion



# Part 1 Introduction

# Introduction



数据挖掘实验室  
Data Mining Lab

## Predicting CTR & RPM

A screenshot of a Taobao search results page. At the top, there's a banner for '狂欢专场' (Double 11 Special) with the tagline '裸价，不看后悔！'. Below it, a section titled '猜你喜欢' (Guess You Like) shows five items: a woman in a cable-knit sweater, a sofa, a woman in a long coat, a computer case, and a pair of sneakers. Each item has its name, price, and a 'View Details' button.

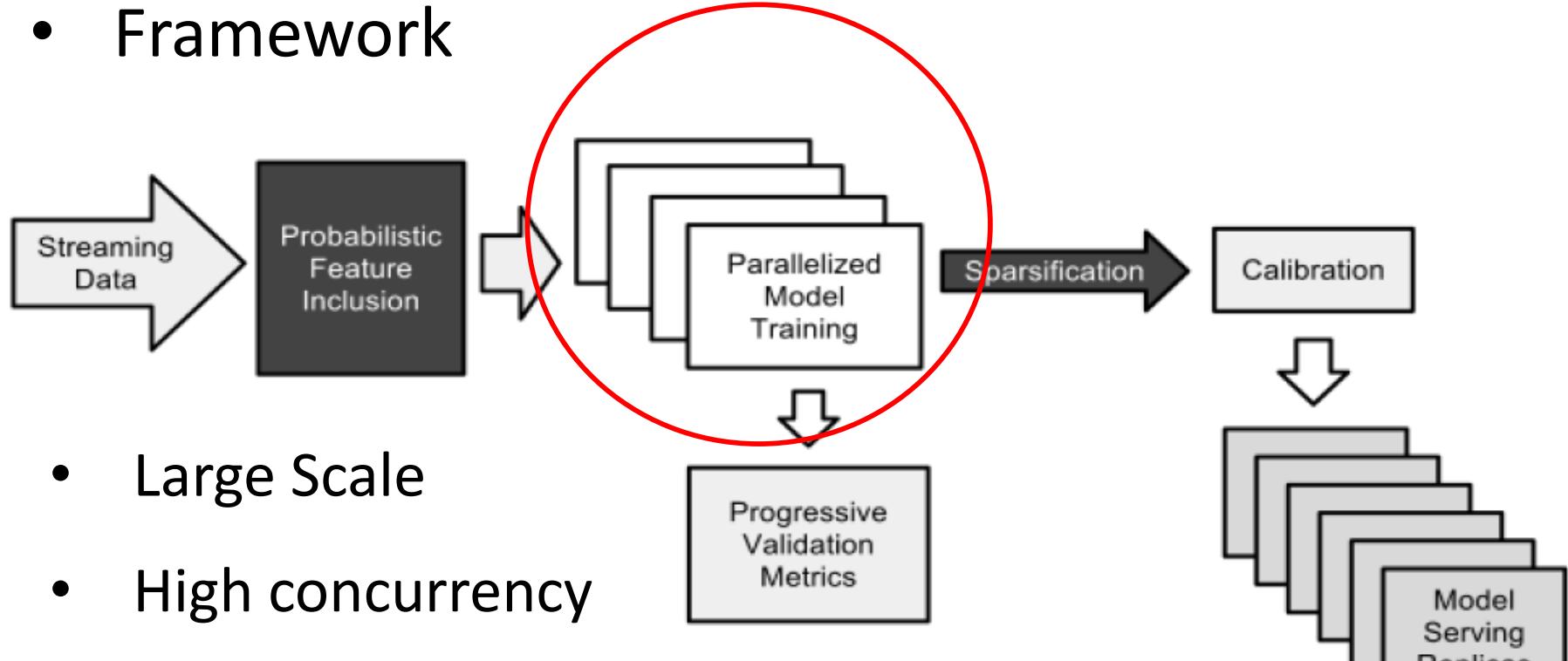
商品名称	价格	状态
高领加厚毛衣女套头毛线衣韩版潮冬季衣服宽松韩国短款学生针织衫	¥62	¥62
布艺沙发组合韩式现代小户型卧室客厅双人位碎花懒人沙发清仓特价	¥100	¥100
原创主题2015韩版秋冬新款休闲拼接棉衣中长款显瘦外套时尚棉服女	¥328	¥328
名龙堂i7 6700/GTX980Ti/M.2水冷概念显卡电脑主机	¥13999	¥13999
阿迪达斯板鞋女鞋2015秋NEO运动鞋复古休闲跑鞋AQ 1571 F	¥375	¥375

A screenshot of a Juhuasuan promotional page. It features a red header '聚划算' (Juhuasuan) and a navigation bar with categories: 全部, 美妆, 童装, 零食, 母婴, 百货, 家居. The main area shows two deals for women's clothing: a red dress and a red hoodie. Each deal includes a price, a 'Buy Now' button, and a 'Sold Out' count.

商品名称	价格	已售
抽红包 [比双11还低23元]秋冬比抢！通勤特显瘦 假两件连...	¥280.00	1383 件已售
金衣紫加绒加厚童装2015秋季新款长袖女童中长款卫衣外...	¥279.00	967 件已售

# Introduction

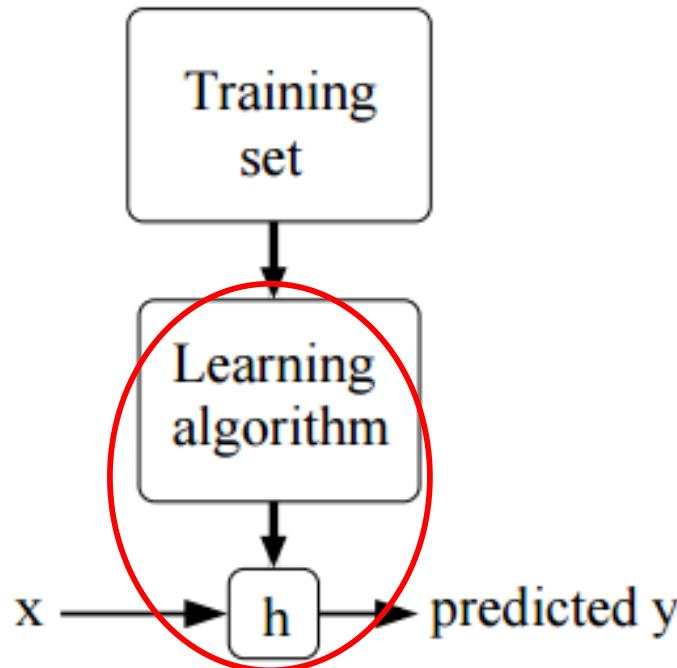
- Framework



- Large Scale
- High concurrency
- Real Time

# Introduction

- Framework
  - \* Solving the optimization problem



$$W = \arg \min_W l(W, Z)$$

$$Z = \{(X_j, y_j) | j = 1, 2, \dots, M\}$$

$$y_j = h(W, X_j)$$

$$l(W, Z) = L(W) + \varphi(W)$$

# Introduction

- From Batch to Online

---

## Algorithm 1. Batch Gradient Descent

---

*Repeat until convergence {*

$$W^{(t+1)} = W^{(t)} - \eta^{(t)} \nabla_W \ell(W^{(t)}, Z)$$

*}*

---

---

## Algorithm 2. Stochastic Gradient Descent

---

*Loop {*

*for j=1 to M {*

$$W^{(t+1)} = W^{(t)} - \eta^{(t)} \nabla_W \ell(W^{(t)}, Z_j)$$

*}*

*}*

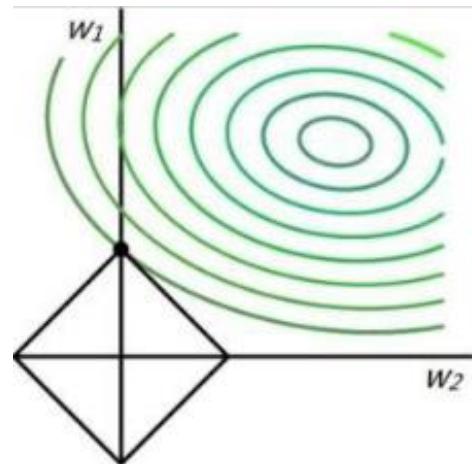
---



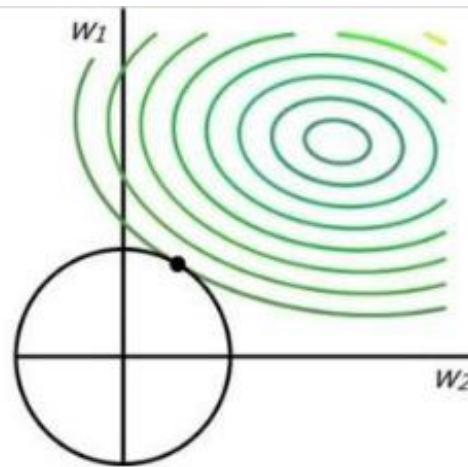
# Part 2 Truncated Gradient, FOBOS

# Truncated Gradient

- Sparsity



(a)  $L1$  Regularization



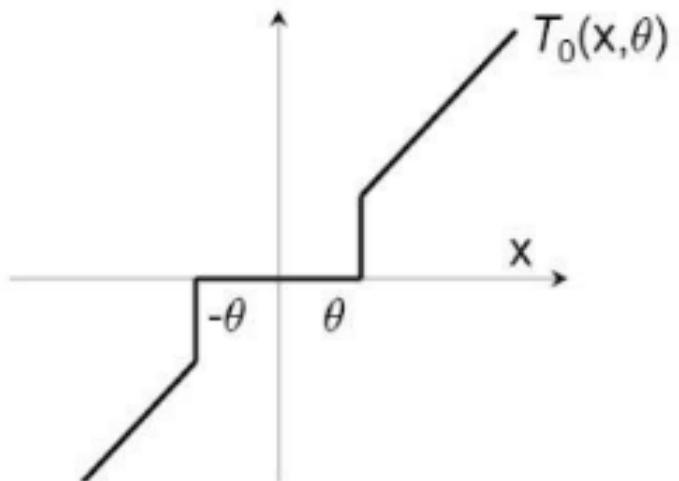
(b)  $L2$  Regularization

# Truncated Gradient

- Aggressive Rounding

$$f(w_i) = T_0(w_i - \eta \nabla_1 L(w_i, z_i), \theta)$$

$$T_0(v_j, \theta) = \begin{cases} 0 & \text{if } |v_j| \leq \theta \\ v_j & \text{otherwise} \end{cases}$$

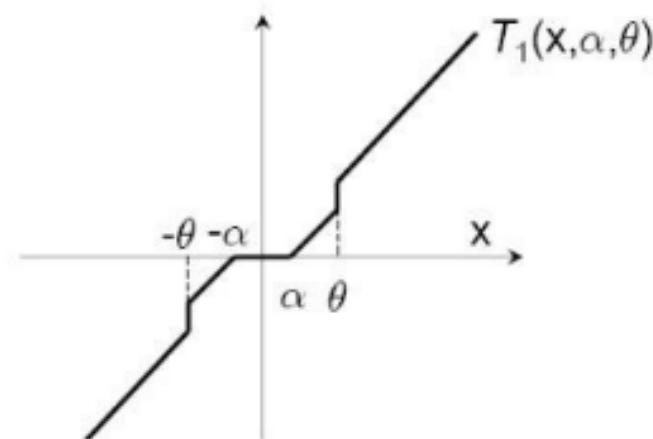


# Truncated Gradient

- Smooth rounding

$$f(w_i) = T_1(w_i - \eta \nabla_1 L(w_i, z_i), \eta g_i, \theta)$$

$$T_0(v_j, \theta) = \begin{cases} \max(0, v_j - \alpha) & \text{if } v_j \in [0, \theta] \\ \min(0, v_j + \alpha) & \text{if } v_j \in [-\theta, 0] \\ v_j & \text{otherwise} \end{cases}$$



# FOBOS

- Empirical gradient decent and optimization

$$W^{(t+\frac{1}{2})} = W^{(t)} - \eta^{(t)} G^{(t)}$$

$$W^{(t+1)} = argmin_W \left\{ \frac{1}{2} \|W - W^{(t+\frac{1}{2})}\|^2 - \eta^{(t+\frac{1}{2})} \psi(W) \right\}$$



$$W^{(t+1)} = argmin_W \left\{ \frac{1}{2} \|W - W^t + \eta^{(t)} G^{(t)}\|^2 - \eta^{(t+\frac{1}{2})} \psi(W) \right\}$$

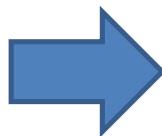
# FOBOS

- Iteration

$$if \quad W^{(t+1)} = argmin_W F(W)$$

$$0 \in \partial F(W) = W - W^{(t)} + \eta^{(t)} G^{(t)} + \eta^{\left(t+\frac{1}{2}\right)} \partial \psi(W)$$

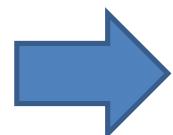
$$0 = \left\{ W - W^{(t)} - \eta^{(t)} G^{(t)} + \eta^{\left(t+\frac{1}{2}\right)} \partial \psi(W) \right\} \Big|_{W=W^{(t+1)}}$$

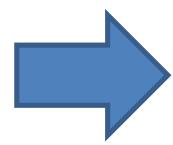
  $W^{(t+1)} = W^{(t)} - \eta^{(t)} G^{(t)} - \eta^{\left(t+\frac{1}{2}\right)} \partial \psi(W^{(t+1)})$

# FOBOS

- L1-norm

Let  $\psi(W) = \lambda \|W\|_1, V = [v_1, v_2, \dots, v_N] \in \mathbf{R}^N$


$$W^{(t+1)} = \operatorname{argmin}_W \sum_{i=1}^N \left( \frac{1}{2} (w_i - v_i)^2 + \tilde{\lambda} |w_i| \right)$$


$$\text{divided } w_i^{t+1} = \operatorname{argmin}_W \left( \frac{1}{2} (w_i - v_i)^2 + \tilde{\lambda} |w_i| \right)$$

# FOBOS

- L1-norm

if  $w_i^*$  is the solution then  $w_i^* v_i \geq 0$

because if  $w_i^* v_i < 0$

$$\frac{1}{2} v_i^2 < \frac{1}{2} v_i^2 - w_i^* v_i + \frac{1}{2} (w_i^*)^2 < \frac{1}{2} (w_i^* - v_i)^2 + \tilde{\lambda} |w_i^*|$$

for  $\text{minimize}_{w_i} \left( \frac{1}{2} (w_i - v_i)^2 + \tilde{\lambda} |w_i| \right)$  let  $-w_i \leq 0$

in KKT condition

# FOBOS

- Case1  $v_i \geq 0, \omega_i \geq 0$

for  $\text{minimize}_{w_i} \left( \frac{1}{2} (w_i - v_i)^2 + \tilde{\lambda} |w_i| \right)$

Let  $\beta$  be the Lagrange factor, then use KKT

condition add  $-w_i \leq 0$

$$\frac{\partial}{\partial w_i} \left( \frac{1}{2} (w_i - v_i)^2 + \tilde{\lambda} w_i - \beta w_i \right) \Big|_{w_i=w_i^*} = 0 \text{ and } \beta w_i = 0$$

$$\Rightarrow w_i^* = v_i - \tilde{\lambda} + \beta$$

# FOBOS

(1)  $w_i^* > 0$  :

$$\begin{aligned}\beta w_i^* &= 0 \Rightarrow \beta = 0 \\ \Rightarrow w_i^* &= v_i - \tilde{\lambda} \\ w_i^* > 0 &\Rightarrow v_i - \tilde{\lambda} > 0\end{aligned}$$

(2)  $w_i^* = 0$  :

$$\begin{aligned}\Rightarrow v_i - \tilde{\lambda} + \beta &= 0 \\ \beta \geq 0 &\Rightarrow v_i - \tilde{\lambda} \leq 0\end{aligned}$$

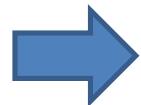
In conclusion  $v_i \geq 0 \Rightarrow w_i^* = \max(0, v_i - \tilde{\lambda})$

# FOBOS

- Case2:  $v_i < 0 \quad w_i^* = -\max(0, -v_i - \tilde{\lambda})$
- Conclusion

$$w_i^{(t+1)} = \begin{cases} 0 & , \quad \text{if } |w_i^{(t)} - \eta^{(t)} g_i^{(t)}| \leq \eta^{(t+\frac{1}{2})} \lambda \\ \left( w_i^{(t)} - \eta^{(t)} g_i^{(t)} - \eta^{(t+\frac{1}{2})} \lambda \cdot \text{sgn}(w_i^{(t)} - \eta^{(t)} g_i^{(t)}) \right) & , \text{otherwise} \end{cases}$$

$$\theta = \infty, k = 1, \lambda_{TG}^{(t)} = \eta^{(t+\frac{1}{2})} \lambda$$



$$f(w_i) = T_1(w_i - \eta \nabla_1 L(w_i, z_i), \eta g_i, \theta)$$

$$T_0(v_j, \theta) = \begin{cases} \max(0, v_j - \alpha) & \text{if } v_j \in [0, \theta] \\ \min(0, v_j + \alpha) & \text{if } v_j \in [-\theta, 0] \\ v_j & \text{otherwise} \end{cases}$$



# Part 3 RDA (Regularized dual averaging)

# RDA

- primal-dual algorithmic schema

$$W^{(t+1)} = \operatorname{argmin}_W \left\{ \frac{1}{t} \sum_{r=1}^t \langle G^{(r)}, W \rangle + \psi(W) + \frac{\beta^{(t)}}{t} h(W) \right\}$$

Let  $\psi(W) = \lambda \|W\|_1$ ,  $h(W) = \frac{1}{2} \|W\|_2^2$ ,  $\{\beta^{(t)} | t \geq 1\}$ ,  $\beta^{(t)} = \gamma \sqrt{t}$

$$W^{(t+1)} = \operatorname{argmin}_W \left\{ \frac{1}{t} \sum_{r=1}^t \langle G^{(r)}, W \rangle + \lambda \|W\|_1 + \frac{\gamma}{2\sqrt{t}} \|W\|_2^2 \right\}$$

# RDA

$$W^{(t+1)} = \operatorname{argmin}_W \left\{ \frac{1}{t} \sum_{r=1}^t \langle G^{(r)}, W \rangle + \lambda \|W\|_1 + \frac{\gamma}{2\sqrt{t}} \|W\|_2^2 \right\}$$

Divided  $\rightarrow \min_{w_i \in R} \{ \bar{g}_i^{(t)} w_i + \lambda |w_i| + \frac{\gamma}{2\sqrt{t}} w_i^2 \}$

$$\rightarrow W_i^{(t+1)} = \begin{cases} 0 & , \quad if \quad |\bar{g}_i^{(t)}| < \lambda \\ \left( -\frac{\sqrt{t}}{\gamma} (\bar{g}_i^{(t)} - \lambda \cdot sgn(\bar{g}_i^{(t)})) \right) & , otherwise \end{cases}$$

FOBOS  $|w_i^{(t)} - \eta^{(t)} g_i^{(t)}| \leq \lambda_{TG}^{(t)} = \eta^{(t+\frac{1}{2})} \lambda \Rightarrow \vartheta \left( \frac{1}{\sqrt{t}} \right) \lambda > \lambda$



# Part 4 FTRL (Follow-the-regularized-Leader)

# FTRL

- The similarity between FOBOS and RDA

$$W^{(t+1)} = \operatorname{argmin}_W \left\{ \frac{1}{2} \|W - W^t + \eta^{(t)} G^{(t)}\|^2 + \eta^{(t)} \lambda \|W\|_1 \right\} \text{--FOBOS}$$

$$\min_{w_i \in R} \left\{ \frac{1}{2} \left( w_i - w_i^{(t)} + \eta^{(t)} g_i^{(t)} \right)^2 + \eta^{(t)} \lambda |w_i| \right\}$$

$$= \min_{w_i \in R} \left\{ \frac{1}{2} \left( w_i - w_i^{(t)} \right)^2 + \frac{1}{2} \left( \eta^{(t)} g_i^{(t)} \right) + w_i \eta^{(t)} g_i^{(t)} + w_i \eta^{(t)} g_i^{(t)} + \eta^{(t)} \lambda |w_i| \right\}$$

$$= \min_{w_i \in R} \left\{ w_i g_i^{(t)} + \lambda |w_i| + \frac{1}{2\eta^{(t)}} \left( w_i - w_i^{(t)} \right)^2 + \left[ \frac{\eta^{(t)}}{2} \left( g_i^{(t)} \right)^2 + w_i^{(t)} g_i^{(t)} \right] \right\}$$

# FTRL

- The similarity between FOBOS and RDA

$$\min_{w_i \in R} \left\{ w_i g_i^{(t)} + \lambda |w_i| + \frac{1}{2\eta^{(t)}} (w_i - w_i^{(t)})^2 + \left[ \frac{\eta^{(t)}}{2} (g_i^{(t)})^2 + w_i^{(t)} g_i^{(t)} \right] \right\}$$

$$\sim \min_{w_i \in R} \left\{ w_i g_i^{(t)} + \lambda |w_i| + \frac{1}{2\eta^{(t)}} (w_i - w_i^{(t)})^2 \right\}$$

$$W^{(t+1)} = \operatorname{argmin}_W \left\{ G^{(t)} \cdot W + \lambda \|W\|_1 + \frac{1}{2\eta^{(t)}} \|W - W^t\|_2^2 \right\} \text{ L1FOBOS}$$

$$W^{(t+1)} = \operatorname{argmin}_W \{ G^{(1:t)} \cdot W + \lambda \|W\|_1 + \frac{1}{2\eta^{(t)}} \|W - 0\|_2^2 \} \text{ L1RDA}$$

# FTRL

- The Combination of FOBOS and RDA ---- FTRL

$$W^{(t+1)} = \operatorname{argmin}_W \left\{ G^{(1:t)} \cdot W + \lambda_1 \|W\|_1 + \right.$$

# FTRL

- Optimization  $w_i^* v_i \geq 0, w_i^* \geq 0$

$$W^{(t+1)} = \operatorname{argmin}_W \left\{ Z^{(t)} \cdot W + \lambda_1 \|W\|_1 + \frac{1}{2} (\lambda_2 + \sum_{s=1}^t \sigma^s) \|W\|_2^2 \right\}$$

Divided  $\underset{w_i \in R}{\operatorname{minimize}} \left\{ z_i^{(t)} \cdot w_i + \lambda_1 |w_i| + \frac{1}{2} \left( \lambda_2 + \sum_{s=1}^t \sigma^s \right) w_i^2 \right\}$

$$W_i^{(t+1)} = \begin{cases} 0 & , \quad if \quad |z_i^{(t)}| < \lambda_1 \\ - \left( \lambda_2 + \sum_{s=1}^t \sigma^s \right)^{-1} (z_i^{(t)} - \lambda_1 \cdot \operatorname{sgn}(z_i^{(t)})) & , \quad otherwise \end{cases}$$

# FTRL

---

**Algorithm 1** Per-Coordinate FTRL-Proximal with  $L_1$  and  $L_2$  Regularization for Logistic Regression

---

# With per-coordinate learning rates of Eq. (2).

**Input:** parameters  $\alpha, \beta, \lambda_1, \lambda_2$

( $\forall i \in \{1, \dots, d\}$ ), initialize  $z_i = 0$  and  $n_i = 0$

**for**  $t = 1$  **to**  $T$  **do**

    Receive feature vector  $\mathbf{x}_t$  and let  $I = \{i \mid x_i \neq 0\}$

    For  $i \in I$  compute

$$w_{t,i} = \begin{cases} 0 & \text{if } |z_i| \leq \lambda_1 \\ -\left(\frac{\beta + \sqrt{n_i}}{\alpha} + \lambda_2\right)^{-1}(z_i - \text{sgn}(z_i)\lambda_1) & \text{otherwise.} \end{cases}$$

Predict  $p_t = \sigma(\mathbf{x}_t \cdot \mathbf{w})$  using the  $w_{t,i}$  computed above

Observe label  $y_t \in \{0, 1\}$

**for** all  $i \in I$  **do**

$g_i = (p_t - y_t)x_i$  #gradient of loss w.r.t.  $w_i$

$\sigma_i = \frac{1}{\alpha} \left( \sqrt{n_i + g_i^2} - \sqrt{n_i} \right)$  #equals  $\frac{1}{\eta_{t,i}} - \frac{1}{\eta_{t-1,i}}$

$z_i \leftarrow z_i + g_i - \sigma_i w_{t,i}$

$n_i \leftarrow n_i + g_i^2$

**end for**

**end for**

---

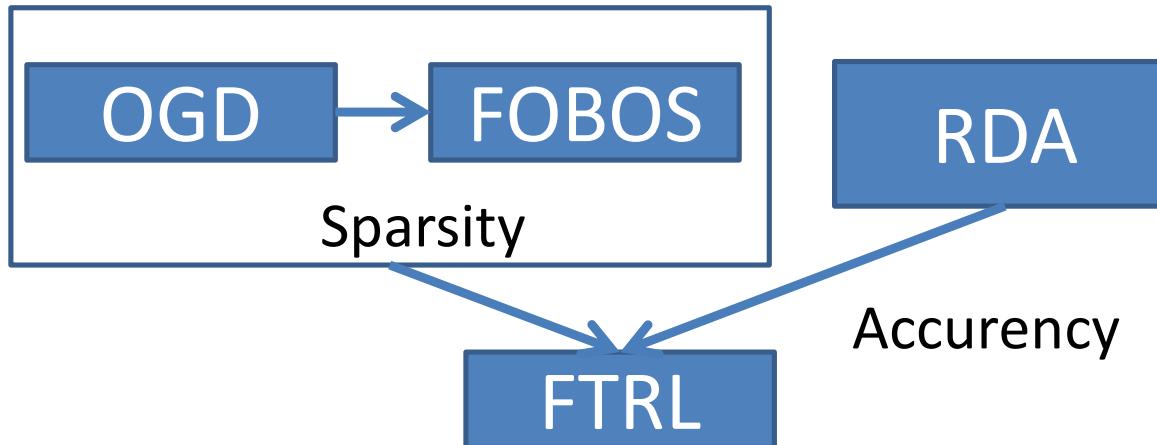
$$\eta_i^{(t)} = \frac{\alpha}{\beta + \sqrt{\sum_{s=1}^t (g_i^{(s)})^2}}, \sigma^{(1:t)} = \frac{1}{\eta^{(t)}},$$

$$\sum_{s=1}^t (\sigma^{(s)}) = \frac{1}{\eta_i^{(t)}} = (\beta + \sqrt{\sum_{s=1}^t (g_i^{(s)})^2}) / \alpha$$



# Part 5 Discussion and Conclusion

# Discussion & Conclusion



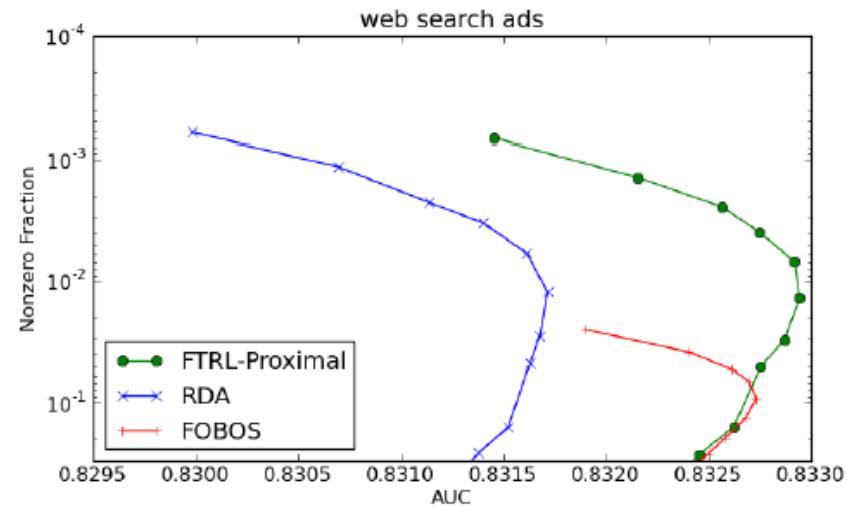
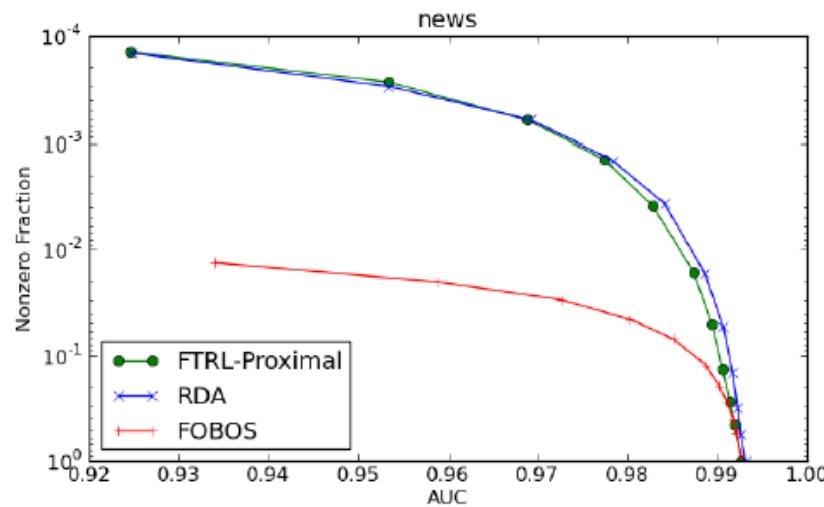
	Num. Non-Zero's	AucLoss Detriment
FTRL-PROXIMAL	baseline	baseline
RDA	+3%	0.6%
FOBOS	+38%	0.0%
OGD-COUNT	+216%	0.0%

# Discussion & Conclusion



数据挖掘实验室  
Data Mining Lab

DATA	FTRL-PROXIMAL	RDA	FOBOS
BOOKS	0.874 (0.081)	0.878 (0.079)	0.877 (0.382)
DVD	0.884 (0.078)	0.886 (0.075)	0.887 (0.354)
ELECTRONICS	0.916 (0.114)	0.919 (0.113)	0.918 (0.399)
KITCHEN	0.931 (0.129)	0.934 (0.130)	0.933 (0.414)
NEWS	0.989 (0.052)	0.991 (0.054)	0.990 (0.194)
RCV1	0.991 (0.319)	0.991 (0.360)	0.991 (0.488)
WEB SEARCH ADS	<b>0.832 (0.615)</b>	0.831 (0.632)	0.832 (0.849)





# Thanks

- Q&A