Adaptive Subgradient Methods for Online Learning and Stochastic Optimization

Farideh Fazayeli

CSCI 8990
ML at Large Scale and High Dimensions

Feb 19, 2014

Reference

J. Duchi, E. Hazan, and Y. Singer. “Adaptive Subgradient Methods for Online Learning and Stochastic Optimization.” Journal of Machine Learning Research, 12, 2121-2159, 2011.

Setting

Composite Objective Function:

$$\min_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x) + \varphi(x)$$

- $f_t(x)$: Closed Convex function
- $\varphi(x)$: Closed Convex function

Setting

Composite Objective Function:

$$\min_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x) + \varphi(x)$$

- $f_t(x)$: Closed Convex function
- $\varphi(x)$: Closed Convex function

$$g_t \in \partial f_t(x_t)$$

$$g_{1:t} = \begin{pmatrix} g_{1,1} & g_{2,1} & \cdots & g_{t,1} \\ g_{1,2} & g_{2,2} & \cdots & g_{t,2} \\ \vdots & \vdots & \ddots & \vdots \\ g_{1,d} & g_{2,d} & \cdots & g_{t,d} \end{pmatrix}$$

Setting

Composite Objective Function:

$$\min_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x) + \varphi(x)$$

- $f_t(x)$: Closed Convex function
- $\varphi(x)$: Closed Convex function

$$g_t \in \partial f_t(x_t)$$

$$g_{1:t} = \begin{pmatrix} g_{1,1} & g_{2,1} & \cdots & g_{t,1} \\ g_{1,2} & g_{2,2} & \cdots & g_{t,2} \\ \vdots & \vdots & \ddots & \vdots \\ g_{1,d} & g_{2,d} & \cdots & g_{t,d} \end{pmatrix}$$

Goal: Minimize the regret:

$$\begin{aligned} R(T) &= \sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - \inf_{x \in \mathcal{X}} \left[\sum_{t=1}^{T} f_t(x) + \varphi(x) \right] \\ &= \sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - f_t(x^*) + \varphi(x^*) \end{aligned}$$

Problem?

- Composite Objective Mirror Descent [OPT 1]

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \{\eta \langle g_t, x \rangle + \eta \varphi(x) + B_\psi(x, x_t)\} \tag{1}$$

- Primal-Dual sub-Gradient Descent [Xiao 2010]

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \{ \underbrace{\frac{\eta}{t} \sum_{\tau=1}^{t} \langle g_\tau, x \rangle}_{\eta \langle \bar{g}_t, x \rangle} + \eta \varphi(x) + \frac{1}{t} \psi(x) \} \tag{2}$$

- Decaying learning rate helps?
- How Proximal Function (ψ) effect learning rate?
- e.g., $\eta_t = \frac{\eta_0}{t} \equiv \psi_t = t\psi$
- Treating all features the same!

Key Insight

Earlier Algorithm:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ \eta \langle g_t, x \rangle + \eta \varphi(x) + \frac{1}{2} \| x - x_t \|_p^2 \right\} \tag{3}$$

Key Insight

Earlier Algorithm:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ \eta \langle g_t, x \rangle + \eta \varphi(x) + \frac{1}{2} \| x - x_t \|_p^2 \right\} \tag{3}$$

Regret Bound:

$$R(T) \leq \frac{1}{2\eta} \| x_1 - x^* \|_p^2 + \varphi(x_1) + \frac{\eta}{2} \boxed{\sum_{t=1}^{T} \| g_t \|_{p*}^2} \tag{4}$$

- $g_t \in \partial f_t(x_t)$.

Key Insight

Earlier Algorithm:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ \eta \langle g_t, x \rangle + \eta \varphi(x) + \frac{1}{2} \| x - x_t \|_p^2 \right\} \tag{3}$$

Adaptive Proximal Function:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ \eta \langle g_t, x \rangle + \eta \varphi(x) + \frac{1}{2} \| x - x_t \|_H^2 \right\} \tag{5}$$

Key Insight

Earlier Algorithm:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ \eta \langle g_t, x \rangle + \eta \varphi(x) + \frac{1}{2} \| x - x_t \|_p^2 \right\} \tag{3}$$

Adaptive Proximal Function:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ \eta \langle g_t, x \rangle + \eta \varphi(x) + \frac{1}{2} \| x - x_t \|_H^2 \right\} \tag{5}$$

- Mahalanobis norm ($H \succcurlyeq 0$):

$$\| x \|_H^2 = \langle x, Hx \rangle = x^T H x$$
$$\| x \|_{H^*}^2 = \langle x, H^{-1} x \rangle = x^T H^{-1} x$$

Key Insight

Earlier Algorithm:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ \eta \langle g_t, x \rangle + \eta \varphi(x) + \frac{1}{2} \| x - x_t \|_p^2 \right\} \tag{3}$$

Adaptive Proximal Function:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ \eta \langle g_t, x \rangle + \eta \varphi(x) + \frac{1}{2} \| x - x_t \|_H^2 \right\} \tag{5}$$

Regret Bound:

$$R(T) \leq \frac{1}{2\eta} \| x_1 - x^* \|_H^2 + \eta \varphi(x_1) + \frac{\eta}{2} \boxed{\sum_{t=1}^{T} \| g_t \|_{H^{-1}}^2} \tag{6}$$

Diagonal Matrices

- In order to minimize regret:

$$\min_{\substack{H \succcurlyeq 0 \\ tr(H) \leq c}} \quad \sum_{t=1}^{T} \| g_t \|^2_{H^{-1}}$$

Diagonal Matrices

- In order to minimize regret:

$$\min_{\substack{H \succcurlyeq 0 \\ tr(H) \leq c}} \quad \sum_{t=1}^{T} \| g_t \|^2_{H^{-1}}$$

$$H = \begin{pmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_d \end{pmatrix}$$

Diagonal Matrices

- In order to minimize regret:

$$\min_{\substack{H \succcurlyeq 0 \\ tr(H) \leq c}} \sum_{t=1}^{T} \| g_t \|^2_{H^{-1}} \equiv \min_{\substack{s \succcurlyeq 0 \\ \langle 1, s \rangle \leq c}} \sum_{i=1}^{d} \frac{\sum_{t=1}^{T} g_{t,i}^2}{s_i}$$

$$H = \begin{pmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_d \end{pmatrix}$$

Diagonal Matrices

- In order to minimize regret:

$$\min_{\substack{H\succcurlyeq 0 \\ tr(H)\leq c}} \sum_{t=1}^{T} \| g_t \|^2_{H^{-1}} \equiv \min_{\substack{s\succcurlyeq 0 \\ \langle 1,s\rangle \leq c}} \sum_{i=1}^{d} \frac{\sum_{t=1}^{T} g_{t,i}^2}{s_i} \qquad H = \begin{pmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_d \end{pmatrix}$$

- Lagrangian of the problem:

$$\mathcal{L}(s,\lambda,\theta) = \sum_{i=1}^{d} \frac{\| g_{1:T,i} \|_2^2}{s_i} - \langle \lambda, s\rangle + \theta(\langle 1, s\rangle - c).$$

Diagonal Matrices

- In order to minimize regret:

$$\min_{\substack{H\succcurlyeq 0 \\ tr(H)\leq c}} \sum_{t=1}^{T} \| g_t \|^2_{H^{-1}} \equiv \min_{\substack{s\succcurlyeq 0 \\ \langle 1,s\rangle \leq c}} \sum_{i=1}^{d} \frac{\sum_{t=1}^{T} g_{t,i}^2}{s_i} \qquad H = \begin{pmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_d \end{pmatrix}$$

- Lagrangian of the problem:

$$\mathcal{L}(s,\lambda,\theta) = \sum_{i=1}^{d} \frac{\| g_{1:T,i} \|_2^2}{s_i} - \langle \lambda, s\rangle + \theta(\langle 1, s\rangle - c).$$

$$\left.\begin{aligned} \frac{\partial \mathcal{L}}{\partial s_i} &= -\frac{\|g_{1:T,i}\|_2^2}{s_i^2} - \lambda + \theta = 0 \\ \lambda &= 0 \end{aligned}\right\} \Rightarrow \boxed{s_i = \frac{c \| g_{1:T,i} \|_2}{\sum_{i=1}^{d} \| g_{1:T,i} \|_2^2}}$$

ADAGRAD with Diagonal Matrices

1. $g_{1:t} = [g_{1:T-1} \;\; g_t]$
2. $s_{t,i} = \| g_{1:t,i} \|_2$
3. $H_t = \delta I + diag(s_t)$
4. $\psi_t(x) = \frac{1}{2}\langle x, H_t x \rangle$
5. Primal-Dual Sub Gradient Update:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \{\eta \left\langle \frac{1}{t} \sum_{\tau=1}^{t} g_\tau, x \right\rangle + \eta\varphi(x) + \frac{1}{t}\psi_t(x)\}$$

6. Composite Mirror Descent Update:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \{\eta\langle g_t, x\rangle + \eta\varphi(x) + B_{\psi_t}(x, x_t)\}$$

ADAGRAD Updates

- First Order Approximation Updates:

$$x_{t+1} = x_t - \eta g_t$$

- ADAGRAD Updates

$$x_{t+1} = x_t - \eta H^{-1} g_t$$

- Capturing the geometry
- Different learning rate for different Features
- Similar to second order approximation

Regret Bound - Corollary 1

ADAGRAD with Composite Mirror Descent Update:

- $\sup_{x,y\in\mathcal{X}} \| x-y \|_\infty \leq D_\infty$
- $\sup_{x,y\in\mathcal{X}} \| x-y \|_2 \leq D_2$
- $\eta = D_\infty/\sqrt{2}$

$$R(T) \leq \sqrt{2d}D_\infty \sqrt{\inf_{s\succcurlyeq 0,\langle 1,s\rangle\leq d} \sum_{t=1}^{T} \| g_t \|^2_{diag(s)^{-1}}} = \sqrt{2}D_\infty \sum_{i=1}^{d} \| g_{1:T,i} \|_2$$

Earlier:

$$R(T) \leq \sqrt{2}D_2 \sqrt{\sum_{t=1}^{T} \| g_t \|_2^2}$$

Proof

Composite Mirror Descent update:

$$R(T) \leq \boxed{\frac{1}{\eta}\sum_{t=1}^{T-1}\left[B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1})\right]}$$

$$\boxed{\frac{1}{\eta}B_{\psi_1}(x^*, x_1)} + \boxed{\frac{\eta}{2}\sum_{t=1}^{T} \| f_t'(x_t) \|_{\psi_t^*}^2}$$

$$\psi_t(x) = \langle x, diag(s_t)x \rangle$$

$$B_\psi(x, y) = \psi(x) - \psi(y) - \langle \bigtriangledown\psi(y), x - y \rangle$$

Proof - Cont.

$$B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1}) = \frac{1}{2}\langle x^* - x_{t+1}, diag(s_{t+1} - s_t)(x^* - x_{t+1})\rangle$$
$$\leq \frac{1}{2} \parallel x^* - x_{t+1} \parallel_\infty^2 \parallel s_{t+1} - s_t \parallel_1$$

$$\frac{1}{\eta}\sum_{t=1}^{T-1}\left[B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1})\right] \leq \frac{1}{2\eta}D_\infty \sum_{i=1}^{d} \parallel g_{1:T,i} \parallel_2 - \frac{1}{2\eta} \parallel x^* - x_1 \parallel_\infty^2 \langle 1, s_1\rangle$$

Proof - Cont.

$$B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1}) = \frac{1}{2}\langle x^* - x_{t+1}, diag(s_{t+1} - s_t)(x^* - x_{t+1})\rangle$$
$$\leq \frac{1}{2} \parallel x^* - x_{t+1} \parallel_\infty^2 \parallel s_{t+1} - s_t \parallel_1$$

$$\frac{1}{\eta}\sum_{t=1}^{T-1}\left[B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1})\right] \leq \frac{1}{2\eta}D_\infty \sum_{i=1}^{d} \parallel g_{1:T,i} \parallel_2 - \frac{1}{2\eta} \parallel x^* - x_1 \parallel_\infty^2 \langle 1, s_1\rangle$$

$$\frac{1}{\eta}B_{\psi_1}(x^*, x_1) \leq \frac{1}{2\eta} \parallel x^* - x_1 \parallel_\infty^2 \langle 1, s_1\rangle$$

Proof - Cont.

$$B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1}) = \frac{1}{2}\langle x^* - x_{t+1}, diag(s_{t+1} - s_t)(x^* - x_{t+1})\rangle$$
$$\leq \frac{1}{2} \parallel x^* - x_{t+1} \parallel_\infty^2 \parallel s_{t+1} - s_t \parallel_1$$

$$\frac{1}{\eta}\sum_{t=1}^{T-1}\left[B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1})\right] \leq \frac{1}{2\eta}D_\infty \sum_{i=1}^{d} \parallel g_{1:T,i} \parallel_2 - \frac{1}{2\eta} \parallel x^* - x_1 \parallel_\infty^2 \langle 1, s_1\rangle$$

$$\frac{1}{\eta}B_{\psi_1}(x^*, x_1) \leq \frac{1}{2\eta} \parallel x^* - x_1 \parallel_\infty^2 \langle 1, s_1\rangle$$

$$\frac{\eta}{2}\sum_{t=1}^{T} \parallel f_t'(x_t) \parallel_{\psi_t^*}^2 \leq \eta \sum_{i=1}^{d} \parallel g_{1:T,i} \parallel_2$$ (Proof by Induction)

Regret Bound - Corollary 1

ADAGRAD with Composite Mirror Descent Update:

- $\sup_{x,y\in\mathcal{X}} \| x-y \|_\infty \leq D_\infty$
- $\sup_{x,y\in\mathcal{X}} \| x-y \|_2 \leq D_2$
- $\eta = D_\infty/\sqrt{2}$

$$R(T) \leq \sqrt{2d}D_\infty \sqrt{\inf_{s\succcurlyeq 0,\langle 1,s\rangle\leq d} \sum_{t=1}^{T} \| g_t \|^2_{diag(s)^{-1}}} = \sqrt{2}D_\infty \sum_{i=1}^{d} \| g_{1:T,i} \|_2$$

Earlier:

$$R(T) \leq \sqrt{2}D_2 \sqrt{\sum_{t=1}^{T} \| g_t \|_2^2}$$

Example - Support Vector Machine

- Hing Loss: $f_t(x) = [1 - y_t \langle z_t, x \rangle]_+$
- Sparse Random Data: $z_t \in \{-1, 0, 1\}^d$
- $z_{t,i} \neq 0$ with probability $\propto i^{-\alpha}$ for $\alpha > 1$
- $\mathcal{X} = \{x : \| x \|_\infty \leq 1\}$ $D_\infty = 2$ $D_2 = 2\sqrt{d}$
- $\| g_t \|_2^2 \geq 1$

Example - Support Vector Machine

- Hing Loss: $f_t(x) = [1 - y_t \langle z_t, x \rangle]_+$
- Sparse Random Data: $z_t \in \{-1, 0, 1\}^d$
- $z_{t,i} \neq 0$ with probability $\propto i^{-\alpha}$ for $\alpha > 1$
- $\mathcal{X} = \{x : \| x \|_\infty \leq 1\}$ $\quad D_\infty = 2$ $\quad D_2 = 2\sqrt{d}$
- $\| g_t \|_2^2 \geq 1$

- Online Gradient Descent: $O(\sqrt{d}\sqrt{T})$

Example - Support Vector Machine

- Hing Loss: $f_t(x) = [1 - y_t \langle z_t, x \rangle]_+$
- Sparse Random Data: $z_t \in \{-1, 0, 1\}^d$
- $z_{t,i} \neq 0$ with probability $\propto i^{-\alpha}$ for $\alpha > 1$
- $\mathcal{X} = \{x : \| x \|_\infty \leq 1\}$ $\quad D_\infty = 2$ $\quad D_2 = 2\sqrt{d}$
- $\| g_t \|_2^2 \geq 1$

- Online Gradient Descent: $O(\sqrt{d}\sqrt{T})$
- ADAGRAD: $O(2max\{\log d, d^{1-\alpha/2}\}\sqrt{T})$

Example - Support Vector Machine

- Hing Loss: $f_t(x) = [1 - y_t \langle z_t, x \rangle]_+$
- Sparse Random Data: $z_t \in \{-1, 0, 1\}^d$
- $z_{t,i} \neq 0$ with probability $\propto i^{-\alpha}$ for $\alpha > 1$
- $\mathcal{X} = \{x : \| x \|_\infty \leq 1\}$ $\quad D_\infty = 2$ $\quad D_2 = 2\sqrt{d}$
- $\| g_t \|_2^2 \geq 1$

- Online Gradient Descent: $O(\sqrt{d}\sqrt{T})$
- ADAGRAD: $O(2max\{\log d, d^{1-\alpha/2}\}\sqrt{T})$

$$\mathbb{E}\sum_{i=1}^{d} \| g_{1:T,i} \|_2 = \sum_{i=1}^{d} \mathbb{E}\sqrt{|\{t : |g_{t,i}| = 1\}|} \leq \sum_{i=1}^{d} \sqrt{\mathbb{E}|\{t : |g_{t,i}| = 1\}|} = \sum_{i=1}^{d} \sqrt{p_i T}$$

$$c\sum_{i=1}^{d} i^{-\alpha/2} = \begin{cases} O(\log d) & \text{if } \alpha \geq 2; \\ O(d^{1-\alpha/2}) & \text{if } \alpha \in (1, 2). \end{cases}$$

ℓ_1-regularization

- $\varphi(x) = \lambda \parallel x \parallel_1$
- $H_{t,ii} = \delta + \parallel g_{1:t,i} \parallel_2$
- ADAGRAD with Primal-Dual sub-Gradient update:

$$x_{t+1,i} = \text{sign}(-\bar{g}_{t,i}) \frac{\eta t}{H_{t,ii}} [|\bar{g}_{t,i}| - \lambda]_+$$

- Standard primal-dual sub-Gradient update:

$$x_{t+1,i} = \text{sign}(-\bar{g}_{t,i}) \eta \sqrt{t} [|\bar{g}_{t,i}| - \lambda]_+$$

ℓ_1-regularization

- $\varphi(x) = \lambda \parallel x \parallel_1$
- $H_{t,ii} = \delta + \parallel g_{1:t,i} \parallel_2$
- ADAGRAD with Primal-Dual sub-Gradient update:

$$x_{t+1,i} = \text{sign}(-\bar{g}_{t,i}) \frac{\eta t}{H_{t,ii}} [|\bar{g}_{t,i}| - \lambda]_+$$

- Standard primal-dual sub-Gradient update:

$$x_{t+1,i} = \text{sign}(-\bar{g}_{t,i}) \eta \sqrt{t} [|\bar{g}_{t,i}| - \lambda]_+$$

- ADAGRAD with Composite Mirror Descent Update:

$$x_{t+1,i} = \text{sign}\left(x_{t,i} - \frac{\eta}{H_{t,ii}} g_{t,i}\right) \left[\left|x_{t,i} - \frac{\eta}{H_{t,ii}} g_{t,i}\right| - \frac{\lambda\eta}{H_{t,ii}}\right]_+$$

ℓ_1-ball projection

- $\mathcal{X} = \{x : \| x \|_1 \leq c\}$
- $x_{t+1} = \arg\min_{x \in \mathcal{X}} \{\langle x, u \rangle\} + \varphi(x) + \frac{1}{2}\langle x, H_t x \rangle\}$
- $u = \eta t \bar{g}_t$
- Assume $\varphi(x) = 0$
- The problem is equivalent to

$$\min_z \| z + H^{-1/2} u \|_2^2 \qquad s.t. \quad \| Az \|_1 \leq c$$

- $z = H^{1/2} x$ and $A = H^{-1/2}$
- $\nu = -H^{-1/2} u$

$$z_i^* = \begin{cases} \nu_i - \theta^* a_i & \text{if } \nu_i \geq \theta^* a_i; \\ 0 & \text{O.W..} \end{cases}$$

- for some $\theta^* \geq 0$
- $O(d \log d)$

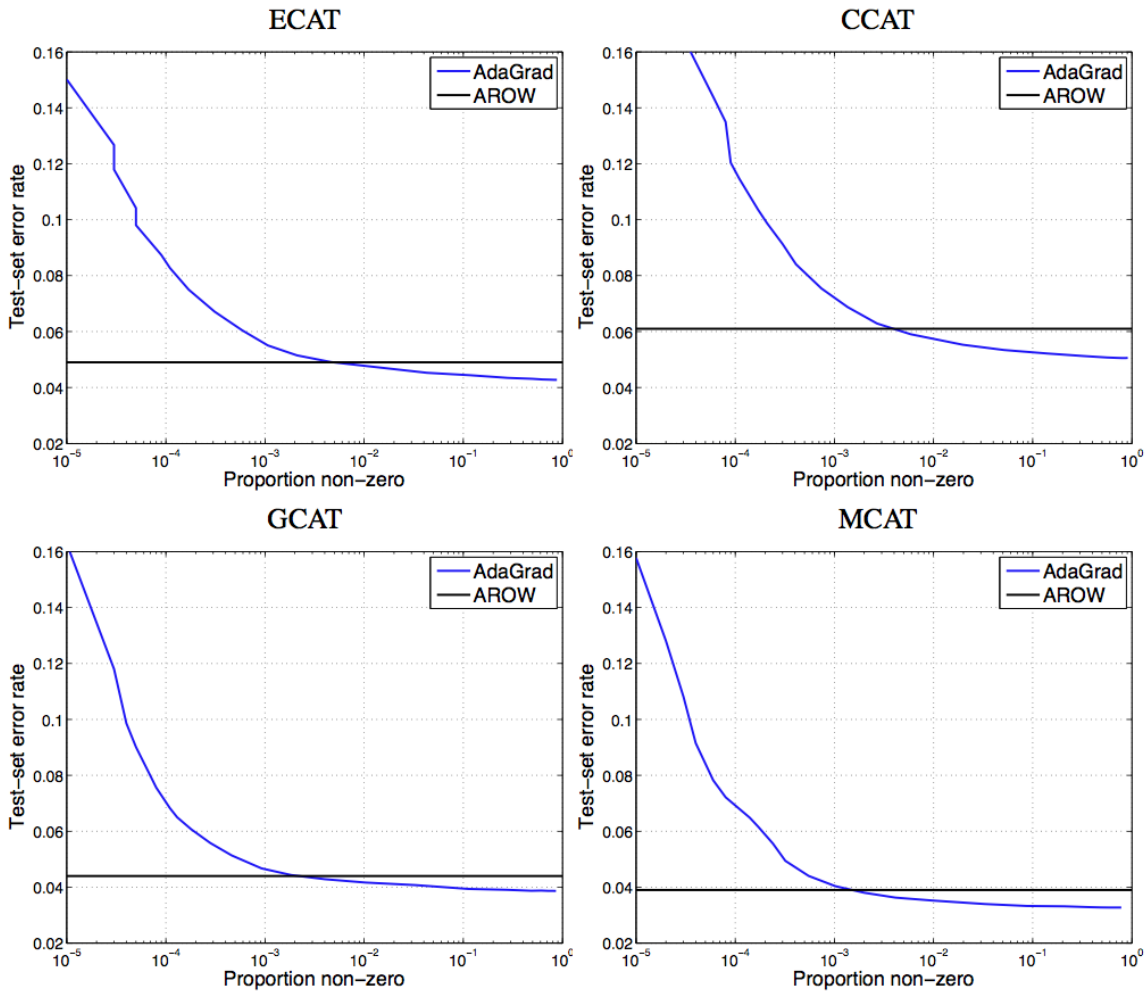
Text Classification

- Reuters RCV1 data set
- 800,000 text article
- 2M features $\in \{0, 1\}$
- < 5000 non-zero features per article

	FOBOS	ADAGRAD	PA	AROW
Economics	.058(.194)	**.044(.086)**	.059	.049
Commerce	.111(**.226**)	**.053** (.276)	.107	.061
Government	.056(**.183**)	**.040**(.225)	.066	.044
Medicine	.056(**.146**)	**.034**(.176)	.053	.039

Sparsity Accuracy Tradeoff

- Varying λ from 10^{-8} to 10^{-1}



ADAGRAD with Full Matrices

- Minimizing regret by:

$$\min_{\substack{A \succcurlyeq 0 \\ tr(A) \leq c}} \sum_{t=1}^{T} \langle g_t, A^{-1} g_t \rangle$$

- Lagrangian of the problem:

$$\mathcal{L}(A, Z, \theta) = \sum_{t=1}^{T} \langle g_t, A^{-1} g_t \rangle - tr(AZ) + \theta(tr(A) - c)$$

- If $G_T = \sum_{\tau=1}^{T} g_\tau g_\tau^T$ is full rank

$$A = \frac{c G_T^{1/2}}{tr(G_T^{1/2})}$$

ADAGRAD with Full Matrices

- Update $G_t = G_{t-1} + g_t g_t^T$
- $S_t = G_t^{\frac{1}{2}}$
- Set $H_t = \delta I + S_t$
- Set $\psi_t(x) = \frac{1}{2}\langle x, H_t x\rangle$
- Primal-Dual Sub Gradient Update:

$$x_{t+1} = \arg\min_{x\in\mathcal{X}}\{\eta\left\langle \frac{1}{t}\sum_{\tau=1}^{t} g_\tau, x\right\rangle + \eta\varphi(x) + \frac{1}{t}\psi_t(x)\} \tag{7}$$

- Composite Mirror Descent Update:

$$x_{t+1} = \arg\min_{x\in\mathcal{X}}\{\eta\langle g_t, x\rangle + \eta\varphi(x) + B_{\psi_t}(x, x_t)\} \tag{8}$$

Regret Bound - Corollary 11

- $\varphi(x_1) = 0$
- $\mathcal{X}$: Compact Set such that $\sup_{x \in \mathcal{X}} \| x - x^* \|_2 \leq D$
- $\eta = D/\sqrt{2}$, $\delta = 0$
- Composite Mirror Descent Update:

$$R_\phi(T) \leq \sqrt{2} D tr(G_T^{1/2}) = \sqrt{2d} D \sqrt{\inf_s \left\{ \sum_{t=1}^{T} g_t^T S^{-1} g_t : S \succcurlyeq 0, tr(S) \leq d \right\}}$$