Composite Objective Optimization and Learning for Large Datasets

John Duchi[1,2] Yoram Singer[2]

[1]University of California, Berkeley

[2]Google Research

Workshop on Massive Modern Datasets, June 15, 2010

Acknowledgments

Elad Hazan

Samy Bengio

Adam Sadovsky

Ambuj Tewari

Shai Shalev-Shwartz

Outline

Online Convex Optimization

Framework for Forward-Backward Splitting

Derived Algorithms and Experiments

Sparse Data: the Need for Adaptivity $\Rightarrow$ AdaGrad

AdaGrad Applications and Experiments

Conclusions

A Brief Review of Online Convex Optimization

Online learning task—repeat:

- Learner plays point x_t
- Receive function f_t
- Suffer loss $f_t(x_t)$

- Weight vector for features
- Receive label y_t, features ϕ_t
- Hinge loss $[1 - y_t \langle \phi_t, x_t \rangle]_+$

Goal: Attain small regret

$$R(T) := \sum_{t=1}^{T} f_t(x_t) - \inf_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x)$$

A Brief Review of Online Convex Optimization

Online learning task—repeat:

- Learner plays point x_t
- Receive function f_t
- Suffer loss $f_t(x_t)$

- Weight vector for features
- Receive label y_t, features ϕ_t
- Hinge loss $[1 - y_t \langle \phi_t, x_t \rangle]_+$

Goal: Attain small regret

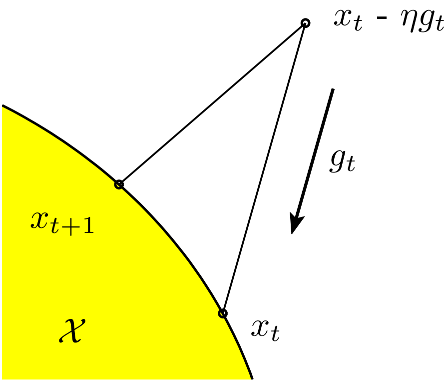
$$R(T) := \sum_{t=1}^{T} f_t(x_t) - \inf_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x)$$

NB: Also works for fixed f, random f_t for stochastic optimization

Common Approach: Online Gradient Descent

Let $g_t = \nabla f_t(x_t)$

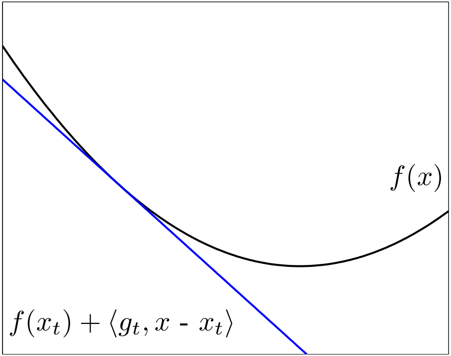
$$x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta_t g_t) = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \left\| x - (x_t - \eta_t g_t) \right\|^2 \right\}$$



Online Gradient Descent: Alternative View

Let $g_t = \nabla f_t(x_t)$. Then

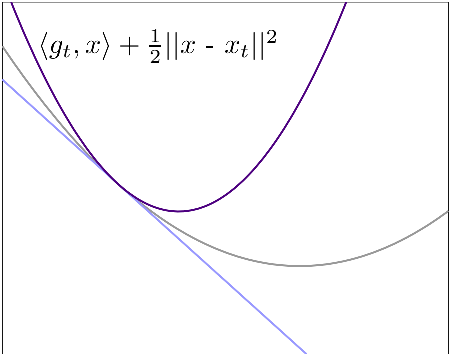
$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|^2 + \eta_t \langle g_t, x \rangle \right\}$$
$$= \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - (x_t - \eta_t g_t)\|^2 \right\}$$



Online Gradient Descent: Alternative View

Let $g_t = \nabla f_t(x_t)$. Then

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|^2 + \eta_t \langle g_t, x \rangle \right\}$$
$$= \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - (x_t - \eta_t g_t)\|^2 \right\}$$



Analysis Idea

- Almost contraction

$$\frac{1}{2}\|x_{t+1} - x^*\|^2 \leq \frac{1}{2}\|x_t - x^*\|^2 + \eta(f_t(x^*) - f_t(x_t)) + \frac{\eta^2}{2}\|g_t\|^2$$

- Sum

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|^2$$

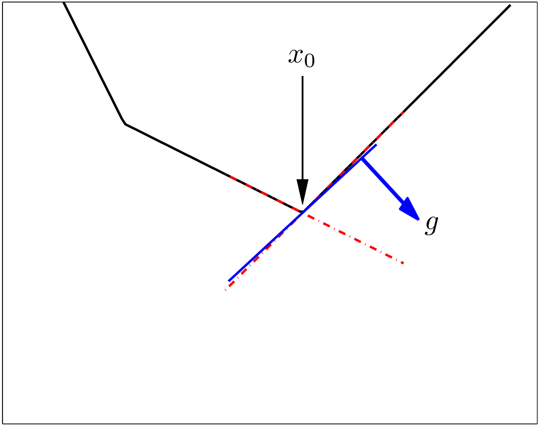
- Set $\eta \propto 1/\sqrt{T}$

$$R(T) = O(\sqrt{T})$$

Subgradients

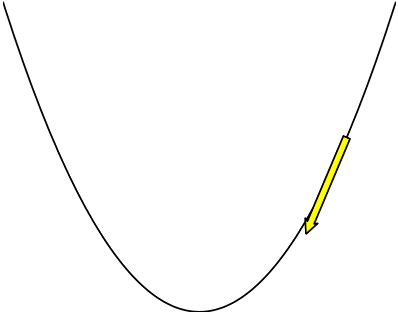
- Subgradient set of a function f

$$\partial f(x_0) = \left\{ g \in \mathbb{R}^d \;\middle|\; f(x) \geq f(x_0) + \langle g, x - x_0 \rangle \right\}$$

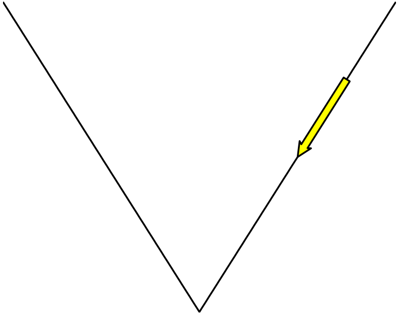


Problems with Subgradient Methods

- Subgradient set is large at singularities



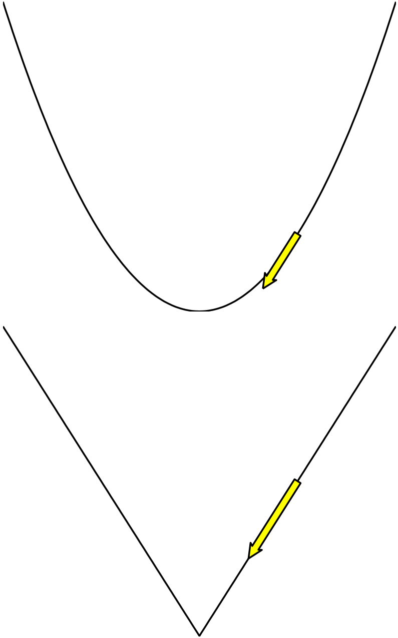
- Subgradients are non-informative at singularities



Problems with Subgradient Methods

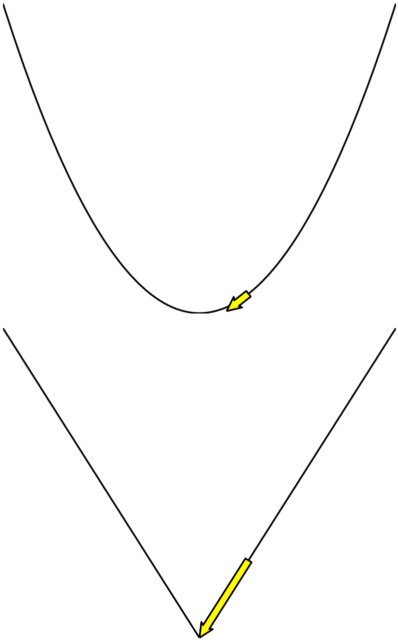
- Subgradient set is large at singularities

- Subgradients are non-informative at singularities



Problems with Subgradient Methods

- Subgradient set is large at singularities
- Subgradients are non-informative at singularities



Structured Regret

Goal: Attain small regret

$$\sum_{t=1}^{T} f_t(x_t) \qquad - \inf_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x)$$

Structured Regret

Goal: Attain small regret

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

- Idea: Exploit structure of known φ

The Fobos Algorithm

Goal: Attain small structured regret

$$R(T) = \sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - \inf_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x) + \varphi(x)$$

The Fobos Algorithm

Goal: Attain small structured regret

$$R(T) = \sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - \inf_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x) + \varphi(x)$$

- Repeat
  I. Unconstrained (stochastic sub) gradient of loss
  II. Incorporate regularization

The Fobos Algorithm

Goal: Attain small structured regret

$$R(T) = \sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - \inf_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x) + \varphi(x)$$

- Repeat
  I. Unconstrained (stochastic sub) gradient of loss
  II. Incorporate regularization
- Similar to forward-backward splitting (Lions and Mercier 79), composite gradient methods (Wright et al. 09, Nesterov 07), dual averaging with regularization (Xiao 09)

Fobos Algorithm

$$R(T) = \sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - \inf_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x) + \varphi(x)$$

- **Earlier:**

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|^2 + \eta_t \langle g_t, x \rangle \right\}$$

Fobos Algorithm

$$R(T) = \sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - \inf_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x) + \varphi(x)$$

- **Earlier:**

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|^2 + \eta_t \langle g_t, x \rangle \right\}$$

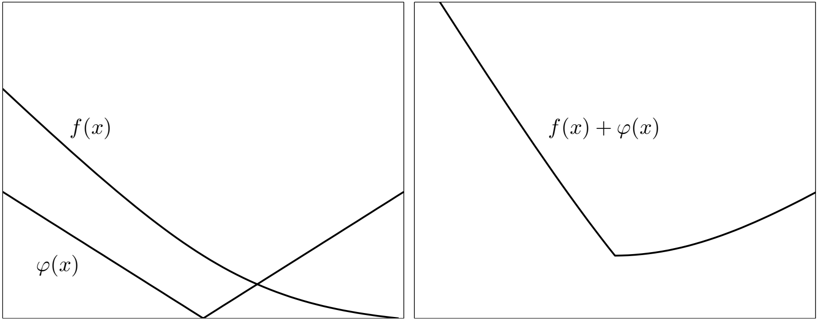
- **Now:**

$$\begin{aligned} x_{t+1} &= \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|^2 + \eta_t \langle g_t, x \rangle + \eta_t \varphi(x) \right\} \\ &= \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - \underbrace{(x_t - \eta_t g_t)}_{\text{Subgradient}} \|^2 + \underbrace{\eta_t \varphi(x)}_{\text{Regularizer}} \right\} \end{aligned}$$

Fobos Algorithm

- Unconstrained gradient $\mathbb{E}g_t \in \partial f(x_t)$ and regularization φ.

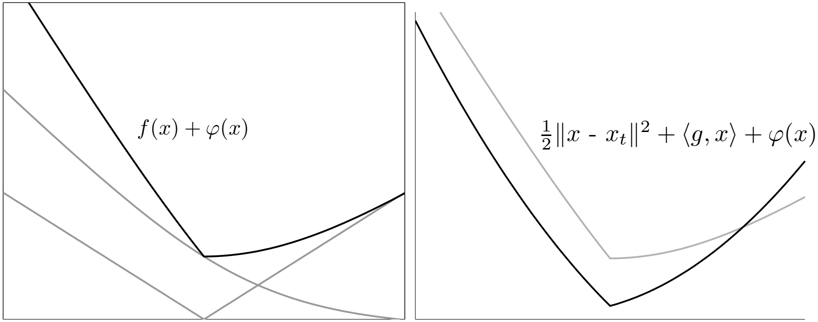
$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|^2 + \eta_t \langle g_t, x \rangle + \eta_t \varphi(x) \right\}$$



Fobos Algorithm

- Unconstrained gradient $\mathbb{E}g_t \in \partial f(x_t)$ and regularization φ.

$$x_{t+1} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|x - x_t\|^2 + \eta_t \langle g_t, x \rangle + \eta_t \varphi(x) \right\}$$



Forward Looking Property

- The optimum x_{t+1} satisfies

$$0 \in x_{t+1} - x_t + \eta_t \partial f_t(x_t) + \eta_t \partial \varphi(x_{t+1})$$

- Pick $g_t^f \in \partial f_t(x_t)$ and $g_{t+1}^{\varphi} \in \partial \varphi(x_{t+1})$

$$x_{t+1} = x_t - \eta_t g_t^f - \eta_t g_{t+1}^{\varphi}$$

$\nearrow$ current loss $\nwarrow$ forward regularization

- *Current* subgradient of loss, *forward* subgradient of regularization

Analysis (same "Contraction" property)

- Before, use $x_{t+1} = x_t - \eta_t g_t$:

$$\frac{1}{2}\|x_{t+1} - x^*\|^2 \le \frac{1}{2}\|x_t - x^*\|^2 + \eta_t\left(f_t(x^*) - f_t(x_t)\right) + \frac{\eta_t^2}{2}\|g_t\|^2$$

- Now, use $x_{t+1} = x_t - \eta_t g_t^f - \eta_t g_{t+1}^\varphi$

$$\frac{1}{2}\|x_{t+1} - x^*\|^2 \le \frac{1}{2}\|x_t - x^*\|^2 + \eta_t\left(f_t(x^*) - f_t(x_t)\right) + \frac{\eta_t^2}{2}\|g_t\|^2 + \eta_t\left(\varphi(x^*) - \varphi(x_{t+1})\right)$$

Stochastic Convergence and Online Regret

- Online (average) regret bounds

$$\mathrm{AvgReg}(T) := \frac{1}{T}\left[\sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - \sum_{t=1}^{T} f_t(x^*) + \varphi(x^*)\right]$$

$$\eta_t \propto \frac{1}{\sqrt{t}} \quad \Rightarrow \quad \mathrm{AvgReg}(T) = O\left(\frac{1}{\sqrt{T}}\right)$$

Stochastic Convergence and Online Regret

- Online (average) regret bounds

$$\mathrm{AvgReg}(T) := \frac{1}{T}\left[\sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - \sum_{t=1}^{T} f_t(x^*) + \varphi(x^*)\right]$$

$$\eta_t \propto \frac{1}{\sqrt{t}} \quad \Rightarrow \quad \mathrm{AvgReg}(T) = O\left(\frac{1}{\sqrt{T}}\right)$$

- Stochastic: when $\mathbb{E} g_t \in \partial f(x_t)$, w.h.p.,

$$f(x_T) + \varphi(x_T) - (f(x^*) + \varphi(x^*)) = O\left(\frac{1}{\sqrt{T}}\right)$$

Derived Algorithms

Break FOBOS update into two parts:

- Step I (unconstrained gradient)

$$x_{t+\frac{1}{2}} = x_t - \eta_t g_t$$

- Step II (correct and project)

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \left\| x - x_{t+\frac{1}{2}} \right\|^2 + \eta_t \varphi(x) \right\}$$

Derived Algorithms

We show step II for

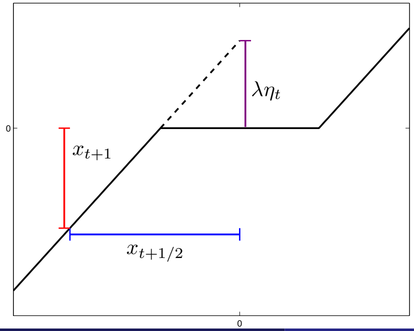
- FOBOS with ℓ_1-regularization
- FOBOS with ℓ_2-regularization
- FOBOS with mixed norms (ℓ_1/ℓ_2 or ℓ_1/ℓ_∞)

FOBOS with ℓ_1

$$\min \; \frac{1}{2}\left\|x - x_{t+\frac{1}{2}}\right\|^2 + \lambda \left\|x\right\|_1$$

- Separable: minimize $\frac{1}{2}\left(x - x_{t+\frac{1}{2},j}\right)^2 + \lambda|x|$.
- Coordinate-wise update yields sparsity:

$$x_{t+1,j} = \operatorname{sign}\left(x_{t+\frac{1}{2},j}\right) \max\left\{\left|x_{t+\frac{1}{2},j}\right| - \lambda\eta_t, 0\right\}$$



Truncated gradient
(Langford et al. 08)
Iterative shrinkage and thresholding
(Donoho 95, Daubechies et al. 04)

FOBOS with ℓ_2

- When $\varphi(x) = \frac{\lambda}{2} \|x\|_2^2$, gradient descent & geometric shrinkage

$$x_{t+1} = \frac{x_{t+\frac{1}{2}}}{1 + \lambda\eta_t} = \frac{x_t - \eta_t g_t}{1 + \lambda\eta_t}$$

- When $\varphi(x) = \lambda \|x\|_2$, all or nothing update

$$x_{t+1} = \left[1 - \frac{\lambda\eta_t}{\left\|x_{t+\frac{1}{2}}\right\|_2}\right]_+ x_{t+\frac{1}{2}}$$

FOBOS with mixed norms

$$\varphi(X) = \|X\|_{\ell_1/\ell_q} = \sum_{j=1}^{d} \|\overline{x}_j\|_q$$

$$X = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_d \end{bmatrix} \quad \Rightarrow \quad \begin{matrix} \|\overline{x}_1\|_q \\ \|\overline{x}_2\|_q \\ \vdots \\ \|\overline{x}_d\|_q \end{matrix}$$

- Separable and solvable using previous methods
- Multitask and multiclass learning
  - $\overline{x}_j$ associated with feature j
  - Penalize $\overline{x}_j$ once

Sparse Gradients

	g				
$t = 1$	[	1	3	0	]
$t = 2$	[	2	**0**	1	]
$t = 3$	[	1	**0**	5	]
$t = 4$	[	1	**0**	2	]
$t = 5$	[	3	**0**	2	]

High Dimensional Efficiency

- Input space is sparse but of very high dimension
- Want update to scale with number of *present* features
  $\Rightarrow$ just in time updates

High Dimensional Efficiency

- Input space is sparse but of very high dimension
- Want update to scale with number of *present* features
  $\Rightarrow$ just in time updates
- **Proposition:** The following are equivalent:

$$x_t = \operatorname*{argmin}_x \|x - x_{t-1}\|^2 + \eta_t \lambda \|x\|_q \quad \text{for } t = 1 \text{ to } T$$

$$x_T = \operatorname*{argmin}_x \|x - x_0\|^2 + \left(\sum_{t=1}^{T-1} \eta_t \lambda \right) \|x\|_q$$

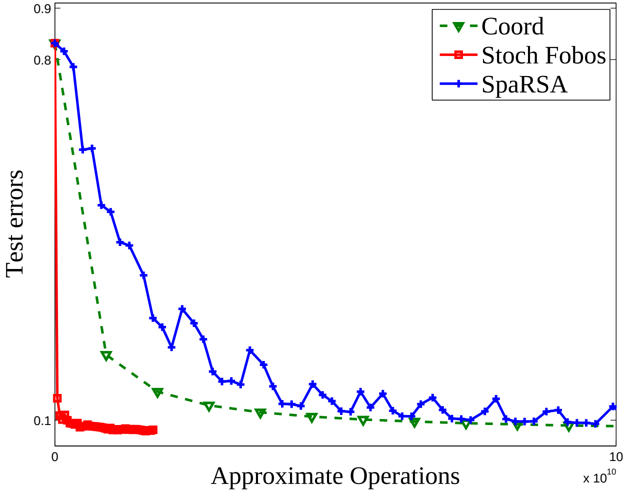
High dimensional update

	g			
$t = 1$	[1	3	0]	
$t = 2$	[2	**0**	.5]	Skip
$t = 3$	[1	**0**	.5]	update
$t = 4$	[.1	**0**	-.25]	(lazy
$t = 5$	[-.5	**0**	.25]	eval)
$t = 6$	[2	1	1]	

- At $t = 6$, FOBOS update with $\lambda = \sum_{t=2}^{6} \lambda_t$

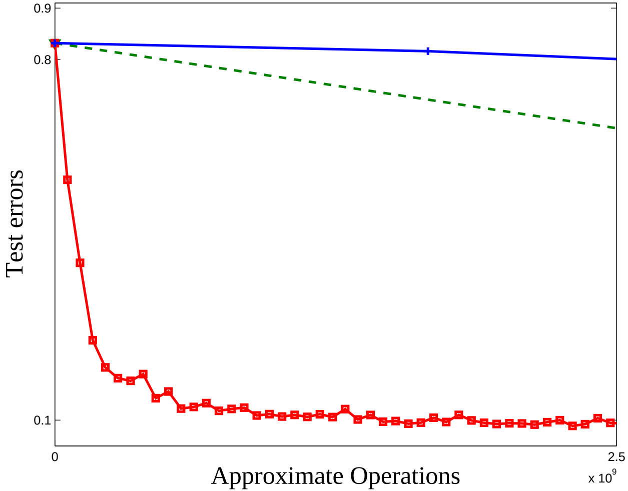
Brief Experimental Results

MNIST experiments



Comparison of test error rate of FOBOS, Sparsa (Wright et al. 2009), coordinate descent (Tseng 2007).

MNIST experiments



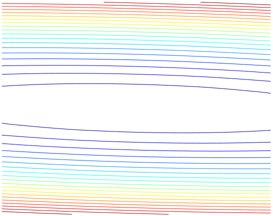
Comparison of test error rate of FOBOS, Sparsa (Wright et al. 2009), coordinate descent (Tseng 2007).

Characteristics of sparse data

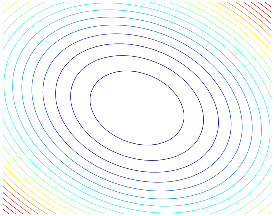
> The most unsung birthday in American business and technological history this year may be the 50th anniversary of the Xerox 914 photocopier.[a]
>
> [a] *The Atlantic*, July/August 2010.

- Some words very infrequent but very predictive

Why adapt to geometry?



Hard



Nice

y_t	$g_{t,1}$	$g_{t,2}$	$g_{t,3}$
1	1	0	0
-1	.5	0	1
1	-.5	1	0
-1	0	0	0
1	.5	0	0
-1	1	0	0
1	-1	1	0
-1	-.5	0	1

1. Frequent, irrelevant
2. Infrequent, predictive
3. Infrequent, predictive

Adapting to Geometry of the Space

- Receive $g_t \in \partial f_t(x_t)$
- Earlier:

$$x_{t+1} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|x - x_t\|^2 + \eta \langle g_t, x \rangle + \varphi(x) \right\}$$

- Now: let $\|x\|_A^2 = \langle x, Ax \rangle$ for $A \succeq 0$. Use

$$x_{t+1} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|x - x_t\|_A^2 + \eta \langle g_t, x \rangle + \varphi(x) \right\}$$

Analysis Idea—Almost Contraction

- Have $g_t \in \partial f_t(x_t)$ (ignore φ for simplicity)
- Before: $x_{t+1} = x_t - \eta g_t$

$$\frac{1}{2}\left\|x_{t+1} - x^*\right\|_2^2 \leq \frac{1}{2}\left\|x_t - x^*\right\|_2^2 + \eta\left(f_t(x^*) - f_t(x_t)\right) + \frac{\eta^2}{2}\left\|g_t\right\|_2^2$$

- Now: $x_{t+1} = x_t - \eta A^{-1} g_t$

$$\begin{aligned}&\frac{1}{2}\left\|x_{t+1} - x^*\right\|_A^2 \\ &\leq \frac{1}{2}\left\|x_t - x^*\right\|_A^2 + \eta\left(f_t(x^*) - f_t(x_t)\right) + \frac{\eta^2}{2}\left\|g_t\right\|_{A^{-1}}^2\end{aligned}$$

$\uparrow$ dual norm to $\|\cdot\|_A$

Meta Learning Problem

- Immediately get regret:

$$\sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - f_t(x^*) - \varphi(x^*)$$

$$\leq \frac{1}{\eta} \|x_1 - x^*\|_A^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|_{A^{-1}}^2$$

- What happens if we choose A in hindsight to minimize above?

$$\min_A \sum_{t=1}^{T} \left\langle g_t, A^{-1} g_t \right\rangle \quad \text{subject to} \quad A \succeq 0, \operatorname{tr}(A) \leq C$$

Hindsight minimization

- Focus on diagonal case (full matrix case similar)

$$\min_s \sum_{t=1}^{T} \langle g_t, \mathrm{diag}(s)^{-1} g_t \rangle \quad \text{subject to} \quad s \succeq 0, \langle 1, s \rangle \leq C$$

- $g_{1:T,j} = [g_{1,j} \; g_{2,j} \; \cdots \; g_{T,j}]$ is vector of jth component of g_t
- Solution:

$$s_j \propto \|g_{1:T,j}\|_2$$

y_t	$g_{t,1}$	$g_{t,2}$	$g_{t,3}$
1	1	0	0
-1	1	0	1
1	-1	1	0
-1	0	0	0
1	1	0	0
-1	1	0	0
$C = 3$	$s_1 = 2$	$s_2 = \frac{1}{2}$	$s_3 = \frac{1}{2}$

Low regret to the best A

- At time t, use

$$s_t = \left[\|g_{1:t,j}\|_2\right]_{j=1}^d \quad \text{and} \quad A_t = \operatorname{diag}(s_t)$$

- AdaGrad step

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|_{A_t}^2 + \eta \langle g_t, x \rangle + \eta \varphi(x) \right\}$$

Low regret to the best A

- At time t, use

$$s_t = \left[\|g_{1:t,j}\|_2\right]_{j=1}^{d} \quad \text{and} \quad A_t = \operatorname{diag}(s_t)$$

- AdaGrad step

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|_{A_t}^2 + \eta \langle g_t, x \rangle + \eta \varphi(x) \right\}$$

- Define $D_\infty = \max_t \|x_t - x^*\|_\infty \leq \sup_{x \in \mathcal{X}} \|x - x^*\|_\infty$

$$\sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - f_t(x^*) - \varphi(x^*)$$

$$\leq \sqrt{2d} D_\infty \sqrt{\inf_s \left\{ \sum_{t=1}^{T} \|g_t\|_{\operatorname{diag}(s)^{-1}}^2 \;\middle|\; s \succeq 0, \langle 1, s \rangle \leq d \right\}}$$

AdaGrad with ℓ_1 regularization

$$\min_x \frac{1}{2} \langle x - x_t, \operatorname{diag}(s_t)(x - x_t) \rangle + \lambda \|x\|_1 + \langle g_t, x \rangle$$

- Coordinate-wise update yields sparsity and adaptivity:

$$x_{t+1,j} = \operatorname{sign}\left(x_{t,j} - \frac{g_{t,j}}{s_{t,j}}\right) \left[\left|x_{t,j} - \frac{g_{t,j}}{s_{t,j}}\right| - \frac{\lambda}{s_{t,j}}\right]_+$$

- Earlier coordinate-wise upate:

$$x_{t+1,j} = \operatorname{sign}(x_{t,j} - \eta_t g_{t,j}) \left[|x_{t,j} - \eta_t g_{t,j}| - \eta_t \lambda\right]_+$$

Experimental Results

Text Classification

Reuters RCV1 document classification task—two million features, approximately 4000 non-zero features per document, one online pass

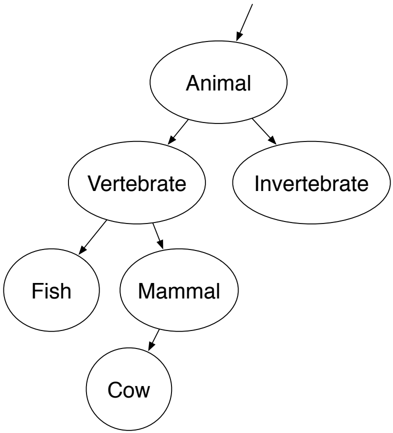
	FOBOS	AdaGrad	PA[1]	AROW[2]
Ecomonics	.058 (.194)	**.044 (.086)**	.059	.049
Corporate	.111 (.226)	**.053 (.105)**	.107	.061
Government	.056 (.183)	**.040 (.080)**	.066	.044
Medicine	.056 (.146)	**.035 (.063)**	.053	.039

Table: Test set classification error rate
(sparsity of final predictor in parenthesis)

[1] Crammer et al., 2006
[2] Crammer et al., 2009

Image Ranking

ImageNet (Deng et al., 2009), large-scale hierarchical image database



Train 15,000 rankers/classifiers to rank images for *each* noun (as in Grangier and Bengio, 2008)

Image Ranking Results

Precision at k: proportion of examples in top k that belong to category. Average precision is average placement of all positive examples. (Variance $\leq 10^{-5}$)

Algorithm	Avg. Prec.	P@1	P@5	P@10	Nonzero
AdaGrad	**0.6022**	0.8502	0.8130	0.7811	**0.7267**
AROW	0.5813	**0.8597**	**0.8165**	**0.7816**	1.0000
PA	0.5581	0.8455	0.7957	0.7576	1.0000
Fobos	0.5042	0.7496	0.6950	0.6545	0.8996

Conclusions and Discussion

- Learning and stochastic optimization with structural assumptions, such as from regularization
- Family of algorithms that adapt to geometry of data. Framework applicable to other algorithms (e.g. regularized dual averaging)
- Efficient algorithms for high-dimensional problems, especially with sparsity

Conclusions and Discussion

- Learning and stochastic optimization with structural assumptions, such as from regularization
- Family of algorithms that adapt to geometry of data. Framework applicable to other algorithms (e.g. regularized dual averaging)
- Efficient algorithms for high-dimensional problems, especially with sparsity
- Future: Put Structural assumptions of problem in regularizer, efficient full-matrix adaptivity, other types of adaptation

Thanks!