Efficient Learning with Forward-Backward Splitting

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

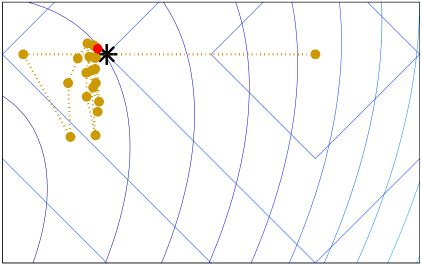
[1]University of California, Berkeley

[2]Google

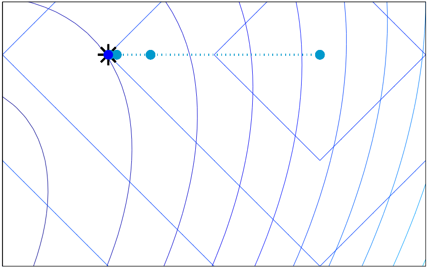
Neural Information Processing Systems, 2009

Motivating Example

Minimize $\frac{1}{2}\boldsymbol{w}^\top A\boldsymbol{w} + \boldsymbol{c}^\top\boldsymbol{w} + \lambda\left\|\boldsymbol{w}\right\|_1$. True solution: $\boldsymbol{w}^* = [-1\ 0]^\top$.



Subgradient

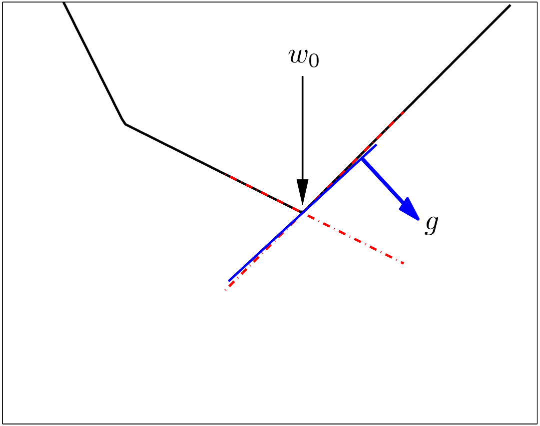


Fobos

Subgradients

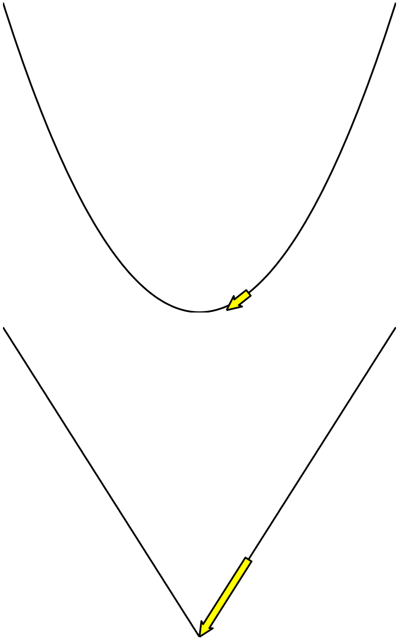
- Subgradient set of a function f

$$\partial f(\boldsymbol{w}_0) = \left\{ \boldsymbol{g} \in \mathbb{R}^n \mid f(\boldsymbol{w}) \geq f(\boldsymbol{w}_0) + \boldsymbol{g}^\top (\boldsymbol{w} - \boldsymbol{w}_0) \right\}$$



What is the problem?

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



Outline

The Fobos Algorithm

Goal: $\min_{\boldsymbol{w}} L(\boldsymbol{w}) + R(\boldsymbol{w}).$

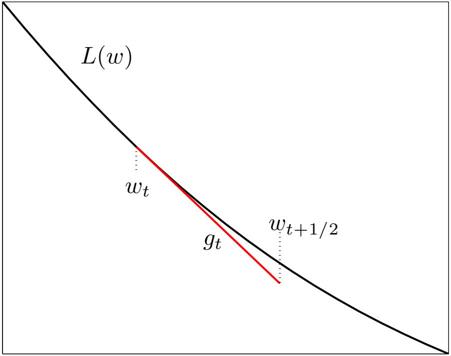
- Repeat
  I. Unconstrained (stochastic sub) gradient of loss
  II. Incorporate regularization
- Similar to forward-backward splitting (Lions and Mercier 79), composite gradient methods (Wright et al. 09, Nesterov 07), dual averaging with regularization (Xiao 09).

Fobos: Step I

$$\textbf{Goal: } \min_{\boldsymbol{w}} L(\boldsymbol{w}) + R(\boldsymbol{w})$$

- Unconstrained (stochastic sub) gradient of loss

$$\boldsymbol{w}_{t+\frac{1}{2}} = \boldsymbol{w}_t - \eta_t \boldsymbol{g}_t \quad \text{where} \quad \mathbb{E}\boldsymbol{g}_t \in \partial L(\boldsymbol{w}_t)$$

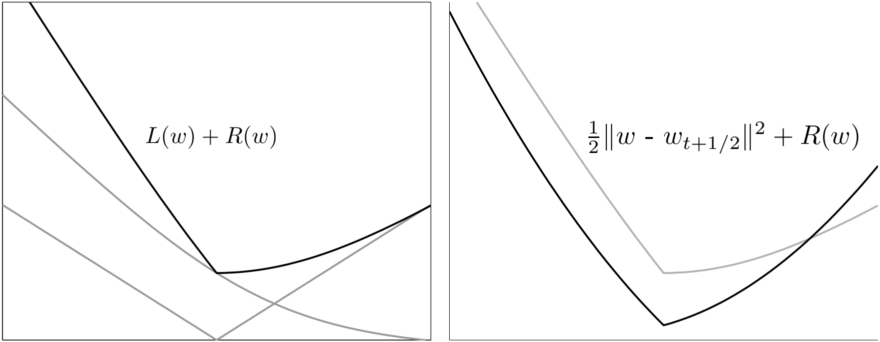


Fobos: Step II

Goal: $\min_{\boldsymbol{w}} L(\boldsymbol{w}) + R(\boldsymbol{w})$

- Incorporate regularization

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \left\{ \frac{1}{2} \left\| \boldsymbol{w} - \boldsymbol{w}_{t+\frac{1}{2}} \right\|^2 + \eta_t R(\boldsymbol{w}) \right\}.$$



Forward Looking Property

- The optimum $\boldsymbol{w}_{t+1}$ satisfies

$$\mathbf{0} \in \boldsymbol{w}_{t+1} - \boldsymbol{w}_t + \eta_t \partial L(\boldsymbol{w}_t) + \eta_t \partial R(\boldsymbol{w}_{t+1})$$

- Pick $\boldsymbol{g}_t^L \in \partial L(\boldsymbol{w}_t)$ and $\boldsymbol{g}_{t+1}^R \in \partial R(\boldsymbol{w}_{t+1})$

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \boldsymbol{g}_t^L - \eta_t \boldsymbol{g}_{t+1}^R$$

current loss   forward regularization

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

Batch Convergence and Online Regret

- Set $\eta_t \propto \frac{1}{\sqrt{T}}$ or $\frac{1}{\sqrt{t}}$ to obtain batch convergence

$$L(\boldsymbol{w}_t) + R(\boldsymbol{w}_t) - (L(\boldsymbol{w}^*) + R(\boldsymbol{w}^*)) = O\left(\frac{1}{\sqrt{T}}\right).$$

- Online (average) regret bounds

$$\mathrm{Regret}(T) \triangleq \frac{1}{T}\left[\sum_{t=1}^{T} L_t(\boldsymbol{w}_t) + R(\boldsymbol{w}_t) - \sum_{t=1}^{T} L_t(\boldsymbol{w}^*) + R(\boldsymbol{w}^*)\right]$$

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

$$\eta_t \propto \frac{1}{t} \quad \Rightarrow \quad \mathrm{Regret}(T) = O\left(\frac{\log T}{T}\right) \quad \text{(strong convexity)}$$

Derived Algorithms

We show step II for

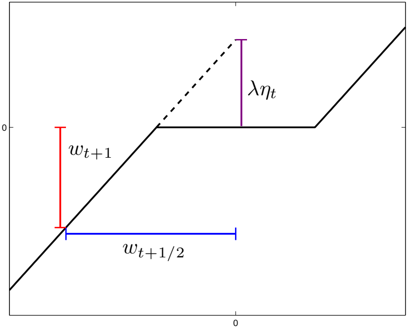
- FOBOS with ℓ_1-regularization
- FOBOS with ℓ_2-regularization
- FOBOS with ℓ_∞-regularization
- FOBOS with mixed norms (ℓ_1/ℓ_2 or ℓ_1/ℓ_∞)

FOBOS with ℓ_1

$$\min \ \frac{1}{2}\left\|\boldsymbol{w} - \boldsymbol{w}_{t+\frac{1}{2}}\right\|^2 + \lambda\left\|\boldsymbol{w}\right\|_1$$

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

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



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

FOBOS with ℓ_2

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

$$w_{t+1} = \frac{w_{t+\frac{1}{2}}}{1 + \lambda\eta_t} = \frac{w_t - \eta_t g_t}{1 + \lambda\eta_t}$$

- When $R(w) = \lambda \|w\|_2$, all or nothing update

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

FOBOS with ℓ_∞

$$\min \frac{1}{2} \left\| \boldsymbol{w} - \boldsymbol{w}_{t+\frac{1}{2}} \right\|^2 + \lambda \left\| \boldsymbol{w} \right\|_\infty$$

- Update is thresholding

$$w_{t+1,j} = \operatorname{sign}(w_{t+1,j}) \min \left\{ \left| w_{t+\frac{1}{2},j} \right|, \theta \right\}$$

- When $\theta = 0$, all zeros
- Dual problem, $\theta = \max_j |w_j - \alpha_j|$ and $\boldsymbol{w}_{t+1} = \boldsymbol{w}_{t+\frac{1}{2}} - \boldsymbol{\alpha}^*$:

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2} \left\| \boldsymbol{\alpha} - \boldsymbol{w}_{t+\frac{1}{2}} \right\|^2 \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_1 \leq \lambda.$$

- Projection onto the ℓ_1-ball (Duchi et al. 2008).

FOBOS with mixed norms

$$r(W) = \|W\|_{\ell_1/\ell_q} = \sum_{j=1}^{d} \|\bar{\boldsymbol{w}}_j\|_q$$

$$W = \begin{bmatrix} \bar{\boldsymbol{w}}_1 \\ \bar{\boldsymbol{w}}_2 \\ \vdots \\ \bar{\boldsymbol{w}}_d \end{bmatrix} \quad \Rightarrow \quad \begin{matrix} \|\bar{\boldsymbol{w}}_1\|_q \\ \|\bar{\boldsymbol{w}}_2\|_q \\ \vdots \\ \|\bar{\boldsymbol{w}}_d\|_q \end{matrix}$$

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

Sparse Gradients

	$\boldsymbol{g}$
$t=1$	[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 huge
- Need to perform lazy updates to $\boldsymbol{w}$
- **Proposition:** The following are equivalent:

$$\boldsymbol{w}_t = \operatorname*{argmin}_{\boldsymbol{w}} \|\boldsymbol{w} - \boldsymbol{w}_{t-1}\|^2 + \eta_t \lambda \|\boldsymbol{w}\|_q \quad \text{for } t = 1 \text{ to } T$$

$$\boldsymbol{w}_T = \operatorname*{argmin}_{\boldsymbol{w}} \|\boldsymbol{w} - \boldsymbol{w}_0\|^2 + \left(\sum_{t=1}^{T-1} \eta_t \lambda\right) \|\boldsymbol{w}\|_q$$

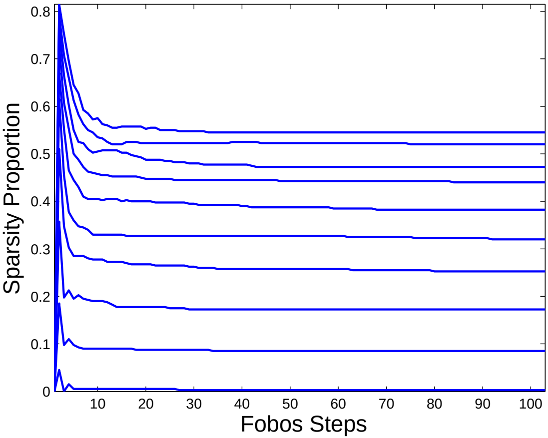
High dimensional update

	$\boldsymbol{g}$		
$t = 1$	[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 $\lambda = \sum_{t=2}^{6} \lambda_t$

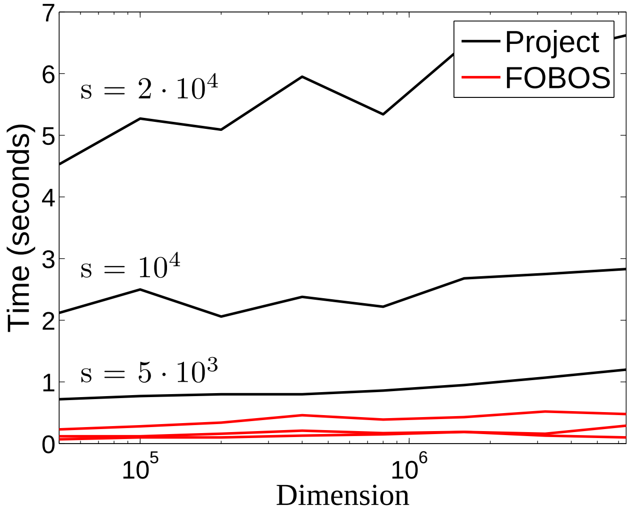
Experimental Results

Sparsity



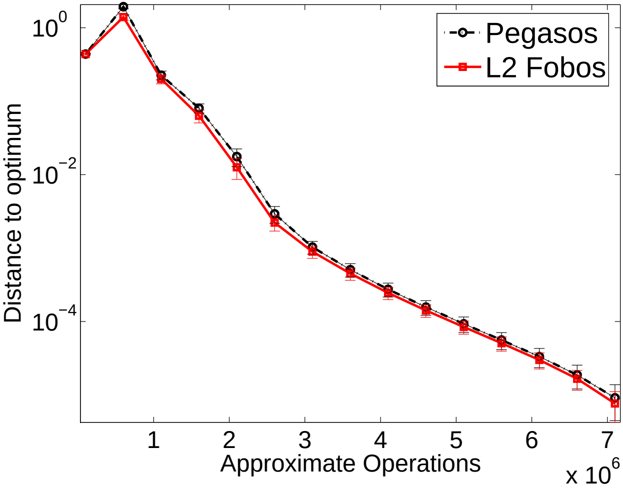
Sparsity as function of FOBOS steps on ℓ_1-regularized logistic regression

Sparse timing experiments



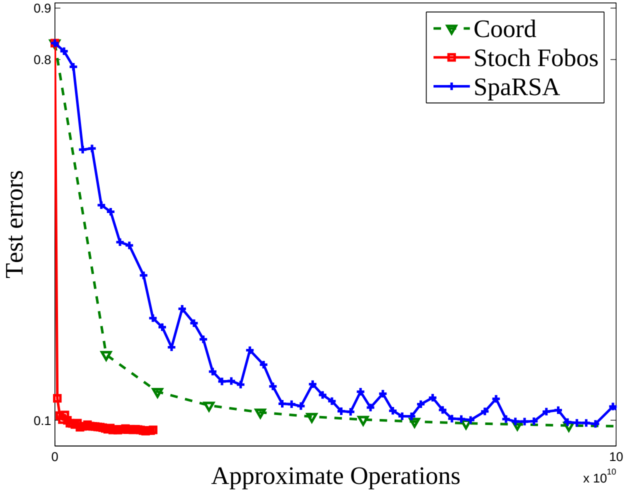
Comparison of ℓ_1-projection to FOBOS lazy update

ℓ_2^2 regularized experiments



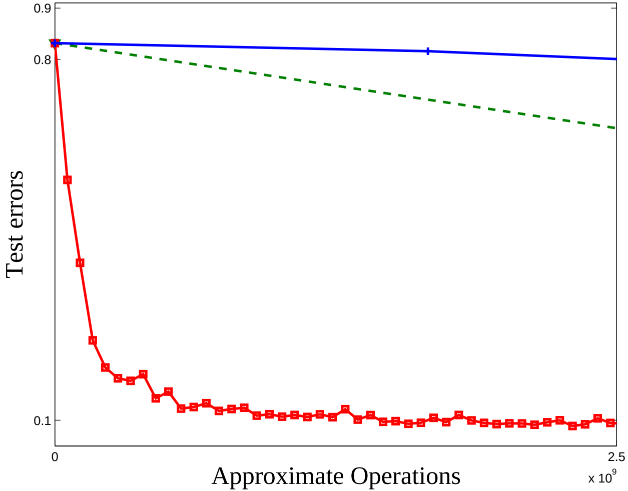
Convergence of FOBOS versus Pegasos on ℓ_2^2 regularized problem

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).

Conclusions

- General framework for stochastic gradient with regularization
- Mixed-norm regularization for multiclass/multitask problems
- Lazy updates for efficency in high dimensions
- Future: Put structural assumptions of problem in regularizer

Thanks!