Boosting with Structural Sparsity

John Duchi* **Yoram Singer**
UC Berkeley **Google**

***Much of this work done at Google**

Outline of talk

- Boosting and coordinate descent
- Losses and mixed norms
- Theory
- Experiments
- Future work and open issues

Coordinate Descent

- Popularized by Friedman, Hastie, Tibshirani
- Add features that do well on a loss function
  $L : \{\mathcal{X} \times \{-1, +1\}\} \times \mathbb{R}^n \rightarrow \mathbb{R}$

$$
\begin{aligned}
(j, \delta) &= \operatorname*{argmin}_{j,\delta} \sum_{i=1}^{m} L(\boldsymbol{x}_i, y_i, \boldsymbol{w} + \delta \boldsymbol{e}_j) \\
\boldsymbol{w}' &= \boldsymbol{w} + \delta \boldsymbol{e}_j
\end{aligned}
$$

What we give

- Family of coordinate descent methods
  - Multiclass, binary, regression
  - Mixed-norm regularization

Loss function

- Multiclass logistic loss

$$L(W) = \sum_{i=1}^{m} \log \left(1 + \sum_{r \neq y_i} \exp \left(\boldsymbol{w}_r \cdot \boldsymbol{x}_i - \boldsymbol{w}_{y_i} \cdot \boldsymbol{x}_i \right) \right)$$

$$W = [\boldsymbol{w}_1 \ \boldsymbol{w}_2 \ \cdots \ \boldsymbol{w}_k]$$

Mixed-norm regularization

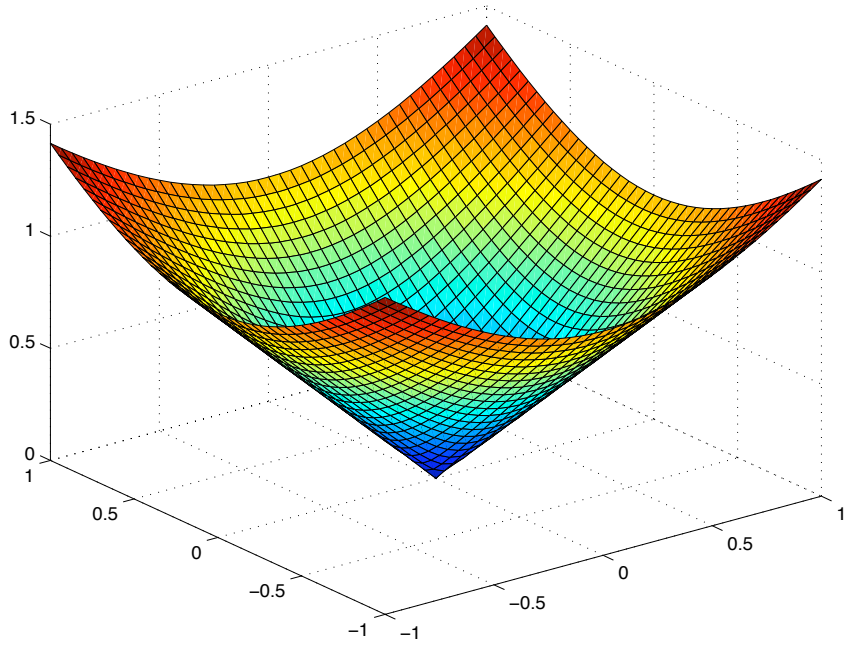
- Goal: group sparsity in rows

$$\|W\|_{\ell_1/\ell_p} = \sum_{j=1}^{n} \|\overline{\boldsymbol{w}}_j\|_p$$

$$W = \begin{bmatrix} \overline{\boldsymbol{w}}_1^T \\ \overline{\boldsymbol{w}}_2^T \\ \vdots \\ \overline{\boldsymbol{w}}_n^T \end{bmatrix}$$

Example

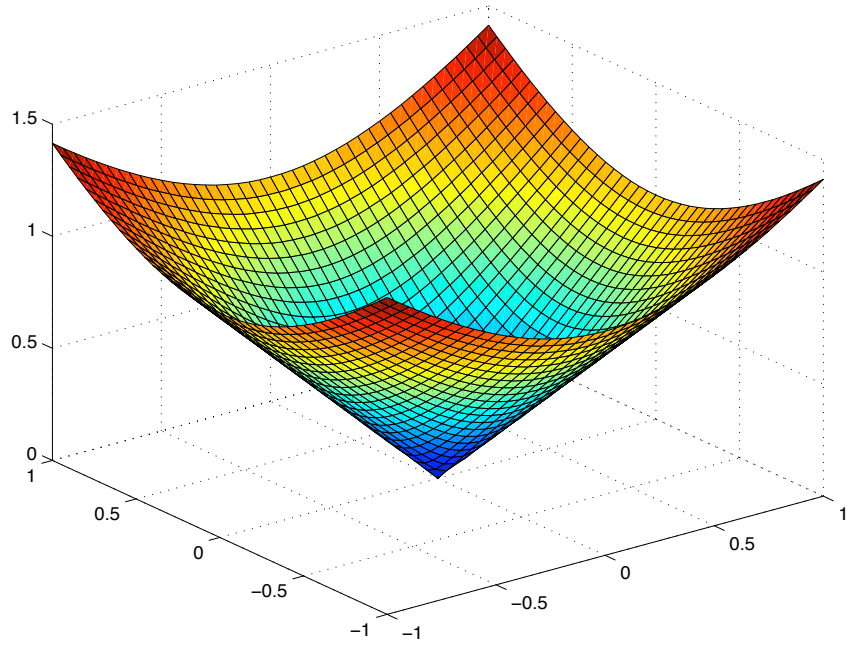
- Contour of $\|w\|_2$



Example

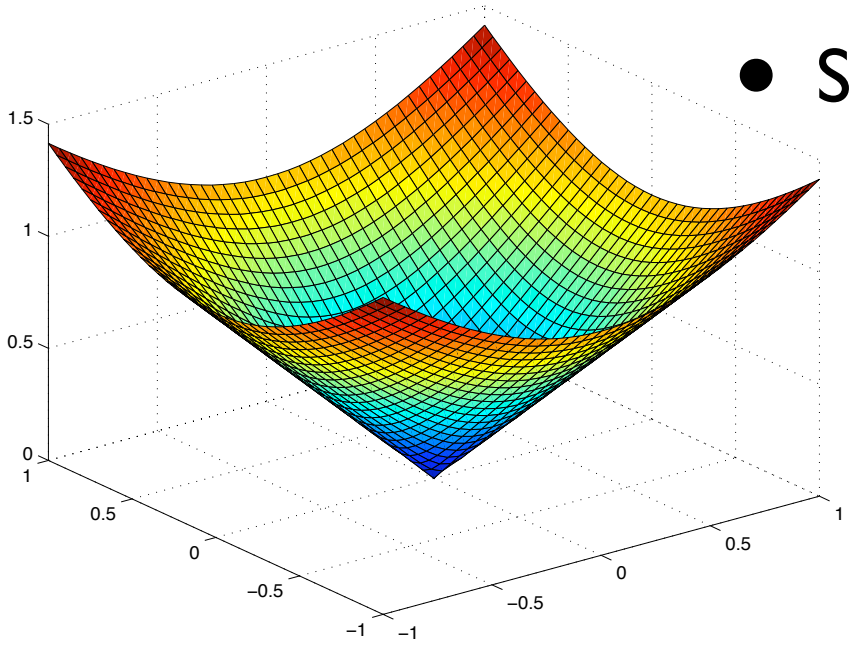
- Contour of $\|w\|_2$
- Subgradients of $\|w\|_2$ at $w = 0$

$$\{z : \|z\|_2 \leq 1\}$$



Example

- Contour of $\|w\|_2$



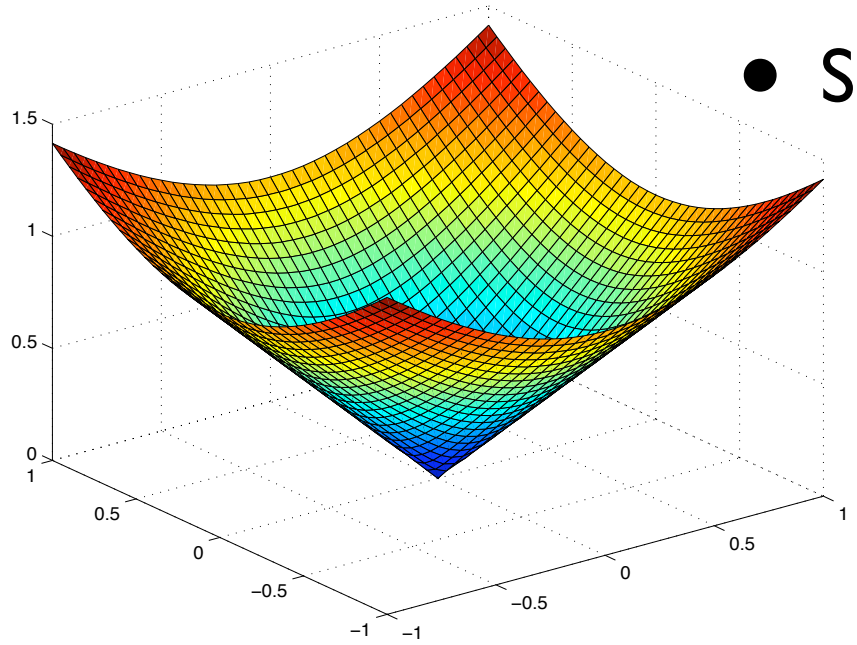
- Subgradients of $\|w\|_2$ at $w = 0$

$$\{z : \|z\|_2 \leq 1\}$$

- So if $\|\nabla L(0)\|_2 \leq \lambda$

Example

- Contour of $\|\boldsymbol{w}\|_2$



- Subgradients of $\|\boldsymbol{w}\|_2$ at $\boldsymbol{w} = 0$

$$\{\boldsymbol{z} : \|\boldsymbol{z}\|_2 \leq 1\}$$

- So if $\|\nabla L(0)\|_2 \leq \lambda$

$$0 = \operatorname*{argmin}_{\boldsymbol{w}} L(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_2$$

Add regularization

- Multiclass logistic loss

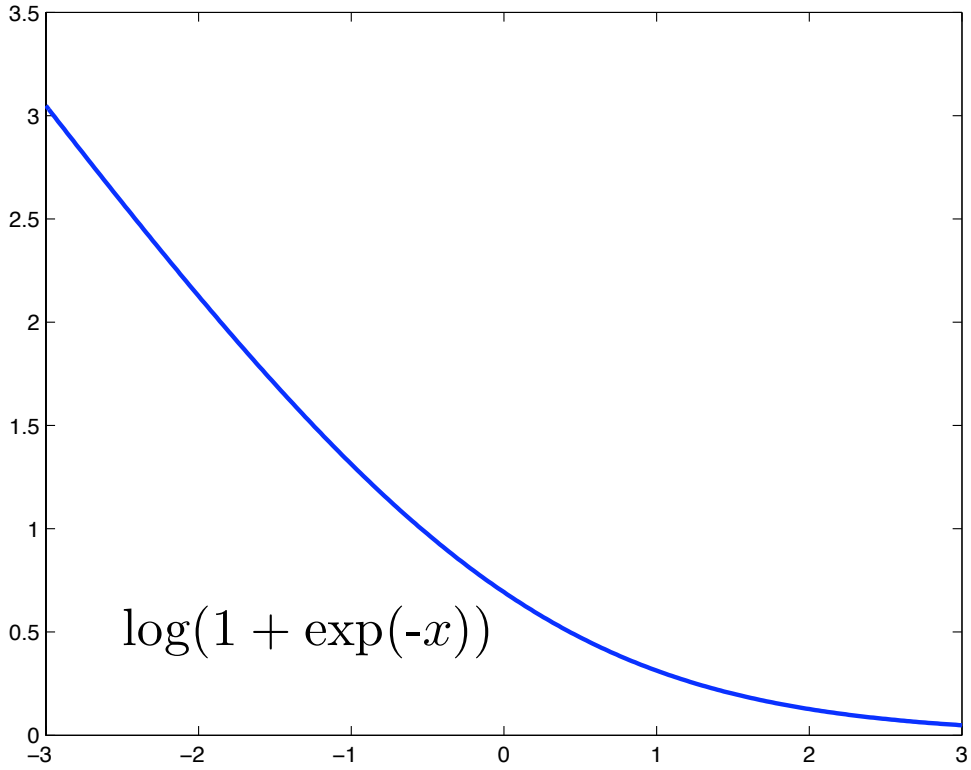
$$L(W) = \sum_{i=1}^{m} \log \left(1 + \sum_{r \neq y_i} \exp \left(\boldsymbol{w}_r \cdot \boldsymbol{x}_i - \boldsymbol{w}_{y_i} \cdot \boldsymbol{x}_i \right) \right) + \lambda \sum_{j=1}^{n} \left\| \overline{\boldsymbol{w}}_j \right\|_p$$

Loss Bounds (Multiclass)

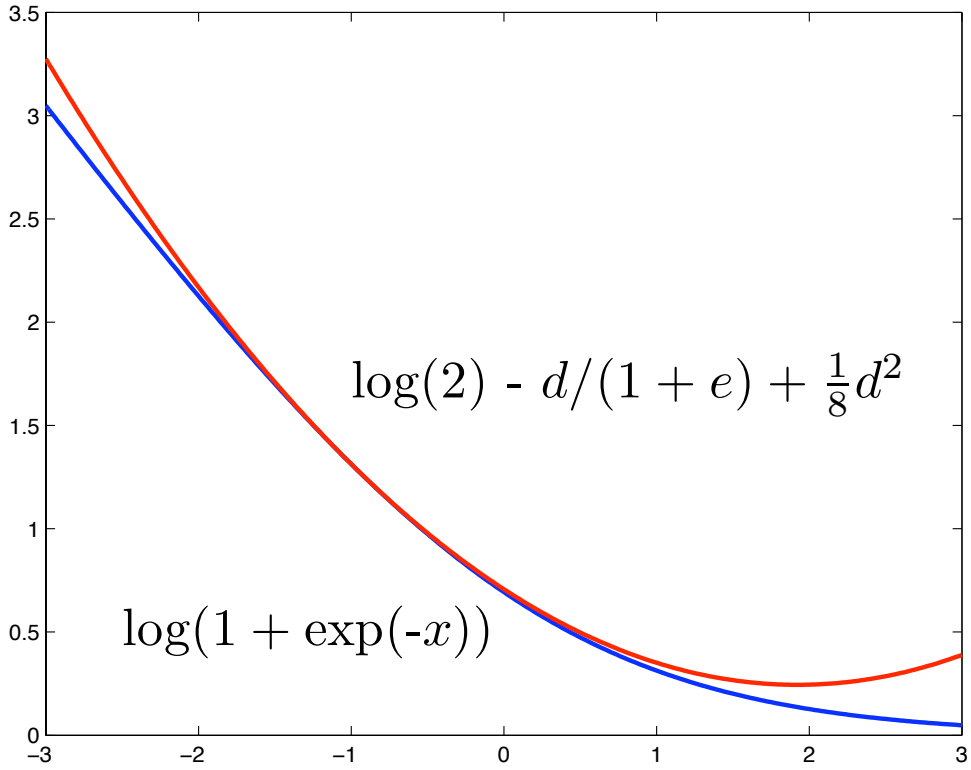
- Idea:

$$L(\boldsymbol{w} + \delta \boldsymbol{e}_j) \leq L(\boldsymbol{w}) + \nabla L(\boldsymbol{w}) \cdot \boldsymbol{e}_j \delta + \frac{1}{2} \delta \boldsymbol{e}_j \cdot D \boldsymbol{e}_j \delta$$

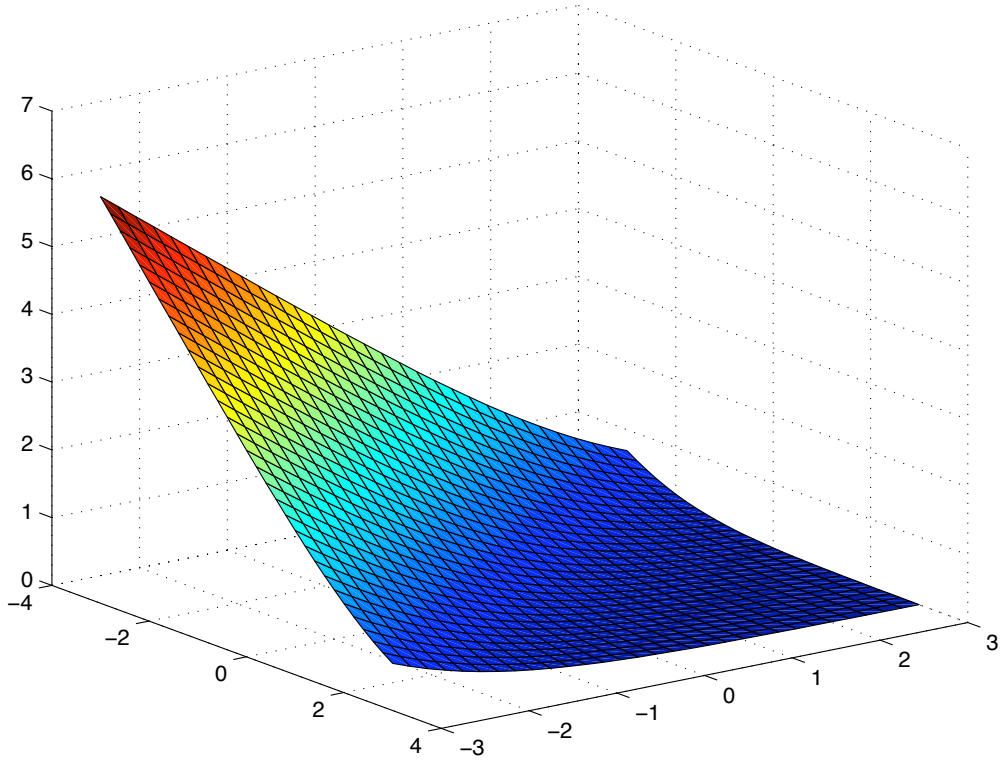
Bound Illustration



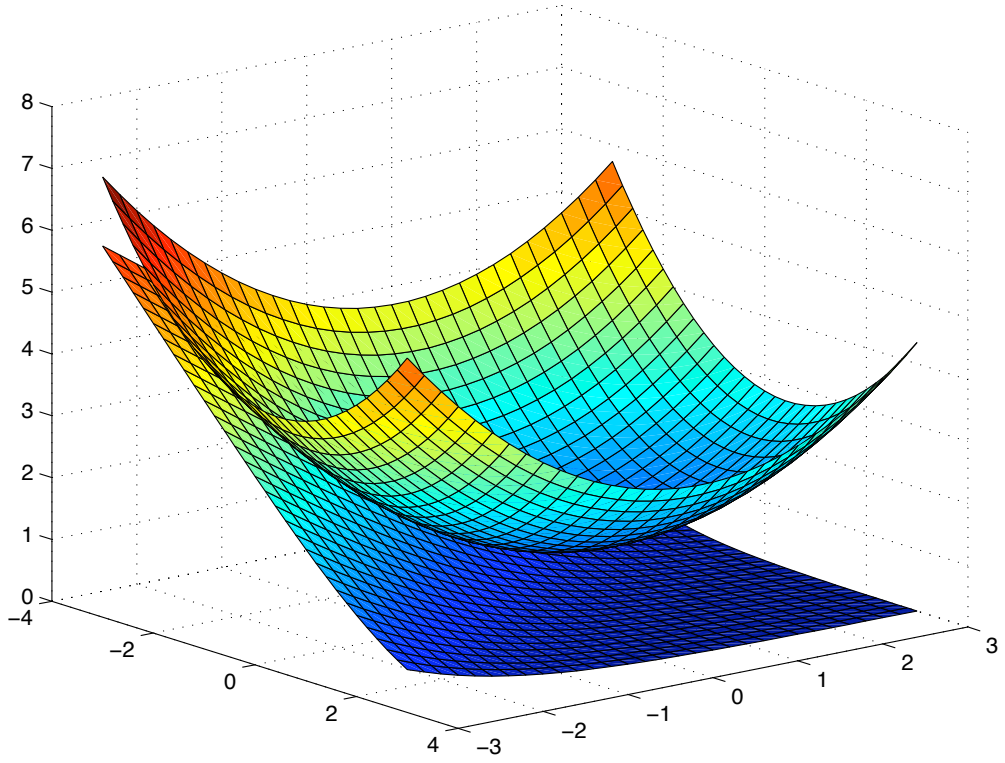
Bound Illustration



Bound Illustration (3D)



Bound Illustration (3D)



Loss Bounds (Multiclass)

- Idea:

$$L(\boldsymbol{w} + \delta \boldsymbol{e}_j) \leq L(\boldsymbol{w}) + \nabla L(\boldsymbol{w}) \cdot \boldsymbol{e}_j \delta + \frac{1}{2} \delta \boldsymbol{e}_j \cdot D \boldsymbol{e}_j \delta$$

Loss Bounds (Multiclass)

- Idea:

$$L(\boldsymbol{w} + \delta \boldsymbol{e}_j) \leq L(\boldsymbol{w}) + \nabla L(\boldsymbol{w}) \cdot \boldsymbol{e}_j \delta + \frac{1}{2} \delta \boldsymbol{e}_j \cdot D \boldsymbol{e}_j \delta$$

- Bound:

$$g_{r,j} = \frac{\partial}{\partial w_{r,j}} L(W), \quad \boldsymbol{\delta} = \begin{bmatrix} & & \mathbf{0} & \\ \delta_1 & \delta_2 & \cdots & \delta_k \\ & & \mathbf{0} & \end{bmatrix}$$

$$L(W + \boldsymbol{\delta}) \leq L(W) + \sum_{r=1}^{k} g_{r,j} \delta_r + \frac{1}{4} \sum_{r=1}^{k} \delta_r^2 x_{i,j}^2$$

The point of bounds

- Add regularization: easy minimization problems, easy updates

$$\left(\text{set } a_j = 1/\sum_i x_{i,j}^2,\ \delta_r = w_r - w_{r,j}^t\right)$$

$$\underset{\boldsymbol{w}}{\text{minimize}} \sum_r \left(g_{r,j} - \frac{w_{r,j}^t}{2a_j}\right) w_r + \sum_{r=1}^{k} \frac{1}{4a_j} w_r^2 + \lambda \left\|\boldsymbol{w}\right\|_p$$

Update example

Multiclass GradBoost with ℓ_2

$$a_j = 1/\sum_i x_{i,j}^2,\ \boldsymbol{g}_j = \left[\frac{\partial}{\partial w_{1,j}} L(W) \quad \cdots \quad \frac{\partial}{\partial w_{k,j}} L(W)\right]^T$$

Update example

Multiclass GradBoost with ℓ_2

$$\overline{\boldsymbol{w}}_j^{t+1} = \left(\overline{\boldsymbol{w}}_j^t - 2a_j \boldsymbol{g}_j\right) \left[1 - \frac{2a_j\lambda}{\left\|\overline{\boldsymbol{w}}_j^t - 2a_j\boldsymbol{g}_j\right\|_2}\right]_+$$

$$a_j = 1/\sum_i x_{i,j}^2,\ \boldsymbol{g}_j = \left[\frac{\partial}{\partial w_{1,j}} L(W) \quad \cdots \quad \frac{\partial}{\partial w_{k,j}} L(W)\right]^T$$

Update Benefits

$$\overline{\boldsymbol{w}}_j^{t+1} = \left(\overline{\boldsymbol{w}}_j^t - 2a_j \boldsymbol{g}_j\right) \left[1 - \frac{2a_j\lambda}{\left\|\overline{\boldsymbol{w}}_j^t - 2a_j \boldsymbol{g}_j\right\|_2}\right]_+$$

Update Benefits

$$\overline{\boldsymbol{w}}_j^{t+1} = \left(\overline{\boldsymbol{w}}_j^t - 2a_j \boldsymbol{g}_j\right) \left[1 - \frac{2a_j\lambda}{\left\|\overline{\boldsymbol{w}}_j^t - 2a_j \boldsymbol{g}_j\right\|_2}\right]_+$$

- Totally corrective: feature pruning

Update Benefits

$$\overline{\boldsymbol{w}}_j^{t+1} = \left(\overline{\boldsymbol{w}}_j^t - 2a_j \boldsymbol{g}_j\right) \left[1 - \frac{2a_j\lambda}{\left\|\overline{\boldsymbol{w}}_j^t - 2a_j \boldsymbol{g}_j\right\|_2}\right]_+$$

- Totally corrective: feature pruning

Update Benefits

$$\overline{\boldsymbol{w}}_j^{t+1} = \left(\overline{\boldsymbol{w}}_j^t - 2a_j \boldsymbol{g}_j\right) \left[1 - \frac{2a_j\lambda}{\left\|\overline{\boldsymbol{w}}_j^t - 2a_j \boldsymbol{g}_j\right\|_2}\right]_+$$

- Totally corrective: feature pruning
- Feature scoring: $\frac{\left(\left[\|\boldsymbol{g}_j\|_2 - \lambda\right]_+\right)^2}{\sum_i x_{i,j}^2}$

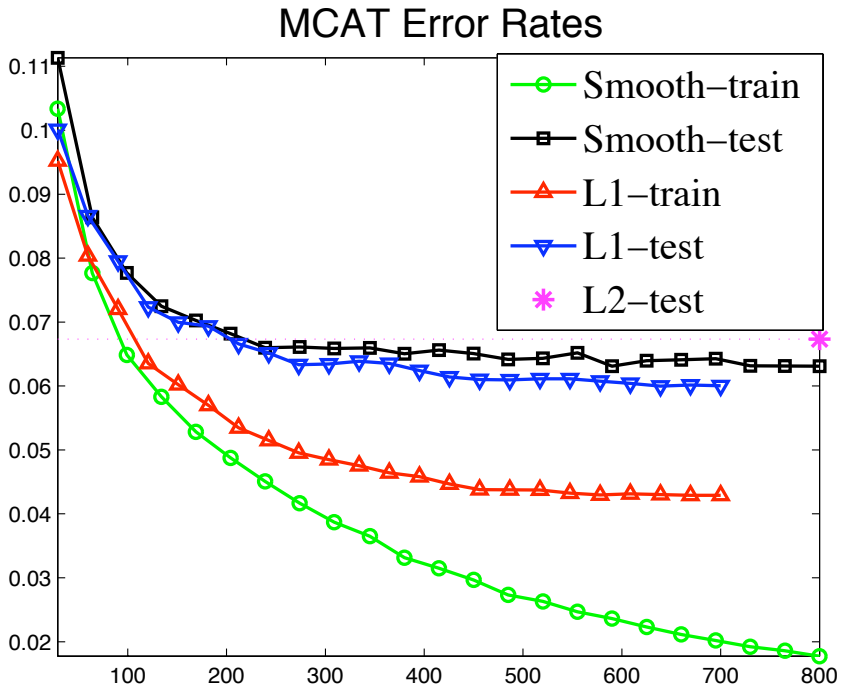
Obligatory Theory Slide

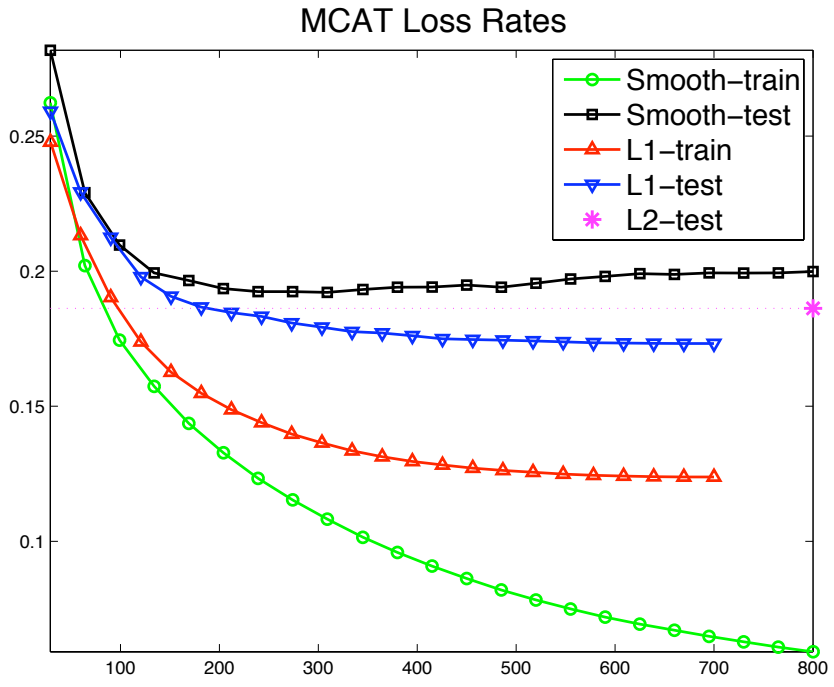
Theorem: If number of features/base hypotheses is finite, all presented algorithms converge to optimum of their respective losses

Experiments

- Binary classification
- Multiclass classification
- Runtime properties

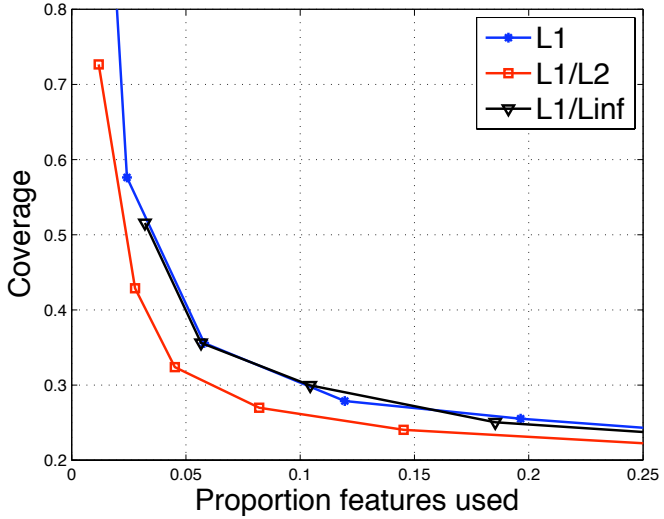
Binary Classification

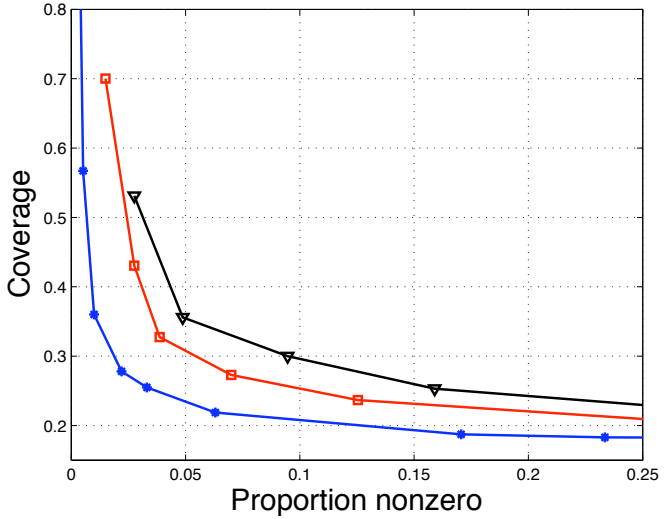


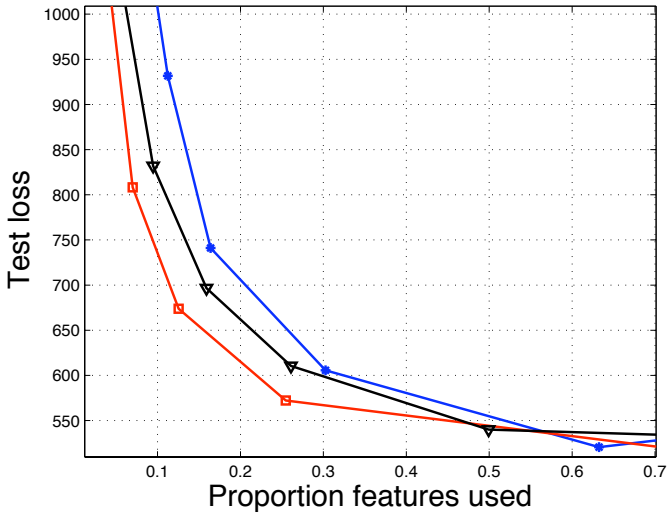


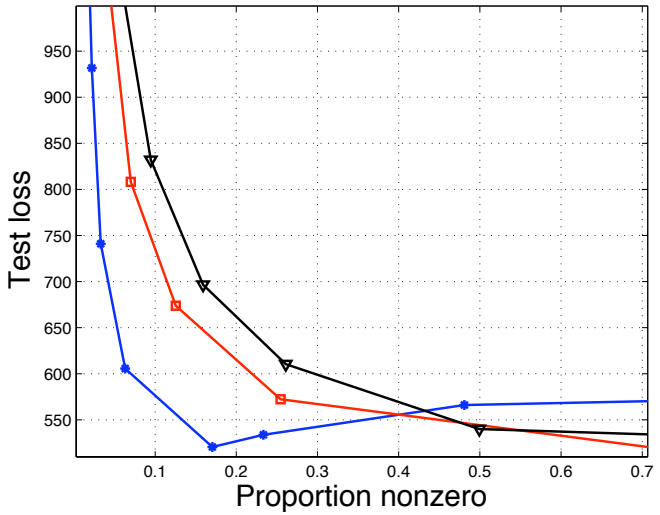
Dataset: Reuters RCV1, Lewis et al. 2004

Multiclass Classification

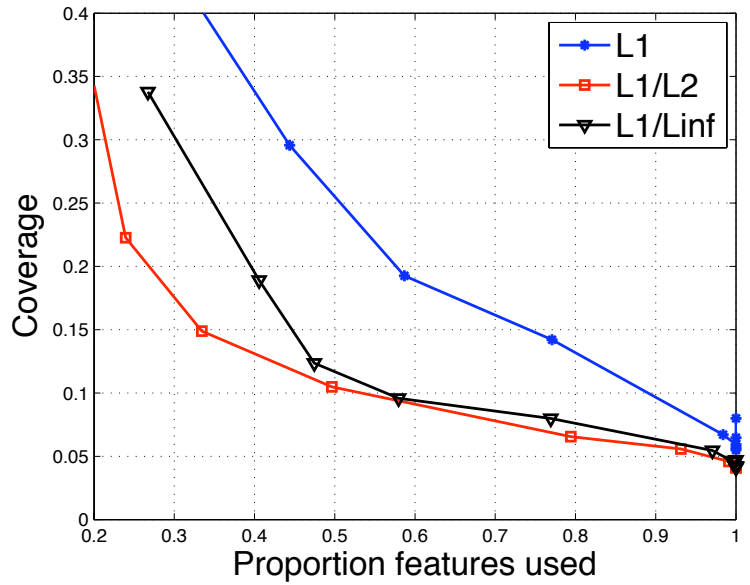


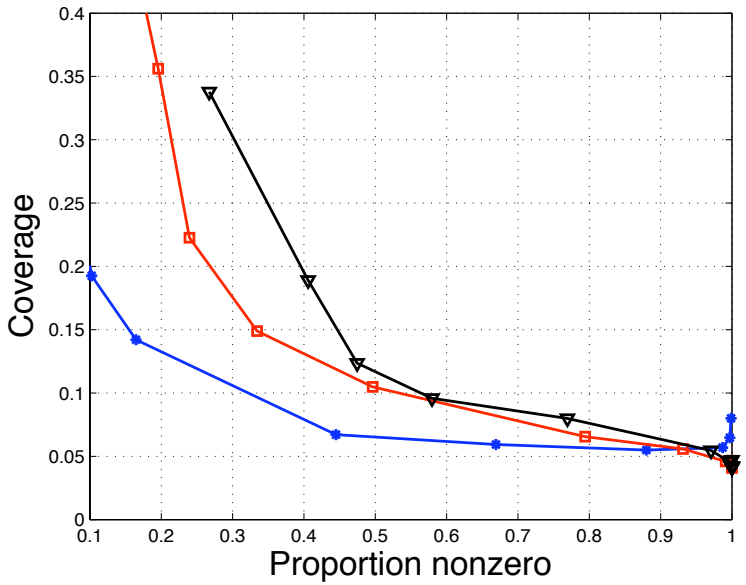


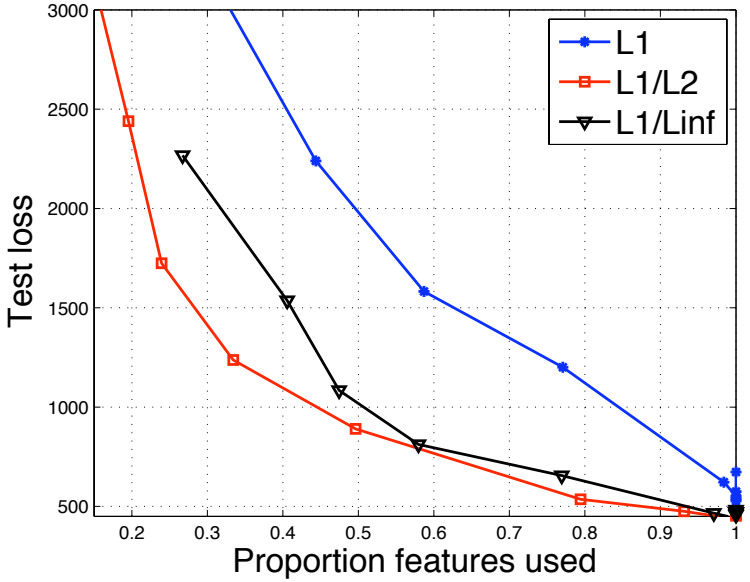


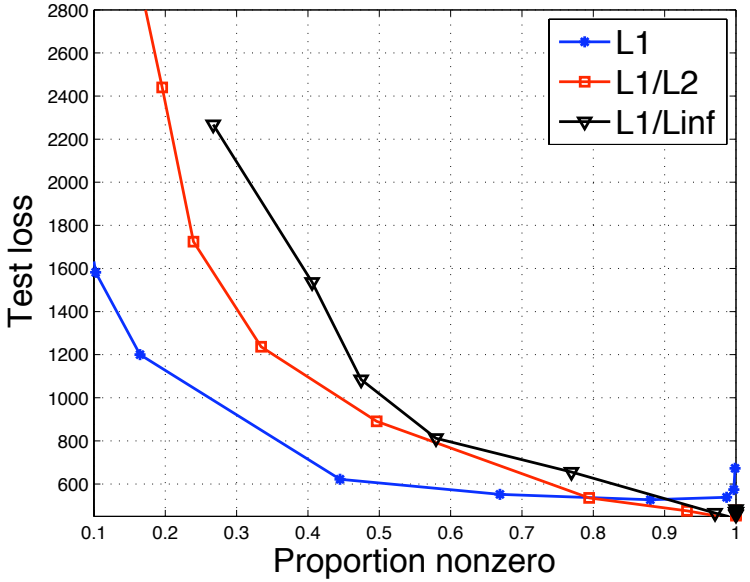


Dataset: StatLog Satellite



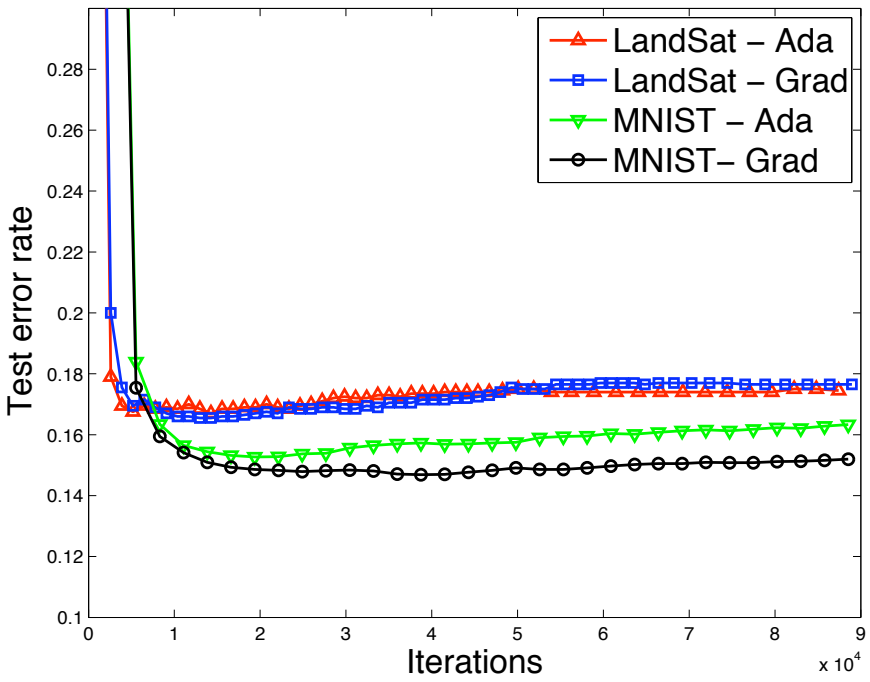




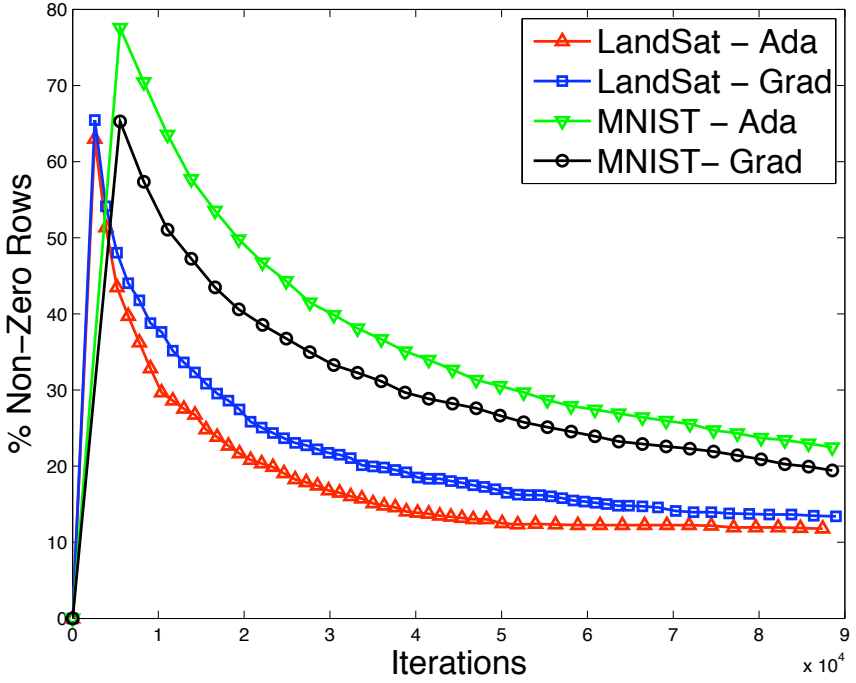


Dataset: PenDigit (UCI ML Repository)

Runtime Behavior



Error rate vs. train time



Non-zeros vs. train time

Concluding Remarks

Concluding Remarks

- Boosting variants for multiclass/binary/ regression

Concluding Remarks

- Boosting variants for multiclass/binary/ regression
- Prune features, perform induction, provably convergent

Concluding Remarks

- Boosting variants for multiclass/binary/ regression
- Prune features, perform induction, provably convergent
- Further work:

Concluding Remarks

- Boosting variants for multiclass/binary/ regression
- Prune features, perform induction, provably convergent
- Further work:
  - Convergence Rates (Shalev-Shwartz and Tewari 2009)

Concluding Remarks

- Boosting variants for multiclass/binary/ regression
- Prune features, perform induction, provably convergent
- Further work:
  - Convergence Rates (Shalev-Shwartz and Tewari 2009)
  - Consistency and generalization (Schapire et al. 1998; Zhang and Yu 2005)

Thanks very much!

- Any questions?