Fast Newton methods for the group fused lasso

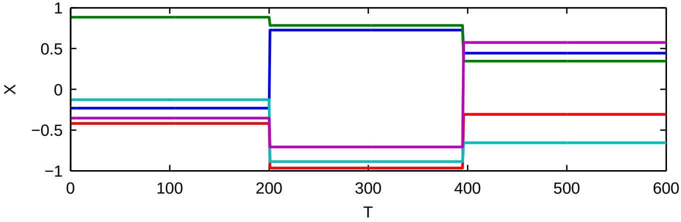
Matt Wytock, Suvrit Sra, and J. Zico Kolter
Machine Learning Department
Carnegie Mellon University

July 26, 2014

The group fused lasso

- Approximates a multivariate input signal $y_1, \ldots, y_T$ ($y_t \in \mathbb{R}^n$) with piecewise constant $x_1, \ldots, x_T$ by solving

$$\underset{x_1,\ldots,x_T}{\text{minimize}} \ \frac{1}{2}\sum_{i=1}^{T} \|x_t - y_t\|_2^2 + \lambda \sum_{t=1}^{T-1} \|x_t - x_{t+1}\|_2$$



Example of piecewise constant structure

- Also referred to as the (multivariate) total variation norm (Bleakley and Vert, 2011), (Alaíz et al, 2013)

Matrix notation

- Equivalently, in matrix notation

$$\underset{X}{\text{minimize}} \frac{1}{2}\|X - Y\|_F^2 + \lambda TV(X)$$

where $X, Y \in \mathbb{R}^{n \times T}$ denote

$$X = \begin{bmatrix} x_1 & \cdots & x_T \end{bmatrix}, \;\; Y = \begin{bmatrix} y_1 & \cdots & y_T \end{bmatrix}$$

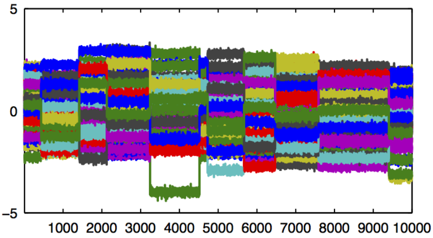
- The multivariate TV norm is defined as

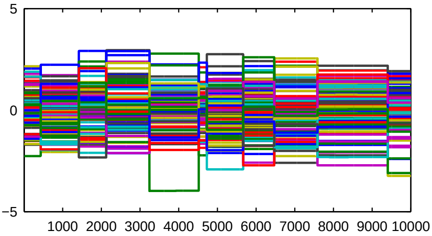
$$TV(X) := \|XD\|_{1,2} = \sum_t^{T-1} \|x_t - x_{t+1}\|_2$$

- Using the first order differencing operator

$$D = \begin{bmatrix} 1 & 0 & 0 & \cdots \\ -1 & 1 & 0 & \cdots \\ 0 & -1 & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Application: Multiple changepoint detection





- Used in place of HMM models, e.g. modeling DNA copy number alterations (Bleakley and Vert, 2011)

Application: Color image denoising

- Penalizes variation across adjacent pixels (Rudin et al., 1992)

$$\underset{X}{\text{minimize}} \;\; \frac{1}{2}\|X - Y\|_F^2 + \lambda \left(\sum_{i=1}^{m} TV(X_{:,i,:}) + \sum_{j=1}^{n} TV(X_{:,:,j}) \right)$$

where $X, Y \in \mathbb{R}^{3 \times m \times n}$ are color images

original image

image with noise

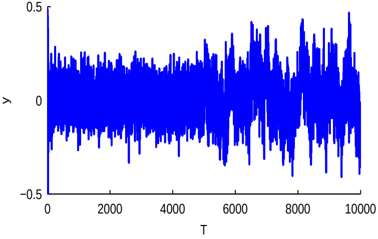
denoised image

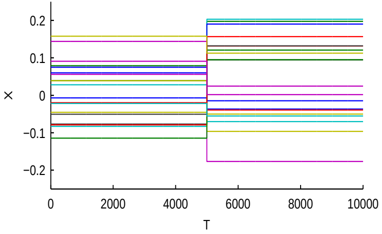
Application: Linear regression segmentation

- Observe a sequence of input/output pairs ($a_t \in \mathbb{R}^n$, $y_t \in \mathbb{R}$) and find x_t such that $y_t \approx a_t^T x_t$ (Ohlsson et al., 2010)

$$\underset{X}{\text{minimize}} \, \|A \operatorname{vec} X - y\|_2^2 + \lambda TV(X)$$

- For example, time-varying AR model with $a_t = (y_{t-1}, \ldots, y_{t-n})$





Optimization, primal problem

- Recall the original problem in matrix notation

$$\underset{X}{\text{minimize}} \, \frac{1}{2}\|X - Y\|_F^2 + \|XD\|_{1,2}$$

- Optimization is complicated by the nonsmooth TV norm

- Note that even when $x_t - x_{t+1}$ is sparse, X will be dense

Dual problem

- Formed by introducing the constraint $V = XD$

$$\begin{aligned} \underset{U}{\text{maximize}} \quad & -\frac{1}{2}\|UD^T\|_F^2 + \operatorname{tr} UD^TY^T \\ \text{subject to} \quad & \|u_t\|_2 \le \lambda, \qquad t = 1, \ldots, T-1 \end{aligned}$$

- Second-order cone program with smooth objective

- Again, U will be dense even when $x_t - x_{t+1}$ is sparse

- Observe that $\|u_t\|_2^2 \le \lambda^2$ is an equivalent constraint

Dual dual problem

- We consider the dual of the dual formed with $\|u_t\|_2^2 \leq \lambda^2$

$$\begin{array}{ll} \underset{z}{\text{minimize}} & \frac{1}{2} Y D(D^T D + Z)^{-1} D^T Y^T + \frac{\lambda^2}{2} 1^T z \\ \text{subject to} & z \geq 0 \end{array}$$

  where $Z = \text{diag}(z)$

- Fewer variables than original problem $z \in \mathbb{R}^{T-1}$ vs. $X \in \mathbb{R}^{T \times Tn}$

- Sparse at the solution, $z^\star$ is nonzero only at change points

- Smooth objective plus simple nonnegative constraints

Active Set Projected Newton (ASPN)

- Apply general projected Newton method for smooth problems with simple constraints (Bertsekas, 1982)
  1. Construct the set of *bound* variables

$$\mathcal{I} := \{i : z_i = 0 \text{ and } (\nabla_z f(z))_i > 0\}$$

  2. Perform a Newton update plus projection on variables *not* bound

$$z_{\bar{\mathcal{I}}} \leftarrow \left[z_{\bar{\mathcal{I}}} - \alpha (\nabla_z^2 f(z))^{-1}_{\bar{\mathcal{I}},\bar{\mathcal{I}}} (\nabla_z f(z))_{\bar{\mathcal{I}}} \right]_+$$

Active Set Projected Newton (ASPN)

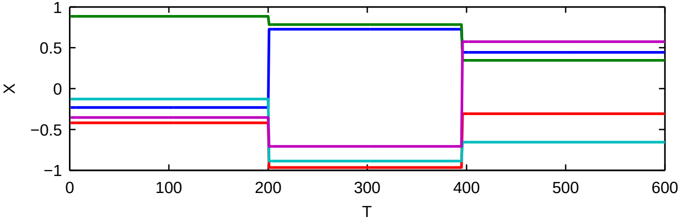
- Apply general projected Newton method for smooth problems with simple constraints (Bertsekas, 1982)
    1. Construct the set of *bound* variables

$$\mathcal{I} := \{i : z_i = 0 \text{ and } (\nabla_z f(z))_i > 0\}$$

    2. Perform a Newton update plus projection on variables *not* bound

$$z_{\bar{\mathcal{I}}} \leftarrow \left[z_{\bar{\mathcal{I}}} - \alpha(\nabla_z^2 f(z))^{-1}_{\bar{\mathcal{I}},\bar{\mathcal{I}}}(\nabla_z f(z))_{\bar{\mathcal{I}}}\right]_+$$

- Active set approach to make Newton step fast by solving a significantly reduced problem



Active Set Projected Newton (ASPN)

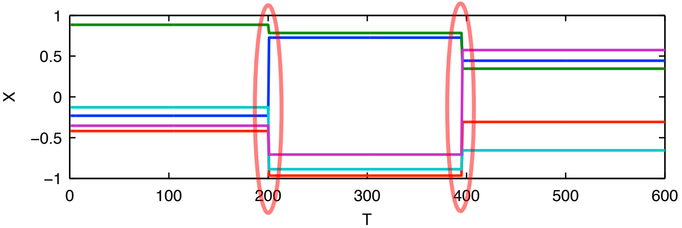
- Apply general projected Newton method for smooth problems with simple constraints (Bertsekas, 1982)
  1. Construct the set of *bound* variables

$$\mathcal{I} := \{i : z_i = 0 \text{ and } (\nabla_z f(z))_i > 0\}$$

  2. Perform a Newton update plus projection on variables *not* bound

$$z_{\bar{\mathcal{I}}} \leftarrow \left[z_{\bar{\mathcal{I}}} - \alpha(\nabla_z^2 f(z))^{-1}_{\bar{\mathcal{I}},\bar{\mathcal{I}}}(\nabla_z f(z))_{\bar{\mathcal{I}}}\right]_+$$

- Active set approach to make Newton step fast by solving a significantly reduced problem



Active Set Projected Newton (ASPN)

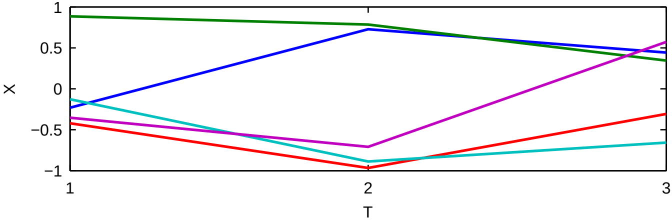
- Apply general projected Newton method for smooth problems with simple constraints (Bertsekas, 1982)
  1. Construct the set of *bound* variables

$$\mathcal{I} := \{i : z_i = 0 \text{ and } (\nabla_z f(z))_i > 0\}$$

  2. Perform a Newton update plus projection on variables *not* bound

$$z_{\bar{\mathcal{I}}} \leftarrow \left[z_{\bar{\mathcal{I}}} - \alpha (\nabla_z^2 f(z))^{-1}_{\bar{\mathcal{I}},\bar{\mathcal{I}}} (\nabla_z f(z))_{\bar{\mathcal{I}}} \right]_+$$

- Active set approach to make Newton step fast by solving a significantly reduced problem

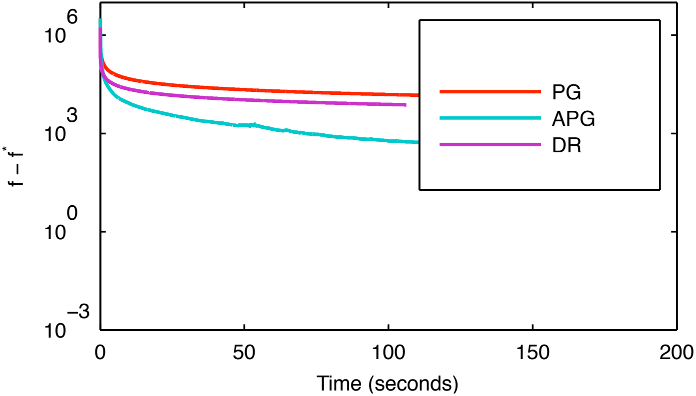


Alternative algorithms

- GFL - coordinate descent on primal (Bleakley and Vert, 2011)
- PG - projected gradient on dual, ISTA
- APG - accelerated projected gradient, FISTA
- DR - Douglas-Rachford splitting, generalization of ADMM (Combettes and Pesquest, 2007)
- LBFGS-B - applied to the dual dual (Byrd, 1995)

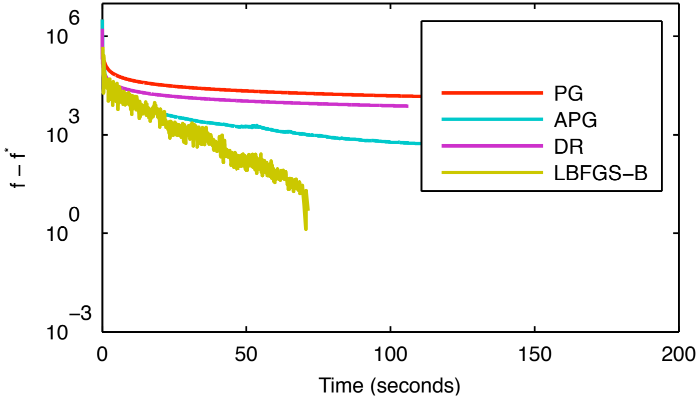
Results on multiple change point detection

- Lung cancer dataset (Bleakley and Vert, 2911), $T = 31708$, $n = 18$



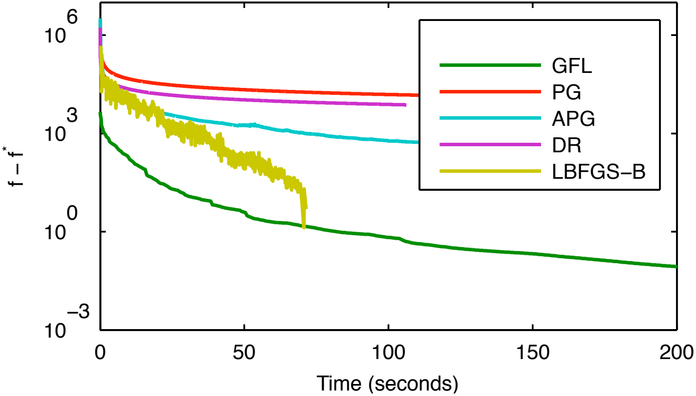
Results on multiple change point detection

- Lung cancer dataset (Bleakley and Vert, 2911), $T = 31708$, $n = 18$



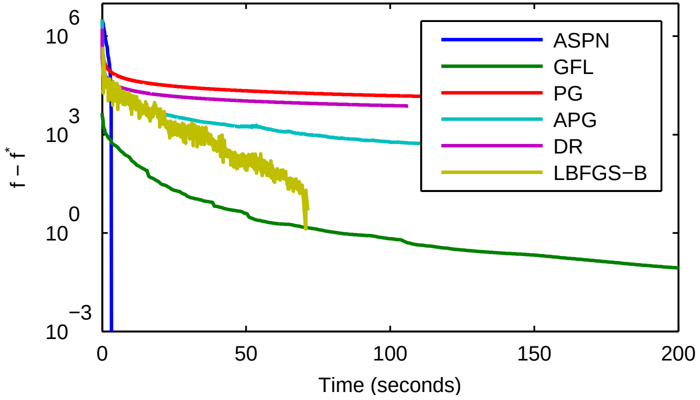
Results on multiple change point detection

- Lung cancer dataset (Bleakley and Vert, 2911), $T = 31708$, $n = 18$

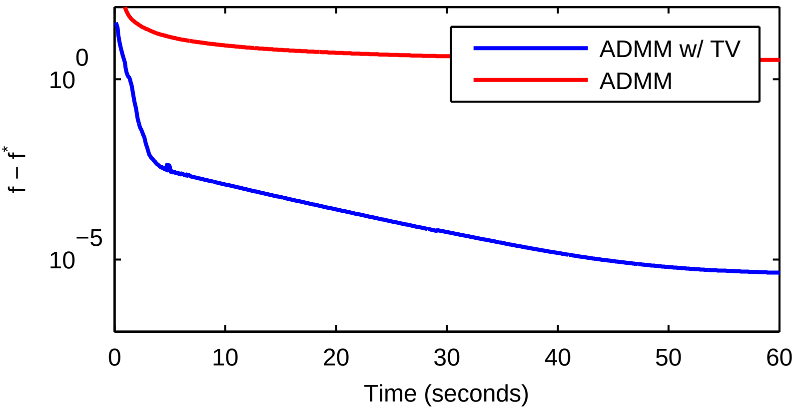


Results on multiple change point detection

- Lung cancer dataset (Bleakley and Vert, 2911), $T = 31708$, $n = 18$

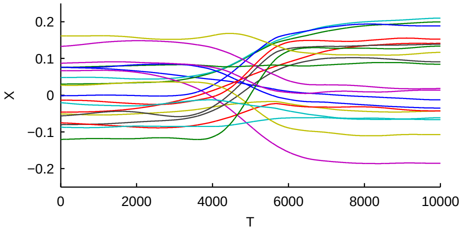


Results on linear regression segmentation

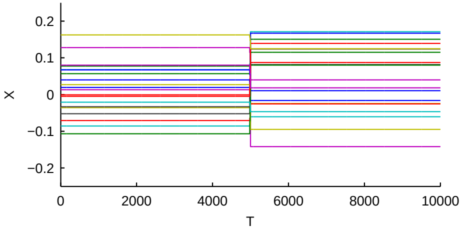


"Simple" ADMM approach converges significantly slower

Comparison of solutions for LR segmentation



Parameters recovered with “simple” ADMM algorithm



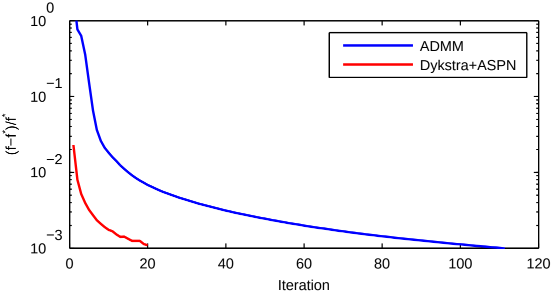
Parameters recovered using ADMM w/ ASPN

Results on color image denoising

original image

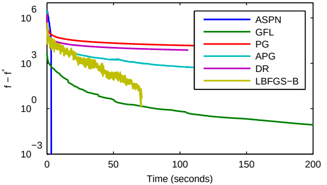
image with noise

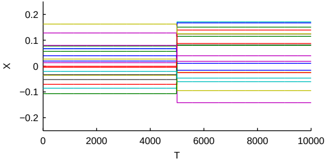
denoised image



Requires significantly fewer iterations for highly accurate solution

Summary and conclusions





- Group fused lasso (total variation norm) used in multiple changepoint detection, color image denoising, linear regression segmentation, etc.
- ASPN algorithm exploits structure for fast convergence to highly accurate solutions
- Code available shortly at `http://www.cs.cmu.edu/~mwytock/gfl`