Second Order Methods for L1-Regularization

Kimon Fountoulakis and Jacek Gondzio

with extra thanks to

Kristian Woodsend and Pavel Zlobich

Edinburgh, May 2, 2013
Outline

• Motivation: Why not \textit{2nd-order} methods?
• Interior Point Methods and Continuation
• Inexact Newton directions
  – Krylov subspace methods
  – Preconditioner is a must
• Computational results
  – Compressed Sensing
  – \textit{Google} Problem
  – Machine Learning Problems
• Linear Algebra viewpoint on $\ell_1$-regularization
• Conclusions
ℓ₁ regularization

Convex optimization problem:

$$\min_x \tau \|x\|_1 + \phi(x),$$

where $\| . \|_1$ is the ℓ₁ norm, and 
$\phi : \mathbb{R}^n \to \mathbb{R}$ is a convex function (often strongly convex).

Usual example:

$$\min_x \tau \|x\|_1 + \frac{1}{2} \|Ax - b\|_2^2$$

where $A \in \mathbb{R}^{m \times n}$ (often $m \geq n$ or $m \gg n$).
Two features:

Difficulty:
non-differentiability of $\|x\|_1$

Triviality:
unconstrained optimization

It is fashionable to use the 1st-order methods to solve these problems. Marketed as *Haute Couture.*

*Prêt-à-porter.* What about the 2nd-order methods???
2nd-order methods for $\ell_1$-regularization

Observation

• First-order methods
  – complexity $O(1/\varepsilon)$ or $O(1/\varepsilon^2)$
  – produce a rough approx. of solution quickly
  – but ... struggle to converge to high accuracy

• IPMs are second-order methods
  (they apply Newton method to barrier subprobs)
    – complexity $O(\log(1/\varepsilon))$
    – produce accurate solution in a few iterations
    – but ... one iteration may be expensive
2nd-order methods for $\ell_1$-regularization

**Just think**

For example, $\varepsilon = 10^{-3}$ gives

$$1/\varepsilon = 10^3 \text{ and } 1/\varepsilon^2 = 10^6, \text{ but } \log(1/\varepsilon) \approx 7.$$ 

For example, $\varepsilon = 10^{-6}$ gives

$$1/\varepsilon = 10^6 \text{ and } 1/\varepsilon^2 = 10^{12}, \text{ but } \log(1/\varepsilon) \approx 14.$$ 

But **ML Community** loves the 1st-order methods.

Stirring up a hornets nest:

**Give 2nd-order/IPMs a serious consideration!**

Edinburgh, May 2, 2013
Serious Issue: nondifferentiability of $\| \cdot \|_1$

Two possible tricks:

- Splitting $x = u - v$ with $u, v \geq 0$
- Huber or pseudo-Huber regression
Splitting: \( x = u - v, \ u \geq 0, \ v \geq 0 \)

Replace \( x_i = u_i - v_i \),
where \( u_i = \max\{x_i, 0\} \) and \( v_i = \max\{-x_i, 0\} \).

Then \( x_i = u_i - v_i \) and \( |x_i| = u_i + v_i \).

Hence \( \|x\|_1 = \sum_{i=1}^{n} (u_i + v_i) \).

Removes nondifferentiability, but:

- doubles the dimension,
- introduces inequality constraints (fine for IPMs).
2nd-order methods for $\ell_1$-regularization

**Huber:** Replace $\|x\|_1$ with $\psi_\mu(x)$

**Huber approximation:** replaces $\|x\|_1$ with $\sum_{i=1}^{n} \left[ \phi_\mu(x) \right]_i$

$$
\left[ \phi_\mu(x) \right]_i = \begin{cases} 
\frac{1}{2} \mu^{-1} x_i^2, & \text{if } |x_i| \leq \mu \\
|x_i| - \frac{1}{2} \mu, & \text{if } |x_i| \geq \mu 
\end{cases} \quad i = 1, 2, \ldots, n
$$

where $\mu > 0$. Only first-order differentiable.

**Pseudo Huber approximation:** replaces $\|x\|_1$ with

$$
\psi_\mu(x) = \mu \sum_{i=1}^{n} (\sqrt{1 + \frac{x_i^2}{\mu^2}} - 1)
$$

Smooth function, has derivatives of any degree.
2nd-order methods for $\ell_1$-regularization

Huber:
2nd-order methods for $\ell_1$-regularization

2nd-order method

Use 2nd-order information (Newton direction).

But, do not waste time on computing exact direction.

Use Inexact Newton Method

Dembo, Eisenstat & Steihaug,

2nd-order methods for $\ell_1$-regularization

Continuation

Embed inexact Newton Meth into a homotopy approach:

- Inequalities $u \geq 0$, $v \geq 0$ $\rightarrow$ use IPM
  replace $z \geq 0$ with $-\mu \log z$ and drive $\mu$ to zero.

- pseudo-Huber regression $\rightarrow$ use continuation
  replace $|x_i|$ with $\mu(\sqrt{1 + \frac{x_i^2}{\mu^2}} - 1)$ and drive $\mu$ to zero.

Theory ???

Edinburgh, May 2, 2013
2nd-order methods for $\ell_1$-regularization

Theory for IPM:


Theory for Continuation:

Fountoulakis and G.
Three examples of simple $\ell_1$ regularization

- Compressed Sensing with K. Fountoulakis and P. Zhlobich
- Google Problem with K. Woodsend
- Machine Learning Problems with K. Fountoulakis
Example One

- Compressed Sensing
  with K. Fountoulakis and P. Zhlobich
2nd-order methods for $\ell_1$-regularization

Compressed Sensing

Relatively small number of random projections of a sparse signal can contain most of its salient information.

If a signal is sparse (or approximately sparse) in some orthonormal basis, then an accurate reconstruction can be obtained from random projections of the original signal. $A$ has the form $A = RW$, where

- $R$ is a low-rank randomised sensing matrix
- $W$ is a basis over which the signal has a sparse representation

Candès, Romberg & Tao,
2nd-order methods for $\ell_1$-regularization

**Compressed Sensing** joint work with

Kimon Fountoulakis and Pavel Zhlobich

Large dense quadratic optimization problem:

$$\min_{x} \tau \|x\|_1 + \frac{1}{2} \|Ax - b\|_2^2,$$

where $A \in \mathbb{R}^{m \times n}$ is a **very special matrix**.

Fountoulakis, G., Zhlobich

Software available at [http://www.maths.ed.ac.uk/ERGO/](http://www.maths.ed.ac.uk/ERGO/)

Edinburgh, May 2, 2013
2nd-order methods for \( \ell_1 \)-regularization

**Two-way Orthogonality of A**

- **rows** of \( A \) are orthogonal to each other (\( A \) is built of a subset of rows of an orthonormal matrix \( U \in \mathbb{R}^{n \times n} \))

\[
AA^T = I_m.
\]

- small subsets of **columns** of \( A \) are nearly-orthogonal to each other: **Restricted Isometry Property (RIP)**

\[
\| \bar{A}^T \bar{A} - \frac{m}{n} I_k \| \leq \delta_k \in (0, 1).
\]

Restricted Isometry Property

Matrix $\bar{A} \in \mathcal{R}^{m \times k}$ ($k \ll n$) is built of a subset of columns of $A \in \mathcal{R}^{m \times n}$.

$$A = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ \end{bmatrix} \quad \rightarrow \quad \bar{A} = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ \end{bmatrix}$$

$$\bar{A}^T \bar{A} = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ \end{bmatrix} = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ \end{bmatrix} \approx \frac{m}{n} I_k.$$  

This yields a very well conditioned optimization problem.

Edinburgh, May 2, 2013
Problem Reformulation

\[ \min_x \tau \|x\|_1 + \frac{1}{2} \|Ax - b\|_2^2 \]

Replace \( x = x^+ - x^- \) to be able to use \( |x| = x^+ + x^- \).

Use \( |x_i| = z_i + z_{i+n} \) to replace \( \|x\|_1 \) with \( \|x\|_1 = 1^T z \).

(Increases problem dimension from \( n \) to \( 2n \).)

\[ \min_{z \geq 0} c^T z + \frac{1}{2} z^T Q z, \]

where

\[ Q = \begin{bmatrix} A^T & A \\ -A^T & A^T \end{bmatrix} \begin{bmatrix} A & -A \end{bmatrix} = \begin{bmatrix} A^T A & -A^T A \\ -A^T A & A^T A \end{bmatrix} \in \mathcal{R}^{2n \times 2n} \]
Preconditioner

Approximate

\[ \mathcal{M} = \begin{bmatrix} A^T A & -A^T A \\ -A^T A & A^T A \end{bmatrix} + \begin{bmatrix} \Theta_1^{-1} \\ \Theta_2^{-1} \end{bmatrix} \]

with

\[ \mathcal{P} = \frac{m}{n} \begin{bmatrix} I_n & -I_n \\ -I_n & I_n \end{bmatrix} + \begin{bmatrix} \Theta_1^{-1} \\ \Theta_2^{-1} \end{bmatrix}. \]

We expect (optimal partition):

- \( k \) entries of \( \Theta^{-1} \to 0, \ k \ll 2n, \)
- \( 2n - k \) entries of \( \Theta^{-1} \to \infty. \)
2nd-order methods for $\ell_1$-regularization

Spectral Properties of $P^{-1}M$

Theorem

- Exactly $n$ eigenvalues of $P^{-1}M$ are 1.
- The remaining $n$ eigenvalues satisfy

$$|\lambda(P^{-1}M) - 1| \leq \delta_k + \frac{n}{m\delta_k L},$$

where $\delta_k$ is the RIP-constant, and $L$ is a threshold of “large” $(\Theta_1 + \Theta_2)^{-1}$.

Fountoulakis, G., Zhlobich
Matrix-free IPM for Compressed Sensing Problems,
2nd-order methods for $\ell_1$-regularization

Preconditioning

Matrix–vector products per CG/PCG call

Spread of $\lambda(M)/\lambda(P^{-1}M)$ per call of CG/PCG

good clustering of eigenvalues

Edinburgh, May 2, 2013
### Computational Results: Comparing MatVecs

<table>
<thead>
<tr>
<th>Prob size</th>
<th>k</th>
<th>NestA</th>
<th>mf-IPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>4k</td>
<td>51</td>
<td>424</td>
<td>301</td>
</tr>
<tr>
<td>16k</td>
<td>204</td>
<td>461</td>
<td>307</td>
</tr>
<tr>
<td>64k</td>
<td>816</td>
<td>453</td>
<td>407</td>
</tr>
<tr>
<td>256k</td>
<td>3264</td>
<td>589</td>
<td>537</td>
</tr>
<tr>
<td>1M</td>
<td>13056</td>
<td>576</td>
<td>613</td>
</tr>
</tbody>
</table>

**NestA**, Nesterov’s smoothing gradient  
**Becker, Bobin and Candés,**  

**mf-IPM**, Matrix-free IPM  
**Fountoulakis, G. and Zhlobich,**  
[http://www.maths.ed.ac.uk/ERGO/](http://www.maths.ed.ac.uk/ERGO/)

Edinburgh, May 2, 2013
2nd-order methods for $\ell_1$-regularization

**SPARCO problems**

Comparison on 18 out of 26 classes of problems (all but 6 complex and 2 installation-dependent ones).

**Solvers compared:**


On 36 runs (noisy and noiseless problems), **mf-IPM**:

- is the fastest on 11,
- is the second best on 14, and
- overall is very robust.

Edinburgh, May 2, 2013
Example Two

- **Google** Problem
  with K. Woodsend
2nd-order methods for $\ell_1$-regularization

Ranking of nodes in networks

PageRank
2nd-order methods for $\ell_1$-regularization

Google Problem joint work with Kristian Woodsend

An adjacency matrix $G \in \mathbb{R}^{n \times n}$ of web-page links is given (web-pages are the nodes). $G$ is column-stochastic.

Teleportation:

$$M = \lambda G + (1 - \lambda) \frac{1}{n} ee^T,$$

with $\lambda \in (0, 1)$, usually $\lambda = 0.85$.

Find the dominant right eigenvector $x$ of $M$ with eigenvalue equal to 1

$$Mx = x,$$ such that $e^T x = 1$, $x \geq 0$.

and use $x$ as a ranking vector.

Edinburgh, May 2, 2013
Google Problem

\[
\begin{align*}
\min & \quad \frac{1}{2} \|Mx - x\|_2^2 \\
\text{s.t.} & \quad e^T x = 1, \quad x \geq 0
\end{align*}
\]

Rearrange:

\[
\|Mx - x\|_2^2 = x^T (M - I)^T (M - I) x
\]

to produce a standard QP formulation with

\[
Q = (M - I)^T (M - I)
\]

A very easy QP problem!
2nd-order methods for $\ell_1$-regularization

Preconditioner for Google Problem

Approximate

$$\mathcal{M} = \begin{bmatrix} Q + \Theta^{-1} & e \\ e^T & 0 \end{bmatrix}$$

with

$$\mathcal{P} = \begin{bmatrix} D_Q & e \\ e^T & 0 \end{bmatrix},$$

where $D_Q = \text{diag}\{Q + \Theta^{-1}\}$.

G., Woodsend
Matrix-free IPM for Google Problems,

Edinburgh, May 2, 2013
2nd-order methods for $\ell_1$-regularization

**Computational Results: mf-IPM**

<table>
<thead>
<tr>
<th>Size</th>
<th>Degree</th>
<th>IPM-itors</th>
<th>MatVecs</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda = 0.85$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16k</td>
<td>20</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>64k</td>
<td>20</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>256k</td>
<td>20</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1M</td>
<td>20</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>$\lambda = 1.0$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16k</td>
<td>20</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>64k</td>
<td>20</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>256k</td>
<td>20</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>1M</td>
<td>20</td>
<td>3</td>
<td>14</td>
</tr>
</tbody>
</table>

**mf-IPM** faster than Nesterov’s coordinate descent. Nesterov (SIOPT 2012) solves them in 45-70 MatVecs.
2nd-order methods for $\ell_1$-regularization

Real-life Networks
Stanford Large Network Dataset Collection
http://snap.stanford.edu/data/

<table>
<thead>
<tr>
<th>Data set</th>
<th>Nodes</th>
<th>Edges</th>
<th>Conjugate</th>
<th>Precond.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Iters</td>
<td>Iters</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>IPM</td>
<td>IPM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CG</td>
<td>PCG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>time</td>
<td>time</td>
</tr>
<tr>
<td>p2p-Gnutella04</td>
<td>10879</td>
<td>50873</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>233</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.7</td>
<td>2.3</td>
</tr>
<tr>
<td>p2p-Gnutella24</td>
<td>26518</td>
<td>91887</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>214</td>
<td>161</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7.3</td>
<td>6.3</td>
</tr>
<tr>
<td>p2p-Gnutella25</td>
<td>22687</td>
<td>77392</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>216</td>
<td>143</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5.4</td>
<td>4.6</td>
</tr>
<tr>
<td>p2p-Gnutella30</td>
<td>36682</td>
<td>125010</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>196</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8.4</td>
<td>7.2</td>
</tr>
<tr>
<td>p2p-Gnutella31</td>
<td>62586</td>
<td>210478</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>205</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>15.6</td>
<td>13.6</td>
</tr>
<tr>
<td>soc-Epinions1</td>
<td>75888</td>
<td>584725</td>
<td>28</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>588</td>
<td>459</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>31.9</td>
<td>48.6</td>
</tr>
<tr>
<td>amazon0601</td>
<td>403394</td>
<td>3790782</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>191</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>76.1</td>
<td>49.3</td>
</tr>
<tr>
<td>web-Google</td>
<td>916428</td>
<td>6021467</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>193</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>197.1</td>
<td>85.6</td>
</tr>
<tr>
<td>wiki-Talk</td>
<td>2394385</td>
<td>7415795</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>110</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>256.8</td>
<td>198.9</td>
</tr>
<tr>
<td>web-BerkStan</td>
<td>685231</td>
<td>8285826</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>204</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>106.7</td>
<td>52.0</td>
</tr>
</tbody>
</table>

The number of CG (PCG) iterations is equal to the number of matrix-vector products.
Example Three

- Machine Learning Problems
  with K. Fountoulakis
2nd-order methods for $\ell_1$-regularization

**Machine Learning Problems** joint work with Kimon Fountoulakis.


2nd-order methods for $\ell_1$-regularization

Huge-Scale LASSO problem

$$\min_x \tau \|x\|_1 + \frac{1}{2} \|Ax - b\|_2^2,$$

where $A \in \mathbb{R}^{m \times n}$ ($m = 2n$: overdetermined system).

Dimensions: $m = 4 \times 10^9$, $n = 2 \times 10^9$.

Very sparse: 20 nonzero entries per column.

- **Parallel CD** (Richtárik and Takáč)
solves it doing 34-37 scans through the matrix 35 iterations, CPU time: 10779s;

- **Truncated Newton** (Fountoulakis and G.)
solves it using 12-13 matrix-vector multiplications 13 iterations, CPU time: 5079s.
2nd-order methods for $\ell_1$-regularization

Trivial problem

$$\min_x \tau \|x\|_1 + \frac{1}{2} \|Ax - b\|_2^2,$$

where $A \in \mathbb{R}^{m \times n}$. Highly overdetermined system: $m = 2n$.

Strongly diagonally dominant matrix $A^T A$.

$$A^T A = \begin{bmatrix}
  x & x & x \\
  x & x & x \\
  x & x & x \\
\end{bmatrix} = \begin{bmatrix}
  d & 0 \\
  0 & d \\
\end{bmatrix}$$
2nd-order methods for $\ell_1$-regularization

More Machine Learning Problems

<table>
<thead>
<tr>
<th>Problem</th>
<th>Features</th>
<th>Training size</th>
<th>CPU time</th>
</tr>
</thead>
<tbody>
<tr>
<td>gisette_scale</td>
<td>5,000</td>
<td>6,000</td>
<td>11.65</td>
</tr>
<tr>
<td>real_sim</td>
<td>20,958</td>
<td>72,309</td>
<td>1.85</td>
</tr>
<tr>
<td>epsilon</td>
<td>2,000</td>
<td>400,000</td>
<td>1,658</td>
</tr>
<tr>
<td>rcv1_train</td>
<td>47,236</td>
<td>20,242</td>
<td>0.77</td>
</tr>
<tr>
<td>news20_binary</td>
<td>1,355,191</td>
<td>19,996</td>
<td>3.30</td>
</tr>
</tbody>
</table>

**CD** Coordinate Descent, **Chih-Jen Lin**, Liblinear: [http://www.csie.ntu.edu.tw/~cjlin/liblinear/](http://www.csie.ntu.edu.tw/~cjlin/liblinear/)

**TN** Truncated Newton Meth, **Fountoulakis and G.**

Edinburgh, May 2, 2013
2nd-order methods for $\ell_1$-regularization

What is going on? Linear Algebra Viewpoint

$$\min_x \tau \|x\|_1 + \frac{1}{2} \|Ax - b\|^2_2,$$

Quadratic Opt. with $Q = A^TA$. For overdetermined systems ($m>n$), $Q$ is likely to be very well conditioned.

**Small exercise:**
Ignore $\ell_1$ term and compute:
$$\nabla \phi(x) = A^T(Ax - b)$$ and $$\nabla^2 \phi(x) = A^TA$$
$$d_{SD} = -\nabla \phi(x)$$ and $$d_N = -(\nabla^2 \phi(x))^{-1} \nabla \phi(x)$$

If $\nabla^2 \phi(x) \approx I$ then $d_{SD} \approx d_N$. 

Edinburgh, May 2, 2013
Conclusions

The **2nd-order information** can (sometimes should) be used also in trivial optimization.

Achievable by using **inexact Newton directions** in:

- IPMs
- continuation approach

Final Comments

- **large/huge** does not always mean **difficult**
- Many **Big Data** problems are trivial!