A Lower Complexity Bound for $\ell_1$-Regularized Least-Squares Problems using a Certain Class of Algorithms

Presentation by Tobias Lindstrøm Jensen

Joint work with Jan Østergaard

29th of June, 2011
Aalborg University, Denmark

SPARS 11, Edinburgh
Outline

• Motivation and Definition
• Algorithm Coverage
• Result and Discussion
Background

- Popular to solve the following convex problem

\[
\text{minimize } f(x) = \frac{1}{2} x^T P x - c^T x + \gamma \|x\|_1
\]  \hspace{1cm} (1)

- Notation: \( f(x) = g(x) + h(x), \) \( g(x) = \frac{1}{2} x^T P x - c^T x \) and \( h(x) = \gamma \|x\|_1 \)

- Application: Anything sparse! (subproblems)

- Focus on large-scale problems
Algorithms - Many

- Many - to name a few

- IST (Daubechies, Defrise & De Mol 2004), GPSR (Figueiredo, Nowak & Wright 2007),

- TwIIST (Bioucas-Dias & Figueiredo 2007), l1–ls (Kim, Koh, Lustig, Boyd & Gorinevsky 2007)

- FPC (Hale, Yin & Zhang 2008), FISTA (Beck & Teboulle 2009)

- FPC–AS (Wen, Yin, Goldfarb & Zhang 2010)

- cont.... can we analyse all these approaches in one framework?
Worst-case Analysis

• Consider all admissible objective functions

\[ \mathcal{F} \]

\[ f \]

• Worst-case analysis: find the “worst” \( f \in \mathcal{F} \), i.e., the most difficult problem to solve
Classical Results

• Minimize a smooth convex function $f$

• Define a class of first-order methods as

$$x^{(k)} \in x^{(0)} + \text{span}\{\nabla f(x^{(0)}), \cdots, \nabla f(x^{(k-1)})\} \subset \mathbb{R}^n$$  \hspace{1cm} (2)

with iteration counter $k$. Convergence $O\left(\frac{1}{k^2}\right)$, $2k + 1 \leq n$ (Nemirovskii & Yudin 1983, Nesterov 2004)

• Interpretation:
  
  – Iterations lie in a subspace defined by the gradients
  – First-order black-box assumption (no additional information than the gradient evaluated at $x^{(j)}$)
Counter Observations

- Many algorithms do use additional knowledge about the problem

- Continuation, active set manipulations, debiasing, proximal-map, reformulations, barrier methods etc.

- Questions
  - How to incorporate this into a model?
  - What is the worst-case iteration complexity?
A Class of Method

- Let \( \mathcal{M} \) be a class of iterative algorithms with \( m_f \in \mathcal{M} \)
- denote the support as \( \text{supp}(x) = \{ i \mid x_i \neq 0 \} \)
- Algorithm

\[
x^{(k)} = m_f \left( x^{(k-1)}; \nabla g(x^{(k-1)}) \right), \quad k = 1, \ldots
\]

with

\[
\text{supp} \left( x^{(k)} \right) \subseteq \text{supp} \left( x^{(k-1)} \right) \cup \text{supp} \left( \nabla g(x^{(k-1)}) \right)
\]

- Interpretation: We use no more “information” about the support than that contained in the gradient
- Is this a sensible definition?
Algorithms Covered I

- IST, TwiIST

\[ x^{(k)} = a_k x^{(k-2)} + b_k x^{(k-1)} + S_{t_k} g(x^{(k-1)} + t_k \nabla g(x^{(k-1)})) \] (5)

- FISTA

\[ x^{(k)} = S_{t_k} g(a_k x^{(k-2)} + b_k x^{(k-1)} + t_k \nabla g(a_k x^{(k-2)} + b_k x^{(k-1)})) \] (6)

- Notice that for the soft thresholding function \( S \)

\[ \text{supp}(S_{\lambda}(x)) \subseteq \text{supp}(x), \quad \lambda \geq 0, x \in \mathbb{R}^n \] (7)

- Interpretation: first-order method with black-box on \( g \) (Nesterov 2007)
Algorithms Covered II

- FPC, FPC–AS

\[
\begin{align*}
1 & \quad k = 1 \\
2 & \quad \text{for } c = 1, 2, \ldots \text{ do} \\
3 & \quad \gamma = \Gamma(c) \\
4 & \quad \text{while stopping criteria do} \\
5 & \quad \begin{cases} 
        x^{(k)} = v(a_k x^{(k-1)} + b_k S t_k \gamma(x^{(k-1)} - t_k \nabla(g^{(k-1)}))) \\
        k = k + 1 
\end{cases} \\
6 & \quad \text{end while} \\
\end{align*}
\]

- Active set manipulations \( v \) (not for all iterations)

\[
\text{supp}(v(x)) \subseteq \text{supp}(x), \quad x \in \mathbb{R}^n \tag{8}
\]
Algorithms Covered III

- l1-ls (truncated Newton method)

\begin{verbatim}
1  c = 0, k = 1
2  while duality gap \leq \epsilon do
3    while \|H_\tau(x^{(c)})x^{(k-1)} - (-\nabla\phi_\tau(x^{(c)}))\|_2 \geq \epsilon_{pcg} do
4      \vdots
5      x^{(k)} = x^{(k-1)} + \alpha_{k-1}p^{(k-1)}
6      \vdots
7    k = k + 1
8  x^{(k)} = x^{(c)} + t_kx^{(k-1)}
9  c = k, k = k + 1
\end{verbatim}

- Only belongs to the class if the preconditioner is “effectively” diagonal
A Problem Instance: The Worst Objective Function

• Similar approach as in (Nesterov 2004)

\[ P = \begin{bmatrix}
  2 & -1 & 0 & 0 & \cdots \\
  -1 & 2 & -1 & 0 \\
  0 & -1 & 2 & \ddots \\
  \vdots & \ddots & \ddots & \ddots & -1 \\
  0 & \cdots & -1 & 2
\end{bmatrix}, \quad c = \begin{bmatrix}
  1 \\
  0 \\
  0 \\
  \vdots \\
  0
\end{bmatrix} \quad (9) \]

• Analytical solution, \( s = |\text{supp}(x^*)| \), select \( \gamma \leftarrow \frac{2}{(s+1)(s+2)} \)

\[ x^*_j = \begin{cases}
  1 - j \frac{2s+3}{(s+1)(s+2)} & \text{for } j = 1, \cdots, s \\
  0 & \text{for } j = s + 1, \cdots, n
\end{cases} \quad (10) \]
Result

• After $k$ iterations with $x^{(0)} = 0$, $\text{supp}(x^{(k)}) \subseteq \{1, \cdots, k\}$ ($P$ is tridiagonal)

• For any

$$k = \begin{cases} \frac{s}{2} & \text{for } s \text{ even} \\ \frac{s-1}{2} & \text{for } s \text{ odd} \end{cases}$$

there exist a function $f(x) = \frac{1}{2}x^T P x - c^T x + \gamma \|x\|_1$, $4 \succeq P \succeq \mu_s$ such that

$$\frac{f(x^{(k)}) - f^*}{\|x^{(0)} - x^*\|_2^2} \geq \frac{1}{6(k+1)^2}$$

for any $m_f \in \mathcal{M}$ with $x^{(0)} = 0$

• Bound is tight (up to a constant) since there exist algorithms that achieve this bound (Nesterov 2007, Beck & Teboulle 2009)
Discussion

• Bound applies to the sparsity degree $s$, not $n$
  – Estimate 1000 non-zeros, then the bound applies up to 500 iterations
  – We need $s$ iterations to be able to fully describe the sparse signal in the standard basis (in worst-case)

• Initialization
  – Let $g(x) = \frac{1}{2} \|Ax - b\|_2^2 \iff P = A^T A, c = A^T b$
  – If $x^{(0)} = A^T b$, this will correspond to a shift of one iteration
  – Problem: The optimization problem (and model) does not allow for arbitrary $x^{(0)}$
Linear Convergence?

- Smooth and strongly convex \( g(x), |S| = s, L \preceq P_S \preceq \mu_s, \kappa(P_S) = \frac{L}{\mu_s} \)

- Why not linear rate for \( g \) smooth, strongly convex objective?
Sub-linear rate is better than linear rate

- for small enough $k$ (and small enough order)

- Example

The provided result bounds $k$, so this is not asymptotic analysis.
Summary

• Provided a common algorithmic framework defined by a class of algorithms

• Shown that many algorithm belongs to this class of methods

• Provided a lower bound of the worst-case complexity
  – Provides information on how well many algorithms can run
  – Warns that it is not possible to obtain better worst-case complexity without designing a method such that $m_f \notin \mathcal{M}$

• How to provide algorithms with better worst-case iteration complexity?
  – Break the model, $m_f \notin \mathcal{M}$
  – Exclude the set of admissible functions
References


