On the analysis of Bregman-surrogate algorithms for large-scale nonconvex optimization

Zhifeng Wang
(*Joint work with Yiyuan She*)

Department of Statistics
Florida State University

Motivation

- Real world challenges:
  - Big data: $p \gg n$
  - Large-scale dataset: millions of unknowns
  - Nonconvexity
  - Second-order methods: unaffordable
- Statistical learning problems $\rightarrow$ optimization
- How to design algorithms to solve **large-scale nonconvex** optimization?

Bregman Surrogate Framework

- **Surrogate function**:

$$g(\boldsymbol{\beta}; \boldsymbol{\beta}^{-}) = f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}, \boldsymbol{\beta}^{-}),$$

  where $\boldsymbol{\Delta}_{\psi}$ is the Bregman notation (define later).

- **Iterative algorithm**:

$$\boldsymbol{\beta}^{(t+1)} \in \arg\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}).$$

  - *We do not assume* $g(\boldsymbol{\beta}; \boldsymbol{\beta}^{-}) \geq f(\boldsymbol{\beta})$.
  - *Any benefit? Step size control and accelerations.*

Challenges

Cheap in each iteration – Numerical stability – Slow convergence

- Convergence rate of **nonconvex** optimization problems?
- Statistical analysis for **high dimensional** problems?
- Accelerations? (not reported today)

Preliminary

Gateaux differential

The (right-hand) *Gateaux differential* of ψ at $\boldsymbol{\beta} \in \Omega$ with increment $\boldsymbol{h}$ is defined as

$$\delta\psi(\boldsymbol{\beta}; \boldsymbol{h}) = \lim_{\epsilon \to 0+} \frac{\psi(\boldsymbol{\beta} + \epsilon \boldsymbol{h}) - \psi(\boldsymbol{\beta})}{\epsilon}.$$

- A relaxed version: $\epsilon \to 0+$
- When $\nabla\psi$ exists, $\delta\psi(\boldsymbol{\beta}, \boldsymbol{h}) = \langle \nabla\psi(\boldsymbol{\beta}), \boldsymbol{h} \rangle$.

The Bregman notation

Our main tool: **Generalized Bregman notation**

$$\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \psi(\boldsymbol{\beta}) - \psi(\boldsymbol{\gamma}) - \delta\psi(\boldsymbol{\gamma}; \boldsymbol{\beta} - \boldsymbol{\gamma}), \ \forall \boldsymbol{\beta}, \boldsymbol{\gamma}.$$

- When $\psi \in \mathcal{C}^{(1)}$, $\boldsymbol{\Delta}_\psi$ becomes the standard *Bregman divergence*:

$$\mathbf{D}_\psi(\boldsymbol{\beta}, \boldsymbol{\gamma}) := \psi(\boldsymbol{\beta}) - \psi(\boldsymbol{\gamma}) - \langle \nabla\psi(\boldsymbol{\gamma}), \boldsymbol{\beta} - \boldsymbol{\gamma} \rangle.$$

- $\mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\gamma}) := \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2/2$.
- In general, $\boldsymbol{\Delta}_\psi$ or $\mathbf{D}_\psi$ may not be symmetric. The following *symmetrized version* turns out to be useful

$$\bar{\boldsymbol{\Delta}}_\psi(\boldsymbol{\beta}, \boldsymbol{\gamma}) := \frac{1}{2}\{\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\gamma}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\gamma}, \boldsymbol{\beta})\}.$$

More intuition on surrogate functions

- Recall the surrogate function:

$$g(\boldsymbol{\beta}; \boldsymbol{\beta}^-) = f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\beta}^-)$$

- Let us assume $\psi \in \mathcal{C}^{(1)}$ for simplicity, then
  - $g(\boldsymbol{\beta}^-; \boldsymbol{\beta}^-) = f(\boldsymbol{\beta}^-)$.
  - $\nabla g(\boldsymbol{\beta}^-; \boldsymbol{\beta}^-) = \nabla f(\boldsymbol{\beta}^-)$.

Some properties of $\boldsymbol{\Delta}_\psi$

Some basic properties of $\boldsymbol{\Delta}$ are given as follows.

1. If ψ is convex, then $\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\gamma}) \geq 0$.
2. $\boldsymbol{\Delta}_{a\psi+b\varphi}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = a\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\gamma}) + b\boldsymbol{\Delta}_\varphi(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\forall a, b \in \mathbb{R}$.
3. **(Idempotence)** $\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot, \boldsymbol{\alpha})}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\gamma})$ under some regularity conditions.

- When $\psi \in \mathcal{C}^{(1)}$, the idempotence property is satisfies.
- The ℓ_1-norm function $\|\cdot\|_1$ has the idempotence property.

Examples of Bregman surrogates

Some examples that can be re-characterized by Bregman surrogates:

1. Gradient Descent & Mirror Descent
2. Iterative Thresholding
3. DC Programming
4. Local Linear Approximation

Example: Gradient Descent

- Gradient descent:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \alpha \nabla f(\boldsymbol{\beta}^{(t)}),$$

  $\alpha > 0$: step size parameter.

- If we use $\boldsymbol{\Delta}_\psi = \rho \mathbf{D}_2 - \boldsymbol{\Delta}_f$, then

$$\begin{aligned}
\boldsymbol{\beta}^{(t+1)} &= \arg\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}) := f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}) \\
&= \arg\min_{\boldsymbol{\beta}} \{ f(\boldsymbol{\beta}^{(t)}) + \langle \nabla f(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(t)} \rangle + \rho \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}) \} \\
&= \boldsymbol{\beta}^{(t)} - \frac{1}{\rho} \nabla f(\boldsymbol{\beta}^{(t)}).
\end{aligned}$$

Example: Mirror Descent

- If φ is strictly convex,

$$g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}) = f(\boldsymbol{\beta}) + (\rho \mathbf{D}_{\varphi} - \boldsymbol{\Delta}_f)(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}).$$

- $\boldsymbol{\beta}^{(t+1)} = \min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)})$

  $\rightarrow$ mirror descent algorithm (Nemirovski and Yudin, 1983):

$$\boldsymbol{\beta}^{(t+1)} = (\nabla \varphi)^{-1}(\nabla \varphi(\boldsymbol{\beta}^{(t)}) - \nabla f(\boldsymbol{\beta}^{(t)})/\rho),$$

Example: Iterative Thresholding

- General penalized problem: $\min_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) + P_\Theta(\varrho\boldsymbol{\beta}; \lambda)$
- Iterative thresholding (Blumensath and Davies, 2009; She, 2009):

$$\boldsymbol{\beta}^{(t+1)} = \frac{1}{\varrho}\Theta(\varrho\boldsymbol{\beta}^{(t)} - \frac{1}{\varrho}\nabla l(\boldsymbol{\beta}^{(t)}); \lambda),$$

  Θ: a thresholding function inducing P_Θ.

- From Bregman surrogate perspective,

$$g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}) = l(\boldsymbol{\beta}) + P(\varrho\boldsymbol{\beta}; \lambda) + (\varrho^2\mathbf{D}_2 - \boldsymbol{\Delta}_l)(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}).$$

Example: DC Programming

- "Difference of convex" (DC) function $f(\boldsymbol{\beta}) = d_1(\boldsymbol{\beta}) - d_2(\boldsymbol{\beta})$.
- d_1 and d_2 are both convex.
- Standard DC algorithm:

$$\boldsymbol{\gamma}^{(t)} \in \partial d_2(\boldsymbol{\beta}^{(t)}),\ \boldsymbol{\beta}^{(t+1)} \in \partial d_1^*(\boldsymbol{\gamma}^{(t)}), \tag{1}$$

  $\partial d(\boldsymbol{\beta})$: the subdifferential of $d(\cdot)$ at $\boldsymbol{\beta}$.

  $d_1^*(\cdot)$: the Fenchel conjugate function of $d_1(\cdot)$.

- Choosing some *special* $\boldsymbol{\beta}^{(t+1)}$ and $\boldsymbol{\gamma}^{(t)}$ can correspond to using the following Bregman surrogate

$$g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}) = f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_{d_2}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}).$$

Example: Local Linear Approximation

- The problem $\min_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) + \sum_j P(\beta_j)$,

  l: Gateaux differentiable,

  P: concave and differentiable over $(0, +\infty)$, and $P(t) = P(-t)$ for any $t \in \mathbb{R}$, $P(0) = 0$.

- The surrogate in LLA (Zou and Li, 2008):
  $l(\boldsymbol{\beta}) + \sum_j \left[P(|\beta_j^{(t)}|) + P'_+(|\beta_j^{(t)}|)(|\beta_j| - |\beta_j^{(t)}|)\right]$.

- The equivalent Bregman surrogate:

$$g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}) = l(\boldsymbol{\beta}) + \sum_j P(\beta_j) + \sum_j \left[\alpha_j \boldsymbol{\Delta}_1(\beta_j, \beta_j^{(t)}) - \boldsymbol{\Delta}_P(\beta_j, \beta_j^{(t)})\right],$$

  with $\boldsymbol{\Delta}_1(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \boldsymbol{\Delta}_{\|\cdot\|_1}(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\alpha_j = |P'_+(\beta_j^{(t)})|$.

Surrogate Algorithm Analysis

- Bregman-surrogate algorithm:

$$\boldsymbol{\beta}^{(t+1)} \in \arg\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}) := f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}), \qquad (2)$$

- f, ψ: can be nonsmooth and nonconvex.
- Assumptions:
  - f: Gateaux differentiable
  - ψ satisfies the idempotence property.

Computational accuracy

This part studies the *numerical error* of $\boldsymbol{\beta}^{(t)}$ in the **optimization process**.

1. General setting
   - Function-value convergence
   - Faster convergence with *strong convexity*
   - Our **relaxed** conclusion using proper error measure
2. Composite setting
   - Iterative thresholding
   - Local linear approximation

Computational accuracy: General setting

Theorem 1

For any $\boldsymbol{\beta}$ satisfying

$$\boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}) + \boldsymbol{\Delta}_{f}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t+1)}) \geq 0, \ 0 \leq t \leq T, \tag{3}$$

we have

$$\underset{0 \leq t \leq T}{\text{avg}} \ f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}) \leq \frac{1}{T+1} [\boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(0)}) - \boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(T+1)})].$$

- Condition (3): step size control.
- $\mathcal{O}(1/T)$: no slower than gradient descent algorithms.

Computational accuracy: General setting

A faster (linear) rate of convergence with strong convexity.

Theorem 2

Assume that $\boldsymbol{\Delta}_\psi \geq 0$. *Let* $\boldsymbol{\beta}^o$ *be any (global) minimizer of* $f(\boldsymbol{\beta})$. *If*

$$2\bar{\boldsymbol{\Delta}}_f \geq \varepsilon \boldsymbol{\Delta}_\psi \tag{4}$$

for some $\varepsilon > 0$, *then*

$$\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(T+1)}) \leq \left(\frac{1}{1+\varepsilon}\right)^{T+1} \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(0)})$$

for any $T \geq 0$. *Alternatively, if for some* $\kappa > 1$,

$$\bar{\boldsymbol{\Delta}}_\phi \geq \frac{\kappa}{\kappa - 1} \boldsymbol{\Delta}_\psi,$$

where $\boldsymbol{\Delta}_\phi = \boldsymbol{\Delta}_\psi + \boldsymbol{\Delta}_f$, *then for any* $T \geq 0$, *we have*

$$\bar{\boldsymbol{\Delta}}_\phi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(T+1)}) \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^{T+1} \bar{\boldsymbol{\Delta}}_\phi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(0)}).$$

Computational accuracy: General setting

- Ben-Tal and Nemirovski (2001) used $\min_{1\le t\le T} \|\nabla f(\boldsymbol{\beta}^{(t)})\|^2$ as the convergence measure in gradient descent studies for nonconvex problems.
- Motivated by this, we choose a proper measure of "stationarity" to state a **more general** result.

Theorem 3

For any $T \ge 1$, the Bregman-surrogate algorithm (2) satisfies

$$\operatorname*{avg}_{0\le t\le T} (2\bar{\boldsymbol{\Delta}}_\psi + \boldsymbol{\Delta}_f)(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) \le \frac{1}{T+1}\left[f(\boldsymbol{\beta}^{(0)}) - f(\boldsymbol{\beta}^{(T+1)})\right].$$

Computational accuracy: Composite setting

- High-dimensional problems

$$\min_{\boldsymbol{\beta}} l_0(\boldsymbol{X}\boldsymbol{\beta}) + P(\varrho\boldsymbol{\beta}; \lambda),$$

  $\boldsymbol{X}$: the predictor matrix,
  l_0: loss function defined on $\boldsymbol{X}\boldsymbol{\beta}$ (and so $l(\boldsymbol{\beta}) = l_0(\boldsymbol{X}\boldsymbol{\beta})$),
  P: sparsity-inducing regularizer.

- As an illustration, we use **Theorem 3** to prove some results for iterative thresholding and LLA.

Computational accuracy: Iterative thresholding

Many popularly used penalty functions are associated with thresholdings, such as the $\ell_r (0 < r \leq 1)$ (Frank and Friedman, 1993; She, 2016), ℓ_2, SCAD (Fan and Li, 2001), MCP (Zhang, 2010a), capped ℓ_1 (Zhang, 2010b), ℓ_0, elastic net (Zou and Hastie, 2005), Berhu (Owen, 2007; He et al., 2013).

Definition 1 (Thresholding function (She, 2009))

A threshold function is a real-valued function $\Theta(t; \lambda)$ *defined for* $-\infty < t < \infty$ *and* $0 \leq \lambda < \infty$ *such that (i)* $\Theta(-t; \lambda) = -\Theta(t; \lambda)$*; (ii)* $\Theta(t; \lambda) \leq \Theta(t'; \lambda)$ *for* $t \leq t'$*; (iii)* $\lim_{t \to \infty} \Theta(t; \lambda) = \infty$*; (iv)* $0 \leq \Theta(t; \lambda) \leq t$ *for* $0 \leq t < \infty$*.*

Computational accuracy: Iterative thresholding

- A critical number $\mathcal{L}_\Theta \leq 1$ can be introduced such that $\mathrm{d}\Theta^{-1}(u;\lambda)\,\mathrm{d}u \geq 1 - \mathcal{L}_\Theta$ for almost every $u \geq 0$, or

$$\mathcal{L}_\Theta = 1 - \operatorname{ess\,inf}\{\mathrm{d}\Theta^{-1}(u;\lambda)/\mathrm{d}u : u \geq 0\},$$

  where ess inf is the essential infimum.
- The corresponding penalty is defined by

$$P_\Theta(t;\lambda) = \int_0^{|t|} (\Theta^{-1}(u;\lambda) - u)\,\mathrm{d}u,\ \forall t \in \mathbb{R}.$$

Computational accuracy: Iterative thresholding

- Recall the problem $\min_{\boldsymbol{\beta}} l_0(\boldsymbol{X}\boldsymbol{\beta}) + P(\varrho\boldsymbol{\beta}; \lambda)$.
- Iterative thresholding: $\boldsymbol{\beta}^{(t+1)} = \Theta(\varrho\boldsymbol{\beta}^{(t)} - \nabla l(\boldsymbol{\beta}^{(t)})/\varrho; \lambda)/\varrho$.
- For any $T \geq 1$

$$\underset{0\leq t\leq T}{\text{avg}} \, (\varrho^2(2 - \mathcal{L}_\Theta)\mathbf{D}_2 - \grave{\boldsymbol{\Delta}}_l)(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) \leq \frac{1}{T+1}\left[f(\boldsymbol{\beta}^{(0)}) - f(\boldsymbol{\beta}^{(T+1)})\right],$$

where $\grave{\boldsymbol{\Delta}}_l\,(\boldsymbol{\beta}, \boldsymbol{\gamma}) := \boldsymbol{\Delta}_l(\boldsymbol{\gamma}, \boldsymbol{\beta})$.

Computational accuracy: LLA

- Recall the LLA surrogate

$$g^{(t)}_{\mathrm{LLA}}(\boldsymbol{\beta};\boldsymbol{\beta}^{(t)}) = l(\boldsymbol{\beta}) + P(\varrho\boldsymbol{\beta}) + \boldsymbol{\Delta}_{\|\boldsymbol{\alpha}^{(t)}\circ(\cdot)\|_1 - P(\cdot)}(\varrho\boldsymbol{\beta}, \varrho\boldsymbol{\beta}^{(t)}),$$

  where $\boldsymbol{\alpha}^{(t)} = [\alpha_j^{(t)}]$ with $\alpha_j^{(t)} = |P'_+(\beta_j^{(t)})|, 1 \le j \le p$.

- We abbreviate $\boldsymbol{\Delta}_{\|\boldsymbol{\alpha}^{(t)}\circ(\cdot)\|_1 - P(\cdot)}$ to $\boldsymbol{\Delta}^{(t)}_{\mathrm{LLA}}$.

- $\boldsymbol{\Delta}^{(t)}_{\mathrm{LLA}}$ does **not** have the general idempotence property, but $\boldsymbol{\Delta}_{\boldsymbol{\Delta}^{(t)}_{\mathrm{LLA}}(\cdot,\boldsymbol{\beta})}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \boldsymbol{\Delta}^{(t)}_{\mathrm{LLA}}(\boldsymbol{\beta},\boldsymbol{\gamma})$.

Computational accuracy: LLA

- P is differentiable over $(0, +\infty)$, $P(t) = P(-t)$ and $P(0) = 0$.
- The LLA iterates satisfy

$$\operatorname*{avg}_{0 \le t \le T} [2\bar{\boldsymbol{\Delta}}^{(t)}_{\mathrm{LLA}}(\varrho\boldsymbol{\beta}^{(t)}, \varrho\boldsymbol{\beta}^{(t+1)}) + \boldsymbol{\Delta}_f(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)})] \le \frac{f(\boldsymbol{\beta}^{(0)}) - f(\boldsymbol{\beta}^{(T+1)})}{T+1}$$

for any $T \ge 1$.

Statistical accuracy

- The *target* is the **statistical truth**, denoted by $\boldsymbol{\beta}^*$.
- We will study the statistical error of
  - an estimate $\hat{\boldsymbol{\beta}}$ as a *fixed point*,
  - the t-*th iterate* $\boldsymbol{\beta}^{(t)}$ as t increases.
- We focus on the high-dimensional sparse problem

$$\min_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) + P_\Theta(\varrho\boldsymbol{\beta}; \lambda),$$

- Bregman-surrogate algorithm:

$$\boldsymbol{\beta}^{(t+1)} \in \arg\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}) := l(\boldsymbol{\beta}) + P_\Theta(\varrho\boldsymbol{\beta}; \lambda) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)})$$

Statistical accuracy: effective noise

- We assume that $l \in \mathcal{C}^{(1)}$.
- Given the statistical truth $\boldsymbol{\beta}^*$, we define the *effective noise* by

$$\boldsymbol{\epsilon} = -\nabla l_0(\boldsymbol{X}\boldsymbol{\beta}^*).$$

- In the noise-free case, $-\nabla l_0(\boldsymbol{X}\boldsymbol{\beta})$ vanishes at the true $\boldsymbol{\beta}^*$.
- In the following theorems, we assume that $\boldsymbol{\epsilon}$ is a *sub-Gaussian* random vector with mean zero and scale bounded by σ.

Statistical accuracy: some notations

- The support of $\boldsymbol{\beta}$: $\mathcal{J}(\boldsymbol{\beta}) = \{j : \beta_j \neq 0\}$.
- The cardinality: $J(\boldsymbol{\beta}) = |\mathcal{J}(\boldsymbol{\beta})| = \|\boldsymbol{\beta}\|_0$.
- We abbreviate $J(\boldsymbol{\beta}^*)$ to J^* and $J(\hat{\boldsymbol{\beta}})$ to $\hat{J}$.
- $P_H(t;\lambda) = (-t^2/2 + \lambda|t|)1_{|t|<\lambda} + (\lambda^2/2)1_{|t|\geq\lambda}$ is the penalty induced by the hard-thresholding $\Theta_H(t;\lambda) = t1_{|t|>\lambda}$.

Statistical accuracy: fixed-point solutions

Fixed points: $\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}})$.

Theorem 4

Suppose that given $\boldsymbol{\beta} \in \mathbb{R}^p$, there exist $\delta > 0$, $\vartheta > 0$ and large enough $K \geq 0$ such that

$$\mathcal{L}_\Theta \mathbf{D}_2(\varrho\boldsymbol{\beta}, \varrho\boldsymbol{\beta}') + \delta \mathbf{D}_2(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\boldsymbol{\beta}') + \vartheta P_H(\varrho(\boldsymbol{\beta}' - \boldsymbol{\beta}); \lambda) + P_\Theta(\varrho\boldsymbol{\beta}; \lambda)$$
$$\leq \boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}') + P_\Theta(\varrho\boldsymbol{\beta}'; \lambda) + K\lambda^2 J(\boldsymbol{\beta})$$

holds for any $\boldsymbol{\beta}' \in \mathbb{R}^p$. Let $\lambda = A\sigma\sqrt{\log(ep)}/\sqrt{(\delta \wedge \vartheta)\vartheta}$ with A a sufficiently large constant. Then the following oracle inequality holds with probability at least $1 - Cp^{-cA^2}$

$$\boldsymbol{\Delta}_l(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \leq \boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + \frac{KA^2}{(\delta \wedge \vartheta)\vartheta}\sigma^2 J(\boldsymbol{\beta}) \log(ep), \tag{5}$$

where C, c are universal positive constants.

Statistical accuracy: fixed-point solutions

- (5) gives a *sharp* oracle inequality (Koltchinskii, 2011).
- In regression, to get the rate result

$$\boldsymbol{\Delta}_l(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \lesssim \boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + \sigma^2 J(\boldsymbol{\beta}) \log(ep),$$

the regularity conditon can be relaxed to (with ε redefined)

$$\begin{aligned} &\mathcal{L}_\Theta \mathbf{D}_2(\varrho\boldsymbol{\beta}, \varrho\boldsymbol{\beta}') + \delta \mathbf{D}_2(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\boldsymbol{\beta}') + \vartheta P_H(\varrho(\boldsymbol{\beta}' - \boldsymbol{\beta}); \lambda) + P_\Theta(\varrho\boldsymbol{\beta}; \lambda) \\ &\leq (2 - \varepsilon)\boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}') + P_\Theta(\varrho\boldsymbol{\beta}'; \lambda) + K\lambda^2 J(\boldsymbol{\beta}). \end{aligned} \tag{6}$$

Statistical accuracy: fixed-point solutions

- To show that (6) is less demanding than some commonly used regularity conditions in the literature, we assume that
  1. $l(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/2$,
  2. P_Θ is sub-additive: $P_\Theta(t+s) \leq P_\Theta(t) + P_\Theta(s)$, which holds when it is concave on $[0, +\infty)$, such as MCP, SCAD and ℓ_r $(0 \leq r \leq 1)$.
- Then (6) is implied by

$$\begin{aligned} &(1+\vartheta)P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}}; \lambda) + \frac{\mathcal{L}_\Theta}{2}\|\varrho\boldsymbol{\gamma}\|_2^2 \\ \leq\ &\frac{2-\varepsilon'}{2}\|\boldsymbol{X}\boldsymbol{\gamma}\|_2^2 + K\lambda^2 J + (1-\vartheta)P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}^c}; \lambda), \end{aligned} \tag{7}$$

where $\mathcal{J} = \mathcal{J}(\boldsymbol{\beta}), \boldsymbol{\gamma} = \boldsymbol{\beta}' - \boldsymbol{\beta}$ and $\varepsilon' = \varepsilon + \delta$.

Statistical accuracy: fixed-point solutions

- In **lasso** problem where $P_\Theta(\boldsymbol{\beta}; \lambda) = \lambda \|\boldsymbol{\beta}\|_1$ and so $\mathcal{L}_\Theta = 0$, a sufficient condition (with ϑ and K redefined) for (7) to hold is

$$(1 + \vartheta)\varrho \|\boldsymbol{\gamma}_{\mathcal{J}}\|_1 \leq K\sqrt{J}\|\boldsymbol{X}\boldsymbol{\gamma}\|_2 + \varrho\|\boldsymbol{\gamma}_{\mathcal{J}^c}\|_1. \tag{8}$$

- We compare (8) with the following restricted eigenvalue (RE) and the compatibility condition (Bickel et al., 2009; van de Geer and Bühlmann, 2009).
- Given $\mathcal{J} \subset [p]$, there exists positive numbers κ_{RE}, ϑ_{RE} such that

$$J\|\boldsymbol{X}\boldsymbol{\gamma}\|_2^2 \geq \kappa_{RE}\|\boldsymbol{\gamma}_{\mathcal{J}}\|_1^2, \quad \text{(compatibility)}$$

or more restrictively,

$$\|\boldsymbol{X}\boldsymbol{\gamma}\|_2^2 \geq \kappa_{RE}\|\boldsymbol{\gamma}_{\mathcal{J}}\|_2^2, \quad \text{(restricted eigenvalue)}$$

for all $\boldsymbol{\gamma} \in \mathbb{R}^p$ falling into the region

$$(1 + \vartheta_{RE})\|\boldsymbol{\gamma}_{\mathcal{J}}\|_1 \geq \|\boldsymbol{\gamma}_{\mathcal{J}^c}\|_1.$$

Statistical accuracy: the sequence of iterates

The statistical error $\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^*)$: linear convergence w.h.p..

Theorem 5

Suppose that the following inequality holds for some $\delta > 0$, $\varepsilon > 0$, $\vartheta > 0$, $K \geq 0$ and any $\boldsymbol{\beta}$

$$\begin{aligned} &\mathcal{L}_\Theta \mathbf{D}_2(\varrho\boldsymbol{\beta}^*, \varrho\boldsymbol{\beta}) + \delta \mathbf{D}_2(\boldsymbol{X}\boldsymbol{\beta}^*, \boldsymbol{X}\boldsymbol{\beta}) + \vartheta P_H(\varrho(\boldsymbol{\beta} - \boldsymbol{\beta}^*); \lambda) + P_\Theta(\varrho\boldsymbol{\beta}^*; \lambda) \\ &\leq 2\bar{\boldsymbol{\Delta}}_l(\boldsymbol{\beta}^*, \boldsymbol{\beta}) - \varepsilon \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*, \boldsymbol{\beta}) + P_\Theta(\varrho\boldsymbol{\beta}; \lambda) + K\lambda^2 J(\boldsymbol{\beta}^*). \end{aligned} \tag{9}$$

Let $\lambda = A\sigma\sqrt{\log(ep)}/\sqrt{(\delta \wedge \vartheta)\vartheta}$ with A sufficiently large and $\kappa = 1/(1+\varepsilon)$. Then we have

$$\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(t)}) \leq \kappa^t \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(0)}) + \frac{\kappa}{1-\kappa}(K\lambda^2 J^* - \min_{1 \leq s \leq t} \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(s)}, \boldsymbol{\beta}^{(s-1)}))$$

for any $t \geq 1$ with probability at least $1 - Cp^{-cA^2}$.

Statistical accuracy: the sequence of iterates

- (9): a large-p extension of (4) in statistical accuracy, it is generally a bit more demanding than (6).
- Therefore, for regular high dimensional problems, one can **terminate** Bregman-surrogate algorithms earlier without sacrificing much statistical accuracy.
- Similar to the discussion of LLA, we can modify Theorem 5 to show the **linear convergence of LLA** in statistical accuracy under some proper regularity conditions.

Simulations

We will present the convergence of computational error and statistical error for various problems, where Bregman surrogates can be easily derived.

1. Computational error:
   - DC programming for *capped* ℓ_1 *penalized SVM*
   - Mirror descent for *entropy maximization*
2. Statistical error:
   - Iterative thresholding
   - Local linear approximation

Simulations: DC programming

- The problem of capped ℓ_1 penalized SVM:

$$\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) := \sum_{i=1}^{n} l(\boldsymbol{x}_i^\top \boldsymbol{\beta}, y_i) + P_{c1}(\boldsymbol{\beta}; \lambda),$$

where $l(\theta, y) = \max(0, 1 - y\theta)$,
$P_{c1}(\boldsymbol{\beta}; \lambda) = \sum_{j=1}^{p} \min(\lambda|\beta_j|, \lambda^2/2)$.

- DC decomposition: $P_{c1}(\boldsymbol{\beta}; \lambda) = d_1(\boldsymbol{\beta}; \lambda) - d_2(\boldsymbol{\beta}; \lambda)$ with

$$d_1(\boldsymbol{\beta}; \lambda) = \lambda\|\boldsymbol{\beta}\|_1, \quad d_2(\boldsymbol{\beta}; \lambda) = \sum_{j=1}^{p} \max(\lambda|\beta_j| - \lambda^2/2, 0).$$

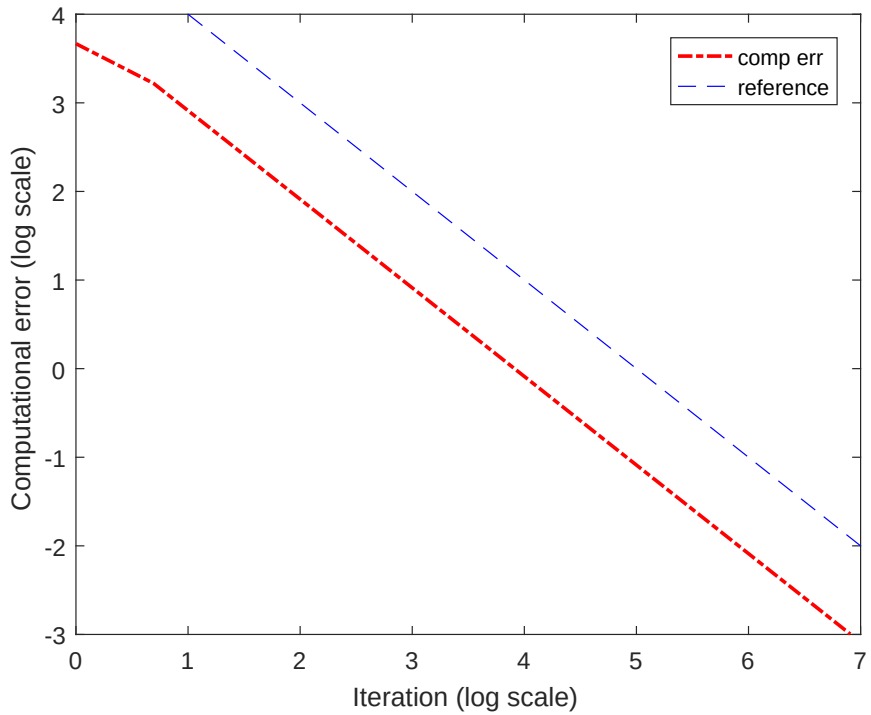
- DC algorithm (Bregman update):

$$\boldsymbol{\beta}^{(t+1)} \in \arg\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_{d_2}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)})$$

$$\in \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \max(0, 1 - y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \left(|\beta_j| - \beta_j 1_{|\beta_j^{(t)}| \geq \lambda/2}\right).$$

Simulations: DC programming

- $\boldsymbol{X} \in \mathbb{R}^{50\times 200} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\Sigma_{ij} = 0.5^{|i-j|}$,
- $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{e}$ with $\boldsymbol{\beta}^* = [3, 5, 7, 0, \ldots, 0]^\top$, $\boldsymbol{e} \sim N(0, 1)$,
- $\lambda = 1$ and $\boldsymbol{\beta}^{(0)} = [0, \ldots, 0]^\top$.



Simulations: Mirror descent

- Nonnegative Burg entropy maximization:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p_+} f(\boldsymbol{\beta}) := \mathrm{IS}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}),$$

  where $\mathrm{IS}(\boldsymbol{a}, \boldsymbol{b}) = \sum_i (a_i/b_i - \log(a_i/b_i) - 1)$ is called the negative cross Burg entropy or Itakura-Saito (IS) divergence.

- A mirror-descent-type surrogate function:

$$g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}) := f(\boldsymbol{\beta}) + (\rho \mathrm{KL} - \mathbf{D}_f)(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}),$$

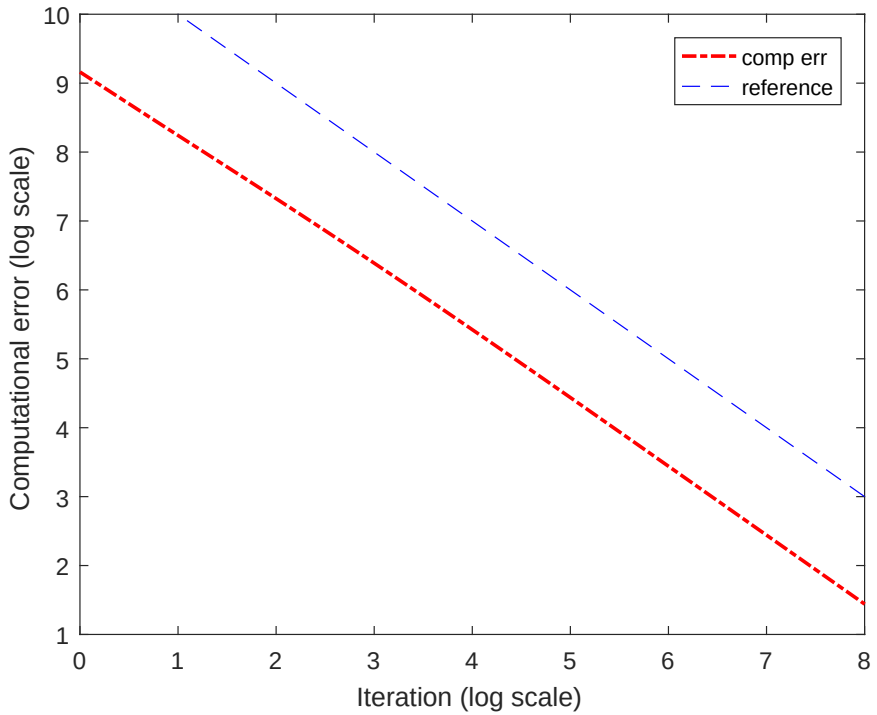
  where $\mathrm{KL}(\cdot, \cdot)$ is the KL divergence.

- Bregman update:

$$\beta_j^{(t+1)} = \beta_j^{(t)} \exp\Big[-\frac{1}{\rho} \sum_i \frac{(\boldsymbol{X}\boldsymbol{\beta}^{(t)})_i - y_i}{(\boldsymbol{X}\boldsymbol{\beta}^{(t)})_i^2} X_{ij} \Big].$$

Simulations: Mirror descent

- $\boldsymbol{X} \in \mathbb{R}^{1000\times 1000}$ with $X_{ij} \sim U(0,1)$,
- $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^*$ with $\boldsymbol{\beta}^* \sim U(0,5)$,
- $\boldsymbol{\beta}^{(0)} \sim U(0,1)$, $\rho = 100$ (large enough for convergence).



Simulations: sparse regression

We consider the following regularized problem using Tukey's loss function and the hard penalty

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} f(\boldsymbol{\beta}) := \sum_{i=1}^{n} l_\psi(\boldsymbol{x}_i^\top \boldsymbol{\beta}, y_i) + \sum_{j=1}^{p} P_H(\varrho\beta_j;\lambda),$$

where $\varrho = \|\boldsymbol{X}\|_2$ is the scaling parameter,

$$P_H(t;\lambda) = \begin{cases} -\frac{1}{2}t^2 + \lambda|t|, & |t| < \lambda \\ \frac{1}{2}\lambda^2, & |t| \geq \lambda, \end{cases}$$

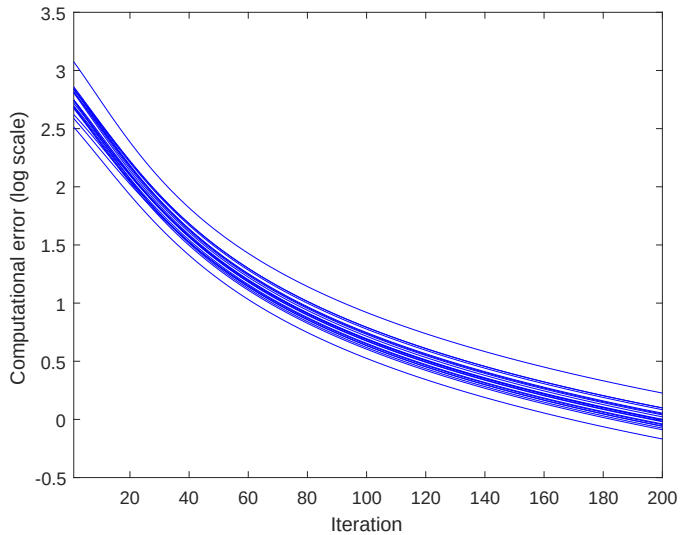
and

$$\psi(t) = \begin{cases} t\left[1 - (\frac{t}{c})^2\right]^2, & |t| \leq c \\ 0, & |t| > c. \end{cases}$$

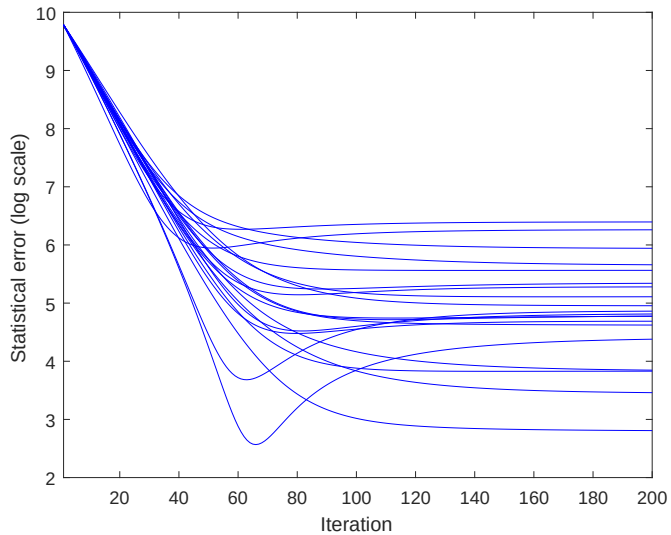
Simulations: sparse regression

Iterative thresholding

- $\boldsymbol{X} \in \mathbb{R}^{n \times p} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ with $n = 1000, p = 15000, \Sigma_{ij} = 0.2^{|i-j|}$.
- $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, 1)$.
- $\boldsymbol{\beta}^* = [0.3, 0.3, 0.4, 0.4, 0.5, 0, \ldots, 0]^\top$ and so $J(\boldsymbol{\beta}^*) = 5$.
- $\lambda = 0.3\sqrt{\log(ep)}$ by extensive experiments.



(a) computational error

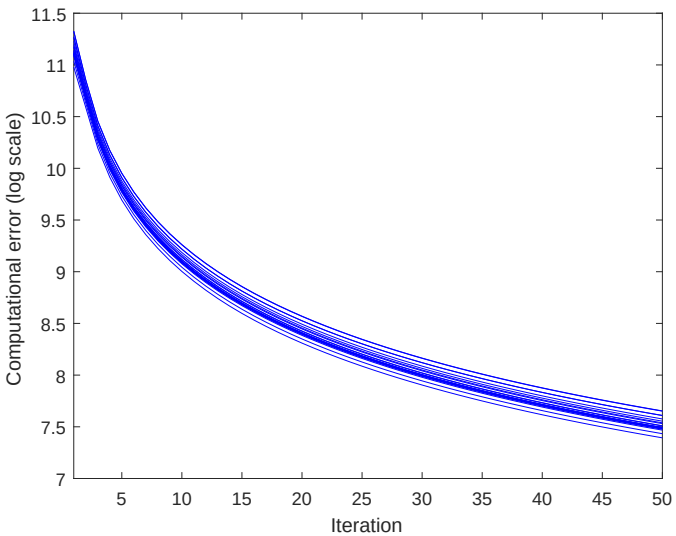


(b) statistical error

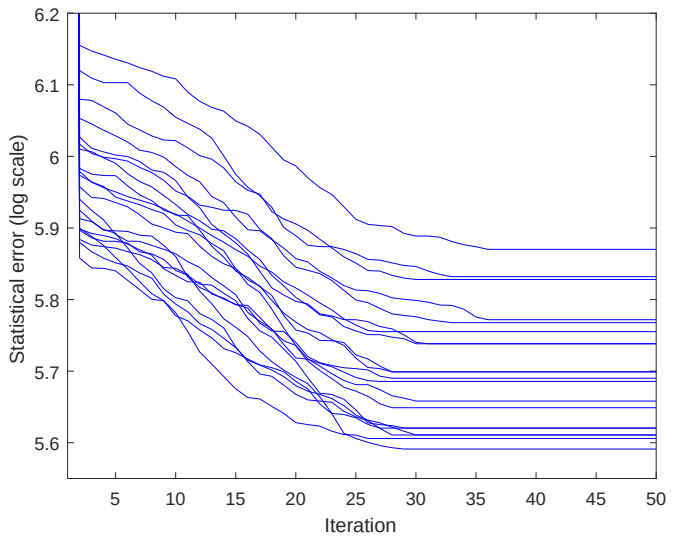
Simulations: sparse regression

Local linear approximation

- $\boldsymbol{X}$ is generated from $N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{ij} = 0.2^{1_{i \neq j}}$.
- $\boldsymbol{\beta}^* = [2, 3, 4, 5, 6, 2, 3, 4, 5, 6, 0, \ldots, 0]^\top$ (so $J^* = 10$).
- $\lambda = 0.5\sqrt{\log(ep)}$ by theoretical results and experiments.
- $\boldsymbol{\beta}^{(0)} = [0 \ldots, 0]^\top$.



(c) computational error



(d) statistical error

References I

Ben-Tal, A. and Nemirovski, A. (2001). *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, pages 1705–1732.

Blumensath, T. and Davies, M. (2009). Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274.

Cichocki, A., ichi Amari, S., Zdunek, R., Kompass, R., Hori, G., and He, Z. (2006). Extended smart algorithms for non-negative matrix factorization . In Rutkowski, L., Tadeusiewicz, R., Zadeh, L. A., and Zurada, J. M., editors, *ICAISC*, volume 4029 of *Lecture Notes in Computer Science*, pages 548–562. Springer.

References II

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96:1348–1360.

Frank, I. E. and Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, 35(2):109–135.

He, Y., She, Y., and Wu, D. (2013). Stationary sparse causality network learning. *J. Mach. Learn. Res.*, 14:3073–3104.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *Amer. Statist*, pages 30–37.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791.

Nemirovski, A. and Yudin, D. (1983). *Problem complexity and method efficiency in optimization*. Wiley-Interscience series in discrete mathematics. Wiley, Chichester, New York.

References III

Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. *Prediction and Discovery (Contemporary Mathematics)*, 443:59–71.

She, Y. (2009). Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic Journal of Statistics*, 3:384–415.

She, Y. (2016). On the finite-sample analysis of Θ-estimators. *Electron. J. Statist.*, 10(2):1874–1895.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392.

Zhang, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942.

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high dimensional sparse estimation problems. *Statist. Sci.*, 27(4):576–593.

References IV

Zhang, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.*, 11:1081–1107.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *JRSSB*, 67(2):301–320.

Zou, H. and Li, R. (2008). One-step Sparse Estimates in Nonconcave Penalized Likelihood Models. *Annals of Statistics*, 36(4):1509–1533.