Super-Linear Convergence of Dual Augmented Lagrangian Algorithm for Sparse Learning

Ryota Tomioka[1], Taiji Suzuki[1], and Masashi Sugiyama[2]

[1]University of Tokyo
[2]Tokyo Institute of Technology

2009-12-12 @ NIPS workshop OPT09

Objective

Develop an optimization algorithm for the optimization problem:

$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{minimize}} \quad \underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w})}_{\text{loss}} + \underbrace{\phi_\lambda(\boldsymbol{w})}_{\text{regularizer}} .$$

For example, lasso:

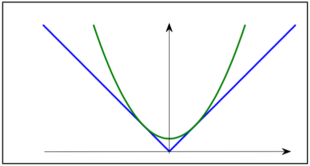
$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{A}\boldsymbol{w}-\boldsymbol{y}\|^2 + \lambda\|\boldsymbol{w}\|_1.$$

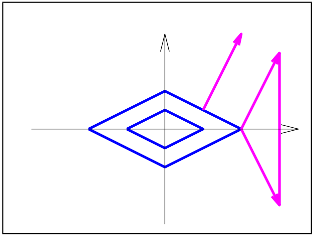
- $\boldsymbol{A} \in \mathbb{R}^{m\times n}$: design matrix ($m$: #observations, n: #unknowns).
- f_ℓ is convex and twice differentiable.
- $\phi_\lambda(\boldsymbol{w})$ is convex but possibly non-differentiable. $\eta\phi_\lambda = \phi_{\eta\lambda}$.
- We are interested in algorithms for general f_ℓ and ϕ_λ ($\leftrightarrow$ LARS).

Where does the difficulty come from?

Conventional view: the non-differentiability of $\phi_\lambda(\boldsymbol{w})$.

- Upper bound the regularizer from above with a differentiable function.
  - FOCUSS (Rao & Kreutz-Delgado, 99)
  - Majorization-Minimization (Figueiredo et al., 07)
  - Iteratively reweighted least squares (IRLS).
- Explicitly handle the non-differentiability.
  - Sub-gradient L-BFGS (Andrew & Gao, 07; Yu et al., 08)





Our view: the coupling between variables introduced by $\boldsymbol{A}$.

Where does the difficulty come from?

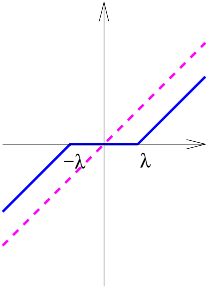
Our view: the coupling between variables introduced by $\boldsymbol{A}$.

In fact, when $\boldsymbol{A} = \boldsymbol{I}_n$

$$\min_{\boldsymbol{w}\in\mathbb{R}^n}\left(\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{w}\|_2^2+\lambda\|\boldsymbol{w}\|_1\right)=\sum_{j=1}^{n}\min_{w_j\in\mathbb{R}}\left(\frac{1}{2}(y_j-w_j)^2+\lambda|w_j|\right).$$

$$\Rightarrow \quad w_j^* = \mathrm{ST}_\lambda(y_j) = \begin{cases} y_j-\lambda & (\lambda\le y_j),\\ 0 & (-\lambda\le y_j\le\lambda),\\ y_j+\lambda & (y_j\le-\lambda).\end{cases}$$

min is obtained analytically!



We focus on ϕ_λ for which the above min can be obtained analytically

Proximation wrt ϕ_λ can be computed analytically

Assumption

Proximation wrt ϕ_λ (soft-thresholding):

$$\mathrm{ST}_\lambda(\boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\phi_\lambda(\boldsymbol{w}) + \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{w}\|_2^2 \right)$$

can be computed analytically.

Outline

1. Introduction
   - Sparse regularized learning.
   - Why is it difficult? *not the non-differentiability*
2. Methods
   - Iterative shrinkage-thresholding (IST)
   - Dual Augmented Lagrangian (porposed)
3. Theoretical results: *super-linear convergence*
   - Exact inner minimization
   - Approximate inner minimization
4. Empirical results
   - Comparison against OWLQN, SpaRSA, and FISTA.
5. Summary

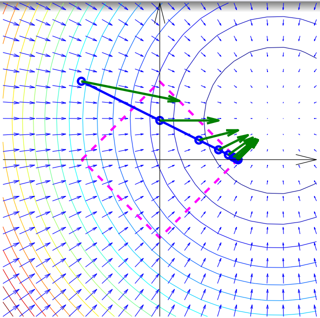
Iterative Shrinkage/Thresholding (IST)

Algorithm (Figueiredo&Nowak, 03; Daubechies et al., 04;...)

1. Choose an initial solution $\boldsymbol{w}^0$.
2. Repeat until some stopping criterion is satisfied:

$$\boldsymbol{w}^{t+1} \leftarrow \underbrace{\mathrm{ST}_{\eta_t\lambda}}_{\text{shrink}}\Bigl(\underbrace{\boldsymbol{w}^t - \eta_t \boldsymbol{A}^\top \nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^t)}_{\text{gradient step}}\Bigr).$$

- Pro: easy to implement.
- Con: bad for poorly conditioned $\boldsymbol{A}$.
- Also known as:
  - Forward-Backward Splitting [Combettes & Wajs, 05]
  - Thresholded Landweber Iteration [Daubechies et al., 04]



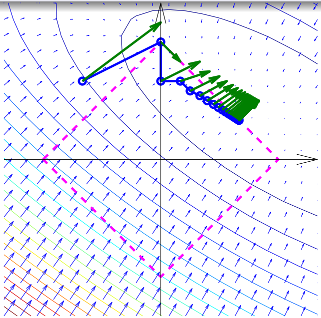
Iterative Shrinkage/Thresholding (IST)

Algorithm (Figueiredo&Nowak, 03; Daubechies et al., 04;...)

1. Choose an initial solution $\boldsymbol{w}^0$.
2. Repeat until some stopping criterion is satisfied:

$$\boldsymbol{w}^{t+1} \leftarrow \underbrace{\mathrm{ST}_{\eta_t\lambda}}_{\text{shrink}}\Big(\underbrace{\boldsymbol{w}^t - \eta_t \boldsymbol{A}^\top \nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^t)}_{\text{gradient step}}\Big).$$

- Pro: easy to implement.
- Con: bad for poorly conditioned $\boldsymbol{A}$.
- Also known as:
  - Forward-Backward Splitting [Combettes & Wajs, 05]
  - Thresholded Landweber Iteration [Daubechies et al., 04]



Dual Augmented Lagrangian (DAL) method

Primal problem

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w}) + \phi_\lambda(\boldsymbol{w})}_{f(\boldsymbol{w})}$$

Proximal minimization:

$$\boldsymbol{w}^{t+1} = \underset{\boldsymbol{w}}{\text{argmin}} \left(f(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right)$$

$$(\eta_0 \leq \eta_1 \leq \cdots)$$

- Easy to analyze.
- $f(\boldsymbol{w}^{t+1}) + \frac{1}{2\eta_t}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2 \leq f(\boldsymbol{w}^t)$.
- Not practical! (as difficult as the original problem)

Dual problem

$$\underset{\boldsymbol{\alpha},\boldsymbol{v}}{\text{maximize}} \quad -f_\ell^*(-\boldsymbol{\alpha}) - \phi_\lambda^*(\boldsymbol{v})$$

$$\text{s.t.} \quad \boldsymbol{v} = \boldsymbol{A}^\top \boldsymbol{\alpha}$$

$\Leftrightarrow$Augmented Lagrangian (Tomioka & Sugiyama, 09):

$$\boldsymbol{w}^{t+1} = \mathrm{ST}_{\lambda\eta_t}(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \boldsymbol{\alpha}^t)$$

$$\boldsymbol{\alpha}^t = \underset{\boldsymbol{\alpha}}{\text{argmin}}\, \varphi_t(\boldsymbol{\alpha})$$

- Minimization of $\varphi_t(\boldsymbol{\alpha})$ is easy (smooth).
- Step-size η_t is increased.
- See Rockafellar 76 for the equivalence.

Dual Augmented Lagrangian (DAL) method

Primal problem

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w}) + \phi_\lambda(\boldsymbol{w})}_{f(\boldsymbol{w})}$$

Proximal minimization:

$$\boldsymbol{w}^{t+1} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left(f(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right)$$

$$(\eta_0 \leq \eta_1 \leq \cdots)$$

- Easy to analyze.
- $f(\boldsymbol{w}^{t+1}) + \frac{1}{2\eta_t}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2 \leq f(\boldsymbol{w}^t)$.
- Not practical! (as difficult as the original problem)

Dual problem

$$\underset{\boldsymbol{\alpha},\boldsymbol{v}}{\text{maximize}} \quad -f_\ell^*(-\boldsymbol{\alpha}) - \phi_\lambda^*(\boldsymbol{v})$$

$$\text{s.t.} \quad \boldsymbol{v} = \boldsymbol{A}^\top \boldsymbol{\alpha}$$

$\Leftrightarrow$Augmented Lagrangian (Tomioka & Sugiyama, 09):

$$\boldsymbol{w}^{t+1} = \mathrm{ST}_{\lambda\eta_t}(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \boldsymbol{\alpha}^t)$$

$$\boldsymbol{\alpha}^t = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\, \varphi_t(\boldsymbol{\alpha})$$

- Minimization of $\varphi_t(\boldsymbol{\alpha})$ is easy (smooth).
- Step-size η_t is increased.
- See Rockafellar 76 for the equivalence.

Dual Augmented Lagrangian (DAL) method

Primal problem

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w}) + \phi_\lambda(\boldsymbol{w})}_{f(\boldsymbol{w})}$$

Proximal minimization:

$$\boldsymbol{w}^{t+1} = \underset{\boldsymbol{w}}{\text{argmin}} \left(f(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right)$$

$$(\eta_0 \le \eta_1 \le \cdots)$$

- Easy to analyze.
- $f(\boldsymbol{w}^{t+1}) + \frac{1}{2\eta_t}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2 \le f(\boldsymbol{w}^t)$.
- Not practical! (as difficult as the original problem)

Dual problem

$$\underset{\boldsymbol{\alpha},\boldsymbol{v}}{\text{maximize}} \quad -f_\ell^*(-\boldsymbol{\alpha}) - \phi_\lambda^*(\boldsymbol{v})$$

$$\text{s.t.} \quad \boldsymbol{v} = \boldsymbol{A}^\top\boldsymbol{\alpha}$$

⇔Augmented Lagrangian (Tomioka & Sugiyama, 09):

$$\boldsymbol{w}^{t+1} = \mathrm{ST}_{\lambda\eta_t}(\boldsymbol{w}^t + \eta_t\boldsymbol{A}^\top\boldsymbol{\alpha}^t)$$

$$\boldsymbol{\alpha}^t = \underset{\boldsymbol{\alpha}}{\text{argmin}}\, \varphi_t(\boldsymbol{\alpha})$$

- Minimization of $\varphi_t(\boldsymbol{\alpha})$ is easy (smooth).
- Step-size η_t is increased.
- See Rockafellar 76 for the equivalence.

Difference: How do we get rid of the couplings?

Proximation wrt f is hard:

$$\boldsymbol{w}^{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \Bigg(\overbrace{\underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w})}_{\text{variables are coupled}} + \phi_\lambda(\boldsymbol{w})}^{f(\boldsymbol{w})} + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \Bigg).$$

- IST: linearly approximates the loss term:

$$f_\ell(\boldsymbol{A}\boldsymbol{w}) \simeq f_\ell(\boldsymbol{A}\boldsymbol{w}^t) + (\boldsymbol{w} - \boldsymbol{w}^t)^\top \boldsymbol{A}^\top \nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^t)$$

$\rightarrow$ tightest at the current point $\boldsymbol{w}^t$

- DAL (proposed): linearly lower-bounds the loss term:

$$f_\ell(\boldsymbol{A}\boldsymbol{w}) = \max_{\alpha \in \mathbb{R}^m} \left(-f_\ell^*(-\alpha) - \boldsymbol{w}^\top \boldsymbol{A}^\top \alpha \right)$$

$\rightarrow$ tightest at the next point $\boldsymbol{w}^{t+1}$

Difference: How do we get rid of the couplings?

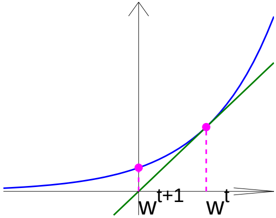
Proximation wrt f is hard:

$$\boldsymbol{w}^{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \Big(\overbrace{\underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w})}_{\text{variables are coupled}} + \phi_\lambda(\boldsymbol{w})}^{f(\boldsymbol{w})} + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \Big).$$

- IST: linearly approximates the loss term:

$$f_\ell(\boldsymbol{A}\boldsymbol{w}) \simeq f_\ell(\boldsymbol{A}\boldsymbol{w}^t) + (\boldsymbol{w} - \boldsymbol{w}^t)^\top \boldsymbol{A}^\top \nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^t)$$

$\rightarrow$ tightest at the current point $\boldsymbol{w}^t$



- DAL (proposed): linearly lower-bounds the loss term:

$$f_\ell(\boldsymbol{A}\boldsymbol{w}) = \max_{\alpha \in \mathbb{R}^m} \left(-f_\ell^*(-\alpha) - \boldsymbol{w}^\top \boldsymbol{A}^\top \alpha \right)$$

$\rightarrow$ tightest at the next point $\boldsymbol{w}^{t+1}$

Difference: How do we get rid of the couplings?

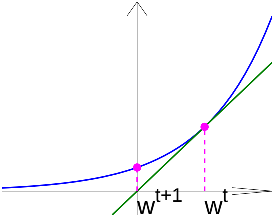
Proximation wrt f is hard:

$$\boldsymbol{w}^{t+1} = \underset{\boldsymbol{w}}{\mathrm{argmin}}\Bigl(\overbrace{\underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w})}_{\text{variables are coupled}} + \phi_\lambda(\boldsymbol{w})}^{f(\boldsymbol{w})} + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2\Bigr).$$

- IST: linearly approximates the loss term:

$$f_\ell(\boldsymbol{A}\boldsymbol{w}) \simeq f_\ell(\boldsymbol{A}\boldsymbol{w}^t) + (\boldsymbol{w} - \boldsymbol{w}^t)^\top \boldsymbol{A}^\top \nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^t)$$

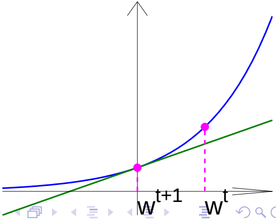
$\rightarrow$ tightest at the current point $\boldsymbol{w}^t$



- DAL (proposed): linearly lower-bounds the loss term:

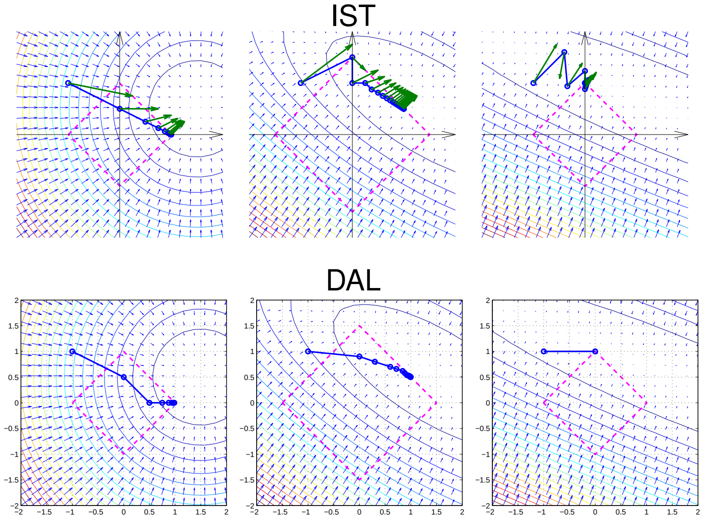
$$f_\ell(\boldsymbol{A}\boldsymbol{w}) = \max_{\boldsymbol{\alpha}\in\mathbb{R}^m}\left(-f_\ell^*(-\boldsymbol{\alpha}) - \boldsymbol{w}^\top \boldsymbol{A}^\top \boldsymbol{\alpha}\right)$$

$\rightarrow$ tightest at the next point $\boldsymbol{w}^{t+1}$



Numerical examples

DAL is better when ***A*** is poorly conditioned.



Theorem 1 (exact minimization)

Definition

- $\boldsymbol{w}^t$: sequence generated by the DAL algorithm with $\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\| = 0$ (exact minimization).
- $\boldsymbol{w}^*$: the unique minimizer of the objective f.

Assumption

There is a constant σ such that

$$f(\boldsymbol{w}^{t+1}) - f(\boldsymbol{w}^*) \geq \sigma\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2 \quad (t = 0, 1, 2, \ldots).$$

Theorem 1

$$\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\| \leq \frac{1}{1 + \sigma\eta_t}\|\boldsymbol{w}^t - \boldsymbol{w}^*\|.$$

I.e., $\boldsymbol{w}^t$ converges super-linearly to $\boldsymbol{w}^*$ if η_t is increasing.

Theorem 2 (approximate minimization)

Definition

- $\boldsymbol{w}^t$: sequence generated by the DAL algorithm with

$$\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\| \leq \sqrt{\frac{\gamma}{\eta_t}}\|\boldsymbol{w}^{t+1}-\boldsymbol{w}^t\| \quad \left(\begin{array}{l}1/\gamma\text{: Lipschitz con-}\\ \text{stant of } \nabla f_\ell.\end{array}\right)$$

Theorem 2

Under the same assumption as in Theorem 1,

$$\|\boldsymbol{w}^{t+1}-\boldsymbol{w}^*\| \leq \frac{1}{\sqrt{1+2\sigma\eta_t}}\|\boldsymbol{w}^t-\boldsymbol{w}^*\|.$$

I.e., $\boldsymbol{w}^t$ converges super-linearly to $\boldsymbol{w}^*$ if η_t is increasing.

Note

- Convergence is slower than the exact case ($\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\| = 0$).
- A faster rate can be obtained if we choose $\frac{\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\|}{\|\boldsymbol{w}^{t+1}-\boldsymbol{w}^t\|} \leq O(1/\eta_t)$.

Theorem 2 (approximate minimization)

Definition

- $\boldsymbol{w}^t$: sequence generated by the DAL algorithm with

$$\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\| \leq \sqrt{\frac{\gamma}{\eta_t}}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\| \quad \left(\begin{array}{l} 1/\gamma\text{: Lipschitz con-} \\ \text{stant of } \nabla f_\ell. \end{array}\right)$$

Theorem 2

Under the same assumption as in Theorem 1,

$$\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\| \leq \frac{1}{\sqrt{1+2\sigma\eta_t}}\|\boldsymbol{w}^t - \boldsymbol{w}^*\|.$$

I.e., $\boldsymbol{w}^t$ converges super-linearly to $\boldsymbol{w}^*$ if η_t is increasing.

Note

- Convergence is slower than the exact case ($\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\| = 0$).
- A faster rate can be obtained if we choose $\frac{\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\|}{\|\boldsymbol{w}^{t+1}-\boldsymbol{w}^t\|} \leq O(1/\eta_t)$.

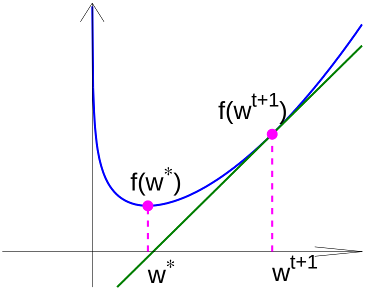
Proof (in essence) of Theorem 1

Since $\boldsymbol{w}^{t+1} = \mathrm{argmin}_{\boldsymbol{w}} \left(f(\boldsymbol{w}) + \frac{1}{2\eta_t} \|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right)$,

$(\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t \in \partial f(\boldsymbol{w}^{t+1})$ (is a subgradient of f). I.e.,

$$f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1}) \geq \left\langle (\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle .$$

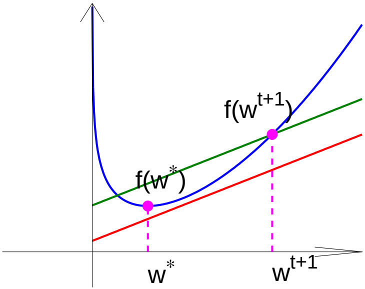
(inspired by Beck & Teboulle 09)



Proof (in essence) of Theorem 2

$$f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1}) \geq \left\langle (\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle \underbrace{-\frac{1}{2\gamma}\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\|^2}_{\text{cost of approximate minimization}} .$$

$1/\gamma$: Lipschitz constant of ∇f_ℓ.



Empirical results: ℓ_1-logistic regression

#samples=$1,024$, #unknowns=$16,384$.

- FISTA

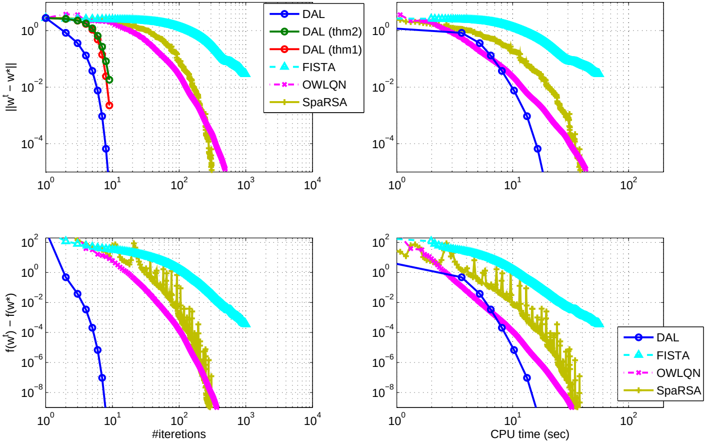
Two-step IST (Beck & Teboulle 09)

- OWLQN

Orthant-wise L-BFGS (Andrew & Gao 07)

- SpaRSA

Step-size improved IST (Wright et al. 09)



Summary

- Why is sparse learning difficult to optimize? – *couplings*
  - Non-differentiability is not bad.
  - Cost of inner minimization $O(m^2 n^+)$ (n^+: number of active variables). Sparsity makes inner minimization efficient.
- How do we get rid of the couplings?
  - Use linear parametric lower bound instead of linear approximation.
- Super-linear convergence for exact/approximate inner minimization.
  - Improved a classic result in optimization by specializing the setting to sparse learning; i.e., proximation wrt ϕ_λ can be performed analytically.
- Empirical results are promising.
  - Faster than OWLQN, SpaRSA, and FISTA with the potential to be generalized further.

Summary

- Why is sparse learning difficult to optimize? – *couplings*
  - Non-differentiability is not bad.
  - Cost of inner minimization $O(m^2 n^+)$ (n^+: number of active variables). Sparsity makes inner minimization efficient.
- How do we get rid of the couplings?
  - Use linear parametric lower bound instead of linear approximation.
- Super-linear convergence for exact/approximate inner minimization.
  - Improved a classic result in optimization by specializing the setting to sparse learning; i.e., proximation wrt ϕ_λ can be performed analytically.
- Empirical results are promising.
  - Faster than OWLQN, SpaRSA, and FISTA with the potential to be generalized further.

Summary

- Why is sparse learning difficult to optimize? – *couplings*
  - Non-differentiability is not bad.
  - Cost of inner minimization $O(m^2 n^+)$ (n^+: number of active variables). Sparsity makes inner minimization efficient.
- How do we get rid of the couplings?
  - Use linear parametric lower bound instead of linear approximation.
- Super-linear convergence for exact/approximate inner minimization.
  - Improved a classic result in optimization by specializing the setting to sparse learning; i.e., proximation wrt ϕ_λ can be performed analytically.
- Empirical results are promissing.
  - Faster than OWLQN, SpaRSA, and FISTA with the potential to be generalized further.

Summary

- Why is sparse learning difficult to optimize? – *couplings*
  - Non-differentiability is not bad.
  - Cost of inner minimization $O(m^2 n^+)$ (n^+: number of active variables). Sparsity makes inner minimization efficient.
- How do we get rid of the couplings?
  - Use linear parametric lower bound instead of linear approximation.
- Super-linear convergence for exact/approximate inner minimization.
  - Improved a classic result in optimization by specializing the setting to sparse learning; i.e., proximation wrt ϕ_λ can be performed analytically.
- Empirical results are promising.
  - Faster than OWLQN, SpaRSA, and FISTA with the potential to be generalized further.

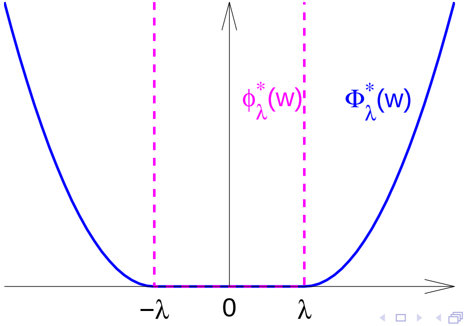
(1) Proximation wrt ϕ_λ is analytic (though non-smooth):

$$\boldsymbol{w}^{t+1} = \mathrm{ST}_{\eta_t\lambda}\left(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \boldsymbol{\alpha}^t\right)$$

(2) Inner minimization is smooth:

$$\boldsymbol{\alpha}^t = \underset{\boldsymbol{\alpha}\in\mathbb{R}^m}{\mathrm{argmin}}\Bigg(\underbrace{f_\ell^*(-\boldsymbol{\alpha})}_{\text{independent of } \boldsymbol{A}.} + \frac{1}{2\eta_t} \underbrace{\|\mathrm{ST}_{\eta_t\lambda}(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \boldsymbol{\alpha})\|_2^2}_{= \Phi_\lambda^*(\cdot)} \Bigg)$$

(linear to the number of active variables)



Comparison to other algorithms

- DAL (this talk)
  $\|\boldsymbol{w}^k - \boldsymbol{w}^*\| = O(\exp(-k))$
- SpaRSA (Step-size improved IST)
  Convergence shown but no rate given. (Wright et al. 09)
- OWLQN (Orthant-wise L-BFGS)
  Convergence shown but no rate given. (Andrew & Gao 07)
- IST (Iterative Soft-thresholding)
  $f(\boldsymbol{w}^k) - f(\boldsymbol{w}^*) = O(1/k)$ (Beck & Teboulle 09)
- FISTA (Two-step IST)
  $f(\boldsymbol{w}^k) - f(\boldsymbol{w}^*) = O(1/k^2)$ (Beck & Teboulle 09)

Comparison to Rockafellar 76

Assumption

The multifunction ∇f^* is (locally) Lipschitz continuous at the origin:

$$\|\nabla f^*(\boldsymbol{\beta}) - \nabla f^*(0)\| \leq L\|\boldsymbol{\beta}\| \quad (\|\boldsymbol{\beta}\| \leq \tau)$$

$\Rightarrow$ Implies our assumption with $\sigma = \frac{1}{2}\min(1/L, \tau/\|\boldsymbol{w}^0 - \boldsymbol{w}^*\|)$.

Convergence (exact minimization) – comparable to Thm 1

$$\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\| \leq \frac{1}{\sqrt{1 + (\eta_t/L)^2}}\|\boldsymbol{w}^t - \boldsymbol{w}^*\|$$

Convergence (approximate minimization) – much worse than Thm 2

$$\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\| \leq \frac{\mu_t + \epsilon_t}{1 - \epsilon_t}\|\boldsymbol{w}^t - \boldsymbol{w}^*\| \quad \left(\mu_t = \frac{1}{\sqrt{1+(\eta_t/L)^2}}\right)$$

(assuming $\|\nabla \varphi_t\| \leq \epsilon_t\sqrt{\gamma/\eta_t}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|$)

Outline of proof of Theorem 1

1. Since $\boldsymbol{w}^{t+1} = \mathrm{argmin}_{\boldsymbol{w}} \left(f(\boldsymbol{w}) + \frac{1}{2\eta_t} \|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right)$, $(\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t$ is a subgradient of f at $\boldsymbol{w}^{t+1}$. I.e.,

$$f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1}) \geq \left\langle (\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle.$$

2. For any $\mu > 0$,

$$\|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\| \|\boldsymbol{w}^t - \boldsymbol{w}^*\| \leq \frac{\mu}{2} \|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^2 + \frac{1}{2\mu} \|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2.$$

3. Combining 1 & 2 and using $f(\boldsymbol{w}^{t+1}) - f(\boldsymbol{w}^*) \geq \sigma \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2$,

$$\frac{1}{2} \|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2 \geq ((1 + \sigma\eta_t)\mu - \frac{\mu^2}{2}) \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2.$$

4. Maximize RHS wrt μ.

Outline of proof of Theorem 1

1. Since $\boldsymbol{w}^{t+1} = \mathrm{argmin}_{\boldsymbol{w}} \left(f(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right)$,
$(\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t$ is a subgradient of f at $\boldsymbol{w}^{t+1}$. I.e.,
$$f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1}) \geq \left\langle (\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle.$$

2. For any $\mu > 0$,
$$\|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|\|\boldsymbol{w}^t - \boldsymbol{w}^*\| \leq \frac{\mu}{2}\|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^2 + \frac{1}{2\mu}\|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2.$$

3. Combining 1 & 2 and using $f(\boldsymbol{w}^{t+1}) - f(\boldsymbol{w}^*) \geq \sigma\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2$,
$$\frac{1}{2}\|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2 \geq ((1+\sigma\eta_t)\mu - \frac{\mu^2}{2})\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2.$$

4. Maximize RHS wrt μ.

Outline of proof of Theorem 1

1. Since $\boldsymbol{w}^{t+1} = \mathrm{argmin}_{\boldsymbol{w}} \left(f(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right)$,
$(\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t$ is a subgradient of f at $\boldsymbol{w}^{t+1}$. I.e.,
$$f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1}) \geq \left\langle (\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle.$$

2. For any $\mu > 0$,
$$\|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|\|\boldsymbol{w}^t - \boldsymbol{w}^*\| \leq \frac{\mu}{2}\|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^2 + \frac{1}{2\mu}\|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2.$$

3. Combining 1 & 2 and using $f(\boldsymbol{w}^{t+1}) - f(\boldsymbol{w}^*) \geq \sigma\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2$,
$$\frac{1}{2}\|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2 \geq ((1 + \sigma\eta_t)\mu - \frac{\mu^2}{2})\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2.$$

4. Maximize RHS wrt μ.

Outline of proof of Theorem 2

1. Let $\boldsymbol{\delta}^t := \nabla\varphi_t(\boldsymbol{\alpha}^t)$,

$$f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1}) \geq \left\langle (\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle - \frac{1}{2\gamma}\|\boldsymbol{\delta}^t\|^2.$$

2. By assumption

$$f(\boldsymbol{w}^{t+1}) - f(\boldsymbol{w}^*) \geq \sigma\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2,$$

$$\|\boldsymbol{\delta}^t\|^2 \leq \frac{\gamma}{\eta_t}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2.$$

3. Combining 1 & 2,

$$\frac{1}{2}\|\boldsymbol{w}^* - \boldsymbol{w}^t\|^2 \geq (\sigma\eta_t + \frac{1}{2})\|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^2.$$

EEG problem – P300 visual speller dataset (subject A)

- Number of samples $m = 2550$.
- 6 class classification.
- $\boldsymbol{w} \in \mathbb{R}^{37\times 64}$.
- Trace-norm regularization.