Iteratively reweighted ℓ_1 algorithms with extrapolation

Ting Kei Pong
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hong Kong

Workshop on Variational Analysis and Stochastic Optimization
December 2017

(Joint work with Peiran Yu)

Motivations

Finding sparse solutions of linear systems:

$$\begin{aligned} \min_x \quad & \|x\|_0 \\ \text{s.t.} \quad & Ax = b. \end{aligned}$$

Motivations

Finding sparse solutions of linear systems:

$$\begin{aligned} \min_x \quad & \|x\|_0 \\ \text{s.t.} \quad & Ax = b. \end{aligned}$$

Approximation models: $\min_{Ax=b} \sum_{i=1}^n \phi(|x_i|)$, or

$$\min_x \quad \frac{1}{2}\|Ax - b\|^2 + \sum_{i=1}^n \phi(|x_i|),$$

where ϕ is a sparsity inducing function, such as:

- ℓ_1 penalty: $\phi(|t|) = \lambda|t|$ (Tibshirani '96);
- log penalty: $\phi(|t|) = \lambda \log(1 + |t|/\epsilon)$ (Nikolova et al. '08);
- smoothly clipped absolute deviation ($\alpha > 2$): $\phi(|t|) = \int_0^{|t|} \min\left\{1, \frac{(\alpha - s/\lambda)_+}{\alpha - 1}\right\} ds$ (Fan, Li '01).

Iterately reweighted ℓ_1 algorithms

- IRL$_1$ algorithms solve $\min_{Ax=b} \sum_{i=1}^n \phi(|x_i|)$ by iteratively solving

$$x^{k+1} \in \operatorname*{Arg\,min}_{Ax=b} \sum_{i=1}^{n} \phi'_{+}(|x_i^k|)|x_i|.$$

Iteratively reweighted ℓ_1 algorithms

- IRL$_1$ algorithms solve $\min_{Ax=b} \sum_{i=1}^n \phi(|x_i|)$ by iteratively solving

$$x^{k+1} \in \operatorname*{Arg\,min}_{Ax=b} \sum_{i=1}^n \phi'_+(|x_i^k|)|x_i|.$$

- Convergence guaranteed under suitable assumptions on ϕ (Candès, Wakin, Boyd '08, Chartrand, Yin '08, etc.).
- A recent variant proposed in Lu '14 minimizes $f(x) + \sum_{i=1}^n \phi(|x_i|)$ for a Lipschitz differentiable f:

$$x^{k+1} = \operatorname*{arg\,min}_x \left\{ \left\langle \nabla f(x^k), x \right\rangle + \frac{L_k}{2}\|x - x^k\|^2 + \sum_{i=1}^n \phi'_+(|x_i^k|)|x_i| \right\};$$

Iteratively reweighted ℓ_1 algorithms

- IRL$_1$ algorithms solve $\min_{Ax=b} \sum_{i=1}^n \phi(|x_i|)$ by iteratively solving

$$x^{k+1} \in \operatorname*{Arg\,min}_{Ax=b} \sum_{i=1}^n \phi'_+(|x_i^k|)|x_i|.$$

- Convergence guaranteed under suitable assumptions on ϕ (Candès, Wakin, Boyd '08, Chartrand, Yin '08, etc.).
- A recent variant proposed in Lu '14 minimizes $f(x) + \sum_{i=1}^n \phi(|x_i|)$ for a Lipschitz differentiable f: $\{L_k\}$ is determined by line-search, for empirical acceleration

$$x^{k+1} = \arg\min_x \left\{ \left\langle \nabla f(x^k), x \right\rangle + \frac{L_k}{2}\|x - x^k\|^2 + \sum_{i=1}^n \phi'_+(|x_i^k|)|x_i| \right\};$$

Iteratively reweighted ℓ_1 algorithms

- IRL$_1$ algorithms solve $\min\limits_{Ax=b} \sum_{i=1}^n \phi(|x_i|)$ by iteratively solving

$$x^{k+1} \in \operatorname*{Arg\,min}_{Ax=b} \sum_{i=1}^n \phi'_+(|x_i^k|)|x_i|.$$

- Convergence guaranteed under suitable assumptions on ϕ (Candès, Wakin, Boyd '08, Chartrand, Yin '08, etc.).
- A recent variant proposed in Lu '14 minimizes $f(x) + \sum_{i=1}^n \phi(|x_i|)$ for a Lipschitz differentiable f: $\{L_k\}$ is determined by line-search, for empirical acceleration

$$x^{k+1} = \operatorname*{arg\,min}_x \left\{ \left\langle \nabla f(x^k), x \right\rangle + \frac{L_k}{2}\|x - x^k\|^2 + \sum_{i=1}^n \phi'_+(|x_i^k|)|x_i| \right\};$$

- Alternative strategies for empirical acceleration?

Extrapolation

- A classical technique for acceleration. Dating back to Polyak's heavy ball method (Polyak '64).

Extrapolation

- A classical technique for acceleration. Dating back to Polyak's heavy ball method (Polyak '64).
- Nesterov's extrapolation techniques (Nesterov '83) led to "optimal" 1st-order methods for minimizing $f + P$, where f is convex Lipschitz differentiable and P is proper closed convex. Some representative algorithms are:
  - ⋆ FISTA (Beck, Teboulle '09, Nesterov '13);
  - ⋆ the method in Auslender, Teboulle '06;
  - ⋆ the method in Lan, Lu, Monteiro '11.

Extrapolation

- A classical technique for acceleration. Dating back to Polyak's heavy ball method (Polyak '64).
- Nesterov's extrapolation techniques (Nesterov '83) led to "optimal" 1st-order methods for minimizing $f + P$, where f is convex Lipschitz differentiable and P is proper closed convex. Some representative algorithms are:
  - ⋆ FISTA (Beck, Teboulle '09, Nesterov '13);
  - ⋆ the method in Auslender, Teboulle '06;
  - ⋆ the method in Lan, Lu, Monteiro '11.
- Extrapolation techniques have been applied to 1st-order methods for some nonconvex problems (Ghadimi, Lan '16, Drusvyatskiy, Paquette '16, Wen, Chen, Pong '17), with good empirical performance.

Aim

Our target:

- Explore how widely used extrapolation techniques can be incorporated to (empirically) accelerate IRL_1 algorithms.
- Analyze convergence of the resulting algorithms: find explicit conditions on the extrapolation parameters for convergence.

Problem setting

$$\min_{x} \; F(x) := f(x) + \delta_C(x) + \Phi(|x|).$$

We assume

- f is convex differentiable, ∇f is Lipschitz with modulus $L > 0$;
- C is a nonempty closed convex set;
- $\Phi(y) = \sum_{i=1}^{n} \phi(y_i)$, where $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies the following properties:
  - ⋆ ϕ is continuous and concave in $\mathbb{R}_+$, and is differentiable on $(0, \infty)$;
  - ⋆ $\phi(0) = 0$ and $\lim_{t \downarrow 0} \phi'(t)$ exists (which imply $\phi'_+(0)$ exists).

Problem setting

$$\min_x \; F(x) := f(x) + \delta_C(x) + \Phi(|x|).$$

We assume

- f is convex differentiable, ∇f is Lipschitz with modulus $L > 0$;
- C is a nonempty closed convex set;
- $\Phi(y) = \sum_{i=1}^n \phi(y_i)$, where $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies the following properties:
  - ⋆ ϕ is continuous and concave in $\mathbb{R}_+$, and is differentiable on $(0, \infty)$;
  - ⋆ $\phi(0) = 0$ and $\lim_{t \downarrow 0} \phi'(t)$ exists (which imply $\phi'_+(0)$ exists).
- F is level-bounded.

$\text{IRL}_1 e_1$: Algorithmic framework

Algorithm 1: $\text{IRL}_1 e_1$

Step 0. Input $x^0 = x^{-1} \in C$, $\{\beta_k\} \subset [0, 1)$. Set $k = 0$.

Step 1. Set

$$y^k = x^k + \beta_k(x^k - x^{k-1});$$

$$x^{k+1} = \arg\min_{y \in C} \left\{ \left\langle \nabla f(y^k), y \right\rangle + \frac{L}{2}\|y - y^k\|^2 + \sum_{i=1}^{n} \phi'_+(|x_i^k|)|y_i| \right\}.$$

Step 2. If a termination criterion isn't met, set $k = k + 1$; go to Step 1.

$\mathrm{IRL}_1 e_1$: Algorithmic framework

Algorithm 1: $\mathrm{IRL}_1 e_1$

Step 0. Input $x^0 = x^{-1} \in C$, $\{\beta_k\} \subset [0,1)$. Set $k = 0$.
Step 1. Set

$$y^k = x^k + \beta_k(x^k - x^{k-1});$$
$$x^{k+1} = \arg\min_{y\in C} \left\{ \left\langle \nabla f(y^k), y \right\rangle + \frac{L}{2}\|y - y^k\|^2 + \sum_{i=1}^n \phi'_+(|x_i^k|)|y_i| \right\}.$$

Step 2. If a termination criterion isn't met, set $k = k+1$; go to Step 1.

- Reduces to FISTA (Beck, Teboulle '09, Nesterov '13) when $\Phi \equiv 0$ and $\{\beta_k\}$ is chosen properly.

$\text{IRL}_1 e_1$: Convergence analysis

Makes extensive use of

$$H_1(x, y) := F(x) + \frac{L}{2}\|x - y\|^2.$$

IRL$_1 e_1$: Convergence analysis

Makes extensive use of

$$H_1(x,y) := F(x) + \frac{L}{2}\|x-y\|^2.$$

Theorem 1. (Yu, P. '17)
Suppose that $\sup \beta_k < 1$. Let $\{x^k\}$ be generated by IRL$_1 e_1$. Then

- $\{H_1(x^k, x^{k-1})\}$ is nonincreasing.
- The sequence $\{x^k\}$ is bounded and any accumulation point is a stationary point of F.
- If H_1 is a Kurdyka-Łojasiewicz function and ϕ'_+ is globally Lipschitz in $\mathbb{R}_+$, then $\{x^k\}$ is convergent.

Recall that x^* is a stationary point of F if $0 \in \partial F(x^*)$.

$\mathrm{IRL}_1 e_2$: Algorithmic framework

Algorithm 2: $\mathrm{IRL}_1 e_2$

Step 0. Input $x^0, z^0 \in C$, $\{\theta_k\} \subset (0, 1]$. Set $k = 0$.
Step 1. Set

$$
\begin{aligned}
y^k &= (1 - \theta_k)x^k + \theta_k z^k; \\
z^{k+1} &= \arg\min_{x \in C} \left\{ \left\langle \nabla f(y^k), x \right\rangle + \frac{L\theta_k}{2}\|x - z^k\|^2 + \sum_{i=1}^{n} \phi'_+(|x_i^k|)|x_i| \right\}; \\
x^{k+1} &= (1 - \theta_k)x^k + \theta_k z^{k+1}.
\end{aligned}
$$

Step 2. If a termination criterion isn't met, set $k = k + 1$; go to Step 1.

$\text{IRL}_1 e_2$: Algorithmic framework

Algorithm 2: $\text{IRL}_1 e_2$

Step 0. Input $x^0, z^0 \in C$, $\{\theta_k\} \subset (0, 1]$. Set $k = 0$.

Step 1. Set

$$
\begin{aligned}
y^k &= (1 - \theta_k)x^k + \theta_k z^k; \\
z^{k+1} &= \arg\min_{x \in C} \left\{ \left\langle \nabla f(y^k), x \right\rangle + \frac{L\theta_k}{2}\|x - z^k\|^2 + \sum_{i=1}^{n} \phi'_+(|x_i^k|)|x_i| \right\}; \\
x^{k+1} &= (1 - \theta_k)x^k + \theta_k z^{k+1}.
\end{aligned}
$$

Step 2. If a termination criterion isn't met, set $k = k + 1$; go to Step 1.

- Reduces to the fast 1st-order method in Auslender, Teboulle '06 when $\Phi \equiv 0$ and $\{\theta_k\}$ is chosen properly.

$\mathrm{IRL}_1 e_2$: Convergence analysis

Also makes extensive use of

$$H_1(x, y) = F(x) + \frac{L}{2}\|x - y\|^2.$$

$\text{IRL}_1 e_2$: Convergence analysis

Also makes extensive use of

$$H_1(x, y) = F(x) + \frac{L}{2}\|x - y\|^2.$$

Theorem 2. (Yu, P. '17)
Suppose that $\{\theta_k\}$ is chosen so that

$$\sup_k \{\theta_k^2(1 - \theta_{k-1})^2 - \theta_{k-1}^2\} < 0,$$

and let $\{x^k\}$ be generated by $\text{IRL}_1 e_2$. Then

- $\{H_1(x^k, x^{k-1})\}$ is nonincreasing.
- The sequence $\{x^k\}$ is bounded and any accumulation point is a stationary point of F.

IRL$_1e_3$: Algorithmic framework

Algorithm 3: IRL$_1e_3$

Step 0. Input $x^0, z^0 \in C$, $\{\theta_k\} \subset (0,1]$. Set $k = 0$.
Step 1. Set

$$y^k = (1-\theta_k)x^k + \theta_k z^k;$$

$$z^{k+1} = \arg\min_{x\in C}\left\{\left\langle \nabla f(y^k), x\right\rangle + \frac{L\theta_k}{2}\|x - z^k\|^2 + \sum_{i=1}^n \phi'(|x_i^k|)|x_i|\right\};$$

$$x^{k+1} = \arg\min_{y\in C}\left\{\left\langle \nabla f(y^k), y\right\rangle + \frac{L}{2}\|y - y^k\|^2 + \sum_{i=1}^n \phi'(|x_i^k|)|y_i|\right\}.$$

Step 2. If a termination criterion isn't met, set $k = k+1$; go to Step 1.

$\text{IRL}_1 e_3$: Algorithmic framework

Algorithm 3: $\text{IRL}_1 e_3$

Step 0. Input $x^0, z^0 \in C$, $\{\theta_k\} \subset (0, 1]$. Set $k = 0$.
Step 1. Set

$$y^k = (1 - \theta_k)x^k + \theta_k z^k;$$

$$z^{k+1} = \arg\min_{x \in C} \left\{ \left\langle \nabla f(y^k), x \right\rangle + \frac{L\theta_k}{2}\|x - z^k\|^2 + \sum_{i=1}^n \phi'(|x_i^k|)|x_i| \right\};$$

$$x^{k+1} = \arg\min_{y \in C} \left\{ \left\langle \nabla f(y^k), y \right\rangle + \frac{L}{2}\|y - y^k\|^2 + \sum_{i=1}^n \phi'(|x_i^k|)|y_i| \right\}.$$

Step 2. If a termination criterion isn't met, set $k = k + 1$; go to Step 1.

- Reduces to the fast 1st-order method in Lan, Lu, Monteiro '11 when $\Phi \equiv 0$ and $\{\theta_k\}$ is suitably chosen.

$\text{IRL}_1 e_3$: Convergence analysis

Makes extensive use of

$$H_3(x, y, w) = F(x) + \frac{L}{2}\|w - x\|^2 + \frac{L}{2}\|w - y\|^2.$$

IRL$_1 e_3$: Convergence analysis

Makes extensive use of

$$H_3(x, y, w) = F(x) + \frac{L}{2}\|w - x\|^2 + \frac{L}{2}\|w - y\|^2.$$

Theorem 3. (Yu, P. '17)
Suppose that $\{\theta_k\}$ is chosen so that for some $\gamma \in (0, 1)$,

$$\sup_k \max\left\{\frac{\theta_k^2(1 - \theta_{k-1})^2}{\gamma} - \theta_{k-1}^2, \frac{\theta_k^2}{1 - \gamma} - 1\right\} < 0.$$

Let $\{x^k\}$ be generated by IRL$_1 e_3$. Then

- The sequence $\{x^k\}$ is bounded and any accumulation point is a stationary point of F.
- If H_3 is a Kurdyka-Łojasiewicz function and ϕ'_+ is globally Lipschitz in $\mathbb{R}_+$, then $\{x^k\}$ is convergent.

Numerical simulations

- Solve
$$\min_x \ \frac{1}{2}\|Ax - b\|^2 + \lambda \sum_{i=1}^{n} \log(1 + |x_i|/\epsilon).$$
- Consider random instances: generate an $m \times n$ matrix A, an r-sparse vector $\tilde{x}$, a noise vector $\hat{n}$ and set $b = A\tilde{x} + \hat{n}$.
- Compare $\mathrm{IRL}_1 e_1$ and $\mathrm{IRL}_1 e_3$ with NPG (Wright et al. '09).
- Initialize at $x^0 = 0$.
- Terminate when $\mathrm{dist}(0, \partial F(x^k)) \leqslant 10^{-4} \max\{1, \|x^k\|\}$.

Numerical simulations

- Solve
$$\min_x \ \frac{1}{2}\|Ax-b\|^2 + \lambda \sum_{i=1}^{n} \log(1+|x_i|/\epsilon).$$
- Consider random instances: generate an $m \times n$ matrix A, an r-sparse vector $\tilde{x}$, a noise vector $\hat{n}$ and set $b = A\tilde{x} + \hat{n}$.
- Compare $\mathrm{IRL}_1 e_1$ and $\mathrm{IRL}_1 e_3$ with NPG (Wright et al. '09).
- Initialize at $x^0 = 0$.
- Terminate when $\mathrm{dist}(0, \partial F(x^k)) \leqslant 10^{-4} \max\{1, \|x^k\|\}$.
- For $\mathrm{IRL}_1 e_1$: $\{\beta_k\}$ chosen as in FISTA with fixed and adaptive restart.
  For $\mathrm{IRL}_1 e_3$: $\{\theta_k\}$ chosen similarly as in Lan, Lu, Monteiro '11 initially, and then set to be constant after 50 iterations.

Numerical simulations

Choose $\lambda = 5 \times 10^{-4}$ and $\epsilon = 0.5$. The t_0 is the time for computing $\lambda_{\max}(A^T A)$. Averaged over 20 instances.[1]

Problem Size				time			fval		
m	n	r	t_0	NPG	IRL$_1$ e_1	IRL$_1$ e_3	NPG	IRL$_1$ e_1	IRL$_1$ e_3
720	2560	80	0.1	1.7	0.7	0.6	3.7918e-2	3.7897e-2	3.7896e-2
1440	5120	160	0.7	7.0	3.3	2.6	7.5904e-2	7.5859e-2	7.5858e-2
2160	7680	240	0.6	15.0	7.2	5.7	1.1443e-1	1.1436e-1	1.1436e-1
2880	10240	320	1.3	25.9	12.6	9.9	1.5224e-1	1.5215e-1	1.5215e-1
3600	12800	400	2.4	39.4	19.9	15.5	1.8805e-1	1.8794e-1	1.8794e-1
4320	15360	480	3.8	56.7	28.1	21.9	2.2774e-1	2.2761e-1	2.2761e-1
5040	17920	560	6.2	75.9	38.4	29.7	2.6491e-1	2.6474e-1	2.6475e-1
5760	20480	640	8.0	99.8	50.4	39.1	3.0627e-1	3.0609e-1	3.0609e-1
6480	23040	720	11.1	124.7	62.8	48.8	3.4231e-1	3.4212e-1	3.4212e-1
7200	25600	800	14.7	157.4	79.2	61.4	3.8133e-1	3.8111e-1	3.8111e-1

[1] Matlab 2015b, 64-bit PC with an Intel(R) Core(TM) i7-4790 CPU (3.60GHz) and 32GB RAM

Conclusion

- Commonly used extrapolation techniques for convex composite optimization can be suitably incorporated into IRL_1 algorithms.
- Explicit conditions on the extrapolation parameters are given to guarantee convergence of the resulting algorithms.

Reference:

- P. Yu and T. K. Pong.
  Iteratively reweighted ℓ_1 algorithms with extrapolation.
  Available at *https://arxiv.org/abs/1710.07886*.

Thanks for coming! :)