Iteratively Re-weighted Least Squares for Sums of Convex Functions

James Burke
University of Washington

Jiashan Wang
LinkedIn

Frank Curtis
Lehigh University

Hao Wang
Shanghai Tech University

Daiwei He
University of Washington

Outline

1. Classical Iterative Re-Weighting
2. Exact Penalization
3. The Iterative Re-Weighting Algorithm (IRWA)
4. Convergence
5. A Numerical Experiment: ℓ_1-SVM
6. Least-Squares and Sums of Convex Functions
7. General Least-Squares Based Algorithms
8. General Methods Applied to $J(x, \varepsilon) = \sum_{i=1}^{\ell} \sqrt{\mathrm{dist}^2(A_i x + b_i \mid C_i) + \epsilon_i^2}$
9. Comparison with IRWA

Classical Iterative Re-Weighting: A mainstay of statistical computing

- ℓ^1-Regression

$$\min_{x\in\mathbb{R}^n} \|Ax+b\|_1 := \sum_{i=1}^{m} |A_i x + b_i| \quad ,$$

where A_i is the ith row of A.

Classical Iterative Re-Weighting: A mainstay of statistical computing

- ℓ^1-Regression

$$\min_{x\in\mathbb{R}^n} \|Ax+b\|_1 := \sum_{i=1}^{m} |A_i x + b_i| \quad ,$$

where A_i is the ith row of A.

- Iterative Least Squares Approach

Having x^k approximate $\|Ax+b\|_1$ by

$$\|Ax+b\|_1 = \sum_{i=1}^{m} |A_i x + b_i| = \sum_{i=1}^{m} \frac{|A_i x + b_i|^2}{|A_i x + b_i|} \approx \sum_{i=1}^{m} \frac{|A_i x + b_i|^2}{|A_i x^k + b_i|}.$$

Solve the linear least squares problem in the approximation for x^{k+1} and iterate.

Classical Iterative Re-Weighting: A mainstay of statistical computing

- ℓ^1-Regression

$$\min_{x\in\mathbb{R}^n} \|Ax+b\|_1 := \sum_{i=1}^{m} |A_i x + b_i| \quad ,$$

where A_i is the ith row of A.

- Iterative Least Squares Approach

Having x^k approximate $\|Ax+b\|_1$ by

$$\|Ax+b\|_1 = \sum_{i=1}^{m} |A_i x + b_i| = \sum_{i=1}^{m} \frac{|A_i x + b_i|^2}{|A_i x + b_i|} \approx \sum_{i=1}^{m} \frac{|A_i x + b_i|^2}{|A_i x^k + b_i|}.$$

Solve the linear least squares problem in the approximation for x^{k+1} and iterate.

- Modified ϵ-Approximate Weighted Least Squares

$$\|Ax+b\|_1 \approx \sum_{i=1}^{m} w(x^k, \epsilon_i^k)|A_i x + b_i|^2$$

where $w(x^k, \epsilon_i^k) := 1/\sqrt{|A_i x^k + b_i|^2 + (\epsilon_i^k)^2}$.

Classical Iterative Re-Weighting

Lawson 1961

Rice and Usow 1968

Karlovitz 1970

Kahng 1972

Fletcher, Grant and Hebden 1972

Schlossmacker 1973

Beaton and Tukey 1974

Wolke and Schwetlick 1988

O'Leary 1990

Burrus and Burreto 1992

Vargas and Burrus 1999

The Exact Penalty Subproblem

$$\min_{x\in X}\; J_0(x) := g^T x + \frac{1}{2}x^T Hx + \sum_{i=1}^{l} \mathrm{dist}_2\left(A_i x + b_i \mid C_i\right),$$

$g \in \mathbb{R}^n$, $H \in \mathcal{S}^n_+$, $A \in \mathbb{R}^{m_i\times n}$, $b \in \mathbb{R}^{m_i}$, and $C_i \subset \mathbb{R}^{m_i}$ are non-empty closed convex, and

$$\mathrm{dist}_2(y_i|C_i) := \inf_{z_i\in C_i} \|y_i - z_i\|_2 = \|y_i - P_{C_i}(y_i)\|_2$$

P_{C_i} is the projection onto C_i.

Examples:

- Equality: $C_i := \{0\}$
- Inequality: $C_i := (-\infty, 0]$
- Inequality: $C_i = [l_i, u_i]$
- ℓ_l-ball: $\|y_i\|_l \le \tau_i$, $l = 1, 2, \infty$
- Trust region: $X = \{x \mid \|x\|_l \le \tau\}, l = 1, 2, \infty$

NLP and Exact Penalties

- Nonlinear Programming (NLP):

$$\begin{gathered}\text{minimize } f(x)\\ \text{subject to } F(x) \in C \text{ and } x \in X\end{gathered}$$

- $f : \mathbb{R}^n \to \mathbb{R}$ and $F : \mathbb{R}^n \to \mathbb{R}^m$ smooth
- $X \subset \mathbb{R}^n$ and $C \subset \mathbb{R}^m$ non-empty, closed, and convex
- $C := C_1 \times C_2 \times \cdots \times C_l$, $F(x) := (F_1(x)\ F_2(x)\ \ldots\ F_l(x))$

$F_i(x) \in C_i, \quad F_i : \mathbb{R}^n \to \mathbb{R}^{m_i}, \quad C_i \in \mathbb{R}^{m_i}$

- Exact penalty formulation:

Given $\alpha > 0$

$$\min_{x \in X} f(x) + \alpha \sum_{i=1}^{l} \text{dist}_2(F_i(x)|C_i)$$

Local direction finding approximation

$$\min_{x \in X} \ J_0(x) := g^T x + \frac{1}{2} x^T H x + \sum_{i=1}^{l} \mathrm{dist}_2\left(A_i x + b_i \mid C_i\right),$$

ϵ-Approximate Re-Weighted Least Squares

Let $P_{C_i}(y) \in C_i$ be the projection of y onto C_i, i.e.

$$\|y - P_{C_i}(y)\|_2 = \mathrm{dist}_2\left(y \mid C_i\right).$$

At (x^k, ϵ^k) approximate $\mathrm{dist}_2\left(A_i x + b_i \mid C_i\right) = \|A_i x + b_i - P_{C_i}(A_i x + b_i)\|_2$ by

$$\mathrm{dist}_2\left(A_i x + b_i \mid C_i\right) \approx \frac{\mathrm{dist}_2\left(A_i x + b_i \mid C_i\right)^2}{\sqrt{\mathrm{dist}_2\left(A_i x^k + b_i \mid C_i\right)^2 + (\epsilon_i^k)^2}}$$

ϵ-Approximate Re-Weighted Least Squares

Let $P_{C_i}(y) \in C_i$ be the projection of y onto C_i, i.e.

$$\|y - P_{C_i}(y)\|_2 = \mathrm{dist}_2\left(y \mid C_i\right).$$

At (x^k, ϵ^k) approximate $\mathrm{dist}_2\left(A_i x + b_i \mid C_i\right) = \|A_i x + b_i - P_{C_i}(A_i x + b_i)\|_2$ by

$$\mathrm{dist}_2\left(A_i x + b_i \mid C_i\right) \approx \frac{\mathrm{dist}_2\left(A_i x + b_i \mid C_i\right)^2}{\sqrt{\mathrm{dist}_2\left(A_i x^k + b_i \mid C_i\right)^2 + (\epsilon_i^k)^2}}$$

$$= \frac{\|A_i x + b_i - P_{C_i}(A_i x + b_i)\|_2^2}{\sqrt{\mathrm{dist}_2\left(A_i x^k + b_i \mid C_i\right)^2 + (\epsilon_i^k)^2}},$$

ϵ-Approximate Re-Weighted Least Squares

Let $P_{C_i}(y) \in C_i$ be the projection of y onto C_i, i.e.

$$\|y - P_{C_i}(y)\|_2 = \operatorname{dist}_2(y \mid C_i).$$

At (x^k, ϵ^k) approximate $\operatorname{dist}_2(A_i x + b_i \mid C_i) = \|A_i x + b_i - P_{C_i}(A_i x + b_i)\|_2$ by

$$\operatorname{dist}_2(A_i x + b_i \mid C_i) \approx \frac{\operatorname{dist}_2(A_i x + b_i \mid C_i)^2}{\sqrt{\operatorname{dist}_2(A_i x^k + b_i \mid C_i)^2 + (\epsilon_i^k)^2}}$$

$$= \frac{\|A_i x + b_i - P_{C_i}(A_i x + b_i)\|_2^2}{\sqrt{\operatorname{dist}_2(A_i x^k + b_i \mid C_i)^2 + (\epsilon_i^k)^2}},$$

$$\approx \frac{\left\|A_i x + b_i - P_{C_i}(A_i x^k + b_i)\right\|_2^2}{\sqrt{\operatorname{dist}_2(A_i x^k + b_i \mid C_i)^2 + (\epsilon_i^k)^2}},$$

The full approximation

$$J_0(x) = g^T x + \frac{1}{2} x^T H x + \sum_{i=1}^{l} \mathrm{dist}_2\left(A_i x + b_i \mid C_i\right)$$

$$\approx g^T x + \frac{1}{2} x^T H x + \frac{1}{2} \sum_{i=1}^{l} w_i(x^k, \epsilon^k) \left\| A_i x + b_i - P_{C_i}(A_i x^k + b_i) \right\|_2^2 \ ,$$

where

$$w_i(x^k, \epsilon^k) = 1 / \sqrt{\mathrm{dist}_2^2(A_i x^k + b_i | C_i) + (\epsilon_i^k)^2}.$$

The full approximation

$$J_0(x) = g^T x + \frac{1}{2} x^T H x + \sum_{i=1}^{l} \mathrm{dist}_2\left(A_i x + b_i \mid C_i\right)$$

$$\approx g^T x + \frac{1}{2} x^T H x + \frac{1}{2} \sum_{i=1}^{l} w_i(x^k, \epsilon^k) \left\| A_i x + b_i - P_{C_i}(A_i x^k + b_i) \right\|_2^2 \ ,$$

where

$$w_i(x^k, \epsilon^k) = 1 / \sqrt{\mathrm{dist}_2^2(A_i x^k + b_i | C_i) + (\epsilon_i^k)^2}.$$

For simplicity, we drop the term $g^T x + \frac{1}{2} x^T H x$ from future consideration. This can be done with (almost) no loss in generality.

The Iterative Re-Weighting Algorithm (IRWA)

IRWA:

$$\boxed{x^{k+1} \leftarrow \arg\min_{x \in X} \;\; \tfrac{1}{2}\sum_{i=1}^{l} w_i(x^k, \epsilon^k) \left\| A_i x + b_i - P_{C_i}(A_i x^k + b_i) \right\|_2^2}$$

with $\epsilon^k \to 0$.

The Iterative Re-Weighting Algorithm (IRWA)

IRWA:

$$\boxed{x^{k+1} \leftarrow \arg\min_{x \in X} \ \frac{1}{2} \sum_{i=1}^{l} w_i(x^k, \epsilon^k) \left\| A_i x + b_i - P_{C_i}(A_i x^k + b_i) \right\|_2^2}$$

with $\epsilon^k \to 0$.

1. Initialize $x^0, \ \epsilon^0, \ M, \nu > 0, \ \eta, \sigma, \sigma' \in (0,1)$.
2. Solve the re-weighted subproblem for x^{k+1}.
3. Set $q_i^k := A_i(x^{k+1} - x^k)$ If

$$\left\| q_i^k \right\|_2 \le M \left[\operatorname{dist}_2^2 \left(A_i x^k + b_i \mid C_i \right) + (\epsilon_i^k)^2 \right]^{\frac{1}{2}+\nu}, \quad \forall\, i = 1, \ldots, l,$$

choose $\epsilon^{k+1} \in (0, \ \eta\epsilon^k]$; otherwise, set $\epsilon^{k+1} := \epsilon^k$.

4. If $\left\| x^{k+1} - x^k \right\|_2 \le \sigma$ and $\left\| \epsilon^k \right\|_2 \le \sigma'$, stop; else $k := k+1$ and go to Step 2.

Convergence

Let $(x^0, \epsilon^0) \in X \times \mathbb{R}_{++}$. Suppose that the sequence $\{(x^k, \epsilon^k)\}_{k=0}^{\infty}$ is generated by IRWA with $\sigma = \sigma' = 0$. Let

$$S := \left\{ k \;\middle|\; \epsilon^{k+1} \leq \eta \epsilon^k \right\} \; .$$

If either $\ker(A) = \{0\}$ or $X = \mathbb{R}^n$, then any cluster point $\bar{x}$ of the subsequence $\{x^k\}_{k \in S}$ satisfies

$$0 \in \partial J_0(\bar{x}) + N\left(\bar{x} \,|\, X\right).$$

If it is assumed that $\ker(A) = \{0\}$, then $x^k \to x^*$ the unique global solution, and in at most $O(1/\varepsilon^2)$ iterations, x^k is an ε-optimal solution, i.e.

$$J_0(x^k) \leq \inf_{x \in X} J_0(x) + \varepsilon.$$

Support Vector Machine Experiment

Consider the exact penalty form of the l_1-SVM problem:

$$\min_{\beta \in \mathbb{R}^n} \sum_{i=1}^{m} \left(1 - y_i \left(\sum_{j=1}^{n} x_{ij} \beta_j \right) \right)_+ + \lambda \|\beta\|_1 ,$$

where $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^{m}$ are the training data points with $\boldsymbol{x}_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$ for each $i = 1, \ldots, m$, and λ is the penalty parameter.

Support Vector Machine Experiment

Consider the exact penalty form of the l_1-SVM problem:

$$\min_{\beta \in \mathbb{R}^n} \sum_{i=1}^{m} \left(1 - y_i \left(\sum_{j=1}^{n} x_{ij} \beta_j \right) \right)_+ + \lambda \|\beta\|_1 ,$$

where $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^{m}$ are the training data points with $\boldsymbol{x}_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$ for each $i = 1, \ldots, m$, and λ is the penalty parameter.

For purposes of numerical comparison, we randomly generate 40 problems where X ranges from a 200×240 matrix to a 1000×1500 matrix where the sparsity of the "true" solution is always 20% of the number of columns. We compare the performance of IRWA with an ADMM implementation whose parameters are pre-optimized for performance on this data set. The least-squares subproblems for both methods are solved using the same CG solver. The effort is measured in terms of the number of CG solves.

ADAL - IRWA Numerical Comparison: Number of CGs

With Nesterov acceleration.

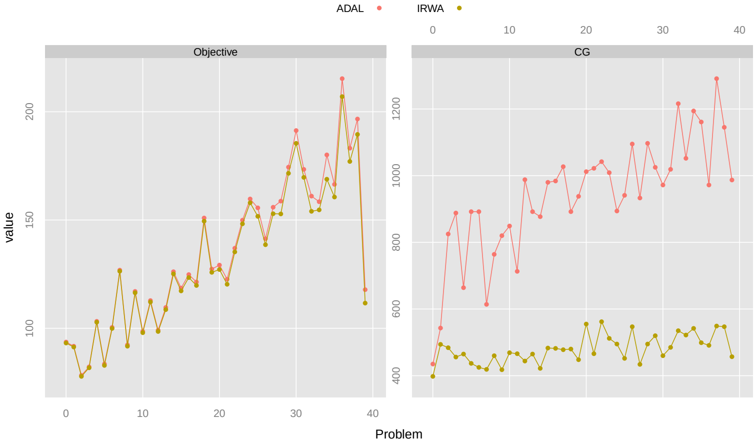

Figure : In all 40 problems, IRWA obtains smaller objective function values with fewer CG steps.

ADAL - IRWA Comparison: Number of False Positives

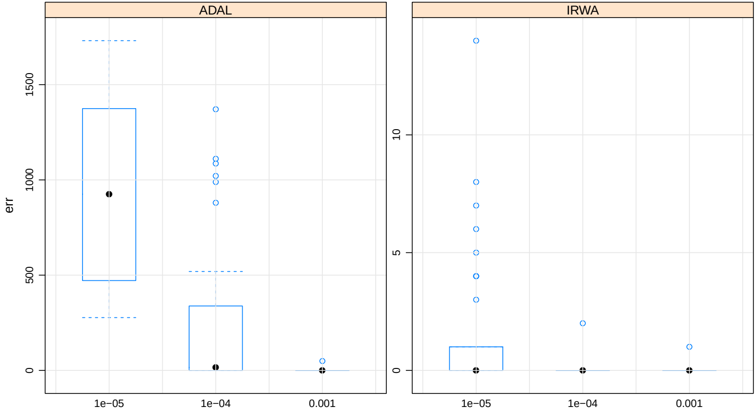

Figure : For both thresholds 10^{-4} and 10^{-5}, IRWA yields fewer false positives in terms of the numbers of "zero" values computed. The numbers of false positives is similar for the threshold 10^{-3}. At the threshold 10^{-5}, the difference in recovery is dramatic with IRWA always having fewer than 14 false positives while ADAL has a median of about 1000 false positives.

Least-Squares and Sums of Convex Functions

$$\min_x \phi(x) := f(Ax + b) = \sum_{i=1}^{\ell} f_i(A_i x + b_i)$$

Goals

1. Minimize ϕ based on properties of the individual f_i only with minimal linkage between the f_i's (splitting). For example, apply Prox to the individual f_i's, but not to the function ϕ as a whole.
2. The A_i's only enter through weighted least-squares:

$$\min_x \frac{1}{2}\sum_{i=1}^{\ell} w_i(x,p)\,\|A_i x + b_i - s_i(p)\|_2^2\,,$$

where $p \in \mathbb{E}_0$ is a parameter vector and $s : \mathbb{E}_0 \to \mathbb{W}$.

Least-Squares and Sums of Convex Functions

$$\min_x \phi(x) := f(Ax+b) = \sum_{i=1}^{\ell} f_i(A_i x + b_i)$$

Goals

1. Minimize ϕ based on properties of the individual f_i only with minimal linkage between the f_i's (splitting). For example, apply Prox to the individual f_i's, but not to the function ϕ as a whole.
2. The A_i's only enter through weighted least-squares:

$$\min_x \frac{1}{2}\sum_{i=1}^{\ell} w_i(x,p)\,\|A_i x + b_i - s_i(p)\|_2^2\,,$$

where $p \in \mathbb{E}_0$ is a parameter vector and $s : \mathbb{E}_0 \to \mathbb{W}$.

The procedures we consider are iterative with x^c and $y^c := Ax^c + b$ representing the current best approximate solutions, and $N(c)$ is the number of iterations to obtain x^c.

Separate the Roles of the f_i's and the A_i's

$$\min_x \phi(x) := f(Ax+b) = \sum_{i=1}^{\ell} f_i(A_i x + b_i)$$

$$\begin{aligned} &\min_x \sum_{i=1}^{\ell} f_i(y_i) \\ &\text{s.t. } \; y \in b + \mathrm{Ran}\, A \end{aligned}$$

Separate the Roles of the f_i's and the A_i's

$$\min_x \phi(x) := f(Ax+b) = \sum_{i=1}^{\ell} f_i(A_i x + b_i)$$

$$\begin{aligned} &\min_x \sum_{i=1}^{\ell} f_i(y_i) \\ &\text{s.t. } \ y \in b + \mathrm{Ran}\, A \end{aligned}$$

$$\mathrm{Prox}_{\gamma_i, f_i} y := \mathrm{argmin}_z \left[f_i(z) + \frac{1}{2\gamma_i} \|z-y\|_2^2 \right]$$

Least-Squares Based Algorithms

Projected Subgradient:

$$\min_x \frac{1}{2}\sum_{i=1}^{\ell} \|A_i(x - x^c) + t_i^c g_i^c\|_2^2\,, \qquad g_i^c \in \partial f_i(y_i^c),\ t_i^c := \frac{R}{L\sqrt{N(c)}}, \quad O(1/\epsilon^2)\ _{\text{no acc}}$$

Least-Squares Based Algorithms

Projected Subgradient:

$$\min_x \frac{1}{2}\sum_{i=1}^{\ell} \|A_i(x-x^c)+t_i^c g_i^c\|_2^2 , \qquad g_i^c \in \partial f_i(y_i^c),\ t_i^c := \frac{R}{L\sqrt{N(c)}}, \quad O(1/\epsilon^2)\ \text{no acc}$$

Projected Prox-Gradient:

$$\min_x \frac{1}{2}\sum_{i=1}^{\ell} \|A_i(x-x^c)+t_i^c(I-\mathrm{Prox}_{\gamma_i,f_i})(A_i x^c+b_i)\|_2^2 ,\ t_j^c := \frac{\gamma_j^{-1}}{\sum_{i=1}^{\ell}\gamma_i^{-1}},\ O(1/\epsilon)\ \text{acc}$$

Least-Squares Based Algorithms

Projected Subgradient:

$$\min_x \frac{1}{2}\sum_{i=1}^{\ell} \|A_i(x - x^c) + t_i^c g_i^c\|_2^2\,, \qquad g_i^c \in \partial f_i(y_i^c),\ t_i^c := \frac{R}{L\sqrt{N(c)}}, \quad O(1/\epsilon^2)_{\text{ no acc}}$$

Projected Prox-Gradient:

$$\min_x \frac{1}{2}\sum_{i=1}^{\ell} \|A_i(x - x^c) + t_i^c(I - \mathrm{Prox}_{\gamma_i,f_i})(A_i x^c + b_i)\|_2^2\,,\ t_j^c := \frac{\gamma_j^{-1}}{\sum_{i=1}^{\ell}\gamma_i^{-1}},\ O(1/\epsilon)_{\text{ acc}}$$

ADAL:

$$\min_x \frac{1}{2}\sum_{i=1}^{\ell}\frac{1}{\gamma_i}\,\|(A_i(x - x^c) + (I - \mathrm{Prox}_{\gamma_i,f_i})(A_i x^c + b_i + \gamma_i u_i^c)\|_2^2\,,\ O(1/\epsilon)_{\text{ acc}}$$

$$u_i^+ := u_i^c + \frac{1}{\gamma_i}(A_i x^+ + b_i - \mathrm{Prox}_{\gamma_i,f_i}(A_i x^c + b_i + \gamma u_i^c)).$$

Application to $J(x,\varepsilon)=\sum_{i=1}^{\ell}\sqrt{\mathrm{dist}^2(A_ix+b_i \mid C_i)+\epsilon_i^2}$

$$\frac{1}{2}\sum_{i=1}^{\ell}\|A_i(x-x^c)+t_i^c(I-P_{C_i})(A_ix^c+b_i)\|_2^2$$

Projected Subgradient for $J(x,0)$

$$t_i^c := \begin{cases} \frac{\mathrm{dist}\left(x^0 \mid \Sigma\right)}{\ell\,\mathrm{dist}(A_ix^c+b_i \mid C_i)\sqrt{N(c)}} & \text{, if } A_ix^c+b_i\notin C_i, \\ 0 & \text{, otherwise.} \end{cases} \qquad (\Sigma \text{ solution set})$$

Projected gradient for $J(x,\varepsilon)$

$$t_j^c := \frac{\epsilon_{\min}}{\sqrt{\mathrm{dist}^2(A_jx^c+b_j \mid C_j)+\epsilon_j^2}}$$

Projected Prox-Gradient for $J(x,0)$

$$t_j^c := \begin{cases} \frac{\gamma_j^{-1}}{\sum_{i=1}^{\ell}\gamma_i^{-1}} & ,\ \mathrm{dist}_2\left(A_jx^c+b_j \mid C_j\right)\le\gamma_j \\ \frac{\mathrm{dist}_2^{-1}(A_jx^c+b_j|C_j)}{\sum_{i=1}^{\ell}\gamma_i^{-1}} & ,\ \mathrm{dist}_2\left(A_jx^c+b_j \mid C_j\right)>\gamma_j \end{cases}$$

ADAL for $J(x,0)$

Same as projected prox-gradient but $t_j^c := 1$ and we include shifts $\gamma_i u_i^c$.

Comparison with IRWA

General Methods:

$$\frac{1}{2}\sum_{i=1}^{\ell} \left\|A_i(x - x^k) + t_i^c(I - P_{C_i})(A_i x^k + b_i)\right\|_2^2 \qquad O(\frac{1}{\epsilon}) \text{ acc}$$

IRWA:

$$\frac{1}{2}\sum_{i=1}^{l} \frac{1}{\sqrt{\operatorname{dist}_2^2\left(A_i x^k + b_i \mid C_i\right) + \epsilon_i^2}} \left\|A_i(x - x^k) + (I - P_{C_i})(A_i x^k + b_i)\right\|_2^2 \; O(\frac{1}{\epsilon^2}) \text{ no acc}$$

Comparison with IRWA

General Methods:

$$\frac{1}{2}\sum_{i=1}^{\ell} \left\| A_i(x - x^k) + t_i^c(I - P_{C_i})(A_i x^k + b_i) \right\|_2^2 \qquad O(\frac{1}{\epsilon}) \text{ acc}$$

IRWA:

$$\frac{1}{2}\sum_{i=1}^{l} \frac{1}{\sqrt{\mathrm{dist}_2^2\left(A_i x^k + b_i \mid C_i\right) + \epsilon_i^2}} \left\| A_i(x - x^k) + (I - P_{C_i})(A_i x^k + b_i) \right\|_2^2 \; O(\frac{1}{\epsilon^2}) \text{ no acc}$$

New Improved IRWA:

$$\frac{1}{2}\sum_{i=1}^{l} \frac{1}{\epsilon_i} \left\| A_i(x - x^k) + \frac{\epsilon_i}{\sqrt{\mathrm{dist}_2^2\left(A_i x^k + b_i \mid C_i\right) + \epsilon_i^2}}(I - P_{C_i})(A_i x^k + b_i) \right\|_2^2$$

$O(\frac{1}{\epsilon})$!! acc

Comparison of Projected Gradient and New IRWA

Projected gradient for $J(x, \varepsilon)$

$$\frac{1}{2}\sum_{i=1}^{\ell}\left\| A_i(x - x^k) + \frac{\epsilon_{\min}}{\sqrt{\mathrm{dist}^2(A_i x^k + b_i \mid C_i) + \epsilon_i^2}}(I - P_{C_i})(A_i x^k + b_i)\right\|_2^2$$

New Improved IRWA:

$$\frac{1}{2}\sum_{i=1}^{l}\frac{1}{\epsilon_i}\left\| A_i(x - x^k) + \frac{\epsilon_i}{\sqrt{\mathrm{dist}_2^2\left(A_i x^k + b_i \mid C_i\right) + \epsilon_i^2}}(I - P_{C_i})(A_i x^k + b_i)\right\|_2^2$$

Comparison of Projected Gradient and New IRWA

Projected gradient for $J(x, \varepsilon)$

$$\frac{1}{2}\sum_{i=1}^{\ell}\left\| A_i(x - x^k) + \frac{\epsilon_{\min}}{\sqrt{\mathrm{dist}^2(A_i x^k + b_i \mid C_i) + \epsilon_i^2}}(I - P_{C_i})(A_i x^k + b_i)\right\|_2^2$$

New Improved IRWA:

$$\frac{1}{2}\sum_{i=1}^{l}\frac{1}{\epsilon_i}\left\| A_i(x - x^k) + \frac{\epsilon_i}{\sqrt{\mathrm{dist}_2^2\left(A_i x^k + b_i \mid C_i\right) + \epsilon_i^2}}(I - P_{C_i})(A_i x^k + b_i)\right\|_2^2$$

When all ϵ_i have the same value ϵ, they are the same algorithm!

Unaccelerated versions with $\epsilon = 0.01$.

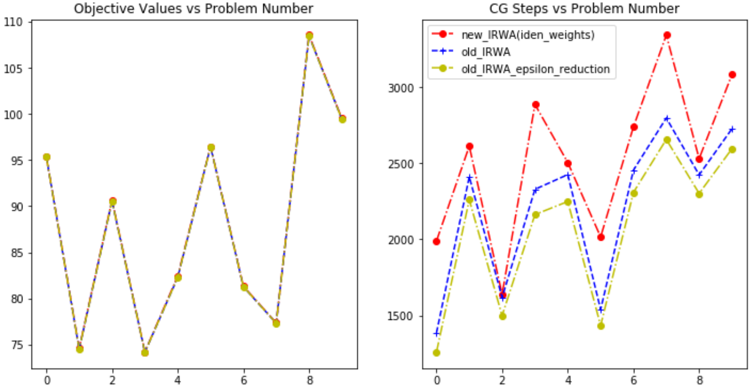

Figure : In all 10 problems, the old IRWA with iterative re-weighting obtains similar objective function values with fewer CG steps.

Accelerated versions with $\epsilon = 0.01$.

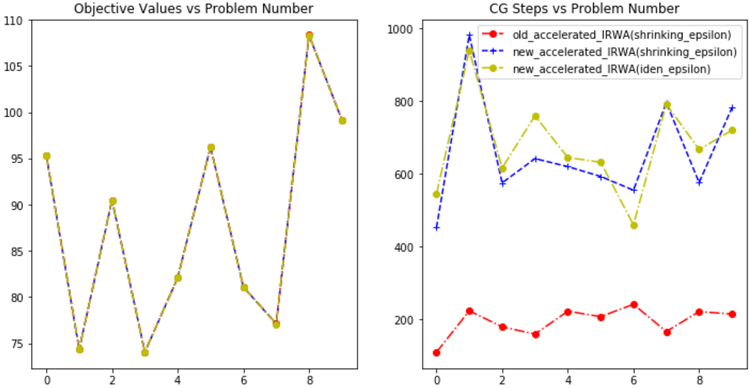

Figure : In all 10 problems, the old IRWA with iterative re-weighting obtains similar objective function values with many fewer CG steps.

Reference

Thank you!

Iteratively Reweighted Linear Least Squares
for
Exact Penalty Subproblems on Product Sets

with F. Curtis, H. Wang and J. Wang. SIAM J. Optim. **25**(2015): 261 - 294.

The Euclidean Huber Distance to a Convex Set

$C \subset \mathbb{E}$ be non-empty closed and convex
$h := \mathrm{dist}_2 \left(\cdot \mid C \right)$ the Euclidean distance to C

The Euclidean Huber distance to C is just the Moreau-Yosida envelope of the distance to C.

$$e_\gamma h(y) = \begin{cases} \frac{1}{2\gamma} \mathrm{dist}_2^2 \left(y \mid C \right) & \text{, if } \mathrm{dist}_2 \left(y \mid C \right) \le \gamma, \\ \mathrm{dist}_2 \left(y \mid C \right) - \frac{\gamma}{2} & \text{, if } \mathrm{dist}_2 \left(y \mid C \right) > \gamma, \end{cases}$$

$$\mathrm{Prox}_{\gamma,h}(y) = \begin{cases} P_C(y) & \text{, if } \mathrm{dist}_2 \left(y \mid C \right) \le \gamma, \\ y - \frac{\gamma}{\mathrm{dist}_2 \left(y \mid C \right)} (y - P_C(y)) & \text{, if } \mathrm{dist}_2 \left(y \mid C \right) \le \gamma. \end{cases}$$