Quadratic Majorization-Minimization in Regularized Radar Imaging

Lee C. Potter
Ohio State University
Columbus, OH, USA

Shannon Blunt
Kansas University
Lawrence, KS, USA

Samuel Pine
Matrix Research, Inc.
Dayton, OH, USA

***Abstract*—Quadratic majorization-minimization is found in a 1937 paper by Weiszfeld and provides an easily accessible iterative optimization algorithm for non-convex and non-smooth optimization tasks; the iteration is often referred to as "iteratively re-weighted least squares" (IRLS). In this manuscript, IRLS is employed as a unifying tutorial description of many regularized imaging techniques previously proposed for radar imaging. The cost function implicit in an IRLS imaging approach is made explicit, yielding a Bayesian interpretation with specific underlying assumptions regarding clutter, noise, and prior distribution on the unknown range-angle-Doppler maps. In particular, the IRLS framework is used to establish the convergence and sparse imaging properties of the reiterative super-resolution (RISR) [1] and background supplemental loading (BaSL) [2] algorithms.**

***Index Terms*—radar imaging, iteratively re-weighted least squares, non-convex optimization**

I. Introduction

Iteratively re-weighted least-squares (IRLS) is a method for solving minimization problems involving non-quadratic cost functions, possibly non-convex or non-smooth. The solution is found by successively determining the minimizer of a quadratic surrogate that locally approximates and bounds the original cost function. The sequence of quadratic problems is easily tackled with numerical linear algebra. The simplicity and generality render the approach familiar and versatile; indeed, Google Scholar reports 43,900 results for "iteratively re-weighted least squares." Forming and minimizing a quadratic surrogate function that upper-bounds the cost function is an example of majorization-minimization. In this manuscript, we use IRLS as a unifying framework for a host of imaging algorithms found in the literature. The framework is used to establish convergence and describe sparse recovery properties. Significantly, the tutorial makes explicit the underlying assumptions on clutter and image priors that are implicit in an imaging algorithm. Codes are available at github.com/ECE36/IRLS.

Let x denote the complex-valued radar image to be recovered from radar data. Any imaging method consists of three components, here expressed in the language of Bayesian estimation. First, there is a choice for the negative log likelihood function, $f(x)$, which incorporates a forward model of the physics relating the unknown scene, x, to the noisy measurements, y. To specify the function $f(x)$, physical assumptions are adopted, such as radar waveforms, propagation characteristics, antenna patterns, and a plane-wave approximation. The function $f(x)$ also includes assumed probability distributions describing clutter and thermal noise. Second, there is a choice of an image prior. A regularization penalty, $\lambda g(x)$, may be interpreted as adopting a prior distribution on the scene proportional to $e^{-\lambda g(x)}$. For example, many priors have been suggested to encourage a sparse scene [3]–[8]. Note that $g(x)$ may be non-quadratic, non-convex, or not everywhere differentiable. Third, there is a choice of numerical procedure for minimizing the risk function $J(x) = f(x) + \lambda g(x)$. The estimated scene, $\widehat{x}$, minimizing the risk may be interpreted as a maximum a posteriori probability (MAP) estimator,

$$\widehat{x} = \arg\min_x \; J(x) = \arg\min_x \; f(x) + \lambda g(x). \tag{1}$$

This work is supported by the Defense Advanced Research Projects Agency under Air Force Research Laboratory contract FA2385-23-C-0002. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

II. Signal Model

Radar scattering is modeled as linear, so discretization of a scene yields the linear measurement model

$$y = Ax + w, \tag{2}$$

where $A \in \mathbb{C}^{m\times n}$, $x \in \mathbb{C}^n$, and $w \in \mathbb{C}^m$ models clutter plus thermal noise as a zero-mean complex Gaussian random vector with covariance R. The vector $y \in \mathbb{C}^m$ combines data samples across fast-time, pulses, and array channels. Generally, we encounter $m < n$. Derivation of the discretized forward operator, A, is found in many texts and papers for a variety of imaging scenarios and is therefore omitted here. For a phased-array radar, x represents sampled bins in range, azimuth, elevation, and range rate (Doppler). Given covariance, R, the Gaussian likelihood function imposes data fidelity, and the negative log likelihood is, to an additive constant

$$f(x) = \tfrac{1}{2}\|R^{-1/2}(Ax - y)\|_2^2. \tag{3}$$

The action of the whitening filter, $R^{-1/2}$, suppresses clutter and noise modeled by the additive Gaussian component, $w \sim \mathcal{CN}(0, R)$, in (2).

III. Regularization Penalty

In radar imaging, the task of inverting f to determine the unknown scene, x, may be ill-conditioned or even ill-posed. Accordingly, a regularization penalty is adopted to stabilize the

TABLE I
SEPARABLE PENALTY FUNCTIONS AND THEIR CORRESPONDING SCALAR WEIGHTS, $0 < q < 2$.

penalty $\phi(x_i^{(k)})$	scalar update	example
$\lvert x_i^{(k)}\rvert^q$	$p_i^{(k)} \leftarrow \frac{1}{q}\lvert x_i^{(k)}\rvert^{2-q}$	[10], [11]
$\left(\lvert x_i^{(k)}\rvert^2+\epsilon\right)^{q/2}$	$p^{(i)} \leftarrow \frac{1}{q}\left(\lvert x_i^{(k)}\rvert^2+\epsilon\right)^{1-q/2}$	[3], [4]
$\left(\lvert x_i^{(k)}\rvert^2+\epsilon^{(k)}\right)^{q/2}$	$p_i^{(k)} \leftarrow \frac{1}{q}\left(\lvert x_i^{(k)}\rvert^2+\epsilon^{(k)}\right)^{1-q/2}$	[12]–[14]
$\log\left(\lvert x_i^{(k)}\rvert^2+\epsilon\right)$	$p_i^{(k)} \leftarrow \frac{1}{2}\left(\lvert x_i^{(k)}\rvert^2+\epsilon\right)$	[9]
$\log\left(\lvert x_i^{(k)}\rvert+\epsilon\right)$	$p_i^{(k)} \leftarrow \lvert x_i^{(k)}\rvert^2+\epsilon\lvert x_i^{(k)}\rvert$	[15]

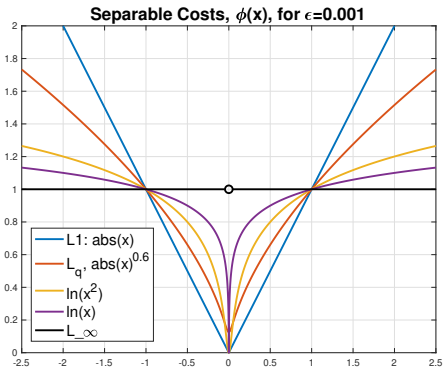

Fig. 1. Illustration of several separable penalty functions, ϕ, each of which gives rise to an implied image prior and diagonal weighting matrix, $P_{(k)}$, in the iteratively re-weighted least-squares algorithm. The penalty functions encourage sparse solutions. The third and fourth lines shown in legend correspond to rows four and five of Table 1 and have been scaled and translated for visualization.

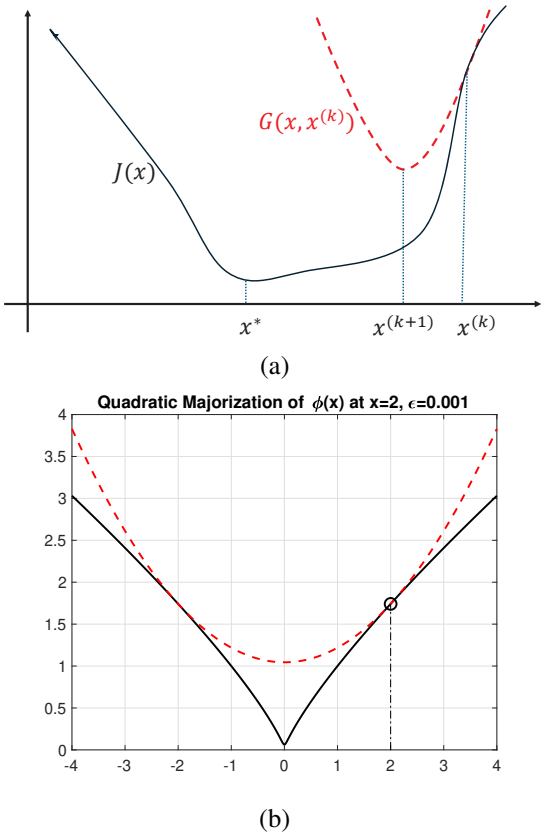

Fig. 2. (a) Illustration of quadratic majorization-minimization of a notional non-convex cost. At each iteration, parameters x are updated by a minimizing quadratic approximation, seen here as the dashed red parabola upper-bounding the solid black cost surface. (b) Quadratic majorization of $\phi(x_i) = (\lvert x_i\rvert^2 + \epsilon)^{q/2}$ at $x_i = 2$ for $q = 0.8$, $\epsilon = 0.001$.

inversion, and the penalty implicitly imposes an assumed prior distribution on the scene, x, in the Bayesian interpretation. For the regularization penalty, $\lambda g(x)$, consider a separable function of the form $\lambda g(x) = \lambda \sum_i \phi(x_i)$, where ϕ is a scalar function operating on a single pixel, or "bin" x_i, of the range/angle/Doppler scene, x. To begin, we highlight the choice

$$g(x) = \sum_{i=1}^{n} \phi(x_i) = \sum_{i=1}^{n} \left(\lvert x_i\rvert^2 + \epsilon\right)^{q/2}. \tag{4}$$

This choice of functional g is an ϵ-smoothed version of the q-norm of the scene. A choice of $0 < q \leq 1$ has been widely adopted to promote a sparse scene [3]–[6]. For $0 < q < 1$, (4) fails the triangle inequality and is a quasi-norm. Further, for $0 < q < 1$, $g(x)$ is neither convex nor everywhere differentiable. The first column of Table I lists five example choices of ϕ found in the literature [3], [4], [9]–[15].

IV. IRLS

To proceed with solution of the optimization task in (1), we follow a quadratic majorization-minimization approach; an early example is found in a 1937 paper by Weiszfeld [16], and the approach is sometimes known as Lawson's method owing to its appearance in a 1961 doctoral dissertation [17].

A. Quadratic majorization-minimization

Consider a current estimate, $x^{(k)}$, of the scene. The cost function $J(x)$ in the vicinity of $x^{(k)}$ is approximated as a quadratic $G(x, x^{(k)})$ that dominates $J(x)$ and is tangent to $J(x)$ at $x = x^{(k)}$. This is illustrated in Fig. 2(a). Then, the unique global minimum of the quadratic $G(x, x^{(k)})$ provides an easily computed update,

$$x^{(k+1)} = \arg\min_x G(x, x^{(k)}). \tag{5}$$

Because $f(x)$ from (3) is already quadratic in x, construction of the quadratic majorizer $G(x, x^{(k)})$ of $J(x)$ requires only quadratic majorization of $g(x)$. Further, $g(x)$ is conveniently separable as a scalar function on each entry of x, so we focus on a quadratic majorizer, $h(z)$, of the function $\phi(z)$. To illustrate, we continue with the case $\phi(z) = \left(\lvert z\rvert^2 + \epsilon\right)^{q/2}$. Because $\phi(z) = \phi(-z)$, the quadratic function is $h(z) = w_0\lvert z\rvert^2 + c_0$. Now, given a point z_0 (e.g, from the previous

iteration), we seek $h(z)$ to match both $\phi(z)$ and its derivative at z_0, as illustrated in Fig. 2(b):

$$h(z_0) = \phi(z_0) \tag{6}$$

$$h'(z_0) = \phi'(z_0). \tag{7}$$

By direct differentiation, (7) implies a weight w_0 given by

$$w_0 = \tfrac{q}{2}\left(|z_0|^2 + \epsilon\right)^{q/2-1}. \tag{8}$$

Thus, we have learned

$$h(z) = z^* p_0^{-1} z + c_0 \tag{9}$$

where $p_0 = 1/w_0 = \frac{2}{q}\left(|z_0|^2 + \epsilon\right)^{1-q/2}$ and some constant c_0 depending on z_0 via (6). Note the exponent $1 - q/2$ is non-negative for $q \leq 2$. Continuing, because $h(z)$ is convex while $\phi(z)$ is concave, (6) and (7) imply that $h(z) \geq \phi(z)$ for all z.

Thus, returning to the penalty function $g(x)$ in (4) defined using the full vector of pixels, x, we have the quadratic majorization of $J(x)$ at iterate $x^{(k)}$

$$G(x, x^{(k)}) = \tfrac{1}{2}\|R^{-1/2}\left(Ax - y\right)\|^2 + \lambda\tfrac{q}{2}x^H P_{(k)}^{-1} x + c_k$$

where ^{H} denotes conjugate transpose, c_k is a constant term depending on $x^{(k)}$, and $P_{(k)}$ is a diagonal matrix with entries

$$P_{(k)}[i,i] = \left(|x_i^{(k)}|^2 + \epsilon\right)^{1-q/2}. \tag{10}$$

Thus, we arrive to the simple update rule,

$$x^{(k+1)} = \arg\min_x \|R^{-1/2}\left(Ax - y\right)\|^2 + \lambda q x^H P_{(k)}^{-1} x. \tag{11}$$

First-order optimality for the quadratic cost in (11) yields

$$x^{(k+1)} = \left(A^H R^{-1} A + \lambda q P_{(k)}\right)^{-1} A^H R^{-1} y, \tag{12}$$

Define $\tilde{\lambda} = \lambda q$. Invertibility of R^{-1} and the matrix inversion lemma[1] provide the alternative formulation,

$$\boxed{x^{(k+1)} = P_{(k)} A^H \left(A P_{(k)} A^H + \tilde{\lambda} R\right)^{-1} y.} \tag{13}$$

The iteration in (13) solves the MAP estimation in (1) using a sequence of quadratic problems. The negative log likelihood term, $f(x)$, adopts the linear data model embodied in A and a multivariate Gaussian model on the clutter plus noise, with covariance R. This clutter plus noise covariance may be estimated from auxiliary data, for example [19], [20]. The iteration in (13) also implicitly adopts a prior distribution on the image; the negative log prior is $\lambda g(x) = \lambda \sum_i \phi(x_i)$. Each choice of scalar penalty function, ϕ, results in a choice of this image prior and in turn defines the diagonal matrix $P_{(k)}$ computed from the reconstructed scene at each iteration. A few example choices of ϕ are depicted in Fig. 1.

The same IRLS procedure producing (13) can be applied to any $g(x)$ that is concave in $|x|^2$. Table I lists the scalar updates for the diagonal entries of $P_{(k)}$ for several choices of regularization function, ϕ. Note also that the smoothing parameter, ϵ, may be varied with iteration number k, resulting in $\lim_{k\to\infty} \epsilon^{(k)} = 0$.

[1]The matrix inversion lemma (e.g., [18]): for invertible matrices A and C, $(A + BCD)^{-1}BC = A^{-1}B(C^{-1} + DA^{-1}B)^{-1}$. Additionally, $(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$.

B. Conjugate gradients

We can approximately solve (13) at each IRLS iteration, k, using a few conjugate gradient (CG) steps. To this end, define

$$C = \begin{bmatrix} P_{(k)}^{1/2} A^H R^{-1/2} \\ \sqrt{\tilde{\lambda}} I \end{bmatrix} \qquad u = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{\lambda}} R^{-1/2} y \end{bmatrix}. \tag{14}$$

Define $\theta^{(k+1)}$ as an intermediate variable in $\mathbb{C}^m$; by direct computation, we have

$$\theta^{(k+1)} = (C^H C)^{-1} C^H u \tag{15}$$

$$= \left(R^{-1/2} A P_{(k)}^{-1} A^H R^{-1/2} + \tilde{\lambda} I\right)^{-1} R^{-1/2} y. \tag{16}$$

And, from (13)

$$x^{(k+1)} = P_{(k)} A^H R^{-1/2} \theta^{(k+1)}. \tag{17}$$

Each CG step requires only two whitening operations, $R^{-1/2}$, and one application each of the forward operator, A, and its adjoint, A^H. The linear model operator, A, often can be computed with fast Fourier transforms without explicitly forming the very large $m \times n$ matrix. Further, a simple Jacobi pre-conditioner may provide nice computational advantage, especially if the same $R^{-1/2}$ is used for many observations.

V. Reiterative MMSE

In a sequence of papers including [1], [2], [21], Blunt, Gerlach, Jones and co-authors arrive at the same approach via a "reiterative minimum mean squared error" (RMMSE) framework, motivated by strategies from code division multiple access multi-user detection. The RMMSE facilitates nulling of self-interference from side-lobes, which in the IRLS framework is seen as a consequence of the sparsity inducing prior, $g(x)$. The "reiterative super resolution" (RISR) algorithm [1] implicitly adopts a covariance R for additive noise and the $\phi(x_i) = \log|x_i| + \epsilon$ penalty in Table I with $\epsilon = 0$. Additionally, the background supplemental loading (BaSL) algorithm [2] extends RISR to include clutter covariance in the $\tilde{\lambda}R$ loading term.

A. Log penalty

We next consider the log penalty seen in line 4 of Table I and show that the log penalty is a limit of q quasi-norms as $q \to 0$. With $g(x) = \frac{1}{q}\|x\|_q^q$, we have

$$\widehat{x} = \arg\min_x \; f(x) + \lambda g(x) \tag{18}$$

$$= \arg\min_x \; f(x) + \lambda g(x) - \lambda\tfrac{n}{q} \tag{19}$$

$$= \arg\min_x \; f(x) + \lambda\left\{\tfrac{1}{q}\sum_{i=1}^{n}\left(|x_i|^q - 1\right)\right\} \tag{20}$$

Now, consider the limit as $q \to 0$.

$$\lim_{q\to 0} \; \tfrac{1}{q}\left(|x_i|^q - 1\right) = \lim_{q\to 0} \; \frac{|x_i|^q - 1}{q} \tag{21}$$

$$= \lim_{q\to 0} \; \frac{\frac{d}{dq}\{|x_i|^q - 1\}}{\frac{d}{dq}\{q\}} \tag{22}$$

$$= \lim_{q\to 0} \; |x_i|^q \ln|x_i| \tag{23}$$

$$= \ln|x_i|. \tag{24}$$

where (22) follows from application of L'Hôpital's rule. Thus, the log penalty function may be interpreted as the limit of the q-norm penalty as q goes to zero.

B. Smoothing

The $\epsilon > 0$ appearing in Table I provides a smoothing of the q-norm, resulting in ϕ differentiable at the origin. This smoothing has been widely employed, playing a role in both effective numerical results [3], [4], [12] and convergence proofs [12], [13]. Experimental evidence [3], [12] suggests that $\epsilon > 0$ improves performance for $q < 1$, enabling IRLS to recover signals with more nonzero components compared to the un-smoothed version.

In the reiterative MMSE framework, a scaling is optionally employed to modify the IRLS iteration in (13)

$$x^{(k+1)} = Q_{(k)} A^H \left(A P_{(k)} A^H + \tilde{\lambda} R \right)^{-1} y, \qquad (25)$$

where the $[i,i]$ element of the diagonal matrix $Q^{(k)}$ is given by

$$Q_{(k)}[i,i] = \frac{1}{a_i^H \left(A P_{(k)} A^H + \sigma^2 I \right)^{-1} a_i}. \qquad (26)$$

Here, a_i is the i^{th} column of A and $\sigma^2 I$ is the noise-only component of the clutter-plus-noise covariance matrix, R. This scaling is similar to forcing weight vectors to have unit energy, as in minimum variance distortionless response (MVDR) beamforming, but omits the clutter component from the covariance matrix.

From the IRLS framework, we learn that the scaling in (25)-(26) performs the same smoothing effect as ϵ in (4). Note firstly that for any matrix A matrix having orthonormal columns, it is easy to verify that $Q_{(k)}[i,i] = |x_i|^2 + \sigma^2$. Hence, for orthonormal A the MVDR-inspired scaling is equivalent to the ϵ-smoothing in (13) for $\epsilon = \sigma^2$. Secondly, for more general $A \in \mathbb{C}^{m \times n}$, consider keeping m rows at random from the unitary discrete Fourier transform (DFT) matrix with $\{k,l\}$ entry $\frac{1}{\sqrt{n}} e^{-j 2\pi kl/n}$. In this case, the $Q_{(k)}[i,i]$ concentrates at $|x_i|^2 + \epsilon$ with $\epsilon = n\sigma^2/m$. (The concentration of measure proof is omitted here due to length.) Hence, the MVDR-inspired scaling in (25)-(26) is very similar to the ϵ-smoothing of the ϕ function in (4) for $\epsilon = n\sigma^2/m$.

VI. MMSE estimator

As an aside, consider the special case of a zero-mean multi-variate Gaussian prior on the scene, with covariance Σ. The MAP image then coincides with the minimum mean squared error (MMSE) estimator and the best linear unbiased estimator:

$$\widehat{x}_{\Sigma} = (A^H R^{-1} A + \underbrace{\sigma^2 \Sigma^{-1}}_{\substack{\text{Tikhonov} \\ \text{regularization} \\ \text{of MLE}}})^{-1} \underbrace{A^H R^{-1} y}_{\substack{\text{whitened} \\ \text{matched filter}}} \qquad (27)$$

$$= \Sigma A^H \left(A \Sigma A^H + \sigma^2 R \right)^{-1} y \qquad (28)$$

where (28) is obtained from (27) by application of the matrix inversion lemma. As indicated by the bracketed notation in (27), the inverse covariance Σ^{-1} of the image prior provides a Tikhonov regularization of the least-squares solution, combating any ill-conditioning and providing existence of an inverse even if A is not full column rank. Additionally, the term $A^H R^{-1} y$ is the whitened matched filter, which employs the inverse covariance of the clutter-plus-noise term. However, the covariance Σ of the image prior is not known in practice.

VII. Convergence and Sparse Recovery

A. Convergence

For any $\phi(x)$ that is concave in $|x|^2$ on $[0,\infty)$, the IRLS sequence gives a descent method for $\arg\min_x \; f(x) + \lambda g(x)$, and thus is convergent to some local minimum [22]. At each step of the IRLS procedure, a quadratic problem is solved, and conjugate gradient steps are used to this end. However, in practice the CG steps must be terminated before convergence. Yet, Fornasier et al. [23] established thresholds such that early termination of CG steps does not prevent convergence of the IRLS procedure.

For $q \geq 1$, the costs in Table I are convex and the IRLS converges to the global optimum. For $0 < q < 1$, three observations are available from the literature. First, as q decreases there is increasing sensitivity to local minima and choice of initialization. Second, for a specific schedule of $\epsilon^{(k)}$ as a function of iteration number k, Daubechies et al. [13] established a post-facto certificate of global convergence, but the certificate may be sensitive to finite precision computation. Third, empirical results [12] exhibit global convergence in many cases, especially when using a schedule of $\epsilon^{(k)} \to 0$ and $q^{(k)} \to q_* < 1$. These empirical results informally suggest: select q less than 1 for a super-linear rate of convergence; and, select a scheduled smoothing constant, $\epsilon^{(k)} \to 0$ to mitigate local minima.

B. Sparse recovery

The literature from compressed sensing provides sufficient conditions under which an under-determined system of linear equations, $y = \Phi x + w$, $m < n$, admits a unique, stable solution when x can be well approximated by only K or fewer non-zero entries. For the radar imaging scenarios considered here, the linear operator Φ corresponds to the whitened forward model, $R^{-1/2} A$.

For sparse recovery guarantees, consider the *q-nullspace property*: no signal in the nullspace of Φ can have half or more of its q-norm "energy" on only K coefficients [24]. If $R^{-1/2} A$ has the q-nullspace property, then all minimizers of $f(x) + \lambda \|x\|_q$, $0 \leq q < 1$, coincide with the (unique) ℓ_1 minimizer.

Given this result, one might ask: why not just use the convex ℓ_1 regularization penalty given that solutions coincide for $0 < q < 1$? The answer lies in two benefits. First, $q < 1$ can offer a super-linear local convergence rate [13], [25]. Second, $q < 1$ has been empirically observed to provide robust recovery for larger K [12].

VIII. Detection Statistics

A solution $\widehat{x}$ in (1) is often sought for the purpose detection. Recall a_i denotes a single column of A. Defi iid secondary data vectors, z_k, $k = 1, ..., N$, of clutter p thermal noise:

$$\begin{aligned} y &= c\,a_i + \sigma^2 w \\ z_k &= w_k. \end{aligned}$$

Consider the binary hypotheses

$$\begin{aligned} \text{no target } H_0 \quad & y \sim \mathcal{CN}(0, \sigma^2 R), \quad z_k \sim \mathcal{CN}(0, R) \text{ ii} \\ \text{target } H_1 \quad & y \sim \mathcal{CN}(ca_i, \sigma^2 R), \quad z_k \sim \mathcal{CN}(0, R) \text{ ii} \end{aligned}$$

This gives rise to a composite test with three unknown deterministic parameters: complex amplitude $c \in \mathbb{C}$, noise scaling $\sigma^2 \in \mathbb{R}_+$, and covariance R. As developed by Kraut, Scharf and Butler [26], the adaptive coherence estimate (ACE) provides a test statistic that is the generalized likelihood ratio test, is a uniformly most powerful invariant test, and has constant false alarm rate (CFAR). The ACE test, for threshold η, is

$$t_i = \frac{|a_i^H S^{-1} y|^2}{\left(a_i^H S^{-1} a_i\right)\left(y^H S^{-1} y\right)} \gtrless \eta, \tag{29}$$

where $S = \frac{1}{N}\sum_{k=1}^{N} z_k z_k^H$ is the sample covariance matrix from the secondary data. Note that S is the unconstrained maximum likelihood estimate of R. Observe that t_i is the whitened matched filter output, magnitude squared and normalized. The hypothesis test is conducted at each coordinate of x by selecting a_i, and this is accomplished without regard to any correlations among the a_i, $i = 1, ..., n$.

If the true signal, x_0, has only a single non-zero element, then the whitened matched filter provides a principled test statistic for detection. But, if x_0 has multiple non-zero entries and the $R^{-1/2}a_i$ candidate response vectors are not orthogonal, then performance of the GLRT degrades. The MAP estimate, $\widehat{x}$, from (1) provides an alternative test statistic to account for the side-lobes of one target biasing the ACE test statistic for another target. Scaling to achieve a CFAR test, we arrive to

$$t_i = \frac{|\widehat{x}_i|^2}{\left(y^H S^{-1} y\right)} \gtrless \eta \tag{30}$$

This is illustrated for one realization in Fig. 3, where A is 170×512 and is generated by randomly drawn rows of the DFT, normalized to have unit-length columns; additive complex noise has unit power. Parameters are $p = 0.8$, $\lambda = 0.1$, $\epsilon = 0.001$, and 5 CG steps per iteration. The matched filter result permits detection of the strong signals, but weaker signals are lost in a myriad of false alarms due to point response sidelobes; in contrast, the MAP solution is able to deconvolve sidelobes, permitting detection of weaker signals, despite locations being off-grid and therefore not exactly matched to locations encoded in the columns of A. For the third and fourth high-energy reflectors, additional low-energy detections appear adjacent to the true positions to account for the model-mismatch of off-grid points.

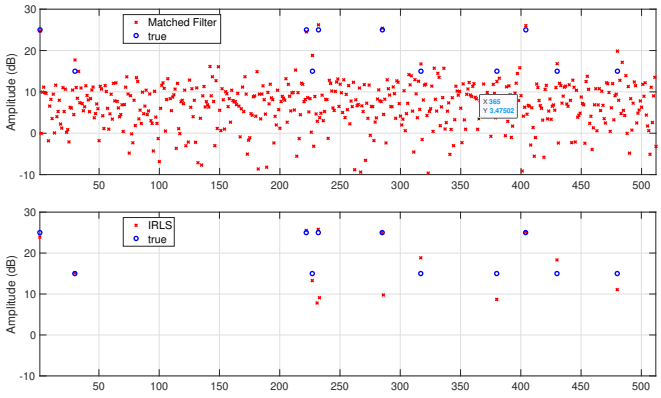

Fig. 3. Matched filter, $A^H R^{-1} y$ (top); MAP solution $\widehat{x}$ from (13) (bottom). For illustration, simulated signals are off-grid and 11-sparse. A dynamic range of 40 dB is displayed.

IX. Partial History

Quadratic majorization-minimization and the resulting IRLS iteration has appeared in many instances, including a 1937 use by Weiszfeld to minimize the weighted sum of Euclidean lengths (the Fermat-Weber problem). Likewise, the approach appears in Lawson's 1961 dissertation for function approximation with Chebyshev polynomials. Variants of the IRLS algorithm have appeared many times under many names. The brief discussion offered here is incomplete and merely meant to be representative.

Katz (1974) [27] and Voss & Eckhardt (1980) [28] established linear convergence rate of Weiszfeld's algorithm. Daubechies et al. (2008) [13] established convergence results and sparse recovery proofs for the constrained case ($y = Ax$) and a specified sequence $\epsilon^{(k)} \to 0$; these results were generalized to the unconstrained case by Lai et al. (2013) [25]. Fornasier et al. (2016) [23] established convergence proofs with incomplete CG iterations.

Variants of the IRLS algorithm have appeared numerous times, including by Holland & Welsch (1977) [29] for robust regression, Lee et al. (1987) [30] for bandlimited extrapolation, and Han et al. (1997) [31] for solving nonlinear partial differential equations. Bouman & Sauer (1993) [32] presented an axiomatic selection of q-norm of the image gradient in the context of edge-preserving regularized imaging. Vogel & Oman (1996) [33] and Charbonnier et al. (1997) [34] similarly adopted an ℓ_1 penalty on the gradient for total variation image denoising, and solved using IRLS. Gorodnitsky & Rao (1997) [10] presented IRLS with $\epsilon_{(k)} = 0$ and $q = 1$ in the "FOCUSS" algorithm and connected the result with sparse recovery, providing sufficient conditions on A and the sparsity, $\|x\|_0$.

With particular application to radar imaging, the IRLS algorithm appears in Cetin & Karl (2001) [3]; Kragh & Kharbouch (2006) [4]; Tan et al. (2011) [11], Blunt, Chan & Gerlach (2011) [1], and Jones et al. (2020) [2].

Interest in (1) with the ℓ_1 norm penalty for deconvolution has long history in seismic applications (e.g., Taylor et al. 1979 [35]). Rudin, Osher, & Fatemi (1992) [36] used ℓ_1 norm on the gradient for image denoising. Interest in a variety of optimization tools for computing $\widehat{x}$ was spurred by the compressive sensing results of Tao, Candes and Romberg (2006) [37], and Donoho (2006) [38], for recovery of sparse signals from linear measurements. Other than IRLS, perhaps the most widely adopted solver is the fast iterative shrinkage-thresholding algorithm (FISTA) by Beck & Teboulle (2009) [39], which is an accelerated proximal gradient descent, and hence a first order method readily applicable to large-scale problems.

Chartrand & Staneva (2008) [40] and Chartrand & Lin (2008) [12] spurred increased interest in $q < 1$, with empirical evidence that the non-convex penalty can provide improved convergence rate and can increase the number of non-zeros in x that can be recovered. Further, Chartrand & Lin argued for a cooling schedule on $\epsilon^{(k)}$ as an empirically effective "smoothing" to avoid local minima in the non-convex case.

REFERENCES

[1] S. D. Blunt, T. Chan, and K. Gerlach, "Robust DOA estimation: The reiterative super-resolution (RISR) algorithm," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 47, no. 1, pp. 332–346, 2011.

[2] C. C. Jones, L. A. Harnett, C. A. Mohr, S. D. Blunt, and C. T. Allen, "Structure-based adaptive radar processing for joint clutter cancellation and moving target estimation," in *2020 IEEE International Radar Conference (RADAR)*, 2020, pp. 413–418.

[3] M. Çetin and W. C. Karl, "Feature-enhanced synthetic aperture radar image formation based on nonquadratic regularization," *IEEE Trans. Image Process.*, vol. 10, pp. 623–631, Apr. 2001.

[4] T. J. Kragh and A. A. Kharbouch, "Monotonic iterative algorithms for SAR image restoration," in *ICIP*, 2006, pp. 645–648.

[5] J. Ender, "On compressive sensing applied to radar," *Signal Process.*, vol. 90, no. 5, pp. 1402–1414, May 2010.

[6] L. C. Potter, E. Ertin, J. T. Parker, and M. Cetin, "Sparsity and compressed sensing in radar imaging," *Proc. IEEE*, vol. 98, no. 6, pp. 1006–1020, 2010.

[7] L. Anitori, A. Maleki, M. Otten, R. Baraniuk, and P. Hoogeboom, "Design and analysis of compressed sensing radar detectors," *IEEE Trans. Signal Process.*, vol. 61, no. 4, pp. 813–827, Feb 2013.

[8] N. A. Goodman and L. C. Potter, "Pitfalls and possibilities of radar compressive sensing," *Appl. Opt.*, vol. 54, no. 8, pp. C1–C13, Mar 2015.

[9] T. Hebert and R. Leahy, "A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors," *IEEE Trans. Med. Imaging*, vol. 8, no. 2, pp. 194–202, 1989.

[10] I. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, 1997.

[11] X. Tan, W. Roberts, J. Li, and P. Stoica, "Sparse learning via iterative minimization with application to MIMO radar imaging," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1088–1101, 2011.

[12] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *2008 IEEE ICASSP*, 2008, pp. 3869–3872.

[13] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Commun. Pure Appl. Math.*, vol. 63, no. 1, pp. 1–38, 2010.

[14] M.-J. Lai and J. Wang, "An unconstrained ℓ_q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems," *SIAM J. Optimiz.*, vol. 21, no. 1, pp. 82–101, 2011.

[15] P. L. Combettes and J.-C. Pesquet, *Proximal Splitting Methods in Signal Processing*. New York, NY: Springer New York, 2011, pp. 185–212.

[16] E. Weiszfeld, "Sur le point pour lequel la somme des distances de n points donnès est minimum," *Tohoku Mathematical Journal, First Series*, vol. 43, pp. 355–386, 1937.

[17] C. L. Lawson, "Contributions to the theory of linear least maximum approximations," Ph.D. dissertation, UCLA, 1961.

[18] L. Guttman, "Enlargement methods for computing the inverse matrix," *The Annals of Mathematical Statistics*, vol. 17, pp. 336–343, 1946.

[19] A. Aubry, A. De Maio, and L. Pallotta, "A geometric approach to covariance matrix estimation and its applications to radar problems," *IEEE Trans. Signal Process.*, vol. 66, no. 4, pp. 907–922, 2018.

[20] A. P. Shikhaliev, L. C. Potter, and Y. Chi, "Low-rank structured covariance matrix estimation," *IEEE Signal Process. Lett.*, vol. 26, no. 5, pp. 700–704, 2019.

[21] S. Blunt and K. Gerlach, "A novel pulse compression scheme based on minimum mean-square error reiteration," in *2003 Proceedings of the International Conference on Radar*, 2003, pp. 349–353.

[22] J. Palmer, K. Kreutz-Delgado, B. Rao, and D. Wipf, "Variational EM algorithms for non-Gaussian latent variable models," in *Advances in Neural Information Processing Systems*, Y. Weiss, B. Schölkopf, and J. Platt, Eds., vol. 18. MIT Press, 2005.

[23] M. Fornasier, S. Peter, H. Rauhut, and S. Worm, "Conjugate gradient acceleration of iteratively re-weighted least squares methods," *Comput Optim Appl*, vol. 65, p. 205–259, 2016.

[24] R. Gribonval and M. Nielsen, "Highly sparse representations from dictionaries are unique and independent of the sparseness measure," *Appl. Comput. Harmon. Anal.*, vol. 22, no. 3, pp. 335–355, 2007.

[25] M.-J. Lai, Y. Xu, and W. Yin, "Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization," *SIAM J. Numer. Anal.*, vol. 51, no. 2, pp. 927–957, 2013.

[26] S. Kraut, L. Scharf, and R. Butler, "The adaptive coherence estimator: a uniformly most-powerful-invariant adaptive detection statistic," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 427–438, 2005.

[27] I. Katz, "Local convergence in Fermat's problem," *Math. Program.*, vol. 6, pp. 89–104, 1974.

[28] H. Voß and U. Eckhardt, "Linear convergence of generalized weiszfeld's method," *Computing*, vol. 25, p. 243–251, 1980.

[29] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares," *Commun. Stat.- Theory Methods*, vol. 6, no. 9, pp. 813–827, 1977.

[30] H. Lee, D. Sullivan, and T. Huang, "Improvement of discrete band-limited signal extrapolation by iterative subspace modification," in *IEEE ICASSP*, 1987, pp. 1569 – 1572.

[31] W. Han, S. Jensen, and I. Shimansky, "The Kačanov method for some nonlinear problems," *Applied Numerical Mathematics*, vol. 24, no. 1, pp. 57–79, 1997.

[32] C. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Trans. Image Process.*, vol. 2, pp. 296–310, Mar. 1993.

[33] C. R. Vogel and M. E. Oman, "Iterative methods for total variation denoising," *SIAM J. Sci. Comput.*, vol. 17, no. 1, pp. 227–238, 1996.

[34] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Deterministic edge-preserving regularization in computed imaging," *IEEE Trans. Image Process.*, vol. 6, no. 2, pp. 298–311, 1997.

[35] H. L. Taylor, S. C. Banks, and J. F. McCoy, "Deconvolution with the ℓ_1 norm," *Geophysics*, vol. 44, pp. 39–52, 1979.

[36] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, no. 5, pp. 259–268, 1992.

[37] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[38] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

[39] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, no. 1, pp. 183–202, 2009.

[40] R. Chartrand and V. Staneva, "Restricted isometry properties and non-convex compressive sensing," *Inverse Probl.*, vol. 24, no. 3, p. 035020, May 2008.