THOUGHTS ON SUPERIORIZATION

Preprint · April 2019

CITATIONS
0

READS
23

1 author:

Charles L. Byrne
University of Massachusetts Lowell
170 PUBLICATIONS **4,823** CITATIONS

SEE PROFILE

THOUGHTS ON SUPERIORIZATION

CHARLES L. BYRNE

ABSTRACT. Let $T : \mathbb{R}^N \to \mathbb{R}^N$ be such that the iterative algorithm $x^{k+1} = Tx^k$ converges to a member of $Fix(T)$, the set of fixed points of T; the T is what is called the "Basic Algorithm" . Let $h : \mathbb{R}^N \to \mathbb{R}_+$. The goal of the superiorization method (SM) is to perturb each iterate of the basic algorithm to obtain a member x of $Fix(T)$ for which $h(x)$ is small, although x need not minimize h over $Fix(T)$. The emphasis so far has been on the "perturbation resilience"of basic algorithms, that is, on whether or not a suitably perturbed version of a convergent basic algorithm will still converge to a member of $Fix(T)$. Because determining when SM achieves its stated objective of reducing h remains open, we take the opportunity to examine the significance of SM and to contrast it with alternative approaches.

1. Overview of Superiorization

Let $T : \mathbb{R}^N \to \mathbb{R}^N$ be such that the iterative algorithm $x^{k+1} = Tx^k$ converges to a member of $Fix(T)$, the set of fixed points of T; the T is what is called the "basic algorithm" in [9]. Let $h : \mathbb{R}^N \to \mathbb{R}_+$. The goal of the superiorization method (SM) is to modify the basic algorithm to obtain a member x of $Fix(T)$ for which $h(x)$ is small, although x need not minimize h over $Fix(T)$ [12, 10]. To achieve this goal the usual SM is to use an iteration of the form

$$x^{k+1} = Tx^k - t_k \nabla h(Tx^k), \tag{1.1}$$

for appropriately chosen parameters $t_k > 0$. If $T = T_I T_{I-1} \cdots T_2 T_1$, we may also consider iterations of the form

$$x^{mI+i} = T_i x^{mI+(i-1)} - t_{mI+i} \nabla h(T_i x^{mI+(i-1)}), \tag{1.2}$$

for $m = 0, 1, ...$ and $i = 1, 2, ..., I$, in which the operators T_i and the superiorization are applied successively.

In [9] Yair Censor distinguishes *weak superiorization*, applied to feasibility-seeking algorithms, from *strong superiorization*. The theory pertaining to weak SM assumes that the set C of vectors satisfying the constraints is nonempty and $Fix(T) = C$. The theory for strong SM allows for C to be empty and considers basic algorithms that minimize a proximity function.

Date: April 28, 2019.

In both cases the SM involves applying the basic algorithm and then perturbing the iterate. The perturbations are bounded and the basic algorithm is *resilient* to such perturbations, as described in [9].

The proximity functions considered here will take the (typical) form

$$p(x) = g(d_1(x), ..., d_I(x)), \tag{1.3}$$

where $g : \mathbb{R}_+^I \to [0, +\infty]$, with $g(x_1, ..., x_I) = 0$ if and only if $x_i = 0$, for all i, and $d_i(x)$ is some measure of distance between the vector x and some projection $P_i(x)$ of x onto the ith constraint set. For example, we could have

$$p(x) = \sum_{i=1}^{I} \|x - P_i^{\perp}(x)\|^2, \tag{1.4}$$

where $P_i^{\perp}$ denotes the orthogonal projection of x onto the (closed, convex) ith constraint set. We shall consider several examples of proximity functions later in this paper.

The focus in SM is to guarantee "perturbation resilience" , which means that the limit of the modified sequence of iterates remains within the fixed-point set of T, while, one hopes, reducing the value of the function h, relative to what it would have been had the SM modifications not been used. There is an interesting connection with Isao Yamada's *hybrid steepest descent* algorithm [19], which uses the iterative step

$$x^{k+1} = Tx^k - t_k F(Tx^k) \tag{1.5}$$

to solve the variational inequality problem relative to the fixed-point set of the operator T and the monotone operator F. Here the sequence $\{t_k\}$ satisfies certain conditions, the operator T is nonexpansive, and F is Lipschitz and strongly monotone. When $F = \nabla h$ is the gradient of a convex differentiable function h, the iteration in Equation(1.5) becomes that of Equation (1.1). The difference now is that, using stricter conditions on T, h, and the sequence $\{t_k\}$, the hybrid steepest descent algorithm seeks to minimize h over the set of fixed points of T, not simply to reduce h.

2. Some Proximity Functions

As we shall show in this section, several, and perhaps all, simultaneous feasibiity-seeking iterative algorithms are also ones that minimize a proximity function, reinforcing the view that the distinction between weak and strong SM is mainly theoretical.

2.1. The Landweber and Projected Landweber Algorithms.

Let A be an M by N real matrix. The Landweber (LW) algorithm for solving $Ax = b$ is $x^{k+1} = L(x^k)$, where

$$L(x) \doteq x - \gamma A^T(b - Ax), \tag{2.1}$$

$0 < \gamma < \frac{2}{\rho(A^TA)}$ and $\rho(A^TA)$ is the largest eigenvalue of A^TA [14]. Theory tells us that the sequence $\{x^k\}$ converges to the minimizer of the function $f(x) \doteq \|b-Ax\|^2$ that minimizes the Euclidean distance $\|x-x^0\|$. Therefore, when $Ax = b$ has solutions, the limit is a solution.

The projected Landweber (PLW) algorithm has the iterative step

$$x^{k+1} = \left(x^k - \gamma A^T(b - Ax^k)\right)_+, \tag{2.2}$$

where $(x)_+$ denotes the orthogonal projection of x onto the nonnegative orthant of $\mathbb{R}^N$. The sequence $\{x^k\}$ defined by Equation (2.2) converges to the minimizer of $f(x)$ over the nonnegative orthant for which $\|x - x^0\|$ is minimized, whenever $f(x)$ has nonnegative minimizers.

The function $f(x)$ can be viewed as a proximity function. For $m = 1, ..., M$ let $H_m \doteq \{x|(Ax)_m = b_m\}$ and P_mx be the orthogonal projection of x onto H_m. Then

$$(P_mx)_n = x_n + \alpha_m^{-1}A_{m,n}(b_m - (Ax)_m), \tag{2.3}$$

where $\alpha_m \doteq \sum_{n=1}^N A_{m,n}^2$. Then it is easy to show that

$$\|b - Ax\|^2 = \sum_{m=1}^M \|x - P_mx\|^2. \tag{2.4}$$

As we shall see, minimizing proximity functions need not produce useful results.

Suppose that the PLW algorithm converges to a nonnegative vector z that minimizes the function $f(x)$ over nonnegative x, but that there is no nonnegative solution to the system $Ax = b$. Assume also that $M < N$. From

$$z = \left(z - \gamma A^T(b - Az)\right)_+$$

it follows that $(A^T(b - Az))_n = 0$ for every index n in S, defined to be the set of all indices n for which $z_n > 0$. Let S have K members and let B be the M by K matrix obtained from A by deleting the nth column of A whenever n is not in S. Then we have $B^T(b - Az) = 0$. If $M \leq K \leq N$ and B has full rank, which is typically the case, then B^T is a one-to-one transformation. Therefore $b = Az$; but this contradicts our assumption that the system has no nonnegative solutions. We conclude then that $K < M$ and that the vector z has at most $M - 1$ positive entries.

This is significant in image processing, where z denotes a vectorized image. In the hope of achieving higher resolution, one often imposes a fine grid of N pixels to account for the $M < N$ values of measured data. Often, however, the data are noisy and there is no nonnegative x consistent with the M measurements. Minimizing $f(x)$ over nonnegative x leads, as we just saw, to an image having at most $M - 1$ positive pixel values. If N is much larger than M these images can resemble stars in the night sky. This "night sky"

phenonmenon, as we shall see shortly, is not limited to the PLW algorithm [2, 3].

3. EMML and SMART

The "expectation maximization maximum likelihood" (EMML) algorithm and the "simultaneous multiplicative algebraic reconstruction technique" (SMART) both involve the cross-entropy, or Kullback–Leibler, distance [13]. For $a > 0$ and $b > 0$ the Kullback–Leibler distance $KL(a, b)$ from a to b is

$$KL(a, b) = a \log a - a \log b + b - a. \tag{3.1}$$

We call this a distance because it is always nonnegative and equals zero if and only if $a = b$. We also define $KL(0, b) = b$ and $KL(a, 0) = +\infty$. Then we extend the KL distance to nonnegative vectors x and z in $\mathbb{R}^J$ componentwise;

$$KL(x, z) = \sum_{j=1}^{J} KL(x_j, z_j). \tag{3.2}$$

Let y be a positive vector in $\mathbb{R}^I$ and $P = [P_{i,j}]$ be an I by J matrix with nonnegative entries and $\sum_{i=1}^{I} P_{i,j} = 1$, for all j. The EMML algorithm [17, 18, 2, 6] minimizes $KL(y, Px)$ over all $x \geq 0$, while the SMART [16, 11, 2, 4, 6] minimizes $KL(Px, y)$ over the same x. When $y = Px$ has nonnegative solutions and $x^0 > 0$ the sequence of SMART iterates converges to the nonnegative solution for which $KL(x, x^0)$ is minimized. The sequence of EMML iterates also converges to a nonnegative solution dependent on x^0, but nothing further about this solution is known.

Let H_i be the set of all nonnegative vectors x such that $(Px)_i = y_i$. For a given nonnegative x the vector z in H_i that minimizes $\sum_{j=1}^{J} P_{i,j} KL(x_j, z_j)$ can be found in closed form. This z, which we denote by $Q_i x$, has entries $(Q_i x)_j = x_j y_i / (Px)_i$. Using these generalized projections onto the subsets H_i we can express both $KL(y, Px)$ and $KL(Px, y)$ as proximity functions:

$$KL(y, Px) = \sum_{i=1}^{I} \left(\sum_{j=1}^{J} P_{i,j} KL((Q_i x)_j, x_j) \right), \tag{3.3}$$

and

$$KL(Px, y) = \sum_{i=1}^{I} \left(\sum_{j=1}^{J} P_{i,j} KL(x_j, (Q_i x)_j) \right). \tag{3.4}$$

The "night sky" phenomenon that we described in terms of the PLW algorithm also is a feature of both the EMML algorithm and the SMART.

3.1. **The EMML Algorithm.** The iterative step of the EMML algorithm is $x^{k+1} = M(x^k)$, where

$$M(x)_j \doteq x_j \left(\sum_{i=1}^{I} P_{i,j} y_i / (Px)_i \right). \tag{3.5}$$

Let $z \geq 0$ be the limit of the sequence $\{x^k\}$. Then for all j for which $z_j > 0$ we have

$$\sum_{i=1}^{I} P_{i,j} y_i / (Pz)_i = 1.$$

Let R be the I by K matrix obtained from P by deleting those nth columns for which $z_n = 0$. Then we have

$$R^T(y/Pz) = u,$$

where y/Pz denotes the vector with entries $y_i/(Pz)_i$ and u is the vector whose entries are all 1. If $K \geq I$ and R has full rank, which is the typical case, then R^T is a one-to-one transformation. But $R^T(u) = u$ also. Consequently, $y/Pz = u$ and $y = Pz$. If, however, the system $y = Px$ has no nonnegative solutions then $K < I$, z is the unique fixed point and z has at most $I - 1$ positive entries.

3.2. **The SMART.** The iterative step for the SMART is $x^{k+1} = S(x^k)$, where

$$S(x)_j \doteq x_j \exp \left(\sum_{i=1}^{I} P_{i,j} \log(y_i/(Px)_i) \right). \tag{3.6}$$

Let $z \geq 0$ be the limit of the sequence $\{x^k\}$. Then for all j for which $z_j > 0$ we have

$$\sum_{i=1}^{I} P_{i,j} \log(y_i/(Pz)_i) = 0.$$

Let R be the I by K matrix obtained from P by deleting those nth columns for which $z_n = 0$. Then we have

$$R^T \log(y/Pz) = 0.$$

As before, if $K \geq I$ and R has full rank, which is the typical case, then R^T is a one-to-one transformation. But $R^T(0) = 0$ also. Consequently, $y/Pz = u$ and $y = Pz$. If $y = Px$ has no nonnegative solutions, it follows that $K < I$, z is the unique fixed point and z has at most $I - 1$ positive entries.

4. Regularization

We see from this discussion of the "night sky" phenomenon that minimizing a proximity function need not lead to useful results. To avoid this phenomenon it is common to include some form of regularization. There are some interesting connections between regularization and SM, as we shall see.

4.1. **Regularizing the Landweber Algorithm.** To regularize the LW algorithm we can minimize $\|b - Ax\|^2 + \epsilon\|x - p\|^2$, for some small $\epsilon > 0$ and some p, perhaps a prior estimate of the correct answer. The regularized iteration is now

$$x^{k+1} = (1 - \alpha)L(x^k) + \alpha p, \tag{4.1}$$

for some $0 < \alpha < 1$. It is interesting to note that, with $h(x) = \frac{1}{2}\|x - p\|^2$, the SM applied to LW gives the iteration

$$x^{k+1} = (1 - t_k)L(x^k) + t_k p. \tag{4.2}$$

In Equation(4.1) the α is constant and the limit is not a fixed point of L, in contrast to Equation (4.2), where the t_k go to zero and the limit, because of perturbation resilience, is a fixed point of L.

4.2. **Regularizing the EMML Algorithm.** To regularize the EMML algorithm while also obtaining the iterates in closed form we can minimize the function $KL(y, Px) + \epsilon KL(p, x)$, for small $\epsilon > 0$ and some positive vector p, perhaps a prior estimate of the correct nonnegative x. The regularized iteration is now

$$x^{k+1} = (1 - \alpha)M(x^k) + \alpha p. \tag{4.3}$$

Once again, if we take $h(x) = \frac{1}{2}\|x - p\|^2$ and apply the SM to the EMML algorithm we get the iteration

$$x^{k+1} = (1 - t_k)M(x^k) + t_k p. \tag{4.4}$$

4.3. **Regularizing the SMART.** To regularize the SMART while also obtaining the iterates in closed form we can minimize the function $KL(Px, y) + \epsilon KL(x, p)$, for a small $\epsilon > 0$ and some positive vector p. The regularized iteration is now

$$x^{k+1} = (S(x^k))^{1-\alpha} p^\alpha. \tag{4.5}$$

We can rewrite Equation (4.5) as

$$\log x^{k+1} = (1 - \alpha)\log S(x^k) + \alpha \log p. \tag{4.6}$$

To relate this to SM we select $h(x) \doteq \frac{1}{2}\|x - \log p\|^2$. Then the logarithmic form of the SMART iteration using SM becomes

$$\log x^{k+1} = \log S(x^k) - t_k \nabla h(\log S(x^k)), \tag{4.7}$$

or

$$\log x^{k+1} = (1 - t_k)\log S(x^k) + t_k \log p, \tag{4.8}$$

which we can write as

$$x^{k+1} = (S(x^k))^{1-t_k} p^{t_k}. \tag{4.9}$$

4.4. **SM as Regularization.** In all three of the examples just presented regularization was achieved by adding a second function to the original objective function and then iteratively minimizing their sum. In each case the second function was chosen carefully so that each iterate could be obtained in closed form. In contrast, SM does not alter the original objective function. However, the formalism of the SM, as expressed in Equation (1.1), suggests a method for regularization that does not involve adding a second function, but rather, perturbing the iterates themselves. This approach would allow for more general regularization methods that still lead to closed-form iterates.

For example, if we chose to regularize the SMART using a second function $h(x)$, at each step of the iteration we would have to solve the equation

$$0 = \sum_{i=1}^{I} P_{i,j} \left(\log x_j^{k+1} - \log x_j^k - \log y_i + \log(Px^k)_i \right) + \nabla h(x^{k+1})$$

for x^{k+1}. Alternatively, we could take as the iterative step

$$x^{k+1} = S(x^k) - \alpha \nabla h(S(x^k)). \tag{4.10}$$

Now we are regularizing by trying to reduce $h(x)$ at each step, rather than by trying to minimize $KL(Px, y) + \alpha h(x)$. We are freer now to select various h, since obtaining the iterate in closed form is no longer a problem.

As an example, we could take a vector $p > 0$ and $h(x) = \frac{1}{2}\|x - p\|^2$, as we did for the LW and EMML algorithms. Then our iterate for regularized SMART is

$$x^{k+1} = (1 - \alpha) S(x^k) + \alpha p. \tag{4.11}$$

5. Some Closing Thoughts on SM

As Yair Censor points out in [9], both weak and strong SM are primarily "research directions" in the sense that by developing theory explicitly for these two cases we can better understand the effects of SM in practice. The focus of weak SM is those situations in which the nonempty set of fixed points of the operator T is exactly the nonempty set of vectors C that satisfy the finitely many constraints. It is assumed that the basic algorithm converges to a fixed point of T and that T is perturbation resilient, so that, under suitably restrictions, the sequence of perturbed iterates also converges to a fixed point of T. Whether or not the fixed point of the perturbed sequence provides a smaller value for h than does the unperturbed sequence with the same starting point is still generally an open question. It is reasonable to ask if concentrating theoretical development on weak and strong SM is going to be useful in obtaining insight into the practical problems of reconstructing from limited data. Here are some reasons for questioning this approach:

(1) Because we never perform infinitely many iterations of the basic algorithm, only theory can tell us if the iterates of the chosen basic

algorithm converge to a fixed point and that such fixed points do satisfy the constraints.

(2) Even theory may be insufficient to tell us if the set C is nonempty. For example, theory may tell us whether or not the linear system of equations $Ax = b$ (almost surely) has solutions, but not if it has nonnegative solutions.

(3) A theory that asserts that, for a given T, h and starting vector x^0, the perturbed iterates will converge to a fixed point of T with smaller h value only helps us if it also tells us what happens after finitely many iterations.

(4) As we have seen, some (all?) simultaneous iterative algorithms that are feasibility-seeking methods, that is, which seek vectors that satisfy all the constraints, are also methods for minimizing a proximity function, so the distinction between weak and strong SM is not as clear as it would seem.

(5) Weak SM can achieve its objective only when it is possible for it to achieve its objective. This obvious statement makes clear that SM is of no use when there is a unique fixed point of T, or when the iterates of the basic algorithm converge to a minimizer of h over the set $Fix(T)$, which can depend of the choice of x^0.

(6) On the other hand, even if the unperturbed iterates converge to a minimizer of h over the constraint set, it may happen that SM produces a sequence of iterates for which h is smaller, at each step, than it would be if SM were not used.

(7) The data we obtain will always include round-off errors and, most likely, some form of noise and model error. Because inverse problems are often ill-conditioned, it is usually not a good idea to seek feasibility, or even to try to minimize a proximity function, since both approaches risk instability due to overfitting to erroneous data. Regularization is usually the safest choice, and, as we have seen, bears close resemblance to SM, particularly when SM is combined with a stopping rule.

(8) The SM formalism suggests new possibilities for regularization, indicating that SM may also be studied as a method for regularization.

(9) In regularization there is always the issue of how much to use. This same issue arises in SM, in the form of deciding how small to make the function h. In an example in [8] the authors use total variation (TV) as the superiorizing function. They note that they have stopped their iteration at a point where the TV of the iterate is actually lower than that in the simulated image they are trying to recapture. As they say, in the effort to get near the simulated original, going further in reducing the TV "is unlikely to be helpful towards achieving this aim". In practice, of course, we never know the actual TV, or much else, about the true image being recovered,

so we would not know how much to reduce h. Had they gone further they might have obtained a less useful image. Knowing how much regularization to use, or, essentially equivalently, when to stop reducing h, is an issue that will certainly need to be investigated.

References

[1] Butnariu, D., Censor, Y., and Reich, S. (eds.) *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, Studies in Computational Mathematics 8. Amsterdam: Elsevier Publ., 2001.

[2] Byrne, C. (1993)"Iterative image reconstruction algorithms based on cross-entropy minimization." *IEEE Trans. Image Proc.*, **IP-2**, pp. 96–103.

[3] Byrne, C. (1995) "Erratum and addendum to 'Iterative image reconstruction algorithms based on cross-entropy minimization.'. " *IEEE Trans. Image Proc.*, **IP-4**, pp. 225–226.

[4] Byrne, C. (1996) "Iterative reconstruction algorithms based on cross-entropy minimization." in *Image Models (and their Speech Model Cousins*, S.E. Levinson and L. Shepp, (eds.), IMA Volumes in Mathematics and its Applications, Volume 80, 1–11. New York: Springer–Verlag.

[5] Byrne, C. (2004) "A unified treatment of some iterative algorithms in signal processing and image reconstruction." *Inverse Problems*, **20**, pp. 103–120.

[6] Byrne, C. (2018) "Auxiliary-function minimization algorithms." *Applied Analysis and Optimization*, **2(2)**, pp. 171–198.

[7] Byrne, C. (2014) *Iterative Optimization in Inverse Problems*. Boca Raton, FL: CRC Press.

[8] Censor, Y., Davidi, R., and Herman, G.T. (2010) "Perturbation resilience and superiorization of iterative algorithms." *Inverse Problems*, **26**, p. 065008.

[9] Censor, Y. (2015) "Weak and strong superiorization: between feasibility seeking and minimization." *Analele Stiint. Univ. Ovidius Constanta– Ser. Mat.*, **23**, pp. 141–154.

[10] Censor, Y. (2017) "Superiorization and perturbation resilience of algorithms: a continuously updated bibliography." Technical Report, Original report: June 13, 2015 contained 41 items. First revision: March 9, 2017 contains 64 items. Available on arXiv at: https://arxiv.org/abs/1506.04219v2 and on YC website.

[11] Darroch, J. and Ratcliff, D. (1972) "Generalized iterative scaling for log-linear models." *Annals of Math. Stat.*, **43**, pp. 1470–1480.

[12] Herman, G.T., Garduño, E., Davidi, R. and Censor, Y. (2012) "Superiorization: An optimization heuristic for medical physics." *Medical Physics*, **39**, pp. 5532–5546. DOI:10.1118/1.4745566.

[13] Kullback, S., and Leibler, R. (1951) "On information and sufficiency. " *Annals of Math. Stat.*, **22**, pp. 79–86.

[14] Landweber, L. (1951) "An iterative formula for Fredholm integral equations of the first kind." *Amer. J. of Math.* **73**, pp. 615–624.

[15] Rockmore, A., and Macovski, A. (1976) "A maximum likelihood approach to emission image reconstruction from projections." *IEEE Trans. Nucl. Sci.*, **NS-23**, pp. 1428–1432.

[16] Schmidlin, P. (1972) "Iterative separation of sections in tomographic scintigrams." *Nuklearmedizin*, **11**, pp. 1–16.

[17] Shepp, L., and Vardi, Y. (1982) "Maximum likelihood reconstruction for emission tomography." *IEEE Trans. Med. Imag.*, **MI-1**, pp. 113–122.

[18] Vardi, Y., Shepp, L., and Kaufman, L. (1985) "A statistical model for positron emission tomography." *J. Amer. Stat. Assoc.* **80**, pp. 8–20.

[19] Yamada, I. (2001) "The hybrid steepest descent method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings." in [1], pp. 473–504.

(C. Byrne) Department of Mathematical Sciences, University of Massachusetts Lowell, Lowell, MA, USA

E-mail address: Charles_Byrne@uml.edu

THOUGHTS ON SUPERIORIZATION II

Preprint · May 2019

CITATIONS
0

READS
64

1 author:

Charles L. Byrne
University of Massachusetts Lowell
174 PUBLICATIONS **4,861** CITATIONS

SEE PROFILE

THOUGHTS ON SUPERIORIZATION II

CHARLES L. BYRNE

Abstract. As I had hoped, my note [11] did elicit several responses from researchers familiar with the superiorization methodology (SM). As I read some of the foundational papers on SM [12, 20, 19, 13, 14], my thinking on the subject is evolving, as I hope to explain in this second note.

1. Overview

To prepare for my discussion here of the superiorization methodology (SM) I have been reading some of the foundational articles on the subject [12, 20, 19, 13, 14]. I have chosen notation somewhat different from that in the cited articles, but with which I am more comfortable. In some places my interpretation of SM differs from ones presented in these papers. I have found that the emphasis these articles place on the novelty of SM obscures important connections between SM and other approaches, such as regularization, and confuses shortcomings of particular algorithms with inadequacies of underlying mathematical problems and models.

In [20], which is devoted to describing the usefulness of SM for medical applications, it is acknowledged that the ultimate objective is to produce a helpful result, and any method should be judged accordingly. Nevertheless, the authors accept the translation of the real-world problem into a mathematical problem and choose to judge algorithms only in so far as they solve the basic mathematical problem; they do not question the suitability of the model itself. This is, of course, their choice, but I do not believe they are consistent in this.

Let me give two quotes from [19] that make clear the view of the authors. The abstract of [19] begins "A reconstructed image in positron emission tomography (PET) should be such that its likelihood, assuming a Poisson model, is high given the observed detector readings." Two sentences later, in discussing the ML-EM algorithm [23, 25, 26] (which I shall call the EMML algoroithm here), they write "An undesirable property of the algorithm is that it produces images with irregular high amplitude patterns as the number of iterations increases." I disagree with both of these statements. The first one fails to distinguish between the real-world problem of reconstructing from PET data and the mathematical problem of maximizing (or at least making high) a likelihood. The second statement is

Date: May 18, 2019.

simply false; poor reconstruction is not a property of the algorithm, but a property of the maximizers of the likelihood function itself, as I proved in [4]. Indeed, the projected Landweber algorithm and the simultaneous multiplicative algebraic reconstruction technique (SMART) [17, 24] also exhibit the same irregularities [4, 5], which also shows that it is not simply the Kullback–Leibler distance [21] that is at fault.

Let A be a real M by N matrix and $f(x) = \frac{1}{2}\|b - Ax\|^2$. The projected Landweber (PLW) algorithm [22] has the iterative step

$$x^{k+1} = \left(x^k - \gamma A^T(b - Ax^k)\right)_+, \tag{1.1}$$

where $(x)_+$ denotes the orthogonal projection of x onto the nonnegative orthant of $\mathbb{R}^N$. The sequence $\{x^k\}$ defined by Equation (1.1) converges to the minimizer of $f(x)$ over the nonnegative orthant for which $\|x - x^0\|$ is minimized, whenever $f(x)$ has nonnegative minimizers.

Suppose now that $z \geq 0$ minimizes $f(x)$ over all $x \geq 0$. Since the gradient of $f(x)$ is $\nabla f(x) = A^T(b - Ax)$ it follows that $\langle A^T(b - Az), x - z\rangle \geq 0$, for all $x \geq 0$. Therefore, we must have $(A^T(b - Az))_n = 0$ for every index n in S, defined to be the set of all indices n for which $z_n > 0$. Let S have K members and let B be the M by K matrix obtained from A by deleting the nth column of A whenever n is not in S. Then we have $B^T(b - Az) = 0$. If $M \leq K \leq N$ and B has full rank, which is typically the case, then B^T is a one-to-one transformation. Therefore $b = Az$; but this contradicts our assumption that the system has no nonnegative solutions. We conclude then that $K < M$ and that the vector z has at most $M - 1$ positive entries. This fact leads to the "irregular high amplitude patterns" observed by the authors of [19]. Clearly, it is not the fault of any algorithm, but is a property of the chosen mathematical problem, namely to find a nonnegative least-squares solution.

The authors of the cited papers clearly want to blame the particular algorithms and to show that by changing the algorithms through the perturbations of SM, as opposed to changing the model or the mathematical problem itself, one can do better. The emphasis on *resilience* throughout the cited papers makes it clear that what the basic algorithm is trying to do is good, at least in theory, but that the algorithm can do better with help from SM. They believe that there is a sharp distinction to be made between the usual approach to regularization, that is, the use of an add-on secondary function, and what SM offers. As I hope to show in this paper, the distinction may only be between incremental and simultaneous optimization.

2. Two Scenarios

The *basic problem* is a mathematical one, to minimize a function $f : \Omega \subseteq \mathbb{R}^J \to \mathbb{R}$, subject to some constraints. It is important, at this stage, to distinguish two scenarios, in preparation for the introduction of the superiorization methodology (SM):

(1) Solving the basic problem, minimizing f subject to constraints, is a good thing, in the sense that doing so will make a positive contribution to solving the real-world problem; or
(2) Solving the basic problem, although originally thought to be a good idea, is revealed to produce results that are not useful.

We shall consider examples of both of these scenarios in what follows. In the cited papers on SM the underlying assumption is that we are in the first scenario, although their example of reconstruction from projections clearly lies within the second scenario.

3. Some Notation

Superiorization is about modifying an iterative algorithm to achieve a better result without significantly increasing the effort required or deviating from the original objective for which the algorithm was designed. With $T : \Omega \subseteq \mathbb{R}^J \to \mathbb{R}^J$ the iterative step of the basic algorithm is

$$x^{k+1} = Tx^k. \tag{3.1}$$

We shall assume that the sequence $\{x^k\}$ converges to a constrained minimizer of the function f, for all starting vectors x^0. A typical iterative step of a superiorized version of the basic algorithm is

$$z^{k+1} = Tz^k + t_k v^k, \tag{3.2}$$

where $t_k > 0$ and $\sum_{k=1}^{\infty} t_k < +\infty$, and the sequence of perturbations $\{v^k\}$ is bounded. An important special case of Equation (3.2) employs $v^k \doteq -\nabla h(Tz^k)$, for some secondary differentiable function h; the iterative step is then

$$z^{k+1} = Tz^k - t_k \nabla h(Tz^k). \tag{3.3}$$

In this case the purpose of the superiorization is to solve the original mathematical problem, while also making h smaller than it would otherwise be without the perturbation. We note that it is also acceptable, in those cases in which the operator T is a product of other operators, that is, $T = \prod_{i=1}^{I} T_i$, for the perturbations to be inserted sequentially, after the application of each T_i.

The basic algorithm is said to be *perturbation resilient* if, whenever the sequence $\{x^k\}$ converges to a solution of the original problem for every starting vector x^0, so does the sequence $\{z^k\}$ [12]. In [13] the author distinguishes between *weak* SM and *strong* SM. Weak SM and the associated notion of resilience refers exclusively to what happens in the limit, after infinitely many iterations, and a solution here is understood to be an exact constrained minimizer of $f(x)$. Strong SM is concerned with the behavior of a proximity function after finitely many steps and acknowledges the significance of approximation. Because we always stop after finitely many iterations, this distinction seems unnecessary. As was shown in [11], many of the functions

$f(x)$ that concern us can also be formulated as proximity functions for some convex feasibility problem.

It is clear from the definition of perturbation resilience and the emphasis on it in the cited papers that we are supposed to assume that solving the mathematical problem is desirable and that will still be possible after the perturbation. Said another way, we like what the basic algorithm does, but we would like to do a bit better. However, the example of likelihood maximization in [19] concerns a situation in which the basic algorithm does something that, while sounding like a good idea initially, is producing an image with irregular high oscillations. The superiorization using total variation, coupled with the necessarily finite number of iterative steps, effects a smoothing of the image. Although the authors of [20] deny that they are changing the mathematical model, the effect is the same as would have been achieved through a *maximum a posteriori* (MAP) regularization. What this suggests is that, in those cases in which the basic algorithm produces an undesirable result, the superiorization provides a generalization of regularization that is simpler than, but just as effective as, a MAP method. It appears to me that the distinction between SM and regularization rests on a rather restrictive notion of regularization. I'll return to this issue later in this note.

4. Some Examples of the First Scenario

In all three of the examples in this section we are in the first scenario, that is, we like what the basic algorithm is doing, but want more.

4.1. **Bauschke's Algorithm.** Consider the problem of finding the point in the nonempty set $C = A \cap B$ that is closest to x in the Euclidean sense, where A and B are closed convex subsets of $\mathbb{R}^N$. The alternating orthogonal projection (AOP) algorithm is the following. Let $y_0 = x$. Having found y_{n-1} let

$$\begin{aligned} z_{n-1} &= P_A y_{n-1} \\ y_n &= P_B z_{n-1}. \end{aligned} \tag{4.1}$$

The sequences $\{y_n\}$ and $\{z_n\}$ both converge to the same member of C, but not necessarily to $P_C x$ [16]. Now we apply sequentially the SM formalism of Equation (3.3) with $h(z) = \frac{1}{2}\|x - z\|^2$ and $y_0 = x$. Having found y_{n-1} we take

$$\begin{aligned} z_{n-1} &= P_A y_{n-1} - t_k \nabla h(P_A y_{n-1}) \\ y_n &= P_B z_{n-1} - t_k \nabla h(P_B z_{n-1}). \end{aligned} \tag{4.2}$$

With $t_k \to 0$, $\sum_{k=1}^{\infty} t_k = \infty$, and $\sum_{k=1}^{\infty} |t_k - t_{k+1}| < +\infty$, the sequences $\{y_n\}$ and $\{z_n\}$ converge to $P_C x$. This is Bauschke's algorithm for $C = A \cap B$ [1]. It is interesting to note that, while $t_k \to 0$, we require that $\sum_{k=1}^{\infty} t_k = \infty$, in contrast to Equation (3.2). In Bauschke's algorithm there is increased emphasis on the ∇h term, which keeps the x alive longer and results in convergence to $P_C x$.

4.2. **A Similar Algorithm.** Here is a similar algorithm for finding $P_C x$, obtained using the SM formalism. Again, let $h(z) = \frac{1}{2}\|x - z\|^2$ and $y_0 = x$. Having found y_{n-1}, we take

$$\begin{aligned} z_{n-1} &= P_A y_{n-1} - \nabla h(y_{n-1}), \\ y_n &= P_B z_{n-1} - \nabla h(z_{n-1}). \end{aligned} \tag{4.3}$$

Although the reader may not recognize it, this is Dykstra's algorithm [18]. The perturbations involved are $\nabla h(y_{n-1}) = y_{n-1} - x$ and $\nabla h(z_{n-1}) = z_{n-1} - x$. The sequences $\{y_n\}$ and $\{z_n\}$ need not converge separately, indeed, they need not be bounded, in which case the perturbations are not bounded. However, the sequences $\{P_A y_n\}$ and $\{P_B z_n\}$ both converge to $P_C x$ and the sequence $\{y_n + z_n\}$ converges to $x + P_C x$. In both these algorithms we have applied the SM formalism to the alternating orthogonal projection iterative algorithm. We can do the same for the alternating Bregman projection iterative algorithm, as was shown in [8].

4.3. **Yamada's Hybrid Steepest Descent Algorithm.** There is an interesting connection between SM and Isao Yamada's *hybrid steepest descent* algorithm [27], which uses the iterative step

$$x^{k+1} = Tx^k - t_k F(Tx^k) \tag{4.4}$$

to solve the variational inequality problem relative to the fixed-point set of the operator T and the monotone operator F. Here the sequence $\{t_k\}$ satisfies certain conditions, the operator T is nonexpansive, and F is Lipschitz and strongly monotone. When $F = \nabla h$ is the gradient of a convex differentiable function h, the iteration in Equation (4.4) becomes that of Equation (3.2). The difference now is that, using stricter conditions on T, h, and the sequence $\{t_k\}$, the hybrid steepest descent algorithm seeks to minimize h over the set of fixed points of T, not simply to reduce h. Once again, we need $t_k \to 0$, but $\sum_{k=1}^{\infty} t_k = +\infty$.

5. Is SM Just Regularization?

The authors of the cited papers on SM would certainly deny this. They point out that regularization, as commonly understood, involves adding to the objective function a secondary function and then optimizing their sum, and SM does no such thing. An example will help to clarify the issue.

5.1. **The EMML Algorithm.** To investigate this point, let's take the example of the EMML algorithm for minimizing

$$KL(y, Px) \doteq \sum_{i=1}^{I} y_i \log y_i - y_i \log(Px)_i + (Px)_i - y_i, \tag{5.1}$$

where y is a positive vector, P is a nonnegative matrix with $\sum_{i=1}^{I} P_{i,j} = 1$, for all $j = 1, ..., J$, and x is a nonnegative vector. With

$$(Mx)_j \doteq x_j \sum_{i=1}^{I} P_{i,j} y_i / (Px)_i, \tag{5.2}$$

the EMML algorithm has the iterative step $x^{k+1} = Mx^k$. For any starting vector $x^0 > 0$ the sequence $\{x^k\}$ converges to a nonnegative minimizer of $KL(y, Px)$ [4, 5]. Minimizing $KL(y, Px)$ over $x \geq 0$ is equivalent to maximizing the likelihood function for the Poisson model in PET image reconstruction.

Let $z \geq 0$ be the limit of the sequence $\{x^k\}$. Then for all j for which $z_j > 0$ we have

$$\sum_{i=1}^{I} P_{i,j} y_i / (Pz)_i = 1.$$

Let R be the I by K matrix obtained from P by deleting those nth columns for which $z_n = 0$. Then we have

$$R^T(y/Pz) = u,$$

where y/Pz denotes the vector with entries $y_i/(Pz)_i$ and u is the vector whose entries are all 1. If $K \geq I$ and R has full rank, which is the typical case, then R^T is a one-to-one transformation. But $R^T(u) = u$ also. Consequently, $y/Pz = u$ and $y = Pz$. If, however, the system $y = Px$ has no nonnegative solutions then $K < I$, z is the unique fixed point and z has at most $I - 1$ positive entries [4]. Once again, we see the irregular high amplitude pattern mentioned in [19], but, as with the PLW algorithm, it is not the algorithm that is at fault.

5.2. **Regularizing the EMML Algorithm.** To regularize the EMML algorithm while also obtaining the iterates in closed form we can minimize the function $KL(y, Px) + \epsilon KL(p, x)$, for small $\epsilon > 0$ and some positive vector p, perhaps a prior estimate of the correct nonnegative x. The regularized iteration is now

$$x^{k+1} = (1 - \alpha)M(x^k) + \alpha p. \tag{5.3}$$

If we take $h(x) = \frac{1}{2}\|x - p\|^2$ and apply the SM to the EMML algorithm we get the iteration

$$x^{k+1} = (1 - t_k)M(x^k) + t_k p. \tag{5.4}$$

In fairness, let it be noted that finding a suitable second function that leads to a closed-form iterate is usually not a simple matter and that the formalism of the SM is certainly simpler. But, again in fairness, it seems that SM is a simpler way to regularize, not something quite different, or, to use the phrase in [14], an "antipodal way of thinking" .

6. Simultaneous or Incremental Optimization?

Let the operator F be such that the sequence $\{F^k x^0\}$ converges to a constrained minimizer of the function $f(x)$. Regularization, as it is commonly understood and described in the papers on SM cited here, involves replacing $f(x)$ with $f(x) + h(x)$ and minimizing this sum iteratively. Unless the $h(x)$ is carefully chosen, it can be difficult to find a function $h(x)$ and an iterative algorithm that produces closed-form iterates. Superiorization is offered as a much different and simpler way to achieve a similar effect. However, we can look at SM in a different way that reveals a much closer connection to regularization.

Suppose that there is an operator H such that the sequence $\{H^k x^0\}$ converges to a constrained minimizer of the function $h(x)$. There is probably no way to employ F and H to perform the regularization, if we insist on using these operators simultaneously. But what if we proceed incrementally [2]? Let $G \doteq H \circ F$, so that $Gx^k = H(F(x^k))$. For example, suppose that H is a gradient descent operator, that is,

$$Hx \doteq x - \gamma \nabla h(x). \tag{6.1}$$

Then we have

$$x^{k+1} = Gx^k = Fx^k - \gamma \nabla h(Fx^k). \tag{6.2}$$

Look familiar?

References

[1] Bauschke, H. (1996) "The approximation of fixed points of compositions of nonexpansive mappings in Hilbert space," *Journal of Mathematical Analysis and Applications*, **202**, pp. 150–159.

[2] Bertsekas, D.P. (1997) "A new class of incremental gradient methods for least squares problems." *SIAM J. Optim.* **7**, pp. 913–926.

[3] Butnariu, D., Censor, Y., and Reich, S. (eds.) *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, Studies in Computational Mathematics 8. Amsterdam: Elsevier Publ., 2001.

[4] Byrne, C. (1993)"Iterative image reconstruction algorithms based on cross-entropy minimization." *IEEE Trans. Image Proc.*, **IP-2**, pp. 96–103.

[5] Byrne, C. (1995) "Erratum and addendum to 'Iterative image reconstruction algorithms based on cross-entropy minimization.'. " *IEEE Trans. Image Proc.*, **IP-4**, pp. 225–226.

[6] Byrne, C. (1996) "Iterative reconstruction algorithms based on cross-entropy minimization." in *Image Models (and their Speech Model Cousins*, S.E. Levinson and L. Shepp, (eds.), IMA Volumes in Mathematics and its Applications, Volume 80, 1–11. New York: Springer–Verlag.

[7] Byrne, C. (2004) "A unified treatment of some iterative algorithms in signal processing and image reconstruction." *Inverse Problems*, **20**, pp. 103–120.

[8] Byrne, C. (2017) "The Dykstra and Bregman–Dykstra algorithms as superiorization–revised version (Dec. 26, 2017)" , *ResearchGate*.

[9] Byrne, C. (2018) "Auxiliary-function minimization algorithms." *Applied Analysis and Optimization*, **2(2)**, pp. 171–198.

[10] Byrne, C. (2014) *Iterative Optimization in Inverse Problems*. Boca Raton, FL: CRC Press.
[11] Byrne, C. (2019) "Thoughts on Superiorization." , *ResearchGate*, April 28, 2019.
[12] Censor, Y., Davidi, R., and Herman, G.T. (2010) "Perturbation resilience and superiorization of iterative algorithms." *Inverse Problems*, **26**, p. 065008.
[13] Censor, Y. (2015) "Weak and strong superiorization: between feasibility seeking and minimization." *Analele Stiint. Univ. Ovidius Constanta– Ser. Mat.*, **23**, pp. 141–154.
[14] Censor, Y., Herman, G.T., and Jiang, M. (2017) "Superiorization: theory and applications." Preface to *Inverse Problems*, **33**.
[15] Censor, Y. (2017) "Superiorization and perturbation resilience of algorithms: a continuously updated bibliography." Technical Report, Original report: June 13, 2015 contained 41 items. First revision: March 9, 2017 contains 64 items. Available on arXiv at: https://arxiv.org/abs/1506.04219v2 and on YC website.
[16] Cheney, W., and Goldstein, A. (1959) "Proximity maps for convex sets." *Proc. Am. Math. Soc.*, **10**, pp. 448–450.
[17] Darroch, J. and Ratcliff, D. (1972) "Generalized iterative scaling for log-linear models." *Annals of Math. Stat.*, **43**, pp. 1470–1480.
[18] Dykstra, R. (1983) "An algorithm for restricted least squares regression." *J. Amer. Statist. Assoc.*, **78 (384)**, pp. 837–842.
[19] Garduño, E. and Herman, G.T. (2014) "Superiorization of the ML–EM algorithm." *IEEE Trans. Nucl. Sci.*, **61**, pp. 162–172.
[20] Herman, G.T., Garduño, E., Davidi, R. and Censor, Y. (2012) "Superiorization: An optimization heuristic for medical physics." *Medical Physics*, **39**, pp. 5532–5546. DOI:10.1118/1.4745566.
[21] Kullback, S., and Leibler, R. (1951) "On information and sufficiency. " *Annals of Math. Stat.*, **22**, pp. 79–86.
[22] Landweber, L. (1951) "An iterative formula for Fredholm integral equations of the first kind." *Amer. J. of Math.* **73**, pp. 615–624.
[23] Rockmore, A., and Macovski, A. (1976) "A maximum likelihood approach to emission image reconstruction from projections." *IEEE Trans. Nucl. Sci.*, **NS-23**, pp. 1428–1432.
[24] Schmidlin, P. (1972) "Iterative separation of sections in tomographic scintigrams." *Nuklearmedizin*, **11**, pp. 1–16.
[25] Shepp, L., and Vardi, Y. (1982) "Maximum likelihood reconstruction for emission tomography." *IEEE Trans. Med. Imag.*, **MI-1**, pp. 113–122.
[26] Vardi, Y., Shepp, L., and Kaufman, L. (1985) "A statistical model for positron emission tomography." *J. Amer. Stat. Assoc.* **80**, pp. 8–20.
[27] Yamada, I. (2001) "The hybrid steepest descent method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings." in [3], pp. 473–504.

(C. Byrne) Department of Mathematical Sciences, University of Massachusetts Lowell, Lowell, MA, USA

E-mail address: Charles_Byrne@uml.edu

THOUGHTS ON SUPERIORIZATION III

CHARLES L. BYRNE

Abstract. The term "proximity function" is used in several of the foundational papers on the superiorization method (SM) in two distinct ways, leading to some confusion and complication in defining resilience. In this note I suggest a modification of the framework for SM and illustrate this modification by applying it to the convex feasibility problem (CFP) and likelihood maximization.

1. My Modified Framework for SM

Let $\mathbb{P}$ denote some mathematical problem whose potential solutions are members of $\mathbb{R}^J$. Let $d : \mathbb{R}^J \to \mathbb{R}_+$ be continuous, with $d(x)$ measuring the distance a vector x is from solving $\mathbb{P}$, such that x is a solution of $\mathbb{P}$ if and only if $d(x) = 0$. Let $T : \Omega \subseteq \mathbb{R}^J \to \Omega$ be an operator and $x^{k+1} = Tx^k$ the iterative step of the *basic algorithm*. For suitable $t_k > 0$ and $v^k \in \mathbb{R}^J$ the *perturbed sequence* is defined by $z^{k+1} = Tz^k + t_k v^k$. We say that T is *perturbation resilient* (pr) if $\{x^k\}$ converges to a solution of $\mathbb{P}$, for each starting vector x^0, and $\{z^k\}$ also converges to a solution of $\mathbb{P}$, for each z^0 and suitable $t_k > 0$ and $v^k \in \mathbb{R}^J$.

In several of the foundational papers on SM [5, 10, 9, 6, 7] the function that I call $d(x)$ is termed *the proximity function*. This is an unfortunate choice, because this term already has a well understood meaning within the context of the convex feasibility problem (CFP). When the CFP is used as an example in these papers, the same function is used as the objective function to be minimized and as the $d(x)$ function that measures how well x does as a potential solution to the minimization problem. Not only is this confusing, but it complicates the discussion of resilience. Let me give several examples to illustrate the modifications I am suggesting.

2. The Convex Feasibility Problem

Let $C_i, i = 1, ..., I$, be closed convex subsets of $\mathbb{R}^J$. The *convex feasibility problem* is to find a member of $C \doteq \cap_{i=1}^I C_i$. We consider two cases: 1) C is not empty; and 2) C is empty.

2.1. C is not empty.

The problem $\mathbb{P}$ is to find a member of C. As our iterative algorithm proceeds we need a way to measure how we are doing. As a measure $d(x)$ of how far x is from solving $\mathbb{P}$ we have several choices.

Date: May 31, 2019.

We let $P_i x$ denote the orthogonal projection of x onto the set C_i. Some of the choices for $d(x)$ are

$$d(x) = \sum_{i=1}^{I} \|x - P_i x\|; \tag{2.1}$$

$$d(x) = \max\{\|x - P_i x\|, i = 1, ..., I\}; \tag{2.2}$$

$$d(x) = \frac{1}{2} \sum_{i=1}^{I} \|x - P_i x\|^2. \tag{2.3}$$

It is often the case in practice that we do not know if the intersection $C \doteq \cap_{i=1}^{I} C_i$ is nonempty; therefore, it makes sense to broaden the problem and to seek a minimizer of a proximity function $F(x)$.

2.2. C **is empty.** Because C is empty, we define our problem $\mathbb{P}$ to be an optimization problem, to minimize an objective function $F(x)$. For the purpose of this discussion we take as our objective function

$$F(x) \doteq \frac{1}{2} \sum_{i=1}^{I} \|x - P_i x\|^2. \tag{2.4}$$

The gradient of $F(x)$ is

$$\nabla F(x) = \sum_{i=1}^{I} x - P_i x, \tag{2.5}$$

and $\hat{x}$ minimizes $F(x)$ if and only if $\nabla F(\hat{x}) = 0$, or, equivalently

$$\hat{x} = \frac{1}{I} \sum_{i=1}^{I} P_i \hat{x}. \tag{2.6}$$

As a measure of progress, of how well x does as a minimizer of $F(x)$, we define

$$d(x) \doteq \|x - \frac{1}{I} \sum_{i=1}^{I} P_i x\|^2. \tag{2.7}$$

We then can say that a vector x is a solution of the original problem $\mathbb{P}$ if and only if $d(x) = 0$. In the cited papers no distinction is made between the objective function $F(x)$ and the function $d(x)$; $F(x)$ is used in both roles, making it difficult to discuss resilience.

3. Likelihood Maximization

In [9] SM is illustrated by taking as the problem $\mathbb{P}$ the maximizing of the likelihood function, assuming the Poisson model for PET data. With y a positive vector whose entries are the measured PET data values, P the system matrix, with $\sum_{i=1}^{I} P_{i,j} = 1$, for all j, and $x \geq 0$ a potential image, maximizing the likelihood is equivalent to minimizing the Kullback–Leibler distance from y to Px, given by

$$KL(y, Px) \doteq \sum_{i=1}^{I} y_i \log y_i - y_i \log(Px)_i + (Px)_i - y_i. \tag{3.1}$$

It was shown in [4] that $KL(y, Px)$ can be rewritten as a proximity function for a CFP. Although $KL(y, Px)$ can be minimized using a variety of iterative algorithms, the authors here choose the ML-EM algorithm, which I have called the EMML algorithm. With

$$(Mx)_j \doteq x_j \sum_{i=1}^{I} P_{i,j} y_i / (Px)_i, \tag{3.2}$$

the EMML algorithm has the iterative step $x^{k+1} = Mx^k$. It is known that, for any starting vector $x^0 > 0$, the sequence $\{x^k\}$ converges to a nonnegative minimizer of $KL(y, Px)$ [1, 2]. It was also shown there that the sequence $\{x^k\}$ is Fejér monotone with respect to the set of all nonnegative minimizers of $KL(y, Px)$, that is, if $\hat{x}$ is a nonnegative minimizer of $KL(y, Px)$, then the sequence $KL(\hat{x}, x^k)$ is decreasing.

Because $\mathbb{P}$ is a minimization problem our function $d(x)$ must measure how far x is from minimizing $KL(y, Px)$. If x minimizes $KL(y, Px)$ then

$$0 = x_j \left(1 - \sum_{i=1}^{I} P_{i,j} y_i / (Px)_i \right), \tag{3.3}$$

for all j. Therefore, we take as the function $d(x)$

$$d(x) = \sum_{j=1}^{J} x_j^2 \left(1 - \sum_{i=1}^{I} P_{i,j} y_i / (Px)_i \right)^2. \tag{3.4}$$

Then $\hat{x}$ is a minimizer of $KL(y, Px)$ if and only if $d(\hat{x}) = 0$, even if $KL(y, P\hat{x})$ is not zero.

4. Two Opposing Views

In [9] the EMML algorithm and SM are studied in the context of reconstruction from PET data. It seems to me that there are two opposing views that one can adopt now.

(1) An image x that maximizes the likelihood will provide a useful reconstruction. Because the EMML algorithm is slow to converge, we would have to take a large number of iterations before we would

obtain a reasonably close approximation of the maximum-likelihood image. Unfortunately, the images produced by the earlier iterates exhibit irregular, high-amplitude oscillations. If we were able to iterate much further, these oscillations would begin to disappear, giving us a decent approximation of the maximum-likelihood image. To overcome this annoying behavior of the EMML algorithm, we apply SM; in the example in the cited papers the authors use total variation. The effect of using SM is the same as if we had been able to iterate much longer; the irregular high-ampitude oscillations have been significantly reduced.

(2) The image produced by a maximizer of the likelihood function may exhibit irregular high-amplitude oscillations and any iterative algorithm that converges to a likelihood maximizer will exhibit similar oscillations as the iterates approach the limit; remember that the sequence is Fejér monotone. The EMML algorithm is not at fault; likelihood maximization is simply a poor choice in some cases.

I believe it is clear that the authors of [5, 10, 9, 6, 7] hold the first view, while I hold the second. In fact, there is theory to back me up on this.

Let $z \geq 0$ be the limit of the sequence $\{x^k\}$. Then for all j for which $z_j > 0$ we have

$$\sum_{i=1}^{I} P_{i,j} y_i / (Pz)_i = 1.$$

Let R be the I by K matrix obtained from P by deleting those nth columns for which $z_n = 0$. Then we have

$$R^T(y/Pz) = u,$$

where y/Pz denotes the vector with entries $y_i/(Pz)_i$ and u is the vector whose entries are all 1. If $K \geq I$ and R has full rank, which is the typical case, then R^T is a one-to-one transformation. But $R^T(u) = u$ also. Consequently, $y/Pz = u$ and $y = Pz$. If, however, the system $y = Px$ has no nonnegative solutions then $K < I$, z is the unique fixed point and z has at most $I - 1$ positive entries [1]. The image then exhibits the irregular high amplitude pattern mentioned in [9], but it is not the algorithm that is at fault. Because of measurement and model error, both hard to avoid, it is often the case that $y = Px$ will have no exact nonnegative solutions. The matrix R will typically have full rank. In [9] $I = 30,300$, while $J = 225,625$, which says that most of the pixels of the likelihood-maximizing image will be zero. As I showed in [4], irregular high-amplitude oscillations are not restricted to the minimizers of $KL(y, Px)$; the limit of the SMART and of the projected Landweber algorithm exhibit similar behavior in some cases.

5. Summary

Throughout the discussion of SM in the cited papers there are several assumptions:

(1) solving the problem $\mathbb{P}$ will be useful in solving the original real-world problem; maximizing likelihood will give us a useful PET reconstructed image;
(2) the basic iterative algorithmic sequence $\{x^k\}$ will eventually converge to this useful solution of $\mathbb{P}$;
(3) the basic algorithm may exhibit *features* that make the iterates obtained early in the iteration not useful; but
(4) using SM we can leap-frog these early unpleasant iterates and obtain iterates that more closely resemble those we would have obtained, had we been able to iterate the basic algorithm longer.

In short, the difficulty is always with the iterative algorithm, not with the choice of $\mathbb{P}$ to be solved. The difficulty can be overcome using SM, and we never change the $\mathbb{P}$. I disagree with these assumptions. As we just saw in discussing the EMML algorithm, there are cases in which the difficulty lies with the choice of $\mathbb{P}$, not with the algorithm. The observed benefits of using SM come from the implicit regularization (and therefore the alteration of $\mathbb{P}$) achieved by the SM perturbations.

References

[1] Byrne, C. (1993)"Iterative image reconstruction algorithms based on cross-entropy minimization." *IEEE Trans. Image Proc.*, **IP-2**, pp. 96–103.
[2] Byrne, C. (1995) "Erratum and addendum to 'Iterative image reconstruction algorithms based on cross-entropy minimization.'. " *IEEE Trans. Image Proc.*, **IP-4**, pp. 225–226.
[3] Byrne, C. (2019) "Thoughts on Superiorization." *ResearchGate*, April 28, 2019.
[4] Byrne, C. (2019) "Thoughts on Superiorization II." *ResearchGate*, May 18, 2019.
[5] Censor, Y., Davidi, R., and Herman, G.T. (2010) "Perturbation resilience and superiorization of iterative algorithms." *Inverse Problems*, **26**, p. 065008.
[6] Censor, Y. (2015) "Weak and strong superiorization: between feasibility seeking and minimization." *Analele Stiint. Univ. Ovidius Constanta– Ser. Mat.*, **23**, pp. 141–154.
[7] Censor, Y., Herman, G.T., and Jiang, M. (2017) "Superiorization: theory and applications." Preface to *Inverse Problems*, **33**.
[8] Censor, Y. (2017) "Superiorization and perturbation resilience of algorithms: a continuously updated bibliography." Technical Report, Original report: June 13, 2015 contained 41 items. First revision: March 9, 2017 contains 64 items. Available on arXiv at: https://arxiv.org/abs/1506.04219v2 and on YC website.
[9] Garduño, E. and Herman, G.T. (2014) "Superiorization of the ML–EM algorithm." *IEEE Trans. Nucl. Sci.*, **61**, pp. 162–172.
[10] Herman, G.T., Garduño, E., Davidi, R. and Censor, Y. (2012) "Superiorization: An optimization heuristic for medical physics." *Medical Physics*, **39**, pp. 5532–5546. DOI:10.1118/1.4745566.

(C. Byrne) Department of Mathematical Sciences, University of Massachusetts Lowell, Lowell, MA, USA

E-mail address: Charles_Byrne@uml.edu

THOUGHTS ON SUPERIORIZATION IV

Preprint · June 2019

CITATIONS
0

READS
19

1 author:

Charles L. Byrne
University of Massachusetts Lowell
174 PUBLICATIONS **4,861** CITATIONS

SEE PROFILE

THOUGHTS ON SUPERIORIZATION IV

CHARLES L. BYRNE

Abstract. The connection between the superiorization mmethod and regularization is further clarified using forward-backward splitting.

1. Can Superiorization Do Regularization?

Although there are some theorems and proofs in the foundational papers on the superiorization method (SM) [5, 12, 11, 6, 7] the authors are clear that they intend the SM to be taken heuristically, at least at this early stage of development. In previous notes [2, 3, 4] I pointed out that it seems worthwhile, in some cases, to view the effects of SM as regularization of an unwisely chosen problem. This is particularly the case in the popular example of reconstructing an image from PET data using likelihood maximization. My suggestions there were also intended to be taken heuristically; I presented some examples, but no theorems or proofs. In this note I attempt to include a bit more mathematical rigor.

In rebuttal to my suggestion to view SM as a type of regularization, some have defended the distinction between SM and regularization by pointing out that the traditional view of regularization is that the original objective function is changed by adding a second function and then optimizing the sum, and that the SM approach does not appear to be doing this. In response, I claimed that perhaps a more general notion of regularization was at work here or the optimization of a sum was being achieved incrementally, rather than simultaneously. In this note I make this more explicit, showing that the *forward-backward splitting* (FBS) algorithm [10], an incremental method for optimizing the sum of two functions, and thereby for doing regularization, leads directly to an iterative step identical to that of SM.

In formulating the SM it is assumed that there is a mathematical problem $\mathbb{P}$ to be solved and that the basic sequence $x^{k+1} = Tx^k$ generated by the operator T converges to a solution of $\mathbb{P}$. For the purpose of this discussion I shall take as $\mathbb{P}$ the minimizing of a convex differentiable function f. The goal of SM is to perturb the iterates so as to maintain convergence to a solution of $\mathbb{P}$ while reducing, if possible, the value of a secondary function h. The iterative step of the SM is

$$z^{k+1} = Tz^k + t_k v^k. \tag{1.1}$$

Date: June 5, 2019.

With $v^k = -\nabla h(x^k)$ we have

$$z^{k+1} = Tz^k - t_k \nabla h(Tx^k). \tag{1.2}$$

Equivalently, we have

$$w^{k+1} = T\left(w^k - t_k \nabla h(w^k)\right), \tag{1.3}$$

for $w^k \doteq Tz^k$.

The cited papers on SM emphasize resilience, which is achieved by requiring that the sequence $\{t_k\}$ of parameters converge to zero and be summable. For resilient operators T the perturbed sequence will still converge to a solution of $\mathbb{P}$. The implication is then that other iterative algorithms that do not require both of these conditions on the $\{t_k\}$ are not truly SM, even though they may have identical iterative steps. While such distinctions are important in theory, they have no significance in practice. Because we never iterate to infinity, we detect no difference between sequences that converge to zero and those that do not, nor between summable sequences and divergent ones. When I point out connections between SM and regularization I am speaking only about the practical effects we observe after finitely many iterations. Differences that show up only at the limits are irrelevant in practice. There is no practical difference between parameters that do not vary with k and those that may vary.

It is my contention that, at least in some cases, the practical effect of using SM is the same as that of doing regularization of an ill-chosen $\mathbb{P}$. Since regularization of $\mathbb{P}$ here would involve the selection of a secondary function h and the minimization of the sum $f + h$, it is helpful to consider situations in which iterates of the SM type do converge to such a minimizer. For that purpose we turn to the *forward-backward splitting* (FBS) method [10].

2. Forward-Backward Splitting

Let $F : \mathbb{R}^J \to \mathbb{R}$ be convex, but not necessarily differentiable. For each $x \in \mathbb{R}^J$ we say that $z = prox_F x$ if $y = z$ is the unique minimizer of the function $F(y) + \frac{1}{2}\|y - x\|^2$. Then we have $0 \in \partial F(z) + z - x$, or

$$x \in z + \partial F(z), \tag{2.1}$$

where $\partial F(z)$ denotes the subdifferential of F at z. We note that the inclusion in Equation (2.1) characterizes z as $prox_F x$. As an example consider the function $F(x) = \frac{1}{2}\|x - p\|^2$; then $z = prox_F x$ if and only if $y = z$ minimizes $\frac{1}{2}\|y - p\|^2 + \frac{1}{2}\|y - x\|^2$. It follows that $z = prox_F x = \frac{1}{2}(x + p)$.

In [1] it was shown that the sequence generated by $x^{k+1} = Tx^k$, for $T = prox_F$, converges to a minimizer of the function F, if minimizers exist. We turn now to the problem of minimizing $F(x) + g(x)$, where F is convex, but not necessarily differentiable, and $g : \mathbb{R}^J \to \mathbb{R}$ is convex, differentiable, and

has gradient ∇g that is L-Lipschitz continuous, that is, $\|\nabla g(a) - \nabla g(b)\| \leq L\|a - b\|$, for all a and b.

Let $0 < \gamma < \frac{1}{L}$. As shown in [1], we can get the iterate x^{k+1} of the FBS method by minimizing the function

$$G_k(x) = F(x) + g(x) + \frac{1}{2\gamma}\|x - x^k\|^2 - D_g(x, x^k), \tag{2.2}$$

where

$$D_g(x, y) \doteq g(x) - g(y) - \langle \nabla g(y), x - y\rangle$$

is the Bregman distance. Then we have

$$0 \in \partial\gamma F(x^{k+1}) + x^{k+1} - x^k + \nabla\gamma g(x^k)$$

or

$$x^k - \gamma\nabla g(x^k) \in x^{k+1} + \partial\gamma F(x^{k+1}.$$

It follows that

$$x^{k+1} = prox_{\gamma F}\left(x^k - \gamma\nabla g(x^k)\right). \tag{2.3}$$

It was shown in [10, 1] that the sequence $\{x^k\}$ converges to a minimizer of $F(x) + g(x)$ whenever this function has minimizers. Using the Baillon–Haddad Theorem and Krasnosel'skii-Mann iteration, Combettes and Wajs show that γ can be selected in the interval $(0, \frac{2}{L})$ [10].

Comparing Equation (1.3) with Equation (2.3) we see that, at least in some cases, SM can be used to minimize the sum of two functions, and therefore to do regularization. To make the comparison we would take F to be the primary function f, g to be the secondary function h and $T = prox_f$.

3. Using SM for Regularization

In most of the optimization problems to which we would apply the SM the primary objective function f is typically more complicated that the secondary one h, since the primary one is dictated largely by the real-world problem, while the secondary function is chosen so as not to further complicate the calculations. In *maximum a posteriori* (MAP) methods the secondary function h is the logarithm of the probability density function attributed to the parameter vector x. It is helpful, then, that the *prox* operator be associated with the secondary function h, which can be conveniently chosen by the user. In [9] we find a number of examples in which the *prox* function is given explicitly for popular log probability density functions. In Equation (1.3) we assume that $x^{k+1} = Tx^k$ minimizes the primary function f, while h is the secondary function. But this is not essential if we are to use SM to do regularization. Even when the operator T in the SM iteration is not a *prox* operator, we can still view SM as providing regularization, at least in a heuristic sense.

4. Summary

To help make the case that SM can, at times, be viewed as regularization, it is helpful to have examples of incremental methods that succeed in minimizing the sum of two function. The FBS method provides just such examples. The iterative optimization in the FBS is incremental, rather than simultaneous, and as such resembles closely the iteration in SM. When the FBS algorithm is used to perform regularization the iterative step is essentially that of the SM. The appearance of t_k in one and γ in the other is not significant, since what happens after finitely many iterations is all that matters in practice. Although the typical problem to which we would apply the SM may not fit precisely into the framework of the FBS, that is, T need not be either $prox_f$ or $prox_h$, the existence of a convergent SM-type incremental optimization method does strengthen the idea that SM can be viewed as regularization. Because the foundational papers on SM clearly invite us to view SM in heuristic terms, the failure, at times, of SM to fall into the FBS framework is not a serious impediment to considering SM in terms of regularization.

References

[1] Byrne, C. (2018) "Auxiliary-function minimization algorithms." *Applied Analysis and Optimization*, **2(2)**, pp. 171–198.

[2] Byrne, C. (2019) "Thoughts on Superiorization." *ResearchGate*, April 28, 2019.

[3] Byrne, C. (2019) "Thoughts on Superiorization II." *ResearchGate*, May 18, 2019.

[4] Byrne, C. (2019) "Thoughts on Superiorization III." *ResearchGate*, May 30, 2019.

[5] Censor, Y., Davidi, R., and Herman, G.T. (2010) "Perturbation resilience and superiorization of iterative algorithms." *Inverse Problems*, **26**, p. 065008.

[6] Censor, Y. (2015) "Weak and strong superiorization: between feasibility seeking and minimization." *Analele Stiint. Univ. Ovidius Constanta– Ser. Mat.*, **23**, pp. 141–154.

[7] Censor, Y., Herman, G.T., and Jiang, M. (2017) "Superiorization: theory and applications." Preface to *Inverse Problems*, **33**.

[8] Censor, Y. (2017) "Superiorization and perturbation resilience of algorithms: a continuously updated bibliography." Technical Report, Original report: June 13, 2015 contained 41 items. First revision: March 9, 2017 contains 64 items. Available on arXiv at: https://arxiv.org/abs/1506.04219v2 and on YC website.

[9] Chaux, C., Combettes, P., Pesquet, J–C., and Wajs, V. (2005) "A variational formulation for frame-based inverse problems." *Inverse Problems*, **23(4)**, 1495.

[10] Combettes, P. and Wajs, V. (2005) "Signal recovery by proximal forward-backward splitting." *Multiscale Modeling and Simulation*, **4(4)**, pp. 1168–1200.

[11] Garduño, E. and Herman, G.T. (2014) "Superiorization of the ML–EM algorithm." *IEEE Trans. Nucl. Sci.*, **61**, pp. 162–172.

[12] Herman, G.T., Garduño, E., Davidi, R. and Censor, Y. (2012) "Superiorization: An optimization heuristic for medical physics." *Medical Physics*, **39**, pp. 5532–5546. DOI:10.1118/1.4745566.

(C. Byrne) Department of Mathematical Sciences, University of Massachusetts Lowell, Lowell, MA, USA

E-mail address: Charles_Byrne@uml.edu

THOUGHTS ON SUPERIORIZATION V

CHARLES L. BYRNE

Abstract. In this note I discuss my evolving understanding of superiorization by posing, and trying to answer, several questions.

1. What is Going On Here?

The problem of reconstructing an image from PET data is discussed in three of the foundational papers on the superiorization method (SM) [9, 19, 18]. There is no question that the improvement exhibited in the reconstructions from simulated data is impressive. What is in question is just what is actually being illustrated by these examples.

The reconstruction problem is translated into the mathematical problem of maximizing the likelihood, given the Poisson model [21, 23, 24], or, equivalently, minimizing the Kullback–Leibler distance $KL(y, Px)$ over $x \geq 0$ [7]. The iterative algorithm is what I call the EMML algorithm [3, 4, 5]. Because the EMML iterates are automatically positive, for any positive starting vector, there is no need for additional constraints; the problem is an unconstrained minimization problem. At first glance this problem does not seem to fit into the SM framework; we are not trying to minimize a function f over the minimizers of $KL(y, Px)$. However, as the iteration proceeds it becomes evident that the constructed images exhibit random high-frequency oscillations and are therefore useless. To improve the images one can attempt to reduce the total variation function, denoted by f, over the minimizers of $h(x) = KL(y, Px)$. In the simulations given in these articles the SM is used, with perturbations coming from the total-variation functon f.

We are told that these examples demonstrate the usefulness of the SM approach, but that is not really the case. As I pointed out in [8], in most cases the maximizer of likelihood will be unique and will have many randomly placed zero pixel values. Any iterative algorithm that converges to the likelihood maximizer will begin to exhibit random high-frequency oscillations as the iterative sequence approaches the limit. It is not the EMML algorithm that is at fault, but the chosen problem itself, that of maximizing likelihood. The improvements made in the examples come from the smoothing effect of the perturbation term, in effect a form of regularization at work. Because we only see the results after finitely many iterations of an algorithm, we can never be sure which algorithm we are watching. Two

Date: June 21, 2019.

methods that have identical iterative steps and differ only in what happens to the parameters at the limit of infinitely many iterations cannot be distinguished after finitely many iterations. Bounded perturbation resilence in SM requires that $\sum_{k=1}^{\infty} t_k < \infty$, while algorithms discussed by Bauschke [2], Yamada [25] and Andersen and Hansen [1] present algorithms whose iterative steps are identical to that of SM, but achieve different objectives by requiring that the sum must diverge. Clearly, we can never decide, looking at any finite number of iterates, which algorithm is being used.

With the matrix P normalized so that $\sum_{i=1}^{I} P_{i,j} = 1$, for all j, the iterative step of the EMML algorithm is $x^{k+1} = Mx^k$, where

$$(Mx)_j = x_j \sum_{i=1}^{I} P_{i,j} y_i / (Px)_i. \tag{1.1}$$

The SM is applied in [18] using the iterative step

$$z^{k+1} = Mz^k + t_k v^k, \tag{1.2}$$

where the v^k are nonascending directions for the total variation. We can regularize the EMML using the iterative step

$$z^{k+1} = (1-\alpha)Mz^k + \alpha p = Mz^k - \alpha \nabla f(Mz^k), \tag{1.3}$$

for $f(x) = \frac{1}{2}\|x - p\|^2$, p some positive vector and $\alpha = \frac{\epsilon}{1+\epsilon}$, for some $\epsilon > 0$. The sequence $\{z^k\}$ converges to the minimizer of the function $KL(y, Px) + \epsilon KL(p, x)$. If the v^k in Equation (1.2) is $v^k = -\nabla f(Mz^k)$ the only difference between SM and the regularization in Equation (1.3) is what happens to the sequence $\{t_k\}$ in the limit. If $t_k = \alpha$ for all k we get regularization, while, if $t_k \to 0$ and $\sum_{k=1}^{\infty} t_k < +\infty$ and the EMML is resilient, the limit minimizes $KL(y, Px)$.

My point here is that SM is used in this example as a convenient form of regularization. Generally speaking, regularization of a minimization problem involves adding a second function and minimizing the sum. Finding a second function such that the iterates can be obtained easily is not itself an easy problem. The choice of $KL(p, x)$ above is particularly convenient and leads to the closed form for the iterate in Equation (1.3). The choice of $KL(x, p)$ would not lead to a closed-form iterate. When SM performs regularization there is no need to worry about this; we simply perturb each iterate and not require that some sum of two functions is being minimized.

2. What is SM?

A number of articles on the superiorization method (SM) have a section that asks this question. I see no reason why this note should be any different. Here is my version of the answer.

We have a real-world problem that we have translated into a mathematical problem that we denote by $\mathbb{P}$. Potential solutions of the mathematical problem are vectors in $\mathbb{R}^J$. There is a discrepancy function $d : \mathbb{R}^J \to [0, +\infty]$

such that $d(x)$ measures how far x is from solving $\mathbb{P}$; x is a solution of $\mathbb{P}$ if and only if $d(x) = 0$. The (possibly empty) solution set S consists of all x with $d(x) = 0$. There is a convenient basic iterative algorithm for solving $\mathbb{P}$, taking the form $x^{k+1} = Tx^k$. It is assumed that the sequence $\{x^k\}$ converges to some solution $\hat{x}$ of $\mathbb{P}$, whenever $\mathbb{P}$ has solutions, and $\{d(x^k)\} \to 0$ in any case. Superiorization involves modifying the basic iterative algorithm at each step to obtain a perturbed sequence

$$z^{k+1} = Tz^k + t_k v^k, \tag{2.1}$$

or, equivalently,

$$w^{k+1} = T\left(w^k + t_k v^k\right), \tag{2.2}$$

that will improve the value of some function f while still having $\{d(z^k)\} \to 0$. Operators T for which $\{d(z^k)\} \to 0$ for suitable t_k and v^k are said to be resilient to such perturbations.

Constrained minimization (CM) is one of the main applications that motivate the study of SM. Suppose that we want to minimize $f : \mathbb{R}^J \to \mathbb{R}$ over all $x \in C$, where C is some nonempty subset of $\mathbb{R}^J$. The forward-backward splitting (FBS) algorithm [15] uses the iterative step

$$x^{k+1} = P_C\left(x^k - \gamma \nabla f(x^k)\right). \tag{2.3}$$

The sequence $\{x^k\}$ converges to a solution if f is convex and differentiable, ∇f is L-Lipschitz continuous, and $0 < \gamma < \frac{2}{L}$. Of course, the hard part will usually be the implementation of P_C, the orthogonal projection onto C. The basic idea of SM is to begin by focusing on the constraint set C and worrying about f later. The basic problem $\mathbb{P}$ is then to find a convenient iterative method $x^{k+1} = Tx^k$ so that $\{x^k\}$ converges to a member $\hat{x}$ of C. Using T, we form a perturbed sequence, as given by Equation (2.1), where v^k is a nonascending direction for f, that is, $f(z^{k+1}) \leq f(Tz^k)$. The hope is that, under the right conditions on T, the t_k and the v^k, the sequence $\{z^k\}$ will converge to a member $\hat{z}$ in C and $f(\hat{z}) < f(\hat{x})$. It is not expected that $\hat{z}$ will actually minimize f over C.

In [10] the authors present two ways in which the constraint set C might be described for CM:

(1) The set C is the nonempty intersection of finitely many closed convex subsets C_i of $\mathbb{R}^J$ and the problem $\mathbb{P}$ is now the *convex feasibility problem* (CFP). The solution set S is now C itself.
(2) The problem $\mathbb{P}$ is to minimize a function $h : \mathbb{R}^J \to \mathbb{R}$, and the solution set S is the nonempty set of all y such that $h(y) \leq h(x)$, for all $x \in \mathbb{R}^J$.

In those cases in which the problem $\mathbb{P}$ is to minimize some function h, it is important not to confuse d with h. For example, suppose that the intersection of the closed convex sets $C_i, i = 1, ..., I$, is empty. It is reasonable, then,

to redefine $\mathbb{P}$ as minimizing

$$h(x) = \sum_{i=1}^{I} \|x - P_i x\|^2, \tag{2.4}$$

where P_i is the orthogonal projection onto C_i. Now we have turned a CFP for which C is empty into a function-minimization problem for which the solution set S need not be empty.

If the problem $\mathbb{P}$ is to minimize h in Equation (2.4) and $\hat{x}$ is in S then

$$\hat{x} = \frac{1}{I} \sum_{i=1}^{I} P_i \hat{x}. \tag{2.5}$$

Therefore, a reasonable choice to measure how far any x is from solving this problem is the discrepancy function

$$d(x) = \|x - \frac{1}{I} \sum_{i=1}^{I} P_i x\|^2, \tag{2.6}$$

not h itself. Note that if the iteration being used is $x^{k+1} = Tx^k$ with

$$Tx = \frac{1}{I} \sum_{i=1}^{I} P_i x, \tag{2.7}$$

then the d in Equation (2.6) is simply $d(x) = \|x - Tx\|^2$. In such cases monitoring $\{h(x^k)\}$, instead of $\{d(x^k)\}$, to see how the iteration is progressing is not a good idea, since we usually have no idea how small the values of h can be.

In [11] the author identifies two research directions for SM: weak SM is when S is assumed to be nonempty, while in strong SM the solution set S may be empty. He claims that strong SM is more practical because in the real world having consistent constraints is unlikely. However, it is not unlikely that h will have minimizers, in which case S is not empty and we will be in the weak SM situation.

3. Should the requirements for the SM Be Weakened?

In [20] the authors point out that it is often the case that iterative algorithms cannot be implemented exactly. The perturbations that are introduced come from unavoidable inexactness and are not put there by the user. When these algorithms are accelerated the hope is that the perturbed iterates will exhibit the same improved rates of convergence as the exact iterates would. In other words, one wants the iteration to be resilient. This suggests to these authors that there may be a role for SM that goes beyond what I have described in the preceding paragraphs. We give some examples to illustrate this point.

3.1. **Some Examples.**

3.1.1. *Bauschke's Algorithm.* In [2] Heinz Bauschke considers the problem of finding a vector that is a common fixed point for a finite family of nonexpansive mappings on Hilbert space. A particular case of his main theorem applies to the CFP in $\mathbb{R}^J$. With T given by Equation (2.7), $f(x) = \frac{1}{2}\|x-p\|^2$, and $v^k = -\nabla f(Tz^k)$, the algorithm with the iterative step given by Equation (2.1) converges to the orthogonal projection of p onto the set of fixed points of T, which is also the set of minimizers of the function h given by Equation (2.4), provided that $\{t_k\} \to 0$, $\sum_{k=1}^{\infty} t_k = +\infty$ and $\sum_{k=1}^{\infty} |t_k - t_{k+1}| < +\infty$.

3.1.2. *Yamada's Algorithm.* There is an interesting connection with Isao Yamada's *hybrid steepest descent* algorithm [25], which uses the iterative step

$$x^{k+1} = Tx^k - t_k F(T(x^k)) \tag{3.1}$$

to solve the variational inequality problem relative to the fixed-point set of the operator T and the monotone operator F. The operator T is nonexpansive, F is Lipschitz and strongly monotone, and, once again, the sum of the sequence $\{t_k\}$ must diverge. When $F = \nabla f$ is the gradient of a convex differentiable function f, the iteration in Equation(3.1) becomes that of Equation (2.1). The difference now is that the hybrid steepest descent algorithm seeks to minimize f over the set of fixed points of T, not simply to reduce f.

3.1.3. *Dykstra's Algorithm.* Because the objective is to design iterative algorithms for projecting onto $C = \cap_{i=1}^{I} C_i$ using only the projections onto the individual C_i, it would help if we had a way to characterize projection onto C in terms of these projections onto the C_i. We do not have such a characterization, but we do have sufficient conditions for c to be $P_C x$. Our lemma [6], relating the orthogonal projection operator P_C to the $P_i \doteq P_{C_i}$, will serve to motivate the Dykstra algorithm.

Lemma 3.1. *If $x = c + p_1 + p_2 + ... + p_I$ and $c = P_i(c + p_i)$, for each i, then $c = P_C x$.*

Proof: Let d be arbitrary in C. Then, for each i,

$$\langle c - (c + p_i), d - c\rangle \geq 0, \tag{3.2}$$

since d is in C_i. Summing the inequalities over i gives

$$\langle c - x, d - c\rangle \geq 0, \tag{3.3}$$

for all d in C. Therefore, $c = P_C x$. ∎

Consider the problem of finding the point in the nonempty set $C = A \cap B$ that is closest to p in the Euclidean sense, where A and B are closed convex subsets of $\mathbb{R}^J$. The alternating orthogonal projection (AOP) algorithm is the following. Let $y_0 = p$. Having found y_{n-1} let

$$z_{n-1} = P_A y_{n-1}$$

$$y_n = P_B z_{n-1}. \tag{3.4}$$

The sequences $\{y_n\}$ and $\{z_n\}$ both converge to the same member of C, but not necessarily to $P_C p$ [14].

Again, let $f(x) = \frac{1}{2}\|x - p\|^2$ and $y_0 = p$. Having found y_{n-1}, we take

$$\begin{aligned} z_{n-1} &= P_A y_{n-1} - \nabla f(y_{n-1}), \\ y_n &= P_B z_{n-1} - \nabla f(z_{n-1}). \end{aligned} \tag{3.5}$$

Although the reader may not recognize it, this is Dykstra's algorithm [17]. The perturbations are $\nabla f(y_{n-1}) = y_{n-1} - p$ and $\nabla f(z_{n-1}) = z_{n-1} - p$. The sequences $\{y_n\}$ and $\{z_n\}$ need not converge separately, indeed, they need not be bounded, in which case the perturbations are not bounded. However, the sequences $\{P_A y_n\}$ and $\{P_B z_n\}$ both converge to $P_C p$ and the sequence $\{y_n + z_n\}$ converges to $p + P_C p$.

4. Is Likelihood Maximization a Linear-Algebra Problem?

For the problem of image reconstruction from PET data the "moment estimator" of the unknown x is found by solving the system of linear equations $y = Px$ for $x \geq 0$. As the authors of [24] point out, noise and model error usually result in this system having no nonnegative solutions. They therefore dismiss the idea that the problem is linear-algebraic and turn to statistical estimation using likelihood maximization and the EMML algorithm. In the Comments that follow that article G.T. Herman et al. point out that one can attempt to solve $y = Px$ approximately, using ART or some other method from numerical linear algebra, and suggest that the statistical approach is not all that different from their linear-algebraic approach. In their response to this Comment the original authors deny that the two approaches are at all similar and that solving, even approximately, any system of linear equations bears no resemblance to their statistical approach.

In [3, 4, 5] I showed that maximizing the likelihood is, in fact, finding an approximate solution of $y = Px$; maximizing likelihood is equivalent to minimizing $KL(y, Px)$. Moreover, the "simultaneous multiplicative algebraic reconstruction technique" (SMART) [16, 22] minimizes $KL(Px, y)$. The SMART has the iterative step $x^{k+1} = Sx^k$, where

$$(Sx)_j = x_j \exp\left(\sum_{i=1}^{I} P_{i,j} \log(y_i/(Px)_i)\right). \tag{4.1}$$

The similarities between Equation (4.1) and Equation (1.1) are striking.

5. Is SM Regularization Made Easy?

The minimizers of $KL(Px, y)$ behave just like those of $KL(y, Px)$ when $y = Px$ has no nonnegative solutions, and so require regularization to produce useful images. We can minimize $KL(Px, y) + \epsilon KL(x, p)$ with the iterative step defined by

$$\log x_j^{k+1} = (1-\alpha)\log(Sx^k)_j + \alpha \log p_j, \tag{5.1}$$

or

$$x_j^{k+1} = (Sx^k)_j^{1-\alpha} p_j^{\alpha}, \tag{5.2}$$

with $\alpha = \frac{\epsilon}{1+\epsilon}$. But we can use SM and $f(x) = \|x-p\|^2$ to regularize SMART more simply, with the iteration

$$x^{k+1} = Sx^k + \epsilon p. \tag{5.3}$$

We are not claiming convergence to any sum of two functions, as typical regularization would require. All we are claiming is that this modification of the iterative step will avoid the random high-frequency oscillations that we may begin to see after some finite number of iterations.

References

[1] Andersen, M., and Hansen, P. C. (2014) "Generalized row-action methods for tomographic imaging." *Numer. Algor.*, **67**, pp. 121–144.

[2] Bauschke, H. (1996) "The approximation of fixed points of compositions of nonexpansive mappings in Hilbert space," *Journal of Mathematical Analysis and Applications*, **202**, pp. 150–159.

[3] Byrne, C. (1993)"Iterative image reconstruction algorithms based on cross-entropy minimization." *IEEE Trans. Image Proc.*, **IP-2**, pp. 96–103.

[4] Byrne, C. (1995) "Erratum and addendum to 'Iterative image reconstruction algorithms based on cross-entropy minimization.'. " *IEEE Trans. Image Proc.*, **IP-4**, pp. 225–226.

[5] Byrne, C. (1996) "Iterative reconstruction algorithms based on cross-entropy minimization." in *Image Models (and their Speech Model Cousins*, S.E. Levinson and L. Shepp, (eds.), IMA Volumes in Mathematics and its Applications, Volume 80, 1–11.

[6] Byrne, C. (2014) *Iterative Optimization in Inverse Problems*. Boca Raton, FL: CRC Press.

[7] Byrne, C. (2019) "Thoughts on Superiorization." *ResearchGate*, April 28, 2019.

[8] Byrne, C. (2019) "Thoughts on Superiorization II." *ResearchGate*, May 18, 2019.

[9] Censor, Y., Davidi, R., and Herman, G.T. (2010) "Perturbation resilience and superiorization of iterative algorithms." *Inverse Problems*, **26**, p. 065008.

[10] Censor, Y., and Zaslavski, A. (2015) "Strict Fejér monotonicity by superiorization of feasibility-seeking projection methods." Journal of Optimization Theory and Applications, **165**, pp. 172–187.

[11] Censor, Y. (2015) "Weak and strong superiorization: between feasibility seeking and minimization." *Analele Stiint. Univ. Ovidius Constanta– Ser. Mat.*, **23**, pp. 141–154.

[12] Censor, Y., Herman, G.T., and Jiang, M. (2017) "Superiorization: theory and applications." Preface to *Inverse Problems*, **33**.

[13] Censor, Y. (2017) "Superiorization and perturbation resilience of algorithms: a continuously updated bibliography." Technical Report, Original report: June 13, 2015

contained 41 items. First revision: March 9, 2017 contains 64 items. Available on arXiv at: https://arxiv.org/abs/1506.04219v2 and on YC website.

[14] Cheney, W., and Goldstein, A. (1959) "Proximity maps for convex sets." *Proc. Am. Math. Soc.*, **10**, pp. 448–450.

[15] Combettes, P. and Wajs, V. (2005) "Signal recovery by proximal forward-backward splitting." *Multiscale Modeling and Simulation*, **4(4)**, pp. 1168–1200.

[16] Darroch, J. and Ratcliff, D. (1972) "Generalized iterative scaling for log-linear models." *Annals of Math. Stat.*, **43**, pp. 1470–1480.

[17] Dykstra, R. (1983) "An algorithm for restricted least squares regression." *J. Amer. Statist. Assoc.*, **78 (384)**, pp. 837–842.

[18] Garduño, E. and Herman, G.T. (2014) "Superiorization of the ML–EM algorithm." *IEEE Trans. Nucl. Sci.*, **61**, pp. 162–172.

[19] Herman, G.T., Garduño, E., Davidi, R. and Censor, Y. (2012) "Superiorization: An optimization heuristic for medical physics." *Medical Physics*, **39**, pp. 5532–5546. DOI:10.1118/1.4745566.

[20] Reem, D., and De Pierro, A. (2017) "A new convergence analysis and perturbation resilience of some accelerated proximal forward-backward algorithms with errors." *Inverse Problems*, **33**, 044001.

[21] Rockmore, A., and Macovski, A. (1976) "A maximum likelihood approach to emission image reconstruction from projections." *IEEE Trans. Nucl. Sci.*, **NS-23**, pp. 1428–1432.

[22] Schmidlin, P. (1972) "Iterative separation of sections in tomographic scintigrams." *Nuklearmedizin*, **11**, pp. 1–16.

[23] Shepp, L., and Vardi, Y. (1982) "Maximum likelihood reconstruction for emission tomography." *IEEE Trans. Med. Imag.*, **MI-1**, pp. 113–122.

[24] Vardi, Y., Shepp, L., and Kaufman, L. (1985) "A statistical model for positron emission tomography." *J. Amer. Stat. Assoc.* **80**, pp. 8–20.

[25] Yamada, I. (2001) "The hybrid steepest descent method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings." in [**?**], pp. 473–504.

(C. Byrne) Department of Mathematical Sciences, University of Massachusetts Lowell, Lowell, MA, USA

E-mail address: Charles_Byrne@uml.edu