Iterative Discretization of Optimization Problems Related to Superresolution

Axel Flinth and Pierre Weiss

Abstract—We study an iterative discretization algorithm for solving optimization problems regularized by the total variation norm over the space $\mathcal{M}(\Omega)$ of Radon measures on a bounded subset Ω of $\mathbb{R}^d$. Our main motivation to study this problem is the recovery of sparse atomic measures from linear measurements. Under reasonable regularity conditions, we arrive at a linear convergence rate guarantee.

I. Superresolution and Semi-Infinite Programming

[1] Imagine a signal consisting of a few localized peaks at points $x_i \in \Omega \subseteq \mathbb{R}^d$, and an inaccurate, linear measurement of it. One may imagine a point light source viewed through a lens with finite aperture. An idealized model of this measurement process is to view the signal as a sparse atomic measure $\mu_0 = \sum_{i=1}^s c_i \delta_{x_i} \in \mathcal{M}(\Omega)$, and the measurement as a linear operator $A : \mathcal{M}(\Omega) \to \mathbb{R}^m$ of the form

$$A\mu = \left(\int_\Omega a_j d\mu \right)_{j=1}^m.$$

A few years ago, several independent teams of researchers [4], [6], [21] realized that for this signal recovery task, the canonical extension of basis pursuit, the *total variation minimization*[2]

$$\min_{\substack{\mu \in \mathcal{M}(\Omega) \\ A\mu=b}} \|\mu\|_{\mathcal{M}} \tag{1}$$

enjoys similar properties to its discrete counterpart. For many measurement operators [18], it can be proven that the solution (1) for $b = A\mu_0$ for a $\mu_0 = \sum_{i=1}^s c_i \delta_{x_i}$ is equal to μ_0, provided the sources $(x_i)_{i=1}^s$ are well-separated. One can also consider other data fidelity terms $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$, leading to a problems of the form

$$\min_{\mu \in \mathcal{M}(\Omega)} \|\mu\|_{\mathcal{M}} + f(A\mu). \tag{$\mathcal{P}$}$$

---

[1]This proceeding is work in progress, and the results should seen as preliminary. More details will be contained in the manuscripts [13], [14], which we aim to submit in near future.

[2]*Total variation* will in this work consequently be used in the sense of measures, i.e. $\|\mu\|_{\mathcal{M}} = \sup_{\substack{f \in \mathcal{C}(\Omega) \\ \|f\|_\infty \leq 1}} \int_\Omega f d\mu$ (and not in the sense of functions of bounded variation, i.e. $\|u\|_{TV} = \sup_{\substack{\phi \in \mathcal{C}^1(\Omega) \\ \|\phi\|_\infty \leq 1}} \int_\Omega u \mathrm{div}\phi dx$).

Generically, $(\mathcal{P})$ will have at least one solution which is a sparse atomic measure [1], [12], [22].

The numerical resolution of $(\mathcal{P})$ is not trivial. Notably, the space over which we optimize is not only infinite-dimensional, but also non-separable. Provided f is convex, one possible strategy for solving (1) is to investigate the dual of $(\mathcal{P})$

$$\sup_{\substack{q \in \mathbb{R}^m \\ \|A^*q\|_\infty \leq 1}} -f^*(q). \tag{$\mathcal{D}$}$$

This problem, although the variable over which we optimize is no longer infinite-dimensional, is also not easy to solve. The reason is that the constraint is still of infinite-dimensional nature. In a few special cases, including the important one of measurement functions of (trigonometrically) polynomial nature [7], [17], the constraint can be rewritten into a finite-dimensional one. In general however, it seems that one needs to resort to discrete approximations.

The simplest way of discretizing $(\mathcal{P})$ (or $(\mathcal{D})$) is to simply restrict it to measures supported on a fixed finite grid Ω_k (or to only control a few of the infinitely many constraints $|(A^*q)(x)| \leq 1$, respectively):

$$\min_{\mu \in \mathcal{M}(\Omega_k)} \|\mu\|_{\mathcal{M}} + f(A\mu) \tag{$\mathcal{P}_k$}$$

$$\sup_{\substack{q \in \mathbb{R}^m \\ \sup_{x \in \Omega_k} |(A^*q)(x)| \leq 1}} -f^*(q). \tag{$\mathcal{D}_k$}$$

If the sources $(x_i)_{i=1}^s$ are guaranteed to lay on the grid, this will of course succeed. However, as is well known, the approach suffers from severe problems if they are not. One speaks of *basis mismatch* [5]. In order to get an accurate approximation, one needs to use a very fine grid, which leads to both computationally heavy and ill-conditioned problems. Knowing that the solution should have a sparse support, it is tempting to try to instead iteratively refine the grid, hoping to obtain a grid which is fine close to the final support and else course. Abstractly, such a scheme would have the following form:

Generic Iterative Discretization
1) Find a primal-dual solution pair (μ_k, q_k) of $(\mathcal{P}_k)$-$(\mathcal{D}_k)$.
2) According to some refinement rule R, calculate $\Omega_{k+1} = R(\Omega_k, \mu_k, q_k)$

3) Repeat until convergence.

These schemes have been known and applied for a long time on the field of semi-infinite programming [19], [16], [20]. Approaches like this are used heuristically in the signal processing community, but their theoretical properties seem to have been overlooked to a great extent. An exception is the instance of the algorithm induced by the refinement rule

$$R(\Omega_k, \mu_k, q_k) = \Omega_k \cup \{\arg\max |A^* q_k|\}.$$

Technically, it has previously been treated under the name of conditional gradient or Frank-Wolfe algorithm [15], e.g. in [3]. It has however been shown to be essentially equivalent to the approach treated here [10]. A convergence rate beyond the one known to hold generically for any instance of the Frank-Wolfe algorithm is yet to be proven. Here should also be mentioned the recent *sliding Frank-Wolfe algorithm*, which consists of alternating between a conjugate gradient and continuously changing the positions and amplitudes of the δ_x-components [2], [8]. The latter are known to converge in a finite, but as of now unbounded.

In this paper, we present and analyse a slightly different refinement rule. We will in particular arrive, under a few regularity assumptions, at a guarantee of linear convergence guarantee, which we believe to be novel. We will also briefly discuss how it could be implemented without having to explicitly find the argmax of the non-convex function $|A^* q_k|$.

II. MAIN RESULTS

Our iterative discretization algorithm is determined by its refinement rule. We assume for simplicity that $\Omega = [0, 1]^d$, and that Ω_k is given as vertices of a collection C_k of *dyadic cells* of the form $n_\ell / 2^{J_\ell} + 2^{-J_\ell} [0, 1]^d$ for some $J_\ell \in \mathbb{N}$ and $n_\ell \in \{0, \ldots, 2^J - 1\}^d$.

The refinement rule we aim to study is now as follows:
Refinement Rule

1) Given q_k, determine the set M of cells $\omega \in C_k$ which have a non-empty intersection with the set

$$X_k = \{x \,|\, x \text{ local max of } |A^* q_k|, |A^* q_k(x)| > 1.\}$$

2) Subdivide each cell $\omega \in M$ into four dyadic subcells of equal size.

3) Define Ω_{k+1} as the vertices of the new collection C_{k+1}.

This rule leads to convergence in great generality.

Theorem II.1. *Assume that f is convex and bounded below, A is weak-∗-continuous and A restricted to measures supported on Ω_0 is surjective. Then, the sequence (μ_k, q_k)*
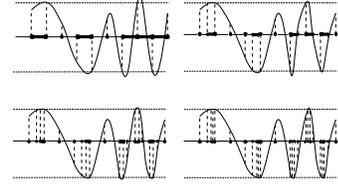
Fig. 1: $A^* q_k$ for the first four iterations.

of dual-primal solutions of $(\mathcal{P}_k)$ and $(\mathcal{D}_k)$ contain subsequences (μ'_k, q'_k) weak-∗-converging to a dual-primal pair of solutions (μ_*, q_*) of $(\mathcal{P})$ and $(\mathcal{D})$. If any of the problems $(\mathcal{P})$, $(\mathcal{D})$ have unique solutions, the whole corresponding sequence converges.

Theorem (II.1) is satisfactory, since it guarantees global convergence in very great generality. It however fails to provide any information about the rate of convergence. In particular, it could be that the refinement takes place uniformly over the whole domain Ω, leading the employment of the exchange algorithm pointless.

Remark 1. *The refinement rule only subdividing the cell containing the* argmax *will also fulfill the generic convergence Theorem II.1. However, for the rest of the analysis, it is (for technical reasons) important that each cell intersecting X_k is subdivided.*

A. Refined Analysis for Smooth Data Fidelity Terms

In order to make a more detailed analysis, we make the following regularity assumptions:

1) f is differentiable, with an L-Lipschitz gradient
2) The measurement functions a_j are all $\mathcal{C}^2$.
3) $(\mathcal{P})$ has a unique solution $\mu_* = \sum_{i=1}^s \alpha_i^* \delta_{\xi_i}$.
4) The associated dual solution $q^\star$ obeys the following regularity conditions: We assume that the only points x for which $|A^* q(x)| = 1$ are the points in ξ, and that the second derivative of $|A^* q|$ is negative definite in each point ξ_i .

In particular assumption (4) seems to hard to control in advance. However, for the important special case of $f(v) = \frac{\tau}{2} \|v - b\|_2^2$ for a $b = A\mu_0$ for an atomic μ_0 with well-separated support, it can furthermore be proven to hold true, at least for large values of τ (see [9], [18]).

To formulate our main result, let us introduce a terminology: we say that the algorithm has entered a δ-*regime* if in all future iterations, each point in X_k lies at a distance at most δ from ξ.

Theorem II.2. *There exists constants C, $K > 0$ with the following property: There exists a finite number N with the so that after iteration N, X_k contains at most s points.*

Furthermore, algorithm enters a τ-regime after at most $N+ k_\tau$ iterations, where

$$k_\tau = s \left(K \log \left(\frac{\bar{\tau}_0}{\tau} \right) + 2^d \lceil \log \left(\tau^{-1} \right) \rceil \right)$$

Additionally, for $k \geq k_\tau + 1$, we will have

$$(\inf (\mathcal{P}_k)) - (\inf (\mathcal{P})) \leq C\tau$$

Remark 2. *The value of N is given by the number of iterations the algorithm needs before $\|q_k - q^\star\|_2$ falls below a certain threshold, and $|\mu_k|(U_i) \neq 0$ for certain neighborhoods U_i of ξ_i. This number is finite due to the generic convergence result and the uniqueness of the solutions $q^\star$ and $\mu^\star$, but its size is hard to determine a-priori.*

Overall, this theorem states after a finite initial number of iterations that we need $C \log(\tau^{-1})$ iterations to pinpoint the location of the support ξ, and the optimal objective function value, up to an error τ. In other words, we obtain an asymptotically linear convergence rate.

The proof of Theorem II.2 is too complicated to present here, and we refer to [13], [14] for details. Let us however sketch its main steps.

Sketch of proof:. Let us make the simplifying assumption that ξ only contains one point (ξ having more than one point leads to additional subtleties), and start by collecting a few (non-trivial) stability facts on $(\mathcal{D})$:

1) If q_k is the solution of $(\mathcal{D}_k)$, we have

$$\sup_{x \in \Omega} |A^* q_k(x)| \leq 1 + \epsilon \Rightarrow \|q - q_*\|_2 \leq K_1 \sqrt{\epsilon}$$

where K_1 is a constant dependent on f^* and q^*. Here, it is crucial that f has an L-Lipschitz smooth gradient, since it causes f^* to be strongly convex, which in turn leads to a favorable geometry of the set $\{q \,|\, f^*(q) \leq f^*(q_*)\}$, in which q_k is contained.

2) If we denote $\widehat{M}_k$ the set of cells ω with $\omega \cap X_k \neq \emptyset$, we have

$$\sup_{\omega \in \widehat{M}_k} \operatorname{diam}(\omega) \leq K_2\epsilon \Rightarrow \|A^* q_k\|_\infty \leq 1 + \epsilon^2.$$

3) Define $\operatorname{dist}(\xi_i, \Omega_k) = \sup_{\xi_i \in \xi} \inf_{x \in \Omega_k} \|x - \xi_i\|_2$. If the algorithm is in a τ-regime, we have

$$\operatorname{dist}(\xi, \Omega_k) \leq K_3\epsilon \Rightarrow \|A^* q_k\|_\infty \leq 1 + \tau\epsilon.$$

4) If q_k is close to q_*, $A^* q_k$ must be close to $A^* q_*$. In particular, the maxima of $|A^* q_k|$ exceeding 1 must be close to points in which $|A^* q_*|$ is equal to 1.

We may now argue as follows: Due to the generic convergence result, we will eventually have $\|A^* q_k\|_\infty \leq 1 + \epsilon$. Choosing ϵ small enough, this will by (1) cause q_k to be close to $q^\star$. This in turn will by (4) imply that X_k is close to ξ, and that $|X_k|$ only contains $s = 1$ point. We will hence enter a $\hat{\tau}_0$-regime for a small enough, but fixed, size $\hat{\tau}_0$ in N iterations.

Being in a $\hat{\tau}_0$-regime, $\omega \cap X_k \neq \emptyset$ will in forthcoming iterations only be true for cells in a region close to ξ. Since those cells are divided in each step, and that region only contains a finite amount of dyadic cells of diameter larger than $K_2\epsilon$, we will at some point have $\operatorname{diam}(\omega) \leq K_2\epsilon$ for all cells in $\widehat{M}_k$. Invoking (2), we may trigger (1) to imply that q_k is very close to q_*. (4) together with the fact that the cell divided is small then guarantees that we add a point to Ω_k which makes $\operatorname{dist}(\xi, \Omega_k)$ smaller. We may then by (3) (via (1) and (2)) secure that for any forthcoming iteration, all points in X_k lie $\hat{\tau}_0/2$-neighborhood of ξ. Hence, we will enter a $\hat{\tau}_0/2$-regime in finite time.

The remaining task is now to carefully repeat the last argument, counting how many iterations of the algorithm we need for each 'argumentation cycle' to be completed. One arrives at the first result.

The second result is proved by arguing that if the optimal solution of $(\mathcal{P})$ is $\mu_* = \sum_{i=1}^s \alpha_i^* \delta_{\xi_i}$, the measure $\widetilde{\mu}_k = \sum_{i=1}^s \alpha_i^* \delta_{x_{k,i}}$ for points $x_{k,i} \in \Omega_k$ close to ξ_i is feasible for $(\mathcal{P}_k)$. Thus, $\inf (\mathcal{P}_k) \leq \|\widetilde{\mu}_k\|_{\mathcal{M}} + f(A\widetilde{\mu}_k) = \|\mu_*\| + f(A\widetilde{\mu}_k)$. By estimating $f(A\widetilde{\mu}_k) - f(A\mu_*)$, we arrive at the second result. $\square$

III. Implementation of the Algorithm

In order to test the empirical performance, we have implemented our algorithm in MATLAB.

The refinement rule we have discussed requires to determine all local maximizers of $|A^* q_k|$, which in general seems impossible. We have instead chosen to implement a rule which in an iterative fashion searches for points where $\nabla |A^* q_k|$ is small and $|A^* q_k|$ is large. If one stops this procedure based on critera relating to the second derivative of A, one can prove that this will refine all cells in $\widehat{M}_k$, while at the same time only refining additional cells which are close to $\widehat{M}_k$. This will lead to slightly more complex problems $(\mathcal{P}_k)$, $(\mathcal{D}_k)$, but not affect the convergence rate in the sense of Theorem (II.2). We refer to [14] for details, as well as a discussion on other ways of implementing the refinement rule.

Let us present the results of our experiments for a toy example: bumped Gaussian measurements on $\Omega = [0, 1]^2$. That is, we choose the measurement functions a_j equal to

$$a_j(x) = \exp(-\Gamma \|x - x_j\|_2^2 / 2)\phi(x),$$

where $x_i \in [0, 1]^2$ and ϕ is a smooth function equal to 1 on $[.1, .9]^2$ and 0 on the boundary of Ω. The inclusion of the latter is necessary to ensure that the operator A
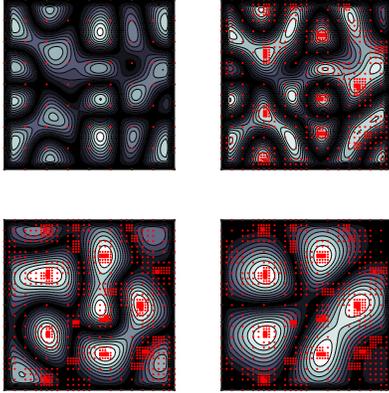
Fig. 2: Evolution of $|A^*q_k|$ and Ω_k, for (from top left to bottom right) $k = 1, 5, 10$ and 15 along a run of iterative discretization algorithm. Notice the concentration of grid points near the vertices of the regular pentagon on the circle of radius .5 around the origin (that is, the support of $\mu^\star$) for late iterations.

is weak-$*$-continuous (weak-$*$-continuity is equivalent to the measurement functions vanishing on the boundary). Excluding ϕ, which should not play a great role if we choose the x_i and the support of the ground truth μ_* far away from the boundary, there are theoretical results [18] indicating that our regularity assumption should be true when the support of $\mu^\star$ is well-separated.

We choose $(x_j)_{j=1}^m$ equal to a uniform grid on $[-3/4, 3/4]^2$ with 6^2 points, $\Gamma = 15$, and generate a sparse atomic ground truth $\mu^\star$ by choosing 5 points ξ_i equally spaced on the circle centered at the origin with radius .5, and assigning them random amplitudes in $\{-1, 1\}$. We start our algorithm with a grid of width 2^{-4} and let it run for 6 iterations, solving

$$\min_{\mu \in \mathcal{M}(\Omega)} \|\mu\|_{TV} + \frac{\tau}{2}\|A\mu - b\|_2^2.$$

with $\tau = 10^7$ and $b = A\mu_0$. The evolutions of the dual certificates A^*q_k and grids Ω_k are displayed in Figure 2. The experiments resonates well with the theory: In early iterations, points are added to the grid which are not close to ξ at all . After a while, although not imminent from the figures, the grids do concentrate around ξ.

Acknowledgements

REFERENCES

[1] Claire Boyer, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric De Gournay, and Pierre Weiss. On representer theorems and convex regularization. *SIAM Optimization*, 2019.

[2] Nicholas Boyd, Geoffrey Schiebinger and Benjamin Recht, The alternating descent conditional gradient method for sparse inverse problems. *SIAM Journal on Optimization*, 27(2):616-639, 2017.

[3] Kristian Bredies and Hanna Katriina Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.

[4] Emmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956, 2014.

[5] Yuejie Chi, Louis L. Scharf, Ali Pezeshki, A. Robert Calderbank. Sensitivity to basis mismatch in dompressed sensing. *IEEE Transactions on Signal Processing*, 59(5):2182-2195, 2011.

[6] Yohann De Castro and Fabrice Gamboa. Exact reconstruction using beurling minimal extrapolation. *Journal of Mathematical Analysis and applications*, 395(1):336–354, 2012.

[7] Yohann De Castro, Fabrice Gamboa, Didier Henrion, and J-B Lasserre. Exact solutions to super resolution on semi-algebraic domains in higher dimensions. *IEEE Transactions on Information Theory*, 63(1):621–630, 2017.

[8] Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The sliding frank-wolfe algorithm and its application to super-resolution microscopy. *arXiv preprint arXiv:1811.06416*, 2018.

[9] Vincent Duval and Gabriel Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.

[10] Armin Eftekhari and Andrew Thompson. Sparse Inverse Problems Over Measures: Equivalence of the Conditional Gradient and Exchange Methods *arXiv preprint arXiv:1804.10243*, 2018.

[11] S. D. Fisher and J. W. Jerome. Spline solutions to L1 extremal problems in one and several variables. *Journal of Approximation Theory*, 13(1):73–83, 1975.

[12] Axel Flinth and Pierre Weiss. Exact solutions of infinite dimensional total-variation regularized problems. *Information and Inference*, 2017. *to appear*.

[13] Axel Flinth, Fréderic de Gournay and Pierre Weiss. Exchange and prosper: Linear rates of convergence for a Frank-Wolfe Algorithm. (Working title) *In preparation*.

[14] Axel Flinth, Fréderic de Gournay and Pierre Weiss. Exchange and prosper II: Efficient and reliable discretization of a Frank-Wolfe Algorithm. (Working title) *In preparation*.

[15] M. Frank and P. Wolfe "An algorithm for quadratic programming". *Naval Research Logistics Quarterly*, 3(1–2): 95–110, 1956.

[16] Rainer Hettich and Kenneth O Kortanek. Semi-infinite programming: theory, methods, and applications. *SIAM review*, 35(3):380–429, 1993.

[17] Jean-Bernard Lasserre. *Moments, positive polynomials and their applications*, volume 1. World Scientific, 2010.

[18] Clarice Poon, Nicolas Keriven, and Gabriel Peyré. Support localization and the fisher metric for off-the-grid sparse regularization. *arXiv preprint arXiv:1810.03340*, 2018.

[19] Rembert Reemtsen. Discretization methods for the solution of semi-infinite programming problems. *Journal of Optimization Theory and Applications*, 71(1):85–103, 1991.

[20] Rembert Reemtsen and Stephan Görner. Numerical methods for semi-infinite programming: A survey. pages 195–262, 1998.

[21] Gongguo Tang, Badri Narayan Bhaskar, Parikshit Shah, and Benjamin Recht. Compressed sensing off the grid. *IEEE transactions on information theory*, 59(11):7465–7490, 2013.

[22] Michael Unser, Julien Fageot, and John Paul Ward. Splines are universal solutions of linear inverse problems with generalized tv regularization. *SIAM Review*, 59(4):769–793, 2017.