

Lasry–Lions Envelopes and Nonconvex Optimization: A Homotopy Approach

Miguel Simões

Dept. Electr. Eng. (ESAT) – STADIUS
KU Leuven
Leuven, Belgium
miguel.simoës@kuleuven.be

Andreas Themelis

Fac. Inf. Sci. Electr. Eng. (ISEE)
Kyushu University
Fukuoka, Japan
andreas.themelis@ees.kyushu-u.ac.jp

Panagiotis Patrinos

Dept. Electr. Eng. (ESAT) – STADIUS
KU Leuven
Leuven, Belgium
panos.patrinis@kuleuven.be

Abstract—In large-scale optimization, the presence of nonsmooth and nonconvex terms in a given problem typically makes it hard to solve. A popular approach to address nonsmooth terms in convex optimization is to approximate them with their respective Moreau envelopes. In this work, we study the use of Lasry–Lions double envelopes to approximate nonsmooth terms that are also not convex. These envelopes are an extension of the Moreau ones but exhibit an additional smoothness property that makes them amenable to fast optimization algorithms. Lasry–Lions envelopes can also be seen as an “intermediate” between a given function and its convex envelope, and we make use of this property to develop a method that builds a sequence of approximate subproblems that are easier to solve than the original problem. We discuss convergence properties of this method when used to address composite minimization problems; additionally, based on a number of experiments, we discuss settings where it may be more useful than classical alternatives in two domains: signal decoding and spectral unmixing.

Index Terms—Nonsmooth nonconvex optimization, hyperspectral imaging, Lasry–Lions smoothing.

I. INTRODUCTION

Many problems in signal processing are formulated as composite problems of the form

$$\underset{\mathbf{x} \in \mathbb{R}^n}{\text{minimize}} \varphi(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x}), \quad (\text{I.1})$$

where $g : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is a data-fitting term and $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ acts as a regularizer or as a constraint. These problems are typically faced with a number of challenges: they involve a very large number of variables, they are ill posed, and their data-fitting and regularizer terms are nonsmooth and nonconvex. As an example, consider that one wishes to solve a regression problem whose solution is known to be sparse. One may use the so-called ℓ_0 pseudo-norm as a regularizer, i.e., $h = \|\cdot\|_0$, which measures the number of nonzero entries of a given vector. Finding global solutions to large-scale problems involving it is impractical, since this function is nonsmooth and nonconvex, and these problems are often approached by replacing it with surrogates, such as the ℓ_1 norm, which allow one to use methods that are more computationally tractable. Under certain conditions on g , both approaches produce equivalent

solutions [1], but these conditions are often violated. Consequently, it is of interest to solve the problem with the nonconvex regularizer. Perhaps surprisingly, in practice it is not always the case that the use of the ℓ_0 pseudo-norm instead of the ℓ_1 surrogate will produce a solution that is more useful, due to either the inherent ill posedness of the problem or wrong modeling assumptions on g . However, recent work has presented experimental evidence that this solution may in fact be more useful in certain noise regimes [2, 3], and has validated the idea that a combination of the ℓ_0 regularizer with additional ℓ_1 or ℓ_2 terms produces the most useful results for the different noise regimes of practical interest [4].

In this work, we discuss a method that makes use of the Lasry–Lions double envelope to build surrogates of nonsmooth nonconvex functions. For a proper, lower semicontinuous (lsc) function $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ and parameters $\lambda > \mu > 0$, its λ -Moreau envelope is

$$h^\lambda(\mathbf{x}) := \inf_{\mathbf{w}} \left\{ h(\mathbf{w}) + \frac{1}{2\lambda} \|\mathbf{w} - \mathbf{x}\|^2 \right\}$$

and its (λ, μ) -Lasry–Lions double envelope is

$$h^{\lambda, \mu}(\mathbf{x}) := -(-h^\lambda)^\mu(\mathbf{x}) = \sup_{\mathbf{w}} \left\{ h^\lambda(\mathbf{w}) - \frac{1}{2\mu} \|\mathbf{w} - \mathbf{x}\|^2 \right\}. \quad (\text{I.2})$$

Replacing h in (I.1) with its envelope $h^{\lambda, \mu}$ —which, e.g., can be made for the ℓ_0 pseudo-norm or the indicator of a nonconvex set—furnishes approximate subproblems arbitrarily close to the original one and that are, in a way to be made clearer in a later section, “less” nonconvex. We also show that these problems are smooth, and hence can be tackled via fast smooth optimization solvers. We make use of the surrogate problems in a way that mimics the behavior of homotopy methods for systems of nonlinear equations or of penalty methods for constrained optimization: we solve a sequence of subproblems that increasingly approximate the original problem, and use the solution of a given subproblem to inform the initial estimate of the following one. As we show in Section III, the sequence of global minimizers of these subproblems converges to the global minimizer of (I.1). We motivate the use of this method in two ways. First, we argue that the use of Lasry–Lions double envelopes provides a way to systematically produce surrogates of nonconvex functions that is useful from a computational perspective, in the sense just described. Second, its use also seems to have benefits in terms of the solutions

This work was supported by the Research Foundation – Flanders (FWO) projects G0A0920N, G086518N, and G086318N, by the Research Council KU Leuven C1 project C14/18/068, and by the Fund for Scientific Research – FNRS and FWO EOS project 30468160 (SeLMA).

that are found through it. In fact, the Lasry–Lions envelope of the ℓ_0 pseudo-norm can be seen as a combination of non-convex and convex terms. This combination is different from the combinations of the ℓ_0 regularizer with convex terms described above but, when compared to the use of the pure ℓ_0 regularizer, can be said to be less “aggressive”, in the sense used in [3] when referring to the bias–variance trade-off issue present in estimation problems.

To the best of our knowledge, the use of the Lasry–Lions envelope as described seems to be new, but there are some connections to existing approaches in the literature. For example, there is a family of methods that addresses the problems of ℓ_0 -regularized least squares and of rank-constrained matrix approximation by replacing the nonconvex terms with others that are still nonconvex but more tractable; additionally, results associating the minimizers of both formulations exist [5]–[7]. The connections to our work come from the fact that these replacements correspond to the proximal hull of certain functions—the λ -proximal hull of function h corresponds to making $\mu = \lambda$ in (I.2), which we denote by $h^{\lambda,\lambda}$. Unlike the Lasry–Lions envelope, the proximal hull is not necessarily smooth.

The structure of this work is as follows. In Section II, we briefly discuss a number of properties of the Lasry–Lions double envelope as well as some examples of interest. In Section III, we study the homotopy approach that makes use of the Lasry–Lions envelope, and discuss a practical implementation of it. In Section IV, we present numerical experiments in signal decoding and spectral unmixing. Section V concludes.

II. THE LASRY–LIONS DOUBLE ENVELOPE

We start by stating our assumptions for (I.1) and then list some results concerning the Lasry–Lions double envelope: Facts II.1 and II.2 are concerned with some of its basic identities and properties, and Proposition II.3 and Corollary II.4 with its smoothness characteristics. We finish the section with two examples of double envelopes: the indicator function of the set $\{-1, 1\}$ in Example II.5 and the ℓ_0 pseudo-norm in Example II.6; we illustrate the latter in Figure 1.

A detailed account of the notions used throughout this work can be found in [8]. Bold lowercase letters denote vectors, bold uppercase letters denote matrices, $[\mathbf{a}]_i$ denotes the i -th element of a vector \mathbf{a} , and $[\mathbf{A}]_{:j}$ denotes the j -th column of a matrix \mathbf{A} . We let $\|\cdot\| := \frac{1}{2}\|\cdot\|^2$.

Assumption I. In problem (I.1), the following hold:

- A1 $g \in C^1(\mathbb{R}^n)$;
- A2 $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is proper, lsc, and prox-bounded;
- A3 a solution exists: $\arg \min \varphi \neq \emptyset$.

Fact II.1 (Basic identities of the Lasry–Lions envelope). *The following hold for $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ and $\lambda \geq \mu > 0$:*

- (i) $h^{\lambda,\mu} = (h^{\lambda,\lambda})^{\lambda-\mu} = (h^{\lambda-\mu})^{\mu,\mu}$;
- (ii) $h^\lambda = \lambda^{-1} \dot{j} - (h + \lambda^{-1} \dot{j})^*(\lambda^{-1} \cdot)$;
- (iii) $h^{\lambda,\lambda} = (h + \lambda^{-1} \dot{j})^{**} - \lambda^{-1} \dot{j}$.

When $\lambda > \mu$, denoting $c := \frac{\lambda(\lambda-\mu)}{\mu}$ it also holds that

- (iv) $h^{\lambda,\mu} = (\lambda - \mu)^{-1} \dot{j} - [(h + \lambda^{-1} \dot{j})^*]^{1/c}(\frac{\cdot}{\lambda-\mu})$;
- (v) $h^{\lambda,\mu} = [(h + \lambda^{-1} \dot{j})^{**}]^c(\frac{\lambda}{\mu} \cdot) - \mu^{-1} \dot{j}$.

Detail. II.1(i)–II.1(iii): follow from [8, Ex.s 1.46, 11.26(c)]; II.1(iv), II.1(v): from II.1(i)–II.1(iii). \square

Fact II.2 (Basic properties of the Lasry–Lions envelope). *Let $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be proper, lsc, and γ_h -prox-bounded. Then, $h^{\lambda,\mu}$ is proper and lsc for every $0 < \mu \leq \lambda < \gamma_h$. Moreover*

- (i) $h^\lambda \leq h^{\lambda,\mu} \leq h^{\lambda-\mu} \leq h$;
- (ii) $h^{\lambda,\mu}$ is pointwise increasing wrt μ and decreasing both wrt λ and $\lambda - \mu$ (i.e., $h^{\lambda,\mu} \geq h^{\lambda',\mu'}$ when $\lambda - \mu \leq \lambda' - \mu'$);
- (iii) for every $\bar{\mathbf{x}} \in \mathbb{R}^n$, $h^{\lambda,\mu}(\mathbf{x}) \rightarrow h(\bar{\mathbf{x}})$ as $\mathbf{x} \rightarrow \bar{\mathbf{x}}$ and $0 < \mu \leq \lambda \rightarrow 0$ in such a way that $\frac{\|\mathbf{x} - \bar{\mathbf{x}}\|}{\lambda}$ is bounded;
- (iv) $\text{dom } h^{\lambda,\lambda} = \text{conv dom } h$;
- (v) $\inf h^{\lambda,\mu} = \inf h$ and $\arg \min h^{\lambda,\mu} = \arg \min h$;
- (vi) let $h : \mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_m} \rightarrow \overline{\mathbb{R}}$ be of the form $h = \sum_{k=1}^m h_k(\mathbf{x}_k)$, where $\mathbf{x}_k \in \mathbb{R}^{n_k}$ for all $k \in \{1, \dots, m\}$; then $h^{\lambda,\mu}(\mathbf{x}) = \sum_{k=1}^m h_k^{\lambda,\mu}(\mathbf{x}_k)$.

Detail. II.2(i), II.2(v): follow from [8, 1.46]; II.2(ii)–II.2(v): from [8, 1.25, 1.44] and (II.1); II.2(vi): from (I.2). \square

Proposition II.3 (Smoothness of the Lasry–Lions envelopes). *Let $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be proper, lsc, and γ_h -prox-bounded. Then, for every $0 < \mu < \lambda < \gamma_h$ the Lasry–Lions envelope $h^{\lambda,\mu}$ is $L_{h^{\lambda,\mu}}$ -Lipschitz-continuously differentiable with gradient*

$$\nabla h^{\lambda,\mu} = \frac{\lambda}{\mu} P - \frac{1}{\mu} \text{id}, \quad (\text{II.1})$$

where $P := \text{prox}_{c^{-1}(h + \lambda^{-1} \dot{j})^*}(\frac{\cdot}{\lambda-\mu})$ and $c := \frac{\lambda(\lambda-\mu)}{\mu}$ is as in Fact II.1. In fact, for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ it holds that

$$\sigma_{h^{\lambda,\mu}} \|\mathbf{x} - \mathbf{y}\|^2 \leq \langle \nabla h^{\lambda,\mu}(\mathbf{x}) - \nabla h^{\lambda,\mu}(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq -\sigma_{-h^{\lambda,\mu}} \|\mathbf{x} - \mathbf{y}\|^2$$

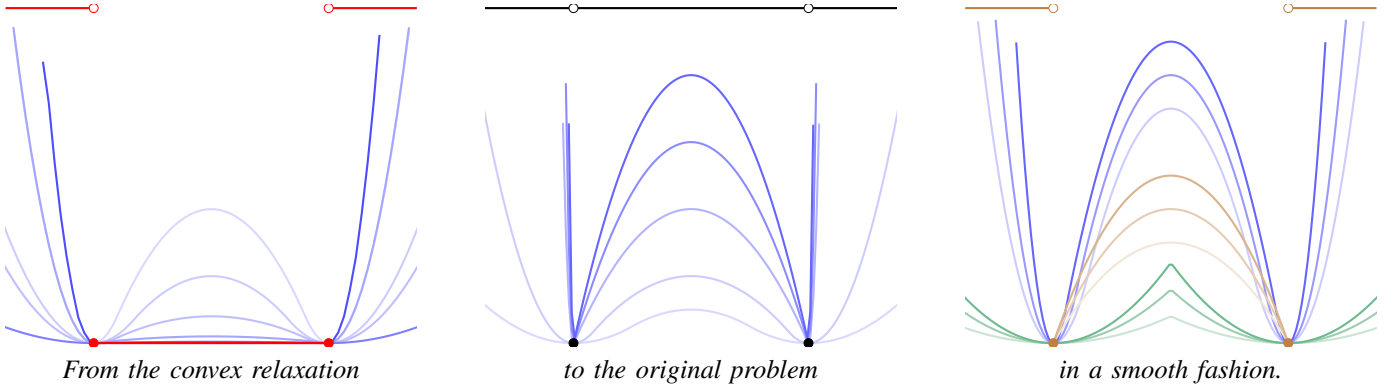
for some $\sigma_{h^{\lambda,\mu}} \geq \frac{1}{\mu}$ and $\sigma_{-h^{\lambda,\mu}} \geq \frac{-1}{\lambda-\mu}$, and in particular $L_{h^{\lambda,\mu}} \leq \max\{\frac{1}{\mu}, \frac{1}{\lambda-\mu}\}$. Moreover, the estimates can be tightened as

- (i) $\sigma_{h^{\lambda,\mu}} \geq \frac{\sigma_h}{1+(\lambda-\mu)\sigma_h}$ and $L_{h^{\lambda,\mu}} \leq \max\{\frac{|\sigma_h|}{1+(\lambda-\mu)\sigma_h}, \frac{1}{\lambda-\mu}\}$ if h is σ_h -hypoconvex;
- (ii) $\sigma_{-h^{\lambda,\mu}} \geq -\frac{L_h}{1+(\lambda-\mu)L_h}$ and $L_{h^{\lambda,\mu}} \leq \frac{L_h}{1+(\lambda-\mu)L_h}$ if h is L_h -smooth.

Proof. The expression for the gradient comes from the identity in Fact II.1(iv) together with the known formula $\nabla \phi^{1/c} = c(\text{id} - \text{prox}_{\phi/c})$, holding for the proper, lsc, convex function $\phi = (h + \lambda^{-1} \dot{j})^*$. In turn, the general bounds on the inner product follow from the monotonicity and nonexpansiveness of $\text{prox}_{\phi/c}$. When h is σ_h -hypoconvex, then ϕ is L_ϕ -smooth (and $(-L_\phi)$ -hypoconvex) with $L_\phi = \frac{\lambda}{1+\lambda\sigma_h}$, and consequently $\text{prox}_{\phi/c}$ is $(1 + c^{-1}L_\phi)^{-1}$ -strongly monotone [9, Lem. 5]. Similarly, when h is L_h -smooth (hence $(-L_f)$ -hypoconvex), then ϕ is σ_ϕ -strongly convex with $\sigma_\phi = \frac{\lambda}{1+\lambda L_h}$, and $\text{prox}_{\phi/c}$ is thus $(1 + c^{-1}\sigma_\phi)$ -contractive. By using these estimates, the tighter values as in the statement are obtained. \square

Corollary II.4. *For $\mathbf{x} \in \mathbb{R}^n$ and $0 < \mu < \lambda < \gamma_h$ it holds that*

$$\mathbf{d} = \nabla h^{\lambda,\mu}(\mathbf{x}) \iff \mathbf{x} - (\lambda - \mu)\mathbf{d} \in \text{conv prox}_{\lambda h}(\mathbf{x} + \mu\mathbf{d}).$$



From the convex relaxation
As $\lambda > \mu \rightarrow \infty$ with $\lambda - \mu \rightarrow 0$, $h^{\lambda, \mu}$ converges to the convex hull of the original function $h^{**} = \delta_{[-1,1]}$ (red).

to the original problem
As $0 < \mu < \lambda \rightarrow 0$, $h^{\lambda, \mu}$ converges to the original function $h = \delta_{\{\pm 1\}}$ (black).

in a smooth fashion.
Approximations of extended-real proximal hulls, continuous Moreau envelopes, and smooth Lasry-Lions envelopes.

Figure 1: Approximations with Lasry-Lions envelopes $h^{\lambda, \mu}$ (blue), proximal hulls $h^{\lambda, \lambda}$ (brown), and Moreau envelopes h^λ (green) of the nonconvex, nonsmooth, extended-real-valued function $h = \delta_{\{\pm 1\}}$ (black).

Proof. Denoting $c := \frac{\lambda(\lambda-\mu)}{\mu}$, it follows from (II.1) that $\mathbf{d} = \nabla h^{\lambda, \mu}(\mathbf{x})$ holds iff

$$\begin{aligned} \frac{\mu \mathbf{d} + \mathbf{x}}{\lambda} &= \text{prox}_{c^{-1}(h + \lambda^{-1}j)^*} \left(\frac{\mathbf{x}}{\lambda - \mu} \right) \\ &\Leftrightarrow c \left(\frac{\mathbf{x}}{\lambda - \mu} - \frac{\mu \mathbf{d} + \mathbf{x}}{\lambda} \right) \in \partial(h + \lambda^{-1}j)^* \left(\frac{\mu \mathbf{d} + \mathbf{x}}{\lambda} \right), \end{aligned}$$

where the last equivalence owes to convexity of $(h + \lambda^{-1}j)^*$. It follows from Fact II.1(ii) that the latter subdifferential expands to $\mu \mathbf{d} + \mathbf{x} + \lambda \partial(-h^\lambda)(\mu \mathbf{d} + \mathbf{x})$, which in turn equals $\text{conv prox}_{\lambda h}(\mathbf{x} + \mu \mathbf{d})$ by virtue of [8, Ex. 10.32]. A direct computation reveals that the left-hand side of the inclusion is $\mathbf{x} - (\lambda - \mu)\mathbf{d}$. \square

Example II.5. Consider $h : \mathbb{R} \rightarrow \mathbb{R} : x \mapsto \delta_{\{\pm 1\}}(x)$. Then, its Moreau envelope and proximal hull are

$$h^\lambda(x) = \frac{(1 - |x|)^2}{2\lambda}, \quad h^{\lambda, \lambda}(x) = \begin{cases} +\infty & |x| > \mu/\lambda \\ \frac{1-x^2}{2\lambda} & \text{otherwise,} \end{cases}$$

and its Lasry-Lions envelope and respective gradient are

$$h^{\lambda, \mu}(x) = \begin{cases} \frac{(1-|x|)^2}{2(\lambda-\mu)} & |x| > \mu/\lambda \\ \frac{1}{2\lambda} - \frac{x^2}{2\mu} & \text{otherwise,} \end{cases} \quad \nabla h^{\lambda, \mu}(x) = \begin{cases} \frac{x - \text{sgn } x}{\lambda - \mu} & |x| > \mu/\lambda \\ -\frac{x}{\mu} & \text{otherwise.} \end{cases}$$

Additionally, from Fact II.2(vi), for $h : \mathbb{R}^n \rightarrow \mathbb{R} : \mathbf{x} \mapsto \delta_{\{\pm 1\}^n}(\mathbf{x})$ we have $h^{\lambda, \mu}(\mathbf{x}) = \sum_{k=1}^n \delta_{\{\pm 1\}}^{\lambda, \mu}(|\mathbf{x}|_k)$. \square

Example II.6. Consider $h : \mathbb{R} \rightarrow \mathbb{R} : x \mapsto |x|_0 = \begin{cases} 1, & x \neq 0 \\ 0, & x = 0. \end{cases}$ Then,

$$\begin{aligned} h^\lambda(x) &= \begin{cases} \frac{1}{2\lambda} x^2 & \text{if } |x| \leq \sqrt{2\lambda}, \\ 1 & \text{if } |x| \geq \sqrt{2\lambda}, \end{cases} \\ h^{\lambda, \lambda}(x) &= \begin{cases} 1 - \frac{1}{2\lambda} (|x| - \sqrt{2\lambda})^2 & \text{if } |x| \leq \sqrt{2\lambda}, \\ 1 & \text{if } |x| \geq \sqrt{2\lambda}, \end{cases} \\ h^{\lambda, \mu}(x) &= \begin{cases} \frac{1}{2(\lambda-\mu)} x^2 & \text{if } |x| \leq (1 - \mu/\lambda) \sqrt{2\lambda}, \\ 1 - \frac{1}{2\mu} (|x| - \sqrt{2\lambda})^2 & \text{if } |x| \in [(1 - \mu/\lambda) \sqrt{2\lambda}, \sqrt{2\lambda}], \\ 1 & \text{if } |x| \geq \sqrt{2\lambda}, \end{cases} \end{aligned}$$

and

$$\nabla h^{\lambda, \mu}(x) = \begin{cases} \frac{1}{(\lambda-\mu)} x & \text{if } |x| \leq (1 - \mu/\lambda) \sqrt{2\lambda}, \\ -\frac{1}{\mu} (|x| - \sqrt{2\lambda}) & \text{if } |x| \in [(1 - \mu/\lambda) \sqrt{2\lambda}, \sqrt{2\lambda}], \\ 0 & |x| \geq \sqrt{2\lambda}. \end{cases}$$

For $h : \mathbb{R}^n \rightarrow \mathbb{R} : \mathbf{x} \mapsto \|\mathbf{x}\|_0$, $h^{\lambda, \mu}(\mathbf{x}) = \sum_{k=1}^n \delta_{\{\pm 1\}}^{\lambda, \mu}(|\mathbf{x}|_k)^{\lambda, \mu}$. \square

III. A HOMOTOPY APPROACH TO NONCONVEX MINIMIZATION

This section describes a new method to address nonsmooth nonconvex composite problems. We state global convergence properties and then discuss a practical implementation of it.

The approximating properties of the Lasry-Lions envelope naturally lead one to consider homotopy approaches to address (I.1). When Assumption I.A2 is satisfied, one has that

$$\varphi(\mathbf{x}) = \lim_{\mu < \lambda \rightarrow 0} \{g(\mathbf{x}) + h^{\lambda, \mu}(\mathbf{x})\}$$

holds for every $\mathbf{x} \in \mathbb{R}^n$, and $h^{\lambda, \mu}$ is a Lipschitz-differentiable function whenever $0 < \mu < \lambda < \gamma_h$. In what follows, it will be assumed without mention that λ and μ comply with these bounds. By replacing the regularizer in (I.1) with its Lasry-Lions envelope, we have:

$$\underset{\mathbf{x} \in \mathbb{R}^n}{\text{minimize}} \varphi_{\lambda, \mu}(\mathbf{x}) := g(\mathbf{x}) + h^{\lambda, \mu}(\mathbf{x}).$$

Theorem III.1. Additionally to Assumption I, suppose that there exist $0 < \bar{\mu} < \bar{\lambda} < \gamma_h$ such that $\varphi_{\bar{\lambda}, \bar{\mu}}$ is level bounded. Let $\lambda_k, \mu_k > 0$ satisfy $\bar{\lambda} \geq \lambda_k > \mu_k \leq \bar{\mu}$ with $\lambda_k \rightarrow 0$ and $\lambda_k - \mu_k \searrow 0$ (so that $\varphi_{\lambda_k, \mu_k} \nearrow \varphi$). The following hold:

- (i) $\min \varphi_{\lambda_k, \mu_k} \rightarrow \min \varphi$ as $k \rightarrow \infty$;
- (ii) any sequence $\mathbf{x}_k^* \in \varepsilon_k\text{-arg min } \varphi_{\lambda_k, \mu_k}$ with $\varepsilon_k \rightarrow 0$ is bounded and has all its cluster points in $\text{arg min } \varphi$.

Proof. Facts II.2(ii) and II.2(iii) ensure through [8, Prop. 7.4(d)] that $\varphi_{\lambda_k, \mu_k}$ epi-converges to φ as $k \rightarrow \infty$. In turn, owing to level boundedness the proof follows by invoking [8, Ex. 7.32(a) and Thm. 7.33]. \square

A conceptual method informed by these results is given in Algorithm 1. For ease of implementation, the ε_k -minimality prescribed in Theorem III.1 is here replaced by approximate stationarity, so that any smooth minimization algorithm can conveniently be employed. Extending Theorem III.1 to account for this relaxation is planned for future work, possibly by mimicking arguments from [10]–[12] made in the context of augmented Lagrangian methods.

In more detail, for the experiments detailed in Section IV, we made the following choices: we let $\tau^1 = 10^{-3}$; Step 1.2 can

Algorithm 1 A homotopy approach to address (I.1) through the use of Lasry–Lions envelopes.

Choose $\mathbf{x}^0 \in \mathbb{R}^n$ and sequences $\tau_k, \lambda_k, \mu_k > 0$ such that $\tau_k \searrow 0$ and $0 < \mu_k < \lambda_k \searrow 0$ with $\lambda^k - \mu^k \searrow 0$
Set $k = 1$
1.1: WHILE stopping criterion is not satisfied DO
1.2: Starting at \mathbf{x}^{k-1} , use a descent method to find \mathbf{x}^k such that $\|\nabla \varphi_{\lambda_k, \mu_k}(\mathbf{x}^k)\| \leq \tau^k$
1.3: $k \leftarrow k + 1$

be implemented by any smooth minimization algorithm, and we used L-BFGS due to its fast properties and low memory requirements; finally, we updated the tolerance as $\tau^{k+1} = 0.9\tau^k$.

IV. NUMERICAL EXPERIMENTS

We discuss a number of experiments performed to evaluate the proposed method of Section III in two problems: signal decoding and spectral unmixing. We make use of the nonconvex functions discussed in Examples II.5 and II.6 to formulate these problems.

A. Signal decoding

We approach the problem of binary-signal decoding through a constrained least-squares formulation:

$$\underset{\mathbf{x} \in \mathbb{R}^P}{\text{minimize}} \quad \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2, \quad \text{subject to} \quad \mathbf{x} \in \{0, 1\}^P, \quad (\text{IV.1})$$

where $\mathbf{x} \in \mathbb{R}^P$, $\mathbf{y} \in \mathbb{R}^N$ are the received and transmitted signals, respectively, and $\mathbf{H} \in \mathbb{R}^{N \times P}$ is the transfer matrix of a given channel. Since this problem's constraint is nonconvex, a convex alternative to this problem is the one given by replacing the constraint with $\mathbf{x} \in [0, 1]^P$. This alternative generates estimates that are not guaranteed to be binary numbers, so they are usually projected into the original problem's feasible set. We randomly generated a number of experiments with varying problem dimensionality (both over- and under-determined formulations), noise levels, and conditioning of the matrix \mathbf{H} . We compared the proposed method with an alternating-direction method of multipliers (ADMM) implementation, with an ADMM implementation of the aforementioned convex relaxation, and with a simple least-squares approach to the problem (i.e., without constraints); we denote the four methods by LL, AN, AR, and LS, respectively. We evaluated their performance by computing the binary-error rate (BER) relative to the transmitted signal. Overall, we found that the proposed method performed similarly or better than the others when $N < P$, and we present the performance of the three methods over examples of such settings in Table I. In more detail, the problem was set as follows: we started by generating the transmitted signal assuming that it was drawn i.i.d. from a Bernoulli distribution, with equal probability of generating both symbols; we then generated the matrix \mathbf{H} assuming that its rows were drawn i.i.d. from a multivariate Gaussian distribution with zero mean and covariance matrix $\Sigma \in \mathbb{R}^{P \times P}$, where $[\Sigma]_{ij} = \rho^{|i-j|}$; finally, we added i.i.d. noise drawn from a Gaussian distribution such that its variance resulted in a given signal-to-noise ratio (SNR). We tested the

Table I: Average BER [%] for estimates of the four methods.

N	P	ρ	SNR [dB]	LS	AR	AN	LL
20	40	0.0	30	47.15	11.35	32.30	15.40
			20	47.45	14.05	33.30	17.20
			10	47.35	21.90	31.30	21.90
		0.5	30	45.75	12.70	32.15	16.50
			20	46.20	13.95	30.85	16.75
			10	47.90	27.80	34.20	27.20
	80	0.0	30	50.00	37.45	40.58	35.95
			20	49.02	35.95	39.42	34.12
			10	49.83	36.12	39.92	35.23
		0.5	30	48.05	34.27	38.35	33.25
			20	49.48	35.73	39.67	34.50
			10	49.08	37.38	39.92	37.27
	100	0.0	30	48.84	37.86	41.26	37.82
			20	49.24	39.24	42.02	39.76
			10	49.54	40.56	43.36	39.98
		0.5	30	49.94	40.44	42.36	40.00
			20	49.38	39.54	41.68	38.28
			10	48.60	39.20	41.56	38.56

three methods by running the experiment 50 times; at the end, we projected all methods' estimates into the set $\{0, 1\}^P$, i.e., we made $[\mathbf{x}]_i^{\text{proj}} = 0$ if $[\mathbf{x}]_i^{\text{est}} < 0.5$ and $[\mathbf{x}]_i^{\text{proj}} = 1$ if $[\mathbf{x}]_i^{\text{est}} \geq 0.5$ for $i = 1, \dots, P$, where \mathbf{x}^{est} denotes a given method's estimate and \mathbf{x}^{proj} denotes its projection. The algorithms' parameters were tuned as follows: for LL, we let $\lambda^1 = 10^5$ and $\mu^1 = 0.999 \times \lambda^1$, and reduced these parameters by 90% every iteration; in Section III, for AR and AN, we followed the heuristic proposed in [13] to tune step sizes. The initial estimate for AN, AN, and LL was the LS estimate, and we stopped them either if the root-mean-squared error (RMSE) relative to the transmitted signal was lower than 10^{-9} or after 10^3 iterations. The experiments were conducted using MATLAB on a machine with a quad-core Intel Core i7 CPU running at 2.5 GHz and with 16 GB of DDR3L RAM.

B. Spectral unmixing

Hyperspectral images are multi-channel images with a relatively large number of channels—usually known as spectral bands—corresponding to short frequency ranges along the electromagnetic spectrum. Frequently, their spatial resolution is low, and it is of interest to disentangle the different spectral components of a given pixel; a pixel typically corresponds to a mixture of different materials. Spectral unmixing techniques produce a set of spectral profiles, one for each material (known as endmember), and a corresponding set of abundances, or percentages of occupation, for each endmember, in each pixel [14]. We consider that the set of spectral profiles is known through a database of spectral signatures (i.e., a database of reflectance profiles as a function of wavelength), and formulate the spectral unmixing problem pixel-wise as an instance of the Lagrangian formulation of the best-subset-selection problem:

$$\underset{\mathbf{a}_j \in \mathbb{R}^P}{\text{minimize}} \quad \|[\mathbf{Y}_h]_{:,j} - \mathbf{U}\mathbf{a}_j\|_2^2 + \beta \|\mathbf{a}_j\|_0, \quad (\text{IV.2})$$

where $\mathbf{a}_j \in \mathbb{R}^P$ is the vector of each endmember's abundances for a given pixel j , to be estimated, $\mathbf{U} \in \mathbb{R}^{N \times P}$ is a matrix corresponding to the spectral database, $\mathbf{Y}_h \in \mathbb{R}^{N \times M}$ corresponds to a matrix representation of a hyperspectral image with M pixels (i.e., corresponds to the lexicographical ordering of a 3-D data cube), and β is a regularization parameter. A typical convex

Table II: Average RMSE, sensitivity [%], specificity [%], and the value $[\times 10^{-3}]$ of the cost function (CF) for varying λ^1 in the LL method.

λ^1	10^4	10^3	10^2	10	1	10^{-1}	10^{-2}
RMSE	0.035	0.032	0.034	0.047	0.056	0.044	0.038
Sens.	38	44	48	54	62	74	88
Spec.	95.89	94.70	90.36	82.10	67.95	52.60	37.08
CF	0.468	0.445	0.432	0.424	0.426	0.464	0.513

formulation of this problem is as the least absolute shrinkage and selection operator (LASSO), where $\|\mathbf{a}_j\|_0$ is replaced by $\|\mathbf{a}_j\|_1$. Constraints such as $[\mathbf{a}_j]_i \geq 0$ for $i = 1, \dots, P$ are also adopted but, for the comparison purposes of this work, we ignored them. We randomly generated a number of experiments with varying problem dimensionality (both over- and underdetermined formulations) and noise levels. We used a real-world spectral dictionary \mathbf{U} : a selection of 498 different mineral types from a USGS library, set up as detailed in [15]. We compared the proposed method to an ADMM implementation, and we denote the two methods by LL and AN, respectively. We evaluated them in two fronts: we computed the RMSE relative to the original vector of abundances, and we evaluated the quality of the estimates' support (i.e., the location of its nonzero elements). The latter was evaluated by computing the sensitivity and specificity of the estimates as follows: we considered a correctly estimated nonzero as a true positive (TP), and an incorrectly assigned one as a false positive (FP); conversely, a correctly estimated zero was considered as a true negative (TN), and an incorrectly assigned one as a false negative (FN); then, the sensitivity is given by $\frac{TP}{TP+FN}$ and the specificity by $\frac{TN}{FP+TN}$. A box plot of the RMSE, sensitivity, and specificity for the two methods for one problem setting is given in Figure 2. This problem was generated as follows: we start by generating a vector of abundances with $P = 224$ and with 5 nonzero elements, where the abundances are drawn from a Dirichlet distribution; we made $N = P$ and added Gaussian noise such that it would result in a SNR of 30 dB. The parameters were tuned as before except that, for LL, the initial points are $\lambda^1 = 10^3$ and $\mu^1 = \frac{\lambda^1}{2}$, and we reduced them by 10% and 82% every iteration, respectively; the initial estimate was considered to be a vector of zeros and the regularization parameter was manually set to $\beta = 10^{-6}$, which we found produced estimates with a support similar to the simulated one. The proposed method, in virtue of its tuning parameters, allows for some flexibility in how "aggressively" it selects nonzero elements. In Table II, and for the problem setting just described, we show how a progressively lower λ^1 —with still making $\mu^1 = \frac{\lambda^1}{2}$ and reducing those as indicated—finds estimates with very different sensitivity and specificity.

V. CONCLUSIONS

We introduced a novel method to address nonsmooth non-convex composite-minimization problems, based on Lasry–Lions double envelopes. This method has very broad applications, and it seems suitable for settings of great practical interest in a number of signal processing problems, as suggested by experiments. When compared to classical methods,

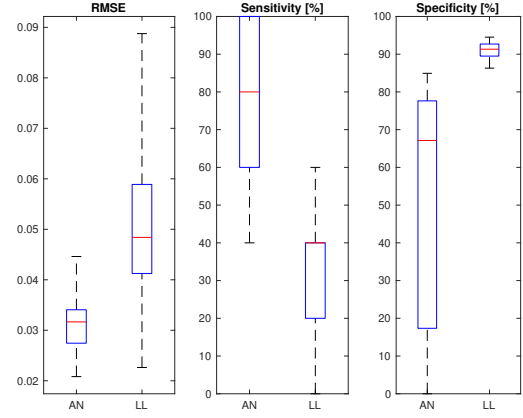


Figure 2: Average RMSE, sensitivity, and specificity.

the tuning of its inner parameters allows for greater flexibility in the choice of the type of solutions that are sought. In future work, we will study techniques to automatically tune these parameters.

REFERENCES

- [1] E. Candès and Y. Plan, "A probabilistic and RIPless theory of compressed sensing," *IEEE Trans Inf Theory*, vol. 57, no. 11, pp. 7235–7254, Nov 2011.
- [2] D. Bertsimas, A. King, and R. Mazumder, "Best subset selection via a modern optimization lens," *Ann Stat*, vol. 44, no. 2, pp. 813–852, 2016.
- [3] T. Hastie, R. Tibshirani, and R. Tibshirani, "Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons," *Stat Sci*, vol. 35, no. 4, pp. 579–592, 2020.
- [4] H. Hazimeh and R. Mazumder, "Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms," *Oper Res*, vol. 68, no. 5, pp. 1517–1537, Oct. 2020.
- [5] E. Soubies, L. Blanc-Féraud, and G. Aubert, "A unified view of exact continuous penalties for ℓ_2 – ℓ_0 minimization," *SIAM J Optim*, vol. 27, no. 3, pp. 2034–2060, 2017.
- [6] M. Carlsson, "On convex envelopes and regularization of non-convex functionals without moving global minima," *J Optim Th Appl*, vol. 183, no. 1, pp. 66–84, 2019.
- [7] C. Grussler, A. Rantzer, and P. Giselsson, "Low-rank optimization with convex constraints," *IEEE Trans Autom Contr*, vol. 63, no. 11, pp. 4000–7, 2018.
- [8] R. Rockafellar and R. Wets, *Variational Analysis*. Springer, 2009.
- [9] A. Themelis, B. Hermans, and P. Patrinos, "A new envelope function for nonsmooth DC optimization," in *Proc IEEE CDC*, 2020, pp. 4697–4702.
- [10] E. Birgin and J. Martínez, *Practical Augmented Lagrangian Methods for Constrained Optimization*. SIAM, 2014.
- [11] G. Grapiglia and Y. Yuan, "On the complexity of an augmented Lagrangian method for nonconvex optimization," *IMA J Numer Anal*, 2020.
- [12] B. Evens, P. Latafat *et al.*, "Neural network training as an optimal control problem: An augmented Lagrangian approach," *ArXiv:2103.14343*, 2021.
- [13] S. Boyd, N. Parikh *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found Trends Mach Learn*, vol. 3, no. 1, pp. 1–122, 2011.
- [14] W.-K. Ma, J. Bioucas-Dias *et al.*, "A signal processing perspective on hyperspectral unmixing: Insights from remote sensing," *IEEE Signal Process Mag*, vol. 31, no. 1, pp. 67–81, 2014.
- [15] J. Bioucas-Dias and M. Figueiredo, "Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing," in *WHISPERS*, Reykjavik, Iceland, 2010, pp. 1–4.