NONLINEAR ACCELERATION OF CONSTRAINED OPTIMIZATION ALGORITHMS

Vien V. Mai and Mikael Johansson

KTH Royal Institute of Technology, School of Electrical Engineering and Computer Science
SE-100 44 Stockholm, Sweden.
Emails: {maivv,mikaelj}@kth.se.

ABSTRACT

This paper introduces a novel technique for nonlinear acceleration of first-order methods for constrained convex optimization. Previous studies of nonlinear acceleration have only been able to provide convergence guarantees for unconstrained convex optimization. In contrast, our method is able to avoid infeasibility of the accelerated iterates and retains the theoretical performance guarantees of the unconstrained case. We focus on Anderson acceleration of the classical projected gradient descent (PGD) method, but our techniques can easily be extended to more sophisticated algorithms, such as mirror descent. Due to the presence of a constraint set, the relevant fixed-point mapping for PGD is not differentiable. However, we show that the convergence results for Anderson acceleration of smooth fixed-point iterations can be extended to the non-smooth case under certain technical conditions.

Index Terms— Anderson acceleration, constrained optimization, projected gradient descent, semi-smoothness

1. INTRODUCTION

Acceleration, i.e. the use of history information to speed up the convergence of an iterative method, is a huge topic in optimization. A notable example is momentum acceleration, which covers well-known methods such as Polyak's heavy ball method [1] and Nesterov's fast gradient method [2]. Unlike momentum acceleration methods, which require knowledge of problem parameters, classical extrapolation techniques, such as Aitken's Δ^2 and Wynn's ϵ-algorithm for scalar sequences, or Anderson acceleration (AA) [3], minimal polynomial extrapolation (MPE) [4], and reduced rank extrapolation (RRE) [5] for vector sequences, estimate the solution directly from the available sequence. The literature on these techniques is vast and we refer to [4, 6] for more comprehensive surveys.

Even though vector extrapolation methods have a wide range of applications across different scientific fields, their development is largely independent of optimization algorithms. For example, AA has been successfully applied to many problems in computational chemistry, physics, and material science, but its connection to quasi-Newton methods was only discovered recently [7]. During the last couple of years, extrapolation methods have started to attract a significant interest in the optimization community (see, e.g., [8–13]). Specifically, a series of papers [10–12] adapt the AA scheme to accelerate several classical algorithms for unconstrained optimization, while [13] extends the reach of AA to non-expansive and possibly non-smooth operators. Several empirical examples are given demonstrating great benefits of nonlinear acceleration even for non-smooth optimization problems.

Although the above-cited methods have been successfully adapted to unconstrained optimization algorithms, none of the proposed techniques are able to guarantee acceleration, or even convergence, for constrained problems. The main difficulty is that the extrapolated point in such methods may lie outside the feasible set. In this paper, we will demonstrate how this issue can be circumvented, and several popular first-order methods for constrained convex optimization can be accelerated. Under certain technical conditions, we extend the convergence results of AA in [14] for smooth fixed-point problems to the nonsmooth ones.

2. NONLINEAR ACCELERATION

In this section, we aim to find an approximate solution to the following fixed-point problem:

$$\text{Find } x \in \mathbb{R}^n \text{ such that } x = g(x), \tag{1}$$

where $g : \mathbb{R}^n \to \mathbb{R}^n$ is a contractive mapping. Let $\{x_i\}_{i=0}^k$ be a (convergent) sequence of iterates generated by the fixed-point iteration:

$$x_{k+1} = g(x_k), \quad k = 0, \ldots, k.$$

We refer the quantity $r_k = g(x_k) - x_k$ as the residual in the kth iteration. Nonlinear acceleration methods seek to combine past iterates into a new point with smaller residual. To this end, it forms

$$x_{\text{ext}} = \sum_{i=0}^{k} \alpha_i x_i \tag{2}$$

This research was sponsored in part by the Knut and Alice Wallenberg Foundation and the Swedish Research Council.

with $\alpha_i \in \mathbb{R}$. The name nonlinear acceleration comes from the fact that the optimal weights α_i are nonlinear functions of x_i. Ideally, we want to have x_{ext} that minimizes the residual among all possible linear combinations of $\{x_i\}_{i=0}^k$:

$$\alpha^\star = \underset{\alpha}{\operatorname{argmin}} \left\| g\left(\sum_{i=0}^k \alpha_i x_i\right) - \sum_{i=0}^k \alpha_i x_i \right\|. \tag{3}$$

However, solving (3) can be hard for a general nonlinear mapping g, hence we solve instead

$$\alpha^\star = \underset{\alpha:\alpha^\top \mathbf{1}=1}{\operatorname{argmin}} \left\| \sum_{i=0}^k \alpha_i g(x_i) - \sum_{i=0}^k \alpha_i x_i \right\| = \underset{\alpha:\alpha^\top \mathbf{1}=1}{\operatorname{argmin}} \left\| \sum_{i=0}^k \alpha_i r_i \right\|, \tag{4}$$

where the constraint on α ensures convergence. It can be verified that problems (3) and (4) are equivalent when g is a linear mapping. If we let $R = [r_0\ r_1\ \ldots\ r_k]$ be the matrix whose columns are the residuals r_i, then the problem (4) can be written as

$$\alpha^\star = \underset{\alpha^\top \mathbf{1}=1}{\operatorname{argmin}} \|R\alpha\|. \tag{5}$$

To keep the cost of evaluating $\alpha^\star$ small, nonlinear acceleration methods typically only use the $m+1$ most recent iterates $\{x_i\}_{i=k-m}^k$ instead of the full past history $\{x_i\}_{i=0}^k$.

In this work, we focus on the AA scheme, which is detailed in Algorithm 1. Note that AA is closely related to other vector extrapolation methods such as MPE [4] and RRE [5]. In the linear case, it has been shown in [8, 15] that AA is essentially equivalent to GMRES method [16] in a certain sense. Since the columns of R_k in Step 4 of Algorithm 1 become increasingly similar as the iterates converge, AA can be highly unstable. To stabilize the algorithm, the authors in [10] add Tikhonov regularization to the least-squares problem (5), resulting in the so-called regularized nonlinear acceleration (RNA). Moreover, the remarkable connection between AA and quasi-Newton methods shown in [7] can also suggest alternative ways to stabilize the AA scheme inspired by the rich history of quasi-Newton methods (see, e.g., [13] and the references therein).

Existing work on AA has focused on unconstrained convex optimization problems:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ f(x). \tag{6}$$

For example, the classical gradient descent (GD) method

$$x_{i+1} = x_i - \gamma \nabla f(x_i),$$

which can be seen as the fixed-point iteration of the mapping $g(x) = x - \gamma\nabla f(x)$. Clearly, a fixed-point of g corresponds to an optimal solution of (6). If f is strongly convex with a Lipschitz continuous gradient, the authors in [10] shown that

Algorithm 1 Anderson Acceleration

Input: x_0, $m \geq 1$

1: $x_1 \leftarrow g(x_0)$
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: $\quad m_k \leftarrow \min(m, k)$
4: $\quad R_k \leftarrow [r_{k-m_k},\ \ldots,\ r_k]$, where $r_i = g(x_i) - x_i$
5: $\quad \alpha^k \leftarrow \operatorname{argmin}_{\alpha^\top \mathbf{1}=1} \|R_k\alpha\|$
6: $\quad x_{k+1} \leftarrow \sum_{i=0}^{m_k} \alpha_i^k g(x_{k-m_k+i})$
7: **end for**

Output: x_K

the RNA scheme (corresponding to AA with $m = \infty$) applied to GD achieves the same convergence rate as the Nesterov's fast gradient descent method, when initialized near the optimal solution. Several extensions have been made recently to handle stochastic gradient methods [11], algorithms with momentum terms such as Nesterov's accelerated gradient, and primal dual methods [12]. It is worth noting that the performance guarantees of RNA requires differentiability of g, which rules out many non-smooth methods such as projection and proximal-type methods. Very recently, the work in [13] extends AA to nonexpansive and possibly non-smooth operators, thereby covering unconstrained proximal-type methods. However, the case with a convex constraint is still left open. The main difficulty is that the extrapolated point may lie outside the feasible set. In the next section, we will show how to use nonlinear acceleration to speed-up the classical projected gradient method in solving constrained convex problems.

3. NONLINEAR ACCELERATION FOR CONSTRAINED OPTIMIZATION

Consider a generic constrained convex optimization problem:

$$\underset{x\in\mathcal{C}}{\text{minimize}}\ f(x), \tag{7}$$

where f is a proper, closed convex function with nonempty interior domain and $\mathcal{C}$ is a closed convex set. A the classical method for solving (7) is projected gradient descent (PGD), where each iteration consists of a gradient step followed by an orthogonal projection onto $\mathcal{C}$. Let $x_0 \in \mathcal{C}$ and let γ be a positive stepsize, then PGD iteration reads:

$$y_{k+1} = x_k - \gamma\nabla f(x_k) \tag{8}$$

$$x_{k+1} = \Pi_{\mathcal{C}}(y_{k+1}). \tag{9}$$

The key idea behind this algorithm is summed up by the following proposition:

Proposition 1. *Let f be a proper closed and convex function and let $\mathcal{C}$ be a closed convex set satisfying $\mathcal{C} \subseteq \operatorname{int}(\operatorname{dom}(f))$. Then, $x^\star$ is an optimal solution to* (7) *if and only if*

$$x^\star = \Pi_{\mathcal{C}}(x^\star - \gamma\nabla f(x^\star)). \tag{10}$$

The proposition implies that the PGD algorithm can be seen as a fixed-point iteration for the mapping $g(x) = \Pi_{\mathcal{C}}(x - \gamma\nabla f(x))$. However, as mentioned before, naively using nonlinear acceleration for this mapping may render iterates infeasible.

Our method hinges on the following simple observation:

$$\begin{aligned}\|x_{k+1} - x^\star\| &= \|\Pi_{\mathcal{C}}(y_{k+1}) - \Pi_{\mathcal{C}}(x^\star - \gamma\nabla f(x^\star))\| \\ &\leq \|y_{k+1} - y^\star\|,\end{aligned}$$

where $y^\star = x^\star - \gamma\nabla f(x^\star)$ and the inequality follows from the nonexpansiveness of the projection operator. The inequality suggests that if one can quickly drive $\|y_k - y^\star\|$ to zero, then the convergence of $\{x_k\}$ will automatically follow. This motivates us to study the following mapping:

$$g(y) = \Pi_{\mathcal{C}}(y) - \gamma\nabla f(\Pi_{\mathcal{C}}(y)). \tag{11}$$

Evidently, if y is a fixed-point of $g(y)$, then $x = \Pi_{\mathcal{C}}(y)$ is an optimal solution to (7) since it satisfies condition (10). Thus, one can instead focus on finding a fixed-point of g defined in (11). Note that the fixed-point iteration $y_{k+1} = g(y_k)$ is exactly the PGD iteration in (8)–(9).

In short, we propose to use nonlinear acceleration for the auxiliary sequence $\{y_k\}$ instead of the primal sequence $\{x_k\}$. This is important since $\{y_k\}$ are not restricted to the constraint set $\mathcal{C}$, hence avoiding the feasibility issue. We emphasize that extending the proposed scheme to other popular methods such as mirror descent is rather straightforward. Due to the limited space, we substantiate this claim in a more complete version of the current paper.

4. CONVERGENCE GUARANTEE

We now discuss the theoretical performance guarantee of the proposed scheme. So far, all convergence rate guarantees for AA rely on linearizing the mapping g around a fixed-point $x^\star$:

$$g(x) = g(x^\star) + G'(x^\star)\,(x - x^\star) + o(\|x - x^\star\|),$$

where $G'(x^\star)$ is the Jacobian matrix of g at $x^\star$. Due to the presence of the projection operator, the mapping g defined in (11) is, in general, non-differentiable. Therefore, the analyses in [10] and [14] are not applicable anymore. However, we observe that the proof of Theorem 2.3 in [14] essentially only needs that the bound

$$\|F(x) - F'(x^\star)(x - x^\star)\| \leq \frac{c}{2}\|x - x^\star\|^2, \tag{12}$$

holds for some constant $c > 0$ and for all x sufficiently close to $x^\star$, where $F(x) \triangleq x - g(x)$. Interestingly, this condition is very similar to assumptions which guarantee convergence of Newton's method for solving non-smooth nonlinear equations (see [17, Chapter 7] for an excellent review). Two key ingredients in the analysis of non-smooth Newton methods are the use of Clarke's generalized Jacobian [18] and the concept of *semi-smoothness* [19, 20], defined below.

Definition 1 (Semi-smoothness)**.** *Let $F : \Omega \to \mathbb{R}^n$ be a locally Lipschitz continuous function and let $\partial F(x)$ be the Clarke generalized Jacobian of F at x. We say that F is semismooth at $x \in \Omega$ if:*
i) F is directionally differentiable at x; and
ii) For any $h \in \Omega$ and $J \in \partial F(x + h)$,

$$\|F(x + h) - F(x) - Jh\| \leq o(\|h\|) \;\; as \;\; h \to 0.$$

Furthermore, F is said to be strongly semi-smooth at $x \in \Omega$ if F is semi-smooth and for any $h \in \Omega$ and $J \in \partial F(x + h)$,

$$\|F(x + h) - F(x) - Jh\| \leq O(\|h\|^2) \;\; as \;\; h \to 0.$$

If F is (strongly) semi-smooth at each point of Ω, then we say that F is (strongly) semi-smooth on Ω.

The class of semi-smooth functions is very broad and includes smooth functions, convex functions, and piecewise smooth functions. For example, differentiable functions with a Lipschitz continuous gradient are strongly semi-smooth; the norm function $\|\cdot\|_p$ is strongly semis-mooth for every $p \in [1, \infty]$; piecewise affine functions such as $\max(x, 0)$ are strongly semi-smooth [17]. Semi-smoothness is closed under scalar multiplication, summation and composition.

Proposition 2 ([17, 21])**.** *Projections onto the nonnegative orthant, second-order cone, positive semidefinite cone, and polyhedral set are all strongly semismooth.*

To extend the results in [14] to nonsmooth mappings, we impose the following assumption.

Assumption 1. *i) There is a $\rho \in (0, 1)$ such that $\|g(x) - g(y)\| \leq \rho\|x - y\|$ for all $x, y \in \mathcal{B}(x^\star, r)$ for some $r > 0$.*
ii) The mapping F is strongly semi-smooth at $x^\star$ and every $J^\star \in \partial F(x^\star)$ is nonsingular.
iii) There is M_α such that $\sum_{i=0}^{m_k} |\alpha_i^k| \leq M_\alpha$ for all $k \geq 0$.

Note that Assumptions 1-i) and 1-iii) are standard in the analysis of AA, while Assumption ii) is new. The following Theorem is the generalization of Theorem 2.3 in [14].

Theorem 1. *Let Assumption 1 hold and let $\rho < \tilde{\rho} < 1$. Then if x_0 is sufficiently close to $x^\star$, the iterates generated by Anderson acceleration converge linearly to $x^\star$:*

$$\|x_k - x^\star\| \leq \frac{1 + \rho}{1 - \rho}\tilde{\rho}^k\|x_0 - x^\star\|.$$

Theorem 1 extends the convergence result for AA of general fixed-point problems presented in [14] to a more general setting, where the relevant mapping g can be non-smooth. The result allows us to establish the following convergence guarantee for AA of PGD proposed in the previous section.

Proposition 3. *Let f be a μ-strongly convex and L-smooth function and let the stepsize $\gamma \in (0, 2/(\mu + L)]$. Suppose that*

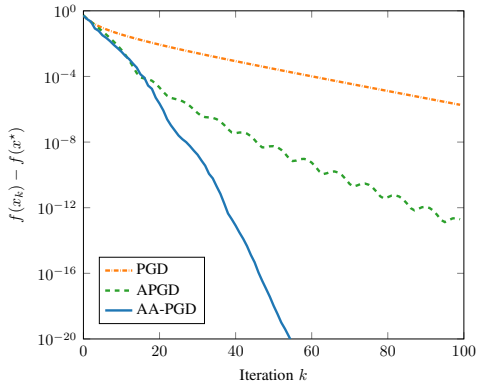

Fig. 1. Error versus the number of iterations for different algorithms for solving the nonnegative least squares problem.

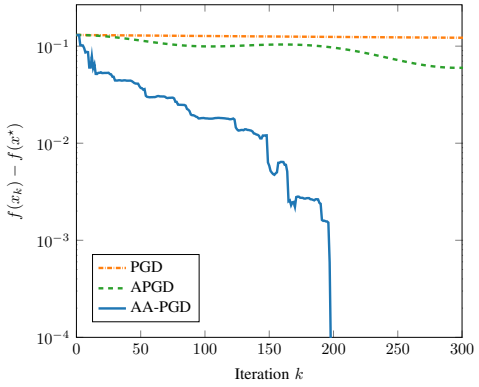

Fig. 2. Error versus the number of iterations for different algorithms for the constrained logistic regression problem.

the projection onto $\mathcal{C}$ is strongly semi-smooth and that x_0 is sufficiently close to $x^\star$. Then, for any $k \in \mathbb{N}_+$, the iterates formed by AA applied to PGD satisfy:

$$\|x_k - x^\star\| \leq \frac{1+\rho}{1-\rho}\tilde{\rho}^k \|x_0 - x^\star\|,$$

where $\rho = \sqrt{1 - \gamma 2\mu L/(\mu + L)}$ *and* $\rho < \tilde{\rho} < 1$.

We make the following remarks:

Remark 1. *The value of ρ can be easily verified using the gradient Lipschitz constant and strong convexity modulus of of f. It is also possible to prove that the mapping $F(x) = x - g(x)$ is strongly monotone, hence the nonsingularity condition of all $J^\star \in \partial F(x^\star)$ in Assumption 1-ii) is always satisfied. Finally, since $\Pi_{\mathcal{C}}(\cdot)$ is assumed to be strongly semi-smooth and since strong semi-smoothness is closed under composition, F is strongly semi-smooth.*

5. NUMERICAL RESULTS

We will now illustrate the performance of our scheme (AA-PGD) on two constrained optimization problems with many applications in signal processing and machine learning. We compare AA-PGD with the original PGD and Nesterov's accelerated projected gradient descent (APGD) method [2].

Nonnegative least squares. Here, we consider

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \frac{1}{2n}\|Ax - b\|^2 \quad \text{subject to } x \geq 0,$$

with $A \in \mathbb{R}^{1000\times 5000}$ and $b \in \mathbb{R}^{1000}$ generated by drawing their elements from a Gaussian distribution with zero mean and unit variance. We set $\gamma = 1/L$, where $L = \|A\|_2^2/n$.

Constrained logistic regression. This problem has the form

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-y_i a_i^\top x)) + \mu\|x\|^2$$
$$\text{subject to } \|x\|_\infty \leq 1,$$

where $a_i \in \mathbb{R}^d$ are training samples and $y \in \{-1, 1\}^n$ are the corresponding labels. We use the UCI Madelon dataset, which contains 2000 training samples and 500 features[1]. We set $\mu = 0.01$, compute $L = \|A\|_2^2/4n$ with $A = [a_1, \ldots, a_n]$, (i.e. the condition number is 3×10^9) and use $\gamma = 2/(L+\mu)$.

For AA-PGD, we simply set $m = 5$ and add a Tikhonov regularization of $10^{-8}\left\|R^\top R\right\|$ to (5) for stabilization, as was done in [11]. For APGD, we use the optimal combination coefficient $\beta = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$ for the second problem. Finally, $f(x^\star)$ is set to the value returned by the package `scipy.optimize`, and all algorithms are initialized at 0.[2]

Figures 1 and 2 show that AA-PGD results in dramatic performance improvements over the vanilla PGD and significantly outperforms the optimal first-order method APGD, even when APGD knows the strong convexity parameter μ. Notably, in Fig. 2, due to the large condition number of the problem, PGD and APGD make very little progress in the first 300 iterations while AA-PGD quickly finds a high accuracy solution within 200 iterations. We remark that the additional computations in AA-PGD are very marginal, hence the speed-up over PGD and APGD in number of iterations translates to a similar acceleration in wall-clock time.

6. CONCLUSIONS

We have proposed a method to accelerate popular first-order methods for constrained convex optimization using vector extrapolation techniques. The method fits nicely to the nonlinear acceleration framework without introducing additional computational burdens or modifications. Using the notion of semi-smoothness from nonsmooth analysis, we extended the convergence guarantees of AA for smooth fixed-point problems to the nonsmooth case, and demonstrated how this result ensures (local) convergence rates for AA applied to PGD.

[1] `https://archive.ics.uci.edu/ml/datasets/Madelon`

[2] `https://docs.scipy.org/doc/scipy/reference/optimize.html`

7. REFERENCES

[1] Boris T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[2] Yurii Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course.*, Springer, New York, USA, 2004.

[3] Donald G. Anderson, "Iterative procedures for nonlinear integral equations," *Journal of the ACM*, vol. 12, no. 4, pp. 547–560, 1965.

[4] David A. Smith, William F. Ford, and Avram Sidi, "Extrapolation methods for vector sequences," *SIAM review*, vol. 29, no. 2, pp. 199–233, 1987.

[5] R. P. Eddy, "Extrapolating to the limit of a vector sequence," in *Information linkage between applied mathematics and industry*, pp. 387–396. Elsevier, 1979.

[6] Claude Brezinski, Michela Redivo-Zaglia, and Yousef Saad, "Shanks sequence transformations and Anderson acceleration," *SIAM Review*, vol. 60, no. 3, pp. 646–669, 2018.

[7] Haw-ren Fang and Yousef Saad, "Two classes of multisecant methods for nonlinear acceleration," *Numerical Linear Algebra with Applications*, vol. 16, no. 3, pp. 197–221, 2009.

[8] Homer F. Walker and Peng Ni, "Anderson acceleration for fixed-point iterations," *SIAM Journal on Numerical Analysis*, vol. 49, no. 4, pp. 1715–1735, 2011.

[9] Nicholas J. Higham and Nataša Strabić, "Anderson acceleration of the alternating projections method for computing the nearest correlation matrix," *Numerical Algorithms*, vol. 72, no. 4, pp. 1021–1042, 2016.

[10] Damien Scieur, Alexandre d'Aspremont, and Francis Bach, "Regularized nonlinear acceleration," in *Advances In Neural Information Processing Systems*, 2016, pp. 712–720.

[11] Damien Scieur, Francis Bach, and Alexandre d'Aspremont, "Nonlinear acceleration of stochastic algorithms," in *Advances in Neural Information Processing Systems*, 2017, pp. 3982–3991.

[12] Raghu Bollapragada, Damien Scieur, and Alexandre d'Aspremont, "Nonlinear acceleration of momentum and primal-dual algorithms," *arXiv preprint arXiv:1810.04539*, 2018.

[13] Junzi Zhang, Brendan O'Donoghue, and Stephen Boyd, "Globally convergent type-I Anderson acceleration for non-smooth fixed-point iterations," *arXiv preprint arXiv:1808.03971*, 2018.

[14] Alex Toth and CT Kelley, "Convergence analysis for Anderson acceleration," *SIAM Journal on Numerical Analysis*, vol. 53, no. 2, pp. 805–819, 2015.

[15] Florian A. Potra and Hans Engler, "A characterization of the behavior of the Anderson acceleration on linear problems," *Linear Algebra and its Applications*, vol. 438, no. 3, pp. 1002–1011, 2013.

[16] Youcef Saad and Martin H. Schultz, "GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems," *SIAM Journal on scientific and statistical computing*, vol. 7, no. 3, pp. 856–869, 1986.

[17] Francisco Facchinei and Jong-Shi Pang, *Finite-dimensional variational inequalities and complementarity problems*, Springer Science & Business Media, 2007.

[18] Frank H. Clarke, *Optimization and nonsmooth analysis*, vol. 5, Siam, 1990.

[19] Robert Mifflin, "Semismooth and semiconvex functions in constrained optimization," *SIAM Journal on Control and Optimization*, vol. 15, no. 6, pp. 959–972, 1977.

[20] Liqun Qi and Jie Sun, "A nonsmooth version of Newton's method," *Mathematical programming*, vol. 58, no. 1-3, pp. 353–367, 1993.

[21] R.Tyrrell Rockafellar and Roger J-B Wets, *Variational analysis*, Springer Science & Business Media, 2009.