Proximal Gradient Algorithms under Local Lipschitz Gradient Continuity*

— A Convergence and Robustness Analysis of PANOC —

Alberto De Marchi† Andreas Themelis‡

Abstract

Composite optimization offers a powerful modeling tool for a variety of applications and is often numerically solved by means of proximal gradient methods. In this paper, we consider fully nonconvex composite problems under only local Lipschitz gradient continuity for the smooth part of the objective function. We investigate an adaptive scheme for PANOC-type methods (Stella et al. in Proceedings of the IEEE 56th CDC, 1939–1944, 2017), namely accelerated linesearch algorithms requiring only the simple oracle of proximal gradient. While including the classical proximal gradient method, our theoretical results cover a broader class of algorithms and provide convergence guarantees for accelerated methods with possibly inexact computation of the proximal mapping. These findings have also significant practical impact, as they widen scope and performance of existing, and possibly future, general purpose optimization software that invoke PANOC as inner solver.

Keywords. Nonsmooth nonconvex optimization · locally Lipschitz gradient · forward-backward splitting · linesearch methods
AMS subject classifications. 49J52 · 65K05 · 90C30

---

*A. Themelis acknowledges the support of the Japan Society for the Promotion of Science (JSPS) KAKENHI grant JP21K17710.

†Universität der Bundeswehr München, Institute for Applied Mathematics and Scientific Computing, Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany. alberto.demarchi@unibw.de, ORCID: 0000-0002-3545-6898

‡Kyushu University, Faculty of Information Science and Electrical Engineering (ISEE), 744 Motooka, Nishi-ku, 819-0395 Fukuoka, Japan. andreas.themelis@ees.kyushu-u.ac.jp, ORCID: 0000-0002-6044-0169

1 Introduction

Problems involving the minimization of the sum of a smooth and a nonsmooth function are of interest for a wide variety of applications ranging from optimal and model predictive control (MPC), signal processing, compressed sensing, machine learning, and many others; see, e.g., [8, 18, 26] and references therein. Structured problems can also arise also as subproblems within other numerical optimization algorithms, e.g., the augmented Lagrangian method (ALM) [9, 25, 14]. These use cases often yield nonconvex and large-scale problems and can pose stringent requirements in terms of both computation and memory.

In the last few years, these considerations led to a renewed interest in algorithms of splitting nature [8, 18] owing to their simple operation oracles and low memory footprint, on top of their amenability to address nonsmooth, possibly nonconvex, constrained problems, making them widely applicable. The price of this flexibility is paid in terms of slow convergence and sensitivity to ill conditioning, hindering their direct employment to real-time applications, such as MPC, where optimal solutions to hard problems have to be retrieved in very limited time. In this very setting, the recently introduced PANOC [27] demonstrated how these downsides within the proximal gradient (PG) algorithm can be overcome while retaining all the favorable features. PANOC is an umbrella framework that includes the PG method as special instance; other variations are obtained by selecting virtually arbitrary update directions, which are suitably dampened in such a way to guarantee convergence. A most prominent use case is the employment of directions stemming from methods of quasi-Newton type, thanks to which considerable speed-ups (in some cases provable and quantifiable, see [27, Thm. III.5]) can be achieved by only performing elementary PG operations, in a nonsmooth and fully nonconvex setting.

Because of these favorable properties, PANOC was originally meant as a nonlinear MPC solver particularly suited for embedded applications subject to limited hardware capabilities, such as land and aerial vehicles [22, 24, 13] and robotics [2, 23, 3]; see also [17, 11] for extensive surveys and comparisons with other popular methods. Its success in the field led to a reconsideration of the spectrum of problems that the solver could be applied to. On a historical note, this evolution was reflected by a swift rebranding of the acronym over the years, originally meant as *Proximal Averaged Newton-type method for Optimal Control* in the original publication [27], but then tacitly reproposed as the same method *for Optimality Conditions* in [1] (and subsequent appearances) to allude to its applicability to the much broader range of composite minimization problems. This flexibility was further exploited in [25], where PANOC is employed as inner solver for ALM minimization subproblems for the general purpose Optimization Engine (OpEn) solver.

This rapid evolution was perhaps neglectful of some aspects, primarily because PG is subject to binding assumptions to guarantee a global Lipschitz differentiability requirement. In the context of MPC, physical bounds on input variables result in optimization problems where the feasible set is bounded, in which case *local* Lipschitzianity can be shown to suffice, making virtually no exclusion to the problems that can be addressed. In more general formulations, and especially so in a fully nonconvex setting, however, all known results are valid under a *global* Lipschitzianity assumption, with the very recent work [12] possibly emerging as unique exception in a vast literature. Other alternatives are to be found in the Bregman setting [6, 16], which are however subject to (and thus limited in applicability by) the identification of a distance-generating function enabling a so-called Lipschitz-like convexity condition and that makes induced proximal operations tractable at the same time. While this may not seem a major issue in composite minimization, it undeniably constitutes a severe drawback in ALM contexts, where constraints relaxation can produce subproblems with unbounded feasible sets, without this necessarily being the case for the original problem. Although adding large box constraints to ensure convergence may be thought of as a viable solution, unsatisfactory practical performance can persist because of poor geometry estimation, as we will show.

This paper addresses the above-mentioned shortcomings of PANOC, and of PG as a byproduct, by investigating an adaptive stepsize selection rule for its PG oracle. This criterion, in a slightly less general form, was first proposed in [19, Alg. 7], but without theoretical guarantees and driven from a different observation, namely the poor performance of PANOC if initial stepsizes are badly estimated. After confirming this claim with a case study example, we provide a complete convergence theory showing that the method, here referred to as PANOC^{+} for clarity, can also cope with *local* Lipschitzianity, while this is not the case for PANOC. Furthermore, we examine the robustness of the improved method with respect to suboptimal solutions of the PG subproblems. These findings will significantly impact on PANOC$^{(+)}$, both in performance and applicability, propagating to all its dependencies, e.g., by removing stringent assumptions of general purpose optimization solvers such as OpEn [25].

A convergence analysis of PG with a locally Lipschitz smooth term and possibly inexact inner minimizations is obtained as simple byproduct of the more general theory here developed. Indeed, a vast class of algorithms is covered by the analysis in this work, thanks to the arbitrarity of the selected update directions within the PANOC framework.

2 Problem Setting and Preliminaries

In this paper we consider structured minimization problems

$$\operatorname*{minimize}_{x\in\mathbb{R}^n} \varphi(x) := f(x) + g(x) \tag{P}$$

under the following standing assumptions, assumed throughout.

Blanket assumption. *The following hold in problem* (P)*:*

A1 $f : \mathbb{R}^n \to \mathbb{R}$ *has locally Lipschitz-continuous gradient.*

A2 $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *is proper, lsc, and* γ_g*-prox-bounded.*

A3 $\inf \varphi > -\infty$.

Motivated by its efficiency and popularity, yet aware of its inaptness to address this general problem formulation, this paper studies a robustified variant of PANOC algorithm with adaptive stepsize selection [27, Rem. III.4], building upon the preliminary work of [19, §6.1]. PANOC and the proposed generalization PANOC$^+$ will be presented and compared in Section 3, after the needed definitions and preliminary material are covered in this section.

2.1 Notational Conventions

With $\mathbb{R}$ and $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ we denote the real and extended-real line. The effective domain of an extended-real-valued function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is denoted by $\operatorname{dom} h := \{x \in \mathbb{R}^n \mid h(x) < \infty\}$, and we say that h is: proper if $\operatorname{dom} h \neq \emptyset$; lower semicontinuous (lsc) if $h(\bar{x}) \leq \liminf_{x\to\bar{x}} h(x)$ for all $\bar{x} \in \mathbb{R}^n$; coercive if $h(x) \to \infty$ as $\|x\| \to \infty$. With $\hat{\partial} h$ and ∂h we indicate the Fréchet and the limiting subdifferential of h at $\bar{x}$, which satisfy $\hat{\partial}(h + h_0) = \hat{\partial} h + \nabla h_0$ and $\partial(h + h_0) = \partial h + \nabla h_0$ for any $h_0 \in C^1(\mathbb{R}^n)$ [21, Ex. 8.8]. With respect to (P), we say that $x^* \in \operatorname{dom}\varphi$ is *stationary* if $0 \in \partial\varphi(x^*)$, which constitutes a necessary optimality condition of x^* for the minimization of φ [21, Thm. 10.1].

Given a parameter value $\gamma > 0$, the *Moreau envelope* function h^γ and the *proximal mapping* $\operatorname{prox}_{\gamma h}$ are defined by

$$h^\gamma(x) \coloneqq \inf_{z \in \mathbb{R}^n} \left\{ h(z) + \tfrac{1}{2\gamma}\|z - x\|^2 \right\}, \tag{2.1}$$

$$\operatorname{prox}_{\gamma h}(x) \coloneqq \operatorname*{arg\,min}_{z \in \mathbb{R}^n} \left\{ h(z) + \tfrac{1}{2\gamma}\|z - x\|^2 \right\}, \tag{2.2}$$

and we say that h is *prox-bounded* if it is proper and $h + \frac{1}{2\gamma}\|\cdot\|^2$ is bounded below on $\mathbb{R}^n$ for some $\gamma > 0$. The supremum of all such γ is the threshold γ_h of prox-boundedness for h. In particular, if h is bounded below by an affine function, then $\gamma_h = \infty$. When h is lsc, for any $\gamma \in (0, \gamma_h)$ the proximal mapping $\operatorname{prox}_{\gamma h}$ is nonempty- and compact-valued, and the Moreau envelope h^γ finite and locally Lipschitz continuous [21, Thm. 1.25 and Ex. 10.32].

2.2 Proximal Gradient Iterations

Given a point $x \in \mathbb{R}^n$, one iteration of the proximal gradient (PG) method for problem (P) consists in selecting

$$\bar{x} \in \mathrm{T}_\gamma(x) \coloneqq \operatorname{prox}_{\gamma g}(x - \gamma \nabla f(x)), \tag{2.3}$$

where $\gamma \in (0, \gamma_g)$ is a stepsize parameter. The necessary optimality condition in the minimization problem defining the proximal mapping then reads

$$\tfrac{1}{\gamma}(x - \bar{x}) - (\nabla f(x) - \nabla f(\bar{x})) \in \hat{\partial}\varphi(\bar{x}), \tag{2.4}$$

and in particular the fixed-point inclusion $x \in \mathrm{T}_\gamma(x)$ implies the stationarity condition $0 \in \partial\varphi(x)$. By interpreting (2.3) as a fixed-point iteration, one can also consider the associated (set-valued) fixed-point residual R_γ, namely

$$\mathrm{R}_\gamma(x) \coloneqq \tfrac{1}{\gamma}(x - \mathrm{T}_\gamma(x)), \tag{2.5}$$

and seek fixed points of T_γ as zeros of the residual R_γ.

2.3 Forward-Backward Envelope

At the heart of PANOC rationale is the observation that, under assumptions, the fixed-point residual R_γ in (2.5) is continuous around and even differentiable at critical points [28, §4], and the inclusion problem $0 \in \mathrm{R}_\gamma(\,\cdot\,)$ reduces to a well-behaved system of equations, when close to solutions. This motivated the adoption of Newton-type directions on R_γ, that enable fast convergence when close to solutions. The key tool enabling convergence regardless of whether or not the initial point happens to be sufficiently close to a solution is the so-called forward-backward envelope (FBE).

Definition 2.1 (forward-backward envelope)**.** *Relative to* (P)*, the FBE with step-size* $\gamma \in (0, \gamma_g)$ *is*

$$\varphi_\gamma^{\text{FB}}(x) \coloneqq \min_{w \in \mathbb{R}^n} \left\{ f(x) + \langle \nabla f(x), w - x \rangle + g(w) + \tfrac{1}{2\gamma}\|w - x\|^2 \right\} \tag{2.6a}$$

$$= f(x) - \tfrac{\gamma}{2}\|\nabla f(x)\|^2 + g^\gamma(x - \gamma \nabla f(x)) \tag{2.6b}$$

or, equivelantly, letting $\bar{x}$ *be any element of* $\mathrm{T}_\gamma(x)$*,*

$$= f(x) + \langle \nabla f(x), \bar{x} - x \rangle + g(\bar{x}) + \tfrac{1}{2\gamma}\|\bar{x} - x\|^2. \tag{2.6c}$$

Owing to its continuity properties, the FBE has been employed to generalize and improve PG-based algorithms that address the general setting of structured nonconvex optimization [15, 28, 7]. The following results are well known when f has globally Lipschitz gradient [28, Prop.s 4.2 and 4.3]. A simple proof in the more general setting addressed here is given for completeness.

Lemma 2.2 (Properties of the FBE)**.** *For any* $\gamma \in (0, \gamma_g)$ *the following hold:*

(i) $\varphi_\gamma^{\text{FB}}$ *is real valued and strictly continuous.*

(ii) $\varphi_\gamma^{\text{FB}}(x) \leq \varphi(x)$ *for any* $x \in \mathbb{R}^n$*, with equality holding iff* $x \in \mathrm{T}_\gamma(x)$*.*

(iii) If $\bar{x} \in \mathrm{T}_\gamma(x)$ *and* $f(\bar{x}) \leq f(x) + \langle \nabla f(x), \bar{x} - x \rangle + \tfrac{L}{2}\|\bar{x} - x\|^2$*, then*

$$\varphi_\gamma^{\text{FB}}(\bar{x}) \leq \varphi(\bar{x}) \leq \varphi_\gamma^{\text{FB}}(x) - \tfrac{1-\gamma L}{2\gamma}\|x - \bar{x}\|^2. \tag{2.7}$$

Proof. Assertion 2.2*(i)* follows from the expression (2.6b), owing to the similar property of the Moreau envelope g^γ, while 2.2*(ii)* is obtained by taking $w = x$ in (2.6a). The first inequality in 2.2*(iii)* owes to item 2.2*(ii)* (independently of L), and the second one follows from the expression (2.6c) of $\varphi_\gamma^{\text{FB}}$. □

Algorithm 1 Original PANOC with "bad" adaptive stepsize γ [27, Rem. III.4]

REQUIRE $x^0 \in \mathbb{R}^n$; $\gamma_0 \in (0, \gamma_g)$; $D \geq 0$; $\alpha, \beta \in (0,1)$

INITIALIZE $k = 0$, compute $\bar{x}^0 \in \mathrm{T}_{\gamma_0}(x^0)$, and start from step 1.6

1.1: Select an update direction $d^k \in \mathbb{R}^n$ with $\|d^k\| \leq D\|\bar{x}^{k-1} - x^{k-1}\|$ and set $\tau_k = 1$

1.2: $x^k = (1 - \tau_k)\bar{x}^{k-1} + \tau_k(x^{k-1} + d^k)$

1.3: Compute $\bar{x}^k \in \mathrm{T}_{\gamma_{k-1}}(x^k)$ and use it to evaluate $\varphi^{\mathrm{FB}}_{\gamma_{k-1}}(x^k)$ as in (2.6c)

1.4: IF $\varphi^{\mathrm{FB}}_{\gamma_{k-1}}(x^k) > \varphi^{\mathrm{FB}}_{\gamma_{k-1}}(x^{k-1}) - \beta\frac{1-\alpha}{2\gamma_{k-1}}\|\bar{x}^{k-1} - x^{k-1}\|^2$ THEN

$\tau_k \leftarrow \tau_k/2$ and go back to step 1.2

1.5: $\gamma_k \leftarrow \gamma_{k-1}$

1.6: WHILE $f(\bar{x}^k) > f(x^k) + \left\langle \nabla f(x^k), \bar{x}^k - x^k \right\rangle + \frac{\alpha}{2\gamma_k}\|\bar{x}^k - x^k\|^2$ DO

$\gamma_k \leftarrow \gamma_k/2$ and recompute $\bar{x}^k \in \mathrm{T}_{\gamma_k}(x^k)$

1.7: $k \leftarrow k + 1$ and start the next iteration at step 1.1

3 Good and Bad Adaptive Stepsize Selection Rules

As briefly mentioned in Section 2.3, the FBE is the key tool for *globalizing* the convergence of fast local methods, such as of quasi-Newton type, applied to the nonlinear equation $\mathrm{R}_\gamma(x) = 0$ encoding necessary optimality conditions for (P). Elaborating on how Newton-type directions can be selected given the nonsmooth, possibly set-valued, nature of R_γ is beyond the scope of this survey, and the interested reader is referred to [28, 27]. The core idea is nevertheless the same as in the familiar context of smooth minimization: trying to enforce (supposedly fast) updates $x \mapsto x + d$ in place of "nominal" updates $x \mapsto \bar{x}$, where $\bar{x}$ would amount to a gradient step or, in our nonsmooth setting, a proximal gradient step $\bar{x} \in \mathrm{T}_\gamma(x)$ as in (2.3). Still in complete analogy with the smooth case, accepting a candidate update $x + d$ must be validated by a "quality check", like an Armijo-type condition, in violation of which d is either discarded or dampened with a smaller stepsize. PANOC is precisely a mechanism to dampen and accept update directions in a nonsmooth setting, using the FBE as validation control. Its steps are given in Algorithm 1.

A needed assumption for this method is that ∇f be globally L_f-Lipschitz, so that a well-known quadratic upper bound, see e.g., [4, Prop. A.24], ensures that

$L = L_f$ can be taken for all $x \in \mathbb{R}^n$ in Lemma 2.2*(iii)*. For any $\alpha \in (0, 1)$ the choice $\gamma_k = {}^{\alpha}/_{L_f}$ then violates step 1.6, meaning that $\gamma_k \equiv \gamma$ is constant. The dampening of the direction occurs at step 1.2, where starting with $\tau_k = 1$ the candidate update $x^{k-1} + d^k$ is pushed towards $\bar{x}^{k-1} \in \mathrm{T}_\gamma(x^{k-1})$ by reducing the steplength τ_k until the value of the FBE is sufficiently reduced, cf. step 1.4. The process terminates, since $\varphi_\gamma^{\text{FB}}$ is continuous (at $\bar{x}^{k-1}$), and it is strictly smaller than $\varphi_\gamma^{\text{FB}}(x^{k-1}) - \beta\frac{1-\alpha}{2\gamma_{k-1}}\|\bar{x}^{k-1} - x^{k-1}\|^2$ there, cf. (2.7).

3.1 PANOC$^+$: the "Good" Adaptive Stepsize Rule

What is presented in Algorithm 1 is actually the "adaptive" variant of PANOC, which still works under the assumption of global Lipschitz differentiability but waives the need of prior knowledge about L_f. The γ-backtracking at step 1.6 decreases (i.e., "adapts") γ_k and terminates as soon as the needed bound as in Lemma 2.2*(iii)* is satisfied. As first noted in [19, §6.1], however, this adaptive criterion may produce bad estimates of the local Lipschitz constant of ∇f and overall result in poor algorithmic performance. The phenomenon can be attributed to an asynchrony between the two backtracking steps, the one dampening the update direction and the one adaptively adjusting the proximal gradient stepsize. This claim can be verified in the iteration mismatch between variable x^k and stepsize γ_{k-1} occurring at step 1.3.

To account for this fact, [19, Alg. 7] proposes to adapt the PG stepsize γ_k within the linesearch on the update direction. As recently showcased in [20], not only does this conservatism prove beneficial in preventing the acceptance of poor quality directions, but it often also reduces the overall computational cost. Although numerical simulations indicate superior performance, this refined linesearch lacks a theoretical analysis of its convergence properties.

This modification, which we allusively call the "good" adaptive variant (or PANOC$^+$ for brevity), is depicted in Algorithm 2. In fact, the method presented here presents a slight, but important generalization, namely in allowing the selection of a new direction d^k every time the stepsize γ_k is reduced, cf. step 2.5, which was not considered in [19, Alg. 7]. This flexibility is crucial: whenever the stepsize γ_k changes so does the PG residual mapping R_{γ_k}, and consistently so should directions using its curvature information. Moreover, we provide theoretical guarantees on the finite termination of the backtracking linesearch procedure, even without global Lipschitz gradient continuity and merely suboptimal proximal computation. These findings uphold the algorithmic framework proposed in

Algorithm 2 PANOC$^+$: the "good" adaptive γ-stepsize rule

REQUIRE $x^0 \in \mathbb{R}^n$; $\gamma_0 \in (0, \gamma_g)$; $D \geq 0$; $\alpha, \beta \in (0, 1)$

INITIALIZE $k \leftarrow 0$, and start from step 2.4

2.1: $\gamma_k \leftarrow \gamma_{k-1}$

2.2: Select an update direction $d^k \in \mathbb{R}^n$ with $\|d^k\| \leq D\|\bar{x}^{k-1} - x^{k-1}\|$ and set $\tau_k = 1$

2.3: $x^k = (1 - \tau_k)\bar{x}^{k-1} + \tau_k(x^{k-1} + d^k)$

2.4: Compute $\bar{x}^k \in \mathrm{T}_{\gamma_k}(x^k)$ and use it to evaluate $\Phi_k := \varphi^{\text{FB}}_{\gamma_k}(x^k)$ as in (2.6c)

2.5: IF $f(\bar{x}^k) > f(x^k) + \left\langle \nabla f(x^k), \bar{x}^k - x^k \right\rangle + \frac{\alpha}{2\gamma_k}\|\bar{x}^k - x^k\|^2$ THEN

$\gamma_k \leftarrow {}^{\gamma_k}/_2$, and go back to step 2.2 if $k > 0$, or step 2.4 if $k = 0$

2.6: IF $k > 0$ AND $\Phi_k > \Phi_{k-1} - \beta\frac{1-\alpha}{2\gamma_{k-1}}\|\bar{x}^{k-1} - x^{k-1}\|^2$ THEN

$\tau_k \leftarrow {}^{\tau_k}/_2$ and go back to step 2.3

2.7: $k \leftarrow k + 1$ and start the next iteration at step 2.1

[27, 19, 20] on two aspects: the adaptive linesearch is shown to terminate, and can cope with a merely locally Lipschitz-differentiable term f. These findings are of high significance also for other methods that rely on PANOC as internal solver, such as the general purpose OpEn [25]. What's more, it will be shown that all this remains true even if the minimization problem defining the PG mapping T_{γ_k} is solved inexactly and/or suboptimally.

The peculiarity of PANOC$^+$ over the *bad* adaptive rule of original PANOC is that the two backtracking steps, the one on the direction τ_k and the one on the PG stepsize γ_k, are tightly intertwined. The intricate structure emerges at steps 2.5 and 2.6: the direction stepsize τ_k resets every time the proximal stepsize γ_k is adjusted and, conversely, the value of γ_k is assessed anew when τ_k changes. This entanglement allows the evaluation of the FBE at step 2.4 with an up-to-date stepsize γ_k, as opposed to (and eliminating) the asynchrony obstructing PANOC's performance. The adaptivity of PANOC$^+$ allows the FBE $\varphi^{\text{FB}}_{\gamma_k}$ to better capture the (local) landscape of φ and, ultimately, to relax the assumption of globally Lipschitz gradient.

To substantiate these claims, in the following Section 3.2 we first showcase the ineffectiveness of PANOC applied to problem (P) where f has only locally Lipschitz-continuous gradient, and then compare the "good" and the "bad" adaptive strategies on a common ground in Section 3.3.

Remark 3.1 (Algorithm notation)**.** Algorithm 2 operates two linesearch steps within each iteration, one on the "proximal" stepsize γ_k at step 2.5 and one on the "direction" stepsize τ_k at step 2.6. Whenever the respective needed conditions are violated, either γ_k or τ_k is reduced and the iteration restarted from a previous step. As a consequence, variables may be *overwritten* within each iteration before being accepted. To avoid a heavy double-index notation, used only within proofs out of full rigor, the sub- and superscript notation is designed to differentiate temporary and permanent variables; specifically, within iteration k only variables indexed with k are updated, whereas those indexed with $k-1$ remain untouched. Similar considerations apply to Algorithm 1. □

3.2 Failure of "Bad" PANOC without Globally Lipschitz Gradient

Let us consider the minimization of the convex, twice continuously differentiable, coercive function $\varphi = f + g$, where $f(x) = \frac{2}{9}|x|^3$ and $g = 0$, namely

$$\operatorname*{minimize}_{x\in\mathbb{R}} \varphi(x) \coloneqq \tfrac{2}{9}|x|^3 + 0, \tag{3.1}$$

and adopt PANOC as given in Algorithm 1. In particular, we choose the directions as $d_k = \frac{9}{2\gamma_{k-1}x_{k-1}}(x_{k-1} - \bar{x}_{k-1})$. As we are about to show, starting from any $x_0 > 0$ this particular choice of directions complies with the bound $\|d_k\| \le D\|x_{k-1} - \bar{x}_{k-1}\|$ for $D = 18$ and satisfies the τ-linesearch with $\tau_k = 1$ for every k. Moreover, the choice $\alpha = {}^{16}/_{27}$ leads to a conveniently simple expression for the γ-linesearch, namely $\gamma_k \le \frac{1}{2x_k}$. As a result, starting from $x_0 > 0$ with $\gamma_0 > \frac{1}{4x_0}$, the algorithm reduces iterating the following lines

$$\begin{cases} \text{halven } \gamma_k \text{ until } \gamma_k \le \frac{1}{2x_k} \\ \;\bar{x}_k \;\;= x_k(1 - \frac{2}{3}\gamma_k x_k) \\ x_{k+1} = x_k + \frac{9}{2\gamma_k x_k}(x_k - \bar{x}_k) = 4x_k \end{cases} \tag{3.2}$$

and thus produces a sequence $x_k = x_0 4^k$ that is diverging, and causes the cost to increase unboundedly. We now show the claims one by one. To this end, denoting $y_k \coloneqq \gamma_k x_k$ throughout, observe that

$$\bar{x}_k = x_k\Big(1 - \tfrac{2}{3}|y_k|\Big) \quad \text{and} \quad \varphi_{\gamma_k}^{\text{FB}}(x) = \tfrac{2}{9}|x|^3(1 - \gamma_k x). \tag{3.3}$$

• *Linesearch on γ.* For $x_k > 0$ the backtracking on γ_k at step 1.6 (after removing a $\frac{2}{9}x_k^3$ factor) terminates when

$$\left|1 - \tfrac{2}{3}y_k\right|^3 \leq 1 - 2y_k + \alpha y_k. \tag{3.4}$$

To simplify the computation, observe that necessarily $y_k \leq 1$ for inequality (3.4) to hold, and in particular the argument of the absolute value is necessarily positive: in fact, since $y_k = \gamma_k x_k > 0$ and $\alpha < 1$, (3.4) implies $\left|1 - \frac{2}{3}y_k\right|^3 \leq 1 - y_k$, hence $y_k \leq 1$. After this simplification and by restricting the analysis to $y_k = \gamma_k x_k > 0$, it can be seen that (3.4) has solution $0 < \gamma_k \leq \frac{9}{4x_k}\left(1 - \sqrt{1 - \frac{2}{3}\alpha}\right)$. For $\alpha = 16/27$, this bound simplifies to $0 < \gamma_k \leq \frac{1}{2x_k}$ as claimed. This shows the validity of the first line in (3.2). Since γ_k is halvened (only) until it enters this range, one also has that

$$y_k \coloneqq \gamma_k x_k > \tfrac{1}{4} \quad \forall k. \tag{3.5}$$

• *Bound on the directions $\|d_{k+1}\| \leq D\|x_k - \bar{x}_k\|$.* Since $d_{k+1} = \frac{9}{2\gamma_k x_k}(x_k - \bar{x}_k)$, one has $\|d_{k+1}\| = \frac{9}{2|\gamma_k x_k|}\|x_k - \bar{x}_k\| \leq 18\|x_k - \bar{x}_k\|$ as it follows from (3.5).

• *Linesearch on τ.* Starting with $x_k > 0$ we show that $x_{k+1} = x_k + d_{k+1} = 4x_k$ satisfies the linesearch condition. Indeed, by using the expression for the FBE in (3.3), according to step 1.4 the iterate $x_{k+1} = 4x_k$ is accepted if

$$\tfrac{2}{9}(4x_k)^3(1 - 4y_k) \leq \tfrac{2}{9}x_k^3(1 - y_k) - \beta(1 - \alpha)\tfrac{2}{9}x_k^3 y_k$$

which is easily reduced to $y_k \geq \frac{4^3-1}{4^4-1-\beta(1-\alpha)}$. Since $\beta(1-\alpha) < 1$, one has $\frac{4^3-1}{4^4-1-\beta(1-\alpha)} \leq \frac{4^3-1}{4^4-2} < \frac{1}{4}$, and (3.5) implies that the inequality always holds.

We stressed that, although we consider an exemplary problem designed to yield simple computations, similar arguments would still apply for C^∞, strongly convex formulations, e.g., $x^4 + x^2$; see also Remark 3.2.

3.3 "Good" PANOC$^+$ vs "Bad" PANOC

In spite of the breakdown demonstrated in Section 3.2, global convergence guarantees for PANOC can be recovered by adding a term g with bounded domain (as is the case of a possibly large but bounded box constraint) and selecting update directions d_k that are bounded, see [27, Rem. III.4]. Nonetheless, as noted in [19, §6.1], this would scarcely help in practice: early iterations would be agnostic to

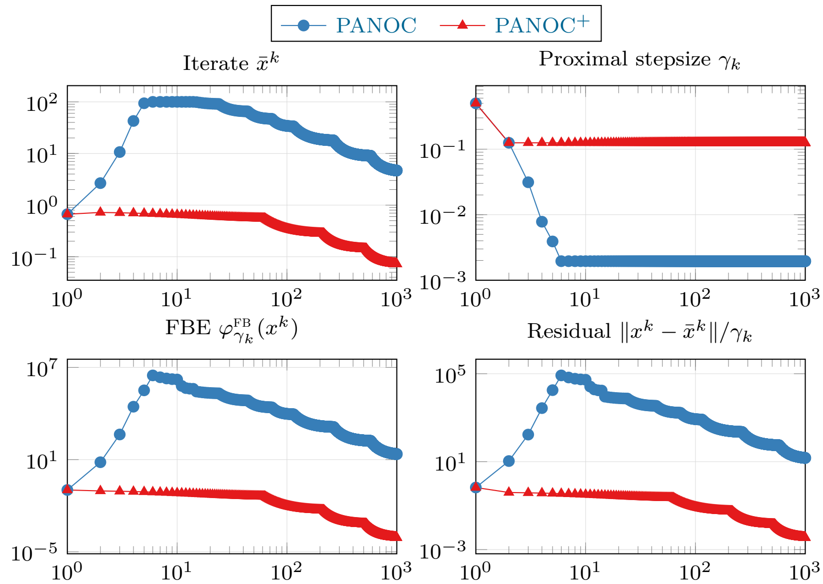

Figure 1: *Comparison of convergence metrics vs iterations for PANOC and PANOC⁺ on an illustrative problem. PANOC's iterates diverge until the (safeguarding) box constraint activates, and only then, with a reduced stepsize γ, slowly recovers*

the large box and exhibit the same diverging behavior until the boundary is approached, at which point a drastically reduced stepsize γ would be the cause of a painfully slow convergence.

We substantiate these claims by considering the example in Section 3.2 with some amendments. In particular, we let g be the indicator function of the interval $[-B, B]$, namely $g(x) = 0$ if $|x| \leq B$ and $g(x) = \infty$ otherwise, and select directions d_k as above if $\|d_k\| \leq E$ and $Ed_k/\|d_k\|$ otherwise, with possibly large but bounded $B, E \geq 0$. Adopting these precautions, PANOC generates iterates that converge to a solution, starting from any initial point. We set $B = E = 100$ for the results displayed in Figure 1 with a comparison against PANOC⁺. Although the latter solves the illustrative problem in its original form (that is, with $B = \infty$), we stress that it would not be affected by the safeguards put in place to guarantee the convergence of "bad" PANOC.

The diverging behavior of PANOC is apparent, until the safeguards activate, as expected from Section 3.2. At step 1.3 PANOC accepts an update x^k based on the sufficient decrease of a merit function defined by the FBE with the *previous* step-

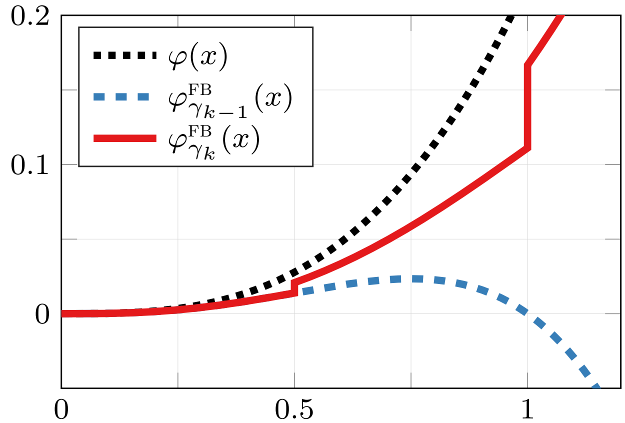

Figure 2: *Comparison of the cost function φ for the illustrative problem* (3.1) *against PANOC's and PANOC$^+$'s merit functions with previous, or initial, estimate $\gamma_{k-1} = 1$*

size γ_{k-1}. Figure 2 illustrates this phenomenon by comparing the merit functions adopted by PANOC and PANOC$^+$ to verify whether a tentative update is to be accepted or not. In this example, PANOC's merit function are lower unbounded (see (3.3)) and full steps along the update directions d_k are accepted, in fact *favored*, leading to diverging iterates. In turn, this results in a temporary departure from the solution, degrading the overall efficiency of the algorithm. Conversely, at step 2.4 PANOC$^+$ verifies sufficient decrease of the FBE with the *current* stepsize γ_k, yielding monotone decrease of the (time varying, but lower bounded) merit function $\varphi^{\mathrm{FB}}_{\gamma_k}$, as depicted in Figure 1. Note that the merit function for PANOC$^+$ in Figure 2 is only piecewise continuous because its evaluation is always preceded by the γ-stepsize backtracking, i.e., the stepsize $\gamma_k = \gamma_k(x^k)$ in $\varphi^{\mathrm{FB}}_{\gamma_k}$ depends on the candidate update x^k being tested. This adaptivity allows PANOC$^+$ to well estimate the geometry of the cost function φ and to construct a tighter merit function.

These simulations also show that, despite the more conservative linesearch, PANOC$^+$ does not necessarily require more iterations nor function evaluations to provide a more consistent performance, nor does it lead to a smaller stepsize. Indeed, considering larger box constraints and update directions, i.e., larger values for B, the limitations and inadequacy of "bad" PANOC in this setting become apparent, while providing support in favor of the (initially) more conservative adaptive scheme of "good" PANOC$^+$.

Remark 3.2. Noticeably, the "bad" PANOC can exhibit this diverging behavior even when the problem admits just one feasible point. To see this, let us consider once again the illustrative example above with $B = 0$, so that $\operatorname{dom} g = \operatorname{dom} \varphi = \{0\}$. Then, patterning the proof in Section 3.2, we obtain that the algorithm produces a sequence $(x_k)_{k\in\mathbb{N}}$ that is diverging, despite the fact that $\bar{x}^k = 0$ for every k, since $\varphi^{\text{FB}}_{\gamma_k}(x) = x^2(\frac{1}{2\gamma_k} - \frac{4}{9}|x|)$ is still lower unbounded for any $\gamma_k > 0$. This also confirms the necessity of imposing bounded $\|d^k\|$ in [27, Rem. III.4], in addition to $\|d^k\| \leq D\|x^{k-1} - \bar{x}^{k-1}\|$ as in step 1.1, not needed in the "good" PANOC$^+$ even with unbounded domains. □

4 Algorithmic Analysis under Inexact Proximal Oracles

In this section we analyze the properties of the iterates generated by PANOC$^+$, starting from their well definedness. As a substantial proof of robustness with respect to inexact prox evaluations, we will generalize the setting to an extent that the oracle of the proximal mapping is not required, and instead only a local solution of the proximal subminimization problem is needed. We will refer to this variant as the *inexact PANOC*$^+$, and emphasize that the exact counterpart described in Algorithm 2 falls as a special case.

The investigation in this section originates essentially from three observations. Firstly, in the inexact scenario we cannot avail ourselves of the FBE, as its evaluation requires global optimality in the solution of the proximal subproblem. Secondly, by considering the equivalent reformulation of (P)

$$\operatorname*{minimize}_{x,z\in\mathbb{R}^n} f(x) + g(z) \quad \text{subject to } x = z$$

and defining the associated augmented Lagrangian function

$$\mathcal{L}_\beta(x, z, y) \coloneqq f(x) + g(z) + \langle y, x - z\rangle + \tfrac{\beta}{2}\|x - z\|^2, \tag{4.1}$$

we remark that

$$\varphi^{\text{FB}}_\gamma(x) = \mathcal{L}_{1/\gamma}(x, \bar{x}, -\nabla f(x)), \tag{4.2}$$

where

$$\bar{x} \in \operatorname{T}_\gamma(x) = \operatorname*{arg\,min} \mathcal{L}_{1/\gamma}(x, \cdot\,, -\nabla f(x)) \tag{4.3}$$

is the result of an exact proximal minimization. Thirdly, in the ALM framework, algorithms can be constructed that converge in some sense to stationary points of the optimization problem, even solving the associated subproblems only approximately [5]. Therefore, we seek relaxed (sub)optimality concepts for the evaluation of the proximal mapping. This viewpoint will ultimately highlight how additionally to being used as a solver within ALMs, as in [25, 10, 20], PANOC$^+$ can operate as an ALM-type solver itself.

In the broadest possible setting, we do not require any (sub)optimality in the proximal minimization subproblem other than improvement with respect to the previous iteration. Clearly, additional conditions are needed for generating meaningful iterates, but as a proof of robustness of PANOC$^+$ we demonstrate that any choice complying with said requirement maintains the well definededness of the algorithm. We will then provide instances of such conditions that, possibly under additional assumptions on the problem, ensure optimality conditions for the limit points of the proposed inexact variant.

Specifically, we consider Algorithm 2 with the following instruction replacing step 2.4 therein, remarking that "exact" $\bar{x}^k \in \mathrm{T}_{\gamma_k}(x^k)$ as prescribed in Algorithm 2 comply with this relaxed requirement (any such $\bar{x}^k$ is a global minimizer of $\mathcal{L}_{1/\gamma_k}(x^k, \cdot, -\nabla f(x^k))$, and $\Phi_k = \varphi^{\text{FB}}_{\gamma_k}(x^k)$ in this case).

Suboptimal prox step for inexact PANOC$^+$

2.4′: Let $\bar{x}^k$ be a suboptimal minimizer of $\mathcal{L}_{1/\gamma_k}(x^k, \cdot, -\nabla f(x^k))$ such that

$$\Phi_k \coloneqq \mathcal{L}_{1/\gamma_k}(x^k, \bar{x}^k, -\nabla f(x^k)) \leq \mathcal{L}_{1/\gamma_k}(x^k, \bar{x}^{k-1}, -\nabla f(x^k)). \tag{4.4}$$

4.1 Well Definedness and Convergence Results

A crucial complication that the stepsize adjustment in the "good" PANOC$^+$ suffers if compared with the original one in the "bad" PANOC, is that it gives rise to a nested dependency between γ_k, τ_k, and d^k that could potentially give rise to infinite recursions. While this is fortunately not the case, as we are about to show, the proof is not as straightforward as in [27]. On top of this, while in the "exact" case local boundedness properties of the PG operator T_{γ_k} could conveniently be exploited, in accounting also for inexactness even for a fixed x^k the set of points $\bar{x}^k$ complying with the relaxed requirement (4.4) may be unbounded. The following result will serve as surrogate of local boundedness for the suboptimal proximal operator.

Lemma 4.1. *Let a constant $c \in \mathbb{R}$, a sequence $(\gamma_j)_{j\in\mathbb{N}} \searrow 0$, and two bounded sequences $(u^j, z^j)_{j\in\mathbb{N}}$ in $\mathbb{R}^n$ be fixed, and for every $j \in \mathbb{N}$ let $\bar{z}^j$ be such that*

$$g(\bar{z}^j) + \langle u^j, \bar{z}^j - z^j \rangle + \tfrac{1}{2\gamma_j}\|\bar{z}^j - z^j\|^2 \leq \tfrac{c}{2\gamma_j}.$$

Then, $(\bar{z}^j)_{j\in\mathbb{N}}$ *is bounded.*

Proof. An application of Young's inequality on the inner product yields

$$2\gamma_j g(\bar{z}^j) \le c + \gamma_j \|u_j\|^2 - (1-\gamma_j)\|\bar{z}^j - z^j\|^2.$$

To arrive to a contradiction, up to extracting if necessary, suppose that $0 < \|\bar{z}^j\| \to \infty$. Since $\liminf_{j\to\infty} g(\bar{z}^j)/\|\bar{z}^j\|^2 > -\infty$ by [21, Ex. 1.24], dividing by $\|\bar{z}^j\|^2$ and passing to the limit leads to the contradiction $0 \le -1$. □

To avoid trivialities, in what follows we assume that $x^k \neq \bar{x}^k$ always holds. This is consistent with stopping criteria based on the PG residual $\frac{1}{\gamma_k}\|x^k - \bar{x}^k\|$, see Section 4.2, in which case $x^k = \bar{x}^k$ would trigger a successfull termination.

Lemma 4.2 (Well definedness of the "good" (inexact) PANOC$^+$)**.** *Consider the iterates generated by Algorithm 2 with inexact proximal evaluation at step 2.4 as given in* (4.4)*. The following hold:*

(i) Well definedness: at every iteration, the number of backtrackings at steps 2.5 and 2.6 is finite.

(ii) At the end of the k-th iteration ($k \ge 1$*), one has*

$$\varphi(\bar{x}^k) + \delta_k \le \Phi_k \le \Phi_{k-1} - \beta\delta_{k-1} \quad \textit{where} \quad \delta_k := \tfrac{1-\alpha}{2\gamma_k}\|\bar{x}^k - x^k\|^2. \tag{4.5}$$

(iii) Every iterate $\bar{x}^k$ *remains within* $\operatorname{lev}_{\le c}\varphi$*, where* $c = \Phi_0 < \infty$*.*

Proof. As observed in Remark 3.1, each iteration k defines or updates only variables indexed with a k sub/superscript, while those defined in previous interations are untouched. In what follows, let us index by k, j the variables defined at the j-th attempt within iteration k. Note further that $\gamma_{k,j}L_{k,j} = \alpha \in (0,1)$ holds for every attempt j within every iteration k, since every time γ_k is halvened the estimate L_k is doubled (cf. step 2.5).

• 4.2*(i)* We proceed by induction on k. If $k = 0$, there is no backtracking on τ, and from Lemma 4.1 we conclude that all the trials $\bar{x}^{0,j}$ remain confined in a bounded set Ω_0, and therefore any stepsize $\gamma_{0,j} < 1/L_{f,\Omega_0}$ is accepted.

Suppose now that $k > 0$ and observe that, by the definition of Φ_k in (4.4) and the failure of the condition at step 2.5, the inequality

$$\varphi(\bar{x}^{k-1}) \le \Phi_{k-1} - \tfrac{1-\alpha}{2\gamma_{k-1}}\|x^{k-1} - \bar{x}^{k-1}\|^2 \tag{4.6}$$

holds. Since $\|d^{k,j}\| \leq D\|\bar{x}^{k-1} - x^{k-1}\|$ and $\tau_{k,j} \in [0,1]$, any attempt $x^{k,j}$ defined at step 2.3 during the k-th iteration satisfies

$$\|x^{k,j} - \bar{x}^{k-1}\| = \tau_{k,j}\|x^{k-1} - \bar{x}^{k-1} + d^{k,j}\| \leq (1+D)\|\bar{x}^{k-1} - x^{k-1}\|$$

and thus remains in a bounded set, be it Ω_k. To arrive to a contradiction, suppose that $\gamma_{k,j} \searrow 0$ as $j \to \infty$. Observe that condition (4.4) reads

$$g(\bar{x}^{k,j}) + \langle \nabla f(x^{k,j}), \bar{x}^{k,j} - \bar{x}^{k-1}\rangle + \tfrac{1}{2\gamma_{k,j}}\|x^{k,j} - \bar{x}^{k,j}\|^2 \leq g(\bar{x}^{k-1}) + \tfrac{1}{2\gamma_{k,j}}\|x^{k,j} - \bar{x}^{k-1}\|^2.$$

Since $(x^{k,j})_{j\in\mathbb{N}}$ is bounded, an application of Lemma 4.1 reveals that $(\bar{x}^{k,j})_{k\in\mathbb{N}}$ too is bounded. Up to possibly enlarging the set, both sequences remain confined in the bounded set Ω_k, implying that the condition at step 2.5 should have terminated in finite time, whence the sought contradiction.

Hence, $\gamma_{k,j}$ is backtracked finitely many times within iteration k; up to discarding early attempts, we may denote $\gamma_{k,j} = \gamma_k$. Condition (4.4) reads

$$\begin{aligned}\mathcal{L}_{1/\gamma_k}(x^{k,j}, \bar{x}^{k,j}, -\nabla f(x^{k,j})) &\leq \mathcal{L}_{1/\gamma_k}(x^{k,j}, \bar{x}^{k-1}, -\nabla f(x^{k,j}))\\ &= f(x^{k,j}) + g(\bar{x}^{k-1}) + \langle \nabla f(x^{k,j}), \bar{x}^{k-1} - x^{k,j}\rangle\\ &\quad + \tfrac{1}{2\gamma}\|x^{k,j} - \bar{x}^{k-1}\|^2.\end{aligned}$$

As $\tau_{k,j} \searrow 0$, one has that $x^{k,j} \to \bar{x}^{k-1}$. Since f and ∇f are continuous, the right-hand side of the inequality converges to $\varphi(\bar{x}^{k-1})$, overall resulting in

$$\limsup_{j\to\infty} \mathcal{L}_{1/\gamma_k}(x^{k,j}, \bar{x}^{k,j}, -\nabla f(x^{k,j})) \leq \varphi(\bar{x}^{k-1}) \overset{(4.6)}{\leq} \Phi_{k-1} - \tfrac{1-\alpha}{2\gamma_{k-1}}\|x^{k-1} - \bar{x}^{k-1}\|^2.$$

Since $\|x^{k-1} - \bar{x}^{k-1}\| > 0$ and $\beta < 1$, for j large enough the condition at step 2.6 will be violated and therefore the k-th iteration successfully terminated.

- 4.2*(ii)* Follows by combining (4.6) with the failure of the condition at step 2.6 at the end of the iteration.

- 4.2*(iii)* Direct consequence of assertion 4.2*(ii)*. □

We next consider an asymptotic analysis of the algorithm.

Theorem 4.3 (Asymptotic analysis of the "good" (inexact) PANOC$^+$)**.** *Consider the iterates generated by Algorithm 2 with inexact proximal evaluation at step 2.4 as given in* (4.4)*. The following hold:*

(i) $(\Phi_k)_{k\in\mathbb{N}}$ *converges to a finite value* $\varphi_\star \geq \inf\varphi$ *from above.*

(ii) $\sum_{k\in\mathbb{N}} \frac{1}{\gamma_k}\|\bar{x}^k - x^k\|^2 < \infty$.

(iii) $\lim_{k\to\infty}\|x^k - \bar{x}^k\| = \lim_{k\to\infty}\|x^k - x^{k-1}\| = \lim_{k\to\infty}\|\bar{x}^k - \bar{x}^{k-1}\| = 0$, *and in particular the set of limit points of* $(x^k)_{k\in\mathbb{N}}$ *is closed and connected, and coincides with that of* $(\bar{x}^k)_{k\in\mathbb{N}}$.

(iv) $\sum_{k\in\mathbb{N}} \gamma_k = \infty$.

(v) $\liminf_{k\to\infty} \frac{1}{\gamma_k}\|x^k - \bar{x}^k\| = 0$.

(vi) Consider the following assertions:

(1) φ *is level bounded;*

(2) $(\bar{x}^k)_{k\in\mathbb{N}}$ *is bounded;*

(3) $(x^k)_{k\in\mathbb{N}}$ *is bounded;*

(4) $(\gamma_k)_{k\in\mathbb{N}}$ *is asymptotically constant, i.e., there exists* $\kappa\in\mathbb{N}$ *such that* $\gamma_k = \gamma_\kappa$ *for every* $k\geq\kappa$.

One has (1) $\Rightarrow$ *(2)* $\Leftrightarrow$ *(3)* $\Rightarrow$ *(4).*

Proof.

- 4.3*(i)* Follows from (4.5).

- 4.3*(ii)* A telescoping argument on (4.5) yields

$$\beta(1-\alpha)\sum_{k\in\mathbb{N}} \tfrac{1}{2\gamma_k}\|\bar{x}^k - x^k\|^2 \leq \Phi_0 - \inf\varphi = \varphi^{\text{FB}}_{\gamma_0}(x^0) - \inf\varphi, \tag{4.7}$$

whence the claimed finite sum.

- 4.3*(iii)* That $\|x^k - \bar{x}^k\| \to 0$ follows from assertion 4.3*(ii)*, since γ_k is upper bounded. Next, by the conditions at steps 2.2 and 2.3, observe that

$$\|x^k - x^{k-1}\| = \left\|(1-\tau_k)(\bar{x}^{k-1} - x^{k-1}) + \tau_k d^k\right\| \leq (1+D)\|\bar{x}^{k-1} - x^{k-1}\| \tag{4.8}$$

and thus $\|x^k - x^{k-1}\|$ vanishes, and in turn so does $\|\bar{x}^k - \bar{x}^{k-1}\|$ since

$$\|\bar{x}^k - \bar{x}^{k-1}\| \leq \|x^k - \bar{x}^k\| + \|\bar{x}^{k-1} - x^{k-1}\| + \|x^k - x^{k-1}\|.$$

- 4.3*(vi)* The first implication follows from Lemma 4.2*(iii)*, and the second one from assertion 4.3*(ii)*. Finally, if $(x^k)_{k\in\mathbb{N}}$ is bounded, and thus so is $(\bar{x}^k)_{k\in\mathbb{N}}$, the set Ω_k in the proof of Lemma 4.2*(i)* can be taken independent of k, and asymptotic constancy of γ_k follows from the same arguments therein.

• 4.3*(iv)* By iteratively applying inequality (4.8), we obtain that

$$\begin{aligned}
\|x^k - x^0\| &\leq (1+D)\sum_{j=0}^{k-1}\|\bar{x}^j - x^j\| \\
&= (1+D)\sum_{j=0}^{k-1}\gamma_j^{-1/2}\|\bar{x}^j - x^j\|\gamma_j^{1/2} \\
&\leq (1+D)\sqrt{\textstyle\sum_{j=0}^{k-1}\gamma_j^{-1}\|\bar{x}^j - x^j\|^2}\sqrt{\textstyle\sum_{j=0}^{k-1}\gamma_j} \\
&\overset{(4.7)}{\leq} (1+D)\sqrt{2\tfrac{\varphi_{\gamma_0}^{\text{FB}}(x^0)-\inf\varphi}{\beta(1-\alpha)}}\sqrt{\textstyle\sum_{j=0}^{k-1}\gamma_j}.
\end{aligned}$$

Contrary to the claim, if $\sum_{k\in\mathbb{N}}\gamma_k < \infty$ holds, then $(x^k)_{k\in\mathbb{N}}$ is bounded. From assertion 4.3*(vi)* proven above we then infer that γ_k is asymptotically constant, thus contradicting the finiteness of $\sum_{k\in\mathbb{N}}\gamma_k$.

• 4.3*(v)* Immediate consequence of assertions 4.3*(ii)* and 4.3*(iv)*. □

If the iterates remain bounded (as is the case when φ is level bounded), owing to Theorem 4.3*(vi)*, Algorithm 2 with exact prox evaluations as in step 2.4 eventually reduces to the original PANOC [27] with constant stepsize, and its convergence results are then readily available, including global convergence (possibly at R-linear rates) under Kurdika-Łojasiewicz assumptions, and superlinear when converging to a strong local minimum with directions satisfying the Dennis-Moré condition, see [27, 28].

Nevertheless, even in accounting for inexact proximal evaluations it is still possible to derive some qualitative guarantees for the limit points, provided that $\bar{x}^k$ satisfies some local suboptimality requirements. We list two such instances in the following definition and later detail a proof validating the claim.

Definition 4.4 (Prox suboptimality criteria)**.** *Relative to the minimization problem* (4.3) *defining the PG mapping, we say that the iterates $\bar{x}^k$ computed at step 2.4′ are:*

(i) δ-stationary *(for some $\delta > 0$) if* $\operatorname{dist}(0, \partial[\mathcal{L}_{1/\gamma_k}(x^k, \cdot, -\nabla f(x^k))](\bar{x}^k)) \leq \delta$, *that is, if there exists $\bar{v}^k \in \partial g(\bar{x}^k)$ such that*

$$\left\|\bar{v}^k + \nabla f(x^k) + \tfrac{1}{\gamma_k}(\bar{x}^k - x^k)\right\| \leq \delta. \tag{4.9}$$

(ii) Uniformly locally optimal *if there exist* $r > 0$ *and a sequence* $\varepsilon_k \searrow 0$ *such that the following local minimality condition holds:*

$$\mathcal{L}_{1/\gamma_k}(x^k, \bar{x}^k, -\nabla f(x^k)) \leq \mathcal{L}_{1/\gamma_k}(x^k, x, -\nabla f(x^k)) + \varepsilon_k \quad \forall x \in \overline{\mathrm{B}}(\bar{x}^k; r). \tag{4.10}$$

Notice that no (approximate) local minimality is required in the approximate stationarity criterion of Definition 4.4*(i)*. Consequently, the output can be retrieved by any descent method starting at the previous iteration and terminating when δ-stationarity is achieved. It is also worth remarking that the prox suboptimality tolerance δ does not need to be small nor fixed for all iterations, and can instead be replaced by a sequence $\delta_k \searrow \delta \geq 0$. The uniform local optimality requirement of Definition 4.4*(ii)* is instead more restrictive, and is possibly subject to prior knowledge on the geometry of the augmented Lagrangian. The uniformity is dictated by the value of $r > 0$, whose role can be appreciated by considering the sequence $z^k = 1/k$ for $k > 0$ which consists of (isolated) local minimizers for the function

$$h(x) = \begin{cases} x & \text{if } x = 1/k,\ k \in \mathbb{N}_{>0} \\ x^2 + x - 1 & \text{if } x \leq 0 \\ \infty & \text{otherwise,} \end{cases}$$

yet the limit $z = 0$ is not stationary for h. The pathology arises from the non uniformity of the radius of local minimality of z^k, which is $r_k < 1/k(k+1) \to 0$.

Theorem 4.5 (Subsequential convergence of inexact PANOC$^+$)**.** *Consider the iterates generated by Algorithm 2 with inexact proximal evaluation at step 2.4 as given in* (4.4)*. Suppose that the iterates remain bounded (as is the case when* φ *is coercive), and let* ω *be the set of limit points of* $(\bar{x}^k)_{k\in\mathbb{N}}$*. Then:*

(i) If $(\bar{x}^k)_{k\in\mathbb{N}}$ *are* δ*-stationary as in Definition 4.4(i) and* $\operatorname{gph} \partial g$ *is closed relative to* $\operatorname{dom} g \times \mathbb{R}^n$ *(as is the case when* g *is subdifferentially continuous), then* ω *is made of* δ*-stationary points for* φ*.*

(ii) If the sequence $(\bar{x}^k)_{k\in\mathbb{N}}$ *is (eventually) uniformly locally optimal as in Definition 4.4(ii) (this being true in case of exact prox evaluations, having* $r = \infty$ *and* $\varepsilon_k = 0$ *in this case), then the set* ω *is made of stationary points for* φ*, and* φ *is constantly equal to* $\varphi_\star$ *as in assertion 4.3(i) there.*

Proof. Up to possibly discarding early iterates, in light of the boundedness of the sequences and the consequent eventual constancy of γ_k by Theorem 4.3*(vi)*, we may assume that $\gamma_k \equiv \gamma > 0$ holds for all k. Let $x^\star \in \omega$ be fixed, and let an infinite set of indices $K \subseteq \mathbb{N}$ be such that $(\bar{x}^k)_{k\in K} \to x^\star$, so that $(x^k)_{k\in K} \to x^\star$ too as if follows from Theorem 4.3*(iii)*.

• 4.5*(i)* Since $\nabla f(x^k) + \frac{1}{\gamma}(\bar{x}^k - x^k) \to \nabla f(x^\star)$ as $K \ni k \to \infty$, up to extracting a subsequence if necessary, it follows from (4.9) that $\bar{v}^k \to \bar{v}^\star$ with $\|\bar{v}^\star + \nabla f(x^\star)\| \leq \delta$. Since $(\Phi_k = \mathcal{L}_{1/\gamma}(x^k, \bar{x}^k, -\nabla f(x^k)))_{k\in\mathbb{N}}$ is bounded, owing to Theorem 4.3*(i)*, and since both f and ∇f are continuous, clearly $(g(\bar{x}^k))_{k\in\mathbb{N}}$ remains bounded, and therefore, by lower semicontinuity, $x^\star \in \operatorname{dom} g$. Since also $(\bar{x}^k)_{k\in K} \subseteq \operatorname{dom} g$, from the assumptions we conclude that $\bar{v}^\star \in \partial g(x^\star)$ and thus $\bar{v}^\star + \nabla f(x^\star) \in \partial\varphi(x^\star)$, proving δ-stationarity of $x^\star$ for φ.

• 4.5*(ii)* Letting $\varphi_\star$ be as in Theorem 4.3*(i)* and invoking (4.5), lsc of φ yields $\varphi(x^\star) \leq \varphi_\star$. For k large enough so that $\bar{x}^k$ is r-close to $x^\star$, we have

$$\begin{aligned}
\varphi_\star = \lim_{k\in K} \Phi_k &= \lim_{k\in K} \mathcal{L}_{1/\gamma}(x^k, \bar{x}^k, -\nabla f(x^k)) \\
&\leq \limsup_{k\in K} \mathcal{L}_{1/\gamma}(x^k, x^\star, -\nabla f(x^k)) + \varepsilon_k \\
&= \mathcal{L}_{1/\gamma}(x^\star, x^\star, -\nabla f(x^\star)) = \varphi(x^\star) \leq \varphi_\star,
\end{aligned}$$

owing to continuity of f and ∇f, and the fact that both ε_k and $\|x^k - \bar{x}^k\|$ vanish (the former by assumption and the latter by Theorem 4.3*(iii)*). From the arbitrarity of $x^\star \in \omega$ we conclude that φ is constant on ω with value $\varphi_\star$. Notice further this also shows that $g(\bar{x}^k) \to g(x^\star)$ as $K \ni k \to \infty$. Ekeland's variational principle [21, Prop. 1.43] with $\delta_k = \sqrt{\varepsilon_k}$ ensures for every $k \in K$ (large enough so that $\sqrt{\varepsilon_k} \leq r$) the existence of $\xi^k \in \overline{\mathrm{B}}(\bar{x}^k; \sqrt{\varepsilon_k})$ together with

$$\eta^k \in \hat{\partial}[\mathcal{L}_{1/\gamma}(x^k, \cdot\,, -\nabla f(x^k))](\xi^k) = \nabla f(x^k) + \hat{\partial} g(\xi^k) + \tfrac{1}{\gamma}(\xi^k - x^k)$$

such that $\mathcal{L}_{1/\gamma}(x^k, \xi^k, -\nabla f(x^k)) \leq \Phi_k$ and $\eta^k \in \overline{\mathrm{B}}(0; \sqrt{\varepsilon_k})$. By lsc of g and since $\xi^k \to x^\star$, necessarily $g(\xi^k) \to g(x^\star)$ and the inclusion $-\nabla f(x^\star) \in \partial g(x^\star)$ is then readily obtained, whence the claimed stationarity of $x^\star$ for φ. □

Closedness of $\operatorname{gph} \partial g$ relative to $\operatorname{dom} g \times \mathbb{R}^n$ as required in Theorem 4.5*(i)* is milder than subdifferential continuity of g, which is however general enough to encompass indicator functions of closed sets. The 0-norm is instead an example of a function which is not subdifferentially contintinuous but that complies with the requirement in Theorem 4.5*(i)*. Indeed, notice that

$$\partial g(x) = \hat{\partial} g(x) = E_1 \times \cdots \times E_n, \quad \text{where} \quad E_i = \begin{cases} \mathbb{R} & \text{if } x_i = 0 \\ \{0\} & \text{if } x_i \neq 0 \end{cases}$$

for $g = \|\cdot\|_0$. Consider a sequence $x^k \to x$ along with $\partial g(x^k) \ni v^k \to v$; we will show that $v \in \partial g(x)$, regardless of whether or not $g(x^k)$ converges to $g(x)$. Indeed, if $x_i = 0$, then trivially $v_i \in \mathbb{R} = E_i$. Otherwise, $x_i^k \neq 0$ holds for large enough k, thus necessarily $v_i^k = 0$, and consequently $v_i \in \{0\} = E_i$. Either way, since this holds for every component, we conclude that $v \in \partial g(x)$.

4.2 Termination Criteria

Algorithm 2 runs indefinitely and generates an infinite sequence of iterates $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$. Along its execution, we are compelled to check some suitable conditions for stopping and returning a $\bar{x}^k$ that, in some sense, satisfactorily minimizes φ. The assertion of Theorem 4.3*(v)* guarantees that the standard termination criterion on the residual

$$\tfrac{1}{\gamma_k}\|x^k - \bar{x}^k\| \leq \tfrac{\varepsilon}{2} \tag{4.11}$$

is verified in finite time. However, considering (2.4), a control on the magnitude of $\|\nabla f(x^k) - \nabla f(\bar{x}^k)\|$ must also be imposed in order to guarantee bounds on $\operatorname{dist}(0, \partial\varphi(\bar{x}^k))$. This calls for a strengthened linesearch condition at step 2.5 ensuring also the satisfaction of

$$\|\nabla f(x^k) - \nabla f(\bar{x}^k)\| \leq \tfrac{1}{\gamma_k}\|x^k - \bar{x}^k\|, \tag{4.12}$$

so that, by a triangular inequality argument on (2.4), ε-stationarity of $\bar{x}^k$ (that is, $\operatorname{dist}(0, \partial\varphi(\bar{x}^k)) \leq \varepsilon$) would be guaranteed by (4.11). On the one hand, owing to Assumption A1 the proof of Lemma 4.2*(i)* (and of all other results) would still verbatim apply, meaning that this criterion would not affect the well definedness of Algorithm 2, or in fact any result presented so far. On the other hand, this would require evaluations of $\nabla f(\bar{x}^k)$, otherwise not needed, and thus affect the overall complexity. To account for this fact, a viable solution is to trigger this strengthened linesearch only after (4.11) is first satisfied, at which point the algorithm can terminate whenever (4.11) is verified again.

Note that the same conclusions can be made under suboptimal prox evaluations complying with the local uniformly of Definition 4.4*(ii)*, as long as $\varepsilon_k = 0$ for all k. In case of δ-stationarity as in Definition 4.4*(i)*, instead, the same criterion would guarantee $(\delta + \varepsilon)$-stationarity of the output.

4.3 Nonmonotone Variant

Nonmonotone linesearch procedures often prove beneficial in practice, as they can reduce conservatism in the linesearch and favor larger steps. By patterning the rationale of the ZeroFPR algorithm [28], a nonmonotone linesearch can be readily integrated in PANOC$^+$ at step 2.6 without affecting the finite termination and asymptotic properties asserted in Lemma 4.2 and Theorem 4.3. This is done by changing the definition of Φ_k at step 2.4 into $\Phi_k = (1 - p_k)\Phi_{k-1} + p_k \varphi^{\text{FB}}_{\gamma_k}(x^k)$ for $k > 0$ (with $\varphi^{\text{FB}}_{\gamma_k}(x^k)$ being replaced by $\mathcal{L}_{1/\gamma_k}(x^k, \bar{x}^k, -\nabla f(x^k))$ in the inexact case), where $(p_k)_{k\in\mathbb{N}} \subset (0, 1]$ is any user-selected sequence bounded away from 0. The key observation enabling the possibility to replicate all the convergence results is the inequality $\varphi^{\text{FB}}_{\gamma_k}(x^k) \leq \Phi_k$, which follows from an elementary induction (cf. [28, Lem. 5.1]).

4.4 Adaptive Proximal Gradient Method

By selecting $d^k = \bar{x}^{k-1} - x^{k-1}$ at step 2.2, PANOC$^+$ reduces to the classical proximal gradient method $x^k \in \mathrm{T}_{\gamma_{k-1}}(x^{k-1})$ with an adaptive stepsize. In fact, the descent condition at step 2.6 need not be checked, as it is always satisfied for any τ_k, having $x^k = (1 - \tau_k)\bar{x}^{k-1} + \tau_k(x^k + d^k) = \bar{x}^{k-1}$ independently of the value of τ_k. For this specific choice of the update direction d^k, the algorithm simplifies and reduces to the proximal gradient method with adaptive stepsize selection given in Algorithm 3. Convergence results developed in the general setting of PANOC$^+$ can thus be readily imported, even in the inexact case.

Corollary 4.6 (Convergence of adaptive PG)**.** *All the assertions of Theorems 4.3 and 4.5 remain valid for the iterates generated by Algorithm 3.*

Algorithm 3 Inexact proximal gradient with adaptive γ-stepsize rule

---

Require $\quad x^0 \in \mathbb{R}^n; \quad \gamma_0 \in (0, \gamma_g); \quad \alpha \in (0, 1)$

Initialize $\; \bar{x}^{-1} = x^0, \quad k \leftarrow 0,$ and start from step 3.2

3.1: $\quad \gamma_k \leftarrow \gamma_{k-1}, \;\; x^k \leftarrow \bar{x}^{k-1}$

3.2: $\quad$ Let $\bar{x}^k$ be as in (4.4) (e.g., $\bar{x}^k \in \mathrm{T}_{\gamma_k}(x^k)$)

3.3: $\quad$ if $f(\bar{x}^k) > f(x^k) + \left\langle \nabla f(x^k), \bar{x}^k - x^k \right\rangle + \frac{\alpha}{2\gamma_k}\|\bar{x}^k - x^k\|^2$ then

$\qquad \gamma_k \leftarrow \gamma_k/2,$ and go back to step 3.2

3.4: $\quad k \leftarrow k + 1$ and start the next iteration at step 3.1

We note that the exact version of Algorithm 3, that is, with $\bar{x}^k \in \mathrm{T}_{\gamma_k}(x^k)$ in step 3.2, can be viewed as the monotone PG method outlined in [12, Alg. 3.1] with a slightly more conservative linesearch, since

$$\begin{aligned}\varphi(\bar{x}^k) &\leq f(x^k) + \left\langle \nabla f(x^k), \bar{x}^k - x^k \right\rangle + \tfrac{\alpha}{2\gamma_k}\|\bar{x}^k - x^k\|^2 + g(\bar{x}^k)\\ &\overset{(2.6c)}{=} \varphi_{\gamma_k}^{\text{FB}}(x^k) - \tfrac{1-\alpha}{2\gamma_k}\|\bar{x}^k - x^k\|^2 \leq \varphi(x^k) - \tfrac{1-\alpha}{2\gamma_k}\|\bar{x}^k - x^k\|^2,\end{aligned}$$

where the inequalities follow from step 3.3 and Lemma 2.2*(ii)*. Remarkably, plain continuous differentiability (as opposed to locally Lipschitzian) suffices in the given reference, under a few other technical assumptions. However, the discussion therein is confined to plain PG iterations as in Algorithm 3, while our analysis is more general and captures plain PG as simple byproduct.

5 Conclusions

We investigated an adaptive scheme to appropriately select the proximal stepsize within solvers for fully nonconvex composite optimization, focusing on (and extending) the PANOC framework. Our convergence analysis demonstrates the well-definedness of the algorithm and characterizes its asymptotic properties, possibly in the absence of (global) Lipschitz gradient continuity for the smooth term. Indeed, witnessing the approach's robustness, we considered a setting with possibly inexact proximal mapping oracle for the nonsmooth term, providing suitable conditions for its approximate computation. By means of a detailed illustrative example, we highlighted weaknesses of previous approaches and the crucial steps undertaken in this work, as well as their benefits in terms of convergence guarantees and efficiency. Our findings indicate that, by better capturing the problem's geometry, a more conservative adaptive scheme can yield superior practical performance under weaker conditions. Comprising also arbitrary acceleration directions and nonmonotone variants, these results significantly enlarge the scope of PANOC, both as stand-alone tool for optimization and internal solver within other algorithms, e.g., in ALM and sequential programming approaches.

References

[1] Niccolò Antonello, Lorenzo Stella, Panagiotis Patrinos, and Toon van Waterschoot. Proximal gradient algorithms: Applications in signal processing. *arXiv:1803.01621*, 2020.

[2] Alejandro Astudillo, Joris Gillis, Wilm Decré, Goele Pipeleers, and Jan Swevers. Towards an open toolchain for fast nonlinear MPC for serial robots. *IFAC-PapersOnLine*, 53(2):9814–9819, 2020.

[3] Jonas Berlin, Georg Hess, Anton Karlsson, William Ljungbergh, Ze Zhang, Knut Åkesson, and Per-Lage Götvall. Trajectory generation for mobile robots in a dynamic environment using nonlinear model predictive control. In *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*, pages 942–947, 2021.

[4] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[5] Ernesto G. Birgin and José Mario Martínez. *Practical Augmented Lagrangian Methods for Constrained Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2014.

[6] Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

[7] Silvia Bonettini, Marco Prato, and Simone Rebegoldi. Convergence of inexact forward–backward algorithms using the forward–backward envelope. *SIAM Journal on Optimization*, 30(4):3069–3097, 2020.

[8] Patrick L. Combettes and Jean-Cristophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, New York, 2011.

[9] Alberto De Marchi. Constrained and sparse switching times optimization via augmented Lagrangian proximal methods. In *2020 American Control Conference (ACC)*, pages 3633–3638, Denver, CO, USA, 2020. IEEE.

[10] Brecht Evens, Puya Latafat, Andreas Themelis, Johan Suykens, and Panagiotis Patrinos. Neural network training as an optimal control problem: An augmented Lagrangian approach. *arXiv:2103.14343*, 2021.

[11] Ben Hermans. *Penalty and Augmented Lagrangian Methods for Model Predictive Control*. PhD thesis, KU Leuven, 2021.

[12] Christian Kanzow and Patrick Mehlitz. Convergence properties of monotone and nonmonotone proximal gradient methods revisited. *arXiv:2112.01798*, 2021.

[13] Alexander Katriniok, Pantelis Sopasakis, Mathijs Schuurmans, and Panagiotis Patrinos. Nonlinear model predictive control for distributed motion planning in road intersections using PANOC. In *2019 IEEE 58th Annual Conference on Decision and Control (CDC)*, pages 5272–5278, 12 2019.

[14] Zichong Li, Pin-Yu Chen, Sijia Liu, Songtao Lu, and Yangyang Xu. Rate-improved inexact augmented Lagrangian method for constrained nonconvex optimization. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2170–2178. PMLR, 4 2021.

[15] Tianxiang Liu and Ting Kei Pong. Further properties of the forward–backward envelope with applications to difference-of-convex programming. *Computational Optimization and Applications*, 67(3):489–520, 2017.

[16] Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[17] Klara Pålsson and Emma Svärling. Nonlinear model predictive control for constant distance between autonomous transport robots. Master's thesis, Chalmers University of Technology, 2020.

[18] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

[19] Pieter Pas. A matrix-free nonlinear solver for embedded and large-scale optimization. Master's thesis, KU Leuven, 2021.

[20] Pieter Pas, Mathijs Schuurmans, and Panagiotis Patrinos. Alpaqa: A matrix-free solver for nonlinear MPC and large-scale nonconvex optimization. *arXiv:2112.02370*, 2021.

[21] R. Tyrrell Rockafellar and Roger J.B. Wets. *Variational analysis*, volume 317. Springer, 1998.

[22] Ajay Sathya, Pantelis Sopasakis, Ruben Van Parys, Andreas Themelis, Goele Pipeleers, and Panagiotis Patrinos. Embedded nonlinear model predictive control for obstacle avoidance using PANOC. In *2018 European Control Conference (ECC)*, pages 1523–1528, 2018.

[23] Ajay S. Sathya, Joris Gillis, Goele Pipeleers, and Jan Swevers. Real-time robot arm motion planning and control with nonlinear model predictive control using augmented Lagrangian on a first-order solver. In *2020 European Control Conference (ECC)*, pages 507–512, 2020.

[24] Elias Small, Pantelis Sopasakis, Emil Fresk, Panagiotis Patrinos, and George Nikolakopoulos. Aerial navigation in obstructed environments with embedded nonlinear model predictive control. In *2019 18th European Control Conference (ECC)*, pages 3556–3563, 6 2019.

[25] Pantelis Sopasakis, Emil Fresk, and Panagiotis Patrinos. OpEn: Code generation for embedded nonconvex optimization. *IFAC-PapersOnLine*, 53(2):6548–6554, 2020. 21st IFAC World Congress.

[26] Giorgos Stathopoulos, Harsh Shukla, Alexander Szucs, Ye Pu, and Colin N. Jones. Operator splitting methods in control. *Foundations and Trends in Systems and Control*, 3(3):249–362, 2016.

[27] Lorenzo Stella, Andreas Themelis, Pantelis Sopasakis, and Panagiotis Patrinos. A simple and efficient algorithm for nonlinear model predictive control. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1939–1944, 2017.

[28] Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM Journal on Optimization*, 28(3):2274–2303, 2018.