arXiv:submit/5298572 [math.OC] 19 Dec 2023

Power Proximal Point and Augmented Lagrangian Method

Konstantinos A. Oikonomidis* Alexander Bodard* Emanuel Laude*
Panagiotis Patrinos*

December 19, 2023

Abstract

In this paper we study an unconventional inexact *Augmented Lagrangian Method* (ALM) for convex optimization problems, wherein the penalty term is a potentially non-Euclidean norm raised to a power between one and two. We analyse the algorithm through the lens of a nonlinear *Proximal Point Method* (PPM) applied to the dual problem. While the scheme is well-understood for Euclidean norms our focus in on the non-Euclidean case which prevents us from using standard tools such as the nonexpansiveness of the proximal mapping. To allow for errors in the primal update, we derive two implementable stopping criteria under which we analyze both the global and the local convergence rates of the algorithm. More specifically, we show that the method enjoys a fast sublinear global rate in general and a local superlinear rate under suitable growth assumptions. We experimentally illustrate the merits of our approach on the task of quadratic programming with equality constraints and linear programming with inequality constraints using an off-the-shelf L-BFGS-B oracle for the subproblems.

Keywords: augmented Lagrangian · duality · proximal point

AMS Subject Classification: 65K05 · 49J52 · 90C30

Contents

1 Introduction **2**

2 Power proximal point and augmented Lagrangian method **4**
2.1 Problem class and duality 4
2.2 Dual power proximal point 4

3 Complexity analysis of the power proximal point method under inexactness **6**
3.1 Global complexity under ε-subgradients 6
3.2 Global complexity under ε-proximal points 10
3.3 Local superlinear rate under function growth conditions 12

*KU Leuven, Department of Electrical Engineering (ESAT-STADIUS), Kasteelpark Arenberg 10, 3001 Leuven, Belgium {konstantinos.oikonomidis,alexander.bodard,emanuel.laude,panos.patrinos}@kuleuven.be

4 The inexact power augmented Lagrangian method **14**
4.1 Inexactness criteria 14
4.2 Convergence of the primal-dual gap 15
4.3 Convergence results for problems with linear equality constraints 17
4.4 Solution of the subproblems of the power augmented Lagrangian method . . 18

5 Numerical evaluation **18**
5.1 Linear programming 19
5.2 Quadratic programming 20

6 Conclusion **22**

A Missing proofs **24**
A.1 Proof of Proposition 4.1 24
A.2 Proof of Theorem 4.4 25

1 Introduction

The *Augmented Lagrangian Method* (ALM) introduced in 1969 by Hestenes and Powell [9, 24] is a popular algorithm that allows one to cast a difficult nonsmooth or constrained problem into a sequence of smooth ones. Rockafellar showed in the seventies that, in the convex regime, the augmented Lagrangian method can be interpreted in terms of the *Proximal Point Method* (PPM) applied to the Lagrangian dual problem [25]. However, the classical PPM requires careful selection of the step-sizes to attain fast convergence to a feasible solution [14]. In this work we consider a variant of the augmented Lagrangian method for convex optimization problems that is more robust towards suboptimal selection of the step-size sequences. For that purpose we consider a different proximal point method that is obtained by replacing the squared Euclidean proximal term (prox-function) with a higher-power of a potentially non-Euclidean norm. We refer to this variant of the PPM as the *power* PPM. Specializing to the Euclidean norm raised to some power this is an instance of Luque's nonlinear proximal point method [16, 18, 15]. Its dual counterpart, the nonlinear augmented Lagrangian method was analyzed by the same author [17] expanding upon the work of Bertsekas [1], who studied the algorithm without its connection to the proximal point method. However, both works are focused on the local convergence of the method only. More recently, Nesterov rediscovered Luque's nonlinear proximal point method in the context of tensor methods [21], analyzing its fast global sublinear convergence rate. However, the results in [21] are specialized to smooth minimization problems and Euclidean norms. Furthermore, error tolerances are incorporated in terms of a tailored relative error bound that is impractical in the context of ALM.

Expanding upon [21, 17, 15], we analyze the global and local complexity of the power proximal point method and its dual counterpart, the power augmented Lagrangian method, under two implementable stopping criteria. Particular emphasis is put on the usage of non-Euclidean norms which goes beyond the existing setup in [15], and comes with the challenge of a lack of nonexpansiveness [11]. As a remedy, we leverage the objective function or its anisotropic Moreau envelope [10] as merit functions that replace the more classical distance to the solution.

More precisely, our contributions are summarized as follows:

- In section 3, we develop a global convergence analysis of the power proximal point method for nonsmooth convex minimization problems where the prox-function is either the ℓ_2-norm or the ℓ_{p+1}-norm raised to the power $p+1$. We show that the suboptimality-gap decreases sublinearly at rate $\mathcal{O}(k^{-p})$, with k being the iteration index, using an absolute error tolarenace based on the notion of ε-subgradients. This generalizes the result of [21] to nonsmooth minimization problems, $(p+1)$-norms and a more practical error tolerance in the context of ALM. As the ε-subgradient error condition requires feasibility of the next iterate, we also analyze the algorithm under a more relaxed error condition based on ε-proximal points [25]. For that purpose we analyze the proximal point method from the viewpoint of an inexact gradient method applied to the anisotropic Moreau envelope [12, 10]. Furthermore, we also prove local superlinear convergence rates of the method under a suitable growth condition of the objective function. In the context of the power augmented Lagrangian method this matches the rates proved by Bertsekas [1].

- In section 4, we specialize our results to the augmented Lagrangian setup. In particular, we provide implementable stopping criteria for the subproblem in the ALM loop, that guarantee validity of the error bounds in the dual proximal point iteration. In addition, we show that the ergodic primal-dual gap enjoys a fast global rate of $\mathcal{O}(k^{-p})$, matching the rate of the dual suboptimality gap. Specializing to problems with linear equality constraints, this implies that the ergodic feasibility and suboptimality gaps also decrease at a rate of $\mathcal{O}(k^{-p})$. In addition, we show that the non-ergodic feasiblity and suboptimality gaps globally decrease at rate $\mathcal{O}(k^{-p^2/(p+1)})$. For $p > 1$ both rates improve upon the known rates for the classical ALM which is obtained when $p = 1$.

- In section 5, we experimentally illustrate the merits of this approach on the task of linearly constrained quadratic programming and linear programming where we use an L-BFGS-B oracle for the subproblems. We show in our simulations that the power ALM consistently outperforms the classical ALM in terms of cumulative inner L-BFGS-B-iterations, when both algorithms are executed with a constant penalty parameter. Furthermore our evaluations show that the power ALM with constant penalty parameter achieves a similar and in many cases better performance than the classical ALM with a fine-tuned adaptive schedule for the penalty parameter. This suggests that our approach can be seen as an automatic step-size selection scheme for the classical ALM.

Notation

We denote by $\langle \cdot, \cdot \rangle$ the standard Euclidean inner product and by $\mathbb{R}^n \times \mathbb{R}^m$ the Cartesian product between $\mathbb{R}^n$ and $\mathbb{R}^m$, while $\mathbb{N}$ is the set of the natural numbers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. The effective domain of an extended real-valued function $f : \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ is denoted by $\operatorname{dom} f := \{x \in \mathbb{R}^n : f(x) < \infty\}$, and we say that f is proper if $\operatorname{dom} f \neq \emptyset$; lower semicontinuous (lsc) if $f(\bar{x}) \leq \liminf_{x \to \bar{x}} f(x)$ for all $\bar{x} \in \mathbb{R}^n$. We define by $\Gamma_0(\mathbb{R}^n)$ the class of all proper, lsc convex functions $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and with $\mathcal{C}^k(\mathbb{R}^n)$ the ones which are k times continuously differentiable on $\mathbb{R}^n$. For any functions $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ we define the infimal convolution or epi-addition of f and g as $(f \,\square\, g)(x) = \inf_{y \in \mathbb{R}^n} g(x-y) + f(y)$. We denote by $\mathbb{R}^n_+ = [0, +\infty)^n$ the nonnegative orthant. Otherwise we adopt the notation from [27].

2 Power proximal point and augmented Lagrangian method

This section first introduces the class of problems considered in this work, and formulates both the Langrangian primal and dual problems. Then, the power augmented Lagrangian method is derived as an instance of the power proximal point method applied to the Lagrangian dual problem, generalizing the Euclidean setup [26]. In addition, connections to generalized and sharp Lagrangians [27] are provided.

2.1 Problem class and duality

In this work we consider convex convex-composite problems that take the form

$$\inf_{x\in\mathbb{R}^n} \big\{\varphi(x) \equiv f(x) + g(\mathcal{A}(x))\big\}, \tag{2.1}$$

where $f \in \Gamma_0(\mathbb{R}^n)$ and $g \in \Gamma_0(\mathbb{R}^m)$ and $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^m$ is a nonlinear mapping. In the Rockafellian perturbartion framework for convex duality [27] the primal problem is embedded in a family of parametric minimization problems of the form $\inf_{x\in\mathbb{R}^n} F(x,u)$ where $F : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ is defined by $F(x,u) := f(x) + g(\mathcal{A}(x) + u)$. Then, (2.1) is recovered by the choice $u = 0$ as $\inf_{x\in\mathbb{R}^n} F(x,0)$. By dualizing the linear equality constraint $u = 0$ with a Lagrange multiplier y we obtain the convex-concave Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ of (2.1):

$$L(x,y) = f(x) + \langle A(x), y\rangle - g^*(y). \tag{2.2}$$

The Lagrange dual function is given as

$$\varrho(y) := \inf_{x\in\mathbb{R}^n} L(x,y), \tag{2.3}$$

and the Lagrange dual problem can be written as

$$\sup_{y\in\mathbb{R}^m} \varrho(y) = \sup_{y\in\mathbb{R}^m} \inf_{x\in\mathbb{R}^n} L(x,y). \tag{2.4}$$

For the remainder of this paper we assume that strong duality, i.e., the identity $\sup_{y\in\mathbb{R}^m} \varrho(y) = \inf_{x\in\mathbb{R}^n} \varphi(x)$, holds and primal and dual solutions exist. Sufficient conditions for convexity of φ and strong duality for nonlinear $\mathcal{A}$ in a general conic optimization context are explored in [3, 4]. A classical example is when g is the indicator function of the nonpositive orthant, and the component-functions of $\mathcal{A}$ are convex and smooth, or when g is the indicator of a closed, convex cone and $\mathcal{A}$ is affine. A classical sufficient condition for strong duality in this case is obtained via Slater's condition.

2.2 Dual power proximal point

The algorithm that is presented in this paper is obtained by applying the power proximal point scheme to the Lagrangian dual problem (2.4). For notational convenience we introduce

$$\phi := \tfrac{1}{p+1}\|\cdot\|_r^{p+1}, \tag{2.5}$$

for some $p \geq 1$, where $\|\cdot\|_r$ is either the Euclidean norm, i.e., $r = 2$ or the $(p+1)$-norm, i.e., $r = p+1$. For brevity we omit r and write $\|\cdot\|_s = \|\cdot\|_*$, for $s \in \{2, q+1\}$ with $\frac{1}{p+1} + \frac{1}{q+1} = 1$, for the dual norm in the remainder of the manuscript. We highlight that in applications the

$(p+1)$-norm can have certain advantages over the 2-norm because it leads to a separable ϕ whereas $\frac{1}{p+1}\|\cdot\|_2^{p+1}$ is nonseparable unless $p=1$. In addition, we introduce the epi-scaling $\lambda \star \phi$ of ϕ, defined as $(\lambda \star \phi)(x) = \lambda\phi(\lambda^{-1}x)$ if $\lambda > 0$ and $(\lambda \star \phi)(x) = \delta_{\{0\}}(x)$ otherwise, whose convex conjugate amounts to $(\lambda \star \phi)^* = \lambda\phi^*$. For our specific choice of ϕ we have $(\lambda \star \phi)(x) = \lambda^{-p}\frac{1}{p+1}\|x\|^{p+1}$ and $(\lambda \star \phi)^* = \lambda\frac{1}{q+1}\|\cdot\|_*^{q+1}$ with $1 = \frac{1}{q+1} + \frac{1}{p+1}$.

The algorithm that is studied in this paper amounts to the proximal point method with prox-term ϕ and step-size λ applied to the negative dual function $\psi := -\varrho$:

$$y^{k+1} = \operatorname{prox}_{\psi}^{\lambda\star\phi}(y^k) = \operatorname*{arg\,max}_{y\in\mathbb{R}^m} \ \varrho(y) - \lambda \star \phi(y^k - y). \tag{2.6}$$

Expanding the definition of the Lagrange dual function ϱ, the exact version of the algorithm can be rewritten:

$$(x^{k+1}, y^{k+1}) = \operatorname*{arg\,minimax}_{x\in\mathbb{R}^n, y\in\mathbb{R}^m} \ L(x,y) - \lambda \star \phi(y^k - y), \tag{2.7}$$

where x^k is the primal variable. We introduce the corresponding augmented Lagrangian with parameter λ

$$L_\lambda(x,y) := \sup_{\eta\in\mathbb{R}^m} \ L(x,\eta) - \lambda \star \phi(y-\eta). \tag{2.8}$$

Note that $-L_\lambda(x^{k+1},\cdot)$ is the infimal convolution of $-L(x^{k+1},\cdot)$ and $\lambda\star\phi$ and thus, by [10, Lemma 2.7(ii)], its gradient w.r.t. y amounts to $-\nabla_y L_\lambda(x^{k+1},y^k) = \nabla\phi(\lambda^{-1}(y^k - y^{k+1}))$ for $y^{k+1} := \arg\max_{y\in\mathbb{R}^m} \ L(x^{k+1},y) - \lambda\star\phi(y^k - y)$. Then (2.7) separates into a primal and a consecutive dual update:

$$x^{k+1} = \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \ L_\lambda(x, y^k) \tag{2.9}$$

$$y^{k+1} = y^k + \lambda\nabla\phi^*(\nabla_y L_\lambda(y^{k+1}, y^k)). \tag{2.10}$$

While the dual update (2.10) can typically be solved in closed form given x^{k+1}, the primal update (2.9) is solved inexactly using an iterative algorithm, e.g., a Quasi-Newton method. Since the dual update depends on x^{k+1} the error in the primal update propagates to the dual update. This requires us to analyze the algorithm under a certain error tolerance. We conclude this section with a simple example for the augmented Lagrangian function.

Example 2.1 (linear equality constraints). Let $g = \delta_{\{0\}}$ and $\mathcal{A}$ be an affine linear mapping $\mathcal{A}(x) = Ax - b$ for matrix $A \in \mathbb{R}^{m\times n}$ and $b\in\mathbb{R}^m$. Then $g(\mathcal{A}(x))$ encodes the equality constraint $\mathcal{A}(x) = b$ and the augmented Lagrangian reads, since $(\lambda\star\phi)^* = \lambda\phi^*$ and $\phi(-x) = \phi(x)$:

$$\begin{aligned} L_\lambda(x,y) &= f(x) + \sup_{\eta\in\mathbb{R}^m} \langle Ax - b, \eta\rangle - \lambda\star\phi(y-\eta) \\ &= f(x) + \langle y, A(x) - b\rangle + \lambda\phi^*(Ax - b), \end{aligned} \tag{2.11}$$

where the penalty $\phi^* = \frac{1}{q+1}\|\cdot\|_*^{q+1}$ with $1 = \frac{1}{q+1} + \frac{1}{p+1}$. The gradient of the augmented Lagrangian w.r.t. y amounts to $\nabla_y L_\lambda(x,y) = Ax - b$ and hence the dual update comes out as

$$y^{k+1} = y^k + \lambda\nabla\phi^*(Ax^{k+1} - b). \tag{2.12}$$

Assume that $\lambda \star \phi = \frac{\lambda}{p+1}\|\lambda^{-1}(\cdot)\|^{p+1}$. Then for $p \to \infty$, $\lambda \star \phi$ converges (in the pointwise sense) to the indicator function $\delta_{\lambda B_{\|\cdot\|}}$ of the unit ball corresponding to the norm $\|\cdot\|$ scaled with radius λ. This limit case can be interpreted in terms of a trust-region constraint with radius λ in the proximal point scheme (2.6). Dually, since $\phi^* = \frac{1}{q+1}\|\cdot\|_*^{q+1}$ with $1 = \frac{1}{q+1} + \frac{1}{p+1}$, $q+1 \to 1$ and hence $\lambda\phi^*$ converges pointwisely to the dual norm $\lambda\|\cdot\|_*$ scaled with λ. In this limit case, the function ϕ^* is a so-called exact penalty [2], which guarantees that for λ sufficiently large the constraints are feasible at the minimizer of the augmented Lagrangian for a certain multiplier y^0. Furthermore, in this case the minimizer of $L_\lambda(\cdot, y^0)$ coincides with the minimizer of the original constrained problem (2.1); see [27, Definition 11.60]. This reflects the dual perspective in which proximal point with a trust-region constraint converges in a single step to the solution if y^0 is close to $y^* \in \arg\max \varrho$ and λ is taken sufficiently large.

3 Complexity analysis of the power proximal point method under inexactness

In this section we analyze the global and local convergence rates of the power proximal point method under certain error tolerances when applied to the problem of minimizing a convex, proper lsc function $\psi \in \Gamma_0(\mathbb{R}^n)$, i.e.,

$$\inf_{y \in \mathbb{R}^m} \psi(y). \tag{3.1}$$

The results provided in this section are independent of the usecase of the augmented Lagrangian method. In particular we show that the algorithm converges globally with a fast sublinear rate and transitions locally to fast superlinear convergence if ψ satisfies a mild growth condition.

Classically, the convergence of the proximal point method is derived using the nonexpansiveness of the proximal mapping which guarantees that the distance to a solution is nonincreasing. This is still valid for the power proximal point method if the prox-function is a higher power of the Euclidean norm [16]. However, in this work our focus is on non-Euclidean norms which lack any form of nonexpansiveness. Hence, the distance to the solution cannot be used to show convergence. Instead we rely on the cost ψ or its anisotropic Moreau envelope [10] as merit functions which are in fact the quantities of interest in minimization problems. Specifically, the fast rates are derived relative to the suboptimality gaps.

3.1 Global complexity under ε-subgradients

In this subsection we analyze the global complexity of the power proximal point method under a certain error tolerance. Note that by first-order optimality the update $y^{k+1} = \operatorname{prox}_\psi^{\lambda\star\phi}(y^k)$ can be expressed in terms of finding a pair (v^k, y^{k+1}) that satisfies the following nonlinear system of inclusions,

$$v^k \in \partial\psi(y^{k+1}) \tag{3.2a}$$
$$y^{k+1} = y^k - \nabla\phi^*(v^k). \tag{3.2b}$$

To incorporate error tolerances we replace $\partial\psi(y^{k+1})$ with the ε_k-subdifferential of ψ at y^{k+1}, which is defined as follows. For $\varepsilon > 0$ the vector $v \in \mathbb{R}^n$ is an ε-subgradient of ψ at $\bar{y} \in \operatorname{dom}\psi$

if the following inequality holds true:

$$\psi(y) \geq \psi(\bar{y}) + \langle v, y - \bar{y} \rangle - \varepsilon, \tag{3.3}$$

for all $x \in \operatorname{dom} \psi$. The set of all ε-subgradients at $\bar{y}$ is called the ε-subdifferential of ψ at $\bar{y}$ and is denoted by $\partial_\varepsilon \psi(\bar{y})$. The algorithm that is analyzed in this section is listed in Algorithm 1, which possibly incorporates an average with an anchor point y^0.

Algorithm 1 Inexact power proximal point method with averaging

Require: Choose $y^0 \in \mathbb{R}^m$ and a sequence $\{\theta_k\}_{k \in \mathbb{N}_0}$ with $\theta_k \in (0, 1]$.

for all $k = 0, 1, \ldots$ **do**

Set $w^k = \theta_k y^k + (1 - \theta_k) y^0$.

Choose $\varepsilon_k \geq 0$ and solve

$$v^k \in \partial_{\varepsilon_k} \psi(y^{k+1}) \tag{3.4a}$$

$$y^{k+1} = w^k - \lambda \nabla \phi^*(v^k). \tag{3.4b}$$

end for

Since $\partial \psi(\bar{y}) \subseteq \partial_\varepsilon \psi(\bar{y})$, Algorithm 1 is a strict generalization of the exact version.

To facilitate our analysis we need the following lemma:

Lemma 3.1. *The following result holds for* $\phi = \frac{1}{p+1} \|\cdot\|^{p+1}$ *with* $p \geq 1$. *For any* $x, y \in \mathbb{R}^m$ *the following inequality holds:*

$$\phi(x) \geq \phi(y) + \langle \nabla \phi(y), x - y \rangle + \tfrac{1}{2^{p-1}} \phi(x - y). \tag{3.5}$$

Proof. In the case $\phi = \frac{1}{p+1} \|\cdot\|_2^{p+1}$ the result follows from [22, Lemma 4.2.3]. If $\phi = \frac{1}{p+1} \|\cdot\|_{p+1}^{p+1}$ is chosen, then the result easily follows by considering ϕ as a separable sum of $\frac{1}{p+1} |\cdot|^{p+1}$ and then applying (3.5) to each summand. □

We first prove the following lemma:

Lemma 3.2. *Let* $w \in \mathbb{R}^m$. *Let* (y^+, v) *such that* $v \in \partial_\varepsilon \psi(y^+)$ *and* $y^+ = w - \lambda \nabla \phi^*(v)$. *Then we have for all* $y \in \mathbb{R}^m$*:*

$$\psi(y^+) \leq \psi(y) + \lambda \star \phi(w - y) - \tfrac{1}{2^p} \lambda \star \phi(y^+ - y) + \varepsilon \tag{3.6}$$

Proof. Let $y \in \mathbb{R}^m$. Since $v \in \partial_\varepsilon \psi(y^+)$ we have by definition of the ε-subdifferential:

$$\psi(y) \geq \psi(y^+) + \langle v, y - y^+ \rangle - \varepsilon.$$

Rearranging yields $\psi(y^+) \leq \psi(y) - \langle v, y - y^+ \rangle + \varepsilon$. Since $v = \nabla \phi(\lambda^{-1}(w - y^+)) = \lambda^{-p} \nabla \phi(w - y^+)$ we obtain

$$\psi(y^+) \leq \psi(y) + \lambda^{-p} \langle \nabla \phi(w - y^+), (w - y) - (w - y^+) \rangle + \varepsilon.$$

In light of Lemma 3.1 we can further bound

$$\psi(y^+) \leq \psi(y) + \lambda^{-p} \phi(w - y) - \lambda^{-p} \phi(w - y^+) - \tfrac{\lambda^{-p}}{2^p} \phi(y^+ - y) + \varepsilon. \tag{3.7}$$

Neglecting $-\lambda^{-p} \phi(w - y^+)$ we obtain the claimed result. □

We define the sequences of coefficients $\{A_k\}_{k=0}^{\infty}$ and $\{a_k\}_{k=1}^{\infty}$ with $A_0 := 0$, $A_k := k^{p+1}$ and $a_{k+1} := A_{k+1} - A_k$ for $k \geq 1$ and for the remainder of this subsection we assume that there exists a $y^\star \in \arg\min \psi$. Using Lemma 3.1, we can describe the global convergence of Algorithm 1 for different choices of the sequence $\{\theta_k\}_{k\in\mathbb{N}_0}$. On the one hand, by choosing $\theta_k := \frac{A_k}{A_{k+1}}$, we obtain a scheme in which the update is given in terms of a convex combination of the proximal point and an anchor point. In this case the convergence proof closely follows [6, Theorem 5] for the inexact tensor method with averaging.

On the other hand, when using $\theta_k := 1$ the update depends only on the previous variable. In this setup our analysis is similar to that of [6, Theorem 1] for the Monotone Inexact Tensor Method I, [6, Algorithm 1]. However, a complication arises due to the fact that the evaluation of ψ is not necessarily cheap. For example, in the context of constrained optimization the Lagrangian dual function might not have a simple closed form and as such evaluating it at each iteration might become very expensive. Consequently, a monotonic decrease can no longer be guaranteed and thus the analysis has to be adapted. For this reason we introduce an enlargement of the diameter of the initial level-set [6, Equation 8]. Assuming that the sequence of errors ε_k is summable we define the constant $\beta := \sum_{k=0}^{\infty} \varepsilon_k < \infty$ and the diameter of the level-set

$$D_\beta := \sup\{\|y^\star - y\| : \psi(y) \leq \psi(y^0) + \beta\}. \tag{3.8}$$

We assume that $D_\beta < \infty$ is finite, which happens to be the case when ψ is level-bounded, which is equivalent to $0 \in \operatorname{int}\operatorname{dom}\psi^*$ [27, Theorem 11.8(c)]. The following theorem describes the global convergence of Algorithm 1 for both choices of θ_k. Note that the derived convergence rate for $\theta_k = 1$ is worse than the one for $\theta_k = \frac{A_k}{A_{k+1}}$ since it depends on the diameter of the initial level-set, which can be large. In practice, however, we observed the opposite behavior: The variant with $\theta_k = 1$ performs better throughout our experiments even for $p = 1$ which corresponds to the classical PPM.

Theorem 3.3. *Choose the sequence of errors $\{\varepsilon_k\}_{k\in\mathbb{N}_0}$ such that $\varepsilon_k = \frac{c}{(k+1)^{p+1}}$, for some $c \geq 0$. Let $y^\star \in \arg\min \psi$. Then, for the sequence $\{y^k\}_{k\in\mathbb{N}}$ generated by Algorithm 1 the following statements hold true:*

(i) *If $\theta_k := \frac{A_k}{A_{k+1}}$, then $\psi(y^k) - \psi(y^\star) \leq \frac{\lambda^{-p}(p+1)^p\|y^0 - y^\star\|^{p+1} + c}{k^p}$.*

(ii) *If $\theta_k := 1$, then $\psi(y^k) - \psi(y^\star) \leq \frac{\lambda^{-p}(p+1)^p D_\beta^{p+1} + c}{k^p}$.*

Proof. Regardless of the choice of θ_k, by definition of Algorithm 1 we have that $v^k \in \partial_{\varepsilon_k}\psi(y^{k+1})$ and $y^{k+1} = w^k - \lambda\nabla\phi^*(v^k)$. Define $y := \frac{a_{k+1}}{A_{k+1}}y^\star + \frac{A_k}{A_{k+1}}y^k$. Invoking Lemma 3.2 and using the convexity of ψ we get:

$$\begin{aligned}\psi(y^{k+1}) &\leq \psi(\tfrac{a_{k+1}}{A_{k+1}}y^\star + \tfrac{A_k}{A_{k+1}}y^k) + \lambda \star \phi(w^k - y) + \varepsilon_k \\ &\leq \tfrac{a_{k+1}}{A_{k+1}}\psi(y^\star) + \tfrac{A_k}{A_{k+1}}\psi(y^k) + \lambda \star \phi(w^k - y) + \varepsilon_k.\end{aligned} \tag{3.9}$$

"3.3(i)": For $\theta_k := \frac{A_k}{A_{k+1}}$, we have by definition of y and w^k:

$$w^k - y = \theta_k y^k + (1-\theta_k)y^0 - (1-\theta_k)y^\star - \theta_k y^k = (1-\theta_k)(y^0 - y^\star).$$

Substituting this identity in (3.9) we obtain, since $\lambda \star \phi$ is positively homogeneous with order $p+1$, i.e., $\lambda \star \phi(\alpha x) = |\alpha|^{p+1}(\lambda \star \phi)(x)$ and $1 - \theta_k = \frac{a_{k+1}}{A_{k+1}}$:

$$\psi(y^{k+1}) \leq \tfrac{a_{k+1}}{A_{k+1}}\psi(y^\star) + \tfrac{A_k}{A_{k+1}}\psi(y^k) + (\tfrac{a_{k+1}}{A_{k+1}})^{p+1}(\lambda \star \phi)(y^0 - y^\star) + \varepsilon_k.$$

Rearranging and multiplying both sides with $A_{k+1} > 0$ yields

$$\begin{aligned} A_{k+1}\big(\psi(y^{k+1}) - \psi(y^\star)\big) \leq\ & A_k\big(\psi(y^k) - \psi(y^\star)\big) \\ & + A_{k+1}(\tfrac{a_{k+1}}{A_{k+1}})^{p+1}(\lambda \star \phi)(y^0 - y^\star) + A_{k+1}\varepsilon_k. \end{aligned}$$

Summing the inequality from $k = 0$ to $k = K-1$ we obtain by telescoping, since $A_0 = 0$, $A_k = k^{p+1}$ and $\varepsilon_k = \frac{c}{A_{k+1}}$: $A_K\big(\psi(y^K) - \psi(y^\star)\big) \leq \lambda \star \phi(y^0 - y^\star)\sum_{k=1}^K \frac{a_k^{p+1}}{A_k^p} + Kc$. In light of [6, Equation 35] we have $\sum_{k=1}^K \frac{a_k^{p+1}}{A_k^p} \leq (p+1)^{p+1}K$ and hence

$$A_K\big(\psi(y^K) - \psi(y^\star)\big) \leq (p+1)^{p+1}K(\lambda \star \phi)(y^0 - y^\star) + Kc.$$

Dividing the inequality by A_K yields:

$$\psi(y^K) - \psi(y^\star) \leq \tfrac{(p+1)^{p+1}(\lambda\star\phi)(y^0-y^\star)}{K^p} + \tfrac{c}{K^p} = \tfrac{(p+1)^{p+1}(\lambda\star\phi)(y^0-y^\star)+c}{K^p}.$$

"3.3(ii)": For $\theta_k := 1$, we have that $w^k = y^k$ and since $y^k - y = \frac{a_{k+1}}{A_{k+1}}(y^k - y^\star)$ using the positive homogeneity of ϕ with order $p+1$, we obtain that

$$\psi(y^{k+1}) \leq \tfrac{a_{k+1}}{A_{k+1}}\psi(y^\star) + \tfrac{A_k}{A_{k+1}}\psi(y^k) + (\tfrac{a_{k+1}}{A_{k+1}})^{p+1}(\lambda \star \phi)(y^k - y^\star) + \varepsilon_k.$$

Multiplying both sides with A_{k+1} we have since $a_{k+1} = A_{k+1} - A_k$

$$A_{k+1}\big(\psi(y^{k+1}) - \psi(y^\star)\big) \leq A_k\big(\psi(y^k) - \psi(y^\star)\big) + \tfrac{a_{k+1}^{p+1}}{A_{k+1}^p}(\lambda \star \phi)(y^k - y^\star) + \varepsilon_k A_{k+1}.$$

Summing the inequality from $k = 0$ to $k = K-1$ we obtain since $A_0 = 0$:

$$A_K\big(\psi(y^K) - \psi(y^\star)\big) \leq \sum_{k=0}^{K-1} \tfrac{a_{k+1}^{p+1}}{A_{k+1}^p}(\lambda \star \phi)(y^k - y^\star) + \varepsilon_k A_{k+1}. \tag{3.10}$$

From Lemma 3.2 for $w = y = y^k$ we have $\psi(y^{k+1}) \leq \psi(y^k) + \varepsilon_k$. Summing the inequality from $k = 0$ to $k = K-1$ we obtain $\psi(y^K) - \psi(y^0) \leq \sum_{k=0}^{K-1}\varepsilon_k \leq \beta$, and hence $\|y^K - y^\star\| \leq D_\beta$ for any $K \geq 0$. This implies that $\phi(y^k - y^\star) = \frac{1}{p+1}\|y^\star - y^k\|^{p+1} \leq \frac{1}{p+1}D_\beta^{p+1}$. Thus we can further bound (3.10):

$$A_K\big(\psi(y^K) - \psi(y^\star)\big) \leq \lambda^{-p}\tfrac{1}{p+1}D_\beta^{p+1}\sum_{k=1}^K \tfrac{a_k^{p+1}}{A_k^p} + \varepsilon_k A_{k+1}.$$

We choose $A_K = K^{p+1}$ and by using the fact that $\sum_{k=1}^K \frac{a_k^{p+1}}{A_k^p} \leq (p+1)^{p+1}K$ we get:

$$A_K\big(\psi(y^K) - \psi(y^\star)\big) \leq \lambda^{-p}D_\beta^{p+1}(p+1)^p K + cK.$$

Dividing by A_K we obtain $\psi(y^K) - \psi(y^\star) \leq \frac{\lambda^{-p}D_\beta^{p+1}(p+1)^p + c}{K^p}$. □

3.2 Global complexity under ε-proximal points

The algorithm studied in the previous section incorporates error tolerances via the notion of ε-subgradients and relies on the cost function ψ as a merit function which is potentially extended-valued. As a consequence, the next iterate has to stay within the domain of ψ which can be restrictive in practice. In this section we therefore consider a more relaxed error tolerance that requires the next iterate y^{k+1} to stay close to the true proximal point $\operatorname{prox}_{\psi}^{\lambda\star\phi}(y^k)$, referred to as an ε-proximal point of ψ at y^k:

$$\lambda \star \phi(y^{k+1} - \operatorname{prox}_{\psi}^{\lambda\star\phi}(y^k)) \leq \varepsilon_k, \tag{3.11}$$

where $\varepsilon_k \geq 0$ and $\sum_{k=0}^{\infty} \varepsilon_k < \infty$. Note that this error criterion generalizes the classical inexactness conditions studied in [14]. For that purpose we leverage the real-valued and smooth anisotropic Moreau envelope [12, 10] as a merit function and interpret the power proximal point method in terms of the anisotropic gradient descent method [10] applied to the merit function.

We define the left anisotropic Moreau envelope adapted from [10, Definition 2.5] which is given as the value function of the power proximal point update (2.6): For $\lambda > 0$ and $y \in \mathbb{R}^m$ the *left anisotropic Moreau envelope* or in short anisotropic envelope of ψ with parameter λ at y is defined as

$$\operatorname{env}_{\psi}^{\lambda\star\phi}(y) := \inf_{\eta\in\mathbb{R}^m} \{\psi(\eta) + \lambda \star \phi(y - \eta)\}. \tag{3.12}$$

Thanks to [10, Lemma 2.7(ii)] we have that

$$\nabla \operatorname{env}_{\psi}^{\lambda\star\phi}(y^k) = \nabla\phi(\lambda^{-1}(y^k - \operatorname{prox}_{\psi}^{\lambda\star\phi}(y^k))). \tag{3.13}$$

Rearranging the identity reveals that the exact proximal point scheme (2.6) is equivalent to the anisotropic gradient descent method [10, Algorithm 1] applied to the anisotropic envelope of ψ, since:

$$y^{k+1} = \operatorname{prox}_{\psi}^{\lambda\star\phi}(y^k) = y^k - \lambda\nabla\phi^*(\nabla \operatorname{env}_{\psi}^{\lambda\star\phi}(y^k)). \tag{3.14}$$

In order to prove the main convergence result of this subsection we exploit [10, Proposition 3.1] which reveals that the anisotropic Moreau envelope $\operatorname{env}_{\psi}^{\lambda\star\phi}$ has the anisotropic descent property [10, Definition 2.1] with reference function ϕ:

Definition 3.4. *Let $h \in \mathcal{C}(\mathbb{R}^m)$ and $\phi = \frac{1}{p+1}\|\cdot\|^{p+1}$. We say that h satisfies the anisotropic descent property relative to ϕ with constant L if for all $\eta \in \mathbb{R}^m$*

$$h(y) \leq h(\eta) + \tfrac{1}{L} \star \phi(y - \eta + L^{-1}\nabla\phi^*(\nabla h(\eta))) - \tfrac{1}{L} \star \phi(L^{-1}\nabla\phi^*(\nabla h(\eta))). \tag{3.15}$$

Note that whenever $p = 1$, and $\|\cdot\|_2$ we obtain the classical Euclidean descent lemma by expanding the square. Based on the anisotropic descent inequality above we can prove the following theorem:

Theorem 3.5. *Let $\{y^k\}_{k\in\mathbb{N}_0}$ be any sequence that satisfies (3.11). Then the following statments hold true:*

(i) $\operatorname{env}_{\psi}^{\lambda\star\phi}(y^{k+1}) - \operatorname{env}_{\psi}^{\lambda\star\phi}(y^k) \leq -\lambda \star \phi(y^k - \operatorname{prox}_{\psi}^{\lambda\star\phi}(y^k)) + \varepsilon_k.$

(ii) *If* $\{\varepsilon_k\}_{k\in\mathbb{N}_0}$ *is summable with* $\sum_{i=1}^{\infty}\varepsilon_k \leq \alpha$*, for* $\alpha \geq 0$*, and the set of solutions* $Y^\star = \arg\min\psi$ *is nonempty, then* $\|y^{k+1}-y^k\| \to 0$.

(iii) *If* $Y^\star$ *is bounded, then* $\{y^k\}_{k\in\mathbb{N}_0}$ *remains bounded and all of its limit points lie in* $Y^\star$*. Moreover, if* $\varepsilon_k = \frac{c}{(k+1)^{p+1}}$ *for some* $c \geq 0$*, then for all* $k \geq 1$

$$\operatorname{env}_\psi^{\lambda\star\phi}(y^k) - \operatorname{env}_\psi^{\lambda\star\phi}(y^\star) \leq \frac{\lambda^{-p}D_\beta^{p+1}(p^2+p)^p+c}{k^p}. \tag{3.16}$$

Proof. "3.5(i)": By [10, Proposition 3.1], $\operatorname{env}_\psi^{\lambda\star\phi}(y)$ is anisotropically smooth relative to ϕ with constant $\frac{1}{\lambda}$. Hence the anisotropic descent inequality (3.15) at points $y = y^{k+1}$ and $\eta = y^k$ yields via the relation (3.13):

$$\begin{aligned}
\operatorname{env}_\psi^{\lambda\star\phi}(y^{k+1}) &\leq \operatorname{env}_\psi^{\lambda\star\phi}(y^k) + \lambda\star\phi\big(y^{k+1}-y^k+\lambda\nabla\phi^*(\nabla\phi(\lambda^{-1}(y^k-\operatorname{prox}_\psi^{\lambda\star\phi}(y^k))))\big)\\
&\quad - \lambda\star\phi\big(\lambda\nabla\phi^*(\nabla\phi(\lambda^{-1}(y^k-\operatorname{prox}_\psi^{\lambda\star\phi}(y^k))))\big)\\
&\leq \operatorname{env}_\psi^{\lambda\star\phi}(y^k) + \lambda\star\phi(y^{k+1}-y^k+y^k-\operatorname{prox}_\psi^{\lambda\star\phi}(y^k)) - \lambda\star\phi(y^k-\operatorname{prox}_\psi^{\lambda\star\phi}(y^k))\\
&\leq \operatorname{env}_\psi^{\lambda\star\phi}(y^k) + \varepsilon_k - \lambda\star\phi(y^k-\operatorname{prox}_\psi^{\lambda\star\phi}(y^k)),
\end{aligned} \tag{3.17}$$

where the last inequality follows by the error bound (3.11).

"3.5(ii)": Since $\arg\min\phi = \{0\}$, we have that $\operatorname{env}_\psi^{\lambda\star\phi}(y^\star) = \psi(y^\star)$ and as such that $\psi^\star \leq \operatorname{env}_\psi^{\lambda\star\phi}(y)$ for all $y \in \mathbb{R}^m$. Thus by summing (3.17) form $k = 0$ to K, we get:

$$\begin{aligned}
-\infty < \psi(y^\star) - \operatorname{env}_\psi^{\lambda\star\phi}(y^0) &\leq \operatorname{env}_\psi^{\lambda\star\phi}(y^{K+1}) - \operatorname{env}_\psi^{\lambda\star\phi}(y^0)\\
&\leq -\sum_{k=0}^{K}\lambda\star\phi(y^k-\operatorname{prox}_\psi^{\lambda\star\phi}(y^k)) + \sum_{k=0}^{K}\varepsilon_k \leq -\sum_{k=0}^{K}\lambda\star\phi(y^k-\operatorname{prox}_\psi^{\lambda\star\phi}(y^k)) + \alpha,
\end{aligned}$$

where the last inequality follows by the assumed bound on ε_k. This result readily implies that $\lambda\star\phi(y^k-\operatorname{prox}_\psi^{\lambda\star\phi}(y^k))$ is summable and therefore $\lambda\star\phi(y^k-\operatorname{prox}_\psi^{\lambda\star\phi}(y^k)) \to 0$. Moreover, from (3.11) we have that $0 \leq \lambda\star\phi(y^{k+1}-\operatorname{prox}_\psi^{\lambda\star\phi}(y^k)) \leq \varepsilon_k \to 0$. Recall that $\lambda\star\phi = \lambda^{-p}\frac{1}{p+1}\|\cdot\|^{p+1}$. Hence, we can combine these two convergence results to show the claimed relationship, using the triangle inequality:

$$\|y^k - y^{k+1}\| \leq \|y^k - \operatorname{prox}_\psi^{\lambda\star\phi}(y^k)\| + \|y^{k+1} - \operatorname{prox}_\psi^{\lambda\star\phi}(y^k)\| \to 0.$$

"3.5(iii)": Regarding the boundedness of the iterates we follow the proof of Theorem 3.3(ii). We have that for any $k \in \mathbb{N}_0$ $\operatorname{env}_\psi^{\lambda\star\phi}(y^{k+1}) - \operatorname{env}_\psi^{\lambda\star\phi}(y^0) \leq \alpha$. We define the diameter of the enlarged initial level-set of the envelope function:

$$D_\alpha := \sup\{\|y^\star - y\| : \operatorname{env}_\psi^{\lambda\star\phi}(y) \leq \operatorname{env}_\psi^{\lambda\star\phi}(y^0) + \alpha\}. \tag{3.18}$$

Since $\operatorname{env}_\psi^{\lambda\star\phi}$ is convex and $Y^\star$ is bounded, D_α is also bounded. Therefore, $\|y^k - y^\star\| \leq D_\alpha$- and this implies that $\{y^k\}_{k\in\mathbb{N}_0}$ remains bounded. The subsequential convergence to $Y^\star$ follows from the fact that $\nabla\operatorname{env}_\psi^{\lambda\star\phi}(y^k) \to 0$.

Now, we can turn our attention to the convergence rate of the anisotropic envelope of ψ. Using the convexity inequality, we get:

$$\operatorname{env}_\psi^{\lambda\star\phi}(y) \geq \operatorname{env}_\psi^{\lambda\star\phi}(y^k) + \langle\nabla\operatorname{env}_\psi^{\lambda\star\phi}(y^k), y - y^k\rangle$$

Rearranging and substituting (3.13), we have that

$$\begin{aligned}\mathrm{env}_{\psi}^{\lambda\star\phi}(y^k) &\leq \mathrm{env}_{\psi}^{\lambda\star\phi}(y) - \lambda^{-p}\langle \nabla\phi(y^k - \mathrm{prox}_{\psi}^{\lambda\star\phi}(y^k)), y - y^k\rangle \\ &\leq \mathrm{env}_{\psi}^{\lambda\star\phi}(y) + \lambda^{-p}(\tfrac{1}{\sigma}\|\nabla\phi(y^k - \mathrm{prox}_{\psi}^{\lambda\star\phi}(y^k))\|_*)(\sigma\|y - y^k\|),\end{aligned}$$

for any $\sigma > 0$, where the last inequality follows by Hölder's inequality. We can further bound the r.h.s. of the previous inequality, by using Young's inequality with exponents $q+1$ and $p+1$, the fact that $\|\nabla\phi(y)\|_* = \|y\|^p$ and $p(q+1) = p+1$:

$$\mathrm{env}_{\psi}^{\lambda\star\phi}(y^k) \leq \mathrm{env}_{\psi}^{\lambda\star\phi}(y) + (\tfrac{1}{\sigma})^{q+1}\tfrac{\lambda^{-p}}{q+1}\|y^k - \mathrm{prox}_{\psi}^{\lambda\star\phi}(y^k)\|^{p+1} + \sigma^{p+1}\tfrac{\lambda^{-p}}{p+1}\|y - y^k\|^{p+1}.$$

Let us choose $\sigma = (\frac{p+1}{q+1})^{\frac{1}{q+1}} = p^{\frac{1}{q+1}}$. Then we have since $(\frac{1}{\sigma})^{q+1}\frac{1}{q+1} = \frac{1}{p+1}$ and $\sigma^{p+1} = p^p$:

$$\mathrm{env}_{\psi}^{\lambda\star\phi}(y^k) \leq \mathrm{env}_{\psi}^{\lambda\star\phi}(y) + \lambda\star\phi(y^k - \mathrm{prox}_{\psi}^{\lambda\star\phi}(y^k)) + p^p\lambda\star\phi(y - y^k),$$

and by putting this result back in (3.17) we obtain:

$$\mathrm{env}_{\psi}^{\lambda\star\phi}(y^{k+1}) \leq \mathrm{env}_{\psi}^{\lambda\star\phi}(y) + \varepsilon_k + p^p\lambda\star\phi(y - y^k). \tag{3.19}$$

The rest of the proof is similar to that of Theorem 3.3(ii). More specifically, if we choose $y := \frac{a_{k+1}y^\star + A_k y^k}{A_{k+1}}$ in (3.19) and use the convexity inequality as well as the fact that ϕ is positively homogeneous with order $p+1$, we have that

$$\mathrm{env}_{\psi}^{\lambda\star\phi}(y^{k+1}) \leq \tfrac{a_{k+1}}{A_{k+1}}\mathrm{env}_{\psi}^{\lambda\star\phi}(y^\star) + \tfrac{A_k}{A_{k+1}}\mathrm{env}_{\psi}^{\lambda\star\phi}(y^k) + \varepsilon_k + \tfrac{a_{k+1}^{p+1}}{A_{k+1}^{p+1}}p^p\lambda\star\phi(y^k - y^\star)$$

Multiplying the inequality above with $A_{k+1} > 0$, rearranging and summing from $k = 0$ to $k = K-1$, we get:

$$A_K\big(\mathrm{env}_{\psi}^{\lambda\star\phi}(y^K) - \mathrm{env}_{\psi}^{\lambda\star\phi}(y^\star)\big) \leq \sum_{k=0}^{K-1}\tfrac{a_{k+1}^{p+1}}{A_{k+1}^p}p^p\lambda\star\phi(y^k - y^\star) + \varepsilon_k A_{k+1}. \tag{3.20}$$

If we choose $A_K = K^{p+1}$ and use the fact that $\sum_{k=1}^K \frac{a_k^{p+1}}{A_k^p} \leq (p+1)^{p+1}K$ we have that $A_K\big(\mathrm{env}_{\psi}^{\lambda\star\phi}(y^k) - \mathrm{env}_{\psi}^{\lambda\star\phi}(y^\star)\big) \leq p^p D_\alpha^{p+1}(p+1)^p K + cK$, since $\|y^k - y^\star\| \leq D_\alpha$. Dividing now with A_K, we obtain the claimed result. □

3.3 Local superlinear rate under function growth conditions

In this subsection we establish the local superlinear convergence of the power proximal point scheme under a mild growth condition of the objective ψ.

In the following, let $Y^\star$ denote the set of minimizers of ψ, which we assume is nonempty and compact, and $\psi^\star$ be the minimum value of ψ. Let $d(y, Y^\star) := \inf_{\eta\in Y^\star}\|y - \eta\|_r$ for $r \in \{2, p+1\}$ denote the distance of y to $Y^\star$. For brevity we again omit r.

The following result captures the local superlinear rate of the algorithm with exact updates. A similar result is provided in [1, Proposition 5.20] for the Augmented Lagrangian method.

Theorem 3.6. *Let $\{y^k\}_{k\in\mathbb{N}_0}$ be the sequence of iterates generated by Algorithm 1 for $\theta_k = 1$, $\varepsilon_k = 0$. Moreover assume that $Y^\star$ is compact and there exists a δ-neighborhood $\mathcal{N}(Y^\star,\delta) := \{y : d(y,Y^\star) < \delta\}$ of $Y^\star$ such that $\psi(y) - \psi^\star \geq \mu d^\nu(y,Y^\star)$, with $\nu > 1$, for all $y \in \mathcal{N}(Y^\star,\delta)$. Let $p > \nu - 1$. Then, $\{d(y^k,Y^\star)\}_{k\in\mathbb{N}_0}$ converges locally Q-superlinearly with order $\omega := \frac{p}{\nu-1} > 1$, i.e., for k sufficiently large we have*

$$\frac{d(y^{k+1},Y^\star)}{d^\omega(y^k,Y^\star)} \leq \Big(\frac{1}{\lambda\mu^q}\Big)^\omega. \tag{3.21}$$

Proof. For a vector $y \in \mathbb{R}^m$, we denote by $y_\star$ its unique projection on $Y^\star$, i.e. $y_\star = \arg\min_{y^\star\in Y^\star}\|y - y^\star\|$. First, let us remark that since we have assumed the exact execution of the algorithm, the results of Theorem 3.5 hold. More specifically, since $Y^\star$ is assumed compact, every subsequence of $\{y^k\}_{k\in\mathbb{N}_0}$ converges to some point in $Y^\star$ and thus $d(y^k,Y^\star) \to 0$. Therefore, there exists a $K \geq 0$ such that for all $k \geq K$, $y^{k+1} \in \mathcal{N}(Y^\star,\delta)$. Hence we have that $\mu d^\nu(y^{k+1},Y^\star) \leq \psi(y^{k+1}) - \psi^\star$. By (3.2) we have that $-\lambda^{-p}\nabla\phi(y^{k+1} - y^k) \in \partial\psi(y^{k+1})$ and hence using the convexity inequality we obtain that $\psi^\star \geq \psi(y^{k+1}) - \lambda^{-p}\langle\nabla\phi(y^{k+1} - y^k), y_\star^{k+1} - y^{k+1}\rangle$. Thus we have:

$$\begin{aligned}
\mu d^\nu(y^{k+1},Y^\star) &\leq -\psi^\star + \psi(y^{k+1}) \\
&\leq \lambda^{-p}\langle\nabla\phi(y^{k+1} - y^k), y_\star^{k+1} - y^{k+1}\rangle \\
&\leq \lambda^{-p}\|\nabla\phi(y^{k+1} - y^k)\|_*\|y^{k+1} - y_\star^{k+1}\| \\
&= \lambda^{-p}\|y^{k+1} - y^k\|^p d(y^{k+1},Y^\star),
\end{aligned} \tag{3.22}$$

where the third inequality follows from Hölder's inequality and the equality follows from the fact that $\|\nabla\phi(x)\|_* = \|x\|^p$ for any $x \in \mathbb{R}^m$. Now, for the update of the algorithm we have that $y^{k+1} = \arg\min_{y\in\mathbb{R}^m}\{\psi(y) + \lambda\star\phi(y - y^k)\}$ and hence:

$$\lambda\star\phi(y^{k+1} - y^k) \leq \psi(y) - \psi(y^{k+1}) + \lambda\star\phi(y^k - y), \qquad \forall y \in \mathbb{R}^m.$$

Thus, by choosing $y = y_\star^k \in Y^\star$, since $\psi^\star - \psi(y^{k+1}) \leq 0$ we obtain since $\lambda\star\phi = \lambda^{-p}\|\cdot\|^{p+1}$:

$$\|y^{k+1} - y^k\| \leq \|y^k - y_\star^k\| = d(y^k,Y^\star).$$

substituting the inequality above in (3.22), we get:

$$d^{\nu-1}(y^{k+1},Y^\star) \leq \tfrac{1}{\lambda^p\mu}d^p(y^k,Y^\star)$$

and by taking both sides to the power of $\frac{1}{\nu-1}$ we obtain the desired result noting that $\frac{1}{\lambda^p\mu}^{\frac{1}{\nu-1}} = (\frac{1}{\lambda\mu^q})^\omega$ and $\frac{1}{p} = q$. □

Although Theorem 3.6 describes the local convergence of the exact version of the algorithm, the extension to the inexact case is straightforward due to the equivalence of the norms in $\mathbb{R}^m$. More specifically, one could follow the analysis of [16] and further specify the inexactness condition (3.11) as follows: $\|y^{k+1} - \mathrm{prox}_\psi^{\lambda\star\phi}(y^k))\| \leq \varepsilon_k\min\{1,\|y^{k+1} - y^k\|^t\}$, for some $t > 0$ and thus obtain a local superlinear rate of at least $\min\{t, \frac{1}{q(\nu-1)}\}$. It is important to note that for $\phi = \frac{1}{p+1}\|\cdot\|_2^{p+1}$ as in [16], it is not necessary to assume that $Y^\star$ is compact. To the best of our knowledge, however, such a condition cannot be avoided when dealing with $(p+1)$-norms.

4 The inexact power augmented Lagrangian method

In this section we specialize the results of the inexact power proximal point method to the augmented Lagrangian setting as described in section 2. In particular, we provide implementable stopping criteria for the primal update x^{k+1} that guarantee that the dual update y^{k+1} is within the error tolerance of the inexact power proximal point method when applied to the Lagrangian dual function. Furthermore, we show that the ergodic primal-dual gap vanishes at the same rate as the dual suboptimality gap established in the previous section and provide global convergence results for important quantities in the constrained optimization framework.

4.1 Inexactness criteria

Recall that $\psi := -\varrho$ is the negative Lagrangian dual function ϱ which is defined in (2.3). This subsection describes stopping criteria for the primal update

$$x^+ \approx \operatorname*{arg\,min}_{x\in\mathbb{R}^n} L_\lambda(x, y), \tag{4.1}$$

which guarantee that the corresponding dual update

$$y^+ = y + \lambda\nabla\phi^*(\nabla_y L_\lambda(x^+, y)). \tag{4.2}$$

is within the error tolerance for the inexact power proximal point scheme applied to ψ. For that purpose we prove the following result:

Proposition 4.1. *The following properties hold for x^+ and y^+:*

(i) *Let x^+ such that $L(x^+, y^+) \le \inf L(\cdot, y^+) + \varepsilon$. Then $-\nabla_y L_\lambda(x^+, y) \in \partial_\varepsilon\psi(y^+)$.*

(ii) *Let x^+ such that $L_\lambda(x^+, y) \le \inf L_\lambda(\cdot, y) + \varepsilon$. Then y^+ is an $2^{p-1}\varepsilon$-proximal point, i.e., $\lambda \star \phi(y^+ - \operatorname{prox}_\psi^\phi(y)) \le 2^{p-1}\varepsilon$.*

(iii) *Assume that $\operatorname{dom} f$ is bounded. Let $e \in \partial_x L_\lambda(x^+, y)$ such that $\|e\| \le \varepsilon/D$, where $D := \sup_{x,y\in\operatorname{dom} f}\|x - y\| < \infty$ is the diameter of $\operatorname{dom} f$. Then the hypotheses of the first and second item are valid.*

A proof of this result is included in Appendix A.1.

Based on the above result we can now specialize Algorithm 1 to the inexact augmented Lagrangian method listed in Algorithm 2.

Algorithm 2 Inexact power augmented Lagrangian method with averaging

Require: Choose $y^0 \in \mathbb{R}^m, x^0 \in \mathbb{R}^n$ and a sequence $\{\theta_k\}_{k\in\mathbb{N}_0}$ with $\theta_k \in (0, 1]$.

for all $k = 0, 1, \dots$ **do**

Set $w^k = \theta_k y^k + (1 - \theta_k)y^0$.

Choose $\varepsilon_k > 0$ and solve

$$x^{k+1} \approx \operatorname*{arg\,min}_{x\in\mathbb{R}^n} L_\lambda(x, w^k) \tag{4.3a}$$

$$y^{k+1} = w^k + \lambda\nabla\phi^*(\nabla_y L_\lambda(x^{k+1}, w^k)). \tag{4.3b}$$

end for

4.2 Convergence of the primal-dual gap

In this subsection we refine the analysis from the previous section in the primal-dual setup. Specifically, we show that the primal-dual gap decreases at the same rate as the dual sub-optimality-gap; see Theorem 3.3. For that purpose we define for any $K \geq 1$ the ergodic primal iterate:

$$\breve{x}^K = \frac{\sum_{k=1}^K a_k x^k}{\sum_{k=1}^K a_k}. \tag{4.4}$$

In the following theorem we show that the ergodic primal-dual gap decreases at the same rate as the negative dual function. The proof is along the lines of the analysis in [28], albeit adapted to the higher-order proximal-point setting where in general nonexpansiveness is not present and as such standard techniques are not directly applicable.

Theorem 4.2. *Choose the sequence of errors $\{\varepsilon_k\}_{k\in\mathbb{N}_0}$ such that $\varepsilon_k = \frac{c}{(k+1)^{p+1}}$, for some $c \geq 0$. Let $(x^k, y^k)_{k\in\mathbb{N}_0}$ be the sequence of iterates generated by Algorithm 2 with $\theta_k := \frac{A_k}{A_{k+1}}$ and termination criterion $L(x^{k+1}, y^{k+1}) \leq \inf L(\cdot, y^{k+1}) + \varepsilon_k$. Then we have for all $K \geq 1$ and any $x \in \mathbb{R}^n$ and any $y \in \mathbb{R}^m$:*

$$L(\breve{x}^K, y) - L(x, y^K) \leq \frac{c+\lambda\star\phi(y^0-y)(p+1)^{p+1}}{K^p}. \tag{4.5}$$

Proof. Choose any $w \in \mathbb{R}^m$. We have by the y-update (2.10) that $v^k = \lambda^{-p}\nabla\phi(w^k - y^{k+1}) = \nabla\phi(\lambda^{-1}(w^k - y^{k+1})) \in \partial_y(-L)(x^{k+1}, y^{k+1})$ and hence via concavity of $L(x^{k+1}, \cdot)$:

$$L(x^{k+1}, w) \leq L(x^{k+1}, y^{k+1}) + \lambda^{-p}\langle \nabla\phi(w^k - y^{k+1}), (w^k - w) - (w^k - y^{k+1})\rangle. \tag{4.6}$$

By uniform convexity of ϕ invoking Lemma 3.1 we can further bound:

$$\begin{aligned} L(x^{k+1}, w) &\leq L(x^{k+1}, y^{k+1}) + \lambda^{-p}\phi(w^k - w) - \lambda^{-p}\phi(w^k - y^{k+1}) - \lambda^{-p}\tfrac{1}{2^p}\phi(y^{k+1} - w) \\ &\leq L(x^{k+1}, y^{k+1}) + \lambda^{-p}\phi(w^k - w). \end{aligned}$$

Let $y \in \mathbb{R}^m$. Plugging $w := (1-\theta_k)y + \theta_k y^k$ in the identity above we obtain by concavity of $L(x^{k+1}, \cdot)$, using $w^k - w = \theta_k y^k + (1-\theta_k)y^0 - (1-\theta_k)y - \theta_k y^k = (1-\theta_k)(y^0 - y)$ and the homgeneity of ϕ with order $p+1$:

$$(1-\theta_k)L(x^{k+1}, y) + \theta_k L(x^{k+1}, y^k) \leq L(x^{k+1}, y^{k+1}) + \lambda^{-p}(1-\theta_k)^{p+1}\phi(y^0 - y).$$

Since $\theta_k = \frac{A_k}{A_{k+1}}$, and $1 - \theta_k = \frac{a_{k+1}}{A_{k+1}}$ we obtain:

$$\tfrac{a_{k+1}}{A_{k+1}}L(x^{k+1}, y) + \tfrac{A_k}{A_{k+1}}L(x^{k+1}, y^k) \leq L(x^{k+1}, y^{k+1}) + \lambda^{-p}(\tfrac{a_{k+1}}{A_{k+1}})^{p+1}\phi(y^0 - y).$$

Multiplication with A_{k+1} yields:

$$a_{k+1}L(x^{k+1}, y) + A_k L(x^{k+1}, y^k) \leq A_{k+1}L(x^{k+1}, y^{k+1}) + \lambda^{-p}\tfrac{a_{k+1}^{p+1}}{A_{k+1}^p}\phi(y^0 - y).$$

Using the relation $a_{k+1} = A_{K+1} - A_k$ the inequality can be rewritten:

$$A_{k+1}\big(L(x^{k+1}, y) - L(x^{k+1}, y^{k+1})\big) \leq A_k\big(L(x^{k+1}, y) - L(x^{k+1}, y^k)\big) + \lambda^{-p}\tfrac{a_{k+1}^{p+1}}{A_{k+1}^p}\phi(y^0 - y). \tag{4.7}$$

Thanks to the error bound for the x-update in Algorithm 2 and using the convention $\varepsilon_{-1} := L(x^0, y^0) - \inf L(\cdot, y^0)$ we have that for all $k \geq 0$, $L(x^k, y^k) \leq \inf L(\cdot, y^k) + \varepsilon_{k-1} \leq L(x^{k+1}, y^k) + \varepsilon_{k-1}$. This yields

$$L(x^{k+1}, y) - L(x^k, y) + L(x^k, y) - L(x^k, y^k) + \varepsilon_{k-1} \geq L(x^{k+1}, y) - L(x^{k+1}, y^k). \tag{4.8}$$

Plugging this inequality in (4.7) we obtain:

$$\begin{aligned} A_{k+1}\big(L(x^{k+1}, y) - L(x^{k+1}, y^{k+1})\big) \leq A_k\big(L(x^k, y) - L(x^k, y^k)\big) + A_k\big(L(x^{k+1}, y) - L(x^k, y)\big) \\ + A_k\varepsilon_{k-1} + \tfrac{a_{k+1}^{p+1}}{A_{k+1}^p}(\lambda \star \phi)(y^0 - y). \end{aligned}$$

Summing the inequality from $k = 0$ to $k = K - 1$ we obtain since $A_0 = 0$ using the fact that $\sum_{k=1}^K \frac{a_k^{p+1}}{A_k^p} \leq (p+1)^{p+1}K$:

$$\begin{aligned} A_K L(x^K, y) - A_k L(x^K, y^K) \leq \sum_{k=0}^{K-1} A_k\big(L(x^{k+1}, y) - L(x^k, y)\big) \\ + \sum_{k=0}^{K-2} A_{k+1}\varepsilon_k + \lambda \star \phi(y^0 - y)(p+1)^{p+1}K. \end{aligned}$$

By reordering the sum using $a_{k+1} = A_{k+1} - A_k$ and $A_0 = 0$ we have the identity:

$$A_K L(x^K, y) - \sum_{k=0}^{K-1} A_k\big(L(x^{k+1}, y) - L(x^k, y)\big) = \sum_{k=1}^{K} a_k L(x^k, y).$$

This allows us to rewrite the previous inequality

$$\sum_{k=1}^{K} a_k L(x^k, y) - A_K L(x^K, y^K) \leq \sum_{k=0}^{K-2} A_{k+1}\varepsilon_k + \lambda \star \phi(y^0 - y)(p+1)^{p+1}K. \tag{4.9}$$

By convexity of $L(\cdot, y)$ since $A_K = \sum_{j=1}^K a_j$ we have:

$$A_K L(\breve{x}^K, y) \leq A_K \sum_{k=1}^{K} \tfrac{a_k}{\sum_{j=1}^K a_j} L(x^k, y) = \sum_{k=1}^{K} a_k L(x^k, y). \tag{4.10}$$

Plugging this relation in (4.9) we can lower bound:

$$A_K\big(L(\breve{x}^K, y) - L(x^K, y^K)\big) \leq \sum_{k=0}^{K-2} A_{k+1}\varepsilon_k + \lambda \star \phi(y^0 - y)(p+1)^{p+1}K.$$

By the x-update we have for any $x \in \mathbb{R}^n$, $L(x^K, y^K) \leq \inf L(\cdot, y^K) + \varepsilon_{K-1} \leq L(x, y^K) + \varepsilon_{K-1}$ and hence we can lower bound since $A_{k+1}\varepsilon_k = c$

$$A_K\big(L(\breve{x}^K, y) - L(x, y^K)\big) \leq cK + \lambda \star \phi(y^0 - y)(p+1)^{p+1}K.$$

Dividing by A_k we obtain $L(\breve{x}^K, y) - L(x, y^K) \leq \frac{c + \lambda \star \phi(y^0 - y)(p+1)^{p+1}}{K^p}$ as claimed. □

Note that this result is not specific to the Lagrangian function as it holds for any convex-concave function, where the inexact higher-order proximal point method is applied to the concave part.

4.3 Convergence results for problems with linear equality constraints

In this subsection we revisit Example 2.1 specializing to the problem of minimizing a convex function f subject to linear equality constraints. First, we provide results for the ergodic iterates, harnessing the analysis of the primal-dual gap that is presented in Theorem 4.2.

Theorem 4.3. *In the situation of Example 2.1 let $\{\breve{x}^K\}_{K\in\mathbb{N}}$ be the ergodic primal iterates of Algorithm 2 as defined in (4.4) with $\theta_k = \frac{A_k}{A_{k+1}}$, $\varepsilon_k = \frac{c}{(k+1)^{p+1}}$ for some $c \geq 0$ and termination criterion $L(x^{k+1}, y^{k+1}) \leq \inf L(\cdot, y^{k+1}) + \varepsilon_k$ for all $k \in \mathbb{N}_0$. Then, the following inequality holds for $K \geq 1$:*

$$\max\{\|A\breve{x}^K - b\|_*, |f(\breve{x}^K) - f(x^\star)|\} \leq \frac{c+\max_{y\in\delta\mathcal{B}} \lambda\star\phi(y^0-y)(p+1)^{p+1}}{K^p}, \tag{4.11}$$

where $\delta = 2\|y^\star\| + 1$, $\mathcal{B} := \{y \in \mathbb{R}^m : \|y\| \leq 1\}$ denotes the closed unit ball, and $(x^\star, y^\star)$ is a saddle-point of the problem.

Proof. From Theorem 4.2 we have that for $K \geq 1$ and $x \in \mathbb{R}^n, y \in \mathbb{R}^m$

$$L(\breve{x}^K, y) - L(x, y^K) \leq \frac{c+\lambda\star\phi(y^0-y)(p+1)^{p+1}}{K^p}$$

and since for $x = x^\star$, $L(\breve{x}^K, y) - L(x^\star, y^K) = f(\breve{x}^K) - f(x^\star) + \langle A\breve{x}^K - b, y\rangle$ we obtain the following inequality for all $y \in \mathbb{R}^m$:

$$f(\breve{x}^K) - f(x^\star) + \langle A\breve{x}^K - b, y\rangle \leq \frac{c+\lambda\star\phi(y^0-y)(p+1)^{p+1}}{K^p}. \tag{4.12}$$

By taking $y = 0$ in the inequality above we obtain

$$f(\breve{x}^K) - f(x^\star) \leq \frac{c+\lambda\star\phi(y^0)(p+1)^{p+1}}{K^p}. \tag{4.13}$$

Now, let $\delta = 2\|y^\star\| + 1$ and take the maximum in (4.12) w.r.t. $y \in \delta\mathcal{B}$:

$$f(\breve{x}^K) - f(x^\star) + \delta\|A\breve{x}^K - b\|_* \leq \frac{c+\max_{y\in\delta\mathcal{B}} \lambda\star\phi(y^0-y)(p+1)^{p+1}}{K^p}. \tag{4.14}$$

Since $(x^\star, y^\star)$ is a saddle point, it holds that $L(x^\star, y) \leq L(x^\star, y^\star) \leq L(x, y^\star)$ for any $x \in \mathbb{R}^n, y \in \mathbb{R}^m$ and as such $L(x, y^\star) - L(x^\star, y) \geq 0$. Choose $x = \breve{x}^K$ and $y = 0$. Then we obtain

$$\begin{aligned} 0 &\leq L(\breve{x}^K, y^\star) - L(x^\star, 0) \\ &= f(\breve{x}^K) - f(x^\star) + \langle A\breve{x}^K - b, y^\star\rangle \\ &\leq f(\breve{x}^K) - f(x^\star) + \|y^\star\|\|A\breve{x}^K - b\|_*, \end{aligned} \tag{4.15}$$

where the last inequality follows by applying Hölder's inequality. Summing now (4.14) and (4.15), we get since $\delta > \|y^\star\|$

$$\|A\breve{x}^K - b\|_* \leq (\delta - \|y^\star\|)\|A\breve{x}^K - b\|_* \leq \frac{c+\max_{y\in\delta\mathcal{B}} \lambda\star\phi(y^0-y)(p+1)^{p+1}}{K^p}. \tag{4.16}$$

From the inequality above together with (4.15) we obtain:

$$\begin{aligned} f(x^\star) - f(\breve{x}^K) \leq \|y^\star\|\|A\breve{x}^K - b\|_* &\leq (\delta - \|y^\star\|)\|A\breve{x}^K - b\|_* \\ &\leq \frac{c+\max_{y\in\delta\mathcal{B}} \lambda\star\phi(y^0-y)(p+1)^{p+1}}{K^p}, \end{aligned} \tag{4.17}$$

which along with (4.13) implies the claimed result. □

Moving on to the analysis of the nonergodic iterates, the following theorem compiles the corresponding results.

Theorem 4.4. *In the situation of Example 2.1 let* $\{x^k\}_{k\in\mathbb{N}_0}$ *be the primal iterates of Algorithm 2 with* $\theta_k = 1$, $\varepsilon_k = \frac{c}{(k+1)^{p+1}}$ *and termination criterion* $L(x^{k+1}, y^{k+1}) \leq \inf L(\cdot, y^{k+1}) + \varepsilon_k$ *for all* $k \in \mathbb{N}_0$. *Then, the following hold for* $k \geq 1$*:*

(i) $\|Ax^k - b\|_* = \mathcal{O}\left(k^{-\frac{p^2}{p+1}}\right)$

(ii) $|f(x^k) - f(x^*)| = \mathcal{O}\left(k^{-\frac{p^2}{p+1}}\right)$.

The proof is presented in Appendix A.2.

Notice that by choosing $p = 1$, we obtain the rates described in [13, Corollary 4.4]. In this sense, it can be considered as a generalization of that result in the power ALM setting. Moreover, such rates can be shown, albeit with different constants, regardless of the specific algorithm.

4.4 Solution of the subproblems of the power augmented Lagrangian method

Again we specialize to the minimization of a convex cost subject to linear equality constraints as in Example 2.1. Although the convergence results of the power ALM in terms of outer iterations seem quite attractive when using $p > 1$, the inner subproblems become harder to solve. Notice that when choosing $\phi = \frac{1}{p+1}\|\cdot\|_2^{p+1}$ the term $\|Ax - b\|_2^{q+1}$ in (2.11) is Hölder smooth of order q. In general, such a term slows down the convergence of first-order methods as shown in [7]. More specifically, if f is an L-smooth function, the augmented Lagrangian (2.11) as a function in x is comprised of functions with varying order of Hölder smoothness. This stands in contrast to the classical ALM, where the augmented Lagrangian is Lipschitz smooth and hence classical first-order methods such as the accelerated proximal gradient method can be applied. Nevertheless, again from [7] we know that the inexact version of this algorithm, specifically the one presented in [20], can be used to minimize (2.11) as well.

In our experiments we chose a different way to tackle the inner problems. It is well-known that although first-order methods have appealing global convergence rates, they can be rather slow in practice. Especially in the setting of the augmented Lagrangian method, where a nonsmooth problem is cast into a sequence of smooth ones, second-order information can substantially accelerate convergence in practice. Therefore, we apply off-the-shelf L-BFGS-B and BFGS solvers. We observed that, even if applied heuristically, these methods behave surprisingly well in practice which often leads to a reduced total number of iterations when compared to the classical Augmented Lagrangian Method.

5 Numerical evaluation

This section compares the performance of power ALM, Algorithm 2 with $\theta_k = 1$, on both linear and quadratic programs for powers $q + 1$, with $q \in \{1/2, 1/3, 1\}$. Note that the choice $q = 1$ is equivalent to the classical ALM. We also compare against classical ALM using an adaptive penalty parameter λ_k, since this is known to yield good performance in practice, see e.g. [23, §17.4] and [8, §5.3]. In particular, define a sequence of errors $\{e_k\}_{k\in\mathbb{N}_0}$ with $e_0 = 1$

and the absolute constraint violation r_k. We consider an adaptive rule that sets $e_{k+1} = e_k/2$, $\lambda_{k+1} = 2\lambda_k$ whenever $r_{k+1} \leq e_k$, and $e_{k+1} = e_k$, $\lambda_{k+1} = \lambda_k$ otherwise. Note that this ensures $\lambda_k \to \infty$ as $k \to \infty$. The solvers are terminated if $|f(x^{k+1}) - f^\star| \leq 10^{-6}$ and $r_{k+1} \leq 10^{-6}$. As a termination criterion for the inner problems we use $\|\nabla_x L_\lambda(x^{k+1}, y^k)\| \leq \frac{c_\varepsilon}{k^{p+1}}$ for some $c_\varepsilon > 0$, i.e., the criterion supported by the theory. This value c_ε is tuned for each solver. The code to reproduce these experiments is available online[1].

5.1 Linear programming

Table 1: Tuned parameters of power ALM for solving the LP.

	$q = 1/2$	$q = 1/3$	q=1 Fixed λ	q=1 Adaptive λ
λ (λ_0 if adaptive)	10	1	1000	100
c_ε	10^{-3}	10^{1}	10^{-6}	10^{-6}

Table 2: Median and 95th percentile (P95) of the total number of BFGS iterations required to solve the LP, computed over $N = 20$ random realizations.

	$q = 1/2$		$q = 1/3$		q=1 Fixed λ		q=1 Adaptive λ	
(m, n)	Med.	P95	Med.	P95	Med.	P95	Med.	P95
(200, 100)	**850.5**	1,001.5	1,078.5	1,321.9	1,271	1,528.1	947	1,168
(400, 200)	**1,359**	1,618.9	1,611.5	2,190.3	1,780	2,540.8	1,505.5	1,841.1
(600, 300)	2,218.5	2,592.8	2,369.5	3,077.9	2,596.5	3,176.6	**2,206**	2,884.3
(300, 100)	**778.5**	919.4	1,079	1,436.7	815.5	1,071.8	919.5	1,061.5
(600, 200)	**1,254.5**	2,641.4	1,864.5	3,361.4	1,577.5	1,836.3	1,526.5	1,771
(400, 100)	**794**	1,011.5	1,294	1,532.3	880	1,489.5	906	1,072.3
(500, 100)	862	1,068	1,408	1,863.5	**840.5**	1,193.3	988.5	1,221.6
(600, 100)	**845.5**	1,102.2	1,514.5	1,970.2	892	1,314.8	1,043.5	1,258.8

Consider a linear program with inequality constraints which is a instance of (2.1) via the choices $f(x) := \langle c, x\rangle$, $g := \delta_{-\mathbb{R}^m_+}$ and $\mathcal{A}(x) = Ax - b$ where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. The problems are randomly generated such that $\text{cond}(A) = 1000$ using the method described in [19]. We used the implementation of the same authors which is available online[2]. The inner problems are solved using BFGS. Tuned parameters λ and c are listed in Table 1.

Table 2 compares power ALM with $(q+1)$-norms in terms of the total number of BFGS iterations. The table reports both the median and the 95th percentile (P95) of the total number of BFGS iterations over $N = 20$ random realizations of (2.1), and this for various dimensions (m, n) of the problem.

[1] https://github.com/alexanderbodard/Power-PPM-and-ALM
[2] https://github.com/QCOL-LU/QIPM.

Table 3: Tuned parameters of power ALM for solving the QP.

	$q = 1/2$	$q = 1/3$	q=1 Fixed λ	q=1 Adaptive λ
λ (λ_0 if adaptive)	0.05	0.05	0.1	0.1
c_ε	10^1	10^2	10^{-2}	10^{-2}

5.2 Quadratic programming

We consider random positive semi-definite quadratic programs. This is an instance of (2.1) via the choices $f(x) = \delta_{[\ell,u]^n}(x) + \frac{1}{2}\langle x, Qx\rangle + \langle c, x\rangle$, $g = \delta_{\{0\}}$ and $\mathcal{A}(x) = Ax - b$. The problems are generated as follows. The vector $c \in \mathbb{R}^n$ and a diagonal matrix $D \in \mathbb{R}^{n\times n}$ are sampled from a normal distribution with unit variance, centered around zero and five respectively. Any negative elements of D are clipped to 0 to ensure positive semi-definiteness. Then, a random integer k is taken uniformly random from the interval $[\lceil \frac{n}{4} \rceil, \lfloor \frac{n}{2} \rfloor]$. Next, k diagonal elements of the matrix D, which are chosen uniformly random without replacement, are set to zero. We generate an orthonormal matrix V as the orthonormal matrix in the QR-decomposition of a random matrix, which is taken from a standard normal distribution. We define the positive semi-definite matrix $Q = VDV^\top$. The constraint matrix $A \in \mathbb{R}^{m\times n}$ is sampled from the standard normal distribution, and the vector $b \in \mathbb{R}^m$ is sampled uniformly random over the interval $[-1, 1]$. The upper and lower bounds are fixed as $u = -\ell = 0.8 \cdot 1_n$.

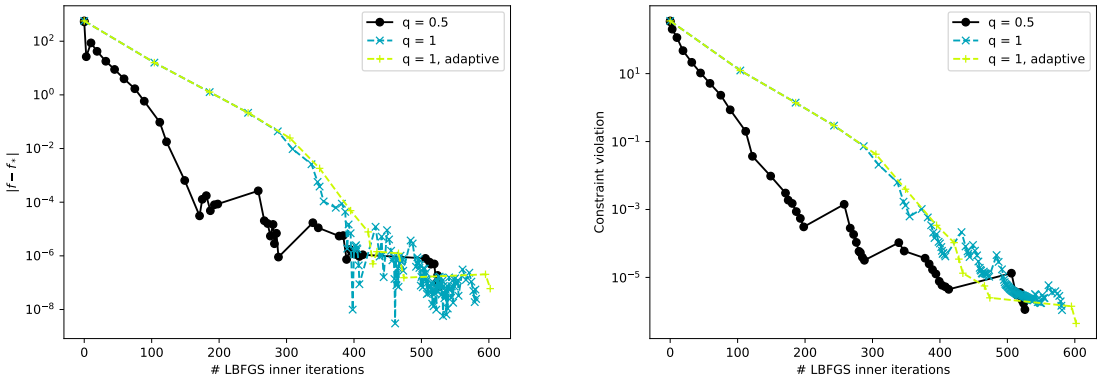
The equality constraint $Ax = b$ is enforced through the augmented Lagrangian, whereas the box-constraints are enforced in the inner problems using the L-BFGS-B algorithm [5, 29]. Tuned parameters λ and c are listed in Table 3. Table 4 compares power ALM with 2-norms raised to a power q in terms of the total number of L-BFGS-B iterations. A reference solution $f^\star$ is computed by QPALM [8] up to an absolute accuracy of (at least) 10^{-10}. The table reports both the median and the 95th percentile (P95) of the total number of L-BFGS-B iterations over $N = 100$ random realizations of the QP, and this for various problem sizes (m, n).

First, observe how classical ALM with an adaptive penalty parameter λ_k consistently performs better than its counterpart with a fixed tuned penalty parameter λ. However, for power ALM with $q \neq 1$, such adaptive schemes do not seem to improve the method and are therefore not included. Second, remark that power ALM outperforms classical ALM if the number of constraints m is relatively large compared to the number of variables n, and vice versa. It is also worth mentioning that this observation seemingly holds irrespective of the size of the problem.

Figure 1 visualizes the suboptimality $|f(x^{k+1}) - f^\star|$ and the constraint violation $\|Ax^{k+1} - b\|_2$ as a function of the performed number of L-BFGS-B iterations for a random realization of the QP with $(m, n) = (400, 800)$. Initially, power ALM ($q = 1/2$) decreases both the suboptimality and the constraint violation significantly faster than the classical variant ($q = 1$). The latter is relatively fast in driving the suboptimality to zero, but takes more iterations to attain the desired constraint violation. This explains why the classical ALM performs best only for the examples in Table 4 where, relative to the problem size, there are fewer constraints. On the other hand, when satsifying the constraint violation becomes harder, power ALM performs better.

Table 4: Median and 95th percentile (P95) of the total number of L-BFGS-B iterations required to solve the QP, over $N = 100$ random realizations.

	$q = 1/2$		$q = 1/3$		q=1 Fixed λ		q=1 Adaptive λ	
(m, n)	Med.	P95	Med.	P95	Med.	P95	Med.	P95
(100, 200)	**449**	560.2	595	811.8	551	698.3	488.5	652.1
(200, 400)	**469.5**	587.2	561	770.3	521.5	619.1	505.5	644
(300, 600)	**484.5**	612.3	641	935.4	535	632.3	504	654.2
(400, 800)	**489.5**	676.7	628.5	978.8	553.5	642.4	511	655.3
(500, 1000)	**536.5**	735.1	729.5	1,029.1	589	703	584	795.4
(600, 1200)	**554**	747	759	1,243.6	610	710.4	582.5	843.4
(100, 300)	392	529.2	509	843.4	403	477.1	**388**	471.3
(200, 600)	469	623.6	728.5	1,200.5	457.5	538.1	**443**	577.3
(300, 900)	513	679.5	848.5	1,383	514.5	595	**473**	648.1
(400, 1200)	550	746.4	973	1,489.2	560	635.1	**510.5**	692.9



(a) Suboptimality (b) Constraint violation

Figure 1: Suboptimality and constraint violation in terms of the performed number of L-BFGS-B iterations for solving the QP with $m = 400$, $n = 800$.

6 Conclusion

In this paper we have considered the inexact power proximal point method and the corresponding Augmented Lagrangian scheme under two different absolute inexactness conditions. The global analysis can be considered an extension of the analysis of [21] for a nonmonotone algorithm under absolute errors. In order to incorporate a different form of inexactness, we have also studied the algorithm as inexact anisotropic gradient descent. Under this framework, we have generalized for the proximal point method the results of [1] which were specialized to the Augmented Lagrangian setting. Moreover, in the Augmented Lagrangian setting we showed ways to enforce the proposed error conditions in practice, while also providing the corresponding global rates for the primal-dual gap and the measure of infeasibility.

References

[1] D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[2] D. P. Bertsekas. "Necessary and sufficient conditions for a penalty method to be exact". *Mathematical programming* 9.1 (1975), pp. 87–99.

[3] J. F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

[4] J. V. Burke, H. Tim, and Q. V. Nguyen. "A study of convex convex-composite functions via infimal convolution with applications". *Mathematics of Operations Research* 46.4 (2021), pp. 1324–1348.

[5] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. "A Limited Memory Algorithm for Bound Constrained Optimization". *SIAM Journal on Scientific Computing* 16.5 (Sept. 1995). Publisher: Society for Industrial and Applied Mathematics, pp. 1190–1208.

[6] N. Doikov and Y. E. Nesterov. "Inexact Tensor Methods with Dynamic Accuracies." In: *ICML*. 2020, pp. 2577–2586.

[7] B. Grimmer. "On optimal universal first-order methods for minimizing heterogeneous sums". *Optimization Letters* (2023), pp. 1–19.

[8] B. Hermans, A. Themelis, and P. Patrinos. *QPALM: A Proximal Augmented Lagrangian Method for Nonconvex Quadratic Programs*. Sept. 2, 2021. arXiv: 2010.02653[math].

[9] M. R. Hestenes. "Multiplier and gradient methods". *Journal of optimization theory and applications* 4.5 (1969), pp. 303–320.

[10] E. Laude and P. Patrinos. "Anisotropic proximal gradient". *arXiv preprint arXiv:2210.15531* (2022).

[11] E. Laude and P. Patrinos. *Anisotropic Proximal Point Algorithm*. Tech. rep. KU Leuven, 2023.

[12] E. Laude, A. Themelis, and P. Patrinos. "Conjugate dualities for relative smoothness and strong convexity under the light of generalized convexity". *arXiv preprint arXiv:2112.08886* (2021).

[13] Y.-F. Liu, X. Liu, and S. Ma. "On the nonergodic convergence rate of an inexact augmented Lagrangian framework for composite convex programming". *Mathematics of Operations Research* 44.2 (2019), pp. 632–650.

[14] F. J. Luque. "Asymptotic convergence analysis of the proximal point algorithm". *SIAM Journal on Control and Optimization* 22.2 (1984), pp. 277–293.

[15] J. Luque. "A nonlinear proximal point algorithm". In: *26th IEEE Conference on Decision and Control*. Vol. 26. IEEE. 1987, pp. 816–817.

[16] J. Luque. "Nonlinear proximal point algorithms". PhD thesis. Massachusetts Institute of Technology, Department of Civil Engineering, 1984.

[17] J. Luque. *The nonlinear poximal point algorithm and multiplier methods*. Tech. rep. Massachusetts Institute of Technology. Laboratory for Information and Decision Systems, 1986.

[18] J. Luque. *The nonlinear proximal point algorithm*. Tech. rep. Massachusetts Institute of Technology. Laboratory for Information and Decision Systems, 1986.

[19] M. Mohammadisiahroudi, R. Fakhimi, B. Augustino, and T. Terlaky. *Generating Linear, Semidefinite, and Second-order Cone Optimization Problems for Numerical Experiments*. Feb. 1, 2023. arXiv: 2302.00711[math].

[20] Y. Nesterov. "Universal gradient methods for convex optimization problems". *Mathematical Programming* 152.1-2 (2015), pp. 381–404.

[21] Y. Nesterov. "Inexact accelerated high-order proximal-point methods". *Mathematical Programming* (2021), pp. 1–26.

[22] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2003.

[23] J. Nocedal and S. J. Wright. *Numerical optimization*. 2nd ed. Springer series in operations research. OCLC: ocm68629100. New York: Springer, 2006. 664 pp.

[24] M. J. Powell. "A method for nonlinear constraints in minimization problems". *Optimization* (1969), pp. 283–298.

[25] R. T. Rockafellar. "Augmented Lagrangians and applications of the proximal point algorithm in convex programming". *Mathematics of operations research* 1.2 (1976), pp. 97–116.

[26] R. T. Rockafellar. "Augmented Lagrange multiplier functions and duality in nonconvex programming". *SIAM Journal on Optimization* 12.2 (1974), pp. 268–285.

[27] R. T. Rockafellar and R. J. Wets. *Variational Analysis*. New York: Springer, 1998.

[28] J. Yang, S. Zhang, N. Kiyavash, and N. He. "A catalyst framework for minimax optimization". *Advances in Neural Information Processing Systems* 33 (2020), pp. 5667–5678.

[29] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization". *ACM Transactions on Mathematical Software* 23.4 (Dec. 1, 1997), pp. 550–560.

A Missing proofs

A.1 Proof of Proposition 4.1

Proof. "4.1(i)": Since $y^+ \in \arg\min_{\eta\in\mathbb{R}^m} \lambda \star \phi(y-\eta) - L(x^+,\cdot)$ we have that

$$v = \nabla\phi(\lambda^{-1}(y-y^+)) \in \partial_y(-L)(x^+,y^+).$$

By concavity of $L(x,\cdot)$ we have that for all $y \in \mathbb{R}^m$

$$\langle v, y^+ - y\rangle \geq L(x^+,y) - L(x^+,y^+) \geq \varrho(y) - L(x^+,y^+),$$

where the second inequality follows by the definition of the dual function ϱ. Since $L(x^+,y^+) \leq \inf L(\cdot,y^+) + \varepsilon$ we can further bound $\langle v, y^+ - y\rangle \geq \varrho(y) - \varrho(y^+) - \varepsilon$. Rearranging this yields $(-\varrho)(y) \geq (-\varrho)(y^+) + \langle v, y - y^+\rangle - \varepsilon$, for any $y \in \mathbb{R}^m$ and hence $v \in \partial_\varepsilon(-\varrho)(y^+)$, i.e. $v \in \partial_\varepsilon\psi(y^+)$.

"4.1(ii)": Using the concavity of the augmented Lagrangian function $L_\lambda(x^{k+1},\cdot)$, we get

$$\begin{aligned} L_\lambda(x^+,y) + \langle \nabla_y L_\lambda(x^+,y), \eta - y\rangle &\geq L_\lambda(x^+,\eta) \geq \inf_{x\in\mathbb{R}^n} L_\lambda(x,\eta) \\ &= \inf_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m}\ L(x,y) - \lambda\star\phi(\eta-y) \\ &= \max_{y\in\mathbb{R}^m}\inf_{x\in\mathbb{R}^n}\ L(x,y) - \lambda\star\phi(\eta-y) \\ &= \max_{y\in\mathbb{R}^m} \varrho(y) - \lambda\star\phi(\eta-y) \\ &\geq \varrho(\operatorname{prox}_\psi^{\lambda\star\phi}(y)) - \lambda\star\phi(\eta - \operatorname{prox}_\psi^{\lambda\star\phi}(y)), \end{aligned} \tag{A.1}$$

where we have used the definitions of the Lagrangian dual function, the properties of inf and max and strong duality. The same reasoning also yields:

$$\begin{aligned} \inf_{x\in\mathbb{R}^n} L_\lambda(x,y) &= \inf_{x\in\mathbb{R}^n}\max_{\eta\in\mathbb{R}^m}\ L(x,\eta) - \lambda\star\phi(y-\eta) = \max_{\eta\in\mathbb{R}^m} \varrho(\eta) - \lambda\star\phi(y-\eta) \\ &= \varrho(\operatorname{prox}_\psi^{\lambda\star\phi}(y)) - \lambda\star\phi(y - \operatorname{prox}_\psi^{\lambda\star\phi}(y)). \end{aligned} \tag{A.2}$$

Now, from the optimality condition in (4.2), we have that $\lambda^{-p}\nabla\phi(y-y^+) = \nabla_y L_\lambda(x^+,y)$. Therefore, by combining (A.1) and (A.2), we obtain the following result:

$$\begin{aligned} &L_\lambda(x^+,y) - \inf_x L_\lambda(x,y) \\ &\quad\geq \lambda\star\phi(y - \operatorname{prox}_\psi^{\lambda\star\phi}(y)) - \lambda\star\phi(\eta - \operatorname{prox}_\psi^{\lambda\star\phi}(y)) + \lambda^{-p}\langle\nabla\phi(y-y^+), \eta - y\rangle, \end{aligned}$$

which holds for all $\eta \in \mathbb{R}^m$ so that we can replace η with its maximum. This is attained if $\lambda^{-p}\nabla\phi(\eta - \operatorname{prox}_\psi^{\lambda\star\phi}(y)) = \lambda^{-p}\nabla\phi(y-y^+)$, i.e., $\eta = y - y^+ + \operatorname{prox}_\psi^{\lambda\star\phi}(y)$. Substituting this result in the inequality above, we finally arrive at the claimed inequality

$$\begin{aligned} L_\lambda(x^+,y) - \inf_x L_\lambda(x,y) &\geq \lambda\star\phi(y - \operatorname{prox}_\psi^{\lambda\star\phi}(y)) - \lambda\star\phi(y-y^+) \\ &\quad - \lambda^{-p}\langle\nabla\phi(y-y^+), y - \operatorname{prox}_\psi^{\lambda\star\phi}(y) - (y-y^+)\rangle \\ &\geq \lambda^{-p}\left(\tfrac{1}{2}\right)^{p-1}\tfrac{1}{p+1}\|y^+ - \operatorname{prox}_\psi^{\lambda\star\phi}(y)\|^{p+1}, \end{aligned}$$

where the final inequality follows from Lemma 3.1.

"4.1(iii)": We have that $e \in \partial_x L(x^+, y^+)$ and $0 \in \partial_y(-L)(x^+, y^+) - \nabla\phi(\lambda^{-1}(y - y^+))$ if and only if

$$(x^+, y^+) \in \operatorname*{arg\,minimax}_{x\in\mathbb{R}^n, \eta\in\mathbb{R}^m} \; L(x, \eta) - \langle e, x\rangle - \lambda \star \phi(y - \eta),$$

and hence

$$e \in \partial_x L(x^+, y^+) \Leftrightarrow x^+ \in \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \; L_\lambda(x, y) - \langle e, x\rangle \Leftrightarrow e \in \partial_x L_\lambda(x^+, y). \tag{A.3}$$

By convexity of $L(\cdot, y^+)$ we have for all $x \in \mathbb{R}^m$ that $L(x, y^+) \geq L(x^+, y^+) + \langle e, x - x^+\rangle$, and hence $\inf L(\cdot, y^+) \geq L(x^+, y^+) - \|e\| \cdot \inf_{x\in\operatorname{dom} f} \|x - x^+\|$ implying that $L(x^+, y^+) \leq \inf L(\cdot, y^+) + \|e\|D \leq \inf L(\cdot, y^+) + \varepsilon$.

For the second part note that by convexity of $L_\lambda(\cdot, y)$, we have that $\inf L_\lambda(\cdot, y) - L_\lambda(x^+, y) \geq \langle e, x - x^+\rangle$, or by rearranging and using the Cauchy–Schwarz inequality $L_\lambda(x^+, y) - \inf L_\lambda(\cdot, y) \leq \|e\|D \leq \varepsilon$, which is the desired result. □

A.2 Proof of Theorem 4.4

Proof. "4.4(i)": The proof is based on the equivalence with the higher-order proximal-point method applied to the negative dual function $\psi = -\varrho$. Invoking Lemma 3.2 for points $y = y^k$ and $w = y^k$:

$$\psi(y^{k+1}) \leq \psi(y^k) - \lambda^{-p}(1 + \tfrac{1}{2^p})\phi(y^{k+1} - y^k) + \varepsilon_k.$$

By rearranging and using the fact that $\psi(y^{k+1}) \geq \inf \psi$ we obtain the following inequality:

$$\lambda^{-p}(1 + \tfrac{1}{2^p})\phi(y^{k+1} - y^k) \leq \psi(y^k) - \inf \psi + \varepsilon_k.$$

Substituting (2.12) into the inequality above and using the fact that $\phi(\nabla\phi^*(x)) = \|x\|_*^q$, we have that

$$\|Ax^{k+1} - b\|_*^{\frac{p+1}{p}} \leq (\lambda(1 + \tfrac{1}{2^p}))^{-1}(\psi(y^k) - \inf \psi + \varepsilon_k).$$

Using now Theorem 3.3, we obtain

$$\|Ax^{k+1} - b\|_* \leq (\sigma\eta_k)^{\frac{p}{p+1}},$$

for $\eta_k = \frac{\lambda^{-p} D_\beta^{p+1}(p+1)^p + c}{k^p}$ and $\sigma = (\lambda(1 + \frac{1}{2^p}))^{-1}$, which implies the claimed result.

"4.4(ii)": To begin with, using the fact that strong duality holds, choosing a saddle point $(x^\star, y^\star)$ and using Hölder's inequality, we have that

$$\begin{aligned} f(x^\star) &\leq L_\lambda(x^{k+1}, y^\star) \\ &\leq f(x^{k+1}) + \|y^\star\|\|Ax^{k+1} - b\|_* + \lambda^{-p}\phi^*(Ax^{k+1} - b), \end{aligned}$$

which using the previous result readily implies that

$$f(x^{k+1}) - f(x^\star) \geq -\|y^\star\|(\sigma\eta_k)^{\frac{p}{p+1}} - \tfrac{\lambda^{-p}}{q+1}\sigma\eta_k \tag{A.4}$$

Now, in order to prove the other side of the inequality for the absolute value, using the error condition of the update and the relation $\inf L(\cdot, y^{k+1}) = \varrho(y^{k+1})$ we have

$$f(x^{k+1}) + \langle y^{k+1}, Ax^{k+1} - b\rangle = L(x^{k+1}, y^{k+1}) = \inf L(\cdot, y^{k+1}) + \varepsilon_k = \varrho(y^{k+1}) + \varepsilon_k,$$

and by using the bound $\|y^{k+1} - y^\star\| \leq D_\beta$, the triangle inequality, as well as the fact that $\varrho(y^{k+1}) \leq f(x^\star)$, we can further write:

$$f(x^{k+1}) - f(x^\star) \leq (\|y^\star\| + D_\beta)(\sigma\eta_k)^{\frac{p}{p+1}} + \varepsilon_k \tag{A.5}$$

The claimed result now follows by combining (A.4) and (A.5). □