SHARPER BOUNDS FOR PROXIMAL GRADIENT ALGORITHMS WITH ERRORS*

ANIS HAMADOUCHE†, YUN WU†, ANDREW M. WALLACE†, AND JOÃO F. C. MOTA†

Abstract. We analyse the convergence of the proximal gradient algorithm for convex composite problems in the presence of gradient and proximal computational inaccuracies. We generalize the deterministic analysis to the quasi-Fejér case and quantify the uncertainty incurred from approximate computing and early termination errors. We propose new probabilistic tighter bounds that we use to verify a simulated Model Predictive Control (MPC) with sparse controls problem solved with early termination, reduced precision and proximal errors. We also show how the probabilistic bounds are more suitable than the deterministic ones for algorithm verification and more accurate for application performance guarantees. Under mild statistical assumptions, we also prove that some cumulative error terms follow a martingale property. And conforming to observations, e.g., in [25], we also show how the acceleration of the algorithm amplifies the gradient and proximal computational errors.

Key words. Convex Optimization, Proximal Gradient Descent, Approximate Algorithms

AMS subject classifications. 49M37, 65K05, 90C25

1. Introduction.

Many problems in science and engineering can be posed as *composite optimization problems*:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\;\; f(x) := g(x) + h(x)\,, \tag{1.1}$$

where the function $g : \mathbb{R}^n \to \mathbb{R}$ is real-valued and differentiable, and the function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is not necessarily differentiable and is possibly infinite-valued, enabling the inclusion of hard constraints in (1.1). Examples include various machine learning frameworks, e.g., logistic regression and support vector machines [11], sparse regression and inference [23, 15, 16], image processing [1], and discrete optimal control [17].

A popular class of algorithms to solve (1.1) is *proximal gradient methods* [4] which, in each iteration, take a gradient step using the function g and, subsequently, evaluate the proximal operator of the function h at the resulting point. Such algorithms have been widely studied under different contexts, and several guarantees have been established, both in the convex [5, 4, 6, 10, 22] and nonconvex [7, 21] cases. Stochastic versions of the proximal gradient algorithm have also been proposed and shown to converge in convex and nonconvex settings, e.g., [2, 29, 20, 24, 12, 30].

All of these results, however, assume that computations are performed with near-infinite precision, which is unrealistic when the computational platform has limitations in power, precision, or both. Examples include applications that are associated with sensing and control of autonomous platforms, often using FPGAs or other finite precision computational hardware. With these applications in mind, we analyze proximal gradient methods when both the gradient and the proximal operator are computed approximately at each iteration, and obtain tight performance bounds.

*Work supported by UK's EPSRC (EP/T026111/1, EP/S000631/1), and the MOD University Defence Research Collaboration.

†Anis Hamadouche, Yun Wu, Andrew M. Wallace, and João F. C. Mota are with the School of Engineering & Physical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK. (e-mail: {ah225,y.wu,a.m.wallace,j.mota}@hw.ac.uk).

While standard proximal gradient methods converge to a solution of (1.1) provided the stepsize s_k is small enough, approximate proximal gradient algorithms require, in addition, that the approximation errors ϵ_1^k and ϵ_2^k satisfy some additional convergence criteria, for example, that they converge to zero along the iterations.

Our goal is then *to characterize the convergence of the approximate proximal gradient* to a solution of (1.1). Differently from prior work, we assume not only deterministic errors, but also probabilistic ones, according to models suited to approximate computing.

1.1. Our approach. In the case of deterministic errors, we get inspiration from [4] to derive, using simple arguments, upper bounds on $f(x^k)$ throughout the iterations. The resulting bounds generalize other bounds [25] in the presence of Lipschitz uncertainty and early termination errors under mild assumptions. In the case of probabilistic errors, our arguments rely on concentration of measure results for martingale sequences and bypass the need to assume that ϵ_1^k and ϵ_2^k converge to zero. The latter yields tighter bounds, and we believe this line of reasoning is novel in the analysis of approximate proximal gradient algorithms.

1.2. Applications. In order to validate our convergence results, we use the proposed error bounds to analyse the convergence of proximal gradient when applied to solve the optimization problem stemming from each time step of Model Predictive Control (MPC) [13] with different levels of injected gradient and proximal computation errors.

1.3. Contributions. We summarize our contributions as follows:

- We establish convergence bounds for the proximal gradient algorithm with deterministic and probabilistic errors. Our deterministic bounds generalize prior bounds to the quasi-Fejér case where we consider approximate iterations and early termination errors and quantify second-order uncertainties. The probabilistic bounds tighten the latter under mild conditions.
- We conduct experiments on a discrete model predictive control problem to verify the sharpness of our bounds and compare them with the bounds in [25]. The models for the errors are inspired by approximate computing techniques suited for low-precision machines, such as reduced-precision accelerators on FPGA and battery-operated devices, in which algorithms are typically run approximately in order to save processing time and/or power.
- We propose new models for the proximal and gradient errors that satisfy martingale properties in accordance with experimental results.

1.4. Organization. We start by reviewing prior work in Section 2. We then describe our approximate computational model, state our assumptions, and present the main results in Section 3. The proofs of the main results are included in Section 4, and some auxiliary results are relegated to the appendix. Section 5 describes our experimental results.

2. Related Work. One year after the seminal work in [5], it was shown that the same nearly optimal rates can still be achieved when the computation of the gradients and proximal operators are approximate [25]. This variant is known as the *approximate* proximal gradient algorithm. The analysis in [25] requires the errors ϵ_1^k and ϵ_2^k to decrease with iterations k at rates $O(1/k^{\varsigma+1})$ for the basic proximal gradient, and $O(1/k^{\varsigma+2})$ for the accelerated proximal gradient, for any $\varsigma > 0$, in order to satisfy the summability assumptions of both error terms. The work in [25] established the following ergodic convergence bound in terms of function values of the averaged iterates for the basic approximate proximal gradient (3.7):

$$f\left(\frac{1}{k}\sum_{i=1}^{k} x^i\right) - f(x^\star) \leq \frac{L}{2k}\left[\left\|x^\star - x^0\right\|_2 + 2A_k + \sqrt{2B_k}\right]^2 \tag{2.1}$$
$$A_k = \sum_{i=1}^{k}\Big(\frac{\|\epsilon_1^i\|_2}{L} + \sqrt{\frac{2\epsilon_2^i}{L}}\Big), \quad B_k = \sum_{i=1}^{k}\frac{\epsilon_2^i}{L},$$

where $x^\star$ is an optimal solution of (1.1), L is the *Lipschitz* constant of the gradient, and x^0 is the initialization vector. The same work also analyzed the *approximate accelerated proximal gradient* and obtained the following convergence result in terms of the function values of the iterates,

$$f(x^i) - f(x^\star) \leq \frac{2L}{(k+1)^2}\left[\left\|x^\star - x^0\right\|_2 + 2\tilde{A}_k + \sqrt{2\tilde{B}_k}\right]^2 \tag{2.2}$$
$$\tilde{A}_k = \sum_{i=1}^{k} i\Big(\frac{\|\epsilon_1^i\|_2}{L} + \sqrt{\frac{2\epsilon_2^i}{L}}\Big), \quad \tilde{B}_k = \sum_{i=1}^{k}\frac{i^2\epsilon_2^i}{L}.$$

This is the most closely related work to ours; however, our work derives similar, yet sharper, convergence bounds. In addition, we derive probabilistic bounds that can be estimated before running the algorithm for given bounded proximal and gradient errors. Specifically, the constants can be computed from the machine representation and software solver tolerances (for the computation of the proximal operator).

The work in [3] extended the analysis of [25] to a more general momentum parameter selection $\alpha_k = ((k+a-1)/a)^d$, where $d \in [0,1]$ and $a > \max(1,(2d)^{\frac{1}{d}})$, which becomes FISTA [5] when $d = 1$. The works in [3, 26] also considered two different types of approximation in the proximal operator computation. For example, [3, Proposition 3.3] makes assumptions similar to ours, but establishes different bounds. The same paper also suggests slowing down the over-relaxations of FISTA to stabilize the algorithm and shows how to obtain a better trade-off between acceleration and error amplification by controlling the approximation errors. In contrast, we show that the basic approximate proximal gradient algorithm (3.7) converges to a constant predictable residual without any assumptions on the gradient error terms (see Theorem 3). We also show that errors in the accelerated proximal gradient method cause the algorithm to eventually diverge as $O(k)$ in the worst case scenario, but to converge sub-optimally, i.e., to a constant error term, using stronger assumptions on the proximal error and under a standard suitable choice of the momentum sequence $\{\beta_k\}$. We also quantify the uncertainties that result from using an inexact optimal reference point (motivated by early termination of practical solvers), inexact Fejér monotonicity (quasi-Fejér monotonicity) and an inexact version of Lipschitz continuity which is associated with approximate gradients with the relative error model 3.8.

3. Main Results.

Before stating our convergence guarantees for the approximate proximal gradient algorithm, we specify our assumptions and describe the class of algorithms that our analysis covers.

3.1. Setup and algorithms.

Recall that we aim to solve convex *composite optimization problems* with the format of (1.1), repeated here for convenience:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\;\; f(x) := g(x) + h(x)\,. \tag{3.1}$$

All of our results assume the following:

ASSUMPTION 1 (Assumptions on the problem).

- *The function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, and convex.*
- *The function $g : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and its gradient $\nabla g : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz-continuous with constant $L > 0$, that is,*

$$\big\|\nabla g(y) - \nabla g(x)\big\|_2 \le L\big\|y - x\big\|_2 \,, \tag{3.2}$$

for all x, $y \in \mathbb{R}^n$, where $\|\cdot\|_2$ stands for the standard Euclidean norm.
- *The set of optimal solutions of (3.1) is nonempty:*

$$X^\star := \big\{x \in \mathbb{R}^n \,:\, f(x) \le f(z), \quad \text{for all } z \in \mathbb{R}^n\big\} \ne \emptyset \,. \tag{3.3}$$

The above assumptions are standard in the analysis of proximal gradient algorithms and are actually required for convergence to an optimal solution from an arbitrary initialization [4, 6].

A consequence of (3.2) that we will often use in our results is that [19, Lem. 1.2.3]

$$g(y) \le g(x) + \nabla g(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2 \,, \tag{3.4}$$

for any x, $y \in \mathbb{R}^n$. Also, as h is closed, proper, and convex, the function $z \mapsto h(z) + (1/2)\|z - y\|_2^2$ is coercive, which implies that the approximate set-valued proximal operator of $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ at $y \in \mathbb{R}^n$, defined as

$$\mathrm{prox}_h^\epsilon(y) := \Big\{x \in \mathbb{R}^n \,:\, h(x) + \frac{1}{2}\|x - y\|_2^2 \le \epsilon + \inf_z \; h(z) + \frac{1}{2}\|z - y\|_2^2\Big\} \ne \emptyset \,,$$

is nonempty for all $\epsilon \ge 0$, and $y \in \mathbb{R}^n$. When $\epsilon = 0$, the proximal operator is computed exactly, and it is single-valued (a singleton) for closed, proper convex functions

$$\mathrm{prox}_h(y) := \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \; h(x) + \frac{1}{2}\|x - y\|_2^2. \tag{3.5}$$

When $\epsilon \ge 0$, this set may contain more than a single element, which results in several possible instances of the accelerated approximate proximal gradient,

$$\begin{aligned} y^k &= x^k + \beta_k(x^k - x^{k-1}), \\ x^{k+1} &\in \mathrm{prox}_{s_k h}^{\epsilon_2^k}\Big[y^k - s_k\big(\nabla g(y^k) + \epsilon_1^k\big)\Big] \,, \end{aligned} \tag{3.6}$$

whenever there exists a k for which $\epsilon_2^k > 0$. However, as we establish bounds on function values [i.e., $f(x^k)$], this ambiguity does not affect our results. By setting $\beta_k = 0$, (3.6) reduces to the basic approximate proximal gradient scheme, i.e.,

$$x^{k+1} \in \mathrm{prox}_{s_k h}^{\epsilon_2^k}\Big[x^k - s_k\big(\nabla g(x^k) + \epsilon_1^k\big)\Big]. \tag{3.7}$$

3.2. Error models and assumptions. In what follows we consider a relative error model for the gradient error ϵ_1.

ERROR MODEL. *Under this model, each evaluation of the gradient of g at a point x is subject to additive noise ϵ_1 whose magnitude is proportional to the magnitude of the gradient $|\nabla g(x)|$. Specifically, the gradient of g in (3.1) is approximated by*

$$\nabla g^{\epsilon_1}(x) = \nabla g(x) + \epsilon_1, \tag{3.8}$$

where

$$|\epsilon_1| \leq \delta |\nabla g(x)|. \tag{3.9}$$

δ is a positive scalar, and $|.|$ stands for the vector componentwise absolute value. This can be used, for example, to model errors in floating-point arithmetic [14].

The parameter δ is known as the machine precision.

For the above error model, our analysis assumes two different scenarios:

1. The sequences of errors $\{\epsilon_1^k\}_{k\geq 1}$ and $\{\epsilon_2^k\}_{k\geq 1}$ are deterministic, or
2. The sequences of errors $\{\epsilon_1^k\}_{k\geq 1}$ and $\{\epsilon_2^k\}_{k\geq 1}$ are random, in which case we use $\epsilon_{1\Omega}^k$ and $\epsilon_{2\Omega}^k$ to denote the respective random vectors/variables of errors at iteration k, where Ω denotes the sample space of a given probability measure.

In scenario 2, the sequences $\{x^k\}_{k\geq 1}$ and $\{y^k\}_{k\geq 1}$ become random as well. And we also use x_Ω^k and y_Ω^k to denote the respective random vectors at iteration k. We make the following assumptions in this case:

ASSUMPTION 2. *In scenario 2, we assume that each random vector $\epsilon_{1\Omega}^k$, for $k \geq 1$, satisfies*

$$\mathbb{E}\big[\epsilon_{1\Omega}^k \,\big|\, \epsilon_{1\Omega}^1, \ldots, \epsilon_{1\Omega}^{k-1}\big] = \mathbb{E}\big[\epsilon_{1\Omega}^k\big] = 0\,, \tag{3.10a}$$

$$\mathbb{P}\big(|\epsilon_{1\Omega}^k| \leq \delta |\nabla g(x_\Omega^k)|\big) = 1, \tag{3.10b}$$

$$\mathbb{E}\big[{\epsilon_{1\Omega}^k}^\top x_\Omega^k \,\big|\, \epsilon_{1\Omega}^1, \ldots, \epsilon_{1\Omega}^{k-1},\, x_{1\Omega}^1, \ldots, x_{1\Omega}^{k-1}\big] = \mathbb{E}\big[{\epsilon_{1\Omega}^k}^\top x_\Omega^k\big] = 0, \tag{3.10c}$$

$$\textit{or} \quad \mathbb{E}\big[\epsilon_{1\Omega}^k | x_\Omega^k\big] = \mathbb{E}\big[\epsilon_{1\Omega}^k\big],$$

where $\delta > 0$ is the machine precision.

ASSUMPTION 3. *Let $\{x^k\}$ denote the sequence produced by (3.6) or (3.7). We define the residual error vector at iteration k as*

$$r^k = x^k - \overline{x}^k, \tag{3.11}$$

where $\overline{x}^k$ stands for the proximal error-free iterate

$$\overline{x}^{k+1} := \operatorname{prox}_{sh}\Big(x^k - s\big(\nabla g(x^k) + \epsilon_1^k\big)\Big). \tag{3.12}$$

In scenario 2, we assume

$$\mathbb{E}\big[r_\Omega^k \,\big|\, r_\Omega^1, \ldots, r_\Omega^{k-1}\big] = \mathbb{E}\big[r_\Omega^k\big] = 0\,, \tag{3.13a}$$

$$\mathbb{E}\big[{r_\Omega^k}^\top x_\Omega^k \,\big|\, r_\Omega^1, \ldots, r_\Omega^{k-1},\, x_{1\Omega}^1, \ldots, x_{1\Omega}^{k-1}\big] = \mathbb{E}\big[{r_\Omega^k}^\top x_\Omega^k\big] = 0\,, \tag{3.13b}$$

Remark 3.1. Lemma 1, stated in the appendix, bounds the norm of the residual vector $\|r^k\|_2$ as a function of ϵ_2^k; therefore, bounding ϵ_2^k implies bounding $\|r^k\|_2$.

3.3. Approximate proximal gradient. In this section, we consider the approximate proximal gradient algorithm in (3.7), i.e., without acceleration. We start by considering deterministic error sequences $\{\epsilon_1^k\}_{k\geq 1}$ and $\{\epsilon_2^k\}_{k\geq 1}$, and then we consider the case in which these sequences are random, as in Assumption 2.

3.3.1. Deterministic errors. Our first result provides a bound on the ergodic convergence of the sequence of function values, and decouples the contribution of the errors in the computation of gradient, ϵ_1^k, and in the computation of the proximal operator, ϵ_2^k and r^k.

THEOREM 1. *Consider problem* (3.1) *and let Assumption* 1 *hold. Suppose we run the approximate proximal gradient in* (3.7) *with a fixed stepsize* $s_k := s$ *satisfying* $s \leq 1/(L+\delta)$, *for all* k, *and under the relative error model in* (3.8). *Let the following stopping criteria hold for* $k \geq k_0$: $\epsilon_2^k \leq c_2 \left\| x^{k+1} - x^k \right\|_2 \leq c_2 \rho$ *and* $\|\epsilon_1^k\|_2 \leq c_1 \left\| \nabla g(x^{k+1}) - \nabla g(x^k) \right\|_2$ *where* ρ, c_1, c_2 *and* k_0 *are constants. Then, for any* $x^\star \in X^\star$ *and* $k \geq k_0$, *the sequence generated by the approximate proximal gradient in* (3.7) *satisfies*

(3.14)
$$f\left(\frac{1}{k+1}\sum_{i=0}^{k} x^{i+1}\right) - f(x^\star) \leq \frac{1}{k+1}\left[\sum_{i=0}^{k}\epsilon_2^i + \sum_{i=0}^{k}\left(\left\|\epsilon_1^i\right\|_2 + \sqrt{\frac{2\epsilon_2^i}{s}}\right)\left\|x^\star - x^0\right\|_2 + \frac{1}{2s}\left\|x^\star - x^0\right\|_2^2\right] + \frac{1}{k+1}\sum_{i=0}^{k}\left(\left\|\epsilon_1^i\right\|_2 + \sqrt{\frac{2\epsilon_2^i}{s}}\right)\left(\sum_{j=1}^{i}E^j + iC_\rho\right),$$

where $E^j = \sqrt{\frac{2\epsilon_2^j}{s}} + s\left\|\epsilon_1^{j-1}\right\|_2$ *and* $C_\rho = \sqrt{2Lc_2\rho} + c_1\rho$.

Proof. See Section 4.1.

Theorem 1 improves over (2.1) by quantifying the uncertainties associated with the Lipschitz and Féjer properties in addition to the ones that stem from proximal and gradient errors.

Remark 3.2. For small perturbations and very small stopping criteria, i.e., $\rho \approx 0$[1], (3.14) can be approximated by

(3.15)
$$f\left(\frac{1}{k+1}\sum_{i=0}^{k} x^{i+1}\right) - f(x^\star) \lessapprox \frac{1}{k+1}\left[\sum_{i=0}^{k}\epsilon_2^i + \sum_{i=0}^{k}\left(\|\epsilon_1^i\|_2 + \sqrt{\frac{2\epsilon_2^i}{s}}\right)\|x^\star - x^0\|_2 + \frac{1}{2s}\|x^\star - x^0\|_2^2\right] - \frac{1}{2s}\sum_{i=0}^{k}\left\|r^{i+1}\right\|_2^2,$$

where we have dropped the second order error terms and kept the residual error vector explicitly, i.e., $-\frac{1}{2s}\sum_{i=0}^{k}\left\|r^{i+1}\right\|_2^2$, which improves the bound progressively with iterations.

This result implies that the $O(1/k)$ convergence rate is still guaranteed with weaker summability assumptions on $\{\epsilon_2^k\}_{k\geq 1}$ and $\{\left\|\epsilon_1^k\right\|_2\}_{k\geq 1}$. For instance, consider the case where both proximal and gradient errors decrease as $O(1/k)$ (i.e., non-summable). Then Theorem 1 yields an overall convergence rate of $O(\log k/k)$ which is less conservative than what would have been obtained from (2.1), i.e, $O(\log^2 k/k)$. Consequently, as a necessary condition for convergence, we only require the partial sums $\sum_{i=1}^{k}\epsilon_2^i$ and $\sum_{i=1}^{k}\|\epsilon_1^i\|_2$ to be in $o(k)$ as compared to the stronger condition $o(\sqrt{k})$ that is implied by (2.1). If we set both errors to zero for all $k \geq 1$, we recover the error-free optimal upper bound $\frac{1}{2sk}\left\|x^\star - x^0\right\|_2^2$ [4].

3.3.2. Random errors. Let us now consider the case in which ϵ_1^k, ϵ_2^k and therefore x^k, are random, and let $\epsilon_{1\Omega}^k$, $\epsilon_{2\Omega}^k$ and x_Ω^k be the corresponding random variables/vectors.

[1] $C_\rho = 0$ if the optimum $x^\star$ is reached.

THEOREM 2 (**Random errors**). *Consider problem (3.1) and let Assumption 1 hold. Assume that the gradient error* $\{\epsilon^k_{1\Omega}\}_{k\geq 1}$ *and residual proximal error* $\{r^k_{\Omega}\}_{k\geq 1}$ *sequences satisfy Assumptions 2, 3 and* $\mathbb{P}\big(\epsilon^k_{2\Omega} \leq \varepsilon_0\big) = 1$*, for all* $k > 0$*, and for some* $\varepsilon_0 \in \mathbb{R}$*. Let* $\{x^i_{\Omega}\}$ *denote a sequence generated by the approximate proximal gradient algorithm in (3.7) with constant stepsize* $s_k = s \leq 1/(L+\delta)$*, for all* k*. Assume that there is a positive scalar* $D_x > 0$ *such that* $\left\|x^k_{\Omega} - x^\star_{\Omega}\right\|_2^2 \leq D_x \left\|x^0_{\Omega} - x^\star_{\Omega}\right\|_2^2$ *holds with probability* p*, for all* k*. Then, for any* $\gamma > 0$*,*

(3.16)
$$f\left(\frac{1}{k}\sum_{i=1}^{k} x^i_{\Omega}\right) - f(x^\star) \leq \frac{1}{k}\sum_{i=1}^{k}\epsilon^i_{2\Omega} + \frac{\gamma}{\sqrt{k}}\left(\sqrt{n}M_{\nabla g}|\delta| + \sqrt{\frac{2\varepsilon_0}{s}}\right) D_x \left\|x^\star - x^0\right\|_2 + \frac{D_x^2}{2sk}\left\|x^\star - x^0\right\|_2^2,$$

with probability at least $p^k\big(1 - 2\exp(-\frac{\gamma^2}{2})\big)$*, where* $x^\star$ *is any solution of (3.1),* $M_{\nabla g} = \sup_{i\in\mathbb{N}_+}\left\{\left\|\nabla g(x^i)\right\|_\infty\right\}$*.*

Proof. See Section 4.2

For large scale problems,[2] we typically have $n \gg \frac{1}{s} \geq L$; therefore, we obtain the following approximated bound

(3.17)
$$f\left(\frac{1}{k}\sum_{i=1}^{k} x^i_{\Omega}\right) - f(x^\star) \lessapprox \frac{1}{k}\sum_{i=1}^{k}\epsilon^i_{2\Omega} + \gamma M_{\nabla g} D_x \sqrt{\frac{n}{k}}|\delta| \left\|x^\star - x^0\right\|_2 + \frac{D_x^2}{2sk}\left\|x^\star - x^0\right\|_2^2,$$

with approximately the same probability. In the absence of computational errors, (3.16) reduces to the deterministic noise-free convergence bound for $D_x = 1$, i.e., $\frac{1}{2sk}\left\|x^\star - x^0\right\|_2^2$.

The following result applies if we assume statistical stationarity[3] of proximal errors.

THEOREM 3 (**Random stationary errors**). *Consider problem (3.1), let Assumptions 1 hold and assume that the rounding error* $\{\epsilon^k_{1\Omega}\}_{k\geq 1}$ *and residual error* $\{r^k_{\Omega}\}_{k\geq 1}$ *sequences satisfy Assumptions 2, 3 and that the proximal computation error is upper bounded, i.e* $\mathbb{P}\big(\epsilon^k_{2\Omega} \leq \varepsilon_0\big) = 1$ *for all* $k \geq 1$ *and stationary with constant mean* $\mathbb{E}[\epsilon_{2\Omega}]$*. Let* $\{x^i_{\Omega}\}$ *denote a sequence generated by the approximate proximal gradient algorithm in (3.7) with constant stepsize* $s_k = s \leq 1/(L+\delta)$*, for all* k*. Assume that there is a positive scalar* $D_x > 0$ *such that* $\left\|x^k_{\Omega} - x^\star_{\Omega}\right\|_2^2 \leq D_x^2 \left\|x^0_{\Omega} - x^\star_{\Omega}\right\|_2^2$ *holds with probability* p*, for all* k*. Then, for any* $\gamma > 0$*,*

$$f\left(\frac{1}{k}\sum_{i=1}^{k} x^i_{\Omega}\right) - f(x^\star) \leq \mathbb{E}\big(\epsilon_{2\Omega}\big) + \frac{\gamma}{\sqrt{k}}\left(\frac{\varepsilon_0}{2} + \sqrt{n}M_{\nabla g}D_x|\delta| \left\|x^\star - x^0\right\|_2\right) + \frac{D_x^2}{2sk}\left\|x^\star - x^0\right\|_2^2, \tag{3.18}$$

[2] And for same levels of error magnitudes δ and ε_0.
[3] Whose ensemble mean and variance are time-invariant.

with probability at least $p^k\big(1-4\exp(-\frac{\gamma^2}{2})\big)$, *where* $x^\star$ *is any solution of* (3.1), $M_{\nabla g} = \sup_{i\in\mathbb{N}_+}\left\{ \left\|\nabla g(x^i)\right\|_\infty \right\}$.

Proof. See Section 4.3

Remark 3.3. D_x could be taken as large as to satisfy $\left\|x_\Omega^k - x_\Omega^\star\right\|_2^2 \leq D_x^2 \left\|x_\Omega^0 - x_\Omega^\star\right\|_2^2$ almost surely, i.e., with probability 1.

Once again, if both errors are forced to zero in (3.18) then the optimal convergence rate is obtained as in Theorem 1 and Theorem 2. (3.18) also implies that we obtain a worst case convergence rate of $O(1)$, i.e., convergence up to a predicted constant residual $\mathbb{E}[\epsilon_{2_\Omega}]$.

3.4. Accelerated Approximate PG.

3.4.1. Deterministic errors. We now analyze the effect of computational inaccuracy on the approximate accelerated PG. In what follows, we establish upper bounds on the convergence of the accelerated PG in the presence of deterministic errors in the computation of the gradient as well as in the proximal operation step.

THEOREM 4 (**Accelerated with deterministic errors**). *Consider problem* (3.1) *and let Assumption 1 hold. Suppose we run the approximate accelerated proximal gradient in* (3.6) *with a fixed stepsize* $s_k := s$ *satisfying* $s \leq 1/(L+\delta)$, *for all* k, *and under the relative error model in* (3.8). *Let the following stopping stopping criteria hold for* $k \geq k_0$: $\epsilon_2^k \leq c_2\left\|x^{k+1}-x^k\right\|_2 \leq c_2\rho$ *and* $\|\epsilon_1^k\|_2 \leq c_1\left\|\nabla g(x^{k+1}) - \nabla g(x^k)\right\|_2$ *where* ρ, c_1, c_2 *and* k_0 *are constants. Assume we have summable iterative displacements* $\left\|x^k - x^{k-1}\right\|_2$. *Let the momentum sequence* $\beta_k = (\alpha_{k-1}-1)/\alpha_k$ *be designed such that* α_k *satisfies the following:*

- $\alpha_k \geq 1 \quad \forall \quad k > 0$ *and* $\alpha_0 = 1$
- $\alpha_k^2 - \alpha_k = \alpha_{k-1}$
- $\{\alpha_k\}_{k=0}^\infty$ *is an increasing sequence and proportional to* k *(*$O(k)$*)*

Then, for any $x^\star \in X^\star$ *and* $k \geq k_0$, *the sequence generated by the approximate accelerated proximal gradient in* (3.6) *satisfies*

$$f(x^{k+1}) - f(x^\star) \leq \frac{1}{\alpha_k^2}\Bigg[\sum_{i=0}^k \alpha_i^2\epsilon_2^i + \sum_{i=0}^k \alpha_i\left\|x^0 - x^\star\right\|_2\left(\left\|\epsilon_1^i\right\|_2 + \sqrt{\frac{2\epsilon_2^i}{s}}\right) + \frac{1}{2s}\left\|x^0 - x^\star\right\|_2^2\Bigg] + \frac{1}{\alpha_k^2}\sum_{i=0}^k \alpha_i\left(\left\|\epsilon_1^i\right\|_2 + \sqrt{\frac{2\epsilon_2^i}{s}}\right)\sum_{j=1}^i \alpha_j(E^j + C_\rho), \tag{3.19}$$

where $x^\star$ *is any solution of* (3.1), $E^j = \sqrt{\frac{2\epsilon_2^j}{s}} + s\left\|\epsilon_1^{j-1}\right\|_2$ *and* $C_\rho = \sqrt{2Lc_2\rho} + c_1\rho$, *and* $C_\rho = \sqrt{2Lc_2\rho} + c_1\rho$.

Proof. See Section 4.4

Remark 3.4. Ignoring second order error terms (for small square summable perturbations and very small suboptimality stopping criterion, i.e., $\rho \approx 0$), (3.19) can be approximated by

$$f(x^{k+1}) - f(x^\star) \lessapprox \frac{1}{\alpha_k^2}\left[\sum_{i=0}^{k}\alpha_i^2\epsilon_2^i + \sum_{i=0}^{k}\alpha_i\left(\left\|\epsilon_1^i\right\|_2 + \sqrt{\frac{2\epsilon_2^i}{s}}\right)\left\|x^0 - x^\star\right\|_2 + \frac{1}{2s}\left\|x^0 - x^\star\right\|_2^2\right]. \tag{3.20}$$

Notice that if we trivially choose $\beta_k = 0$ we recover back the nonaccelerated basic scheme. In the noise-free case, (3.19) reduces to $\frac{1}{2s\alpha_k^2}\left\|x^\star - x^0\right\|_2^2$, which coincides with the convergence rate of the accelerated proximal gradient algorithm [4, Thm. 10.34], i.e., $O(1/k^2)$ if α_k is in the order of $O(k)$.

3.4.2. Random errors. The following result gives an estimate of the convergence rate when both errors are stochastic and bounded following a probabilistic analysis approach.

THEOREM 5 (**Accelerated with random errors**). *Consider problem (3.1) and let Assumption 1 hold. Suppose that the rounding error $\{\epsilon_{1_\Omega}^k\}_{k\geq 1}$ and residual error $\{r_\Omega^k\}_{k\geq 1}$ sequences satisfy Assumptions 2 and 3, respectively. Let the norm of the iterative difference $\left\|x_\Omega^k - x_\Omega^{k-1}\right\|_2$ be summable. Define a new sequence $u_\Omega^k := x^\star - x_\Omega^k + (1-\alpha_{k-1})(x_\Omega^k - x_\Omega^{k-1})$. Assume that there is a positive scalar $D_u > 0$ such that $\left\|u_\Omega^i\right\|_2^2 \leq D_u^2\left\|x^0 - x^\star\right\|_2^2$ holds with probability p. Let ε_0 be an upper bound on the proximal error, i.e., $\epsilon_{2_\Omega}^k \leq \varepsilon_0$ for all k. Then, for all $\gamma > 0$, the sequence generated by the approximate APG in (3.6) with constant stepsize $s_k := s \leq 1/(L+\delta)$, for all k, under error models (3.10) and (3.13), and with the following choices:*

- $\beta_k = \frac{\alpha_{k-1}-1}{\alpha_k}$
- $\alpha_k \geq 1 \quad \forall \quad k > 0$ *and* $\alpha_0 = 1$
- $\alpha_k^2 - \alpha_k = \alpha_{k-1}$
- $\{\alpha_k\}_{k=0}^\infty$ *increases as* $o(k)$

satisfies

$$f(x_\Omega^{k+1}) - f(x^\star) \leq \frac{1}{\alpha_k^2}\left[S_{\epsilon_{2_\Omega}} + S_{r_\Omega} + S_{\epsilon_{1_\Omega}} + \frac{1}{2s}\left\|x^\star - x^0\right\|_2^2\right], \tag{3.21}$$

where

$$S_{\epsilon_{2_\Omega}} = \varepsilon_0\sum_{i=0}^{k} i^2 + \frac{\gamma}{2}\sqrt{\sum_{i=1}^{k} i^4(\epsilon_{2_\Omega}^i)^2}, \tag{3.22}$$

$$S_{\epsilon_{1_\Omega}} = \gamma|\delta|M_{\nabla g}D_u^2\left\|x^0 - x^\star\right\|_2^2\sqrt{n\sum_{i=1}^{k} i^2}, \tag{3.23}$$

$$S_{r_\Omega} = \gamma D_u^2\left\|x^0 - x^\star\right\|_2^2\sqrt{\frac{2}{s}\sum_{i=1}^{k} i^2\epsilon_2^i} \tag{3.24}$$

with probability at least $p^k\left(1 - 4\exp(-\gamma^2/2)\right)$, where $x^\star$ is any solution of (3.1), $M_{\nabla g} = \sup_{i\in\mathbb{N}_+}\left\{\left\|\nabla g(x^i)\right\|_\infty\right\}$, and $\mathbb{E}[.]$ stands for the expectation operator.

Proof. See Section 4.5

Remark 3.5. D_u could be taken as large as to satisfy $\left\|u_{\Omega}^i\right\|_2^2 \leq D_u^2 \left\|x^0 - x^\star\right\|_2^2$ with probability 1.

The following corollary results from the substitution of partial sums by their corresponding closed forms and using the worst case upper bound ε_0 on $\epsilon_{2_\Omega}^i$ for all $i = 1, \ldots, k$.

COROLLARY 5.1 (**Accelerated with random errors**). *Consider problem* (3.1) *and let the assumptions of Theorem* 5 *hold. Define a new sequence* $u_{\Omega}^k := x^\star - x_{\Omega}^k + (1 - \alpha_{k-1})(x_{\Omega}^k - x_{\Omega}^{k-1})$. *Assume that there is a positive scalar* $D_u > 0$ *such that* $\left\|u_{\Omega}^i\right\|_2^2 \leq D_u^2 \left\|x^0 - x^\star\right\|_2^2$ *holds with probability* p. *Then we have, for all* k. *Let* ε_0 *be an upper bound on the proximal error, i.e.,* $\epsilon_{2_\Omega}^k \leq \varepsilon_0$ *for all* k. *Then we have, for all* k,

$$f(x_{\Omega}^{k+1}) - f(x^\star) \leq \frac{1}{\alpha_k^2}\left[\overline{S}_{\epsilon_{2_\Omega}} + \overline{S}_{r_\Omega} + \overline{S}_{\epsilon_{1_\Omega}} + \frac{1}{2s}\left\|x^\star - x^0\right\|_2^2\right], \tag{3.25}$$

where

$$\overline{S}_{\epsilon_{2_\Omega}} = \varepsilon_0 \frac{k(k+1)(2k+1)}{6} + \frac{\gamma}{2}\varepsilon_0\sqrt{\frac{k(k+1)(2k+1)(3k^2+3k-1)}{30}}, \tag{3.26}$$

$$\overline{S}_{\epsilon_{1_\Omega}} = \gamma|\delta| D_u M_{\nabla g}\left\|x^0 - x^\star\right\|_2\sqrt{\frac{nk(k+1)(2k+1)}{6}}, \tag{3.27}$$

$$\overline{S}_{r_\Omega} = \gamma D_u \left\|x^0 - x^\star\right\|_2\sqrt{\frac{2s\varepsilon_0 k(k+1)(2k+1)}{6}}. \tag{3.28}$$

with probability at least $1 - 4\exp(-\gamma^2/2)$, *where* $x^\star$ *is any solution of* (3.1), $M_{\nabla g} = \sup_{i \in \mathbb{N}_+}\left\{\left\|\nabla g(x^i)\right\|_\infty\right\}$.

Proof. Substituting

$$\sum_{i=1}^k i^2 = \frac{k(k+1)(2k+1)}{6}, \tag{3.29}$$

and substituting

$$\sum_{i=1}^k i^4 = \frac{k(k+1)(2k+1)(3k^2+3k-1)}{30}, \tag{3.30}$$

and using $\left\|u_{\Omega}^i\right\|_2 \leq D_u\left\|x^0 - x^\star\right\|_2$, $\left\|\epsilon_{1_\Omega}^i\right\|_2 \leq |\delta| M_{\nabla g}\sqrt{n}$ in Theorem 5 completes the proof. ∎

In the absence of errors, both probabilistic and deterministic analyses lead to the optimal convergence rate of $O(1/k^2)$ for the accelerated scheme (3.19)-(3.21). However, as stated previously in Theorem 5, under the influence of computational inaccuracies and due to error amplification, acceleration has a counter-effect in the Nesterov's sense [18] and the method becomes more sensitive to gradient and proximal errors whenever we want to speed up the algorithm.

Although computational errors are deterministic in nature [14], probabilistic results such as (3.21) give us practical convergence bounds when errors cannot be measured or are undetectable but with known upper bounds. If the ensemble mean $\mathbb{E}[\epsilon_{2_\Omega}^k]$

is constant for all $k \geq 1$ in (3.21), i.e., the error sequence $\{\epsilon^k_{2_\Omega}\}$ is stationary, then (3.21) becomes totally independent from the instantaneous running errors $\epsilon^k_{1_\Omega}$, $\epsilon^k_{2_\Omega}$ as well as from the running iterates x^k_Ω and would be only determined by the machine precision δ, the tolerance $\mathbb{E}[\epsilon_{2_\Omega}]$ and the given probability parameter γ. The factor α_k is designed to be proportional to the iteration counter $o(k)$.

Although boundedness of the gradient error is sufficient for the gradient error term $S_{\epsilon_{1_\Omega}}$ to asymptotically vanish, the algorithm fails to converge without the summability of the proximal error term $\{\alpha_k^2 \mathbb{E}(\epsilon^k_{2_\Omega})\}$.

4. Proofs.

4.1. Proof of Theorem 1. Recall the definition of ϵ-suboptimal proximal operator in (3.1):

$$\operatorname{prox}^\epsilon_u(y) := \left\{ x \in \mathbb{R}^n \,:\, u(x) + \frac{1}{2}\|x-y\|_2^2 \leq \epsilon + \inf_z\, u(z) + \frac{1}{2}\|z-y\|_2^2 \right\}. \tag{4.1}$$

Because this is a set, the point x^{k+1} in approximate proximal gradient (3.7) is not defined uniquely. To bound the effect of the error ϵ^k_2, we will therefore compute its difference with respect to the case where $\epsilon^k_2 = 0$, as measured by a function that we will define shortly. Recall that $\overline{x}^{k+1}$ is the noiseless computation of the proximal operator in (3.7) at x^k with constant stepsize s:

$$\overline{x}^{k+1} := \operatorname{prox}_{sh}\left[x^k - s\big(\nabla g(x^k) + \epsilon^k_1\big)\right], \tag{4.2}$$

$$= \operatorname{prox}_{sh}\left[x^k - s\nabla^{\epsilon^k_1} g(x^k)\right] \tag{4.3}$$

$$= \arg\min_x\; g(x^k) + \nabla^{\epsilon^k_1} g(x^k)^\top (x - x^k) + \frac{1}{2s}\left\|x - x^k\right\|_2^2 + h(x) \tag{4.4}$$

$$:= \arg\min_x\; G\big(x,\, x^k\big)\,. \tag{4.5}$$

From (4.2) to (4.3), we used $\nabla^{\epsilon^k_1} g(x^k) := \nabla g(x^k) + \epsilon^k_1$ as the inexact gradient of g at x^k. From (4.3) to (4.4), we developed the squared ℓ_2-norm term in the definition of the proximal operator [cf. (3.5)] and added $g(x^k)$ to the objective function. Finally, from (4.4) to (4.5), we defined

$$G\big(x,\, x^k\big) := g(x^k) + \nabla^{\epsilon^k_1} g(x^k)^\top (x - x^k) + \frac{1}{2s}\left\|x - x^k\right\|_2^2 + h(x)\,. \tag{4.6}$$

As h is convex [cf. Assumption 1], the quadratic term in (4.6) makes the function $G(\cdot,\, x^k)$ strongly convex with parameter $1/s$ [4].

Recall that $\overline{x}^{k+1}$ is the optimal solution of (4.5) and that x^{k+1} is the actual, noisy iterate in (3.7). Therefore, according to (3.7) and to the definition of the ϵ-suboptimal proximal operator in (4.1),

$$h\big(x^{k+1}\big) + \frac{1}{2s}\left\|x^{k+1} - x^k + s\nabla^{\epsilon^k_1} g(x^k)\right\|_2^2 \leq \epsilon^k_2 + h\big(\overline{x}^{k+1}\big) + \frac{1}{2s}\left\|\overline{x}^{k+1} - x^k + s\nabla^{\epsilon^k_1} g(x^k)\right\|_2^2 \tag{4.7}$$

$$\iff \quad h\big(x^{k+1}\big) + \frac{1}{2s}\left\|x^{k+1} - x^k\right\|_2^2 + \nabla^{\epsilon^k_1} g(x^k)^\top \big(x^{k+1} - x^k\big) \leq \epsilon^k_2 + h\big(\overline{x}^{k+1}\big) + \frac{1}{2s}\left\|\overline{x}^{k+1} - x^k\right\|_2^2 + \nabla^{\epsilon^k_1} g(x^k)^\top \big(\overline{x}^{k+1} - x^k\big) \tag{4.8}$$

$$\iff \quad G\big(x^{k+1},\, x^k\big) - G\big(\overline{x}^{k+1},\, x^k\big) \leq \epsilon^k_2\,. \tag{4.9}$$

From (4.7) to (4.8), we developed the squared-norm terms and cancelled the common term. From (4.8) to (4.9), we added the constant $g(x^k) - \frac{s}{2}\left\|\nabla g(x^k)\right\|_2^2$ to both sides and used the definition (4.6). Notice that (4.9) bounds the distance between x^{k+1} and $\overline{x}^{k+1}$ as measured by $G(\cdot,\, x^k)$.

Because $G(\cdot,\, x^k)$ is strongly convex, [4, Theorem. 5.25] establishes that

$$G\big(x,\, x^k\big) - G\big(\overline{x}^{k+1},\, x^k\big) \geq \frac{1}{2s}\left\|x - \overline{x}^{k+1}\right\|_2^2, \tag{4.10}$$

for any $x \in \mathbb{R}^n$. In particular, it holds for any optimal solution $x^\star$ of (3.1).

Thus, subtracting (4.10) with $x = x^\star$ from (4.9) yields

$$\begin{aligned}
&(4.11) && G\big(x^{k+1},\, x^k\big) - G\big(x^\star,\, x^k\big) \leq \epsilon_2^k - \frac{1}{2s}\left\|x^\star - \overline{x}^{k+1}\right\|_2^2 \\
&(4.12) \iff && g(x^k) + \nabla^{\epsilon_1^k} g(x^k)^\top \big(x^{k+1} - x^k\big) + \frac{1}{2s}\left\|x^{k+1} - x^k\right\|_2^2 + h\big(x^{k+1}\big) \\
& && - G\big(x^\star,\, x^k\big) \leq \epsilon_2^k - \frac{1}{2s}\left\|x^\star - \overline{x}^{k+1}\right\|_2^2 \\
&(4.13) \iff && g(x^k) + \nabla g(x^k)^\top \big(x^{k+1} - x^k\big) + {\epsilon_1^k}^\top \big(x^{k+1} - x^k\big) \\
& && + \frac{1}{2s}\left\|x^{k+1} - x^k\right\|_2^2 + h\big(x^{k+1}\big) - G\big(x^\star,\, x^k\big) \leq \epsilon_2^k - \frac{1}{2s}\left\|x^\star - \overline{x}^{k+1}\right\|_2^2.
\end{aligned}$$

From (4.11) to (4.12), we simply used the definition of $G(x,\, x^k)$ in (4.6) with $x = x^{k+1}$ and we also used $\nabla^{\epsilon_1^k} g(x^k) := \nabla g(x^k) + \epsilon_1^k$ in (4.13).

Applying (3.4) to (4.13) (with $s \leq 1/L$) and using $f := g + h$, we obtain

$$\begin{aligned}
&(4.14) && g\big(x^{k+1}\big) + h\big(x^{k+1}\big) - G\big(x^\star,\, x^k\big) \leq \epsilon_2^k - \frac{1}{2s}\left\|x^\star - \overline{x}^{k+1}\right\|_2^2 \\
& && + {\epsilon_1^k}^\top \big(x^k - x^{k+1}\big), \\
&(4.15) \iff && f\big(x^{k+1}\big) - G\big(x^\star,\, x^k\big) \leq \epsilon_2^k - \frac{1}{2s}\left\|x^\star - \overline{x}^{k+1}\right\|_2^2 + {\epsilon_1^k}^\top \big(x^k - x^{k+1}\big).
\end{aligned}$$

We now expand $G(x^\star,\, x^k)$ in (4.15) as follows

$$\begin{aligned}
f(x^{k+1}) - g(x^k) - \nabla^{\epsilon_1^k} g(x^k)^\top (x^\star - x^k) - \frac{1}{2s}\left\|x^\star - x^k\right\|_2^2 - h(x^\star) \\
\leq \epsilon_2^k - \frac{1}{2s}\left\|x^\star - \overline{x}^{k+1}\right\|_2^2 + {\epsilon_1^k}^\top \big(x^k - x^{k+1}\big).
\end{aligned} \tag{4.16}$$

Rearranging and subtracting $g(x^\star)$ from both sides yields

$$\begin{aligned}
f(x^{k+1}) - h(x^\star) - g(x^\star) \leq -g(x^\star) + \epsilon_2^k - \frac{1}{2s}\left\|x^\star - \overline{x}^{k+1}\right\|_2^2 + g(x^k) \\
+ \nabla^{\epsilon_1^k} g(x^k)^\top (x^\star - x^k) + \frac{1}{2s}\left\|x^\star - x^k\right\|_2^2 + {\epsilon_1^k}^\top \big(x^k - x^{k+1}\big).
\end{aligned} \tag{4.17}$$

Using the definitions $f := g + h$ and $\nabla^{\epsilon_1^k} g(x^k) = \nabla g(x^k) + \epsilon_1^k$ in (4.17), we obtain

$$\begin{aligned}
& f(x^{k+1}) - f(x^\star) \leq \epsilon_2^k - g(x^\star) + g(x^k) + \nabla g(x^k)^\top \big(x^\star - x^k\big) \\
& - \frac{1}{2s}\left\|x^\star - \overline{x}^{k+1}\right\|_2^2 + \frac{1}{2s}\left\|x^\star - x^k\right\|_2^2 + {\epsilon_1^k}^\top \big(x^\star - x^k\big) + {\epsilon_1^k}^\top \big(x^k - x^{k+1}\big) \\
& \leq \epsilon_2^k - \frac{1}{2s}\left\|x^\star - \overline{x}^{k+1}\right\|_2^2 + \frac{1}{2s}\left\|x^\star - x^k\right\|_2^2 + {\epsilon_1^k}^\top \big(x^\star - x^{k+1}\big),
\end{aligned} \tag{4.18}$$

where in the second inequality we used the fact that g is convex, i.e., $g(x^\star) \geq g(x^k) + \nabla g(x^k)^\top (x^\star - x^k)$. Summing both sides of (4.18) from 0 to k,

$$\begin{aligned}
\sum_{i=0}^{k} \left[f(x^{i+1}) - f(x^\star)\right] &\leq \sum_{i=0}^{k} \epsilon_2^i + \sum_{i=0}^{k} {\epsilon_1^i}^\top \left(x^\star - x^{i+1}\right) \\
&\quad + \frac{1}{2s} \sum_{i=0}^{k} \left[\left\|x^\star - x^i\right\|_2^2 - \left\|x^\star - \overline{x}^{i+1}\right\|_2^2\right], \\
&= \sum_{i=0}^{k} \epsilon_2^i + \sum_{i=0}^{k} {\epsilon_1^i}^\top \left(x^\star - x^{i+1}\right) + \frac{1}{2s} \sum_{i=0}^{k} \Big[\left\|x^\star - x^i\right\|_2^2 \\
&\quad - \left(\left\|x^\star - x^{i+1}\right\|_2^2 + \left\|x^{i+1} - \overline{x}^{i+1}\right\|_2^2 \right. \\
&\quad \left. + 2(x^{i+1} - \overline{x}^{i+1})^\top (x^\star - x^{i+1})\right)\Big], \\
&= \sum_{i=0}^{k} \epsilon_2^i + \sum_{i=0}^{k} {\epsilon_1^i}^\top \left(x^\star - x^{i+1}\right) + \frac{1}{2s} \sum_{i=0}^{k} \Big[\left\|x^\star - x^i\right\|_2^2 \\
&\quad - \left(\left\|x^\star - x^{i+1}\right\|_2^2 + \left\|r^{i+1}\right\|_2^2 + 2(r^{i+1})^\top (x^\star - x^{i+1})\right)\Big], \\
&= \sum_{i=0}^{k} \epsilon_2^i + \sum_{i=0}^{k} (\epsilon_1^i - \frac{1}{s} r^{i+1})^\top \left(x^\star - x^{i+1}\right) + \frac{1}{2s}\Big[\left\|x^\star - x^0\right\|_2^2 \\
&\quad - \left\|x^\star - x^{k+1}\right\|_2^2\Big] - \frac{1}{2s} \sum_{i=0}^{k} \left\|r^{i+1}\right\|_2^2,
\end{aligned} \tag{4.19}$$

where in the second-to-last equality we used the definition of r^i in (3.11), and in the last equality we noticed that the quadratic terms involving $x^\star$ formed a telescopic sequence. Rearranging and moving negative terms to the left hand side results in

$$\begin{aligned}
\sum_{i=0}^{k} \left[f(x^{i+1}) - f(x^\star)\right] + \frac{1}{2s} \sum_{i=0}^{k} \left\|r^{i+1}\right\|_2^2 + \frac{1}{2s} \left\|x^\star - x^{k+1}\right\|_2^2 \leq \sum_{i=0}^{k} \epsilon_2^i \\
+ \sum_{i=0}^{k} (\epsilon_1^i - \frac{1}{s} r^{i+1})^\top \left(x^\star - x^{i+1}\right) + \frac{1}{2s} \left\|x^\star - x^0\right\|_2^2.
\end{aligned} \tag{4.20}$$

Since f is a convex function, Jensen's inequality implies

$$f\left(\frac{1}{k+1} \sum_{i=0}^{k} x^{i+1}\right) - f(x^\star) \leq \frac{1}{k+1} \sum_{i=0}^{k} \left[f(x^{i+1}) - f(x^\star)\right],$$

which, applied to (4.20) and together with the fact that the last two terms of the left-hand side of (4.20) are nonnegative, yields

$$f\left(\frac{1}{k+1} \sum_{i=0}^{k} x^{i+1}\right) - f(x^\star) + \frac{1}{2(k+1)s} \sum_{i=0}^{k} \left\|r^{i+1}\right\|_2^2 + \frac{1}{2(k+1)s} \left\|x^\star - x^{k+1}\right\|_2^2 \leq$$

$$\frac{1}{k+1} \Big[\sum_{i=0}^{k} \epsilon_2^i + \sum_{i=0}^{k} (\epsilon_1^i - \frac{1}{s} r^{i+1})^\top (x^\star - x^{i+1}) + \frac{1}{2s} \left\|x^\star - x^0\right\|_2^2\Big]. \tag{4.21}$$

Using Lemma 1 to bound the norm of the residual error $r^k = x^k - \overline{x}^k$ resulting from the proximal error ϵ_2^k, Cauchy-Schwarz yields

$$(\epsilon_1^i - \frac{1}{s}r^{i+1})^\top (x^\star - x^{i+1}) \le \left(\left\|\epsilon_1^i\right\|_2 + \frac{1}{s}\left\|r^{i+1}\right\|_2 \right) \left\|x^\star - x^{i+1}\right\|_2 \le \left(\left\|\epsilon_1^i\right\|_2 + \sqrt{\frac{2\epsilon_2^i}{s}} \right) \left\|x^\star - x^{i+1}\right\|_2. \tag{4.22}$$

Using (4.22) in (4.21) yields

$$f\left(\frac{1}{k+1}\sum_{i=0}^{k} x^{i+1}\right) - f(x^\star) \le \frac{1}{k+1}\sum_{i=0}^{k}\epsilon_2^i + \frac{1}{k+1}\sum_{i=0}^{k}\left(\left\|\epsilon_1^i\right\|_2 + \sqrt{\frac{2\epsilon_2^i}{s}} \right) \left\|x^\star - x^{i+1}\right\|_2 + \frac{1}{2s(k+1)}\left\|x^\star - x^0\right\|_2^2 \tag{4.23}$$

Applying Quasi-Féjer (Theorem 6 in the appendix) recursively gives

$$f\left(\frac{1}{k+1}\sum_{i=0}^{k} x^{i+1}\right) - f(x^\star) \le \frac{1}{k+1}\sum_{i=0}^{k}\epsilon_2^i + \frac{1}{2s(k+1)}\left\|x^\star - x^0\right\|_2^2 + \frac{1}{k+1}\sum_{i=0}^{k}\left(\left\|\epsilon_1^i\right\|_2 + \sqrt{\frac{2\epsilon_2^i}{s}} \right) \left\|x^\star - x^0\right\|_2 + \frac{1}{k+1}\sum_{i=0}^{k}\left(\left\|\epsilon_1^i\right\|_2 + \sqrt{\frac{2\epsilon_2^i}{s}} \right)(\sum_{j=1}^{i} E^j + iC_\rho), \tag{4.24}$$

where $E^j = \left\|r^j\right\|_2 + s_{j-1}\left\|\epsilon_1^{j-1}\right\|_2$ and $C_\rho = 0$ if the optimum $x^\star$ is reached. This completes the proof of Theorem 1.

4.2. Proof of Theorem 2. This result is about the basic version of approximate PGD, but with random proximal computation error ϵ_{2_Ω}, component-wise bounded gradient error ϵ_{1_Ω} and bounded residuals $\left\|x_\Omega^k - x^\star\right\|_2$. As the algorithm generates a sequence of random vectors $\{x_\Omega^k\}$, the residual vector sequence $\{r_\Omega^k\}$ will also be a random.

Let T_k denote the second error term in the bound of (3.14) [Theorem 1], i.e.,

$$T_k = \begin{cases} 0 & , \ k = 0 \\ \sum_{i=1}^{k}(\epsilon_{1_\Omega}^{i-1} - \frac{1}{s}r_\Omega^i)^\top (x^\star - x_\Omega^i) & , \ k = 1, 2, \dots, \end{cases} \tag{4.25}$$

The first step is to show that $\{T_k\}$ is a martingale. Recall that a sequence of random variables $T_0, T_1, \dots$ is a martingale with respect to the sequence $X_0, X_1, \dots$ if, for all $k \ge 0$, the following conditions hold:

- T_k is a function of $X_0, X_1, \dots, X_k$;
- $\mathbb{E}[|T_k|] < \infty$;
- $\mathbb{E}[T_{k+1}|X_0, X_1, \dots, X_k] = T_k$.

A sequence of random variables $T_0, T_1, \dots$ is called a martingale when it is a martingale with respect to itself. That is, $\mathbb{E}[|T_k|] < \infty$, and $\mathbb{E}[T_{k+1}|T_0, T_1, \dots, T_k] = T_k$.

Let $\nu_\Omega^k = \epsilon_{1\Omega}^{k-1} - \frac{1}{s} r_\Omega^k$ and recall the definition of r_Ω^k in (3.11):

$$r^k = x^k - \overline{x}^k. \tag{4.26}$$

Rewriting (4.25) in terms of ν_Ω^k yields

$$T_k = T_{k-1} + {\nu_\Omega^k}^\top (x^\star - x_\Omega^k). \tag{4.27}$$

We now show that Assumptions 2 and 3 imply that $\{T_k\}_{k\geq 0}$ is a martingale. Specifically, (3.10a) and (3.13a), we have

$$\mathbb{E}\big[\nu_\Omega^k \big| \nu_\Omega^1 \dots \nu_\Omega^{k-1}\big] = \mathbb{E}[\nu_\Omega^k] = 0.$$

And from (3.10c) and (3.13b), we have

$$\mathbb{E}\big[{\nu_\Omega^k}^\top x_\Omega^k \big| \nu_\Omega^1 \dots \nu_\Omega^{k-1}, x_\Omega^1 \dots x_\Omega^{k-1}\big] = \mathbb{E}\big[{\nu_\Omega^k}^\top x_\Omega^k\big] = 0.$$

Taking the expected value of both sides of (4.27) conditioned on $\{T_i\}_{i=1}^{k-1}$ gives

$$\begin{aligned}
\mathbb{E}\big[T_k \big| T_1 \dots T_{k-1}\big] &= \mathbb{E}\big[T_{k-1} + {\nu_\Omega^k}^\top (x^\star - x_\Omega^k) \big| T_1 \dots T_{k-1}\big] \\
&= \mathbb{E}\big[T_{k-1} \big| T_1 \dots T_{k-1}\big] + \mathbb{E}\big[{\nu_\Omega^k}^\top (x^\star - x_\Omega^k) \big| T_1 \dots T_{k-1}\big] \\
&= T_{k-1} + \mathbb{E}\big[{\nu_\Omega^k}^\top (x^\star - x_\Omega^k) \big| T_1 \dots T_{k-1}\big] \\
&= T_{k-1} + \mathbb{E}\big[{\nu_\Omega^k}^\top x^\star \big| T_1 \dots T_{k-1}\big] - \mathbb{E}\big[{\nu_\Omega^k}^\top x_\Omega^k \big| T_1 \dots T_{k-1}\big] \\
&= T_{k-1} + \mathbb{E}\big[\nu_\Omega^k \big| T_1 \dots T_{k-1}\big]^\top x^\star - \mathbb{E}\big[{\nu_\Omega^k}^\top x_\Omega^k \big| T_1 \dots T_{k-1}\big] \\
&= T_{k-1} + \mathbb{E}\big[\nu_\Omega^k \big| \nu_\Omega^1 \dots \nu_\Omega^{k-1}, x_\Omega^1 \dots x_\Omega^{k-1}\big]^\top x^\star \\
&\quad - \mathbb{E}\big[{\nu_\Omega^k}^\top x_\Omega^k \big| \nu_\Omega^1 \dots \nu_\Omega^{k-1}, x_\Omega^1 \dots x_\Omega^{k-1}\big] && (4.28) \\
&= T_{k-1} + \mathbb{E}\left[\epsilon_{1\Omega}^{k-1} - \frac{1}{s} r_\Omega^k\right]^\top x^\star - \mathbb{E}\left[(\epsilon_{1\Omega}^{k-1} - \frac{1}{s} r_\Omega^k)^\top x_\Omega^k\right] && (4.29) \\
&= T_{k-1} + \mathbb{E}\left[\epsilon_{1\Omega}^{k-1} - \frac{1}{s} r_\Omega^k\right]^\top x^\star - \mathbb{E}\left[\mathbb{E}\left[\epsilon_{1\Omega}^{k-1} - \frac{1}{s} r_\Omega^k \big| x_\Omega^k\right]^\top x_\Omega^k\right] && (4.30) \\
&= T_{k-1} - \mathbb{E}\left[\epsilon_{1\Omega}^{k-1} - \frac{1}{s} r_\Omega^k\right]^\top x_\Omega^k && (4.31) \\
&= T_{k-1}. && (4.32)
\end{aligned}$$

From (4.28) to (4.29), we used the error mean independence assumption, i.e., $E\big[\nu_\Omega^k \big| \nu_\Omega^1 \dots \nu_\Omega^{k-1}\big] = E\big[\nu_\Omega^k\big]$ as well as the data mean independence assumption (or the less restrictive statistical orthogonality in high dimensional problems), i.e., $E\big[{\nu_\Omega^k}^\top x_\Omega^k \big| \nu_\Omega^1 \dots \nu_\Omega^{k-1}, x_\Omega^1 \dots x_\Omega^{k-1}\big] = E\big[{\nu^k}^\top x_\Omega^k\big]$. From (4.31) to (4.32), we used the zero mean error assumption, i.e., $E\big[\nu_\Omega^k\big] = 0$. Therefore, $T_1, T_2, \dots, T_k$ is a martingale.

In what follows, we establish upper bounds on the absolute value of the martingale $\{T_k\}$. To do that, we use the Azuma-Hoeffding inequality in [27, p. 36], noticing that $\big|T_k - T_{k-1}\big| = \big|{\nu_\Omega^k}^\top (x^\star - x_\Omega^k)\big| \leq \Big(\sqrt{n}\delta M_{\nabla g} + \sqrt{2\epsilon_2^k/s}\Big) \big\|x_\Omega^\star - x_\Omega^k\big\|_2$, where we have used Cauchy-Schwarz, etc. Corollary [27, Corollary 2.20] then yields

$$\Pr\left(|T_k - T_0| \geq \gamma \sqrt{\sum_{i=1}^k \big(\sqrt{n} M_{\nabla g} |\delta| + \sqrt{\frac{2\epsilon_2^i}{s}}\big)^2 \big\|x_\Omega^\star - x_\Omega^i\big\|_2^2}\right) \leq 2\exp(-\frac{\gamma^2}{2}). \tag{4.33}$$

Since $\epsilon_2^k \leq \varepsilon_0$, then the following also holds

$$\Pr\Big(|T_k - T_0| \geq \gamma\big(\sqrt{n}M_{\nabla g}|\delta| + \sqrt{\frac{2\varepsilon_0}{s}}\big)\sqrt{\sum_{i=1}^{k}\left\|x_\Omega^\star - x_\Omega^i\right\|_2^2}\Big) \leq 2\exp(-\frac{\gamma^2}{2}). \tag{4.34}$$

And since $T_0 = 0$ we obtain

$$\Pr\Big(|T_k| \geq \gamma\big(\sqrt{n}M_{\nabla g}|\delta| + \sqrt{\frac{2\varepsilon_0}{s}}\big)\sqrt{\sum_{i=1}^{k}\left\|x_\Omega^\star - x_\Omega^i\right\|_2^2}\Big) \leq 2\exp(-\frac{\gamma^2}{2}). \tag{4.35}$$

Or, equivalently, that

$$|T_k| \leq \gamma\big(\sqrt{n}M_{\nabla g}|\delta| + \sqrt{\frac{2\varepsilon_0}{s}}\big)\sqrt{\sum_{i=1}^{k}\left\|x_\Omega^\star - x_\Omega^i\right\|_2^2} \tag{4.36}$$

holds for all $k \geq 1$ with probability at least $1 - 2\exp(-\frac{\gamma^2}{2})$. Expanding T_k we obtain

$$\left|\sum_{i=1}^{k}(\epsilon_{1\Omega}^{i-1} - \frac{1}{s}r_\Omega^i)^\top(x_\Omega^\star - x_\Omega^i)\right| \leq \gamma\left(\sqrt{n}M_{\nabla g}|\delta| + \sqrt{\frac{2\varepsilon_0}{s}}\right)\sqrt{\sum_{i=1}^{k}\left\|x_\Omega^\star - x_\Omega^i\right\|_2^2}. \tag{4.37}$$

By assumption, we have that $\left\|x_\Omega^\star - x_\Omega^i\right\|_2^2 \leq D_x\left\|x_\Omega^\star - x_\Omega^0\right\|_2^2$ holds with probability p, for each i. Then,

$$\left|\sum_{i=1}^{k}(\epsilon_{1\Omega}^{i-1} - \frac{1}{s}r_\Omega^i)^\top(x_\Omega^\star - x_\Omega^i)\right| \leq \gamma\left(M_{\nabla g}\sqrt{nk}|\delta| + \sqrt{\frac{2k\varepsilon_0}{s}}\right)D_x\left\|x_\Omega^\star - x_\Omega^0\right\|_2 \tag{4.38}$$

holds with probablity $p^k\big(1 - 2\exp(-\frac{\gamma^2}{2})\big)$. Substituting (4.38) into (3.14) completes the proof of Theorem 2.

4.3. Proof of Theorem 3. Here ϵ_{2_Ω} is bounded almost surely and has stationary mean. Specifically, we have $0 \leq \epsilon_{2_\Omega}^k \leq \varepsilon_0$, with probability 1. By Hoeffding's inequality ([27, Proposition 2.5]), we can write,

$$\Pr\left(|\sum_{i=1}^{k}\epsilon_{2_\Omega}^i - \mathbb{E}\left(\sum_{i=1}^{k}\epsilon_{2_\Omega}^i\right)| \geq t\right) \leq 2\exp\left(\frac{-2t^2}{k\varepsilon_0^2}\right), \quad \text{for all} \quad t > 0. \tag{4.39}$$

Defining the constant mean $\mathbb{E}[\epsilon_{2_\Omega}^k] = \mathbb{E}[\epsilon_{2_\Omega}]$ and substituting in (4.39) yields

$$\Pr\left(|\sum_{i=1}^{k}\epsilon_{2_\Omega}^i - k\mathbb{E}[\epsilon_{2_\Omega}]| \geq t\right) \leq 2\exp\left(\frac{-2t^2}{k\varepsilon_0^2}\right), \quad \text{for all} \quad t > 0. \tag{4.40}$$

By choosing $t = \frac{\gamma\sqrt{k}\varepsilon_0}{2}$, for some $\gamma > 0$, we obtain

$$\Pr\left(|\sum_{i=1}^{k}\epsilon_{2_\Omega}^i - k\mathbb{E}[\epsilon_{2_\Omega}]| \geq \frac{\gamma\sqrt{k}\varepsilon_0}{2}\right) \leq 2\exp\left(\frac{-\gamma^2}{2}\right) \quad \text{for all} \quad \gamma > 0. \tag{4.41}$$

Equivalently,

$$\sum_{i=1}^{k} \epsilon_{2\Omega}^{i} \leq k\mathbb{E}[\epsilon_{2\Omega}] + \frac{\gamma\sqrt{k}\varepsilon_0}{2} \tag{4.42}$$

holds with probability at least $1 - 2\exp(-\frac{\gamma^2}{2})$. Using the last inequality (4.42) in (3.16) and applying the probability union bound completes the proof of Theorem 3.

4.4. Proof of Theorem 4. Following the same line of proof of Section 4.1 but with $y_k = (1+\beta_k)x^k - \beta_k x^{k-1}$, where $\{\beta_k\} \in [0,1]$ is the momentum sequence, and using the approximate accelerated PG iteration scheme 3.6, we obtain

$$\begin{aligned} f(x^{k+1}) - f(x) \leq \epsilon_2^k + {\epsilon_1^k}^\top(x - x^{k+1}) - \frac{1}{2s}\left\|x - x^{k+1}\right\|_2^2 \\ - \frac{1}{2s}(r^{k+1})^\top(x - x^{k+1}) + \frac{1}{2s}\left\|x - y^k\right\|_2^2. \end{aligned} \tag{4.43}$$

Let us now substitute y^k and x by,

$$y^k = x^k + \beta_k(x^k - x^{k-1}) \tag{4.44}$$

$$x = \alpha_k^{-1}x^\star + (1 - \alpha_k^{-1})x^k, \tag{4.45}$$

where (4.44) follows from the definition of the acceleration scheme (3.6), and (4.45) is a choice that we make to simplify the analysis.[4] $\{\alpha_k\}_{k\geq 1}$ is a given parameter sequence that satisfies $\alpha_0 = 1$, $\alpha_k \geq 1$ and $\beta_k = \frac{\alpha_{k-1}-1}{\alpha_k}$. (4.43) can now be expanded as

$$\begin{aligned} f(x^{k+1}) - f(\alpha_k^{-1}x^\star + (1-\alpha_k^{-1})x^k) \leq \epsilon_2^k + {\epsilon_1^k}^\top(\alpha_k^{-1}x^\star + (1-\alpha_k^{-1})x^k - x^{k+1}) \\ - \frac{1}{2s}\left\|\alpha_k^{-1}x^\star + (1-\alpha_k^{-1})x^k - x^{k+1}\right\|_2^2 \\ + \frac{1}{2s}\left\|\alpha_k^{-1}x^\star + (1-\alpha_k^{-1})x^k - y^k\right\|_2^2 \\ - \frac{1}{2s}(r^{k+1})^\top(\alpha_k^{-1}x^\star + (1-\alpha_k^{-1})x^k - x^{k+1}). \end{aligned} \tag{4.46}$$

Since $\alpha_k^{-1} \in]0,1]$, and from the convexity of f, we have

$$\begin{aligned} f(x^{k+1}) - f(\alpha_k^{-1}x^\star + (1-\alpha_k^{-1})x^k) &\geq f(x^{k+1}) + (1-\alpha_k^{-1})f(x^\star) \\ &\quad - (1-\alpha_k^{-1})f(x^k) - f(x^\star) \\ &= f(x^{k+1}) - f(x^\star) - (1-\alpha_k^{-1})(f(x^k) - f(x^\star)). \end{aligned} \tag{4.47}$$

Let us now define the new sequences $\{v^k\}$ and $\{u^k\}$ by

$$u^k := x^\star + (\alpha_k - 1)x^k - \alpha_k y^k = x^\star - (x^k + (\alpha_{k-1} - 1)(x^k - x^{k-1})) \tag{4.48}$$

$$v^k = f(x^k) - f(x^\star). \tag{4.49}$$

[4]Note that $y^k \to x^k$ as $x^k \to x^\star$.

From these we can obtain

$$u^{k+1} := x^\star + (\alpha_k - 1)x^k - \alpha_k x^{k+1} = x^\star - (x^{k+1} + (\alpha_k - 1)(x^{k+1} - x^k)), \tag{4.50}$$

by using $\beta_k = (\alpha_{k-1} - 1)/\alpha_k$ and $y^k = (1+\beta_k)x^k - \beta_k x^{k-1}$.

Rewriting (4.46) in terms of the newly defined sequences, $\{u^k\}$ and $\{v^k\}$, and using (4.47) with $c_k := 1 - \alpha^{-1}$, as well as (4.48) and (4.50) we obtain

$$\begin{aligned} v^{k+1} - c_k v^k \leq \epsilon_2^k + \frac{1}{\alpha_k}{\epsilon_1^k}^\top u^{k+1} - \frac{1}{2s\alpha_k^2}\left\|u^{k+1}\right\|_2^2 + \frac{1}{2s\alpha_k^2}\left\|u^k\right\|_2^2 \\ - \frac{1}{2s}\left\|r^{k+1}\right\|_2^2 - \frac{1}{2s\alpha_k}(r^{k+1})^\top u^{k+1}. \end{aligned} \tag{4.51}$$

Rearranging (4.51) we obtain

$$\begin{aligned} v^{k+1} + \frac{1}{2s}\left\|r^{k+1}\right\|_2^2 + \frac{1}{2s\alpha_k^2}\left\|u^{k+1}\right\|_2^2 \leq \epsilon_2^k + \frac{1}{\alpha_k}{\epsilon_1^k}^\top u^{k+1} + c_k v^k \\ + \frac{1}{2s\alpha_k^2}\left\|u^k\right\|_2^2 - \frac{1}{2s\alpha_k}(r^{k+1})^\top u^{k+1}. \end{aligned} \tag{4.52}$$

Multiplying both sides by α_k^2,

$$\begin{aligned} \alpha_k^2 v^{k+1} + \frac{\alpha_k^2}{2s}\left\|r^{k+1}\right\|_2^2 + \frac{1}{2s}\left\|u^{k+1}\right\|_2^2 \leq \alpha_k^2\epsilon_2^k + \alpha_k{\epsilon_1^k}^\top u^{k+1} + \alpha_k^2 c_k v^k \\ + \frac{1}{2s}\left\|u^k\right\|_2^2 - \frac{\alpha_k}{2s}(r^{k+1})^\top u^{k+1}. \end{aligned} \tag{4.53}$$

Applying (4.53) recursively, and substituting $\alpha_k^2 c_k = \alpha_k^2 - \alpha_k = \alpha_{k-1}$ yields

$$\begin{aligned} \alpha_k^2 v^{k+1} + \frac{\alpha_k^2}{2s}\left\|r^{k+1}\right\|_2^2 + \frac{1}{2s}\left\|u^{k+1}\right\|_2^2 \leq \alpha_k^2\epsilon_2^k + \alpha_k{\epsilon_1^k}^\top u^{k+1} + \alpha_{k-1} v^k \\ + \frac{1}{2s}\left\|u^k\right\|_2^2 - \frac{\alpha_k}{2s}(r^{k+1})^\top u^{k+1}, \end{aligned} \tag{4.54}$$

$$\ldots,$$

$$\begin{aligned} \alpha_1^2 v^2 + \frac{\alpha_1^2}{2s}\left\|r^2\right\|_2^2 + \frac{1}{2s}\left\|u^2\right\|_2^2 \leq \alpha_1^2\epsilon_2^2 + \alpha_1{\epsilon_1^2}^\top u^2 + \alpha_0 v^1 \\ + \frac{1}{2s}\left\|u^1\right\|_2^2 - \frac{\alpha_1}{2s}(r^2)^\top u^2. \end{aligned} \tag{4.55}$$

Adding both sides of all inequalities,

$$\begin{aligned} \alpha_k^2 v^{k+1} + \sum_{i=0}^k \frac{\alpha_i^2}{2s}\left\|r^{i+1}\right\|_2^2 + \frac{1}{2s}\left\|u^{k+1}\right\|_2^2 + \sum_{i=0}^k(\alpha_{i-1}^2 - \alpha_{i-1})v^i \\ \leq \sum_{i=0}^k \alpha_i^2\epsilon_2^i + \frac{1}{2s}\left\|u^1\right\|_2^2 + \sum_{i=0}^k \alpha_i{\epsilon_1^i}^\top u^{i+1} + \alpha_0 v^1 - \sum_{i=0}^k \frac{\alpha_i}{2s}(r^{i+1})^\top u^{i+1}. \end{aligned} \tag{4.56}$$

Substituting $\alpha_{i-1}^2 - \alpha_{i-1} = \alpha_{i-2}^2$ and $\alpha_0 = 1$ gives,

$$\begin{aligned} \alpha_k^2 v^{k+1} + \sum_{i=0}^k \frac{\alpha_i^2}{2s}\left\|r^{i+1}\right\|_2^2 + \frac{1}{2s}\left\|u^{k+1}\right\|_2^2 + \sum_{i=0}^k \alpha_{i-2} v^i \\ \leq \sum_{i=0}^k \alpha_i^2\epsilon_2^i + \sum_{i=0}^k \alpha_i{\epsilon_1^i}^\top u^{i+1} + v^1 + \frac{1}{2s}\left\|u^1\right\|_2^2 - \sum_{i=0}^k \frac{\alpha_i}{2s}(r^{i+1})^\top u^{i+1}. \end{aligned}$$

For a positive sequence $\{\alpha_k\}_{k\geq 0}$ and because $x^\star$ is a (global) minimizer, $\sum \alpha_{i-2} v^i \geq 0$ is always satisfied; hence the following holds

$$\begin{aligned}\alpha_k^2 v^{k+1} &\leq \alpha_k^2 v^{k+1} + \sum_{i=0}^{k} \frac{\alpha_i^2}{2s} \left\|r^{i+1}\right\|_2^2 + \frac{1}{2s}\left\|u^{k+1}\right\|_2^2 + \sum_{i=0}^{k} \alpha_{i-2} v^i \\ &\leq \sum_{i=0}^{k} \alpha_i^2 \epsilon_2^i + \sum_{i=0}^{k} \alpha_i \left(\epsilon_1^i - \frac{1}{s} r^{i+1}\right)^\top u^{i+1} + v^1 + \frac{1}{2s}\left\|u^1\right\|_2^2 .\end{aligned} \tag{4.57}$$

From (4.43) with $k = 0$ and $x = x^\star$, we have

$$\begin{aligned}v^1 = f(x^1) - f(x^\star) \leq \epsilon_2^0 + \left(\epsilon_1^0 - \frac{1}{2s} r^1\right)^\top (x^\star - x^1) - \frac{1}{2s}\left\|x^\star - x^1\right\|_2^2 \\ + \frac{1}{2s}\left\|x^\star - x^0\right\|_2^2 ,\end{aligned} \tag{4.58}$$

since $y^0 = x^0$. From the definition of $\{u^k\}$ in (4.50) we have

$$\begin{aligned}\frac{1}{2s}\left\|u^1\right\|_2^2 &= \frac{1}{2s}\left\|x^\star + (\alpha_0 - 1)x^0 - \alpha_0 x^1\right\|_2^2 , \\ &= \frac{1}{2s}\left\|x^\star - x^1\right\|_2^2 ,\end{aligned} \tag{4.59}$$

where we have used the initialization $\alpha_0 = 1$. Substituting for v^{k+1} and combining (4.58) and (4.59) with (4.57) yields

$$\alpha_k^2 (f(x^{k+1}) - f(x^\star)) \leq \sum_{i=0}^{k} \alpha_i^2 \epsilon_2^i + \sum_{i=0}^{k} \alpha_i \left(\epsilon_1^i - \frac{1}{s} r^{i+1}\right)^\top u^{i+1} + \frac{1}{2s}\left\|x^\star - x^0\right\|_2^2 . \tag{4.60}$$

Dividing both sides by α_k^2 and applying Cauchy-Schwarz inequality yields

$$\begin{aligned}f(x^{k+1}) - f(x^\star) \leq \frac{1}{\alpha_k^2}\Bigg[\sum_{i=0}^{k} \alpha_i^2 \epsilon_2^i + \Big[\sum_{i=0}^{k} \alpha_i \Big(\left\|\epsilon_1^i\right\|_2 + \frac{1}{s}\left\|r^{i+1}\right\|_2\Big)\Big]\left\|u^{i+1}\right\|_2 \\ + \frac{1}{s}\left\|x^\star - x^0\right\|_2^2\Bigg].\end{aligned} \tag{4.61}$$

We have by definition 4.48 and 4.50

$$u^k = x^\star + (\alpha_k - 1)x^k - \alpha_k y^k = x^\star - (x^k + (\alpha_{k-1} - 1)(x^k - x^{k-1})), \tag{4.62}$$

$$u^{k+1} = x^\star + (\alpha_k - 1)x^k - \alpha_k x^{k+1} = x^\star - (x^{k+1} + (\alpha_k - 1)(x^{k+1} - x^k)). \tag{4.63}$$

By triangle inequality of the vector norm, we have

$$\begin{aligned}\left\|u^k\right\|_2 &\leq \left\|(\alpha_k - 1)(x^k - x^\star)\right\|_2 + \alpha_k \left\|y^k - x^\star\right\|_2 , \\ \left\|u^{k+1}\right\|_2 &\leq |\alpha_k - 1| \left\|x^k - x^\star\right\|_2 + \alpha_k \left\|x^{k+1} - x^\star\right\|_2\end{aligned}$$

By the nonexpansivity of the displacement operator, i.e., $\mathbf{I} - s\nabla g$, where $\mathbf{I}$ is the identity operator, we obtain

$$\begin{aligned}\left\|u^{k+1}\right\|_2 - \left\|u^k\right\|_2 &\leq \alpha_k \Big|\left\|x^{k+1} - x^\star\right\|_2 - \left\|y^k - x^\star\right\|_2\Big| , \\ &\leq \alpha_k \Big|\left\|r^{k+1}\right\|_2 + s_k \left\|\epsilon_1^k\right\|_2 + C_{\rho, s_{k_0}}\Big| , \quad \forall s_k \leq \frac{1}{L},\end{aligned} \tag{4.64}$$

where we have used inequality (A.18). Rearranging and taking into account that all the terms inside the absolute value are nonnegative, we obtain

$$\left\|u^{k+1}\right\|_2 \leq \left\|u^k\right\|_2 + \alpha_k\left(\left\|r^{k+1}\right\|_2 + s_k\left\|\epsilon_1^k\right\|_2 + C_{\rho,s_{k_0}}\right), \quad \forall s_k \leq \frac{1}{L}. \tag{4.65}$$

Using the bound $\left\|r^{i+1}\right\|_2 \leq \sqrt{2s\epsilon_2^i}$ from Lemma 1, by induction and backward substitution

$$\left\|u^{k+1}\right\|_2 \leq \left\|u^0\right\|_2 + \sum_{j=1}^{k}\alpha_j\left(\sqrt{2s\epsilon_2^j} + s_j\left\|\epsilon_1^j\right\|_2 + C_{\rho,s_{k_0}}\right), \quad \forall s_j \leq \frac{1}{L}. \tag{4.66}$$

where $\left\|u^0\right\|_2 = \|x^0 - x^\star\|_2$. By multiplying we obtain the bound of Theorem 4.

4.5. Proof of Theorem 5. This result is about the accelerated version of approximate PGD, but with random proximal computation error ϵ_{2_Ω}, component-wise bounded gradient error ϵ_{1_Ω} and bounded residuals $\left\|x_\Omega^k - x^\star\right\|_2$. As the algorithm generates a sequence of random vectors $\{x_\Omega^k\}$, the residual vector sequence $\{r_\Omega^k\}$ will also be a random. Let $\nu_\Omega = \epsilon_1^{i-1} - \frac{1}{s}r^i$ and let $\{T_k\}$ denote the second error term in (3.14) [Theorem 4], i.e.,

$$T_k = \begin{cases} 0, & k=0 \\ \sum_{i=1}^k \alpha_i {\nu_\Omega^i}^\top u_\Omega^i, & k = 1,2,\dots, \end{cases} \tag{4.67}$$

where

$$u_\Omega^i = x^\star - x_\Omega^i + (1-\alpha_{i-1})(x_\Omega^i - x_\Omega^{i-1}). \tag{4.68}$$

The first step is to show that $\{T_k\}$ is a martingale. Recall that a sequence of random variables $T_0, T_1, \dots$ is a martingale with respect to the sequence $X_0, X_1, \dots$ if, for all $k \geq 0$, the following conditions hold:

- T_k is a function of $X_0, X_1, \dots, X_k$;
- $\mathbb{E}[|T_k|] < \infty$;
- $\mathbb{E}[T_{k+1}|X_0, X_1, \dots, X_k] = T_k$.

A sequence of random variables $T_0, T_1, \dots$ is called a martingale when it is a martingale with respect to itself. That is, $\mathbb{E}[|T_k|] < \infty$, and $\mathbb{E}[T_{k+1}|T_0, T_1, \dots, T_k] = T_k$. We now show that Assumptions 2 and 3 imply that $\{T_k\}_{k\geq 0}$ is a martingale. Specifically, (3.10a) and (3.13a), we have

$$\mathbb{E}\left[\nu_\Omega^k | \nu_\Omega^1 \dots \nu_\Omega^{k-1}\right] = \mathbb{E}\left[\nu_\Omega^k\right] = 0.$$

And from (3.10c) and (3.13b), we have

$$\mathbb{E}\left[{\nu_\Omega^k}^\top x_\Omega^k | \nu_\Omega^1 \dots \nu_\Omega^{k-1}, x_\Omega^1 \dots x_\Omega^{k-1}\right] = \mathbb{E}\left[{\nu_\Omega^k}^\top x_\Omega^k\right] = 0.$$

We have from (4.67),

$$T_k = T_{k-1} + \alpha_k {\nu_\Omega^k}^\top u_\Omega^k. \tag{4.69}$$

Substituting for u_Ω^k using (4.68) gives,

$$T_k = T_{k-1} + \alpha_k\alpha_{k-1}{\nu_\Omega^k}^\top(x^\star - x_\Omega^k) + \alpha_k(1-\alpha_{k-1}){\nu_\Omega^k}^\top(x^\star - x^{k-1}). \tag{4.70}$$

Taking the conditional expectation from both sides and proceeding as in Section 4.2, we obtain $\mathbb{E}\big[T_k|T_1 \dots T_{k-1}\big] = T_{k-1}$, i.e., $T_1, T_2, \dots, T_k$ is a martingale.

In what follows, we establish upper bounds on the absolute value of the martingale $\{T_k\}$. By noticing that $\big|T_k - T_{k-1}\big| = \big|{\nu_\Omega^k}^\top u_\Omega^k\big| \leq \alpha_k\Big(\sqrt{n}\delta M_{\nabla g} + \sqrt{2\epsilon_2^k/s}\Big)\left\|u_\Omega^k\right\|_2$, where we have used Cauchy-Schwarz, etc. [27, Corollary 2.20] then yields

$$|T_k| \leq \gamma|\delta|M_{\nabla g}\sqrt{n\sum_{i=1}^{k} i^2\left\|u_\Omega^i\right\|_2^2} + \gamma\sqrt{2s}\sqrt{\sum_{i=1}^{k} i^2\left\|u_\Omega^i\right\|_2^2\epsilon_2^i} \tag{4.71}$$

$$\leq \gamma|\delta|M_{\nabla g}\sqrt{n}\sum_{i=1}^{k} i\left\|u_\Omega^i\right\|_2 + \gamma\sqrt{2s}\sum_{i=1}^{k} i\left\|u_\Omega^i\right\|_2\sqrt{\epsilon_2^i}$$

where $M_{\nabla g} = \sup_{i\in\mathbb{N}_+}\Big\{\left\|\nabla g(x^i)\right\|_\infty\Big\}$ is the upper bound on the elements of the gradient. Let $\{S_k\}$ denote the first error term in (4) [Theorem 4] i.e.,

$$S_k = \sum_{i=1}^{k}\alpha_i^2\epsilon_{2_\Omega}^i. \tag{4.72}$$

If $0 \leq \epsilon_{2_\Omega}^k \leq \varepsilon_0$ and $\alpha_k \leq k$, then applying [27, Proposition 2.5] to $S_k = \sum_{i=1}^{k}\alpha_i^2\epsilon_{2_\Omega}^i$ with $0 \leq \epsilon_{2_\Omega}^k \leq \varepsilon_0$ and $\alpha_k \leq k$ yields

$$S_k \leq \mathbb{E}\big[\sum_{i=1}^{k}\alpha_i^2\epsilon_{2_\Omega}^i\big] + \frac{\gamma}{2}\sqrt{\sum_{i=1}^{k} i^4(\epsilon_{2_\Omega}^i)^2} \leq \mathbb{E}\big[\sum_{i=1}^{k}\alpha_i^2\epsilon_{2_\Omega}^i\big] + \frac{\gamma}{2}\sum_{i=1}^{k} i^2\epsilon_{2_\Omega}^i, \tag{4.73}$$

with probability at least $1 - 2\exp(-\frac{\gamma^2}{2})$. Applying the probability union bound and assuming that $\left\|u_\Omega^i\right\|_2^2 \leq D_u^2\left\|x^0 - x^\star\right\|_2^2$ holds with probability p completes the proof of Theorem 5.

5. Experimental Results.

We now experimentally assess the proposed bounds on an ℓ_1-regularized model predictive control (MPC) problem. We consider a discrete linear time invariant (LTI) state space model of a spacecraft [13]. The approximation errors are simulated error sequences generated from a truncated Gaussian distribution.

5.1. Model Predictive Control (MPC).

The ℓ_1-regularized MPC can be formulated as

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\;\; f(x) := g(x) + h(x)\,, \tag{5.1}$$

where $g : \mathbb{R}^n \to \mathbb{R}$ is the following real-valued, convex and differentiable function,

$$g(x) := \left\|\big(\Phi^\top Q\Phi + R\big)^{\frac{1}{2}}x - \big(\Phi^\top Q\Phi + R\big)^{-\frac{1}{2}}\Phi^\top Q\big(R_s - \Psi x(k)\big)\right\|_2^2,$$

and $h : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ is the nondifferentiable convex ℓ_1-norm

$$h(x) := \lambda\left\|x\right\|_1,$$

with $x \in \mathbb{R}^{p\cdot N_c\times 1}$ being the vectorized differential control $\Delta u = u^k - u^{k-1} \in \mathbb{R}^{p\times N_c}$, where p is the input dimension of the system and N_c is the control horizon. The regularization parameter $\lambda \in \mathbb{R}^+$ is a positive scalar. $Q \in \mathbb{R}^{N_p \cdot m \times m \cdot N_p}$ and $R \in \mathbb{R}^{p \cdot N_c \times p \cdot N_c}$ are positive semi-definite design matrices where m is the output dimension and N_p is the prediction horizon. $R_s \in \mathbb{R}^{m \cdot N_p \times 1}$ is the vectorization of the matrix that is constructed by N_p times stacking of the set-point vector $r(k)$. $\Phi \in \mathbb{R}^{m \cdot N_p \times p \cdot N_c}$ and $\Psi \in \mathbb{R}^{m \cdot N_p \times n}$ are augmented matrices which can be obtained from the spacecraft LTI discrete state-space model (A, B, C) of [13] using a standard formula [28, Eq. 1.12]. For simulation, we select the problem's matrices as follows,

$$Q = \text{diag}(500.0, 500.0, 500.0, 10^{-7}, 1.0, 1.0, 1.0, 500.0, 500.0, 500.0, 10^{-7}, 1.0, 1.0, 1.0);$$
$$R = \text{diag}(200.0, 200.0, 200.0, 1.0, 200.0, 200.0, 200.0, 1.0),$$

and set the regularization parameter $\lambda = 2.5021$. The control and prediction horizons are set to $N_c = N_p = 5$. The quadratic term of the ℓ_1-regularized MPC problem, $g(x)$, has a gradient's Lipschitz constant of $L = 11539$, and therefore, a stepsize of $s = \frac{1}{L}$ is adopted.

For the simulated errors, we use $\epsilon_{1_\Omega}^k = \nabla g(x^k) \odot \text{trand}(-\delta, \delta)$ where $\text{trand}(a, b)$ is the doubly truncated normal distribution [8] with lower and upper truncation points a and b, respectively. δ is the gradient element-wise precision, which is a scalar upper bound on the gradient error. $\epsilon_2^k = \text{trand}(0, \epsilon_0)$ where ϵ_0 is a scalar upper bound on the proximal computation error. The output of the distribution function $\text{trand}(l, u)$ is a vector randomly generated from the standard multivariate normal distribution truncated over the region $[l, u]$.

5.2. Results. The deterministic and probabilistic bounds are plotted and superimposed with the bound (2.1) of [25] in Figure 1 and Figure 2. The latter is denoted by `Schmidt_1` (`Schmidt_2` in the accelerated case) and the proposed bounds are denoted by `Thrm_1` and `Thrm_2` (`Thrm_4` and `Thrm_5` in the accelerated case), respectively.

Notice that we expect the effect of ϵ_1^k to be negligible near the optimum since, according to model (3.8), ϵ_1^k is proportional to the magnitude of the gradient. However, depending on the choice of the upper bound of ϵ_2^k in the proximal operation step (3.7), the effect of the error ϵ_2 can still be significant and sometimes permanent even near the optimum as we will see next.

In the presence of small gradient and proximal computation errors, the bounds in Theorem 1, Theorem 2 practically coincide with (2.1). Therefore, in order to emphasize the sharpness of the proposed bounds, we run the simulation with $|\epsilon_1^k| \leq 2.2\times 10; \epsilon_2^k \leq 10$ for the nonaccelerated case (Figure 1), and with $|\epsilon_1^k| \leq 2.2\times 10^{-4}; \epsilon_2^k \leq 10^{-4}$ for the accelerated case (Figure 2).

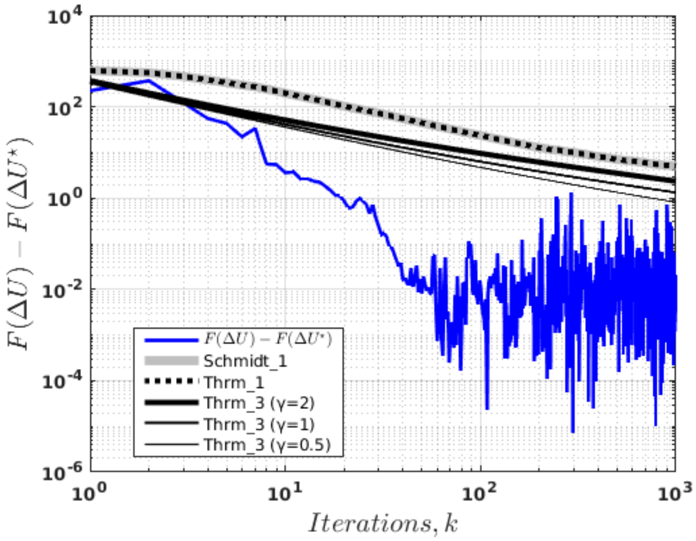

Fig. 1: Upper bounds based on Theorems 1 & 3 vs Proposition 1 ((2.1)) in Schmidt et al. 2010 [25] (with $\delta = 2.2 \times 10^1; \epsilon_0 = 10^1$).

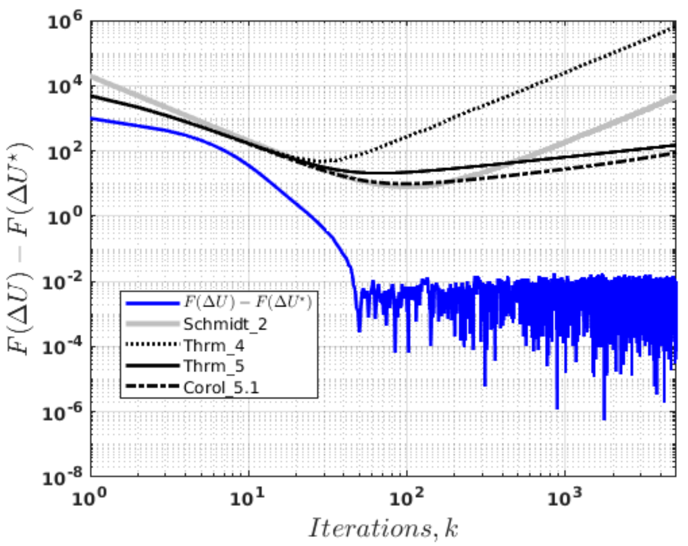

Fig. 2: Upper bounds based on Theorems 4 & 5 vs Proposition 2 in Schmidt et al. 2010 [25] (with $\delta = 2.2 \times 10^{-4}$; $\epsilon_0 = 10^{-4}$).

Figure 1 suggests that by using the proposed probabilistic bounds, we can predict the suboptimality, i.e., $f - f^\star$, more accurately and the improvement is more significant with lower values of γ (with lower probabilities). Note that the probabilistic bounds can possibly drop below the suboptimality plot $(f - f^\star)$ during some iterations; however, this would not present any conflict with the theory as this is what can be expected from probabilistic statements (dependent on the parameter γ) which do not hold 100% of the algorithm's execution time.

From Figure 2, we can see that none of the bounds can successfully estimate the function values suboptimality in the accelerated case, however, the probabilistic bound of Theorem 5 gives the best estimate and the slowest divergence rate. The bound of Corollary 5.1 slightly improves on Theorem 5 but still diverges, although at the slowest rate.

6. Conclusions. We have analysed the convergence of the proximal gradient descent under computational errors. We derived deterministic and probabilistic upper bounds on the objective function value which we used as an assertion for convergence test. We considered the special case in which the gradient $\nabla g(x^k)$ of g is computed with errors as well as the proximal operator prox_h (with respect to h) is evaluated approximately. We also considered accelerated versions of the proximal gradient descent, which is known to converge faster in the error-free case, but we have shown that this comes at a price of amplified perturbations, which may lead to divergence. We proved that the effect of each contributing error term can be decoupled under mild assumptions. We also obtained probabilistic bounds with three main advantages:

- The bounds are sharper (i.e., reflect practical performance better);
- The bounds are simpler to interpret and predict *a priori*;
- The contribution of each error term is decoupled.

We have also shown that some error terms follow martingale sequences when error conditional mean independence and data conditional mean independence assumptions both hold. Finally, we have perceived that in the accelerated case, the algorithm actually converges to some suboptimal level around the optimum, however, the latter could not be determined using the current convergence bounds. This opens the possibility of other types of analyses with different error models.

Appendix A. Supplementary results. The following Lemma establishes bounds on the norm of the residual error vector due to proximal error (forward error).

LEMMA 1. *Consider problem (3.1) and let Assumption 1 hold. For $L, s > 0$, define $G\colon \mathbb{R}^n \times \mathbb{R}^n \to (-\infty, \infty]$ as the proper, closed, and L-strongly convex function*

$$G\big(y,\, x\big) := g(y) + \nabla g(y)^\top (x - y) + \frac{1}{2s}\left\|x - y\right\|_2^2 + h(x),$$

Define $\widehat{y}^\star := \arg\min G\big(y,\, x\big)$ as the minimizer of G with respect to y when x is fixed, and $y^\star \in \{y : G(y, x) - G(\widehat{y}^\star, x) \le \epsilon_2\}$ as an ϵ_2-approximate solution of the same problem. Then,

$$\left\|\widehat{y}^\star - y^\star\right\|_2 \le \sqrt{2s\epsilon_2}.$$

THEOREM 6 (**Quasi-Fejér monotonicity of the sequence generated by the proximal gradient method**). *Let $\{x^k\}_{k\ge 0}$ be the sequence generated by the approximate proximal gradient (3.7) for solving problem (3.1) under Assumption 1 and with $s_k \le \frac{1}{L}$. Assume that, for $k \ge k_0$, we have $\epsilon_2^k \le c_2 \left\|x^{k+1} - x^k\right\|_2 \le c_2\rho$ and $\|\epsilon_1^k\|_2 \le c_1 \left\|\nabla g(x^{k+1}) - \nabla g(x^k)\right\|_2$. Then for any $x^\star \in X^\star$ and $k \ge 0$ we have*

$$\left\|x^{k+1} - x^\star\right\|_2 \le \left\|x^k - x^\star\right\|_2 + \left\|r^{k+1}\right\|_2 + s_k \left\|\epsilon_1^k\right\|_2 + C_{\rho,1/L}, \tag{A.1}$$

where $C_{\rho,1/L} = \sqrt{2Lc_2\rho} + c_1\rho$. If $E^{k+1} := \left\|r^{k+1}\right\|_2 + s_k \left\|\epsilon_1^k\right\|_2$ is a positive and *absolutely summable sequence, then* $\{x^k\}_{k\geq 0}$ *is a quasi-Féjer sequence relative to the set* $X^\star$*.*

Proof. we have

$$\left\|x^{k+1}-x^{k_0+1}\right\|_2 = \left\|\operatorname{prox}^{\epsilon_2^k}_{s_k h}(x^k - s_k \nabla^{\epsilon_1^k} g(x^k)) - \operatorname{prox}^{\epsilon_2^{k_0}}_{s_{k_0} h}(x^{k_0} - s_{k_0}\nabla^{\epsilon_1^{k_0}} g(x^{k_0}))\right\|_2 . \tag{A.2}$$

Writing $\operatorname{prox}^{\epsilon_2^k}_{s_k h}(x)$ as $\operatorname{prox}_{s_k h}(x) + r^k$ and $\nabla^{\epsilon_1^k} g(x)$ as $\nabla g(x) + \epsilon_1^k$ for any suboptimal solution x^{k_0} of (3.1), we obtain

$$\begin{aligned}\left\|x^{k+1}-x^{k_0+1}\right\|_2 = \Big\| &\operatorname{prox}_{s_k h}(x^k - s_k\nabla g(x^k) - s_k\epsilon_1^k) \\ &- \operatorname{prox}_{s_{k_0} h}(x^{k_0} - s_{k_0}\nabla g(x^{k_0}) - s_{k_0}\epsilon_1^{k_0}) + r^{k+1} - r^{k_0}\Big\|_2 .\end{aligned} \tag{A.3}$$

By assumption we have $\epsilon_2^k \leq c_2\left\|x^{k+1}-x^k\right\|_2$, or equivalently, $\left\|r^{k+1}\right\|_2 \leq \sqrt{2c_2\left\|x^{k+1}-x^k\right\|_2/s}$ and $\epsilon_2^k \leq c_1\left\|\nabla g(x^{k+1}) - \nabla g(x^k)\right\|_2$ $\leq c_1 L\left\|x^{k+1}-x^k\right\|_2$ for $k \geq k_0$. By the triangle inequality we have

$$\begin{aligned}\left\|x^{k+1}-x^{k_0+1}\right\|_2 \leq \Big\| &\operatorname{prox}_{s_k h}(x^k - s_k\nabla g(x^k) - s_k\epsilon_1^k) \\ &- \operatorname{prox}_{s_{k_0} h}(x^{k_0} - s_{k_0}\nabla g(x^{k_0}) - s_{k_0}\epsilon_1^{k_0})\Big\|_2 + \left\|r^{k+1}\right\|_2 + \left\|r^{k_0+1}\right\|_2 \\ \leq \Big\| &\operatorname{prox}_{s_k h}(x^k - s_k\nabla g(x^k) - s_k\epsilon_1^k) \\ &- \operatorname{prox}_{s_{k_0} h}(x^{k_0} - s_{k_0}\nabla g(x^{k_0}) - s_{k_0}\epsilon_1^{k_0})\Big\|_2 + \left\|r^{k+1}\right\|_2 + \sqrt{\frac{2c_2\rho}{s}}\end{aligned} \tag{A.4}$$

where we have used $\left\|x^{k_0+1}-x^{k_0}\right\|_2 \leq \rho$.

By the nonexpansivity of the proximal operator we have

$$\begin{aligned}\left\|x^{k+1}-x^{k_0+1}\right\|_2 &\leq \left\|[x^k - s_k\nabla g(x^k)] - [x^{k_0} - s_{k_0}\nabla g(x^{k_0})]\right\|_2 + \left\|r^{k+1}\right\|_2 + \sqrt{\frac{2c_2\rho}{s_{k_0}}} \\ &\quad + s_k\left\|\epsilon_1^k\right\|_2 + s_{k_0}\left\|\epsilon_1^{k_0}\right\|_2 \\ &\leq \left\|[x^k - s_k\nabla g(x^k)] - [x^{k_0} - s_{k_0}\nabla g(x^{k_0})]\right\|_2 + \left\|r^{k+1}\right\|_2 + s_k\left\|\epsilon_1^k\right\|_2 \\ &\quad + \sqrt{\frac{2c_2\rho}{s_{k_0}}} + s_{k_0}c_1 L\rho\end{aligned} \tag{A.5}$$

By the nonexpansivity of the gradient descent operator, i.e., $\mathbf{I} - s\nabla g$, we obtain

$$\left\|x^{k+1}-x^{k_0+1}\right\|_2 \leq \left\|x^k - x^{k_0}\right\|_2 + \left\|r^{k+1}\right\|_2 + s_k\left\|\epsilon_1^k\right\|_2 + C_\rho, \quad \forall s_k \leq \frac{1}{L} \tag{A.6}$$

$$= \left\|x^k - x^{k_0}\right\|_2 + E^{k+1} + C_\rho, \tag{A.7}$$

$$\left\|x^{k+1}-x^{k_0}-(x^{k_0+1}-x^{k_0})\right\|_2 \leq \left\|x^k-x^{k_0}\right\|_2 + E^{k+1} + C_\rho, \tag{A.8}$$

where $C_\rho = \sqrt{\frac{2c_2\rho}{s_{k_0}}} + s_{k_0}c_1L\rho$ and $E^{k+1} = \left\|r^{k+1}\right\|_2 + s_k\left\|\epsilon_1^k\right\|_2$.

By the triangle difference inequality we have

$$\left|\left\|x^{k+1}-x^{k_0}\right\|_2 - \left\|x^{k_0+1}-x^{k_0}\right\|_2\right| \leq \left\|x^k-x^{k_0}\right\|_2 + E^{k+1} + C_\rho. \tag{A.9}$$

For $x^{k_0+1} \approx x^{k_0} = x^\star$ we have

$$\left\|x^{k+1}-x^\star\right\|_2 \leq \left\|x^k-x^\star\right\|_2 + E^{k+1} + C_\rho, \tag{A.10}$$

From (A.10) and by [9, Definition 1.1], the sequence $\{x^k\}_{k\geq 1}$ is quasi-Féjer relative to the set $X^\star$ if $\{E^k\}_{k\geq 1}$ is positive and absolutely summable.
□

THEOREM 7 (**Quasi-Fejér monotonicity of the sequence generated by the accelerated proximal gradient method**). *Let $\{x^k\}_{k\geq 0}$ be the sequence generated by the approximate accelerated proximal gradient (3.6) for solving problem (3.1) under Assumption 1 and with $s_k \leq \frac{1}{L}$. Assume we have summable iterative displacements $\left\|x^k-x^{k-1}\right\|_2$ and that, for $k \geq k_0$, we have $\epsilon_2^k \leq c_2\left\|x^{k+1}-x^k\right\|_2 \leq c_2\rho$ and $\|\epsilon_1^k\|_2 \leq c_1\left\|\nabla g(x^{k+1})-\nabla g(x^k)\right\|_2^\top$, then for any $x^{k_0} \in X^{k_0}$ and $k \geq 0$ we have*

$$\left\|x^{k+1}-x^{k_0}\right\|_2 \leq \left\|x^k-x^{k_0}\right\|_2 + \left\|x^k-x^{k-1}\right\|_2 + E^{k+1} + C_{\rho,1/L} \tag{A.11}$$

where $C_{\rho,1/L} = \sqrt{2Lc_2\rho} + c_1\rho$, $E^{k+1} = \left\|r^{k+1}\right\|_2 + s_k\left\|\epsilon_1^k\right\|_2$. If $E^{k+1} := \left\|r^{k+1}\right\|_2 + s_k\left\|\epsilon_1^k\right\|_2$ is a positive and absolutely summable sequence, then $\{x^k\}_{k\geq 0}$ is a quasi-Féjer sequence relative to the set X^{k_0}.

Proof. For any optimal solution x^{k_0} of (3.1), we have

$$\left\|x^{k+1}-x^{k_0+1}\right\|_2 = \left\|\text{prox}_{s_kh}^{\epsilon_2^k}(y^k - s_k\nabla^{\epsilon_1^k}g(y^k)) - \text{prox}_{s_{k_0}h}^{\epsilon_2^{k_0}}(x^{k_0} - s_{k_0}\nabla^{\epsilon_1^{k_0}}g(x^{k_0}))\right\|_2. \tag{A.12}$$

Rewriting $\text{prox}_{s_kh}^{\epsilon_2^k}(y)$ as $\text{prox}_{s_kh}(y) + r^k$ and $\nabla^{\epsilon_1^k}g(y)$ as $\nabla g(y) + \epsilon_1^k$ we obtain

$$\begin{aligned}\left\|x^{k+1}-x^{k_0+1}\right\|_2 = \Big\| &\text{prox}_{s_kh}(y^k - s_k\nabla g(y^k) - s_k\epsilon_1^k) \\ &- \text{prox}_{s_{k_0}h}(x^{k_0} - s_{k_0}\nabla g(x^{k_0}) - s_{k_0}\epsilon_1^{k_0}) + r^{k+1} - r^{k_0}\Big\|_2.\end{aligned} \tag{A.13}$$

By assumption we have $\epsilon_2^k \leq c_2\left\|x^{k+1}-x^k\right\|_2$ and $\epsilon_2^k \leq c_1\left\|\nabla g(x^{k+1})-\nabla g(x^k)\right\|_2 \leq c_1L\left\|x^{k+1}-x^k\right\|_2$ for $k \geq k_0$. By the triangle inequality we have

$$\begin{aligned}\|x^{k+1} - x^{k_0+1}\|_2 \leq & \Big\| \text{prox}_{s_k h}(y^k - s_k \nabla g(y^k) - s_k \epsilon_1^k) \\ & - \text{prox}_{s_{k_0} h}(x^{k_0} - s_{k_0} \nabla g(x^{k_0}) - s_{k_0} \epsilon_1^{k_0}) \Big\|_2 + \|r^{k+1}\|_2 + \|r^{k_0+1}\|_2 \\ \leq & \Big\| \text{prox}_{s_k h}(y^k - s_k \nabla g(y^k) - s_k \epsilon_1^k) \\ & - \text{prox}_{s_{k_0} h}(x^{k_0} - s_{k_0} \nabla g(x^{k_0}) - s_{k_0} \epsilon_1^{k_0}) \Big\|_2 + \|r^{k+1}\|_2 + \sqrt{\frac{2c_2\rho}{s}}\end{aligned} \tag{A.14}$$

where we have used $\|x^{k_0+1} - x^{k_0}\|_2 \leq \rho$.

By the nonexpansivity of the proximal operator we have

$$\begin{aligned}\|x^{k+1} - x^{k_0+1}\|_2 \leq & \|[y^k - s_k \nabla g(y^k)] - [x^{k_0} - s_{k_0} \nabla g(x^{k_0})]\|_2 + \|r^{k+1}\|_2 \\ & + \sqrt{\frac{2c_2\rho}{s_{k_0}}} + s_k \|\epsilon_1^k\|_2 + s_{k_0} \|\epsilon_1^{k_0}\|_2 \\ \leq & \|[y^k - s_k \nabla g(y^k)] - [x^{k_0} - s_{k_0} \nabla g(x^{k_0})]\|_2 + \|r^{k+1}\|_2 \\ & + s_k \|\epsilon_1^k\|_2 + \sqrt{\frac{2c_2\rho}{s_{k_0}}} + s_{k_0} c_1 L \rho\end{aligned} \tag{A.15}$$

By the nonexpansivity of the gradient descent operator, i.e., $\mathbf{I} - s\nabla g$, we obtain

$$\begin{aligned}\|x^{k+1} - x^{k_0+1}\|_2 \leq & \|y^k - x^{k_0}\|_2 + \|r^{k+1}\|_2 + s_k \|\epsilon_1^k\|_2 + C_{\rho,s_{k_0}}, \quad \forall s_k \leq \frac{1}{L} \\ = & \|x^k - x^{k_0} + \beta_k(x^k - x^{k-1})\|_2 + E^{k+1} + C_{\rho,s_{k_0}} \\ = & \|x^k - x^{k_0}\|_2 + \|x^k - x^{k-1}\|_2 + E^{k+1} + C_{\rho,s_{k_0}},\end{aligned} \tag{A.16}$$

where $C_{\rho,s_{k_0}} = \sqrt{\frac{2c_2\rho}{s_{k_0}}} + s_{k_0} c_1 L \rho$, $E^{k+1} = \|r^{k+1}\|_2 + s_k \|\epsilon_1^k\|_2$ and we used $\beta_k \leq 1$. By the triangle difference inequality we have

$$\big| \|x^{k+1} - x^{k_0}\|_2 - \|x^{k_0+1} - x^{k_0}\|_2 \big| \leq \|x^k - x^{k_0}\|_2 + \|x^k - x^{k-1}\|_2 + E^{k+1} + C_{\rho,s_{k_0}}, \tag{A.17}$$

For $x^{k_0+1} \approx x^{k_0} = x^\star$ we have

$$\|x^{k+1} - x^\star\|_2 \leq \|x^k - x^\star\|_2 + \|x^k - x^{k-1}\|_2 + E^{k+1} + C_{\rho,s_{k_0}}, \tag{A.18}$$

From (A.18) and by [9, Definition 1.1], the sequence $\{x^k\}_{k\geq 1}$ is quasi-Féjer relative to the set $X^\star$ if $\{E^k\}_{k\geq 1}$ is positive and absolutely summable provided we have summable iterative displacements $\|x^k - x^{k-1}\|_2$. ∎

Acknowledgements. This work was supported by the Engineering and Physical Research Council (EPSRC) grants (EP/T026111/1 and EP/S000631/1), and the MOD University Defence Research Collaboration (UDRC).

REFERENCES

[1] M. V. Afonso, J. M. Bioucas-Dias, and M. A. Figueiredo, *Fast image recovery using variable splitting and constrained optimization*, IEEE transactions on image processing, 19 (2010), pp. 2345–2356.

[2] Y. F. Atchade, G. Fort, and E. Moulines, *On stochastic proximal gradient algorithms*, arXiv preprint arXiv:1402.2365, 23 (2014).

[3] J.-F. Aujol and C. Dossal, *Stability of over-relaxations for the forward-backward algorithm, application to fista*, SIAM Journal on Optimization, 25 (2015), pp. 2408–2433.

[4] A. Beck, *First-order methods in optimization*, vol. 25, SIAM, 2017.

[5] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM journal on imaging sciences, 2 (2009), pp. 183–202.

[6] D. P. Bertsekas and A. Scientific, *Convex optimization algorithms*, Athena Scientific Belmont, 2015.

[7] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd, *First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems*, SIAM Journal on Optimization, 28 (2018), pp. 2131–2151.

[8] J. Cha, B. R. Cho, and J. L. Sharp, *Rethinking the truncated normal distribution*, International Journal of Experimental Design and Process Optimisation, 3 (2013), pp. 327–363.

[9] P. L. Combettes, *Quasi-fejérian analysis of some optimization algorithms*, in Studies in Computational Mathematics, vol. 8, Elsevier, 2001, pp. 115–152.

[10] P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Modeling & Simulation, 4 (2005), pp. 1168–1200.

[11] C. Cortes and V. Vapnik, *Support-vector networks*, Machine learning, 20 (1995), pp. 273–297.

[12] D. Davis, B. Edmunds, and M. Udell, *The sound of apalm clapping: Faster nonsmooth nonconvex optimization with stochastic asynchronous palm*, in Advances in Neural Information Processing Systems, 2016, pp. 226–234.

[13] Ø. Hegrenæs, J. T. Gravdahl, and P. Tøndel, *Spacecraft attitude control using explicit model predictive control*, Automatica, 41 (2005), pp. 2107–2114.

[14] N. J. Higham, *Accuracy and stability of numerical algorithms*, SIAM, 2002.

[15] N. Lawrence, M. Seeger, and R. Herbrich, *Fast sparse Gaussian process methods: The informative vector machine*, in Proceedings of the 16th annual conference on neural information processing systems, no. CONF, 2003, pp. 609–616.

[16] N. D. Lawrence and R. Herbrich, *A sparse Bayesian compression scheme-the informative vector machine*, in NIPS 2001 workshop on kernel methods, Citeseer, 2001.

[17] M. Nagahara, D. E. Quevedo, and D. Nešić, *Maximum hands-off control: a paradigm of control effort minimization*, IEEE Transactions on Automatic Control, 61 (2015), pp. 735–747.

[18] Y. Nesterov, *A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2)*, in Doklady an ussr, vol. 269, 1983, pp. 543–547.

[19] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, 2004.

[20] A. Nitanda, *Stochastic proximal gradient descent with acceleration techniques*, in Advances in Neural Information Processing Systems, 2014, pp. 1574–1582.

[21] P. Ochs, J. Fadili, and T. Brox, *Non-smooth non-convex bregman minimization: Unification and new algorithms*, Journal of Optimization Theory and Applications, 181 (2019), pp. 244–278.

[22] D. P. Palomar and Y. C. Eldar, *Convex optimization in signal processing and communications*, Cambridge university press, 2010.

[23] J. Quinonero-Candela and C. E. Rasmussen, *A unifying view of sparse approximate Gaussian process regression*, The Journal of Machine Learning Research, 6 (2005), pp. 1939–1959.

[24] L. Rosasco, S. Villa, and B. C. Vũ, *Convergence of stochastic proximal gradient algorithm*, Applied Mathematics & Optimization, (2019), pp. 1–27.

[25] M. Schmidt, N. L. Roux, and F. R. Bach, *Convergence rates of inexact proximal-gradient methods for convex optimization*, in Advances in neural information processing systems, 2011, pp. 1458–1466.

[26] S. Villa, S. Salzo, L. Baldassarre, and A. Verri, *Accelerated and inexact forward-backward algorithms*, SIAM Journal on Optimization, 23 (2013), pp. 1607–1633.

[27] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48, Cambridge University Press, 2019.

[28] L. Wang, *Model predictive control system design and implementation using MATLAB®*, Springer Science & Business Media, 2009.

[29] Y. Zhou, Y. Liang, Y. Yu, W. Dai, and E. P. Xing, *Distributed proximal gradient algorithm*

for partially asynchronous computer clusters, The Journal of Machine Learning Research, 19 (2018), pp. 733–764.

[30] Y. Zhou, Y. Yu, W. Dai, Y. Liang, and E. Xing, *On convergence of model parallel proximal gradient algorithm for stale synchronous parallel system*, in Artificial Intelligence and Statistics, PMLR, 2016, pp. 713–722.