New nonasymptotic convergence rates of stochastic proximal point algorithm for convex optimization problems with many constraints

Andrei Patrascu
University Bucharest, Str. Academiei 14, 010014, Bucharest, Romania

ARTICLE HISTORY
Compiled August 8, 2019

ABSTRACT
Significant parts of the recent stochastic optimization literature focused on analyzing the theoretical and practical behaviour of stochastic first order schemes under various convexity properties. Due to its simplicity, the traditional method of choice for most supervised machine learning problems is the stochastic gradient descent (SGD) method, which is known to have a relatively slow convergence. Many iteration improvements and accelerations have been added to the pure SGD in order to boost its convergence under different (strong) convexity conditions when constraints are present. However, full projections on complicated feasible set, smoothness or strong convexity assumptions are an essential requirement for these improved stochastic first-order schemes. In this paper novel convergence results are presented for the stochastic proximal point (SPP) algorithm for (non-)strongly convex optimization with many constraints. We show that a prox-quadratic growth assumption is sufficient to guarantee for SPP $O\left(\frac{1}{k}\right)$ convergence rate, in terms of the distance to the optimal set, using only projections onto a simple component set. Furthermore, linear convergence is obtained for interpolation setting, when the optimal set of the expected cost is included into the optimal sets of each functional component.

KEYWORDS
Stochastic proximal point, randomized alternating projections, quadratic growth, nonsmooth optimization, linear convergence, sublinear convergence rate

1. Introduction

In this paper we consider the following constrained stochastic convex optimization problem:

$$\min_{x \in \mathbb{R}^n} \, f(x) := (\mathbb{E}[f(x; \xi)]) ,$$

$$\text{s.t.} \quad x \in \bigcap_{\zeta \in \Omega_2} X_\zeta ,$$

where $\xi$ and $\zeta$ are a random variables associated with probability spaces $(\mathbb{P}, \Omega_1)$ and $(\mathbb{P}, \Omega_2)$, respectively. Each function $f(\cdot, \xi) : \mathbb{R}^n \mapsto (-\infty, +\infty]$ is proper convex and lower-semicontinuous, each set $X_\zeta$ is closed and convex and $\mathbb{E}_{\xi \in \Omega_1} [\cdot]$ is the expectation over $\xi$. In general, many existing primal schemes encounter computational difficulties when a large (possibly infinite) number of constraints are present, since they require projections onto complicated feasible set. Clearly, when $f = 0$, the model (1) reduces to a convex feasibility probl-
lm (CFP):

$$\min_{x \in \mathbb{R}^n} \mathbb{E}[\mathbb{1}_{X_\xi}(x)],$$

(2)

asking to find a feasible point in the intersection $X := \bigcap_{\xi \in \Omega_2} X_\xi$. There exist plenty of iterative algorithms which solve CFPs efficiently under various regularity conditions on the feasible sets, from which we mention only the (randomized) alternating projections schemes due to their relevance to our paper, see [4,6,15,17].

On the other hand, for the unconstrained case $X = \mathbb{R}^n$, when the model reduces to

$$\min_{x \in \mathbb{R}^n} \mathbb{E}[f(x; \xi)],$$

(accelerated) stochastic gradient schemes are available under suitable assumptions on the smoothness and convexity properties of the functions $f(\cdot; \xi)$. However, in general, the nonsmooth components defined by the indicator functions bring several difficulties in the behavior of traditional stochastic schemes. These algorithms typically tend to compute projections on the entire feasible set $X$, which can be computationally prohibitive. The method of choice for general stochastic optimization, confirmed by the literature, is the stochastic gradient descent (SGD). Usually regarded as very simple and intuitive, SGD randomly samples $\xi$ at each iteration and takes a step along the gradient of the chosen individual function. In the constrained case, projected SGD performs additionally a full projection step onto the entire feasible set [13,16,29]. The theoretical guarantees of SGD and of its improved or accelerated variants show sublinear convergence rates under appropriate continuity or strong convexity assumptions, see [3,8,13,16,19,21,22,25,27,29]. However, different modern applications request minimization of generic non-smooth non-strongly convex stochastic objective functions given by regularized expected risk, with complicated constraints [5,20,29].

In the particular setting (2), when $X$ is nonempty, the RAP schemes are advantageous for large scale CFP, since at each iteration they compute a projection onto a simple set $X_\xi$. It is well-known, by using the Moreau smoothing for each indicator function component:

$$\text{dist}^2_{X_\xi}(x) := \min_z \mathbb{1}_{X_\xi}(z) + \frac{1}{2}\|z - x\|^2,$$

then applying SGD on the new surrogate problem

$$\min_{x \in \mathbb{R}^n} \mathbb{E}[\text{dist}^2_{X_\xi}(x)]$$

yields the alternating projections (RAP). Sublinear and linear convergence rates have been obtained under typical assumptions [6,15,17]. However, for general problems (1), the natural extension of RAP leads to stochastic schemes based on the Moreau smoothing envelope (see [24]). For this purpose, consider the Moreau envelope of each individual component $f(\cdot; \xi)$ and $\mathbb{1}_{X_\xi}$ which produces a modified smooth approximation of the original objective function $F$:

$$\min_{x \in \mathbb{R}^n} \mathbb{E}[f_\mu(x; \xi)] + \mathbb{E}\left[\frac{1}{2\mu}\text{dist}^2_{X_\xi}(x)\right] := \mathbb{E}\left[\min_z f(z; \xi) + \frac{1}{2\mu}\|z - x\|^2\right] + \mathbb{E}\left[\frac{1}{2\mu}\text{dist}^2_{X_\xi}(x)\right],$$

let smoothing parameter $\mu > 0$. Following a similar algorithmic reasoning as in the alternating projections and SGD, using stepsize sequence $\{\mu_k\}_{k \geq 0}$ the stochastic proximal point (SPP) algorithm is obtained by randomly choosing a sample $\xi_k$ and further computing one of the following iterations:

$$x^{k+1} = x^k - \mu_k \nabla f_\mu(x^k; \xi_k) \quad \text{or} \quad x^{k+1} = \pi_{X_{\xi_k}}(x^k).$$

Recently its convergence behaviour has been analyzed under various assumptions and several
advantages have been theoretically and empirically illustrated over the standard or modified SGD schemes \cite{3,20,26,28}. Moreover, the optimal convergence guarantees for SPP have been obtained under strong convexity and smoothness properties on the objective function $F$. In this paper we address these issues. The main contributions of this paper are:

(i) We offer a unified theoretical perspective over SPP and AP schemes, using a unified convergence rate analysis based on simple novel arguments. Up to our knowledge, this is the first unified analysis providing complexity results for stochastic proximal point and randomized alternating projections.

(ii) We provide sublinear/linear convergence rates for SPP scheme on constrained convex optimization under . The key structural assumption allowing this general result is the prox-quadratic growth property, which proves to be a natural generalization of classical linear regularity property of convex sets.

(iii) Our analysis applies to nonsmooth constrained optimization with complicated constraints (see Section 4), unlike to most of related papers on stochastic schemes. In our analysis SPP requires only simple projections onto individual sets, while most projected stochastic first order scheme require full projections onto the entire feasible set, which are computationally expensive.

(iv) The new proof techniques based on the prox-quadratic growth property are simpler than previous approaches. They allow us to show a sublinear $O\left(\frac{1}{k}\right)$ convergence rate for the stochastic proximal point algorithm in terms of the distance from the optimal set. Moreover, in the particular interpolation case when the functional components share minimizers, linear convergence is obtained.

1.1. Related work

Significant parts of the tremendous literature on stochastic optimization algorithms focused on the theoretical and practical behaviour of stochastic first order schemes under different convexity properties, see \cite{8,13,16,19,21,25,27}. Due to its simplicity, the traditional method of choice for most supervised machine learning problems is the SGD method. At each iteration $k$, the vanilla SGD algorithm randomly samples a functional component $\xi$ and takes a step along the gradient of the chosen individual function. Interesting results regarding SGD’s nonasymptotic theoretical complexity has been given in \cite{27}, where a sublinear $O\left(\frac{\log(k)}{k}\right)$ iteration complexity was provided for strongly convex objective functions with bounded gradients (in the SGD iterates). These assumptions matches the main $\ell_2$-regularized Support Vector Machine (SVM) application considered in \cite{27}, but are uncertainly satisfied by more general models. Further, in \cite{21} the clear $O\left(\frac{1}{k}\right)$ convergence rate of average SGD was established, in context of stochastic smooth (Lipschitz gradient continuity) strongly convex objective functions, while in the bounded gradients (of the iterates) case a simple modification in the averaging step of average SGD improved the previously known $O\left(\frac{\log(k)}{k}\right)$ rate to the better $O\left(\frac{1}{k}\right)$ estimate. Similar complexity results has been provided in \cite{3} in context of the stochastic and online, convex and strongly convex functions, but the authors further assumes computation of full projection on the feasible set and bounded gradients on the generated iterates. All these previously mentioned papers approached the classical choice of decreasing stepsize $\mu_k = \frac{1}{k}$ in the SGD algorithm. A recent extensive nonasymptotic analysis of the SGD scheme for more general decreasing stepsizes $\mu_k = \frac{\mu_0}{k^\gamma}, \gamma \in [0, 1]$, has been provided in \cite{13}, under various differentiability and convexity assumptions on the objective function. The theoretical estimates for smooth case obtained in \cite{13} highlights a mandatory limitation of the stepsize to small values through an exponential term which naturally appears in the convergence rate. Thus, general vanishing stepsize SGD was proved to converge as...
that the SPP scheme attain an O(\frac{\gamma^2}{\kappa} + \frac{C_\kappa}{\kappa}) in terms of average distance from the optimal point, under strong convexity and gradient Lipschitz assumptions on the objective function $F$. However in the complexity estimate corresponding to the strongly convex nonsmooth case, when the bounded gradients condition holds, the exponential term vanishes but in order to avoid a contradictory setting an additional limitation of the domain was considered. It can be easily seen, and also observed by authors of \cite{13,19,29}, that the strong convexity and bounded gradient assumptions are somehow contradicting on the unbounded domain. Thus, to avoid this situation, the authors of \cite{19} analyzed some (distributed) SGD variants under the combination of Lipschitz continuity and strong convexity properties. Also to avoid this situation, in \cite{30} the strong convexity assumption is relaxed to a local growth condition and a domain restriction is imposed.

In \cite{30} the constrained stochastic optimization problem $\min_{x \in X} F(x) := \mathbb{E}[f(x; \xi)]$ is analyzed and an accelerated SGD (ASSG) method has been devised, avoiding the typical strong convexity and gradient continuity assumptions. On short, ASSG represents a restarted projected variant of the vanilla SGD scheme involving an additional ball constraint to the subproblem: $x^{k+1} = \pi_{X \cap B(x^k, D)} \left[ x^k - \mu_k g(x^k; \xi_k) \right]$. It is worth to observe that the projection step impose to consider particularly simple feasible sets, otherwise the projection would involve an impractical computational burden which can be further augmented by the intersection with local ball. Note that when many constraints are involved, our analysis of the SPP scheme allows us to consider only projection onto a single set per iteration. The authors of \cite{30} suppose the objective function $F$ has $G-$bounded gradients and satisfies the functional $\theta-$local growth in the $\epsilon-$sublevel set. Under these circumstances the ASSG algorithm guarantees with high probability an impressive rate $O\left(\frac{1}{\kappa^4}\right)$. However, this convergence rate order requires bounded gradients condition, which is violated, for example, for smooth quadratically growing functions ($\theta = 1/2$), such as the linear regression cost $\|Ax - b\|^2$. We show in Section \ref{section:4} that SPP scheme attain $O\left(\frac{1}{\kappa}\right)$ rate on this type of functions.

The stochastic proximal point algorithm has been recently analyzed using various differentiability assumptions, see \cite{1,12,20,26,28,31}. In \cite{28} is considered the typical stochastic learning model involving the expectation of random particular components $f(x; \xi)$ defined by the composition of a smooth function and a linear operator, i.e.: $f(x; \xi) = \ell(a_\xi^T x)$, where $a_\xi \in \mathbb{R}^n$. The complexity analysis requires the linear composition form, i.e. $\ell(a_\xi^T x)$, and that the objective function $\mathbb{E}[\ell(a_\xi^T x)]$ to be smooth and strongly convex. The nonasymptotic convergence of the SPP with decreasing stepsize $\mu_k = \frac{\mu_0}{\kappa^\gamma}$, with $\gamma \in (1/2, 1]$, has been analyzed in the quadratic mean and an $O\left(\frac{1}{\kappa^\gamma}\right)$ convergence rate has been derived. The generalization of these convergence guarantees is undertaken in \cite{20}, where no linear composition structure is required and an (in)finite number of constraints are included in the stochastic model. However, the stochastic model from \cite{20} requires strong convexity and Lipschitz gradient continuity for each functional component $f(\cdot; \xi)$. Furthermore, it is explicitly specified that their analysis do not extend to certain models, such as those with nonsmooth functional components $\hat{f}(x; \xi) := f(x; \xi) + \mathbb{1}_{X_\epsilon}(x)$, where $f(\cdot; \xi)$ is smooth and convex. Note that our analysis surpasses these restrictions and provides a natural generalization of \cite{20} to nonsmooth constrained models.

In \cite{26}, the SPP scheme with decreasing stepsize $\mu_k = \frac{\mu_0}{\kappa^\gamma}$ has been applied to problems with the objective function having Lipschitz continuous gradient and the restricted strong convexity property, and its asymptotic global convergence is derived. A sublinear $O\left(\frac{1}{\kappa^\gamma}\right)$ asymptotic convergence rate in the quadratic mean has been given. In this paper we make more general assumptions on the objective function, which hold for restricted strongly convex functions, and provide nonasymptotic convergence analysis of the SPP for a more general stepsize $\mu_k = \frac{\mu_0}{\kappa^\gamma}$, with $\gamma > 0$. Further, in \cite{3} a general asymptotic convergence analysis of slightly modified
SPP scheme has been provided, under mild convexity assumptions on a finitely constrained stochastic problem. Although this scheme is very similar to the SPP algorithm, only the almost sure asymptotic convergence has been provided in \cite{3}.

Recently, in \cite{1}, the authors analyze SPP schemes for shared minimizers stochastic optimization and strongly convex functions. Remarkably, they eliminate any continuity assumption for the sublinear rate in the strongly convex case, which allows indicator functions. However, our analysis uses non-trivially the linear regularity of feasible set for obtaining better convergence constants.

Moreover, we use quadratic growth relaxations of strong convexity assumption which allow a unified treatment of SPP and AP schemes.

**Notations.** We use notation \([m] = \{1, \cdots, m\}\). For \(x, y \in \mathbb{R}^n\) denote the scalar product \(\langle x, y \rangle = x^T y\) and Euclidean norm by \(\|x\| = \sqrt{x^T x}\). The projection operator onto set \(X\) is denoted by \(\pi_X\) and the distance from \(x\) to the set \(X\) is denoted \(\text{dist}_X(x) = \min_{z \in X} \|x - z\|\).

The indicator function of a set \(X\) is denoted: \(\mathbb{I}_X(x) = \begin{cases} 0, & \text{if } x \in X \\ \infty, & \text{otherwise} \end{cases}\). For function \(f(\cdot; \xi)\), we use notations \(\partial f(x; \xi)\) the subdifferential set at \(x\) and \(g_f(x; \xi)\) for a subgradient of \(f\) at \(x\). If \(f(\cdot; \xi)\) is differentiable we use the gradient notation \(\nabla f(\cdot; \xi)\). Also we use \(g_X(x; \xi) \in \mathcal{N}_X(x)\) for a subgradient of \(\mathbb{I}_X(x)\). Finally, we define the function \(\varphi_\alpha : (0, \infty) \rightarrow \mathbb{R}\) as:

\[
\varphi_\alpha(x) = \begin{cases} (x^\alpha - 1)/\alpha, & \text{if } \alpha \neq 0 \\ \log(x), & \text{if } \alpha = 0. \end{cases}
\]

**1.2. Problem formulation**

In order to analyze in a unifying manner the stochastic projection and proximal point algorithms, we unify the two expectation terms from the composite model \cite{1}, i.e.

\[
\min_{x \in \mathbb{R}^n} \mathbb{E}[f(x; \xi)] + \mathbb{E}[\mathbb{I}_X(x)].
\]

into a single one, resulting the following simple reformulation of the composite model:

\[
F^* = \min_{x \in \mathbb{R}^n} F(x) \quad (: = \mathbb{E}[F(x; \xi)])
\]

where \(F(\cdot; \xi) : \mathbb{R}^n \rightarrow (-\infty, +\infty)\) are proper convex and lower-semicontinuous functions. The random variable \(\xi\) has its associated probability space \((\Omega, \mathbb{P})\), where \(\Omega = \Omega_1 \cup \Omega_2\). If \(\xi \in \Omega_2\), then \(f(x; \xi) = \mathbb{I}_X(x)\). In the sequel, some results require the composite formulation, thus we will return to it when it is necessary. We denote the set of optimal solutions with \(X^*\) and \(x^*\) any optimal point for \((3)\).

**Assumption 1.1.** The central problem \cite{3} satisfies:

(i) The optimal set \(X^*\) is nonempty.

(ii) There exists subgradient mapping \(g_F : \mathbb{R}^n \times \Omega \rightarrow \mathbb{R}^n\) such that \(g_F(x; \xi) \in \partial F(x; \xi)\) and \(\mathbb{E}[g_F(x; \xi)] \in \partial F(x)\).

(iii) \(F(\cdot; \xi)\) has bounded gradients on the optimal set: there exists \(S_F^* \geq 0\) such that \(\mathbb{E} \left[\|g_F(x^*; \xi)\|^2\right] \leq S_F^* < \infty\) for all \(x^* \in X^*\).

The first part of the above assumption is natural in the stochastic optimization problems. The Assumption \cite{1} ii) guarantee the existence of a subgradient mapping. Moreover, since \(0 \in
\[ \partial F(x^*) \text{ for any } x^* \in X^*, \text{ then we assume in this paper that } g(x^*) := \mathbb{E}[g_F(x^*; \xi)] = 0. \] We also denote \( g_f(x; \xi) \in \partial f(x; \xi) \) a subgradient of \( f(x; \xi) \) at \( x \). Also the third part Assumption \( (iii) \) is standard in the literature related to stochastic algorithms.

As we motivated in the introduction, we approximate the functional components \( F(\cdot; \xi) \) through their Moreau envelope, that is:

\[ F_\mu(x; \xi) := \min_{z \in \mathbb{R}^n} F(z; \xi) + \frac{1}{2\mu} \| z - x \|^2 \]

for some smoothing parameter \( \mu > 0 \). The approximate \( F_\mu(\cdot; \xi) \) keeps the convexity properties of \( F(\cdot; \xi) \) and additionally has Lipschitz continuous gradient with constant \( \frac{1}{\mu} \), see \[24\]. By this smoothing we obtain at a new stochastic smooth optimization problem:

\[ F^*_\mu = \min_{x \in \mathbb{R}^n} F_\mu(x) := \mathbb{E}[F_\mu(x; \xi)] \] (4)

in a tight connection with the original one. Features of this connection are presented in Section 3 For this resulting problem we denote \( X^*_\mu = \arg \min_x F_\mu(x) \), which in general \( X^*_\mu \neq X^* \). As we have pointed in the Introduction, this approach arises naturally in the CFPs where for finite \( \Omega: F(x; \xi) = \mathbb{I}_{X_\xi}(x) \) and \( X^* = \bigcap_{\xi \in \Omega} X_\xi \). In this particular case the smooth approximation becomes \( F_\mu(x; \xi) = \text{dist}_{X_\xi}^2(x) \) and the objective of (4) becomes:

\[ F_\mu(x) = \frac{1}{2\mu} \mathbb{E}[\text{dist}_{X_\xi}^2(x)]. \] (5)

When \( X^* \neq \emptyset \) then \( X^*_\mu = X^* \) for all \( \mu > 0 \) which yields that, in this particular setting, the problem (4) is equivalent with the original nonsmooth problem (3).

### 2. Stochastic Proximal Point algorithm

In the following section we propose a stochastic iterative scheme for solving the problem (4) and analyze its convergence behaviour towards the optimal set of the original problem (3). For this purpose, we denote the prox operator corresponding to \( f(\cdot; \xi) \) with:

\[ z_\mu(x; \xi) = \arg \min_{z \in \mathbb{R}^n} F(z; \xi) + \frac{1}{2\mu} \| z - x \|^2. \]

In particular, when \( F(x; \xi) = \mathbb{I}_{X_\xi}(x) \) the prox operator becomes the projection operator \( z_\mu(x; \xi) = \pi_{X_\xi}(x) \). Indeed, given a fixed \( \mu > 0 \), by applying the pure constant stepsize SGD to solve (4) it is easy to observe that:

\[ x^{k+1} = x^k - \mu \nabla F_\mu(x^k; \xi_k) = z_\mu(x^k; \xi_k). \]

Following the equivalence of the prox and projection operators on feasibility problems, the above algorithm might be interpreted as a natural generalization of RAP (see e.g. [4]). Since in general \( X^*_\mu \neq X^* \), the smoothing parameter should be decreased in order to guarantee convergence towards the minimizer of the original problem.

Let \( x^0 \in \mathbb{R}^n \) be a starting point and \( \{\mu_k\}_{k \geq 0} \) be a nonincreasing positive sequence of stepsizes.


Stochastic Proximal Point (SPP) \((x_0, \{\mu_k\}_{k \geq 0})\): For \(k \geq 1\) compute
1. Choose randomly \(\xi_k \in \Omega\) w.r.t. probability distribution \(P\)
2. Update: \(x^{k+1} = z_{\mu_k}(x^k; \xi_k)\).

Note that there are many practical cases when the prox operator \(z_{\mu}\) can be computed easily or even has a closed form. To exemplify a few:

(i) the least-square loss \(F(x; \xi) = \frac{1}{2}(a^T\xi x - b\xi)^2\), \(z_{\mu}(x; \xi) = x - \frac{\mu(a^T\xi - b\xi)}{1+\mu\|a\|^2}a\xi;\)

(ii) regularized hinge-loss \(F(x; \xi) = \max\{0, a^T\xi x - b\xi\} + \frac{\mu}{2}\|x\|^2\), \(z_{\mu}(x; \xi) = \frac{1}{1+\lambda\mu}(x - \mu a\xi s)\), where \(s = \pi_{[0,1]}\left(\frac{1-\lambda\mu}{\mu(1+\lambda\mu)}a^T\xi - \frac{1+\lambda\mu}{\mu\|a\|^2}b\right)\).

(iii) halfspace: \(H_\xi = \{x : a^T\xi x \leq b\xi, F(x; \xi) = I_{H_\xi}(x), z_{\mu}(x; \xi) = x - \max\{0,a^T\xi - b\xi\}\}

\frac{\mu a\xi}{{\|a\|^2}}a\xi.

2.1. Preliminary results

We derive first some simple auxiliary results, we will be intensively used in the sequel. By returning to the CFP framework \([5]\), we define the linear regularity property.

Definition 2.1. Let \(\{X_\xi\}_{\xi \in \Omega}\) be convex sets with nonempty intersection \(X = \bigcap_{\xi \in \Omega} X_\xi\). They are linearly regular with constant \(\kappa > 0\) if:

\[
\kappa \text{dist}_X^2(x) \leq \mathbb{E}\left[\text{dist}_{X_\xi}^2(x)\right] \quad \forall x \in \mathbb{R}^n.
\]

Further we present bounds on the gap between the smooth approximations and the optimal values of the objective function.

Lemma 2.2. Given \(\mu > 0\), let \(\{X_\xi\}_{\xi \in \Omega}\) be some convex sets satisfying linear regularity with constant \(\kappa_x > 0\) and recall \(F(x) = f(x) + \mathbb{E}\left[I_{X_\xi}(x)\right]\). Then the following relations hold:

(i) \(F_\mu(x) \leq F(x) \quad \forall x \in \mathbb{R}^n\),

(ii) \(F^* - F_\mu(x) \leq \frac{\mu}{\kappa} \mathbb{E}\left[\|g_F(x^*; \xi)\|^2\right] \leq \frac{\mu}{\kappa} S_F^2 \quad \forall x \in \mathbb{R}^n\).

(iii) \(F^* - F_\mu(x) \leq \mathbb{E}\left[\frac{\mu}{\kappa} \|g_f(x^*; \xi)\|^2\right] + \frac{\mu}{2\kappa} \|g_f(x^*)\|^2 \quad \forall x \in \mathbb{R}^n\).

Proof. It is straightforward that

\[
F_\mu(x; \xi) = \min_{z \in \mathbb{R}^n} F(z; \xi) + \frac{1}{2\mu}\|z - x\|^2 \leq F(x; \xi) \quad \forall x \in \mathbb{R}^n.
\]

By taking expectation w.r.t. \(\xi\) in both sides we get (i). In order to prove (ii), let \(z \in \mathbb{R}^n\). Then,
given \( x^* \in X^* \) and \( g_F(x^*; \xi) \in \partial F(x^*; \xi) \), by convexity of \( f(\cdot; \xi) \) we have:

\[
F^* - F_\mu(x) = \mathbb{E} \left[ F(x^*; \xi) - F(z_\mu(x; \xi); \xi) - \frac{1}{2\mu} \| z_\mu(x; \xi) - x \|^2 \right] \\
\leq \mathbb{E} \left[ \langle g_F(x^*; \xi), x^* - z_\mu(x; \xi) \rangle - \frac{1}{2\mu} \| z_\mu(x; \xi) - x \|^2 \right] \\
\leq \mathbb{E} \left[ \langle g_F(x^*; \xi), x^* - x \rangle + \max_{z} \langle g_F(x^*; \xi), x - z \rangle - \frac{1}{2\mu} \| z - x \|^2 \right] \\
\leq \mathbb{E} \left[ \langle g_F(x^*; \xi), x^* - x \rangle + \mathbb{E} \left[ \frac{\mu}{2} \| g_F(x^*; \xi) \|^2 \right] \right] \quad \forall x^* \in X^*,
\]

where we recall that we consider \( \mathbb{E} [g_F(x^*; \xi)] = 0 \). Therefore, we finally obtain

\[
F^* - F_\mu(x) \leq \frac{\mu}{2} \mathbb{E} \left[ \| g_F(x^*; \xi) \|^2 \right] \leq S^*_F,
\]

which confirms result (ii). For the third part (iii), denote \( D_\mu(x) := \mathbb{E} \left[ \frac{1}{2\mu} \text{dist}^2_{X_\xi}(x; \xi) \right] \) and \( g_f(x^*) := \mathbb{E} [g_f(x^*; \xi)] \). Then we derive that:

\[
F_\mu(x) - F(x^*) = f_\mu(x) - f(x^*) + D_\mu(x) \\
\geq \mathbb{E} \left[ \langle g_f(x^*; \xi), z_\mu(x; \xi) - x^* \rangle + \frac{1}{2\mu} \| z_\mu(x; \xi) - x \|^2 \right] + D_\mu(x) \\
\geq \mathbb{E} \left[ \langle g_f(x^*; \xi), z_\mu(x; \xi) - x \rangle + \frac{1}{2\mu} \| z_\mu(x; \xi) - x \|^2 \right] + \mathbb{E} [\langle g_f(x^*; \xi), x - x^* \rangle] + D_\mu(x) \\
\geq -\mathbb{E} \left[ \frac{\mu}{2} \| g_f(x^*; \xi) \|^2 \right] + \langle g_f(x^*), x - x^* \rangle + D_\mu(x) \\
\geq -\mathbb{E} \left[ \frac{\mu}{2} \| g_f(x^*; \xi) \|^2 \right] + \langle g_f(x^*), \pi_X(x) - x^* \rangle + \langle g_f(x^*), x - \pi_X(x) \rangle + D_\mu(x) \\
\geq -\mathbb{E} \left[ \frac{\mu}{2} \| g_f(x^*; \xi) \|^2 \right] + \langle g_f(x^*), x - \pi_X(x) \rangle + D_\mu(x) \\
\geq -\mathbb{E} \left[ \frac{\mu}{2} \| g_f(x^*; \xi) \|^2 \right] - \| g_f(x^*) \| \text{dist}_X(x) + \frac{\kappa x}{2\mu} \text{dist}^2_{X_\xi}(x) \\
\geq -\mathbb{E} \left[ \frac{\mu}{2} \| g_f(x^*; \xi) \|^2 \right] - \frac{\mu}{2\kappa} \| g_f(x^*) \|^2, \tag{8}
\]

where in the fifth inequality we used the optimality conditions: \( \langle g_f(x^*), z - x^* \rangle \geq 0 \) for all \( z \in X \).

A simple key inequality for the convergence rate results is given in the following lemma.

**Lemma 2.3.** For any \( x \in \mathbb{R}^n, \mu > 0, \xi \in \Omega \), the following relation holds:

\[
\mathbb{E} \left[ \| z_\mu(x; \xi) - z \|^2 \right] \leq \| x - z \|^2 + 2\mu \left( F(z) - F_\mu(x) \right).
\]

**Proof.** Note that \( F(\cdot; \xi) + \frac{1}{2\mu} \| \cdot - x \|^2 \) is strongly convex with constant \( \frac{1}{\mu} \), which further
yields:
\[
F(z; \xi) + \frac{1}{2\mu} \|z - x\|^2 \geq F(z_{\mu}(x; \xi); \xi) + \frac{1}{2\mu} \|z_{\mu}(x; \xi) - x\|^2 + \frac{1}{2\mu} \|z_{\mu}(x; \xi) - z\|^2 \\
= F_{\mu}(x; \xi) + \frac{1}{2\mu} \|z_{\mu}(x; \xi) - z\|^2, \quad \forall z \in \mathbb{R}^n.
\]

By taking expectation in both sides, the last relation leads to the above result.

3. Iteration complexity of SPP under prox-quadratic growth

In this section we derive sublinear and linear convergence rates of stochastic proximal point scheme under various convexity and regularity conditions of the objective function.

Further we derive convergence rate of stochastic proximal point scheme under a general regularity assumption similar with the functional quadratic growth [5].

Assumption 3.1. The objective function \( F \) satisfies prox-quadratic growth property if there exists positive constants \( \sigma_{F,\mu} \) and \( \beta \) such that for any \( \mu > 0 \):

\[
\frac{\sigma_{F,\mu}}{2} \text{dist}_{X^*}^2(x) \leq F_{\mu}(x) - F^* + \mu \beta \quad \forall x \in \mathbb{R}^n.
\]

Moreover, the mapping \( \mu \mapsto \sigma_{F,\mu} \) is nonincreasing in \( \mu \).

Assumption 3.1 can be interpreted as a generalized quadratic growth since for \( \mu = 0 \) reduces to the well-known pure quadratic growth property for the objective function \( F \), which has been extensively analyzed in the deterministic setting, see for example [4,14,29,30]. Although in many practical applications the strong convexity does not hold, first-order algorithms exhibit linear convergence under pure quadratic growth and certain additional smoothness conditions [14]. However, for general stochastic first order algorithms the geometric convergence feature cannot be attained, due to the variance of the chosen stochastic direction. In [29,30], sublinear convergence of the restarted SGD has been shown under the quadratic growth property and bounded gradients. A simple look at the particular instance of CFP case \( f = 0 \) will provide more intuition about the Assumption 3.1. Since the right hand side of (6) is equivalent with \( F_1(x) \) then, in general, the linear regularity is also recovered by particular prox-quadratic growth with constants \( \sigma_{F,\mu} = \frac{\sigma}{\mu} \) and \( \beta = 0 \). In Section 4 we properly analyze several well-known classes of functions and prove that they satisfy the Assumption 3.1.

The recurrence which establishes the SPP iteration convergence is given by the following lemma.

Theorem 3.2. Let Assumptions 1.1 and 3.1 hold. The sequence \( \{x^k\} \) generated by SPP satisfies:

\[
E \left[ \text{dist}_{X^*}^2(x^{k+1}) \right] \leq \prod_{i=0}^{k} (1 - \mu_i \sigma_{F,\mu_0}) \text{dist}_{X^*}^2(x^0) + (S^* + 2\beta) \sum_{i=0}^{k} \prod_{j=i+1}^{k} (1 - \mu_j \sigma_{F,\mu_0}) \mu_i^2.
\]

Proof. By taking \( z = \pi_{X^*}(x^k), \mu = \mu_k, x = x^k \) in Lemma 2.3 then for any \( x^* \in \mathbb{X}^* \) we
obtain:

\[
\mathbb{E} \left[ \| x^{k+1} - \pi_{X^*}(x^k) \|^2 \right] \leq \| x^k - \pi_{X^*}(x^k) \|^2 + 2\mu_k \left( F(\pi_{X^*}(x^k)) - F_{\mu_k}(x^k) \right)
\]

\[
= \text{dist}_{X^*}^2(x^k) + 2\mu_k \left( F^* - F_{\mu_k}(x^k) \right) + 2\mu_k \left( F_{\mu_k}^* - F_{\mu_k}(x^k) \right)
\]

Lemma 2.2 \(\ast\) \(\ast\) \(\ast\) \(\ast\) \(\ast\) \(\ast\) \(\ast\) \(\ast\) \(\ast\) \(\ast\) \(\ast\)

\[
\leq \text{dist}_{X^*}^2(x^k) + \mu_k^2 \mathbb{E} \left[ \| g_F(x^*; \xi) \|^2 \right] + 2\mu_k \left( F_{\mu_k}^* - F_{\mu_k}(x^k) \right)
\]

Assumption 3.1 \(\ast\) \(\ast\) \(\ast\) \(\ast\) \(\ast\) \(\ast\) \(\ast\) \(\ast\) \(\ast\) \(\ast\) \(\ast\)

\[
\leq (1 - \mu_k \sigma_{F, \mu_k}) \text{dist}_{X^*}^2(x^k) + \mu_k^2 \left( \mathbb{E} \left[ \| g_F(x^*; \xi) \|^2 \right] + 2\beta \right).
\]

By observing that \( \| x^{k+1} - \pi_{X^*}(x^k) \| \geq \text{dist}^2_{X^*}(x^{k+1}) \) then by taking full expectation in both sides of (10) we get:

\[
\mathbb{E} \left[ \text{dist}_{X^*}^2(x^{k+1}) \right] \leq (1 - \mu_k \sigma_{F, \mu_k}) \mathbb{E} \left[ \text{dist}_{X^*}^2(x^k) \right] + \mu_k^2 \left( \mathbb{E} \left[ \| g_F(x^*; \xi) \|^2 \right] + 2\beta \right)
\]

\[
\leq (1 - \mu_k \sigma_{F, \mu_0}) \mathbb{E} \left[ \text{dist}_{X^*}^2(x^k) \right] + \mu_k^2 \left( \mathbb{E} \left[ \| g_F(x^*; \xi) \|^2 \right] + 2\beta \right),
\]

where we used that \( \sigma_{F, \mu_k} \geq \sigma_{F, \mu_0} \), since \( \sigma_{F, \mu} \) is nonincreasing in \( \mu \). Now for simplicity if we denote \( \theta_k = (1 - \mu_k \sigma_{F, \mu_0}) \), then we can further derive:

\[
\mathbb{E} \left[ \text{dist}_{X^*}^2(x^{k+1}) \right] \leq \theta_k \mathbb{E} \left[ \text{dist}_{X^*}^2(x^k) \right] + \mu_k^2 (S^*_F + 2\beta)
\]

\[
\leq \left( \prod_{i=0}^k \theta_i \right) \text{dist}_{X^*}^2(x^0) + (S^*_F + 2\beta) \sum_{i=0}^k \left( \prod_{j=i+1}^k \theta_j \right) \mu_i^2,
\]

which confirms our result. \(\square\)

Remark 1. The universal upper bound provided in Theorem 3.2 will be used to generate sublinear rate for non-interpolation context, i.e. \( S^*_F + \beta > 0 \), and linear convergence rates for constant stepsize SPP under interpolation assumption, i.e. \( S^*_F = \beta = 0 \).

3.1. Sublinear convergence rate

The sublinear convergence rate for SPP under the prox-quadratic growth property can be easily obtained from Theorem 3.2

Theorem 3.3. Let Assumptions 1.1, 3.1 hold. Also let the decreasing stepsize sequence \( \mu_k = \frac{\mu_0}{k} \) and \( \{x^k\}_{k \geq 0} \) be the sequence generated by SPP(\( x^0 \), \( \{\mu_k\}_{k \geq 0} \)). Then, for any \( k \geq 0 \), the following relation holds:

(i) If \( \gamma \in (0, 1) : \quad \mathbb{E} [\| x^k - x^* \| ^2] \leq O \left( \frac{1}{k^\gamma} \right) \)

(ii) If \( \gamma = 1 : \quad \mathbb{E} [\| x^k - x^* \| ^2] \leq \begin{cases} O \left( \frac{1}{k^\gamma} \right) & \text{if } \mu_0 \sigma_{F, \mu_0} > e - 1 \\ O \left( \frac{\ln k}{k} \right) & \text{if } \mu_0 \sigma_{F, \mu_0} > e - 1 \\ O \left( \frac{1}{k^2 \ln(1 + \mu_0 \sigma_{F, \mu_0})} \right) & \text{if } \mu_0 \sigma_{F, \mu_0} < e - 1. \end{cases} \)

10
**Proof.** For simplicity denote \( \theta_k = (1 - \mu_k \sigma_{F,\mu_0}) \), then Theorem 3.2 implies that:

\[
\mathbb{E} \left[ \text{dist}_{X^*}^2(x_{k+1}) \right] \leq \left( \prod_{i=0}^{k} \theta_i \right) \text{dist}_{X^*}^2(x^0) + S_F^* \sum_{i=0}^{k} \left( \prod_{j=i+1}^{k} \theta_j \right) \mu_i^2.
\]

By using the Bernoulli inequality \( 1 - tx \leq \frac{1}{1+tx} \leq (1 + x)^{-t} \) for \( t \in [0, 1] \), \( x \geq 0 \), then we have:

\[
\prod_{i=l}^{u} \theta_i = \prod_{i=l}^{u} \left( 1 - \frac{\mu_0}{\sigma_{F,\mu_0}} \right) \leq \prod_{i=l}^{u} (1 + \mu_0 \sigma_{F,\mu_0})^{-1/\gamma} = (1 + \mu_0 \sigma_{F,\mu_0})^{-\sum_{i=l}^{u} \frac{1}{\gamma}}.
\]

On the other hand, if we use the lower bound

\[
\sum_{i=l}^{u} \frac{1}{\gamma} \geq \int_{l}^{u+1} \frac{1}{\gamma} d\tau = \varphi_{1-\gamma}(u + 1) - \varphi_{1-\gamma}(l).
\]

then we can finally derive:

\[
\sum_{i=0}^{k} \left( \prod_{j=i+1}^{k} \theta_j \right) \mu_i^2 = \sum_{i=0}^{m} \left( \prod_{j=i+1}^{k} \theta_j \right) \mu_i^2 + \sum_{i=m+1}^{k} \left( \prod_{j=i+1}^{k} \theta_j \right) \mu_i^2
\]

\[
\leq (1 + \mu_0 \sigma_{F,\mu_0}) \varphi_{1-\gamma}(m) \sum_{i=0}^{m} \mu_i^2 + \frac{\mu_{m+1}}{\sigma_{F,\mu_0}} \sum_{i=m+1}^{k} \left[ \prod_{j=i+1}^{k} (1 - \mu_j \sigma_{F,\mu_0}) \right] (1 - (1 - \varphi_{1-\gamma}(k)))
\]

\[
= (1 + \mu_0 \sigma_{F,\mu_0}) \varphi_{1-\gamma}(m) \sum_{i=0}^{m} \mu_i^2 + \frac{\mu_{m+1}}{\sigma_{F,\mu_0}} \sum_{i=m+1}^{k} \left[ \prod_{j=i+1}^{k} (1 - \mu_j \sigma_{F,\mu_0}) \right] (1 - (1 - \varphi_{1-\gamma}(k)))
\]

\[
\leq (1 + \mu_0 \sigma_{F,\mu_0}) \varphi_{1-\gamma}(m) \sum_{i=0}^{m} \mu_i^2 + \frac{\mu_{m+1}}{\sigma_{F,\mu_0}} \sum_{i=m+1}^{k} \left[ \prod_{j=i+1}^{k} (1 - \mu_j \sigma_{F,\mu_0}) \right] (1 - (1 - \varphi_{1-\gamma}(k)))
\]

By denoting the second constant \( \tilde{\theta}_0 = \frac{1}{1+\mu_0 \sigma_{F,\mu_0}} \), then the last relation implies the following bound:

\[
\mathbb{E} \left[ \text{dist}_{X^*}^2(x_{k+1}) \right] \leq \tilde{\theta}_0 \varphi_{1-\gamma}(k) \text{dist}_{X^*}^2(x^0) + \tilde{\theta}_0 \varphi_{1-\gamma}(k) \varphi_{1-2\gamma}(m) S_F^* + \frac{\mu_{m+1}}{\sigma_{F,\mu_0}} S_F^*.
\]

To derive an explicit convergence rate order we analyze upper bounds on function \( \phi \).
(i) First assume that $\gamma \in (0, \frac{1}{2})$. This implies that $1 - 2\gamma > 0$ and that:

$$\varphi_{1-2\gamma}\left(\frac{k}{2}\right) \leq \varphi_{1-2\gamma}\left(\frac{k}{2}\right) = \frac{(k/2)^{1-2\gamma} - 1}{1 - 2\gamma} \leq \frac{(k/2)^{1-2\gamma}}{1 - 2\gamma}.$$  \hspace{1cm} (13)

On the other hand, by using the inequality $e^{-x} \leq \frac{1}{1+x}$ for all $x \geq 0$, we obtain:

$$\theta_0^{\varphi_{1-\gamma}(k)} - \varphi_{1-\gamma}(k) \leq e^{(\varphi_{1-\gamma}(k) - \varphi_{1-\gamma}(k/2))} \ln \theta_0 \varphi_{1-2\gamma}\left(\frac{k}{2}\right) \leq \frac{\varphi_{1-2\gamma}\left(\frac{k}{2}\right)}{1 + [\varphi_{1-\gamma}(k) - \varphi_{1-\gamma}(k/2 - 1)] \ln \frac{1}{\theta_0} \varphi_{1-2\gamma}\left(\frac{k}{2}\right)} \leq \frac{\theta_0^{k^{1-2\gamma}}}{2^{1-2\gamma}(1-2\gamma)}$$

$$= \frac{k^{1-2\gamma}}{1-\gamma} \left[1 - \left(\frac{1}{2}\right)^{1-\gamma}\right] \ln \frac{1}{\theta_0} \varphi_{1-2\gamma}\left(\frac{k}{2}\right) = 1 - \gamma \frac{2^{1-2\gamma} \theta_0^{k^{1-2\gamma}}}{2^{1-2\gamma}(1-2\gamma)} \ln \frac{1}{\theta_0} \varphi_{1-2\gamma}\left(\frac{k}{2}\right) = O\left(\frac{1}{k^{\gamma}}\right).$$

Therefore, in this case, the overall rate will be given by:

$$r_{k+1}^2 \leq \theta_0^{O(k^{1-\gamma})} r_0^2 + O\left(\frac{1}{k^{\gamma}}\right) \approx O\left(\frac{1}{k^{\gamma}}\right).$$

If $\gamma = \frac{1}{2}$, then the definition of $\varphi_{1-2\gamma}(k/2)$ provides that:

$$r_{k+1}^2 \leq \theta_0^{O(\sqrt{k})} r_0^2 + \theta_0^{O(\sqrt{k})} O(\ln k) + O\left(\frac{1}{\sqrt{k}}\right) \approx O\left(\frac{1}{\sqrt{k}}\right).$$

When $\gamma \in (\frac{1}{2}, 1)$, it is obvious that $\varphi_{1-2\gamma}(k/2) \leq \frac{1}{2\gamma-1}$ and therefore the order of the convergence rate changes into:

$$r_{k+1}^2 \leq \theta_0^{O(k^{1-\gamma})} r_0^2 + O(1) + O\left(\frac{1}{k^{\gamma}}\right) \approx O\left(\frac{1}{k^{\gamma}}\right).$$

(ii) Lastly, if $\gamma = 1$, by using $\theta_0^{\ln k+1} \leq (\frac{1}{k})^{\ln \frac{1}{\theta_0}}$, we obtain the second part of our result. □

A similar convergence rate result can be found under the Lipschitz gradient and strong convexity assumptions. However, our analysis is much simpler and requires only prox-quadratic growth, which holds even for some particular nonsmooth non-strongly convex objective functions.

### 3.2. Linear convergence rate

In this section we show that we can improve further the convergence rate of SPP scheme under an additional stronger assumption related to the interpolation setting.

**Assumption 3.4.** The functional components $F(\cdot; \xi)$ share common minimizers, i.e. for any $x^* \in X^*$

$$0 \in \partial F(x^*; \xi) \quad \forall \xi \in \Omega.$$
The interpolation condition is typical for CFPs, where is aimed to find a common point of a collection of convex sets, i.e. \( f(\cdot; \xi) = \mathbb{I}_{X_\xi}(\cdot) \) and \( X^* = \bigcap_{\xi \in \Omega} X_\xi \). For example in [10], for the interpolation least-squares problem the linear rate behaviour of SGD has been extensively analyzed. Notice that an immediate consequence of Assumption 3.4 is that given any optimal \( x^* \) we can find subgradients \( g_F(x^*; \xi) \) for each \( \xi \) such that \( \mathbb{E} \left[ \|g_F(x^*; \xi)\|^2 \right] = S_F^* = 0, \forall x^* \in X^* \). Further by taking into account that the Moreau envelope preserves the set of minimizers corresponding to each functional component, then we have

\[
X^* = X^*_\mu \quad \text{and} \quad \mathbb{E} \left[ \|\nabla f_{\mu}(x^*; \xi)\|^2 \right] = 0 \quad \forall x^* \in X^*, \mu > 0.
\]

This fact implies that the decaying stepsize of the SPP iteration is not necessary any more. A straightforward application of Theorem 3.2 leads to the following constant decrease:

**Corollary 3.5.** Let Assumption 3.1 hold with \( \beta = 0 \). If also the Assumption 3.4 holds, then Theorem 3.2 implies that the sequence \( \{x^k\}_{k \geq 0} \) generated by constant stepsize SPP satisfies:

\[
\mathbb{E} \left[ \text{dist}_{X^*}(x^k) \right] \leq (1 - \mu \sigma_{F,\mu})^k \mathbb{E} \left[ \text{dist}_{X^*}(x^0) \right].
\]

As proved in Section 4, the indicator functions w.r.t. linearly regular sets, the restricted strongly convex functions and some particularly structured quadratically growing functions satisfy the Assumption 3.1 with \( \beta = \mathcal{O} \left( \mathbb{E} \left[ \|g_F(x^*; \xi)\|^2 \right] \right) \), which is possibly vanishing in the interpolation context when the Assumption 3.4 holds. It seems that using other analysis from [20,28] cannot be guaranteed that SPP converges linearly in the interpolation settings. The work [1] is centered on behaviour of vanishing stepsize SPP for interpolation ("easy") problems and obtain impressive complexity and stability results. However, we focus on quadratic growth relaxations of strong convexity assumption which allow the a unified treatment of interpolation and non-interpolation contexts.

Let us consider CFP case \( f = 0 \). In this case the SPP iteration becomes the vanilla RAP: given \( x^0 \in \mathbb{R}^n \), then for \( k \geq 0 \)

1. Choose randomly set \( \xi_k \)
2. Compute: \( x^{k+1} = \pi_{X_{\xi_k}}(x^k) \).

As we have shown in this case

\[
\mu \sigma_{F,\mu} = \kappa,
\]

which proves that Corollary 3.5 recovers the widely known linear convergence rate (see [15, 17]):

\[
\mathbb{E} \left[ \text{dist}_{X^*}(x^k) \right] \leq (1 - \kappa)^k \mathbb{E} \left[ \text{dist}_{X^*}(x^0) \right].
\]

**4. Function classes satisfying prox-quadratic growth**

Further we will enumerate some classes of functions which often proved empirical utility in the optimization and machine learning literature, and then show for each class that satisfies
the prox-quadratic growth property. For brevity, we recall our original constrained model (1):

\[
\min f(x) = \mathbb{E}[f(x; \xi)] \\
\text{s.t. } x \in X = \bigcap_{\zeta \in \Omega_2} X_\zeta.
\]

As in the previous sections, we will refer to \( F(x) = \mathbb{E}[F(x; \xi)] = f(x) + \mathbb{E}[\mathbb{I}_{X_\zeta}(x)]. \)

### 4.1. Quadratically growing functions with linearly regular constraints

A well known strong convexity relaxation which is often used in the linear convergence analysis of the deterministic first-order methods is the quadratic growth property. In this section we prove that usual quadratic growth together with a smoothness assumption further implies the prox-quadratic growth.

**Definition 4.1.** The function \( f \) satisfies quadratic growth with constant \( \sigma_f \) if the following relation holds:

\[
f(x) - f^* \geq \frac{\sigma_f}{2} \text{dist}^2_X(x) \quad \forall x \in X.
\]

From convexity of function \( f \) we have

\[
\langle g_f(x), x - \pi_X^*(x) \rangle \geq f(x) - f^* \geq \frac{\sigma_f}{2} \text{dist}^2_X(x),
\]

which by Cauchy-Schwartz inequality implies:

\[
\|g_f(x)\| \geq \frac{\sigma_f}{2} \text{dist}^2_X(x) \quad \forall x \in X. \tag{15}
\]

**Assumption 4.2.** Each function \( f(\cdot; \xi) \) has Lipschitz continuous gradients with constant \( L_\xi \), i.e. there exists \( L_\xi \) such that the following relation holds:

\[
\|g_f(x; \xi) - g_f(y; \xi)\| \leq L_\xi \|x - y\|, \quad \forall x, y \in \mathbb{R}^n.
\]

It is easy to see that under Lipschitz continuity, there exist a relation between norms of gradients \( g_f \) and \( \nabla f_\mu \).

**Lemma 4.3.** Let Assumption 4.2 then the following relation holds:

\[
\mathbb{E} \left[ \frac{\|g_f(x; \xi)\|^2}{(1 + L_\xi \mu)^2} \right] \leq \mathbb{E} \left[ \|\nabla f_\mu(x)\|^2 \right] \quad \forall x \in \mathbb{R}^n.
\]

**Proof.** Based on the Lipschitz gradient assumption and the triangle inequality we have:

\[
\|g_f(x; \xi)\| \leq \|g_f(x; \xi) - g_f(z_\mu(x; \xi); \xi)\| + \|g_f(z_\mu(x; \xi); \xi)\| \leq L_\xi \|x - z_\mu(x; \xi)\| + \|g_f(z_\mu(x; \xi); \xi)\| = (1 + L_\xi \mu) \|\nabla f_\mu(x; \xi)\|.
\]

By taking expectation in both sides we obtain the result. \( \square \)
Further we derive the prox-quadratic growth relation for the extended function $F$.

**Lemma 4.4.** Let Assumption [4.2] hold and suppose that $f$ satisfies quadratic growth with constant $\sigma_f$. Also assume that the constraints sets $\{X_\xi\}_{\xi \in \Omega}$ are linearly regular with constant $\kappa$. Then, the composite objective $F$ satisfies prox-quadratic growth (Assumption [8.1]) with constants:

$$
\sigma_{F, \mu} = \frac{\sigma_f \mu \kappa}{\sigma_f \mu^2 + 8(1 + 2\kappa)(1 + \mu L_{\max})^2}
$$

$$
\beta = \mathbb{E} \left[ \|g_f(x^*; \xi)\|^2 \right] + \left[ \frac{1}{\kappa} + \frac{1}{4\kappa^3} \left(1 + \frac{\sigma_f}{4L_{\max}^2}\right)^2 \right] \|g_f(x^*)\|^2,
$$

where $L_{\max} = \max_{\xi \in \Omega} L_\xi$.

**Proof.** We make two central observations. First, using the linear regularity of the feasible set, it can be easily seen that:

$$
F_\mu(x) = f_\mu(x) + \mathbb{E} \left[ \frac{1}{2\mu} \text{dist}_X^2(x) \right]
$$

$$
\geq \mathbb{E} \left[ f(x^*; \xi) + \langle g_f(x^*; \xi), z_\mu(x; \xi) - x^* \rangle + \frac{1}{2\mu} \|z_\mu(x; \xi) - x\|^2 \right] + \frac{\kappa}{2\mu} \text{dist}_X^2(x)
$$

$$
= f^* + \mathbb{E} \left[ \langle g_f(x^*; \xi), z_\mu(x; \xi) - x \rangle + \frac{1}{2\mu} \|z_\mu(x; \xi) - x\|^2 \right]
$$

$$
\quad + \langle g_f(x^*), \pi_X(x) - x^* \rangle + \langle g_f(x^*), x - \pi_X(x) \rangle + \frac{\kappa}{2\mu} \text{dist}_X^2(x)
$$

$$
\geq f^* - \mu \mathbb{E} \left[ \|g_f(x^*; \xi)\|^2 \right] + \frac{1}{4\mu} \mathbb{E} \left[ \|z_\mu(x; \xi) - x\|^2 \right]
$$

$$
\quad - \|g_f(x^*)\| \text{dist}_X(x) + \frac{\kappa}{2\mu} \text{dist}_X^2(x)
$$

$$
\geq f^* - \mu \mathbb{E} \left[ \|g_f(x^*; \xi)\|^2 \right] + \frac{\mu}{4} \mathbb{E} \left[ \|\nabla f_\mu(x; \xi)\|^2 \right] - \frac{\mu}{\kappa} \|g_f(x^*)\|^2 + \frac{\kappa}{4\mu} \text{dist}_X^2(x), \quad (16)
$$

for all $x \in \mathbb{R}^n$. In the first inequality we used convexity of $f(\cdot; \xi)$ and in the second the optimality conditions: $\langle g_f(x^*), z - x^* \rangle \geq 0$ for all $z \in X$, and in the third the inequality $ab \leq \frac{a}{2\alpha} + \frac{ab}{2}$, for $a = \|g_f(x^*)\|$, $b = \text{dist}_X(x)$, $\alpha = \frac{\kappa}{2\mu}$. Now we derive two auxiliary inequalities, useful for the final constant bounds. First, using the smoothing gradient inequality from Lemma [4.3] we obtain:

$$
\frac{\mu}{4} \mathbb{E} \left[ \|\nabla f_\mu(x; \xi)\|^2 \right] \geq \frac{\mu}{8} \mathbb{E} \left[ \|\nabla f_\mu(\pi_X(x); \xi)\|^2 \right] - \frac{\mu}{4} \mathbb{E} \left[ \|\nabla f_\mu(x; \xi) - \nabla f_\mu(\pi_X(x); \xi)\|^2 \right]
$$

$$
\geq \frac{\mu}{8} \mathbb{E} \left[ \|\nabla f_\mu(\pi_X(x); \xi)\|^2 \right] - \frac{1}{4\mu} \text{dist}_X^2(x)
$$

$$
\text{Lemma 4.3}\quad \geq \frac{\mu}{8} \mathbb{E} \left[ \|g_f(\pi_X(x); \xi)\|^2 \right] - \frac{1}{4\mu} \text{dist}_X^2(x)
$$

$$
\geq \frac{\sigma_f \mu}{16(1 + \mu L_{\max})^2} \text{dist}_X^2(\pi_X(x)) - \frac{1}{4\mu} \text{dist}_X^2(x)
$$

$$
\geq \frac{\sigma_f \mu}{32(1 + \mu L_{\max})^2} \text{dist}_X^2(x) - \left(\frac{1}{4\mu} + \frac{\sigma_f \mu}{16(1 + \mu L_{\max})^2}\right) \text{dist}_X^2(x). \quad (17)
$$
by the Cauchy-Schwartz inequality and the first order optimality conditions. By combining (16)-(17)-(18), then we have:

\[ f(\mu) \geq f^* - \frac{\mu}{2} \mathbb{E} \left[ \| g_f(x^\mu; \xi) \|_2^2 + \langle g_f(x^\mu), x - \pi(x) \rangle + \langle g_f(x^\mu), \pi(x) - x^* \rangle \right] \]

\[ \geq f^* - \frac{\mu}{2} \mathbb{E} \left[ \| g_f(x^\mu; \xi) \|_2^2 \right] - \| g_f(x^\mu) \| \text{dist}_\pi(x), \tag{18} \]

where in the last inequality we used linear regularity. By transferring all the terms containing \( F_\mu \) in the left hand side and denoting \( c = \frac{1}{2\mu} + \frac{\sigma_f \mu^2}{8\kappa(1 + \mu L_{\max})^2} \), then we finally obtain:

\[ (1 + c)(F_\mu(x) - F^*) \geq \frac{\sigma_f \mu}{32(1 + \mu L_{\max})^2} \text{dist}_\pi^2(x) \]

\[ - (1 + c)\mu \mathbb{E} \left[ \| g_f(x^\mu; \xi) \|_2^2 \right] - (1 + c^2)\frac{\mu}{\kappa} \| g_f(x^* \|_2^2, \tag{19} \]

which immediately confirms our above result. \( \Box \)

**Remark 2.** In the unconstrained finite case, when \( X = \mathbb{R}^n, \Omega = [m] \), a well-known class of problems having quadratic growth is given by:

\[ F(x) := \mathbb{E}[g(A_x; \xi)], \]

where \( g(\cdot; \xi) \) is strongly convex with constant \( \sigma_\xi > 0 \) and has Lipschitz gradients. However, for completeness we provide the main arguments for proving the property is satisfied. From the strong convexity property of \( g(\cdot; \xi \) we derive the restricted-strong convexity for each \( \xi \):

\[ g(A_x; \xi) \geq g(A_y; \xi) + \langle \nabla A_\xi^T g(A_y; \xi), x - y \rangle + \frac{\sigma_\xi}{2} \| A_x(x - y) \|_2^2, \quad \forall x, y \in \mathbb{R}^n. \tag{20} \]
Let \( x_1^*, x_2^* \in X^* \), then setting \( x = x_1^* \) and \( x = x_2^* \) and by taking expectation in both sides, results:

\[
F(x_1^*) \geq F(x_2^*) + \langle \nabla F(x_2^*), x_1^* - x_2^* \rangle + \frac{1}{2} \langle x_1^* - x_2^*, \mathbb{E}[\sigma_\xi A_\xi^T A_\xi](x_1^* - x_2^*) \rangle \\
\geq F(x_2^*) + \frac{1}{2} \langle x_1^* - x_2^*, \mathbb{E}[\sigma_\xi A_\xi^T A_\xi](x_1^* - x_2^*) \rangle.
\]

Since \( F(x_1^*) = F(x_2^*) \), the relation yields that there are unique \( y_\xi^* = A_\xi x^* \) for all \( x^* \in X^*, \xi \in \Omega \). Therefore, we clearly have that \( X^* = \{ x : A_\xi x = y_\xi^* \} \) and we can use the Hoffman’s bound to derive: there is \( \kappa_c \) such that \( \mathbb{E}[\| A_\xi x - y_\xi^* \|^2] \geq \kappa_c \text{dist}_{X^*}^2(x) \). Using this fact in (20) with \( y = x^* \) then we reach our conclusion.

For example, linear regression can be casted by the above particular model, i.e.

\[
F(x) = \mathbb{E}[(a_\xi^T x - b_\xi)^2].
\]

Notice that we do not assume the interpolation property and thus the model admits systems \( Ax = b \) without solution.

Based on standard arguments from literature, it can be shown that the above unconstrained model can be further generalized to linear constraints and polyhedral regularization (see \([5, 14, 30]\)). Many practical applications can be casted into one these models (see \([5, 30]\)), such as constrained linear regression, LASSO-regularized regression, support vector machine with polyhedral regularization etc.

4.2. Restricted strongly convex function with general constraints

A slightly more restrictive class of functions is described by the restricted strong convexity (RSC) property, which has been extensively analyzed in \([1, 26, 28, 29]\). Notice that in \([1]\) the authors derive complexity of SPP under RSC of component \( f(\cdot; \xi) \) and strong convexity of \( F \), which might indirectly allow indicator functions. However, our analysis allows using the linear regularity and deriving better constant bounds.

**Definition 4.5.** The function \( f(\cdot; \xi) \) is \( M_\xi \)-restricted strongly convex if there exists \( M_\xi \geq 0 \) such that

\[
f(x; \xi) \geq f(y; \xi) + \langle g_f(y; \xi), x - y \rangle + \frac{1}{2} \langle x - y, M_\xi(x - y) \rangle, \quad \forall x, y \in \mathbb{R}^n.
\]

Although it describes the behaviour of each component \( f(\cdot; \xi) \), this restricted convexity allows the elimination of other smoothness assumptions, which for the previous quadratic growth is not the case. Next, we show that the smoothing function also inherits the restricted strong convexity with specific parameter.

**Lemma 4.6.** Let \( f(\cdot; \xi) \) be \( M_\xi \)-restricted strongly convex (RSC). Then, given \( \mu > 0 \), the approximation \( f_\mu \) is \( \mathbb{E} \left[ \frac{M_\xi}{\lambda_{\max}(I + \mu M_\xi)} \right] \)-RSC:

\[
f_\mu(x) \geq f_\mu(y) + \langle \nabla f_\mu(y), x - y \rangle + \frac{1}{2} \langle x - y, \mathbb{E} \left[ \frac{M_\xi}{\lambda_{\max}(I + \mu M_\xi)} \right] (x - y) \rangle \quad \forall x, y \in \mathbb{R}^n.
\]

**Proof.** The proof can be found in the appendix. \( \square \)
Recall the fact that for a strongly convex function with constant $\sigma$, its Moreau smoothing remains strongly convex with constant $\frac{\sigma}{1 + \mu \sigma}$, see \cite{23}. Thus it is obvious that the RSC matrix $\frac{M_\xi}{\lambda_{\max}(I + \mu M_\xi)}$ might be regarded as a natural generalization of the previous fact. The $M_\xi$-RSC property do not require that the functional component $f(\cdot; \xi)$ to be strongly convex since $M_\xi \succeq 0$. However, if $M_f = \mathbb{E}[M_\xi] \succ 0$, then $f$ is $\lambda_{\min}(M_f)$–strongly convex. Although we are more interested in the non-strongly convex objective functions, further we also analyze for completeness the strongly convex case. Moreover, we show that the linear regularity brings significant advantages when $F$ is strongly convex.

Now we provide the result stating that, under RSC, the extended objective function $F$ satisfies the prox-quadratic growth property.

**Theorem 4.7.** Let $f(\cdot; \xi)$ be $M_\xi$–RSC and denote $\hat{M}_f = \mathbb{E} \left[ \frac{M_\xi}{\lambda_{\max}(I + \mu M_\xi)} \right]$. Then the composite function $F(x) := \mathbb{E} [f(x; \xi)] + \mathbb{E} \left[ ||X_\xi(x)|| \right]$ satisfies the following properties:

(i) Assume that $M_f \succ 0$, $y^* = M_f^{1/2} x^*$ is unique for all optimal points $x^*$ and $\{X_\xi\}_{\xi \in \Omega}$ are linearly regular with constant $\kappa$. By denoting $\hat{X} = \{ x : M_f^{1/2} x = y^* \}$, assume that the optimal set is defined by $X^* = X \cap \hat{X}$ and $\{X_\xi\}_{\xi \in \Omega}$ are linearly regular with constant $\kappa_{f,X}$. Then $F$ satisfies Assumption \ref{assump:prox} with constants:

\[
\sigma_{F,\mu} = \min \{ 1, \lambda_{\max}(M_{\max}) \} \kappa_{f,X}, \quad \beta = \frac{\mu}{2} \left( 1 + \frac{2}{\kappa} \right) \mathbb{E} [ ||g_f(x^*; \xi)||^2 ] + \frac{\mu}{2\kappa} ||g_f(x^*)||^2,
\]

where $M_{\max} = \max_{\xi \in \Omega} \lambda_{\max}(M_\xi)$.

(ii) If $M_f \succeq 0$, then $F$ satisfies Assumption \ref{assump:prox} with constants:

\[
\sigma_{F,\mu} = \lambda_{\min}(\hat{M}_f), \quad \beta = \frac{1}{2} \mathbb{E} [ ||g_F(x^*; \xi)||^2 ].
\]

(iii) If $M_f \succ 0$ and $\{X_\xi\}_{\xi \in \Omega}$ are linearly regular with constant $\kappa$, then $F$ satisfies prox-quadratic growth assumption \ref{assump:prox} with constants:

\[
\sigma_{F,\mu} = \lambda_{\min}(\hat{M}_f), \quad \beta = \frac{1}{2} \mathbb{E} [ ||g_f(x^*; \xi)||^2 ] + \frac{1}{2\kappa} ||\mathbb{E}[g_f(x^*; \xi)]||^2.
\]

**Proof.** Recall that $F_\mu(x) = f_\mu(x) + D_\mu(x)$, where $D_\mu(x) := \mathbb{E}[\text{dist}_{X_\xi}(x)]$. Let $x^*_\mu \in \arg\min_x F_\mu(x)$, $\tilde{x}^*_\mu = \arg\min_{x \in X} f_\mu(x)$ and denote $\hat{M}_f = \mathbb{E} \left[ \frac{M_\xi}{\lambda_{\max}(I + \mu M_\xi)} \right]$. Then, by using
Lemma 4.6 for $F_\mu$, we obtain:

$$F_\mu(x) = f_\mu(x) + D_\mu(x)$$

$$\geq f_\mu(\tilde{x}^*_\mu) + \langle \nabla f_\mu(\tilde{x}^*_\mu), x - \tilde{x}^*_\mu \rangle + \frac{1}{2} \langle x - \tilde{x}^*_\mu, M_\mu(x - \tilde{x}^*_\mu) \rangle + D_\mu(x)$$

$$= F_\mu(\tilde{x}^*_\mu) + \langle \nabla f_\mu(\tilde{x}^*_\mu), \pi_X(x) - \tilde{x}^*_\mu \rangle + \langle \nabla f_\mu(\tilde{x}^*_\mu), x - \pi_X(x) \rangle +$$

$$\frac{1}{2} \langle x - \tilde{x}^*_\mu, M_\mu(x - \tilde{x}^*_\mu) \rangle + D_\mu(x)$$

C.S. $$\geq F_\mu(\tilde{x}^*_\mu) + \langle \nabla f_\mu(\tilde{x}^*_\mu), \pi_X(x) - \tilde{x}^*_\mu \rangle + \langle \nabla f_\mu(\tilde{x}^*_\mu), x - \pi_X(x) \rangle +$$

$$\frac{1}{2} \langle x - \tilde{x}^*_\mu, M_\mu(x - \tilde{x}^*_\mu) \rangle + D_\mu(x)$$

l.r. $$\geq F_\mu(\tilde{x}^*_\mu) + \langle \nabla f_\mu(\tilde{x}^*_\mu), \pi_X(x) - \tilde{x}^*_\mu \rangle + \langle \nabla f_\mu(\tilde{x}^*_\mu), x - \pi_X(x) \rangle +$$

$$\frac{1}{2} \langle x - \tilde{x}^*_\mu, M_\mu(x - \tilde{x}^*_\mu) \rangle + D_\mu(x) + \frac{\kappa}{4\mu} \text{dist}_X(x)$$

$$\geq F_\mu(\tilde{x}^*_\mu) - \frac{\mu}{\kappa} \|\nabla f_\mu(\tilde{x}^*_\mu)\|^2 + \frac{1}{2} \langle x - \tilde{x}^*_\mu, M_\mu(x - \tilde{x}^*_\mu) \rangle + \frac{1}{2} D_\mu(x), \quad (21)$$

where in the second inequality we used Cauchy-Schwartz inequality and in the third we used linear regularity of feasible sets $\{X_\xi\}_{\xi \in \Omega}$. To bound further the right hand side, we first have from Lemma 6.1:

$$\|\nabla f_\mu(\tilde{x}^*_\mu)\|^2 \leq \frac{2}{\mu} (f_\mu(\tilde{x}^*_\mu) - f^*_\mu) \leq \frac{2}{\mu} (f_\mu(x^*) - f^*_\mu)$$

Lemma 2.2 (i) $$2 \|f(x^*) - f^*_\mu\| \leq \mathbb{E} \left[ \|g_f(x^*; \xi)\|^2 \right] + \frac{2}{\mu} \|g_f(x^*)\| \text{dist}_X(x), \quad (22)$$

where in the last inequality we applied Lemma 2.2 (ii) with $F = f$. Second, by applying (21) with $x = x^*$ and by using (22), we obtain:

$$\frac{1}{2} \langle x^* - \tilde{x}^*_\mu, M_\mu(x^* - \tilde{x}^*_\mu) \rangle \leq F_\mu(x^*) - F_\mu(\tilde{x}^*_\mu) + \frac{\mu}{\kappa} \mathbb{E} \left[ \|g_f(x^*; \xi)\|^2 \right]$$

$$\leq F(x^*) - F_\mu(\tilde{x}^*_\mu) + \frac{\mu}{\kappa} \mathbb{E} \left[ \|g_f(x^*; \xi)\|^2 \right]$$

Lemma 2.2 (iii) $$\frac{\mu}{2} \left( 1 + \frac{2}{\kappa} \right) \mathbb{E} \left[ \|g_f(x^*; \xi)\|^2 \right] + \frac{\mu}{2\kappa} \|g_f(x^*)\|^2 \quad (23)$$
Lastly, by taking into account that $X^* = \{ x^* : M_f^{1/2}x^* = y^*, x^* \in X_\xi, \xi \in \Omega \}$, then:

\[
\| M_f^{1/2}(x - x^*) \|^2 + D_\mu(x) \\
= \frac{1}{1 + \mu \lambda_{\max}(M_{\max})} \left( \langle x - x^*, [1 + \mu \lambda_{\max}(M_{\max})] M_f(x - x^*) \rangle \\
+ \frac{1 + \mu \lambda_{\max}(M_{\max})}{2\mu} \mathbb{E}[\text{dist}^2_{X_\xi}(x)] \right) \\
\geq \frac{1}{1 + \mu \lambda_{\max}(M_{\max})} \left( \langle x - x^*, \mathbb{E} \left[ \frac{1 + \mu \lambda_{\max}(M_{\max}) \lambda_{\max}(I_n + \mu M_{\xi})}{\lambda_{\max}(I_n + \mu M_{\xi})} \right] (x - x^*) \rangle \\
+ \frac{\lambda_{\max}(M_{\max})}{2} \mathbb{E}[\text{dist}^2_{X_\xi}(x)] \right) \\
\geq \frac{\min\{1, \lambda_{\max}(M_{\max})/2\}}{1 + \mu \lambda_{\max}(M_{\max})} \left( \langle x - x^*, M_f(x - x^*) \rangle + \mathbb{E}[\text{dist}^2_{X_\xi}(x)] \right) \\
\geq \frac{\min\{1, \lambda_{\max}(M_{\max})/2\}}{1 + \mu \lambda_{\max}(M_{\max})} \left( \| M_f^{1/2}x - y^* \|^2 + \mathbb{E}[\text{dist}^2_{X_\xi}(x)] \right) \\
l.r. \geq \frac{\min\{1, \lambda_{\max}(M_{\max})/2\}}{1 + \mu \lambda_{\max}(M_{\max})} \kappa_f, x \text{ dist}^2_{X_\xi}(x), \ \forall x \in \mathbb{R}^n, \ (24)
\]

where in the last inequality we have used that $X^* = \hat{X} \cap X$ and the linear regularity. This
last argument leads to the final lower bound:

\[
F_\mu(x) \geq F_\mu(\bar{x}_\mu) - \frac{\mu}{2} \left(1 + \frac{2}{\kappa}\right) \mathbb{E} \left[ \|g_f(x^*; \xi)\|^2 \right] - \frac{\mu}{\kappa} \left(\frac{1}{2} + \frac{2}{\kappa^2}\right) \|g_f(x^*)\|^2 \\
+ \frac{1}{4} \langle x - x^*, \bar{M}_f(x - x^*) \rangle + \frac{1}{4} D_\mu(x)
\]

which confirms the constants from part (i). For (ii) and (iii) we commonly derive

\[
F_\mu(x) = f_\mu(x) + D_\mu(x) \\
\geq F_\mu(x^*_\mu) + \langle \nabla F_\mu(x^*_\mu), x - x^*_\mu \rangle + \frac{1}{2} \langle x - x^*_\mu, \bar{M}_f(x - x^*_\mu) \rangle \\
\geq F_\mu(x^*_\mu) - \frac{1}{2} \langle x^* - x^*_\mu, \bar{M}_f(x^* - x^*_\mu) \rangle + \frac{1}{4} \langle x - x^*, \bar{M}_f(x - x^*) \rangle
\]

On the other hand, by taking \( x = x^* \) in (25), we get

\[
\frac{1}{2} \langle x^* - x^*_\mu, \bar{M}_f(x^* - x^*_\mu) \rangle \leq F_\mu(x^*) - F_\mu(x^*_\mu)
\]

\[
\frac{1}{2} \mathbb{E}[\|g_f(x^*; \xi)\|^2].
\]

From (25) and (26) we obtain the prox-quadratic growth relation:

\[
F_\mu(x) \geq F^*_\mu + \frac{\mu}{4} \langle x - x^*, \bar{M}_f(x - x^*) \rangle - \frac{1}{4} \langle x^* - x^*_\mu, \bar{M}_f(x^* - x^*_\mu) \rangle \\
\geq F^*_\mu + \frac{\lambda_{\min} (\bar{M}_f)}{4} \|x - x^*\|^2 - \frac{\mu}{2} \mathbb{E}[\|g_f(x^*; \xi)\|^2] \quad \forall x \in \mathbb{R}^n,
\]

which confirms result (ii). Lastly if \( X \) is linearly regular then, using Lemma 2.2 (iii), (26) transforms into:

\[
\frac{1}{2} \langle x^* - x^*_\mu, \bar{M}_f(x^* - x^*_\mu) \rangle \leq F^*_\mu - F_\mu(x^*_\mu) \\
\frac{1}{2} \mathbb{E}[\|g_f(x^*; \xi)\|^2] + \frac{\mu}{2\kappa} \|g_f(x^*)\|^2,
\]

and following the same lines as in the previous result (ii), we immediately obtain the constants from (iii). □

**Remark 3.** It is clear that the assumptions of Theorem 4.7 (i) hold for similar models as in the previous quadratic growth case:

\[
\min_{x \in \mathbb{R}^n} \mathbb{E}[g(A_x x; \xi)] \\
\text{s.t.} \quad C_x x \leq d_x, \quad \forall x \in \Omega.
\]
where \( g(\cdot; \xi) \) is strongly convex with constant \( \sigma_\xi > 0 \) and \( \Omega = \{1, \ldots, m\} \). However, the Lipschitz gradient assumption is not necessary any more. Since \( M_\xi = A_\xi^T A_\xi \), using similar arguments as in Remark 2 we have that there are unique \( y_\xi^* = A_\xi x^* \) and \( y^* = M_f^{1/2} x^* \) for all \( x^* \in X^*, \xi \in \Omega \). By observing that

\[
M_f^{1/2} x = y^* \iff \langle x - x^*, \mathbb{E}[A_\xi^T A_\xi(x - x^*)] \rangle = 0 \iff A_\xi x = y_\xi^*
\]

we clearly have that \( X^* = \{x^*: M_f^{1/2} x^* = y^*, C_\xi x \leq d_\xi, \xi \in \Omega\} \).

**Remark 4.** The other two cases (ii) and (iii) hold for more general optimization problems such as:

\[
\min_{x \in \mathbb{R}^n} \mathbb{E}[g(A_\xi x; \xi)] + \mathbb{E}[h(x; \xi)]
\]

s.t. \( x \in X = \bigcap_{\xi \in \Omega} X_\xi \),

where \( \mathbb{E}[A_\xi^T A_\xi] \succ 0 \) and \( h(\cdot; \xi) \) are general convex functions. It is important to observe that, under restricted strong convexity with \( M_f \succ 0 \), the prox-quadratic growth hold for general convex constraints, i.e. even if the constraints are not linearly regular. However, we provide a brief argument for the fact that when the linear regularity condition do not hold, the variance term \( \mathbb{E}[\|g(x^*; \xi)\|^2] \) might be arbitrarily large.

Assume that \( \Omega = [m] \) and \( X_\xi = \{x : h_\xi(x) \leq 0\} \). Then from the first-order stationarity conditions we have:

\[
0 \in \partial f(x^*) + N_{X_1}(x^*) + \cdots + N_{X_m}(x^*) \quad \forall x^* \in X^*.
\]

Notice that from these conditions one can derive the KKT optimality conditions ([23]), and thus yields the primal-dual optimality relation:

\[
\exists \lambda^* \in \mathbb{R}^m \text{ such that } 0 \in \partial f(x^*) + \sum_{i=1}^m \lambda_i^* \partial h_i(x^*) \quad \forall x^* \in X^*,
\]

which implies that the term \( \mathbb{E}[\|g(x^*; \xi)\|^2] \) might be dependent on the norm of the Lagrange multipliers. In the case when constraint qualifications do not hold, the optimal Lagrange multipliers might be significantly large, making the \( \beta \) constant from (iii) significantly safer and smaller over the \( \beta \) from (ii).

In [20], the linear regularity property of the constraint sets was essential to get the sublinear convergence rates. Also in [29] no regularity assumption is made on the feasible set, but note that a full projection on the entire feasible set is required at each iteration. This fact can be prohibitive when many constraints are present.

### 4.3. Convex feasibility problems

The most intuitive function class in our framework proves to be the indicator functions class, where \( f(x; \xi) = 0 \). Let \( \{X_\xi\}_{1 \leq \xi \leq m} \) be a finite collection of convex sets and \( X = \cap_\xi X_\xi \neq \emptyset \).
Under these terms yields that:

\[ F_\mu(x) = \mathbb{E} \left[ \text{dist}^2_{X_\xi}(x) \right], \quad F^*_\mu = F^* = 0, \quad X^*_\mu = X^* = X. \]

Then, it is easy to see that under the linear regularity assumption, the prox-quadratic growth property is immediately implied

\[ F_\mu(x) - F^*_\mu = \mathbb{E} \left[ \frac{1}{2\mu} \text{dist}^2_{X_\xi}(x) \right] \geq \frac{\kappa}{2\mu} \text{dist}^2_{X}(x) \quad \forall x \in \mathbb{R}^n, \]

with corresponding constants \( \sigma_{F,\mu} = \frac{\kappa}{\mu} \) and \( \beta = 0 \). Most common example of linearly regular sets are the polyhedral sets. Also in the case when the intersection has nonempty interior and contains a ball of radius \( \delta \), then the linear regularity holds with constant \( \kappa \) dependent on \( \delta \), see [15].

5. Conclusions

We presented novel convergence results for stochastic proximal point algorithm in the (non-)strongly convex setting. In particular, we show that the prox-quadratic growth assumption is sufficient to guarantee \( O\left(\frac{1}{k}\right) \) convergence rate, in terms of the distance to the optimal set. Also, linear convergence is recovered for interpolation setting, when SPP becomes RAP scheme.

Funding

This work was supported by BRD Groupe Societe Generale through Data Science Research Fellowships of 2019.

References

[1] H. Asi and J. C. Duchi, *Stochastic (Approximate) Proximal Point Methods: Convergence, Optimality, and Adaptivity*, Arxiv, 2018.
[2] H.H. Bauschke, F. Deutsch, H. Hundal, and S.-H. Park, *Accelerating the convergence of the method of alternating projections*, Transactions of the American Mathematical Society 355(9), pp. 3433-3461, 2003.
[3] P. Bianchi, *Ergodic convergence of a stochastic proximal point algorithm*, SIAM Journal on Optimization, 26(4): 2235–2260, 2016.
[4] Y. Censor, W. Chen, P. L. Combettes, R. Davidi and G. T. Herman, *On the effectiveness of projection methods for convex feasibility problems with linear inequality constraints*, Computational Optimization and Applications, 51(3) : 1065–1088 , 2012.
[5] D. Drusvyatskiy, A. S. Lewis, *Error Bounds, Quadratic Growth, and Linear Convergence of Proximal Methods*, Mathematics of Operations Research, 43(3): 919-948, 2018.
[6] L.G. Gubin, B.T. Polyak and E.V. Raik, *The method of projections for finding the common points of convex sets*, USSR Comp. Math. Phys. 7 (1967) l-24.
[7] O. Guler, *On the Convergence of the Proximal Point Algorithm for Convex Minimization*, SIAM Journal on Control and Optimization, 29(2) : 403 - 419, 1991.
[8] E. Hazan and S. Kale, *Beyond the Regret Minimization Barrier: Optimal Algorithms for Stochastic Strongly-Convex Optimization*, Journal of Machine Learning Research, 15:2489–2512, 2014.
[9] R. A. Horn, C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1990.

[10] S. Ma, R. Bassily and M. Belkin, *The Power of Interpolation: Understanding the Effectiveness of SGD in Modern Over-parametrized Learning*, arXiv:1712.06559, 2018.

[11] J. Koshal and A. Nedic and U. V. Shanbhag, *Regularized Iterative Stochastic Approximation Methods for Stochastic Variational Inequality Problems*, IEEE Transactions on Automatic Control, 58(3): 594 - 609, 2013.

[12] S. Lacoste-Julien, M. Schmidt and F. Bach, *A simpler approach to obtaining an O(1/t) convergence rate for projected stochastic subgradient descent*, CoRR, abs/1212.2002, 2012.

[13] E. Moulines and F. R. Bach, *Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning*, Advances in Neural Information Processing Systems 24 (NIPS), 451-459, 2011.

[14] I. Necoara, Yu. Nesterov and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, Mathematical Programming, https://doi.org/10.1007/s10107-018-1232-1, 2018.

[15] A. Nedic, *Random projection algorithms for convex set intersection problems*, 49th IEEE Conference on Decision and Control (CDC), 7655-7660, 2010.

[16] A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on Optimization, 19(4):1574–1609, 2009.

[17] I. Necoara, P. Richtarik and A. Patrascu, *Randomized projection methods for convex feasibility problems: conditioning and convergence rates*, submitted, arXiv:1801.04873, 2018.

[18] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, Springer, 2004.

[19] L. Nguyen, P. H. NGUYEN, M. Dijk, P. Richtarik, K. Scheinberg, M. Takac, *SGD and Hogwild! Convergence Without the Bounded Gradients Assumption*, Proceedings of the 35th International Conference on Machine Learning, PMLR 80:3750-3758, 2018.

[20] A. Patrascu, I. Necoara, *Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization*, Journal of Machine Learning Research, 19:1-42, 2018.

[21] A. Rakhlin, O. Shamir and K. Sridharan, *Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization*, Proceedings of the 29th International Conference on Machine Learning 1571–1578, 2012.

[22] A. Ramdas and A. Singh, *Optimal rates for stochastic convex optimization under Tsybakov noise condition*, Proceedings of the 30th International Conference on Machine Learning, 28(1):365–373, 2013.

[23] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1998.

[24] R.T. Rockafellar and R.J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin Heidelberg, 1998.

[25] L. Rosasco, S. Villa and B. C. Vu, *Convergence of Stochastic Proximal Gradient Algorithm*, Arxiv, https://arxiv.org/abs/1403.5074, 2014.

[26] E. Ryu and S. Boyd, *Stochastic Proximal Iteration: A Non-Asymptotic Improvement Upon Stochastic Gradient Descent*, http://web.stanford.edu/~eryu/, 2016.

[27] S. Shalev-Shwartz, Y. Singer, N. Srebro, A. Cotter, *Pegasos: primal estimated sub-gradient solver for SVM*, Mathematical Programming, 127(1):3–30, 2011.

[28] P. Toulis, D. Tran and E. M. Airoldi, *Towards stability and optimality in stochastic gradient descent*, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, PMLR 51:1290-1298, 2016.

[29] T. Yang and Q. Lin, *RSG: Beating Subgradient Method without Smoothness and Strong Convexity*, Journal of Machine Learning Research 19 : 1 - 33, 2018.

[30] Y. Xu, Q. Lin and T. Yang, *Stochastic Convex Optimization: Faster Local Growth Implies Faster Global Convergence*, International Conference on Machine Learning (ICML), 2017.

[31] M. Wang and D. P. Bertsekas, *Stochastic First-Order Methods with Random Constraint Projection*, SIAM Journal on Optimization, 26(1):681717, 2016.
6. Appendix

Lemma 6.1. Let \( h : \mathbb{R}^n \to \mathbb{R} \) be convex and having Lipschitz continuous gradient with constant \( L_h \), then the following relation hold:

\[
\| \nabla h(x) \| \leq 2L_h (h(x) - h(x^*)) \quad \forall x \in \mathbb{R}^n,
\]

where \( x^* \in \arg \min_x h(x) \).

**Proof.** From Lipschitz continuity, we have:

\[
h(y) \leq h(x) + \langle \nabla h(x), y - x \rangle + \frac{L_h}{2} \| x - y \|^2.
\]

By minimizing both sides over \( y \), we obtain:

\[
h(x^*) \leq h(x) - \frac{1}{2L_h} \| \nabla h(x) \|^2,
\]

which confirms the result. \(\square\)

Lemma 6.2. Let \( f \) be continuously differentiable, then \( f \) is \( M \)-restricted strongly convex if and only if:

\[
\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \langle x - y, M(x - y) \rangle \quad \forall x, y \in \mathbb{R}^n.
\] (28)

**Proof.** Assume that \( f \) is \( M \)-restricted strongly convex, then by adding the relation

\[
f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2}(x - y, M(x - y))
\]

with the same but with interchanged \( x \) and \( y \) then we obtain the first implication. Next, assume that (28) holds. By the Mean Value Theorem we have:

\[
f(x) = f(y) + \int_0^1 \langle \nabla f(\tau x + (1 - \tau)y), x - y \rangle d\tau
\]

\[
= f(y) + \langle \nabla f(y), x - y \rangle + \int_0^1 \frac{1}{\tau} \langle \nabla f(\tau x + (1 - \tau)y) - \nabla f(y), \tau(x - y) \rangle d\tau
\]

\[
\geq f(y) + \langle \nabla f(y), x - y \rangle + \int_0^1 \frac{1}{2} \langle x - y, M(x - y) \rangle d\tau
\]

\[
= f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2}(x - y, M(x - y))
\]

which confirms the second implication. \(\square\)

**proof of Lemma 4.6** From the \( M_\xi \)-restricted strong convexity assumption we have:

\[
\langle \nabla f(x; \xi) - \nabla f(y; \xi), x - y \rangle \geq \| x - y \|^2_{M_\xi}.
\]
By taking \( x = z_\mu(x; \xi) \) and \( y = z_\mu(y; \xi) \) then the above relation implies:

\[
\|z_\mu(x; \xi) - z_\mu(y; \xi)\|_{M_\xi}^2 \leq \langle \nabla f(z_\mu(x; \xi); \xi) - \nabla f(z_\mu(y; \xi); \xi), z_\mu(x; \xi) - z_\mu(y; \xi) \rangle \\
\leq \frac{1}{\mu} \langle x - z_\mu(x; \xi) - (y - z_\mu(y; \xi)), z_\mu(x; \xi) - z_\mu(y; \xi) \rangle \\
\leq \frac{1}{\mu} \langle x - y, z_\mu(x; \xi) - z_\mu(y; \xi) \rangle - \frac{1}{\mu} \|z_\mu(x; \xi) - z_\mu(y; \xi)\|^2.
\]

(29)

After simple manipulations, using the Cauchy-Schwartz inequality the last inequality further implies:

\[
\langle z_\mu(x; \xi) - z_\mu(y; \xi), (I_n + \mu M_\xi)(z_\mu(x; \xi) - z_\mu(y; \xi)) \rangle \leq \langle x - y, z_\mu(x; \xi) - z_\mu(y; \xi) \rangle \\
= \langle (I_n + \mu M_\xi)^{-1/2}(x - y), (I_n + \mu M_\xi)^{1/2}(z_\mu(x; \xi) - z_\mu(y; \xi)) \rangle \\
\leq \|I_n + \mu M_\xi\|^{-1/2}(x - y)\|((I_n + \mu M_\xi)^{1/2}(z_\mu(x; \xi) - z_\mu(y; \xi))\|
\]

(30)

An important consequence of (30) is the following contraction property:

\[
\|(I_n + \mu M_\xi)^{1/2}(z_\mu(x; \xi) - z_\mu(y; \xi))\| \leq \|(I_n + \mu M_\xi)^{-1/2}(x - y)\| \quad \forall x, y \in \mathbb{R}^n, \xi \in \Omega.
\]

(31)

Now by using the particular structure of \( \nabla f_\mu(\cdot; \xi) \) and that fact that \( I_n + \mu M_\xi \) is invertible, we have:

\[
\langle \nabla f_\mu(x; \xi) - \nabla f_\mu(y; \xi), x - y \rangle = \frac{1}{\mu} \|x - y\|^2 - \frac{1}{\mu} \langle z_\mu(x; \xi) - z_\mu(y; \xi), x - y \rangle \\
= \frac{1}{\mu} \|x - y\|^2 - \frac{1}{\mu} \|(I_n + \mu M_\xi)^{1/2}(z_\mu(x; \xi) - z_\mu(y; \xi)), (I_n + \mu M_\xi)^{-1/2}(x - y)\rangle.
\]

By taking expectation in both sides and also using the Cauchy-Schwartz inequality and the contraction property (31) we get:

\[
\langle \nabla F_\mu(x) - \nabla F_\mu(y), x - y \rangle \\
= \frac{1}{\mu} \|x - y\|^2 - \frac{1}{\mu} \mathbb{E} \left[ \langle (I_n + \mu M_\xi)^{1/2}(z_\mu(x; \xi) - z_\mu(y; \xi)), (I_n + \mu M_\xi)^{-1/2}(x - y) \rangle \right] \\
\geq \frac{1}{\mu} \|x - y\|^2 - \frac{1}{\mu} \mathbb{E} \left[ \|(I_n + \mu M_\xi)^{-1/2}(x - y)\|^2 \right] \\
\geq \frac{1}{\mu} \|x - y\|^2 - \frac{1}{\mu} (x - y, \mathbb{E} [(I_n + \mu M_\xi)^{-1}] (x - y)) \\
= \frac{1}{\mu} \|x - y\|^2 - \frac{1}{\mu} (x - y, \mathbb{E} [(I_n + \mu M_\xi)^{-1}] (x - y)) \\
= \frac{1}{\mu} \langle x - y, I - \mathbb{E} [(I_n + \mu M_\xi)^{-1}] (x - y) \rangle.
\]

(32)
We further deduce that:

\[
I_n - (I_n + \mu M \xi)^{-1} = \left[ I_n - (I_n + \mu M \xi)^{-1} \right]^{1/2} \left[ I_n - (I_n + \mu M \xi)^{-1} \right]^{1/2} \\
= (I_n + \mu M \xi - I_n)^{1/2} (I_n + \mu M \xi)^{-1/2} (I_n + \mu M \xi) (I_n + \mu M \xi - I_n)^{1/2} \\
= \mu M^{1/2} (I_n + \mu M \xi)^{-1} M^{1/2}.
\] (33)

By using this bound into (32), then we finally obtain the strong convexity relation:

\[
\langle \nabla F_{\mu}(x) - \nabla F_{\mu}(y), x - y \rangle \geq \frac{1}{\mu} \langle x - y, I_n - \mathbb{E} [(I_n + \mu M \xi)^{-1}] (x - y) \rangle \\
\geq \langle x - y, \mathbb{E} \left[ M^{1/2} (I_n + \mu M \xi)^{-1} M^{1/2} \right] (x - y) \rangle \\
\geq \langle x - y, \mathbb{E} \left[ M \xi \lambda_{\min} \left( (I_n + \mu M \xi)^{-1} \right) \right] (x - y) \rangle.
\] (34)

As the last step of the proof, by observing \( \lambda_{\min} \left( (I_n + \mu M \xi)^{-1} \right) = \frac{1}{\lambda_{\text{max}} (I_n + \mu M \xi)} \) and by applying Lemma 6.2 with \( f = F_{\mu} \) and \( M = \mathbb{E} \left[ M \xi \lambda_{\min} \left( (I_n + \mu M \xi)^{-1} \right) \right] \), makes the connection between (34) and the above result.