Minibatch and Momentum Model-based Methods for Stochastic Weakly Convex Optimization

Qi Deng[1] **Wenzhi Gao**[2]

School of Information Management and Engineering
Shanghai University of Finance and Economics

[1]qideng@sufe.edu.cn [2]gwz@163.shufe.edu.cn

Abstract

Stochastic model-based methods have received increasing attention lately due to their appealing robustness to the stepsize selection and provable efficiency guarantee. We make two important extensions for improving model-based methods on stochastic weakly convex optimization. First, we propose new minibatch model-based methods by involving a set of samples to approximate the model function in each iteration. For the first time, we show that stochastic algorithms achieve linear speedup over the batch size even for non-smooth and non-convex (particularly, weakly convex) problems. To this end, we develop a novel sensitivity analysis of the proximal mapping involved in each algorithm iteration. Our analysis appears to be of independent interests in more general settings. Second, motivated by the success of momentum stochastic gradient descent, we propose a new stochastic extrapolated model-based method, greatly extending the classic Polyak momentum technique to a wider class of stochastic algorithms for weakly convex optimization. The rate of convergence to some natural stationarity condition is established over a fairly flexible range of extrapolation terms.

While mainly focusing on weakly convex optimization, we also extend our work to convex optimization. We apply the minibatch and extrapolated model-based methods to stochastic convex optimization, for which we provide a new complexity bound and promising linear speedup in batch size. Moreover, an accelerated model-based method based on Nesterov's momentum is presented, for which we establish an optimal complexity bound for reaching optimality.

1 Introduction

In this paper, we are interested in the following stochastic optimization problem:

$$\min_{x \in \mathcal{X}} \quad f(x) = \mathbb{E}_{\xi \sim \Xi}\big[f(x,\xi)\big] \tag{1}$$

where $f(\cdot,\xi)$ stands for the loss function, sample ξ follows certain distribution Ξ, and $\mathcal{X}$ is a closed convex set. We assume that $f(\cdot,\xi)$ is weakly convex, namely, the sum of $f(x,\xi)$ and a quadratic function $\frac{\lambda}{2}\|x\|^2$ is convex ($\lambda > 0$). This type of non-smooth non-convex functions can be found in a variety of machine learning applications, such as phase retrieval, robust PCA and low rank decomposition [9]. To solve problem (1), we consider the stochastic model-based method (`SMOD`, [15, 10, 2]), which comprises a large class of stochastic algorithms (including stochastic (sub)gradient descent, proximal point, among others). Recent work [15, 10] show that `SMOD` exhibits promising convergence property: both asymptotic convergence and rates of convergence to certain stationarity

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

measure have been established for the SMOD family. In addition, empirical results [10, 16] indicate that SMOD exhibits remarkable robustness to hyper-parameter tuning and often outperforms SGD.

Despite much recent progress, our understanding of model-based methods for weakly convex optimization is still quite limited. Particularly, it is still unknown whether SMOD is competitive against modern SGD used in practice. We highlight some important remaining questions. First, despite the appealing robustness and stable convergence, the SMOD family is sequential in nature. It is unclear whether minibatching, which is immensely used in training learning models, can improve the performance of SMOD when the problem is non-smooth. Particularly, the current best complexity bound $\mathcal{O}(\frac{L^2}{\varepsilon^4})$ from [10], which is regardless of batch size, is unsatisfactory. Were this bound tight, a sequential algorithm (using one sample per iteration) would be optimal: it offers the highest processing speed per iteration as well as the best iteration complexity. Therefore, it is crucial to know whether minibatching can improve the complexity bound of the SMOD family or the current bound is tight. Second, in modern applications, momentum technique has been playing a vital role in large-scale non-convex optimization (see [34, 31]). In spite of its effectiveness, to the best of our knowledge, momentum technique has been provably efficient only in **1)** unconstrained smooth optimization [25, 11, 20] and **2)** non-smooth optimization with a simple constraint [27], which constitute only a portion of the interesting applications. From the practical aspect, it is peculiarly desirable to know whether momentum technique is applicable beyond in SGD and whether it can benefit the SMOD algorithm family in the non-smooth and non-convex setting.

Contributions. Our work is motivated by the aforementioned challenge to make SMOD more practically efficient. We summarize the contributions as follows. First, we extend SMOD to the minibatch setting and develop sharper rates of convergence to stationarity. Leveraging the tool of algorithm stability ([7, 30, 21]), we provide a nearly complete recipe on when minibatching would be helpful even in presence of non-smoothness. Our theory implies that stochastic proximal point and stochastic prox-linear are inherently parallelizable: both algorithms achieve linear speedup over the minibatch size. To the best of our knowledge, this is the first time that these minibatch stochastic algorithms are proven to exhibit such an acceleration even for *non-smooth* and *non-convex* (particularly, *weakly convex*) optimization. Moreover, our theory recovers the complexity of minibatch (proximal) SGD in [10], showing that (proximal) SGD enjoys the same linear speedup by minibatching for smooth composite problems with non-smooth regularizers or with constrained domain.

Second, we present new extrapolated model-based methods by incorporating a Polyak-type momentum term. We develop a unified Lyapunov analysis to show that a worst-case complexity of $\mathcal{O}(1/\varepsilon^4)$ holds for all momentum SMOD algorithms. To the best of our knowledge, these are the first complexity results of momentum stochastic prox-linear and stochastic proximal point for non-smooth non-convex optimization. Since our analysis offers complexity guarantees for momentum SGD and its proximal variant, our work appears to be more general than a recent study [27], which only proves the convergence of momentum projected SGD. Proximal SGD is more advantageous in composite optimization, where the non-smooth term is often involved via its proximal operator rather than the subgradient. For example, in the Lasso problem, it is often favorable to invoke the proximal operator of ℓ_1 function (Soft-Thresholding) to enhance solution sparsity. We summarize the complexity results in Table 1.

Third, we develop new convergence results of SMOD for convex optimization, showing that minibatch extrapolated SMOD achieves a promising linear speedup over the batch size under some mild condition. Specifically, to obtain some ε-optimal solution, our proposed method exhibits an $\mathcal{O}(1/\varepsilon + 1/(m\varepsilon^2))$ complexity bound in the worst case. Moreover, we develop a new minibatch SMOD based on Nesterov's momentum, achieving the $\mathcal{O}(1/\varepsilon^{1/2} + 1/(m\varepsilon^2))$ optimal complexity bound. Note that a similar complexity result, explicitly relying on the smoothness assumption, has been shown in a recent study [8]. Compared to this work, our analysis makes weaker assumptions, showing that smoothness is not a must-have for many model-based algorithms, such as SPL and SPP, to get sharper complexity bound.

Other related work. For smooth and composite optimization, it is well known that SGD can be linearly accelerated by minibatching (c.f. [12, 19, 32]). Minibatch model-based methods have been studied primarily in the convex setting. Asi et al. [3] investigates the speedups of minibatch stochastic model-based methods in the convex smooth, restricted strongly convex and convex interpolation settings, respectively. Since their assumptions differ from ours, the technique does not readily apply to the non-convex setting. Chadha et al. [8] studies the accelerated minibatch model-based methods

Table 1: Complexity of SMOD to reach $\mathbb{E}\,\|\nabla_{1/\rho} f\| \leq \varepsilon$ (M: minibatch; E: Extrapolation, m: batch size)

Algorithms	Problem	Current Best	Ours
M + SGD	f: non-smooth	$\mathcal{O}(1/\varepsilon^4)$[10]	$\mathcal{O}(1/\varepsilon^4)$
M + Prox. SGD	$f = \ell + \omega$; ℓ:smooth	$\mathcal{O}(1/(m\varepsilon^4) + 1/\varepsilon^2)$[10]	$\mathcal{O}(1/(m\varepsilon^4) + 1/\varepsilon^2)$
M + SPL/SPP	f: non-smooth	$\mathcal{O}(1/\varepsilon^4)$[10]	$\mathcal{O}(1/(m\varepsilon^4) + 1/\varepsilon^2)$
E + SGD	f: non-smooth	$\mathcal{O}(1/\varepsilon^4)$[27]	$\mathcal{O}(1/\varepsilon^4)$
E + Prox. SGD	$f = \ell + \omega$; ℓ:smooth	—	$\mathcal{O}(1/\varepsilon^4)$
E + SPL/SPP	f: non-smooth	—	$\mathcal{O}(1/\varepsilon^4)$
M + E + SGD	f: non-smooth	$\mathcal{O}(1/\varepsilon^4)$[27]	$\mathcal{O}(1/\varepsilon^4)$
M + E + Prox. SGD	$f = \ell + \omega$; ℓ:smooth	—	$\mathcal{O}(1/(m\varepsilon^4) + 1/\varepsilon^2)$
M + E + SPL/SPP	f: non-smooth	—	$\mathcal{O}(1/(m\varepsilon^4) + 1/\varepsilon^2)$

for convex smooth and convex interpolated problems. The interpolation setting, where the model can perfectly fit the data, is not considered in our paper. Algorithm stability [7, 30]—an important technique for analyzing the generalization performance of stochastic algorithms [21, 4], is the key tool to obtain some of our convergence results. In contrast to the traditional work, our paper employs the stability argument to obtain sharper optimization convergence rates (with respect to the batch size). See Section 3. As noted by an anonymous reviewer, a similar idea of using stability analysis was proposed by Wang et al. [33], albeit with a different motivation from distributed stochastic optimization. Robustness and fast convergence of model-based methods have been shown on various statistical learning problems [9, 16, 2, 5, 17, 6]. Drusvyatskiy and Paquette [14] give a complete complexity analysis of the accelerated proximal-linear methods for deterministic optimization. Zhang and Xiao [35] further improve the convergence rates of prox-linear methods on certain finite-sum and stochastic problems by using variance-reduction. Momentum and accelerated methods for convex stochastic optimization can be referred from [26, 29]. The study [11, 25, 34] develop the convergence rate of stochastic momentum method for smooth non-convex optimization.

2 Background

Throughout the paper, we use $\|\cdot\|$ to denote the Euclidean norm and $\langle\cdot,\cdot\rangle$ to denote the Euclidean inner product. We assume that $f(x)$ is bounded below. i.e., $\min_x f(x) > -\infty$. The subdifferential $\partial f(x)$ of function $f(x)$ is the set of vectors $v \in \mathbb{R}^d$ that satisfy: $f(y) \geq f(x) + \langle v, y - x\rangle + o(\|x - y\|)$, as $y \to x$. Any such vector in $\partial f(x)$ is called a subgradient and is denoted by $f'(x) \in \partial f(x)$ for simplicity. We say that a point x is stationary if $0 \in \partial f(x) + N_{\mathcal{X}}(x)$, where the normal cone $N_{\mathcal{X}}(x)$ is defined as $N_{\mathcal{X}}(x) \triangleq \{d : \langle d, y - x\rangle \leq 0, \forall y \in \mathcal{X}\}$. For a set S, define the set distance to 0 by: $\|\mathcal{S}\|_- \triangleq \inf\{\|x - 0\|, x \in \mathcal{S}\}$. It is natural to use the quantity $\|\partial f(x) + N_{\mathcal{X}}(x)\|_-$ to measure the stationarity of point x.

Moreau-envelope. The μ-Moreau-envelope of f is defined by $f_\mu(x) \triangleq \min_{y\in\mathcal{X}} \left\{f(y) + \frac{1}{2\mu}\|x - y\|^2\right\}$ and the proximal mapping associated with $f(\cdot)$ is defined by $\mathrm{prox}_{\mu f}(x) \triangleq \mathrm{argmin}_{y\in\mathcal{X}} \left\{f(y) + \frac{1}{2\mu}\|x - y\|^2\right\}$. Assume that $f(x)$ is λ-weakly convex, then for $\mu < \lambda^{-1}$, the Moreau envelope $f_\mu(\cdot)$ is differentiable and its gradient is $\nabla f_\mu(x) = \mu^{-1}(x - \mathrm{prox}_{\mu f}(x))$.

The SMOD family iteratively computes the proximal map associated with a model function $f_{x^k}(\cdot, \xi_k)$:

$$x^{k+1} = \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ f_{x^k}(x, \xi_k) + \frac{\gamma_k}{2}\|x - x^k\|^2 \right\}, \tag{2}$$

where $\{\xi_k\}$ are i.i.d. samples. Typical algorithms and the accompanied models are described below.

Stochastic (Proximal) Gradient Descent: consider the composite function $f(x, \xi) = \ell(x, \xi) + \omega(x)$ where $\ell(x, \xi)$ is a data-driven and weakly-convex loss term and $\omega(x)$ is a convex regularizer such as ℓ_1-penalty. SGD applies the model function:

$$f_y(x, \xi) = \ell(y, \xi) + \left\langle \ell'(y, \xi), x - y \right\rangle + \omega(x). \tag{3}$$

Stochastic Prox-linear (SPL): consider the composition function $f(x, \xi) = h(C(x, \xi))$ where $h(\cdot, \xi)$ is convex continuous and $C(x, \xi)$ is a continuously differentiable map. We perform partial

linearization to obtain the model

$$f_y(x,\xi) = h\big(C(y,\xi) + \langle \nabla C(y,\xi), x-y\rangle\big). \tag{4}$$

Stochastic Proximal Point (SPP): compute (2) with full stochastic function:

$$f_y(x,\xi) = f(x,\xi). \tag{5}$$

Throughout the paper, we assume that $f(x,\xi)$ is continuous and μ-weakly convex, and that the model function $f_x(\cdot,\cdot)$ satisfies the following assumptions [10].

- **A1:** For any $\xi \sim \Xi$, the model function $f_x(y,\xi)$ is λ-weakly convex in y ($\lambda \geq 0$).
- **A2:** Tightness condition: $f_x(x,\xi) = f(x,\xi),\ \forall x \in \mathcal{X},\ \xi \sim \Xi$.
- **A3:** One-sided quadratic approximation: $f_x(y,\xi) - f(y,\xi) \leq \frac{\tau}{2}\|x-y\|^2,\ \forall x,y \in \mathcal{X}, \xi \sim \Xi$.
- **A4:** Lipschitz continuity: There exists $L > 0$ that $f_x(z,\xi) - f_x(y,\xi) \leq L\,\|z-y\|$, for any $x,y,z \in \mathcal{X},\ \xi \sim \Xi$.

Remark 1. Assumption A2 is quite standard and will be used only in the convergence proof. Combining A1 and A3, we immediately have that $f(x,\xi)$ is $(\lambda+\tau)$-weakly convex. Thus, it suffices to assume that $\mu < \tau + \lambda$. Assumptions A2-A4 can be slightly relaxed by replacing the uniform bound with a bound on expectation over ξ, leading to only a minor adjustment to the analysis.

Denote $\hat{x} \triangleq \operatorname{prox}_{f/\rho}(x) = \operatorname{argmin}_y \left\{ f(y) + \frac{\rho}{2}\|y-x\|^2 \right\}$ for some $\rho > \mu$. Davis and Drusvyatskiy [10] revealed a striking feature of Moreau envelope to characterize stationarity:

$$\|\hat{x} - x\| = \rho^{-1}\|\nabla f_{1/\rho}(x)\|, \text{ and } \|\partial f(\hat{x}) + N_{\mathcal{X}}(\hat{x})\|_- \leq \|\nabla f_{1/\rho}(x)\|.$$

Namely, a point x with small gradient norm $\|\nabla f_{1/\rho}(x)\|$ stays in the proximity of a nearly-stationary point $\hat{x}$. With this observation, they show the first complexity result of SMOD for non-smooth non-convex optimization: $\min_{1\leq k\leq K} \mathbb{E}[\|\nabla f_{1/\rho}(x^k)\|]^2 \leq \mathcal{O}(\frac{L}{\sqrt{K}})$. Note that this rate is regardless of the size of minibatches since it does not explicitly use any information of the samples other than the Lispchitzness of the model function. Due to this limitation, it remains unclear whether minibatching can further improve the convergence rate of SMOD.

3 SMOD with minibatches

In this section, we present a minibatch SMOD method which takes a small batch of i.i.d. samples to estimate the model function. The overall procedure is detailed in Algorithm 1. Within each iteration, Algorithm 1 forms a stochastic model function $f_{x^k}(\cdot, B_k) = \frac{1}{m_k}\sum_{i=1}^{m_k} f_{x^k}(x,\xi_{k,i})$ parameterized at x^k by sampling over m_k i.i.d. samples $B_k = \xi_{k,1},\ldots,\xi_{k,m_k}$. Then it performs proximal update to get the next iterate x^{k+1}. We will illustrate the main convergence results of Algorithm 1 and leave all the proof details in Appendix sections. But first, let us present an additional assumption.

- **A5:** Two-sided quadratic bound: for any $x,y \in \mathcal{X}, \xi \sim \Xi$, $\big|f_x(y,\xi) - f(y,\xi)\big| \leq \frac{\tau}{2}\|x-y\|^2$.

Remark 2. Assumption A5 is vital for our improved convergence analysis. While it is slightly stronger than A3, A5 is indeed satisfied by the SMOD family in most contexts: **1)** For SPP, A5 is trivially satisfied by taking $f_x(y,\xi) = f(y,\xi)$. **2)** For SPL, we minimize a composition function $f(x,\xi) = h(C_\xi(x))$ where $h(\cdot)$ is a c_1-Lipschitz convex function and $C_\xi(\cdot)$ is a c_2-Lipschitz smooth map. In view of (4), A5 is verified with $|f_x(y,\xi) - f(y,\xi)| \leq c_1\big\|C_\xi(y) - C_\xi(x) - \nabla C_\xi(x)^{\mathrm{T}}(y-x)\big\| \leq \frac{c_1c_2}{2}\|x-y\|^2$. **3)** For SGD, A5 is satisfied if $\ell(\cdot,\xi)$ is c_3-Lipschitz smooth for some $c_3 > 0$, as $|f_x(y,\xi) - f(y,\xi)| \leq |\ell(y,\xi) - \ell(x,\xi) - \nabla\ell(x,\xi)^{\mathrm{T}}(y-x)| \leq \frac{c_3}{2}\|x-y\|^2$. We note that A5 is not satisfied by SGD when the loss $\ell(\cdot,\xi)$ is also non-smooth. Unfortunately, there seems to be little hope to accelerate SGD in such a case since the convergence rate of SGD already matches the rate of deterministic subgradient method.

We present an improved complexity analysis of SMOD by leveraging the framework of algorithm stability [7, 30]. In stark contrast to its standard application in characterizing the algorithm generalization performance, stability analysis is applied to determine how the variation of a minibatch affects the *estimation of the model function* in each algorithm iteration.

Algorithm 1 Stochastic Model-based Method with Minibatches (SMOD)

Input: x^1

for $k = 1$ **to** K **do**

Sample a minibatch $B_k = \{\xi_{k,1}, \ldots, \xi_{k,m_k}\}$ and update x^{k+1} by solving

$$\min_{x \in \mathcal{X}} \left\{ \frac{1}{m_k} \sum_{i=1}^{m_k} f_{x^k}(x, \xi_{k,i}) + \frac{\gamma_k}{2} \|x - x^k\|^2 \right\} \tag{6}$$

end for

Notations. Let $B = \{\xi_1, \xi_2, \ldots, \xi_m\}$ be a batch of i.i.d. samples and $B_{(i)} = B \setminus \{\xi_i\} \cup \{\xi_i'\}$ by replacing ξ_i with an i.i.d. copy ξ_i', and $B' = \{\xi_1', \xi_2', \ldots, \xi_m'\}$. Let $h(\cdot, \xi)$ be a stochastic model function, and denote $h(y, B) = \frac{1}{m} \sum_{i=1}^m h(y, \xi_i)$. The stochastic proximal mapping associated with $h(\cdot, B)$ is defined by $\mathrm{prox}_{\rho h}(x, B) \triangleq \mathrm{argmin}_{y \in \mathcal{X}} \left\{ h(y, B) + \frac{1}{2\rho} \|y - x\|^2 \right\}$ for some $\rho > 0$. We denote $x_B^+ \triangleq \mathrm{prox}_{\rho h}(x, B)$ for brevity. We say that the stochastic proximal mapping $\mathrm{prox}_{\rho h}$ is *ε-stable* if, for any $x \in \mathcal{X}$, we have

$$\left| \mathbb{E}_{B, B', i} \left[h(x_{B_{(i)}}^+, \xi_i') - h(x_B^+, \xi_i') \right] \right| \leq \varepsilon, \tag{7}$$

where i is an index chosen from $\{1, 2, \ldots, m\}$ uniformly at random.

The next lemma exploits the stability of proximal mapping associated with the model function.

Lemma 3.1. *Let $f_z(\cdot, B)$ be a stochastic model function under the assumptions A1-A4. For $\gamma \in (\lambda, \infty)$, vectors z and y, the proximal mapping $\mathrm{prox}_{f_z/\gamma}(y, B) = \mathrm{argmin}_{x \in \mathcal{X}} \left\{ f_z(x, B) + \frac{\gamma}{2} \|x - y\|^2 \right\}$ is ε-stable with $\varepsilon = \frac{2L^2}{m(\gamma - \lambda)}$.*

Applying Lemma 3.1, we obtain the error bound for approximating the full model function in the next theorem.

Theorem 3.2. *Under all the assumptions of Lemma 3.1, we have*

$$\left| \mathbb{E}_{B_k} \left[f_{x^k}(x^{k+1}, B_k) - \mathbb{E}_\xi f_{x^k}(x^{k+1}, \xi) | \sigma_k \right] \right| \leq \varepsilon_k, \ \varepsilon_k = \frac{2L^2}{m_k(\gamma_k - \lambda)}. \tag{8}$$

where σ_k is the σ-algebra generating $\{B_i\}_{1 \leq i \leq k-1}$.

Note that since x^{k+1} is dependent on B_k, $f_{x^k}(x^{k+1}, B_k)$ is not an unbiased estimator of $\mathbb{E}_\xi[f_{x^k}(x^{k+1}, \xi)]$. However, the stability argument identifies that the expected approximation error is a decreasing function of batch size m_k. This observation is the key to the sharp analysis of minibatch stochastic algorithms. With all the tools at our hands, we obtain the key descent property in the following theorem.

Theorem 3.3. *Suppose that $\rho > \lambda + \tau$, $\gamma_k \geq \rho + \tau$, A5 and all the assumptions in Lemma 3.1 hold. Let $\mathbb{E}_k[\cdot]$ abbreviates $\mathbb{E}_{B_k}[\cdot | \sigma_k]$ and ε_k be given by (8), then we have*

$$\frac{(\rho - \lambda - \tau)}{\rho(\gamma_k + \rho - 2\lambda - \tau)} \|\nabla f_{1/\rho}(x^k)\|^2 \leq f_{1/\rho}(x^k) - \mathbb{E}_k \left[f_{1/\rho}(x^{k+1}) \right] + \frac{\rho \varepsilon_k}{\gamma_k + \rho - 2\lambda - \tau}. \tag{9}$$

Next, we specify the rate of convergence to stationarity using a constant stepsize policy.

Theorem 3.4. *Under the assumptions of Theorem 3.3, let $\Delta = f_{1/\rho}(x^1) - \min_x f(x)$, $m_k = m$, and $\gamma_k = \gamma = \max\{\rho + \tau, \lambda + \eta\}$ where $\eta = \frac{\sqrt{K}}{\alpha_0 \sqrt{m}}$ and $\alpha_0 \in (0, \infty)$. Let k^* be an index chosen in $\{1, 2, \ldots, K\}$ uniformly, then we have*

$$\mathbb{E}\left[\|\nabla f_{1/\rho}(x^{k^*})\|^2 \right] \leq \frac{\rho}{\rho - \lambda - \tau} \left[\frac{(2\rho - \lambda)\Delta}{K} + \left(\frac{\Delta}{\alpha_0} + 2\alpha_0 \rho L^2 \right) \frac{1}{\sqrt{mK}} \right]. \tag{10}$$

Remark 3. The performance of SMOD depends on α_0 and batch size m. (10) implies that when batch size is fixed, the best rate is obtained at $\alpha_0^* = \sqrt{\frac{\Delta}{2\rho}} \frac{1}{L}$. Since both Δ and L are unknown,

Algorithm 2 Stochastic Extrapolated Model-Based Method (SEMOD)

Input: x^0, x^1, β, γ;
for $k = 1$ **to** K **do**
Sample data ξ^k and update:

$$y^k = x^k + \beta(x^k - x^{k-1}) \tag{11}$$

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ f_{x^k}(x, \xi^k) + \frac{\gamma}{2}\|x - y^k\|^2 \right\} \tag{12}$$

end for

hyper-parameter tuning over α_0 is required to obtain good empirical performance. For the simplicity of theoretical analysis, let us take $\alpha_0 = \alpha_0^*$. Hence, to obtain an iterate whose Moreau envelop has expected gradient norm smaller than ε, the total iteration count is $\mathcal{T}_\varepsilon = \max\left\{\mathcal{O}(\frac{\Delta}{\varepsilon^2}), \mathcal{O}(\frac{L^2\Delta}{m\varepsilon^4})\right\}$. For small batch size m (i.e. $m = o(1/\varepsilon^2)$), the second term in $\max(,)$ dominates the bound $\mathcal{T}_\varepsilon$, yielding a total complexity of $\mathcal{O}(\frac{L^2\Delta}{m\varepsilon^4})$. Note that this complexity bound is better than the $\mathcal{O}(\frac{L^2\Delta}{\varepsilon^4})$ bound [10] by a factor of m.

Remark 4. Theorem 3.4 implies that SGD can be accelerated by minibatching on the smooth composite problems (3) but leaves out the more general problems where $\ell(x, \xi)$ is non-smooth and weakly convex. In the latter case, showing any improved rate of minibatch SGD is substantially more challenging. Without additional knowledge, the $\mathcal{O}(\frac{L^2\Delta}{\varepsilon^4})$ complexity of SGD already matches the best result for deterministic subgradient method (c.f. [10]). It remains unknown whether such $\mathcal{O}(1/\varepsilon^4)$ bound is tight or not, and a possible direction to obtain sharper complexity bound is by exploiting the non-smooth structure information such as sharpness.

Solving the subproblems. SGD is embarrassingly parallelizable by simply averaging the stochastic subgradients. We highlight how to solve the proximal subproblems for SPL and SPP. Consider the composition function $f(x, \xi) = h(C(x, \xi))$ where $h(a) = |a|$. For SPL, it is easy to transform the corresponding subproblem to an $\mathcal{O}(m_k)$-dimensional quadratic program (QP) in the dual space (e.g. [3]). The dual QP can be efficiently solved in parallel, for example, by a fast interior point solver. For SPP, we show that the subproblem can be solved by a deterministic prox-linear method at a rapid linear convergence rate. Note that the SPP subproblem is especially well-conditioned because our stepsize policy ensures a large strongly convex parameter $\gamma - \lambda$. We refer to the appendix for more technical details.

4 SMOD with momentum

We present a new model-based method by incorporating an additional extrapolation term, and we record this stochastic extrapolated model-based method in Algorithm 2. Each iteration of Algorithm 2 consists of two steps, first, an extrapolation step is performed to get an auxiliary update y^k. Then a random sample ξ_k is collected and the proximal mapping, associated with the model function $f_{x^k}(\cdot, \xi_k)$, is computed at y^k to obtain the new point x^{k+1}. For ease of exposition, we take constant values of stepsize and extrapolation term.

Note that Algorithm 2 can be interpreted as an extension of the momentum SGD by replacing the gradient descent step with a broader class of proximal mappings. To see this intuition, we combine (11) and (12) to get

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ f_{x^k}(x, \xi^k) + \gamma\beta\langle x^{k-1} - x^k, x - x^k\rangle + \frac{\gamma}{2}\|x - x^k\|^2 \right\}, \tag{13}$$

If we choose the linear model (3), i.e., $f_{x^k}(x, \xi^k) = f(x^k, \xi^k) + \langle f'(x^k, \xi^k), x - x^k\rangle$, and assume $\mathcal{X} = \mathbb{R}^d$, then the update (13) has the following form:

$$x^{k+1} = x^k - \gamma^{-1} f'(x^k, \xi^k) - \beta(x^{k-1} - x^k). \tag{14}$$

Define $v^k \triangleq \gamma(x^{k-1} - x^k)$ and apply it to (14), then Algorithm 2 reduces to the heavy-ball method

$$v^{k+1} = f'(x^k, \xi^k) + \beta v^k, \tag{15}$$

$$x^{k+1} = x^k - \gamma^{-1} v^{k+1}. \tag{16}$$

Despite such relation, the gradient averaging view (15) only applies to SGD for unconstrained optimization, which limits the use of standard analysis of heavy-ball method ([34]) for our problem. To overcome this issue, we present a unified convergence analysis which can deal with all the model functions and is amenable to both constrained and composite problems.

Our theoretical analysis of Algorithm 2 relies on a different potential function from the one in the previous section. Let us define the auxiliary variable

$$z^k \triangleq x^k + \frac{\beta}{1-\beta}(x^k - x^{k-1}). \tag{17}$$

The following lemma proves some approximate descent property by adopting the potential function $f_{1/\rho}(z^k) + \frac{\rho(\gamma\beta+\rho\beta^2\theta^{-2})}{2(\gamma\theta-\lambda\theta)}\|x^k - x^{k-1}\|^2$ and measuring the quantity of $\|\nabla f_{1/\rho}(z^k)\|$.

Lemma 4.1. *Assume that $\rho \geq 2(\tau+\lambda)$ and $\beta \in [0,1)$. Let $\theta = 1-\beta$. Then we have*

$$\begin{aligned}\frac{(\rho-\lambda\theta)}{2\rho(\gamma\theta-\lambda\theta)}\|\nabla f_{1/\rho}(z^k)\|^2 \leq f_{1/\rho}(z^k) - \mathbb{E}_k\big[f_{1/\rho}(z^{k+1})\big] + \frac{\rho L^2}{(\gamma\theta^2-\rho\beta^2\theta^{-1})(\gamma\theta^2-\lambda\theta^2)} \\ + \frac{\rho(\gamma\beta+\rho\beta^2\theta^{-2})}{2(\gamma\theta-\lambda\theta)}\big(\|x^k - x^{k-1}\|^2 - \mathbb{E}_k[\|x^{k+1}-x^k\|^2]\big) \\ - \frac{\rho(\gamma-\rho\beta^2\theta^{-3})}{4(\gamma-\lambda)}\mathbb{E}_k[\|x^{k+1}-x^k\|^2].\end{aligned} \tag{18}$$

Invoking Lemma 4.1 and specifying the stepsize policy, we obtain the main convergence result of Algorithm 2 in the following theorem.

Theorem 4.2. *Under assumptions of Lemma 4.1, if we choose $x^1 = x^0$, and set $\gamma = \gamma_0\theta^{-1}\sqrt{K} + \lambda + \rho\beta^2\theta^{-3}$ for some $\gamma_0 > 0$, then*

$$\mathbb{E}[\|\nabla f_{1/\rho}(z^{k^*})\|^2] \leq \frac{2\rho}{\rho-\lambda}\left[\frac{\rho\beta^2\theta^{-2}\Delta}{K} + \Big(\gamma_0\Delta + \frac{\rho L^2}{\theta\gamma_0}\Big)\frac{1}{\sqrt{K}}\right] \tag{19}$$

where k^* is an index chosen in $\{1,2,\ldots,K\}$ uniformly at random.

Remark 5. Despite the fact that convergence is established for all $\gamma_0 > 0$, we can see that the optimal γ_0 would be $\gamma_0 = \sqrt{\frac{\rho}{\Delta\theta}}L$, which gives the bound $\mathbb{E}[\,\|\nabla f_{1/\rho}(z^{k^*})\|^2] \leq \frac{2\rho}{\rho-\lambda}\big(\frac{\rho\beta^2\theta^{-2}\Delta}{K} + 2L\sqrt{\frac{\rho\Delta}{\theta K}}\big)$. In practice, we can set γ_0 to a suboptimal value and obtain a possibly loose upper-bound.

Remark 6. Since z^k is an extrapolated solution, it may not be feasible. It is desirable to show optimality guarantee at iterates x^k. Note that using Lemma 4.1 and the parameters in Theorem 4.2, it is easy to show that $\mathbb{E}[\|x^{k^*} - x^{k^*-1}\|^2] = \mathcal{O}(\frac{1}{K})$. Based on (17) we have $\|z^{k^*} - x^{k^*}\|^2 = \beta^2\theta^{-2}\mathbb{E}[\|x^{k^*} - x^{k^*-1}\|^2] = \mathcal{O}(\frac{1}{K})$. Using Lipschitz smoothness of Moreau envelop, we can show $\mathbb{E}[\|\nabla f_{1/\rho}(x^{k^*})\|^2]$ converges at the same $\mathcal{O}(\frac{1}{\sqrt{K}})$ rate as is shown in Theorem 4.2.

Combining momentum and minibatching, we develop a minibatch version of Algorithm 2 that takes a batch of samples B_k in each iteration. The convergence analysis of this minibatch SEMOD is more involving. We leave the details in the Appendix but informally state the main result below.

Theorem 4.3 (Informal)**.** *In the minibatch SEMOD, suppose that A5 holds, the batch size $|B_k| = m$ and $\gamma = \mathcal{O}(\sqrt{\frac{K}{m}})$, then $\mathbb{E}[\|\nabla f_{1/\rho}(z^{k^*})\|^2] = \mathcal{O}\big(\frac{1}{K} + \sqrt{\frac{1}{mK}}\big)$.*

5 SMOD for convex optimization

Besides the study on non-convex optimization, we also apply model-based methods to stochastic convex optimization. Due to the space limit, we highlight main theoretical results but defer all the technical details to the Appendix section. We show that if certain assumption adapted from A5 for the convex setting holds, then the function gap of minibatching SEMOD converges at a rate of $\mathcal{O}\Big(\frac{1}{K} + \frac{1}{\sqrt{mK}}\Big)$. In view of this result, the deterministic part of our rate is consistent with the best $\mathcal{O}(\frac{1}{K})$ rate for the heavy-ball method. For example, see [13, 18]. Moreover, the stochastic part of the rate is improved from the $\mathcal{O}(\frac{1}{\sqrt{K}})$ rate of Theorem 4.4 [10] by a factor of $\sqrt{m}$.

An important question arises naturally: Can we further improve the convergence rate of model-based methods for stochastic convex optimization? Due to the widely known limitation of heavy-ball type momentum, it would be interesting to consider Nesterov's acceleration. To this end, we present a model-based method with Nesterov type momentum. Thanks to the stability argument, we obtain the following improved rate of convergence: $\mathcal{O}\Big(\frac{1}{K^2}+\frac{1}{\sqrt{mK}}\Big)$. We note that a similar convergence rate for minibatching model-based methods is obtained in a recent paper [8]. However, their result requires the assumption that the stochastic function is Lipschitz smooth while our assumption is much weaker.

6 Experiments

In this section, we examine the empirical performance of our proposed methods through experiments on the problem of robust phase retrieval. (Additional experiments on blind deconvolution are given in Appendix section). Given a set of vectors $a_i \in \mathbb{R}^d$ and nonnegative scalars $b_i \in \mathbb{R}_+$, the goal of phase retrieval is to recover the true signal x^* from the measurement $b_i = |\langle a_i, x^*\rangle|^2$. Due to the potential corruption in the dataset, we consider the following penalized formulation

$$\min_{x\,\in\,\mathbb{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n}\big|\langle a_i, x\rangle^2 - b_i\big| \tag{20}$$

where we impose ℓ_1-loss to promote robustness and stability (cf. [16, 10, 27]).

Data Preparation. We conduct experiments on both synthetic and real datasets.

1) Synthetic data. Synthetic data is generated following the setup in [27]. We set $n = 300, d = 100$ and select x^* from unit sphere uniformly at random. Moreover, we generate $A = QD$ where $Q \in \mathbb{R}^{n\times d}, q_{ij} \sim \mathcal{N}(0,1)$ and $D \in \mathbb{R}^d$ is a diagonal matrix whose diagonal entries are evenly distributed in $[1/\kappa, 1]$. Here $\kappa \geq 1$ plays the role of condition number (large κ makes problem hard). The measurements are generated by $b_i = \langle a_i, x^*\rangle^2 + \delta_i\zeta_i$ $(1 \leq i \leq n)$ with $\zeta_i \sim \mathcal{N}(0, 25)$, $\delta_i \sim \mathrm{Bernoulli}(p_{\text{fail}})$, where $p_{\text{fail}} \in [0, 1]$ controls the fraction of corrupted observations on expectation.

2) Real data. We consider `zipcode`, a dataset of 16×16 handwritten digits collected from [22]. Following the setup in [16], let $H \in \mathbb{R}^{256\times 256}$ be a normalized Hadamard matrix such that $h_{ij} \in \big\{\frac{1}{16}, -\frac{1}{16}\big\}, H = H^{\mathrm{T}}$ and $H = H^{-1}$. Then we generate $k = 3$ diagonal sign matrices S_1, S_2, S_3 such that each diagonal element of S_k is uniformly sampled from $\{-1, 1\}$. Last we set $A = [HS_1, HS_2, HS_3]^{\mathrm{T}} \in \mathbb{R}^{(3\times 256)\times 256}$. As for the true signal and measurements, each image is represented by a data matrix $X \in \mathbb{R}^{16\times 16}$ and gets vectorized to $x^* = \mathrm{vec}(X)$. To simulate the case of corruption, we set measurements $b = \phi_{p_{\text{fail}}}(Ax^*)$, where $\phi_{p_{\text{fail}}}(\cdot)$ denotes element-wise squaring and setting a fraction p_{fail} of entries to 0 on expectation.

In the first experiment, we illustrate that `SMOD` methods enjoy linear speedup in the size of minibatches and exhibit strong robustness to the stepsize policy. We conduct comparison on `SPL` and `SGD` and describe the detailed experiment setup as follows.

1) Dataset generation. We generate four testing cases: the synthetic datasets with $(\kappa, p_{\text{fail}}) = (10, 0.2)$, and $(10, 0.3)$; `zipcode` with digit images of id 2 and 24;

2) Initial point. We set the initial point $x^1(= x^0) \sim \mathcal{N}(0, I_d)$ for synthetic data and $x^1 = x^* + \mathcal{N}(0, I_d)$ for `zipcode`;

3) Stopping criterion. We set the stopping criterion to be $f(x^k) \leq 1.5\hat{f}$, where $\hat{f} = f(x^*)$ is the corrupted objective evaluated at the true signal x^*;

3) Stepsize. We set the parameter $\gamma = \alpha_0^{-1}\sqrt{K/m}$ where m is the batch size; For synthetic dataset, we test 10 evenly spaced α_0 values in range $[10^{-1}, 10^2]$ on logarithmic scale, and for `zipcode` dataset we set such range of α_0 to $[10^1, 10^3]$;

4) Maximum iteration. We set the maximum number of epochs to be 200 and 400 respectively for minibatch and momentum related tests;

5) Batch size. We take minibatch size m from the range $\{1, 4, 8, 16, 32, 64\}$;

6) Sub-problems The solution to the proximal sub-problems is left in the appendix.

For each algorithm, speedup from minibatching is quantified as T_1^*/T_m^* where T_m^* is the total number of iterations for reaching the desired accuracy, with batch size m and the best initial stepsize α_0 among values specified above. Specially, if an algorithm fails to reach desired accuracy after running out of 400 epochs, we set its iteration number to the maximum.

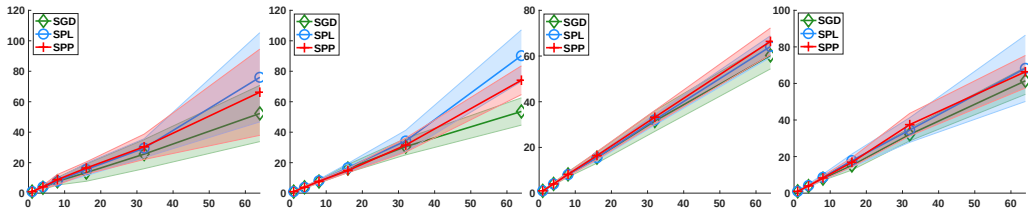

Figure 1: Speedup over minibatch sizes. The left two are for synthetic datasets $\kappa = 10, p_{\text{fail}} \in \{0.2, 0.3\}$; Digit datasets: digit image (id:24) with $p_{\text{fail}} \in \{0.2, 0.3\}$.

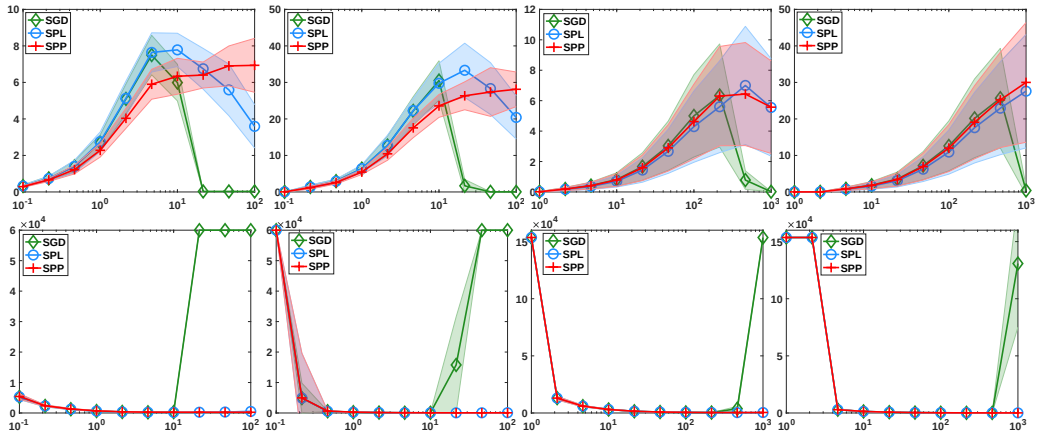

Figure 2: From left to right: synthetic datasets with $m \in \{8, 32\}$ and `zipcode` image (id=24) with $m \in \{8, 32\}$. x-axis: initial stepsize α_0. y-axis (first row): speedup over the sequential version: $T_1^*/T_m^*(\alpha_0)$ where $T_m^*(\alpha_0)$ stands for the number of iterations when using batch size m and initial stepsize α_0. y-axis (second row): Total number of iterations.

Figure 1 plots the speedup of each algorithm over different values of batch size according to the average of 20 independent runs. It can be seen that SPL exhibits a linear acceleration over the batch size, which confirms our theoretical analysis. Moreover, we find SGD admits considerable acceleration using minibatches, and sometimes the speedup performance matches that of SPL and SPP. This observation seems to suggest the effectiveness of minibatch SGD in practice, despite the lack of theoretical support.

Next, we investigate the sensitivity of minibatch acceleration to the choice of initial stepsizes. We plot the algorithm speedup over the initial stepsize α_0 in Figure 2 (1st row). It can be readily seen that SGD, SPL and SPP all achieve considerable minibatch acceleration when choosing the initial stepsize properly. However, SPL and SPP enjoy a much wider range of initial stepsizes for good speedup performance, and hence, lays more robust performance than SGD. To further illustrate the robustness of SPL and SPP, we compare the efficiency of both algorithms in the minibatch setting. In contrast to the previous comparison on the relative scale, we directly compare the iteration complexity of the two algorithms. We plot the total iteration number over the choice of initial stepsizes in Figure 2 (2nd row) for batch size $m = 8$ and 32. We observe that minibatch SPL(SPP)s exhibits promising performance for a wide range of stepsize policies, while minibatch SGD quickly diverges for large stepsizes. Overall, our experiment complements the recent work [10], which shows that SPL (SPP) is more robust than SGD in the sequential setting.

Our second experiment investigates the performance of the proposed momentum methods. We compare three model-based methods (SGD, SPL, SPP) and extrapolated model-based methods (SEGD, SEPL, SEPP). We generate four testing cases: the synthetic datasets with $(\kappa, p_{\text{fail}}) = (10, 0.2)$ and

$(10, 0.3)$; `zipcode` with digit images of id 2 and $p_{\text{fail}} \in \{0.2, 0.3\}$. We set $\alpha_0 \in [10^{-2}, 10^0], \beta = 0.6$ for synthetic data, and set $\alpha_0 \in [10^0, 10^1], \beta = 0.9$ for `zipcode` dataset. The rest of settings are the same as in minibatch with $m = 1$.

Figure 3 plots the number of epochs to ε-accuracy over initial stepsize α_0. It can be seen that with properly selected momentum parameters (`SEGD`, `SEPL`, `SEPP`) all suggest improved convergence when stepsize is relatively small.

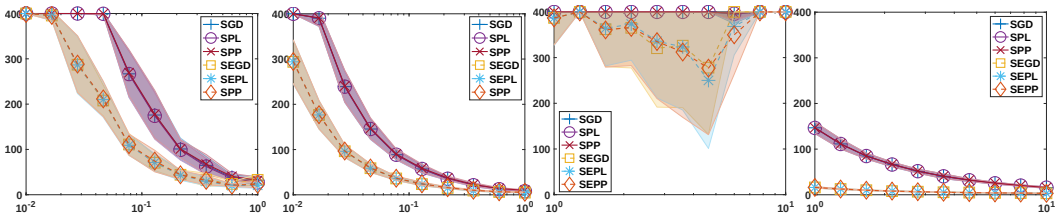

Figure 3: From left to right: synthetic datasets with $\kappa = 10, p_{\text{fail}} \in \{0.2, 0.3\}, \beta = 0.6$ and `zipcode` image (id=2) with $p_{\text{fail}} \in \{0.2, 0.3\}, \beta = 0.9$. x-axis: initial stepsize α_0. y-axis: number of epochs on reaching desired accuracy

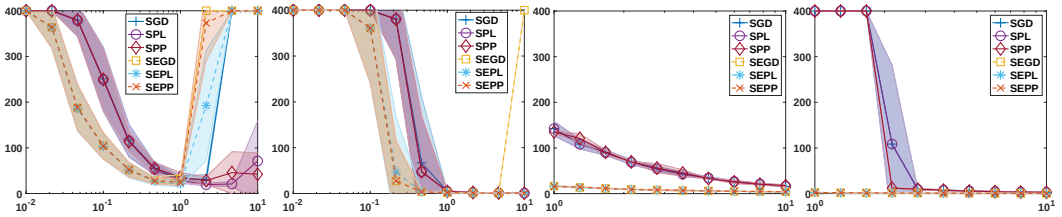

Figure 4: From left to right: synthetic datasets with $\kappa = 10, p_{\text{fail}} = 0.2, \beta = 0.6, m \in \{1, 32\}$ and `zipcode` image (id=24) with $p_{\text{fail}} = 0.3, \beta = 0.9, m \in \{1, 32\}$. x-axis: initial stepsize α_0. y-axis: number of epochs for reaching desired accuracy

In the last experiment, we attempt to exploit the performance of the compared algorithms when minibatching and momentum are applied simultaneously. The parameter setting is the same as that of the second experiment, except that we choose $m \in \{1, 32\}$. Results are plotted in Figure 4 and it can be seen that minibatch, when combined with momentum, exhibits even better convergence and robustness.

7 Discussion

On a broad class of non-smooth non-convex (particularly, weakly convex) problems, we make stochastic model-based methods more efficient by leveraging minibatching and momentum—two techniques that are well-known only for `SGD`. Applying algorithm stability for optimization analysis is a key step to achieving improved convergence rate over the batch size. This perspective appears to be interesting for stochastic optimization in a much broader context. Although some progress is made, we are unable to show whether minibatches can accelerate `SGD` when the objective does not have a smooth component. Note that the complexity of `SGD` already matches the best bound of full subgradient method. It would be interesting to know whether this bound for `SGD` is tight or improvable. It would also be interesting to study the lower bound of `SGD` (and other stochastic algorithms) in the non-smooth setting. Some interesting recent results can be referred from [23, 36].

8 Acknowledgement and disclosure of funding

The authors are grateful to the Area Chairs and the anonymous reviewers for their constructive suggestions. QD was partially supported by National Natural Science Foundation of China (Grant 11831002, 72150001).

References

[1] H. Asi and J. C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *Siam Journal on Optimization*, 29(3):2257–2290, 2019.

[2] H. Asi and J. C. Duchi. The importance of better models in stochastic optimization. *Proceedings of the National Academy of Sciences*, 116(46):22924–22930, 2019.

[3] H. Asi, K. Chadha, G. Cheng, and J. C. Duchi. Minibatch stochastic approximate proximal point methods. *Advances in Neural Information Processing Systems*, 33, 2020.

[4] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

[5] L. Berrada, A. Zisserman, and M. P. Kumar. Deep frank-wolfe for neural network optimization. In *ICLR 2019 : 7th International Conference on Learning Representations*, 2019.

[6] A. Botev, H. Ritter, and D. Barber. Practical gauss-newton optimisation for deep learning. In *International Conference on Machine Learning*, pages 557–565. PMLR, 2017.

[7] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[8] K. Chadha, G. Cheng, and J. C. Duchi. Accelerated, optimal, and parallel: Some results on model-based stochastic optimization. *arXiv preprint arXiv:2101.02696*, 2021.

[9] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *arXiv preprint arXiv:1904.10020*, 2019.

[10] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *Siam Journal on Optimization*, 29(1):207–239, 2019.

[11] A. Defazio. Understanding the role of momentum in non-convex optimization: Practical insights from a lyapunov analysis. *arXiv preprint arXiv:2010.00406*, 2020.

[12] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1), 2012.

[13] J. Diakonikolas and M. I. Jordan. Generalized momentum-based methods: a hamiltonian perspective. *SIAM Journal on Optimization*, 31(1):915–944, 2021.

[14] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, pages 1–56, 2018.

[15] J. C. Duchi and F. Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

[16] J. C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2019.

[17] T. Frerix, T. Möllenhoff, M. Moeller, and D. Cremers. Proximal backpropagation. In *International Conference on Learning Representations*, 2018.

[18] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.

[19] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. ISSN 1052-6234.

[20] I. Gitman, H. Lang, P. Zhang, and L. Xiao. Understanding the role of momentum in stochastic gradient methods. In *Advances in Neural Information Processing Systems*, volume 32, pages 9633–9643, 2019.

[21] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

[22] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

[23] G. Kornowski and O. Shamir. Oracle complexity in nonsmooth nonconvex optimization. *arXiv preprint arXiv:2104.06763*, 2021.

[24] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

[25] Y. Liu, Y. Gao, and W. Yin. An improved analysis of stochastic gradient descent with momentum. *arXiv preprint arXiv:2007.07989*, 2020.

[26] N. Loizou and P. Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710, 2020.

[27] V. Mai and M. Johansson. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6630–6639, 2020.

[28] 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.

[29] O. Sebbouh, R. M. Gower, and A. Defazio. On the convergence of the stochastic heavy ball method. *arXiv preprint arXiv:2006.07867*, 2020.

[30] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

[31] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147, 2013.

[32] M. Takáč, P. Richtárik, and N. Srebro. Distributed mini-batch sdca. *arXiv preprint arXiv:1507.08322*, 2015.

[33] J. Wang, W. Wang, and N. Srebro. Memory and communication efficient distributed stochastic optimization with minibatch prox. In *Conference on Learning Theory*, pages 1882–1919. PMLR, 2017.

[34] Y. Yan, T. Yang, Z. Li, Q. Lin, and Y. Yang. A unified analysis of stochastic momentum methods for deep learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 2955–2961, 2018.

[35] J. Zhang and L. Xiao. Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization. *Mathematical Programming*, pages 1–43, 2021.

[36] J. Zhang, H. Lin, S. Jegelka, A. Jadbabaie, and S. Sra. Complexity of finding stationary points of nonsmooth nonconvex functions. *arXiv preprint arXiv:2002.04130*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See remarks 3, 4 in Section 3 and Discussion section
   (c) Did you discuss any potential negative societal impacts of your work? [No] The paper addresses theoretical questions on algorithm complexity, which, to the best of our knowledge, has no negative social impact
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See assumptions in Section 2 and 3
   (b) Did you include complete proofs of all theoretical results? [Yes] Proof is left in the appendix
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code is supplied in the supplemental materials

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 6 for details of experiments

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] From the experiments the error bars are relatively thin and the results are presented by taking average.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] The main goal of experiments is to demonstrate our theoretical foundings, thereby only showing the iteration complexity of algorithms.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] `zipcode` dataset is referenced from [22].

(b) Did you mention the license of the assets? [No] The dataset used is published on an open site without license.

(c) Did you include any new assets either in the supplemental material or as a URL? [No] The experiments do not involve new datasets.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] An open dataset is used.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The dataset has been open for years and only involves zipcode digits.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No] No crowdsourcing or human object is involved.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No] No crowdsourcing or human object is involved.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] No crowdsourcing or human object is involved.

Appendix

Table of Contents

Contents of the appendix

In the appendix, we present additional convergence analyses of the proposed algorithms. Appendix A and B respectively give the convergence results for minibatching and momentum SMOD. Convergence results of SMOD with both minibatching and momentum are formally presented in Appendix B.3. Besides the missing proofs from the main article, in Appendix C we also give some new results of SMOD for stochastic convex optimization, and show how to achieve and to possibly improve the state-of-the-art complexity rates. Particularly, SMOD with Nesterov acceleration, which achieves the best complexity rate, is developed in Appendix C.3. Last, we provide details on how to solve the subproblems from the experiments in Section D. Additional experiments on blind deconvolution are available in Appendix E.

A Proof of results in Section 3

Our paper will make use of the following elementary result, we refer to [3] for proof details.

Lemma A.1. *A function $f(x)$ is λ-weakly convex if and only if for any x, y and $f'(x) \in \partial f(x)$, we have $f(y) \geq f(x) + \langle f'(x), y - x\rangle - \frac{\lambda}{2}\|y - x\|^2$.*

We state an important result which generalizes the well-known three-point lemma to handle nonconvex function.

Lemma A.2. *Let $g(x)$ be a η-weakly convex function, and $\kappa > \eta$. If*

$$z^+ = \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ g(x) + \frac{\kappa}{2}\|x - z\|^2 \right\},$$

then for any $x \in \mathcal{X}$, we have

$$g(z^+) + \frac{\kappa}{2}\|z^+ - z\|^2 \leq g(x) + \frac{\kappa}{2}\|x - z\|^2 - \frac{\kappa - \eta}{2}\|x - z^+\|^2. \tag{21}$$

Proof. Since $g(x)$ is η-weakly convex, $g(x) + \frac{\kappa}{2}\|x - z\|^2 = \big[g(x) + \frac{\eta}{2}\|x - z\|^2\big] + \frac{\kappa-\eta}{2}\|x - z\|^2$ is strongly convex with parameter $\kappa - \eta$. Using the optimality condition $0 \in \partial\big[g(z^+) + \frac{\kappa}{2}\|z^+ - z\|^2\big]$ and strong convexity of $g(\cdot) + \frac{\kappa}{2}\|\cdot - z\|^2$, we immediately obtain

$$g(x) + \frac{\kappa}{2}\|x - z\|^2 \geq g(z^+) + \frac{\kappa}{2}\|z^+ - z\|^2 + \langle 0, x - z^+\rangle + \frac{\kappa - \eta}{2}\|x - z^+\|^2.$$

□

Before getting down to the proof, first recall that in Section 3, we let $B = \{\xi_1, \xi_2, \ldots, \xi_m\}$ be the i.i.d. samples and $B_{(i)} = \{\xi_1, \ldots, \xi_{i-1}, \xi_i', \xi_{i+1} \ldots, \xi_m\}$ by replacing ξ_i with an i.i.d. copy ξ_i'. We denote $B' = \{\xi_1', \xi_2', \ldots, \xi_{m-1}', \xi_m'\}$.

A.1 Proof of Lemma 3.1

For brevity, for $i = 1, 2, \ldots, m$, we denote

$$\hat{y} = \arg\min_{x\in\mathcal{X}} \left\{ f_z(x, B) + \frac{\gamma}{2}\|x - y\|^2 \right\},$$
$$\hat{y}_i = \arg\min_{x\in\mathcal{X}} \left\{ f_z(x, B_{(i)}) + \frac{\gamma}{2}\|x - y\|^2 \right\}.$$

Using triangle inequality and Jensen's inequality, we deduce

$$\begin{aligned}
&\big|\mathbb{E}_{B,B',i}\big[f_z(\hat{y}_i, \xi_i') - f_z(\hat{y}, \xi_i')\big]\big| \\
&= \Big|\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{B,\xi_i'}\big[f_z(\hat{y}_i, \xi_i') - f_z(\hat{y}, \xi_i')\big]\Big| \\
&\leq \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{B,\xi_i'}\big|f_z(\hat{y}_i, \xi_i') - f_z(\hat{y}, \xi_i')\big| \\
&\leq \frac{L}{m}\sum_{i=1}^m \mathbb{E}_{B,\xi_i'}\|\hat{y}_i - \hat{y}\|,
\end{aligned} \tag{22}$$

where the last inequality follows from A4.

Next we bound $\|\hat{y} - \hat{y}_i\|$. Due to λ-weak convexity of $f_z(x, B)$ and by Lemma A.2, for any $i \in \{1, 2, \ldots, m\}$, we obtain

$$f_z(\hat{y}, B) + \frac{\gamma}{2}\|\hat{y} - y\|^2 \leq f_z(\hat{y}_i, B) + \frac{\gamma}{2}\|\hat{y}_i - y\|^2 - \frac{\gamma - \lambda}{2}\|\hat{y}_i - \hat{y}\|^2,$$
$$f_z(\hat{y}_i, B_{(i)}) + \frac{\gamma}{2}\|\hat{y}_i - y\|^2 \leq f_z(\hat{y}, B_{(i)}) + \frac{\gamma}{2}\|\hat{y} - y\|^2 - \frac{\gamma - \lambda}{2}\|\hat{y}_i - \hat{y}\|^2.$$

Summing up the above two relations, we deduce that

$$\begin{aligned}&(\gamma-\lambda)\|\hat{y}_i-\hat{y}\|^2\\&\le f_z(\hat{y},B_{(i)})-f_z(\hat{y},B)+f_z(\hat{y}_i,B)-f_z(\hat{y}_i,B_{(i)})\\&=\frac{1}{m}\big[f_z(\hat{y},\xi_i')-f_z(\hat{y}_i,\xi_i')+f_z(\hat{y}_i,\xi_i)-f_z(\hat{y},\xi_i)\big]\end{aligned}\tag{23}$$

Next, we use Assumption A4 and (23) to obtain $(\gamma-\lambda)\|\hat{y}_i-\hat{y}\|^2\le\frac{2L}{m}\|\hat{y}_i-\hat{y}\|$, which implies that

$$\|\hat{y}_i-\hat{y}\|\le\frac{2L}{m(\gamma-\lambda)}.\tag{24}$$

In view of (22) and (24), we have

$$\big|\mathbb{E}_{B,B',i}\big[f_z(\hat{y}_i,\xi_i')-f_z(\hat{y},\xi_i')\big]\big|\le\frac{2L^2}{m(\gamma-\lambda)}=\varepsilon.$$

A.2 Proof of Theorem 3.2

Theorem 3.2 is an immediate consequence of Lemma 3.1 and the following theorem which indicates that stability bounds the error of approximating the full model function on expectation.

Theorem A.3. *Assume that* $\mathrm{prox}_{\rho h}(\cdot,\cdot)$ *is* ε*-stable and denote* $x_B^+=\mathrm{prox}_{\rho h}(x,B)$*. Then, we have*

$$\big|\mathbb{E}_B\big\{h\big(x_B^+,B\big)-\mathbb{E}_\xi\big[h\big(x_B^+,\xi\big)\big]\big\}\big|\le\varepsilon.$$

Proof of Theorem A.3 The proof resembles the argument of Lemma 11 [7]. For brevity we denote $\hat{x}=\mathrm{prox}_{\rho h}(x,B)$ and $\hat{x}_i=\mathrm{prox}_{\rho h}(x,B_{(i)})$. Since ξ_i' is independent of B, we have $\mathbb{E}_\xi\big[h\big(\hat{x},\xi\big)\big]=\mathbb{E}_{\xi_i'}\big[h\big(\hat{x},\xi_i'\big)\big]$ for any $i\in\{1,\dots,m\}$. Therefore, we have

$$\mathbb{E}_\xi\big[h\big(\hat{x},\xi\big)\big]=\frac{1}{m}\sum_{j=1}^m\mathbb{E}_{\xi_j'}\big[h\big(\hat{x},\xi_j'\big)\big].\tag{25}$$

Similarly, due to the independence assumption, we have

$$\mathbb{E}_B\big[h\big(\hat{x},\xi_i\big)\big]=\mathbb{E}_{B_{(i)}}\big[h(\hat{x}_i,\xi_i')\big],\tag{26}$$

which implies that

$$\mathbb{E}_B\big[h\big(\hat{x},B\big)\big]=\frac{1}{m}\sum_{i=1}^m\mathbb{E}_B\big[h(\hat{x},\xi_i)\big]=\frac{1}{m}\sum_{i=1}^m\mathbb{E}_{B_{(i)}}\big[h(\hat{x}_i,\xi_i')\big]\tag{27}$$

In view of (25) and (27), we deduce

$$\begin{aligned}&\mathbb{E}_B\Big\{h\big(\hat{x},B\big)-\mathbb{E}_\xi\big[h\big(\hat{x},\xi\big)\big]\Big\}\\&=\frac{1}{m}\sum_{i=1}^m\mathbb{E}_{B_{(i)}}\big[h(\hat{x}_i,\xi_i')\big]-\frac{1}{m}\sum_{i=1}^m\mathbb{E}_{B,\xi_i'}\big[h\big(\hat{x},\xi_i'\big)\big]\\&=\frac{1}{m}\sum_{i=1}^m\mathbb{E}_{B,\xi_i'}\big[h(\hat{x}_i,B_{(i)})-h\big(\hat{x},\xi_i'\big)\big]\\&=\mathbb{E}_{B,B',i}\big[h(\hat{x}_i,B_{(i)})-h\big(\hat{x},\xi_i'\big)\big].\end{aligned}$$

Appealing to the stability assumption, we complete the proof.

A.3 Proof of Theorem 3.3

First, due to the weak convexity of $f_{x^k}(\cdot,B_k)$ and Lemma A.2, we have

$$f_{x^k}(x^{k+1},B_k)+\frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2\le f_{x^k}(x,B_k)+\frac{\gamma_k}{2}\|x-x^k\|^2-\frac{\gamma_k-\lambda}{2}\|x^{k+1}-x\|^2,\quad\forall x\in\mathcal{X}.\tag{28}$$

For simplicity, we denote $\hat{x}^k = \mathrm{prox}_{f/\rho}(x^k) = \mathrm{argmin}_{x\in\mathcal{X}} \left\{f(x)+\frac{\rho}{2}\|x-x^k\|^2\right\}$. Then substituting $x = \hat{x}^k$ in (28), we have

$$f_{x^k}(x^{k+1}, B_k)+\frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2 \le f_{x^k}(\hat{x}^k, B_k)+\frac{\gamma_k}{2}\|\hat{x}^k-x^k\|^2-\frac{\gamma_k-\lambda}{2}\|x^{k+1}-\hat{x}^k\|^2. \quad (29)$$

Analogously, since $f(x)$ is $(\lambda+\tau)$-weakly convex, applying Lemma A.2 with $g(x) = f(x)$, $\eta = \lambda+\tau$ and $\kappa = \rho$, we have

$$f(\hat{x}^k)+\frac{\rho}{2}\|\hat{x}^k-x^k\|^2 \le f(x^{k+1})+\frac{\rho}{2}\|x^{k+1}-x^k\|^2-\frac{\rho-\lambda-\tau}{2}\|\hat{x}^k-x^{k+1}\|^2. \quad (30)$$

Summing up (29) and (30) gives

$$\begin{aligned}
&\frac{\gamma_k-\rho}{2}\|x^{k+1}-x^k\|^2+\frac{\gamma_k+\rho-2\lambda-\tau}{2}\|\hat{x}^k-x^{k+1}\|^2-\frac{\gamma_k-\rho}{2}\mathbb{E}_k\|\hat{x}^k-x^k\|^2\\
&\le f(x^{k+1})-f_{x^k}(x^{k+1}, B_k)+f_{x^k}(\hat{x}^k, B_k)-f(\hat{x}^k)\\
&=\left\{f(x^{k+1})-\mathbb{E}_\xi\left[f_{x^k}(x^{k+1},\xi)\right]\right\}+\left\{\mathbb{E}_\xi\left[f_{x^k}(x^{k+1},\xi)\right]-f_{x^k}(x^{k+1}, B_k)\right\}\\
&\quad+\left[f_{x^k}(\hat{x}^k, B_k)-f(\hat{x}^k)\right]\\
&\le \frac{\tau}{2}\|x^k-x^{k+1}\|^2+\frac{\tau}{2}\|x^k-\hat{x}^k\|^2+\mathbb{E}_\xi\left[f_{x^k}(x^{k+1},\xi)\right]-f_{x^k}(x^{k+1}, B_k), \quad (31)
\end{aligned}$$

where the last inequality uses the Assumption A5. Moreover, note that Theorem 3.2 implies

$$\mathbb{E}_k\left\{\mathbb{E}_\xi\left[f_{x^k}(x^{k+1},\xi)\right]-f_{x^k}(x^{k+1}, B_k)\right\} \le \varepsilon_k. \quad (32)$$

Taking expectation over B_k in (31) and combining the result with (32), we obtain

$$\begin{aligned}
&\frac{\gamma_k-\rho}{2}\,\mathbb{E}_k\left[\|x^{k+1}-x^k\|^2\right]+\frac{\gamma_k+\rho-2\lambda-\tau}{2}\,\mathbb{E}_k\left[\|\hat{x}^k-x^{k+1}\|^2\right]-\frac{\gamma_k-\rho}{2}\|\hat{x}^k-x^k\|^2\\
&\le \frac{\tau}{2}\,\mathbb{E}_k\left[\|x^k-x^{k+1}\|^2\right]+\frac{\tau}{2}\|\hat{x}^k-x^k\|^2+\varepsilon_k,
\end{aligned}$$

which, by rearranging terms, implies

$$\begin{aligned}
&\mathbb{E}_k\left[\|x^{k+1}-\hat{x}^k\|^2\right]\\
&\le \frac{\gamma_k-\rho+\tau}{\gamma_k+\rho-2\lambda-\tau}\|\hat{x}^k-x^k\|^2-\frac{\gamma_k-\rho-\tau}{\gamma_k+\rho-2\lambda-\tau}\mathbb{E}_k\left[\|x^k-x^{k+1}\|^2\right]+\frac{2\varepsilon_k}{\gamma_k+\rho-2\lambda-\tau}\\
&\le \|\hat{x}^k-x^k\|^2-\frac{2(\rho-\lambda-\tau)}{\gamma_k+\rho-2\lambda-\tau}\|\hat{x}^k-x^k\|^2+\frac{2\varepsilon_k}{\gamma_k+\rho-2\lambda-\tau}, \quad (33)
\end{aligned}$$

Above, the last inequality in (33) uses the assumption $\gamma_k-\rho-\tau \ge 0$.

Moreover, following the result (33) and the definition of Moreau envelope, we have

$$\begin{aligned}
&\mathbb{E}_k\left[f_{1/\rho}(x^{k+1})\right]\\
&=\mathbb{E}_k\left[f(\hat{x}^{k+1})+\frac{\rho}{2}\|\hat{x}^{k+1}-x^{k+1}\|^2\right]\\
&\le f(\hat{x}^k)+\mathbb{E}_k\left[\frac{\rho}{2}\|\hat{x}^k-x^{k+1}\|^2\right]\\
&\le f(\hat{x}^k)+\frac{\rho}{2}\|\hat{x}^k-x^k\|^2-\frac{\rho(\rho-\lambda-\tau)}{\gamma_k+\rho-2\lambda-\tau}\|\hat{x}^k-x^k\|^2+\frac{\rho\varepsilon_k}{\gamma_k+\rho-2\lambda-\tau}\\
&= f_{1/\rho}(x^k)-\frac{\rho(\rho-\lambda-\tau)}{\gamma_k+\rho-2\lambda-\tau}\|\hat{x}^k-x^k\|^2+\frac{\rho\varepsilon_k}{\gamma_k+\rho-2\lambda-\tau}.
\end{aligned}$$

Finally, applying the identity $\|\hat{x}^k-x^k\|^2 = \rho^{-2}\|\nabla f_{1/\rho}(x^k)\|^2$ and rearranging the terms, we get (9).

A.4 Proof of Theorem 3.4

First, summing up (9) over $k = 1, 2, \ldots, K$, and taking expectation over all randomness, we have

$$\begin{aligned}
&\sum_{k=1}^{K} \frac{\rho - \lambda - \tau}{\rho(\gamma_k + \rho - 2\lambda - \tau)} \mathbb{E}[\|\nabla f_{1/\rho}(x^k)\|^2] \\
&\leq f_{1/\rho}(x^1) - \mathbb{E}\big[f_{1/\rho}(x^{K+1})\big] + \sum_{k=1}^{K} \frac{\rho\varepsilon_k}{\gamma_k + \rho - 2\lambda - \tau} \\
&\leq \Delta + \sum_{k=1}^{K} \frac{\rho\varepsilon_k}{\gamma_k + \rho - 2\lambda - \tau},
\end{aligned}$$

where the second inequality uses $-f_{1/\rho}(x^{K+1}) \leq -\min_x f(x)$. Plugging in $\gamma_k = \gamma$ and $m_k = m$ in above and appealing to the definition of x^{k^*}, we have

$$\begin{aligned}
\frac{\rho - \lambda - \tau}{\rho} \mathbb{E}\big[\|\nabla f_{1/\rho}(x^{k^*})\|^2\big] &= \frac{\rho - \lambda - \tau}{\rho K} \sum_{k=1}^{K} \mathbb{E}\big[\|\nabla f_{1/\rho}(x^k)\|^2\big] \\
&\leq \frac{(\gamma + \rho - 2\lambda - \tau)\Delta}{K} + \frac{\rho}{K} \sum_{k=1}^{K} \varepsilon_k \\
&\leq \frac{(2\rho - \lambda)\Delta}{K} + \frac{\eta\Delta}{K} + \frac{2\rho L^2}{m(\gamma - \lambda)} \\
&\leq \frac{(2\rho - \lambda)\Delta}{K} + \frac{\eta\Delta}{K} + \frac{2\rho L^2}{m\eta}, \qquad (34)
\end{aligned}$$

where the second inequality uses $\gamma \leq \rho + \tau + \lambda + \eta$, the third inequality uses $\gamma - \lambda \geq \eta$. Dividing both sides of (34) by $\frac{\rho - \lambda - \tau}{\rho}$ gives (10).

B Proof of results in Section 4

B.1 Proof of Lemma 4.1

Denote $\bar{x} = \beta x^k + (1 - \beta)x$ for $x \in \mathcal{X}$. Then $\bar{x}$ is also feasible due to the convexity of $\mathcal{X}$. Noting that $\theta = 1 - \beta$, we have the following identities:

$$\bar{x} - x^k = \theta(x - x^k), \qquad (35)$$

$$\bar{x} - y^k = \theta(x - z^k), \qquad (36)$$

$$\bar{x} - x^{k+1} = \theta(x - z^{k+1}). \qquad (37)$$

Applying Lemma A.2 and using the optimality of x^{k+1}, we have

$$\begin{aligned}
&f_{x^k}(x^{k+1}, \xi^k) + \frac{\gamma}{2}\|x^{k+1} - y^k\|^2 \\
&\leq f_{x^k}(\bar{x}, \xi^k) + \frac{\gamma}{2}\|\bar{x} - y^k\|^2 - \frac{\gamma - \lambda}{2}\|x^{k+1} - \bar{x}\|^2 \\
&= f_{x^k}(\bar{x}, \xi^k) + \frac{\gamma\theta^2}{2}\|x - z^k\|^2 - \frac{(\gamma - \lambda)\theta^2}{2}\|x - z^{k+1}\|^2 \qquad (38)
\end{aligned}$$

Since $f_{x^k}(\cdot, \xi^k) + \frac{\lambda}{2}\|\cdot - x^k\|^2$ is convex, we have

$$\begin{aligned}
f_{x^k}(\bar{x}, \xi^k) &\leq (1 - \theta)\big[f_{x^k}(x^k, \xi^k)\big] + \theta\big[f_{x^k}(x, \xi^k) + \frac{\lambda}{2}\|x - x^k\|^2\big] - \frac{\lambda}{2}\|\bar{x} - x^k\|^2 \\
&\leq (1 - \theta) f(x^k, \xi^k) + \theta\big[f(x, \xi^k) + \frac{\lambda + \tau}{2}\|x - x^k\|^2\big] - \frac{\lambda\theta^2}{2}\|x - x^k\|^2 \qquad (39)
\end{aligned}$$

where the second inequality uses Assumptions A2, A3 and (35). Summing up (38) and (39), we get

$$\begin{aligned}
&f_{x^k}(x^{k+1},\xi^k)+\frac{\gamma}{2}\|x^{k+1}-y^k\|^2\\
\le\ &(1-\theta)f(x^k,\xi^k)+\theta\big[f(x,\xi^k)+\frac{\lambda+\tau}{2}\|x-x^k\|^2\big]-\frac{\lambda\theta^2}{2}\|x-x^k\|^2\\
&+\frac{\gamma\theta^2}{2}\|x-z^k\|^2-\frac{(\gamma-\lambda)\theta^2}{2}\|x-z^{k+1}\|^2
\end{aligned}\tag{40}$$

Moreover, appealing to Assumption A2 and A4, we have

$$f(x^k,\xi^k)-L\|x^{k+1}-x^k\|=f_{x^k}(x^k,\xi^k)-L\|x^{k+1}-x^k\|\le f_{x^k}(x^{k+1},\xi^k).\tag{41}$$

Next, Putting (40) and (41) together, we have

$$\begin{aligned}
&-L\|x^{k+1}-x^k\|+\frac{\gamma}{2}\|x^{k+1}-y^k\|^2\\
\le\ &-\theta f(x^k,\xi^k)+\theta\big[f(x,\xi^k)+\frac{\lambda+\tau}{2}\|x-x^k\|^2\big]-\frac{\lambda\theta^2}{2}\|x-x^k\|^2\\
&+\frac{\gamma\theta^2}{2}\|x-z^k\|^2-\frac{(\gamma-\lambda)\theta^2}{2}\|x-z^{k+1}\|^2
\end{aligned}\tag{42}$$

Denote $\hat{z}^k=\mathrm{prox}_{f/\rho}(z^k)$. Note that z^k may be infeasible, but the feasibility of $\hat{z}^k$ is always guaranteed. Substituting $x=\hat{z}^k$ in the above result and then taking expectation over ξ^k, we have

$$\begin{aligned}
&-L\mathbb{E}_k[\|x^{k+1}-x^k\|]+\theta f(x^k)\\
\le\ &\theta f(\hat{z}^k)+\frac{\theta(\lambda+\tau)}{2}\|\hat{z}^k-x^k\|^2-\frac{\lambda\theta^2}{2}\|\hat{z}^k-x^k\|^2\\
&+\frac{\gamma\theta^2}{2}\|\hat{z}^k-z^k\|^2-\frac{(\gamma-\lambda)\theta^2}{2}\mathbb{E}_k[\|\hat{z}^k-z^{k+1}\|^2]-\frac{\gamma}{2}\mathbb{E}_k[\|x^{k+1}-y^k\|^2]
\end{aligned}\tag{43}$$

Next we apply Lemma A.2 and use the optimality condition for $\hat{z}^k$, noting that $f(x)$ is $(\tau+\lambda)$-weakly convex, we get

$$f(\hat{z}^k)+\frac{\rho}{2}\|\hat{z}^k-z^k\|^2\le f(x^k)+\frac{\rho}{2}\|x^k-z^k\|^2-\frac{\rho-\tau-\lambda}{2}\|x^k-\hat{z}^k\|^2.\tag{44}$$

Multiplying (44) by θ and then adding the result to (43), we deduce

$$\begin{aligned}
&-L\mathbb{E}_k[\|x^{k+1}-x^k\|]\\
\le\ &\frac{\rho\theta}{2}\|x^k-z^k\|^2-\frac{\theta(\rho-\tau-\lambda)}{2}\|x^k-\hat{z}^k\|^2-\frac{\rho\theta}{2}\|\hat{z}^k-z^k\|^2\\
&+\frac{\theta(\lambda+\tau)}{2}\|\hat{z}^k-x^k\|^2-\frac{\lambda\theta^2}{2}\|\hat{z}^k-x^k\|^2\\
&+\frac{\gamma\theta^2}{2}\|\hat{z}^k-z^k\|^2-\frac{(\gamma-\lambda)\theta^2}{2}\mathbb{E}_k[\|\hat{z}^k-z^{k+1}\|^2]-\frac{\gamma}{2}\mathbb{E}_k[\|x^{k+1}-y^k\|^2]\\
=\ &\frac{\gamma\theta^2-\lambda\theta^2}{2}\big(\|\hat{z}^k-z^k\|^2-\mathbb{E}_k[\|\hat{z}^k-z^{k+1}\|^2]\big)-\frac{\rho\theta-\lambda\theta^2}{2}\mathbb{E}_k[\|\hat{z}^k-z^k\|^2]\\
&-\frac{\theta((\rho-2(\lambda+\tau))+\lambda\theta)}{2}\|\hat{z}^k-x^k\|^2\\
&-\frac{\gamma}{2}\mathbb{E}_k[\|x^{k+1}-y^k\|^2]+\frac{\rho\beta^2\theta^{-1}}{2}\|x^k-x^{k-1}\|^2.
\end{aligned}\tag{45}$$

where the last equality uses the identity $z^k-x^k=\beta\theta^{-1}(x^k-x^{k-1})$.

Moreover, we can bound the term $\mathbb{E}_k[\|x^{k+1}-y^k\|^2]$ using the following relation

$$\begin{aligned}
&\|x^{k+1}-y^k\|^2\\
=\ &\|x^{k+1}-x^k\|^2+\beta^2\|x^k-x^{k-1}\|^2-2\beta\langle x^{k+1}-x^k,x^k-x^{k-1}\rangle\\
\ge\ &\|x^{k+1}-x^k\|^2+\beta^2\|x^k-x^{k-1}\|^2-\beta\|x^{k+1}-x^k\|^2-\beta\|x^k-x^{k-1}\|^2\\
=\ &\theta^2\|x^{k+1}-x^k\|^2+\beta\theta\big(\|x^{k+1}-x^k\|^2-\|x^k-x^{k-1}\|^2\big).
\end{aligned}\tag{46}$$

Next, adding $L\mathbb{E}_k[\|x^{k+1}-x^k\|]$ to both sides of (45), using the non-negativity of $\rho-2(\lambda+\tau)$ and the bound (46), we deduce

$$\begin{aligned}
0 \leq & \frac{\gamma\theta^2-\lambda\theta^2}{2}\big(\|\hat{z}^k-z^k\|^2-\mathbb{E}_k[\|\hat{z}^k-z^{k+1}\|^2]\big)-\frac{\rho\theta-\lambda\theta^2}{2}\|\hat{z}^k-z^k\|^2 \\
& -\frac{\gamma\beta\theta+\rho\beta^2\theta^{-1}}{2}\mathbb{E}_k[\|x^{k+1}-x^k\|^2]+\frac{\gamma\beta\theta+\rho\beta^2\theta^{-1}}{2}\|x^k-x^{k-1}\|^2 \\
& +\mathbb{E}_k\Big[L\|x^{k+1}-x^k\|-\frac{\gamma\theta^2-\rho\beta^2\theta^{-1}}{2}\|x^{k+1}-x^k\|^2\Big] \\
\leq & \frac{\gamma\theta^2-\lambda\theta^2}{2}\big(\|\hat{z}^k-z^k\|^2-\mathbb{E}_k[\|\hat{z}^k-z^{k+1}\|^2]\big)-\frac{\rho\theta-\lambda\theta^2}{2}\|\hat{z}^k-z^k\|^2 \\
& -\frac{\gamma\beta\theta+\rho\beta^2\theta^{-1}}{2}\mathbb{E}_k[\|x^{k+1}-x^k\|^2]+\frac{\gamma\beta\theta+\rho\beta^2\theta^{-1}}{2}\|x^k-x^{k-1}\|^2 \\
& +\frac{L^2}{(\gamma\theta^2-\rho\beta^2\theta^{-1})}-\frac{\gamma\theta^2-\rho\beta^2\theta^{-1}}{4}\mathbb{E}_k[\|x^{k+1}-x^k\|^2]
\end{aligned}$$

where the last inequality identifies the fact that $bx-\frac{a}{4}x^2\leq\frac{b^2}{a}$ for $a,b>0,\forall x\in\mathbb{R}$. It then follows that

$$\begin{aligned}
& \mathbb{E}_k[\|\hat{z}^k-z^{k+1}\|^2] \\
\leq & \|\hat{z}^k-z^k\|^2-\frac{\rho-\lambda\theta}{\gamma\theta-\lambda\theta}\|\hat{z}^k-z^k\|^2+\frac{2L^2}{(\gamma\theta^2-\rho\beta^2\theta^{-1})(\gamma\theta^2-\lambda\theta^2)} \\
& -\frac{\gamma\beta+\rho\beta^2\theta^{-2}}{\gamma\theta-\lambda\theta}\big(\mathbb{E}_k[\|x^{k+1}-x^k\|^2]-\|x^k-x^{k-1}\|^2\big) \\
& -\frac{\gamma-\rho\beta^2\theta^{-3}}{2(\gamma-\lambda)}\mathbb{E}_k[\|x^{k+1}-x^k\|^2] \qquad (47)
\end{aligned}$$

In view of (47) and the definition of Moreau envelope, we have

$$\begin{aligned}
& \mathbb{E}_k\left[f_{1/\rho}(z^{k+1})\right] \\
= & \mathbb{E}_k\big[f(\hat{z}^{k+1})+\frac{\rho}{2}\|z^{k+1}-\hat{z}^{k+1}\|^2\big] \\
\leq & \mathbb{E}_k\big[f(\hat{z}^k)+\frac{\rho}{2}\|z^{k+1}-\hat{z}^k\|^2\big] \\
\leq & f_{1/\rho}(z^k)-\frac{\rho(\rho-\lambda\theta)}{2(\gamma\theta-\lambda\theta)}\|z^k-\hat{z}^k\|^2+\frac{\rho L^2}{(\gamma\theta^2-\rho\beta^2\theta^{-1})(\gamma\theta^2-\lambda\theta^2)} \\
& +\frac{\rho(\gamma\beta+\rho\beta^2\theta^{-2})}{2(\gamma\theta-\lambda\theta)}\big\{\|x^k-x^{k-1}\|^2-\mathbb{E}_k\big[\|x^{k+1}-x^k\|^2\big]\big\}. \\
& -\frac{\rho(\gamma-\rho\beta^2\theta^{-3})}{4(\gamma-\lambda)}\mathbb{E}_k[\|x^{k+1}-x^k\|^2] \qquad (48)
\end{aligned}$$

In view of the above result and the relation $\|z^k-\hat{z}^k\|^2=\rho^{-2}\|\nabla_{1/\rho}f(z^k)\|^2$, we obtain (18).

B.2 Proof of Theorem 4.2

Unfolding the relation (18) and then taking expectation over all the randomness, we have

$$\begin{aligned}
& \frac{\rho-\lambda\theta}{2\rho(\gamma\theta-\lambda\theta)}\sum_{k=1}^{K}\mathbb{E}[\|\nabla f_{1/\rho}(z^k)\|^2] \\
\leq & f_{1/\rho}(z^1)-\mathbb{E}\big[f_{1/\rho}(z^{K+1})\big]+\frac{\rho(\gamma\beta+\rho\beta^2\theta^{-2})}{2(\gamma\theta-\lambda\theta)}\|x^1-x^0\|^2 \\
& +\frac{\rho L^2K}{(\gamma\theta^2-\rho\beta^2\theta^{-1})(\gamma\theta^2-\lambda\theta^2)} \\
\leq & \Delta+\frac{\rho L^2K}{(\gamma\theta^2-\rho\beta^2\theta^{-1})(\gamma\theta^2-\lambda\theta^2)}, \qquad (49)
\end{aligned}$$

where the last inequality uses $x^1 = x^0 = z^1$ and that $f_{1/\rho}(z^1) - f_{1/\rho}(z^{K+1}) \le f(z^1) - \min_x f(x) = \Delta$. Appealing to the definition of k^* and relation (64), we have

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f_{1/\rho}(z^{k^*})\|^2\right] \\
&= \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\|\nabla f_{1/\rho}(z^k)\|^2\right] \\
&\le \frac{2\rho(\gamma\theta - \lambda\theta)\Delta}{(\rho - \lambda\theta)K} + \frac{2\rho^2 L^2}{\theta(\rho - \lambda\theta)(\gamma\theta - \rho\beta^2\theta^{-2})} \\
&\le \frac{2\rho}{\rho - \lambda}\left[\frac{(\gamma\theta - \lambda\theta)\Delta}{K} + \frac{\rho L^2}{\theta(\gamma\theta - \rho\beta^2\theta^{-2})}\right] \\
&= \frac{2\rho}{\rho - \lambda}\left[\frac{(\rho\beta^2\theta^{-2} + \gamma_0\sqrt{K})\Delta}{K} + \frac{\rho L^2}{\theta(\gamma_0\sqrt{K} + \lambda\theta)}\right] \\
&\le \frac{2\rho}{\rho - \lambda}\left[\frac{\rho\beta^2\theta^{-2}\Delta}{K} + \Big(\gamma_0\Delta + \frac{\rho L^2}{\theta\gamma_0}\Big)\frac{1}{\sqrt{K}}\right].
\end{aligned}$$

where the first inequality uses the fact that $(\rho - \lambda\theta)^{-1} \le (\rho - \lambda)^{-1}$ for $\theta \in (0, 1]$ and that $\gamma = \gamma_0\theta^{-1}\sqrt{K} + \lambda + \rho\beta^2\theta^{-3}$. Therefore, (19) immediately follows.

B.3 SMOD with momentum and minibatching

We present a new model-based method by combining the momentum and minibatching techniques in a single framework.

Algorithm 3 Stochastic Extrapolated Model-Based Method with Minibatching

Input: x^0, x^1, β, γ
for $k = 1$ **to** K **do**
Sample a minibatch $B_k = \{\xi_{k,1}, \ldots, \xi_{k,m}\}$ and update:

$$y^k = x^k + \beta(x^k - x^{k-1}) \tag{50}$$

$$x^{k+1} = \operatorname*{argmin}_{x\in\mathcal{X}} \Big\{ f_{x^k}(x, B_k) + \frac{\gamma}{2}\|x - y^k\|^2 \Big\} \tag{51}$$

end for

The convergence analysis of Algorithm 3 is more complicated than that of the sequential extrapolated SMOD. We require a different design of potential function:

$$f_{1/\rho}(z^k) + \alpha f(x^k) + \beta\|x^k - x^{k-1}\|^2$$

where α and β are some constants and z_k is defined as in Section 4. We summarize the approximate descent property in the following function.

Lemma B.1. *In Algorithm 3, Assume that A5 holds and $\rho > 3(\tau + \lambda)$, then we have*

$$\begin{aligned}
&\frac{\rho - \lambda\theta}{2\theta\rho(\gamma - \lambda)}\|\nabla f_{1/\rho}(z^k)\|^2 \\
&\le f_{1/\rho}(z^k) - \mathbb{E}_k\big[f_{1/\rho}(z^{k+1})\big] + \frac{\rho\beta}{2\theta^2(\gamma - \lambda)}\big[f(x^k) - \mathbb{E}_k[f(x^{k+1})]\big] \\
&\quad - \frac{\rho(\gamma\theta^2 - \zeta)}{4\theta^2(\gamma - \lambda)}\|x^{k+1} - x^k\|^2 + \frac{\rho\varepsilon}{2\theta^2(\gamma - \lambda)} \\
&\quad + \frac{\rho(\gamma\beta + 2\rho\beta^2\theta^{-2})}{2\theta(\gamma - \lambda)}\big\{\|x^k - x^{k-1}\|^2 - \mathbb{E}_k[\|x^{k+1} - x^k\|^2]\big\}.
\end{aligned} \tag{52}$$

where $\zeta = 2\theta(\rho + \lambda\beta + \tau) + \tau + 2\rho\beta^2\theta^{-1}$ and $\varepsilon = \frac{2L^2}{m(\gamma-\lambda)}$.

Proof. Analogous to the relation (40), we have

$$\begin{aligned}
&f_{x^k}(x^{k+1}, B_k) + \frac{\gamma}{2}\|x^{k+1} - y^k\|^2 \\
&\leq (1-\theta) f(x^k, B_k) + \theta\big[f(x, B_k) + \frac{\lambda+\tau}{2}\|x - x^k\|^2\big] - \frac{\lambda\theta^2}{2}\|x - x^k\|^2 \\
&\quad + \frac{\gamma\theta^2}{2}\|x - z^k\|^2 - \frac{(\gamma-\lambda)\theta^2}{2}\|x - z^{k+1}\|^2
\end{aligned} \tag{53}$$

Placing the value $x = \hat{z}^k$, we arrive at

$$\begin{aligned}
&f_{x^k}(x^{k+1}, B_k) + \frac{\gamma}{2}\|x^{k+1} - y^k\|^2 \\
&\leq (1-\theta) f(x^k, B_k) + \theta f(\hat{z}^k, B_k) + \frac{(\lambda+\tau)\theta - \lambda\theta^2}{2}\|\hat{z}^k - x^k\|^2 \\
&\quad + \frac{\gamma\theta^2}{2}\|\hat{z}^k - z^k\|^2 - \frac{(\gamma-\lambda)\theta^2}{2}\|\hat{z}^k - z^{k+1}\|^2 \\
&\leq (1-\theta) f(x^k, B_k) + \theta f(\hat{z}^k, B_k) + \theta(\lambda\beta+\tau)\big[\|\hat{z}^k - x^{k+1}\|^2 + \|x^k - x^{k+1}\|^2\big] \\
&\quad + \frac{\gamma\theta^2}{2}\|\hat{z}^k - z^k\|^2 - \frac{(\gamma-\lambda)\theta^2}{2}\|\hat{z}^k - z^{k+1}\|^2
\end{aligned} \tag{54}$$

where the last inequality uses the fact $(\lambda+\tau)\theta - \lambda\theta^2 = \theta(\lambda\beta+\tau)$ and applies $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ with $a = \hat{z}^k - x^{k+1}$ and $b = x^{k+1} - x^k$.

Recall that $\hat{z}^k = \mathrm{prox}_{f/\rho}(z^k)$. In view of Lemma A.2 and the $(\tau+\lambda)$-weak convexity of $f(\cdot)$, we have

$$\theta f(\hat{z}^k) + \frac{\rho\theta}{2}\|\hat{z}^k - z^k\|^2 \leq \theta f(x^{k+1}) + \frac{\rho\theta}{2}\|x^{k+1} - z^k\|^2 - \frac{\theta(\rho-\tau-\lambda)}{2}\|x^{k+1} - \hat{z}^k\|^2. \tag{55}$$

Summing up (54) and (55) and rearranging the terms, we arrive at

$$\begin{aligned}
&\frac{\gamma}{2}\|x^{k+1} - y^k\|^2 \\
&\leq (1-\theta)\big[f(x^k, B_k) - f(x^{k+1})\big] + \theta\big[f(\hat{z}^k, B_k) - f(\hat{z}^k)\big] + f(x^{k+1}) - f_{x^k}(x^{k+1}, B_k) \\
&\quad + \theta(\lambda\beta+\tau)\|x^k - x^{k+1}\|^2 \\
&\quad + \frac{\gamma\theta^2 - \rho\theta}{2}\|\hat{z}^k - z^k\|^2 - \frac{(\gamma-\lambda)\theta^2}{2}\|\hat{z}^k - z^{k+1}\|^2 \\
&\quad + \frac{\rho\theta}{2}\|x^{k+1} - z^k\|^2 - \frac{\theta(\rho - 3(\tau+\lambda) + 2\lambda\theta)}{2}\|x^{k+1} - \hat{z}^k\|^2
\end{aligned} \tag{56}$$

On both sides of the above inequality, we take expectation over B_k conditioned on all the randomness that generates $B_1, B_2, \ldots, B_{k-1}$. Noting that $\mathbb{E}_k\big[f(x^k, B_k)\big] = f(x^k)$ and $\mathbb{E}_k\big[f(\hat{z}^k, B_k)\big] = f(\hat{z}^k)$, it follows that

$$\begin{aligned}
&\frac{\gamma}{2}\mathbb{E}_k\big[\|x^{k+1} - y^k\|^2\big] \\
&\leq (1-\theta)\big[f(x^k) - \mathbb{E}_k[f(x^{k+1})]\big] + \mathbb{E}_k\big[f(x^{k+1}) - f_{x^k}(x^{k+1}, B_k)\big] \\
&\quad + \theta(\lambda\beta+\tau)\mathbb{E}_k\|x^k - x^{k+1}\|^2 + \frac{\gamma\theta^2 - \rho\theta}{2}\|\hat{z}^k - z^k\|^2 - \frac{(\gamma-\lambda)\theta^2}{2}\mathbb{E}_k\|\hat{z}^k - z^{k+1}\|^2 \\
&\quad + \frac{\rho\theta}{2}\mathbb{E}_k\|x^{k+1} - z^k\|^2 - \frac{\theta(\rho - 3(\tau+\lambda) + 2\lambda\theta)}{2}\mathbb{E}_k\|x^{k+1} - \hat{z}^k\|^2
\end{aligned} \tag{57}$$

Moreover, similar to the analysis for minibatch SMOD, we apply Theorem A.3 and Lemma 3.1 to show that

$$\mathbb{E}_k\big\{\mathbb{E}_\xi\big[f_{x^k}(x^{k+1}, \xi)\big] - f_{x^k}(x^{k+1}, B_k)\big\} \leq \varepsilon.$$

In view of this result and Assumption A5, we arrive at

$$\begin{aligned}
&\mathbb{E}_k\big[f(x^{k+1}) - f_{x^k}(x^{k+1}, B_k)\big] \\
&= \mathbb{E}_k\big[f(x^{k+1}) - \mathbb{E}_\xi[f_{x^k}(x^{k+1}, \xi)]\big] + \mathbb{E}_k\big\{\mathbb{E}_\xi\big[f_{x^k}(x^{k+1}, \xi)\big] - f_{x^k}(x^{k+1}, B_k)\big\} \\
&\leq \frac{\tau}{2}\mathbb{E}_k[\|x^k - x^{k+1}\|^2] + \varepsilon.
\end{aligned} \tag{58}$$

Putting (57) and (58) together and using the assumption $\rho > 3(\tau + \lambda)$, we have

$$\begin{aligned}
&\frac{\gamma}{2}\mathbb{E}_k\big[\|x^{k+1} - y^k\|^2\big] \\
&\leq (1-\theta)\big[f(x^k) - \mathbb{E}_k[f(x^{k+1})]\big] + \frac{2\theta(\lambda\beta+\tau)+\tau}{2}\mathbb{E}_k[\|x^k - x^{k+1}\|^2] + \varepsilon \\
&\quad + \frac{\gamma\theta^2 - \rho\theta}{2}\|\hat{z}^k - z^k\|^2 - \frac{(\gamma-\lambda)\theta^2}{2}\mathbb{E}_k[\|\hat{z}^k - z^{k+1}\|^2] \\
&\quad + \frac{\rho\theta}{2}\mathbb{E}_k[\|x^{k+1} - z^k\|^2] \qquad (59)
\end{aligned}$$

Moreover, we can bound the term $\mathbb{E}_k[\|x^{k+1} - y^k\|^2]$

$$\begin{aligned}
&\|x^{k+1} - y^k\|^2 \\
&= \|x^{k+1} - x^k\|^2 + \beta^2\|x^k - x^{k-1}\|^2 - 2\beta\langle x^{k+1} - x^k, x^k - x^{k-1}\rangle \\
&\geq \|x^{k+1} - x^k\|^2 + \beta^2\|x^k - x^{k-1}\|^2 - \beta\|x^{k+1} - x^k\|^2 - \beta\|x^k - x^{k-1}\|^2 \\
&= \theta\|x^{k+1} - x^k\|^2 - \beta\theta\|x^k - x^{k-1}\|^2, \qquad (60)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\rho\theta}{2}\|x^{k+1} - z^k\|^2 &= \frac{\rho\theta}{2}\|x^{k+1} - x^k - \beta\theta^{-1}(x^k - x^{k-1})\|^2 \\
&\leq \rho\theta\|x^{k+1} - x^k\|^2 + \rho\beta^2\theta^{-1}\|x^k - x^{k-1}\|^2 \qquad (61)
\end{aligned}$$

where the inequality comes from the fact that $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$.
Putting (59), (60) and (61) together, we have

$$\begin{aligned}
&\frac{\gamma\theta^2 - 2\theta(\rho+\lambda\beta+\tau) - \tau - 2\rho\beta^2\theta^{-1}}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] \\
&\leq (1-\theta)\big[f(x^k) - \mathbb{E}_k[f(x^{k+1})]\big] + \varepsilon \\
&\quad + \frac{\gamma\theta^2 - \rho\theta}{2}\|\hat{z}^k - z^k\|^2 - \frac{(\gamma-\lambda)\theta^2}{2}\mathbb{E}_k[\|\hat{z}^k - z^{k+1}\|^2] \\
&\quad + \frac{\gamma\beta + 2\rho\beta^2\theta^{-1}}{2}\mathbb{E}_k\big[\|x^k - x^{k-1}\|^2 - \|x^{k+1} - x^k\|^2\big]
\end{aligned}$$

It then follows that

$$\begin{aligned}
&\mathbb{E}_k[\|\hat{z}^k - z^{k+1}\|^2] \\
&\leq \|\hat{z}^k - z^k\|^2 - \frac{\rho - \lambda\theta}{\gamma\theta - \lambda\theta}\|\hat{z}^k - z^k\|^2 + \frac{\varepsilon}{(\gamma-\lambda)\theta^2} \\
&\quad + \frac{\beta}{(\gamma-\lambda)\theta^2}\big[f(x^k) - \mathbb{E}_k[f(x^{k+1})]\big] \\
&\quad - \frac{\gamma\theta^2 - 2\theta(\rho+\lambda\beta+\tau) - \tau - 2\rho\beta^2\theta^{-1}}{2(\gamma-\lambda)\theta^2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] \\
&\quad - \frac{\gamma\beta + 2\rho\beta^2\theta^{-2}}{\gamma\theta - \lambda\theta}\big(\mathbb{E}_k\big[\|x^{k+1} - x^k\|^2 - \|x^k - x^{k-1}\|^2\big]\big) \qquad (62)
\end{aligned}$$

In view of (62) and the definition of Moreau envelope, we have

$$
\begin{aligned}
&\mathbb{E}_k\left[f_{1/\rho}(z^{k+1})\right] \\
&= \mathbb{E}_k\big[f(\hat{z}^{k+1}) + \frac{\rho}{2}\|z^{k+1}-\hat{z}^{k+1}\|^2\big] \\
&\le \mathbb{E}_k\big[f(\hat{z}^{k}) + \frac{\rho}{2}\|z^{k+1}-\hat{z}^{k}\|^2\big] \\
&\le f_{1/\rho}(z^k) - \frac{\rho(\rho-\lambda\theta)}{2(\gamma\theta-\lambda\theta)}\|z^k-\hat{z}^k\|^2 + \frac{\rho\varepsilon}{2(\gamma\theta^2-\lambda\theta^2)} + \frac{\rho\beta}{2(\gamma-\lambda)\theta^2}\big[f(x^k)-\mathbb{E}_k[f(x^{k+1})]\big] \\
&\quad - \frac{\rho(\gamma\theta^2-2\theta(\rho+\lambda\beta+\tau)-\tau-2\rho\beta^2\theta^{-1})}{4(\gamma-\lambda)\theta^2}\|x^{k+1}-x^k\|^2 \\
&\quad + \frac{\rho(\gamma\beta+2\rho\beta^2\theta^{-2})}{2(\gamma\theta-\lambda\theta)}\big\{\|x^k-x^{k-1}\|^2 - \mathbb{E}_k\big[\|x^{k+1}-x^k\|^2\big]\big\}. \qquad (63)
\end{aligned}
$$

In view of the above result and the relation $\|z^k-\hat{z}^k\|^2=\rho^{-2}\|\nabla f_{1/\rho}(z^k)\|^2$, we obtain (52). □

Theorem B.2. *Suppose we choose $\gamma=\gamma_0\sqrt{\frac{K}{m}}+\theta^{-2}\zeta+\lambda$, where ζ is defined in B.1. Then we have*

$$
\mathbb{E}\left[\|\nabla f_{1/\rho}(z^{k^*})\|^2\right] \le \frac{\rho}{\rho-\theta\lambda}\left[\frac{\theta^{-1}(\rho\beta+2\zeta)\Delta}{K} + \Big(\theta\gamma_0\Delta+\frac{\rho L^2}{\theta\gamma_0}\Big)\frac{2}{\sqrt{mK}}\right].
$$

Proof. Unfolding the relation (52) and then taking expectation over all the randomness, we have

$$
\begin{aligned}
&\frac{\rho-\lambda\theta}{2\theta\rho(\gamma-\lambda)}\sum_{k=1}^{K}\mathbb{E}[\|\nabla f_{1/\rho}(z^k)\|^2] \\
&\le f_{1/\rho}(z^1) - \mathbb{E}\big[f_{1/\rho}(z^{K+1})\big] + \frac{\rho\beta}{2\theta^2(\gamma-\lambda)}\big[f(x^1)-\mathbb{E}_k[f(x^{K+1})]\big] \\
&\quad + \frac{\rho\varepsilon K}{2\theta^2(\gamma-\lambda)} + \frac{\rho(\gamma\beta+2\rho\beta^2\theta^{-2})}{2\theta(\gamma-\lambda)}\|x^1-x^0\|^2. \\
&\le \Big(1+\frac{\rho\beta}{2\theta^2(\gamma-\lambda)}\Big)\Delta + \frac{L^2\rho K}{\theta^2 m(\gamma-\lambda)^2}, \qquad (64)
\end{aligned}
$$

where we use the assumption $x^1=x^0=z^1$ and that

$$
\max\left\{f_{1/\rho}(z^1)-f_{1/\rho}(z^{K+1}),\, f_{1/\rho}(x^1)-f_{1/\rho}(x^{K+1})\right\}\le\Delta.
$$

Appealing to the definition of k^*, γ and then using relation (64), we arrive at

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla f_{1/\rho}(z^{k^*})\|^2\right] \\
&\le \frac{\rho}{\rho-\theta\lambda}\left[\frac{\rho\beta\theta^{-1}\Delta}{K} + \frac{2\theta(\gamma-\lambda)\Delta}{K} + \frac{2\rho L^2}{\theta m(\gamma-\lambda)}\right] \\
&\le \frac{\rho}{\rho-\theta\lambda}\left[\frac{\theta^{-1}(\rho\beta+2\zeta)\Delta}{K} + \Big(\theta\gamma_0\Delta+\frac{\rho L^2}{\theta\gamma_0}\Big)\frac{2}{\sqrt{mK}}\right].
\end{aligned}
$$

□

Remark 7. While the convergence result in Theorem B.2 is established for all $\gamma_0>0$, we can see that the optimal γ_0 would be $\gamma_0=\theta^{-1}\sqrt{\frac{\rho}{\Delta}}L$, which gives the bound $\mathbb{E}[\,\|\nabla f_{1/\rho}(z^{k^*})\|^2]=\mathcal{O}\big(\frac{\Delta}{K}+L\sqrt{\frac{\rho\Delta}{mK}}\big)$. In practice we can set γ_0 to a suboptimal value and obtain a possibly loose upper-bound.

C SMOD for convex optimization

In this section, we develop new complexity results of model-based methods for stochastic convex optimization. To provide the sharpest convergence rate possible, we replace Assumption A5 with the following assumption

A6: For any $x \in \mathcal{X}$, $f_x(\cdot, \xi)$ is a convex function, and

$$-\frac{\tau}{2}\|x-y\|^2 \leq f_x(y,\xi) - f(y,\xi) \leq 0, \quad \xi \in \Xi, y \in \mathcal{X}. \tag{65}$$

It is easy to see that Assumption A6 ensures the convexity of $f(y, \xi)$. More specifically, let $\bar{x} = (1-\alpha)x + \alpha y$ where $x, y \in \mathcal{X}$ and $\alpha \in [0, 1]$, we have

$$\begin{aligned} f(\bar{x},\xi) &= f_{\bar{x}}(\bar{x},\xi) \\ &\leq (1-\alpha) f_{\bar{x}}(x,\xi) + \alpha f_{\bar{x}}(y,\xi) \\ &\leq (1-\alpha) f(x,\xi) + \alpha f(y,\xi) \end{aligned}$$

where the equality comes from Assumption A2, the first inequality follows from convexity of $f_{\bar{x}}(\cdot, \xi)$ and the second inequality uses (65).

Outline of this section. Since convergence to global optimality can be guaranteed in convex optimization, it is favorable to describe convergence rates with respect to the optimality gap. To this end, we conduct new convergence analysis of SMOD with minibatching and momentum for stochastic convex optimization. In subsection C.1, we show that under the additional Assumption A6, after K iterations of the extrapolated minibatch method (Algorithm 3), the expected optimality gap converges at rate

$$\mathcal{O}\Big(\frac{1}{K} + \frac{1}{\sqrt{mK}}\Big).$$

In view of the above result, the deterministic part of our rate is consistent with the best $\mathcal{O}(\frac{1}{K})$ rate for heavy-ball method. For example, see [4, 6]. Moreover, the stochastic part of the rate is improved from the result $\mathcal{O}(\frac{1}{\sqrt{K}})$ of Theorem 4.4 [3] by a factor of $\sqrt{m}$.

As is mentioned in the main article, one major advantage of SMOD methods is the robustness to stepsize selection (see [1]). In other words, compared to SGD, SPL and SPP tend to admit a wider range of stepsizes. In subsection C.2, we show that the extrapolated model-based method inherits the merits of robustness from the model-based method.

An important question arises naturally: Can we further improve the convergence rate of model-based methods? Due to the widely known limitation of heavy-ball type momentum, it would be interesting to consider Nesterov's acceleration. In subsection C.3, we present a model-based method with Nesterov type momentum. Thanks to the stability argument, we obtain the following improved rate of convergence:

$$\mathcal{O}\Big(\frac{1}{K^2} + \frac{1}{\sqrt{mK}}\Big).$$

We note that a similar convergence rate for minibatch model-based methods is obtained in a recent paper [2]. However, their result requires the assumption that the stochastic function is Lipschitz smooth while our assumption is much weaker. The full complexity results are presented in Table 2.

Table 2: Complexity of stochastic algorithms to reach ε-accuracy: $\mathbb{E}[f(x) - f(x^*)] \leq \varepsilon$. (M: minibatching; E: Extrapolation (Polyak type); N: Nesterov acceleration

Algorithms	Problems	Current Best	Ours
M + SMOD	f: smooth composite	$\mathcal{O}(1/\varepsilon^2)$ [3]	$\mathcal{O}(1/\varepsilon + 1/(m\varepsilon^2))$
M + E + SMOD	f: non-smooth	$\mathcal{O}(1/\varepsilon^2)$ [3]	$\mathcal{O}(1/\varepsilon + 1/(m\varepsilon^2))$
M + N + SMOD	f: smooth composite	$\mathcal{O}(1/\varepsilon^{1/2} + 1/(m\varepsilon^2))$ [2]	$\mathcal{O}(1/\varepsilon^{1/2} + 1/(m\varepsilon^2))$
M + N + SMOD	f: non-smooth	—	$\mathcal{O}(1/\varepsilon^{1/2} + 1/(m\varepsilon^2))$

C.1 Convergence of extrapolated SMOD

The following Lemma summarizes some important convergence property of Extrapolated SMOD for convex stochastic optimization.

Lemma C.1. *Under Assumption A6, let $\theta = 1-\beta$ in Algorithm 3. Then for any $\hat{x} \in \mathcal{X}$ and $k = 1, 2, 3, \ldots$, we have*

$$\begin{aligned}
&\mathbb{E}_k\big[f(x^{k+1}) - f(\hat{x})\big] - (1-\theta)\big[f(x^k) - f(\hat{x})\big] \\
\leq\ & \frac{2L^2}{m\gamma} + \frac{\gamma\theta^2}{2}\|\hat{x} - z^k\|^2 - \frac{\gamma\theta^2}{2}\mathbb{E}_k[\|\hat{x} - z^{k+1}\|^2] \\
&+ \frac{\gamma\beta(1-\beta)}{2}\|x^k - x^{k-1}\|^2 - \frac{\gamma(1-\beta)-\tau}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2]
\end{aligned} \tag{66}$$

Proof. Applying three point lemma, for any $x \in \mathcal{X}$, we have

$$f_{x^k}(x^{k+1}, B_k) - f_{x^k}(x, B_k) \leq \frac{\gamma}{2}\|x - y^k\|^2 - \frac{\gamma}{2}\|x - x^{k+1}\|^2 - \frac{\gamma}{2}\|y^k - x^{k+1}\|^2. \tag{67}$$

Based on Assumption A6, we have

$$\begin{aligned}
&f(x^{k+1}) - f_{x^k}(x^{k+1}, B_k) \\
=\ & \mathbb{E}_\xi\big[f(x^{k+1}, \xi)\big] - f_{x^k}(x^{k+1}, B_k) \\
=\ & \mathbb{E}_\xi\big[f(x^{k+1}, \xi) - f_{x^k}(x^{k+1}, \xi)\big] + \mathbb{E}_\xi\big[f_{x^k}(x^{k+1}, \xi) - f_{x^k}(x^{k+1}, B_k)\big] \\
\leq\ & \frac{\tau}{2}\|x^k - x^{k+1}\|^2 + \mathbb{E}_\xi\big[f_{x^k}(x^{k+1}, \xi) - f_{x^k}(x^{k+1}, B_k)\big].
\end{aligned} \tag{68}$$

Plugging the above into (67), we have that

$$\begin{aligned}
f(x^{k+1}) - f_{x^k}(x, B_k) \leq\ & \frac{\gamma}{2}\|x - y^k\|^2 - \frac{\gamma}{2}\|x - x^{k+1}\|^2 - \frac{\gamma}{2}\|y^k - x^{k+1}\|^2 \\
&+ \frac{\tau}{2}\|x^k - x^{k+1}\|^2 + \mathbb{E}_\xi\big[f_{x^k}(x^{k+1}, \xi) - f_{x^k}(x^{k+1}, B_k)\big].
\end{aligned}$$

Let $x = (1-\theta)x^k + \theta\hat{x}$ and $z^k = x^k + \theta^{-1}\beta(x^k - x^{k-1})$. Then we have

$$\begin{aligned}
x - y^k &= \theta(\hat{x} - z^k), \\
x - x^{k+1} &= \theta(\hat{x} - z^{k+1}),
\end{aligned}$$

and by convexity, we obtain that

$$\begin{aligned}
&f(x^{k+1}) - f(\hat{x}, B_k) - (1-\theta)\big[f(x^k, B_k) - f(\hat{x}, B_k)\big] \\
\leq\ & f(x^{k+1}) - f_{x^k}(x, B_k) \\
\leq\ & \frac{\gamma\theta^2}{2}\|\hat{x} - z^k\|^2 - \frac{\gamma\theta^2}{2}\|\hat{x} - z^{k+1}\|^2 - \frac{\gamma}{2}\|y^k - x^{k+1}\|^2 \\
&+ \frac{\tau}{2}\|x^k - x^{k+1}\|^2 + \mathbb{E}_\xi\big[f_{x^k}(x^{k+1}, \xi) - f_{x^k}(x^{k+1}, B_k)\big].
\end{aligned} \tag{69}$$

Then we have

$$\begin{aligned}
&-\frac{\gamma}{2}\|y^k - x^{k+1}\|^2 + \frac{\tau}{2}\|x^k - x^{k+1}\|^2 \\
=\ & -\frac{\gamma}{2}\|x^{k+1} - x^k\|^2 + \gamma\beta\big\langle x^{k+1} - x^k, x^k - x^{k-1}\big\rangle - \frac{\gamma\beta^2}{2}\|x^k - x^{k-1}\|^2 + \frac{\tau}{2}\|x^k - x^{k+1}\|^2 \\
\leq\ & \frac{\gamma\beta(1-\beta)}{2}\|x^k - x^{k-1}\|^2 - \frac{\gamma(1-\beta)-\tau}{2}\|x^{k+1} - x^k\|^2,
\end{aligned} \tag{70}$$

where the last inequality is by Cauchy-Schwarz and we deduce that

$$\begin{aligned}
&f(x^{k+1}) - f(\hat{x}, B_k) - (1-\theta)\big[f(x^k, B_k) - f(\hat{x}, B_k)\big] \\
\leq\ & \frac{\gamma\theta^2}{2}\|\hat{x} - z^k\|^2 - \frac{\gamma\theta^2}{2}\|\hat{x} - z^{k+1}\|^2 \\
&+ \frac{\gamma\beta(1-\beta)}{2}\|x^k - x^{k-1}\|^2 - \frac{\gamma(1-\beta)-\tau}{2}\|x^{k+1} - x^k\|^2 \\
&+ \mathbb{E}_\xi\big[f_{x^k}(x^{k+1}, \xi) - f_{x^k}(x^{k+1}, B_k)\big].
\end{aligned}$$

Next, we take expectation over B_k conditioned on $B_1, B_2, \ldots, B_{k-1}$. Note that $\mathbb{E}_k[f(\hat{x}, B_k)] = f(\hat{x})$, $\mathbb{E}_k[f(x^k, B_k)] = f(x^k)$ and

$$\begin{aligned}
&\mathbb{E}_k\big[f(x^{k+1}) - f(\hat{x})\big] - (1-\theta)\big[f(x^k) - f(\hat{x})\big] \\
\leq\ & \frac{\gamma\theta^2}{2}\|\hat{x} - z^k\|^2 - \frac{\gamma\theta^2}{2}\mathbb{E}_k[\|\hat{x} - z^{k+1}\|^2] \\
&+ \frac{\gamma\beta(1-\beta)}{2}\|x^k - x^{k-1}\|^2 - \frac{\gamma(1-\beta)-\tau}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] \\
&+ \mathbb{E}_k\big\{\mathbb{E}_\xi\big[f_{x^k}(x^{k+1}, \xi) - f_{x^k}(x^{k+1}, B_k)\big]\big\}.
\end{aligned} \tag{71}$$

Moreover, based on the stability of the proximal mapping, we have

$$\mathbb{E}_k\big\{\mathbb{E}_\xi\big[f_{x^k}(x^{k+1}, \xi) - f_{x^k}(x^{k+1}, B_k)\big]\big\} \leq \varepsilon_k, \text{ where } \varepsilon_k = \frac{2L^2}{m\gamma}. \tag{72}$$

Combining (71) and (72) gives the desired result (66). □

By specifying a constant stepsize and batch size, we develop the convergence rate of SEMOD in the following Theorem.

Theorem C.2. *Let $x^1 = x^0$, x^* be an optimal solution and $\gamma = \gamma_0\sqrt{\frac{K}{m}} + \theta^{-2}\tau$, where $\gamma_0 = \frac{2\theta^{-1}L}{\tilde{D}}$ and $\tilde{D} \geq \|x^0 - x^*\|$, then we have*

$$\mathbb{E}\big[f(x^{k^*}) - f(x^*)\big] \leq \frac{f(x^0) - f(x^*)}{K} + \frac{\theta^{-1}\tau\tilde{D}^2}{2K} + \frac{2\tilde{D}L}{\sqrt{mK}}. \tag{73}$$

where k^* is an index chosen in $\{1, 2, \ldots, K\}$ uniformly at random.

Proof. Let us denote $\Delta_k = \mathbb{E}[f(x^k) - f(x^*)]$ for the sake of simplicity. Following Lemma C.1 (with $\hat{x} = x^*$), we sum up (66) over $k = 1, 2, \ldots, K$ and then take expectation over all the randomness, then we have

$$\Delta_{K+1} + \theta\sum_{k=1}^{K}\Delta_k \leq \Delta_1 + \frac{\gamma\theta^2}{2}\|\hat{x} - z^1\|^2 + \frac{\gamma\beta(1-\beta)}{2}\|x^1 - x^0\|^2 + \frac{2L^2K}{m\gamma},$$

where the inequality holds since $\gamma \geq \theta^{-2}\tau$. Using $x^1 = z^1 = x^0$, we have

$$\begin{aligned}
\mathbb{E}\big[f(x^{k^*}) - f(x^*)\big] &= \frac{1}{K}\sum_{k=1}^{K}\Delta_k \\
&\leq \frac{\Delta_1}{K} + \frac{\gamma\theta}{2K}\|x^* - x^0\|^2 + \frac{2L^2}{m\theta\gamma} \\
&\leq \frac{\Delta_1}{K} + \frac{\gamma\theta\tilde{D}^2}{2K} + \frac{2L^2}{m\theta\gamma} \\
&\leq \frac{\Delta_1}{K} + \frac{\theta^{-1}\tau\tilde{D}^2}{2K} + \frac{\theta\gamma_0\tilde{D}^2}{2\sqrt{mK}} + \frac{2L^2}{\sqrt{mK}\theta\gamma_0} \\
&= \frac{\Delta_1}{K} + \frac{\theta^{-1}\tau\tilde{D}^2}{2K} + \frac{2\tilde{D}L}{\sqrt{mK}}.
\end{aligned}$$

Therefore, we complete the proof. □

C.2 Robustness of the extrapolated SMOD

As is mentioned in the main article, one major advantage of SMOD methods is the robustness to stepsize selection (see [1]). In other words, compared to SGD, SPL and SPP tend to admit a wider range of stepsizes. We show that the extrapolated model-based method inherit the merits of robustness. For the sake of the asymptotic analysis, stepsize parameter γ_k in SEMOD is now indexed by k. We present the main convergence property in the following theorem.

Theorem C.3. *Suppose that Assumption A6 holds, $x^1 = x^0$ and the stepsize γ_k satisfies $\gamma_k \geq 2\tau\theta^{-2}$. Then we have*

$$\mathbb{E}[\|x^* - x^{K+1}\|^2] \leq \|x^* - x^1\|^2 + 2\theta^{-2}\sum_{k=1}^{K}\gamma_k^{-2}\mathbb{E}[\|f'(x^*, B_k)\|^2].$$

Proof. First, Assumption A6 implies that

$$\begin{aligned} f_{x^k}(x^{k+1}, B_k) &\geq f(x^{k+1}, B_k) - \frac{\tau}{2}\|x^{k+1} - x^k\|^2, \\ -f_{x^k}(x, B_k) &\geq -f(x, B_k). \end{aligned}$$

Summing up the above two relations gives

$$f_{x^k}(x^{k+1}, B_k) - f_{x^k}(x, B_k) \geq f(x^{k+1}, B_k) - f(x, B_k) - \frac{\tau}{2}\|x^{k+1} - x^k\|^2.$$

In view of (67), we have

$$\begin{aligned} f(x^{k+1}, B_k) - f(x, B_k) &\leq f_{x^k}(x^{k+1}, B_k) - f_{x^k}(x, B_k) + \frac{\tau}{2}\|x^{k+1} - x^k\|^2 \\ &\leq \frac{\gamma_k}{2}\|x - y^k\|^2 - \frac{\gamma_k}{2}\|x - x^{k+1}\|^2 - \frac{\gamma_k}{2}\|y^k - x^{k+1}\|^2 + \frac{\tau}{2}\|x^{k+1} - x^k\|^2, \end{aligned}$$

which implies that

$$\frac{\gamma_k}{2}\|x - x^{k+1}\|^2 \leq \frac{\gamma_k}{2}\|x - y^k\|^2 + \frac{\tau}{2}\|x^{k+1} - x^k\|^2 - \frac{\gamma_k}{2}\|y^k - x^{k+1}\|^2 - [f(x^{k+1}, B_k) - f(x, B_k)].$$

By the convexity of $f(\cdot, B_k)$, we have, for any $\eta > 0$ that

$$\begin{aligned} & f(x^{k+1}, B_k) - f(x, B_k) \\ &\geq \langle f'(x, B_k), x^{k+1} - x\rangle \\ &= \langle f'(x, B_k), x^k - x\rangle + \langle f'(x, B_k), x^{k+1} - x^k\rangle \\ &\geq \langle f'(x, B_k), x^k - x\rangle - \|f'(x, B_k)\|\|x^{k+1} - x^k\| \\ &\geq \langle f'(x, B_k), x^k - x\rangle - \frac{1}{2\eta\gamma_k}\|f'(x, B_k)\|^2 - \frac{\eta\gamma_k}{2}\|x^{k+1} - x^k\|^2. \end{aligned}$$

Recalling the identities $x - y^k = \theta(\hat{x} - z^k)$ and $x - x^{k+1} = \theta(\hat{x} - z^{k+1})$, we have

$$\begin{aligned} & \frac{\gamma_k\theta^2}{2}\|\hat{x} - z^{k+1}\|^2 \\ &= \frac{\gamma_k}{2}\|x - x^{k+1}\|^2 \\ &\leq \frac{\gamma_k}{2}\|x - y^k\|^2 + \frac{\tau}{2}\|x^{k+1} - x^k\|^2 - \frac{\gamma_k}{2}\|y^k - x^{k+1}\|^2 - [f(x^{k+1}, B_k) - f(x, B_k)] \\ &\leq \frac{\gamma_k}{2}\|x - y^k\|^2 + \frac{\tau}{2}\|x^{k+1} - x^k\|^2 - \frac{\gamma_k}{2}\|y^k - x^{k+1}\|^2 + \frac{\eta\gamma_k}{2}\|x^{k+1} - x^k\|^2 \\ &\quad - \langle f'(x, B_k), x^k - x\rangle + \frac{1}{2\eta\gamma_k}\|f'(x, B_k)\|^2 \\ &= \frac{\gamma_k\theta^2}{2}\|\hat{x} - z^k\|^2 + \frac{\tau + \eta\gamma_k}{2}\|x^{k+1} - x^k\|^2 - \frac{\gamma_k}{2}\|y^k - x^{k+1}\|^2 \\ &\quad - \langle f'(x, B_k), x^k - x\rangle + \frac{1}{2\eta\gamma_k}\|f'(x, B_k)\|^2. \end{aligned}$$

Moreover, using an argument of (70), we obtain

$$\begin{aligned} & \frac{\gamma_k}{2}\|y^k - x^{k+1}\|^2 + \frac{\tau + \eta\gamma_k}{2}\|x^{k+1} - x^k\|^2 \\ &= -\frac{\gamma_k}{2}\|y^k - x^{k+1}\|^2 + \frac{\tau + \eta\gamma_k}{2}\|x^{k+1} - x^k\|^2 \\ &\leq \frac{\gamma_k\beta(1-\beta)}{2}\|x^k - x^{k-1}\|^2 - \frac{\gamma_k(1 - \beta - \eta) - \tau}{2}\|x^{k+1} - x^k\|^2. \end{aligned}$$

Combining the above two results and taking expectation $\mathbb{E}_k[\cdot]$, we have that

$$\begin{aligned}
&\frac{\gamma_k\theta^2}{2}\mathbb{E}_k[\|\hat{x}-z^{k+1}\|^2] \\
&\leq \frac{\gamma_k\theta^2}{2}\|\hat{x}-z^k\|^2 - \mathbb{E}_k[\langle f'(x,B_k), x^k - x\rangle] + \frac{1}{2\eta\gamma_k}\mathbb{E}_k[\|f'(x,B_k)\|^2] \\
&\quad + \frac{\tau+\eta\gamma_k}{2}\mathbb{E}_k[\|x^{k+1}-x^k\|^2] - \frac{\gamma_k}{2}\mathbb{E}_k[\|y^k - x^{k+1}\|^2] \\
&\leq \frac{\gamma_k\theta^2}{2}\|\hat{x}-z^k\|^2 - \mathbb{E}_k[\langle f'(x,B_k), x^k - x\rangle] + \frac{1}{2\eta\gamma_k}\|f'(x,B_k)\|^2 + \frac{\gamma_k\beta(1-\beta)}{2}\|x^k - x^{k-1}\|^2 \\
&\quad - \frac{\gamma_k(1-\beta-\eta)-\tau}{2}\mathbb{E}_k[\|x^{k+1}-x^k\|^2].
\end{aligned}$$

Dividing both sides of the above relation by $\gamma_k\theta^2/2$ and taking take $x = x^*$ gives

$$\begin{aligned}
&\mathbb{E}_k[\|\hat{x}-z^{k+1}\|^2] \\
&\leq \|\hat{x}-z^k\|^2 - \frac{2}{\gamma_k\theta^2}\mathbb{E}_k[\langle f'(x,B_k), x^k - x^*\rangle] + \frac{1}{\eta\gamma_k^2\theta^2}\mathbb{E}_k[\|f'(x^*,B_k)\|^2] \\
&\quad + \frac{\beta(1-\beta)}{\theta^2}\|x^k - x^{k-1}\|^2 - \frac{(1-\beta-\eta)-\tau/\gamma_k}{\theta^2}\mathbb{E}_k[\|x^{k+1}-x^k\|^2] \\
&\leq \|\hat{x}-z^k\|^2 + \frac{1}{\eta\gamma_k^2\theta^2}\mathbb{E}_k[\|f'(x^*,B_k)\|^2] + \frac{\beta(1-\beta)}{\theta^2}\|x^k - x^{k-1}\|^2 \\
&\quad - \frac{(1-\beta-\eta)-\tau/\gamma_k}{\theta^2}\mathbb{E}_k[\|x^{k+1}-x^k\|^2].
\end{aligned}$$

where the last inequality uses the property $\mathbb{E}_k[\langle f'(x^*,B_k), x^k - x^*\rangle] = \langle f'(x^*), x^k - x^*\rangle \leq 0$, which is derived from optimality condition.

Last we take $\eta = \theta^2/2$, $\gamma_k \geq \frac{\tau}{(1-\beta)^2-\eta}$ such that $\beta(1-\beta) \leq (1-\beta-\eta) - \tau/\gamma_k$ and sum over $k = 1,\ldots,K$ to obtain

$$\mathbb{E}[\|\hat{x}-z^{K+1}\|^2] \leq \|\hat{x}-z^1\|^2 + \frac{\beta(1-\beta)}{\theta^2}\|x^1-x^0\|^2 + \frac{2}{\theta^4}\sum_{k=1}^{K}\gamma_k^{-2}\mathbb{E}[\|f'(x^*,B_k)\|^2].$$

Plugging $\|\hat{x}-z^{K+1}\|^2 = \frac{1}{\theta^2}\|x^*-x^{K+1}\|^2$ in the above inequality and then multiplying both sides by θ^2, we obtain the desired result. □

Remark 8. Let $\mathcal{X}^*$ be the set of optimal solutions and assume that $\sup_{x\in\mathcal{X}^*}\mathbb{E}_k[\|f'(x^*,B_k)\|^2] < \infty$. Using an argument of Cor 3.2 [1], we can show that when $\sum_{k=1}^{\infty}\gamma_k^{-2} < \infty$, then $\sup_k \text{dist}(x^k, X^*) < \infty$ with probability one. This completes our proof of the boundedness of the iterates.

It is interesting to compare SEMOD and SGD in terms of the robustness to the stepsize policy. Consider that SGD takes the form $x^{k+1} = \text{argmin}_x\langle f'(x^k,B_k), x\rangle + \frac{\gamma_k}{2}\|x-x^k\|^2$. Using the argument of [28], it is easy to show that SGD exhibits the bound

$$\mathbb{E}[\|x^*-x^{K+1}\|^2] \leq \|x^*-x^1\|^2 + \sum_{k=1}^{K}\gamma_k^{-2}\mathbb{E}[\|f'(x^k,B_k)\|^2],$$

which explicitly depends on the subgradients of iterates $\{x^k\}$. When $\|f'(x^k,B_k)\|$ is large, (e.g. f is a high order polynomial or an exponential function) we need sufficiently large $\{\gamma_k\}$ (i.e. small stepsize $1/\gamma_k$) to ensure the boundedness of iterates. However, in contrast to SGD, SEMOD has a bound only depending on the subgradient over the optimal solutions. For many problems, (e.g. interpolation problems), $\|f'(x^*,\xi)\|$ can be substantially smaller than $\sup_x\|f'(x,\xi)\|$.

We also note that the best bound for SMOD is when $\theta = 1$ (i.e. $\beta = 0$). It appears that adding momentum encourages more exploration of the parameter space, however, at the cost of potentially departing from the original solution path.

C.3 Improved convergence using Nesterov acceleration

It is known that the heavy-ball type stochastic gradient does not give an optimal rate of convergence. Next we show that our proposed stability analysis can be combined with Nesterov's acceleration [24], yielding an accelerated SMOD method which achieves the best complexity for convex stochastic optimization.

Algorithm 4 Stochastic Model-based Method with Minibatching and Nesterov's Acceleration

Input: $x^0 = z^0$;
for $k = 0$ **to** K **do**
Sample a minibatch $B_k = \{\xi_{k,1}, \dots, \xi_{k,m_k}\}$ and update y^k, z^{k+1}, x^{k+1} by

$$\begin{aligned} y^k &= (1-\theta_k)x^k + \theta_k z^k, \\ z^{k+1} &= \operatorname*{argmin}_{x\in\mathcal{X}} \Big\{ f_{y^k}(x, B_k) + \frac{\gamma_k}{2}\|x - z^k\|^2 \Big\}, \\ x^{k+1} &= (1-\theta_k)x^k + \theta_k z^{k+1}. \end{aligned}$$

end for

Lemma C.4. *Let $\Delta_k \triangleq f(x^k) - f(x)$ for some $x \in \mathcal{X}$. For $k = 0, 1, 2, \dots$ we have*

$$\begin{aligned} &\mathbb{E}_k\big[\Delta_{k+1}\big] - (1-\theta_k)\Delta_k \\ &\le \frac{2L^2\theta_k}{m_k\gamma_k} + \frac{\gamma_k\theta_k}{2}\|x - z^k\|^2 - \frac{\gamma_k\theta_k}{2}\mathbb{E}_k[\|x - z^{k+1}\|^2] \\ &\quad - \frac{\gamma_k\theta_k - \tau\theta_k^2}{2}\mathbb{E}_k[\|z^k - z^{k+1}\|^2]. \end{aligned} \tag{74}$$

Proof. First, recall that $f_y(x) = \mathbb{E}_\xi[f_y(x,\xi)]$. Assumption A6 implies that for any $x, y \in \mathcal{X}$, we have

$$f(x) = \mathbb{E}_\xi[f(x,\xi)] \le \mathbb{E}_\xi\big[f_y(x,\xi) + \frac{\tau}{2}\|x-y\|^2\big] = f_y(x) + \frac{\tau}{2}\|x-y\|^2.$$

Therefore, we deduce that

$$\begin{aligned} f(x^{k+1}) &\le f_{y^k}(x^{k+1}) + \frac{\tau}{2}\|x^{k+1} - y^k\|^2 \\ &= f_{y^k}\big((1-\theta_k)x^k + \theta_k z^{k+1}\big) + \frac{\tau\theta_k^2}{2}\|z^{k+1} - z^k\|^2 \\ &\le (1-\theta_k)f_{y^k}(x^k) + \theta_k f_{y^k}(z^{k+1}) + \frac{\tau\theta_k^2}{2}\|z^{k+1} - z^k\|^2 \\ &\le (1-\theta_k)f(x^k) + \theta_k f_{y^k}(z^{k+1}) + \frac{\tau\theta_k^2}{2}\|z^{k+1} - z^k\|^2 \\ &= (1-\theta_k)f(x^k) + \theta_k f_{y^k}(z^{k+1}, B_k) + \frac{\tau\theta_k^2}{2}\|z^{k+1} - z^k\|^2 \\ &\quad + \theta_k\big[f_{y^k}(z^{k+1}) - f_{y^k}(z^{k+1}, B_k)\big] \end{aligned} \tag{75}$$

where the equality uses the fact $\theta_k(z^{k+1} - z^k) = x^{k+1} - y^k$, the third inequality uses Assumption A6 again. Moreover, due to the optimality of z^{k+1} for the subproblem, for any $x \in \mathcal{X}$, we have

$$\begin{aligned} f_{y^k}(z^{k+1}, B_k) &\le f_{y^k}(x, B_k) + \frac{\gamma_k}{2}\|x - z^k\|^2 - \frac{\gamma_k}{2}\|x - z^{k+1}\|^2 - \frac{\gamma_k}{2}\|z^k - z^{k+1}\|^2 \\ &\le f(x, B_k) + \frac{\gamma_k}{2}\|x - z^k\|^2 - \frac{\gamma_k}{2}\|x - z^{k+1}\|^2 - \frac{\gamma_k}{2}\|z^k - z^{k+1}\|^2 \end{aligned} \tag{76}$$

where the second inequality uses Assumption A6. Following (76) and (75), we obtain

$$\begin{aligned} f(x^{k+1}) &\le (1-\theta_k)f(x^k) + \theta_k f(x, B_k) + \theta_k\big[f_{y^k}(z^{k+1}) - f_{y^k}(z^{k+1}, B_k)\big] \\ &\quad + \frac{\gamma_k\theta_k}{2}\|x - z^k\|^2 - \frac{\gamma_k\theta_k}{2}\|x - z^{k+1}\|^2 - \frac{\gamma_k\theta_k - \tau\theta_k^2}{2}\|z^k - z^{k+1}\|^2. \end{aligned} \tag{77}$$

On both sides of (77), we take expectation over B_k conditioned on $B_1, B_2, \ldots, B_{k-1}$. Noting that $\mathbb{E}_k[f(x, B_k)] = f(x)$, we have that

$$\begin{aligned}
&\mathbb{E}_k\big[f(x^{k+1}) - f(x)\big] - (1-\theta_k)\big[f(x^k) - f(x)\big] \\
\leq\ & \frac{\gamma_k\theta_k}{2}\|x - z^k\|^2 - \frac{\gamma_k\theta_k}{2}\mathbb{E}_k[\|x - z^{k+1}\|^2] - \frac{\gamma_k\theta_k - \tau\theta_k^2}{2}\mathbb{E}_k[\|z^k - z^{k+1}\|^2] \\
&+ \theta_k\mathbb{E}_k\big[f_{y^k}(z^{k+1}) - f_{y^k}(z^{k+1}, B_k)\big].
\end{aligned} \tag{78}$$

Moreover, based on the stability of proximal mapping, we have that

$$\mathbb{E}_k\big[f_{y^k}(z^{k+1}) - f_{y^k}(z^{k+1}, B_k)\big] = \mathbb{E}_k\big\{\mathbb{E}_\xi\big[f_{y^k}(z^{k+1}, \xi) - f_{y^k}(z^{k+1}, B_k)\big]\big\} \leq \frac{2L^2}{m_k\gamma_k}. \tag{79}$$

Combining the above two results together immediately gives us the desired result (74). □

Theorem C.5. *In Algorithm 4, let the sequence* $\{\Gamma_k\}$*,*

$$\Gamma_k = \begin{cases} (1-\theta_k)^{-1}\Gamma_{k-1} & \text{if } k > 0 \\ 1 & \text{if } k = 0 \end{cases} \tag{80}$$

and assume that Γ_k*,* γ_k*, and* θ_k *satisfy*

$$\Gamma_k\gamma_k\theta_k \geq \Gamma_{k+1}\gamma_{k+1}\theta_{k+1}, \tag{81}$$

$$\gamma_k \geq \tau\theta_k, \tag{82}$$

then we have

$$\Gamma_K\mathbb{E}[\Delta_{K+1}] \leq (1-\theta_0)\Delta_0 + \frac{\Gamma_0\gamma_0\theta_0^2}{2}\|x - z^0\|^2 + \sum_{k=0}^{K}\frac{2L^2\Gamma_k\theta_k}{m_k\gamma_k}. \tag{83}$$

Moreover, if we take $x = x^*$ *be an optimal solution, and assume that* $m_k = m$*,* $\theta_k = \frac{2}{k+2}$*,* $\gamma_k = \frac{\gamma}{k+1}$*,* $\gamma = 2\tau + \eta$*,* $\eta = \frac{2L}{\sqrt{3m}\tilde{D}}(K+2)^{\frac{3}{2}}$ *where* $\tilde{D} \geq \|x^0 - x^*\|$*, then we have*

$$\mathbb{E}\big[f(x^{K+1}) - f(x^*)\big] \leq \frac{2\tau\tilde{D}^2}{(K+1)(K+2)} + \frac{4\sqrt{2}L\tilde{D}}{\sqrt{3m(K+1)}}. \tag{84}$$

Proof. First of all, it can be easily checked that conditions (81) and (82) are satisfied by the proposed setting of θ_k and γ_k. Next, multiplying both sides of (74) by Γ_k, and then dropping out the negative term $-\frac{\gamma_k\theta_k - \tau\theta_k^2}{2}\Gamma_k\mathbb{E}_k[\|z^k - z^{k+1}\|^2]$ in the result, we have

$$\begin{aligned}
&\Gamma_k\mathbb{E}_k\big[\Delta_{k+1}\big] - \Gamma_{k-1}\Delta_k \\
\leq\ & \frac{2L^2\Gamma_k\theta_k}{m_k\gamma_k} + \frac{\Gamma_k\gamma_k\theta_k}{2}\|x - z^k\|^2 - \frac{\Gamma_k\gamma_k\theta_k}{2}\mathbb{E}_k[\|x - z^{k+1}\|^2]
\end{aligned}$$

Summing up the above result over $k = 0, 1, 2, ..., K$ and taking expectation over all the randomness, we obtain the desired result (83).

Moreover, note that $\theta_0 = 1$, $\Gamma_k = \frac{(k+2)(k+1)}{2}$, hence we have

$$\sum_{k=0}^{K}\frac{2L^2\Gamma_k\theta_k}{m_k\gamma_k} = \sum_{k=0}^{K}\frac{2L^2(k+1)^2}{m\gamma} \leq \frac{2L^2}{m\gamma}\int_1^{K+2} s^2\mathrm{d}s \leq \frac{2L^2}{3m\gamma}(K+2)^3. \tag{85}$$

Placing $x = x^*$, then we have

$$\begin{aligned}
\mathbb{E}\big[f(x^{K+1}) - f(x^*)\big] &\le \Gamma_K^{-1}\Big\{(1-\theta_0)\Delta_0 + \frac{\Gamma_0\gamma_0\theta_0^2}{2}\|x - z^0\|^2 + \sum_{k=0}^{K}\frac{2L^2\Gamma_k\theta_k}{m_k\gamma_k}\Big\} \\
&\le \Gamma_K^{-1}\Big\{\frac{\gamma}{2}\tilde{D}^2 + \frac{2L^2}{3m\gamma}(K+2)^3\Big\} \\
&= \frac{1}{K+1}\Big\{\frac{\gamma\tilde{D}^2}{K+2} + \frac{4L^2(K+2)^2}{3m\gamma}\Big\} \\
&\le \frac{2\tau\tilde{D}^2}{(K+1)(K+2)} + \frac{1}{K+1}\Big\{\frac{\eta\tilde{D}^2}{K+2} + \frac{4L^2(K+2)^2}{3m\eta}\Big\} \\
&= \frac{2\tau\tilde{D}^2}{(K+1)(K+2)} + \frac{4L\tilde{D}}{K+1}\sqrt{\frac{K+2}{3m}} \\
&\le \frac{2\tau\tilde{D}^2}{(K+1)(K+2)} + \frac{4\sqrt{2}L\tilde{D}}{\sqrt{3m(K+1)}}.
\end{aligned}$$

where the second inequality uses (85), and $\tilde{D} \ge \|x^0 - x^*\|$, the third inequality uses the fact $\gamma = 2\tau + \eta$ and $\frac{1}{\gamma} \le \frac{1}{\eta}$, and the last inequality uses $K + 2 \le 2(K+1)$ for $K \ge 1$. This completes the proof. □

D Solving the subproblems

In this section, we describe how to solve the subproblems arising from (SGD), (SPL) and(SPP). For the sake of simplicity, we abstract the SMOD subproblems by

$$\min_{x \in \mathbb{R}^d} \quad \frac{1}{m}\sum_{i=1}^{m}\varphi_z(x, \xi_i) + \frac{\gamma}{2}\|x - y\|^2 \tag{86}$$

D.1 Phase retrieval

We first state the expressions for the sequential updates (i.e. $m = 1$). More technical derivations can be referred from [3]. Given current iterate x, we denote x^+ to be the output of SMOD update and suppress all the iteration indices.

In terms of phase retrieval, let $\xi = (a, b)$ for $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$ and we have

$$\begin{aligned}
x_{\text{sgd}}^+ &= \operatorname*{argmin}_x \Big\{\langle v, x - z\rangle + \frac{\gamma}{2}\|x - y\|^2\Big\} \\
x_{\text{spl}}^+ &= \operatorname*{argmin}_x \Big\{\big|\langle a, z\rangle^2 + 2\langle a, z\rangle\langle a, x - z\rangle - b\big| + \frac{\gamma}{2}\|x - y\|^2\Big\} \\
x_{\text{spp}}^+ &= \operatorname*{argmin}_x \Big\{|\langle a, x\rangle^2 - b| + \frac{\gamma}{2}\|x - y\|^2\Big\}.
\end{aligned}$$

The above three subproblems admit closed-form solutions

$$\begin{aligned}
x_{\text{sgd}}^+ &= y - \gamma^{-1}v \\
x_{\text{spl}}^+ &= y + \text{Proj}_{[-1,1]}\Big(-\tfrac{\delta}{\|\zeta\|^2}\Big)\zeta \\
x_{\text{spp}}^+ &\in \Big\{y - \tfrac{2\langle a,y\rangle a}{2\|a\|^2 \pm \gamma}, y - \tfrac{\langle a,y\rangle \pm \sqrt{b}}{\|a\|^2}a\Big\},
\end{aligned}$$

where

$$\begin{aligned}
v &\in \partial_x(|\langle a, z\rangle^2 - b|) \\
&= 2\langle a, z\rangle a \cdot \begin{cases} \text{sign}(\langle a, z\rangle^2 - b), & \text{if } \langle a, z\rangle^2 - b \ne 0 \\ [-1, 1], & \text{o.w.} \end{cases} \\
\delta &= \gamma^{-1}(\langle a, z\rangle^2 + 2\langle a, z\rangle\langle a, x - z\rangle - b), \\
\zeta &= 2\gamma^{-1}\langle a, z\rangle a
\end{aligned}$$

and $\mathrm{Proj}_{[-1,1]}(\cdot)$ denotes the orthogonal projection operator onto $[-1,1]$.

For minibatching, we let $y = z$ and

$$
\begin{aligned}
x_{\text{sgd}}^{+} &= \operatorname*{argmin}_{x} \left\{ \frac{1}{m}\sum_{i=1}^{m} \langle v_i, x - z\rangle + \frac{\gamma}{2}\|x-z\|^2 \right\} \\
x_{\text{spl}}^{+} &= \operatorname*{argmin}_{x} \left\{ \frac{1}{m}\sum_{i=1}^{m} \left|\langle a_i, z\rangle^2 - b_i + 2\langle a_i, z\rangle\langle a_i, x-z\rangle\right| + \frac{\gamma}{2}\|x-z\|^2 \right\} \\
x_{\text{spp}}^{+} &= \operatorname*{argmin}_{x} \left\{ \frac{1}{m}\sum_{i=1}^{m} |\langle a_i, x\rangle^2 - b_i| + \frac{\gamma}{2}\|x-z\|^2 \right\},
\end{aligned}
$$

where $v_i \in \partial_x(|\langle a_i, z\rangle^2 - b_i|)$. Minibatch subproblems can be reformulated as standard convex programs.

$$
x_{\text{sgd}}^{+} = z - \frac{1}{m\gamma}\sum_{i=1}^{m} v_i \tag{87}
$$

$$
\begin{aligned}
\left(x_{\text{spl}}^{+}, *\right) = \operatorname*{argmin}_{(x,t)} & \left\{ \frac{1}{m}\sum_{i=1}^{m} t_i + \frac{\gamma}{2}\|x-z\|^2 \right\} \\
\text{subject to} \quad & \langle a_i, z\rangle^2 - b_i + 2\langle a_i, z\rangle\langle a_i, x-z\rangle \geq -t_i \qquad (88) \\
& \langle a_i, z\rangle^2 - b_i + 2\langle a_i, z\rangle\langle a_i, x-z\rangle \leq t_i \; i = 1, 2, \ldots, m. \qquad (89)
\end{aligned}
$$

$$
\begin{aligned}
\left(x_{\text{spp}}^{+}, *\right) = \operatorname*{argmin}_{(x,t)} & \left\{ \frac{1}{m}\sum_{i=1}^{m} t_i \right\} \\
\text{subject to} \quad & x^{\mathrm{T}}\left(\frac{\gamma}{2}I - a_i a_i^{\mathrm{T}}\right)x - \gamma\langle z, x\rangle + \frac{\gamma}{2}\|z\|^2 + b_i \leq t_i \\
& x^{\mathrm{T}}\left(\frac{\gamma}{2}I + a_i a_i^{\mathrm{T}}\right)x - \gamma\langle z, x\rangle + \frac{\gamma}{2}\|z\|^2 - b_i \leq t_i, \; i = 1, 2, \ldots, m \qquad (90)
\end{aligned}
$$

Remark 9. We make a few comments. First, the update (`SGD`) (87) admits a simple closed-form solution by directly using the average subgradients over the minibatches. Second, the `SPL` subproblem (89) can be further transformed into an $O(m)$-dimensional quadratic program in the dual form, which can be efficiently solved in parallel. (See [1]). Third, the `SPP` subproblem (90) is solvable by interior point methods for quadratically constrained quadratic programming (QCQP).

However, despite the fast theoretical convergence, interior point methods are potentially unscalable the growing number of nonlinear constraints. In our initial experiments, we apply Gurobi for solving (90) but fail to get an accurate solution to subproblems when $m > 5$. Therefore, we alternatively utilize the strong convexity of (90) and adopt deterministic prox-linear algorithm to obtain an accurate solution (up to `1e-08` accuracy) by solving several QPs as in (89). The theoretical linear convergence of this method is verified in D.3. Finally, similar observations can be made for the experiments of blind deconvolution.

D.2 Blind deconvolution

The detailed formulation of blind deconvolution is deferred to Section 6 and we focus on its proximal subproblems here. For convenience we use $(x; y)$ to denote the vertical concatenation of two column

vectors. Given current iterate $w = (w_x; w_y)$, the subproblems are given by

$$
\begin{aligned}
w_{\text{sgd}}^+ &= \operatorname*{argmin}_{(x;y)} \left\{ \langle s, (x - z_x; y - z_y) \rangle + \frac{\gamma}{2}\|x - w_x\|^2 + \frac{\gamma}{2}\|y - w_y\|^2 \right\} \\
w_{\text{spl}}^+ &= \operatorname*{argmin}_{(\Delta_x;\Delta_y)} \Big\{ |\langle u, z_x\rangle\langle v, z_y\rangle + \langle v, z_y\rangle\langle u, \Delta_x\rangle + \langle u, z_x\rangle\langle v, \Delta_y\rangle \\
&\qquad + \langle v, z_y\rangle\langle u, w_x - z_x\rangle + \langle u, z_x\rangle\langle v, w_y - z_y\rangle - b| + \frac{\gamma}{2}[\|\Delta_x\|^2 + \|\Delta_y\|^2] \Big\} + w \\
w_{\text{spp}}^+ &= \operatorname*{argmin}_{(x;y)} \left\{ |\langle u, x\rangle\langle v, y\rangle - b| + \frac{\gamma}{2}\|x - w_x\|^2 + \frac{\gamma}{2}\|y - w_y\|^2 \right\}
\end{aligned}
$$

and we have

$$
\begin{aligned}
w_{\text{sgd}}^+ &= w - \gamma^{-1} s \\
w_{\text{spl}}^+ &= w + \operatorname{Proj}_{[-1,1]}\left(-\tfrac{\delta}{\|\zeta\|^2}\right)\zeta
\end{aligned}
$$

where

$$
\begin{aligned}
s &\in \partial_{(x;y)}(|\langle u, z_x\rangle\langle v, z_y\rangle - b|) \\
&= (\langle v, z_y\rangle u; \langle u, z_x\rangle v) \cdot \begin{cases} \operatorname{sign}(\langle u, z_x\rangle\langle v, z_y\rangle - b), & \text{if } \langle u, z_x\rangle\langle v, z_y\rangle - b \neq 0 \\ [-1, 1], & \text{o.w.} \end{cases} \\
\delta &= \gamma^{-1}\left[\langle u, z_x\rangle\langle v, z_y\rangle + \langle v, z_y\rangle\langle u, w_x - z_x\rangle + \langle u, z_x\rangle\langle v, w_y - z_y\rangle - b\right] \\
\zeta &= \gamma^{-1}\left(\langle v, z_y\rangle u; \langle u, z_x\rangle v\right).
\end{aligned}
$$

As for SPP, we consider the following two cases.

Case 1. If $\langle u, w_x\rangle\langle v, w_y\rangle - b \neq 0$, then

$$
w_x^+ = w_x - \left\{ \tfrac{\pm\gamma\langle v, w_y\rangle - \|v\|^2\langle u, w_x\rangle}{\gamma^2 - \|u\|^2\|v\|^2} \right\} u, \quad w_y^+ = w_y - \left\{ \tfrac{\pm\gamma\langle u, w_x\rangle - \|u\|^2\langle v, w_y\rangle}{\gamma^2 - \|u\|^2\|v\|^2} \right\} v.
$$

Case 2. If $\langle u, w_x\rangle\langle v, w_y\rangle - b = 0$, then

$$
w_x^+ = w_x - \zeta\left(\tfrac{b}{\eta}\right) u, \quad w_y^+ = w_y - \zeta\eta v,
$$

where $\zeta = \frac{\eta\langle u, w_x\rangle - \eta^2}{b\|u\|^2}$ and η is determined by

$$
\eta^4\|v\|^2 - \eta^3\|v\|^2\langle u, w_x\rangle + b\eta\|u\|^2\langle v, w_y\rangle - b^2\|u\|^2 = 0.
$$

Moreover, for the minibatch variants, we set $w = z$ and get the following subproblems

$$
\begin{aligned}
w_{\text{sgd}}^+ &= \operatorname*{argmin}_{(x;y)} \left\{ \frac{1}{m}\sum_{i=1}^m \langle s_i, (x - z_x; y - z_y) \rangle + \frac{\gamma}{2}\|x - z_x\|^2 + \frac{\gamma}{2}\|y - z_y\|^2 \right\} \\
w_{\text{spl}}^+ &= \operatorname*{argmin}_{(\Delta_x;\Delta_y)} \left\{ \frac{1}{m}\sum_{i=1}^m |\langle u_i, z_x\rangle\langle v_i, z_y\rangle + \langle v_i, z_y\rangle\langle u_i, \Delta_x\rangle + \langle w_i, z_x\rangle\langle v_i, \Delta_y\rangle - b_i| \right. \\
&\qquad \left. + \frac{\gamma}{2}\|\Delta_x\|^2 + \frac{\gamma}{2}\|\Delta_y\|^2 \right\} + z \\
w_{\text{spp}}^+ &= \operatorname*{argmin}_{(x;y)} \left\{ \frac{1}{m}\sum_{i=1}^m |\langle u_i, x\rangle\langle v_i, y\rangle - b_i| + \frac{\gamma}{2}\|x - z_x\|^2 + \frac{\gamma}{2}\|y - z_y\|^2 \right\},
\end{aligned}
$$

where $s_i \in \partial_{(x;y)}(|\langle u_i, z_x\rangle\langle v_i, z_y\rangle - b_i|)$. Then we solve the subproblems by

$$w_{\text{sgd}}^+ = z - \frac{1}{m\gamma}\sum_{i=1}^m s_i,$$

$$\begin{aligned}(x_{\text{spl}}^+; y_{\text{spl}}^+, *) = \operatorname*{argmin}_{(x,y,t)} & \left\{\frac{1}{m}\sum_{i=1}^m t_i + \frac{\gamma}{2}\|x - z_x\|^2 + \frac{\gamma}{2}\|y - z_y\|^2\right\} \\ \text{subject to} \quad & \langle u_i, z_x\rangle\langle v_i, z_y\rangle + \langle v_i, z_y\rangle\langle u_i, x - z_x\rangle + \langle u_i, z_x\rangle\langle v_i, y - z_y\rangle - b_i \le t_i \\ & \langle u_i, z_x\rangle\langle v_i, z_y\rangle + \langle v_i, z_y\rangle\langle u_i, x - z_x\rangle + \langle u_i, z_x\rangle\langle v_i, y - z_y\rangle - b_i \ge -t_i, \\ & i = 1, 2, \ldots, m\end{aligned}$$

$$\begin{aligned}(x_{\text{spp}}^+; y_{\text{spp}}^+, *) = \operatorname*{argmin}_{(x,y,t)} & \left\{\frac{1}{m}\sum_{i=1}^m t_i\right\} \\ \text{subject to} \quad & \frac{\gamma}{2}[\|x - z_x\|^2 + \|y - z_y\|^2] + \langle u_i; x\rangle\langle v_i, y\rangle - b_i \le t_i \\ & \frac{\gamma}{2}[\|x - z_x\|^2 + \|y - z_y\|^2] - \langle u_i, x\rangle\langle v_i, y\rangle + b_i \le t_i, \quad i = 1, 2, \ldots, m,\end{aligned}$$

where the last two problems are solved by either QP (QCQP) optimizers or prox-linear iterations as in phase retrieval.

D.3 Solving the SPP subproblem by Prox-linear algorithm

Suppose that the objective admits a composition form $h(c(\cdot))$. We show that when applied to the SPP subproblem, the deterministic prox-linear algorithm obtains a linear convergence rate. Without loss of generality, consider the SPP subproblem

$$\min_{x \in \mathcal{X}} \quad \frac{1}{m}\sum_{i=1}^m \varphi_{\bar{x}}(x, \xi_i) + \frac{\gamma}{2}\|x - \bar{x}\|^2 \tag{91}$$

where $\varphi_{\bar{x}}(x, \xi_i) = h(c(x, \xi_i))$ and we apply deterministic prox-linear method to solve the above subproblem. For clarity we denote z^t to be the iterate of the subproblems and define $\varphi_{z^t}(z) := \frac{1}{m}\sum_{i=1}^m h(c(z^t, \xi_i) + \langle\varphi(\nabla c(z^t, \xi_i), (z - z^t)\rangle)$, $\varphi(z) := \frac{1}{m}\sum_{i=1}^m h(c(x, \xi_i))$. In each prox-linear iteration, we tate $\eta \ge \tau$ and compute

$$z^{t+1} = \arg\min_z \left\{\varphi(z) + \frac{\gamma}{2}\|z - \bar{x}\|^2 + \frac{\eta}{2}\|z - z^t\|^2\right\}.$$

First by A5 we have

$$\begin{aligned}\varphi_{z^t}(z) - \varphi(z) &\le \frac{\tau}{2}\|z - z^t\|^2 \\ \varphi(z^{t+1}) - \varphi_{z^t}(z^{t+1}) &\le \frac{\tau}{2}\|z^{t+1} - z^t\|^2\end{aligned}$$

and by the strong convexity of subproblems we have

$$\begin{aligned}&\varphi_{z^t}(z^{t+1}) + \frac{\gamma}{2}\|z^{t+1} - \bar{x}\|^2 + \frac{\eta}{2}\|z^{t+1} - z^t\|^2 \\ &\le \varphi_{z^t}(z) + \frac{\gamma}{2}\|z - \bar{x}\|^2 + \frac{\eta}{2}\|z - z^t\|^2 - \frac{\gamma + \eta - \lambda}{2}\|z^{t+1} - z\|^2,\end{aligned}$$

which implies

$$\begin{aligned}&\varphi_{z^t}(z^{t+1}) + \frac{\gamma}{2}\|z^{t+1} - \bar{x}\|^2 + \frac{\eta}{2}\|z^{t+1} - z^t\|^2 + \varphi_{z^t}(z) - \varphi(z) + \varphi(z^{t+1}) - \varphi_{z^t}(z^{t+1}) \\ &\le \varphi_{z^t}(z) + \frac{\gamma}{2}\|z - \bar{x}\|^2 + \frac{\eta}{2}\|z - z^t\|^2 - \frac{\gamma + \eta - \lambda}{2}\|z^{t+1} - z\|^2 + \frac{\tau}{2}\|z - z^t\|^2 + \frac{\tau}{2}\|z^{t+1} - z^t\|^2.\end{aligned}$$

Re-arranging the terms, we have

$$\begin{aligned}
&\Big[\varphi(z^{t+1}) + \frac{\gamma}{2}\|z^{t+1} - \bar{x}\|^2\Big] - \Big[\varphi(z) + \frac{\gamma}{2}\|z - \bar{x}\|^2\Big] \\
&\leq \frac{\eta+\tau}{2}\|z - z^t\|^2 - \frac{\gamma+\eta-\lambda}{2}\|z^{t+1} - z\|^2 + \frac{\tau-\eta}{2}\|z^{t+1} - z^t\|^2 \\
&\leq \frac{\eta+\tau}{2}\|z - z^t\|^2 - \frac{\gamma+\eta-\lambda}{2}\|z^{t+1} - z\|^2, \qquad (92)
\end{aligned}$$

where the last inequality is by $\eta \geq \tau$. Define $\alpha = \frac{\eta+\tau}{\gamma+\eta-\lambda}$ and divide both sides of the inequality by $\left(\frac{\eta+\tau}{2}\right)\alpha^t$, we obtain

$$\begin{aligned}
&\frac{2}{\alpha^t(\eta+\tau)}\left\{\Big[\varphi(z^{t+1}) + \frac{\gamma}{2}\|z^{t+1} - \bar{x}\|^2\Big] - \Big[\varphi(z) + \frac{\gamma}{2}\|z - \bar{x}\|^2\Big]\right\} \\
&\leq \frac{1}{\alpha^t}\|z^t - z\|^2 - \frac{1}{\alpha^t}\cdot\frac{\gamma+\eta-\lambda}{\eta+\tau}\|z^{t+1} - z\|^2 \\
&= \frac{1}{\alpha^t}\|z^t - z\|^2 - \frac{1}{\alpha^{t+1}}\|z^{t+1} - z\|^2.
\end{aligned}$$

Last we define $\Delta_t := \big[\varphi(z^{t+1}) + \frac{\gamma}{2}\|z^{t+1} - \bar{x}\|^2\big] - \big[\varphi(z^*) + \frac{\gamma}{2}\|z^* - \bar{x}\|^2\big]$ and by (92) we can verify that $\{\Delta_k\}$ is monotonically decreasing. By taking $z = z^*$ and summing over $t = 0, \ldots, T$, we get

$$\begin{aligned}
\sum_{t=0}^{T}\frac{2\Delta_T}{\alpha^t(\eta+\tau)} &\leq \sum_{t=0}^{T}\frac{2\Delta_t}{\alpha^t(\eta+\tau)} \\
&\leq \sum_{t=0}^{T}\frac{1}{\alpha^t}\|z^t - z^*\|^2 - \frac{1}{\alpha^{t+1}}\|z^{t+1} - z^*\|^2 \\
&= \|z^0 - z^*\|^2 - \frac{1}{\alpha^{T+1}}\|z^{T+1} - z^*\|^2 \\
&\leq \|z^0 - z^*\|^2
\end{aligned}$$

and we have

$$\Delta_T \leq \frac{(\eta+\tau)\|z^0 - z^*\|^2}{2\big(\sum_{t=0}^{T} 1/\alpha^t\big)} \leq \frac{(\eta+\tau)\|z^0 - z^*\|^2}{2}\left(\frac{\eta+\tau}{\gamma+\eta-\lambda}\right)^T$$

and this implies linear convergence.

E Additional experiments

This section presents the experiments that were not displayed in the main article due to space limit.

E.1 Blind deconvolution

Blind deconvolution aims to separate two unknown signals from their convolution, resulting in the following non-smooth biconvex problem

$$\min_{x,\,y\,\in\,\mathbb{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n}\big|\langle u_i, x\rangle\langle v_i, y\rangle - b_i\big|. \qquad (93)$$

Data preparation. We conduct experiments over synthetic dataset.

1) Synthetic data. We choose n, d and the signal x^* in the same way as in phase retrieval. Namely we generate $U = Q_1 D_1, V = Q_2 D_2$ where $q_{ij} \sim \mathcal{N}(0,1)$ and D_1, D_2 are diagonal matrices whose diagonal entries evenly distribute between 1 and $1/\kappa$; Measurements $\{b_i\}$ are generated by $b_i = \langle u_i, x^*\rangle\langle v_i, x^*\rangle + \delta_i\zeta_i$ with $\zeta_i \sim \mathcal{N}(0, 25)$ and $\delta \sim \mathrm{Bernoulli}(p_{\mathrm{fail}})$

The detailed experiment setup is given as follows

1) Dataset generation. We test $\kappa \in \{1, 10\}$ and $p_{\text{fail}} \in \{0.2, 0.3\}$;

2) Initial point. For all algorithms, we set the initial point $x^1 (= x^0)$ and $y^1 (= y^0) \sim \mathcal{N}(0, I_d)$;

3) Stepsize. We set the parameter $\gamma = \alpha_0^{-1}\sqrt{K/m}$ where m is the batch size; we test 10 evenly spaced α_0 values in range $[10^{-1}, 10^2]$ for `SGD`, `SPL` and in range $[10^{-2}, 10^1]$ for `SGD`, `SPL` and `SPP`;

4) Others. The rest of the experiment setup are the same as in synthetic phase retrieval, which can be referred from Section 6.

In our setup, when we take $\alpha_0 \geq 10$, the resultant `SPP` subproblem will remain nonconvex. Thus we present the results with α_0 in two ranges for `SPP` and the other two `SMOD` algorithms.

In Figure 5 we plot the the algorithm speedup over the size of minibatches for two different settings $p_{\text{fail}} \in \{0.2, 0.3\}$. We find that both `SPL` and `SGD` enjoy linear speedup over the size of minibatches. Figure 6 shows the algorithm speedup over different values of α_0. In comparison with `SGD`, `SPL` has significant acceleration over a much wider range of stepsize values. Figure 7 shows the total iteration number over different values of α_0. The result suggests that momentum can further improve the performance of both stochastic algorithms, particularly if algorithms are initiated with small stepsizes.

Figure 5: Speedup vs. batch size m. $\kappa = 10$. From left to right: $(\alpha_0, p_{\text{fail}}) = ([10^{-2}, 10], 0.2), ([10^{-1}, 10^2], 0.2), ([10^{-2}, 10], 0.3), ([10^{-1}, 10^2], 0.3)$.

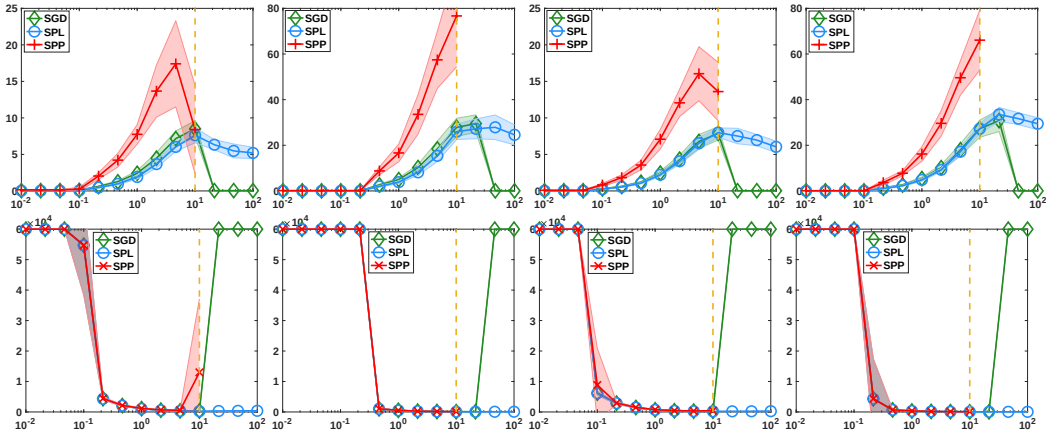

Figure 6: First row: Speedup vs. Stepsize α_0. Second row: Iteration number on reaching desired accuracy vs. Stepsize α_0. From left to right: $\kappa = 10, (p_{\text{fail}}, m) = (0.2, 8), (0.2, 32), (0.3, 8), (0.3, 32)$.

E.2 Phase retrieval

We complement the experiments in Section 6 by visualizing the effectiveness of image recovery on `zipcode` datasets. Details on data processing and parameter settings are available in Section 6.

More detailedly, we conduct experiments on the test images of digit 6 and illustrate the results of `SPL` and `SGD` in Figure 8 and Figure 9, respectively. We fix $\alpha_0 = 100$ and run each algorithm over 200 epochs (number of passes over the data). Then we report the results over the earliest 600 iterations and plot the recovered digits for different batch sizes $m \in \{1, 4, 8, 16, 32, 48, 64\}$. It can be seen that

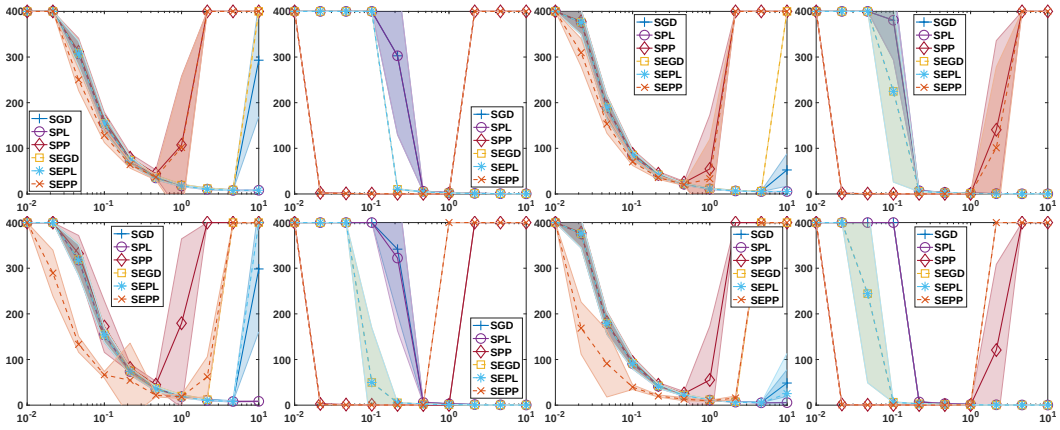

Figure 7: Epoch number on reaching desired accuracy vs. Stepsize α_0. First row: $\beta = 0.2$. Second row: $\beta = 0.6$. From left to right: $\kappa = 10, (p_{\text{fail}}, m) = (0.2, 1), (0.2, 32), (0.3, 1), (0.3, 32)$.

with larger batch size, both methods exhibit improved performance and generate images with better quality, which suggests the practical advantage of using large batch size. Moreover, SPL outperforms SGD by giving a much better recovered image quality. This observation confirms the earlier study about the superior performance of prox-linear methods [5].

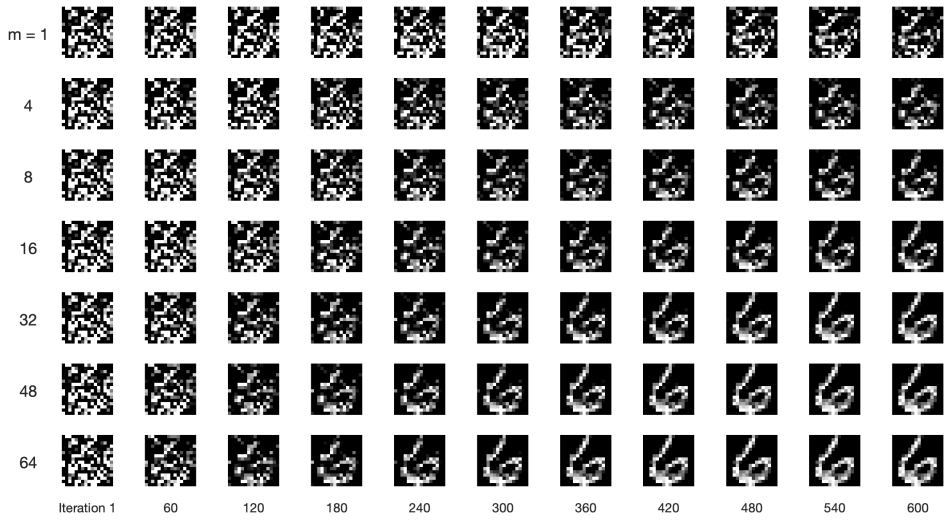

Figure 8: Reconstruction of real image (digit 6) for stochastic prox-linear method. Rows correspond to recovery results of different minibatch size $m \in \{1, 4, 8, 16, 32, 48, 64\}$. Columns correspond to recovery results after different number of iterations $T \in \{1, 60, 120, 180, 240, 300, 360, 420, 480, 540, 600\}$.

Reference

[1] H. Asi, K. Chadha, G. Cheng, and J. C. Duchi. Minibatch stochastic approximate proximal point methods. *Advances in Neural Information Processing Systems*, 33, 2020.

[2] K. Chadha, G. Cheng, and J. C. Duchi. Accelerated, optimal, and parallel: Some results on model-based stochastic optimization. *arXiv preprint arXiv:2101.02696*, 2021.

[3] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *Siam Journal on Optimization*, 29(1):207–239, 2019.

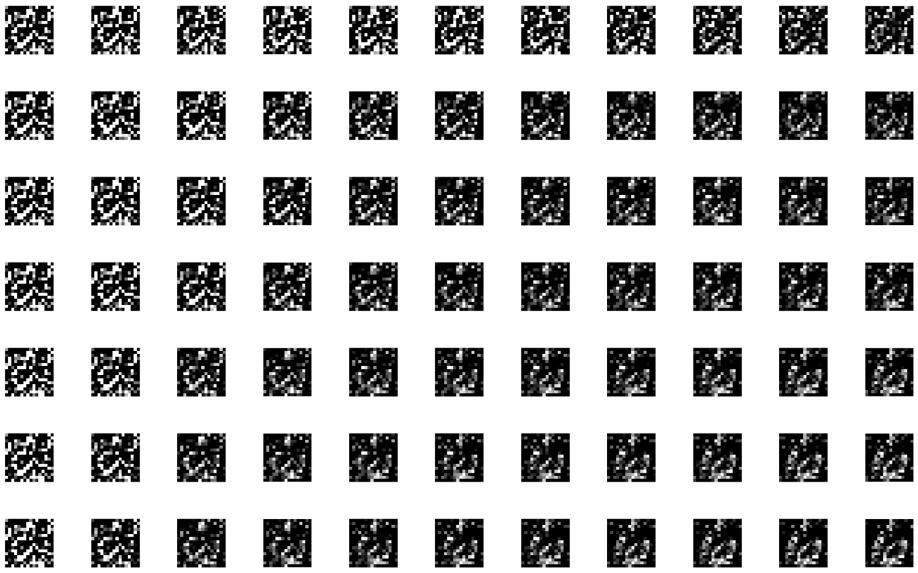

Figure 9: Reconstruction of real image (digit 6) for stochastic (sub)gradient descent.

[4] J. Diakonikolas and M. I. Jordan. Generalized momentum-based methods: a hamiltonian perspective. *SIAM Journal on Optimization*, 31(1):915–944, 2021.

[5] J. C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2019.

[6] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.

[7] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.