Convergence and Stability of the Stochastic Proximal Point Algorithm with Momentum

Junhyung Lyle Kim JLYLEKIM@RICE.EDU
Department of Computer Science, Rice University

Panos Toulis PANOS.TOULIS@CHICAGOBOOTH.EDU
Booth School of Business, University of Chicago

Anastasios Kyrillidis ANASTASIOS@RICE.EDU
Department of Computer Science, Rice University

Abstract

Stochastic gradient descent with momentum (SGDM) is the dominant algorithm in many optimization scenarios, including convex optimization instances and non-convex neural network training. Yet, in the stochastic setting, momentum interferes with gradient noise, often leading to specific step size and momentum choices in order to guarantee convergence, set aside acceleration. Proximal point methods, on the other hand, have gained much attention due to their numerical stability and elasticity against imperfect tuning. Their stochastic accelerated variants though have received limited attention: how momentum interacts with the stability of (stochastic) proximal point methods remains largely unstudied. To address this, we focus on the convergence and stability of the stochastic proximal point algorithm with momentum (SPPAM), and show that SPPAM allows a faster linear convergence to a neighborhood compared to stochastic proximal point algorithm (SPPA) with a better contraction factor, under proper hyperparameter tuning. In terms of stability, we show that SPPAM depends on problem constants more favorably than SGDM, allowing a wider range of step size and momentum that lead to convergence.

Keywords: empirical risk minimization, stochastic proximal point algorithm, momentum, stability.

1. INTRODUCTION

Background. We focus on unconstrained empirical risk minimization instances (Robbins and Monro, 1951; Polyak and Juditsky, 1992; Bottou, 2012; Bottou and Bousquet, 2011; Shalev-Shwartz et al., 2011; Nemirovski et al., 2009; Moulines and Bach, 2011; Bach and Moulines, 2013), as in:

$$\min_{x\in\mathbb{R}^p} \; f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x). \tag{1}$$

To solve (1), stochastic gradient descent (SGD) is the de facto method used by the machine learning community, mainly due to its computational efficiency (Zhang, 2004; Bottou, 2012; Bottou et al., 2018). For completeness, SGD iterates as follows:

$$x_{t+1} = x_t - \eta \nabla f_{i_t}(x_t), \tag{2}$$

where η is the step size, and $\nabla f_i(\cdot)$ is the gradient computed at the i-th data point.

Properties of SGD and Its Momentum Extension. While computationally efficient, stochastic methods often suffer from two major limitations: (i) slow convergence, and (ii) numerical instability. Due to gradient noise, SGD could take longer to converge, in terms of iterations (Moulines and Bach, 2011; Gower et al., 2019). Moreover, SGD suffers from numerical instabilities both in theory (Nemirovski et al., 2009) and practice (Bottou, 2012), allowing a small range of η values that lead to convergence (Moulines and Bach, 2011), but often depend on unknown quantities.

With respect to slow convergence, many variants of accelerated methods have been proposed, along with analyses (Su et al., 2014; Defazio, 2019; Laborde and Oberman, 2019; Allen-Zhu and Orecchia, 2017; Lessard et al., 2016; Hu and Lessard, 2017; Wibisono et al., 2016; Bubeck et al., 2015). Most notable cases include the Polyak's momentum method (Polyak, 1964, 1987) and Nesterov's acceleration (Nesterov, 2018; Ahn, 2020; Nesterov, 1983). These methods allow faster (sometimes optimal) convergence rates, while having virtually the same computational cost as SGD. In particular, SGD with momentum (SGDM) (Polyak, 1964, 1987) iterates as follows:

$$x_{t+1} = x_t - \eta \nabla f_{i_t}(x_t) + \beta(x_t - x_{t-1}), \tag{3}$$

where $\beta \in [0, 1)$ is the momentum parameter. The intuition is that, if the direction from x_{t-1} to x_t was "correct," SGDM utilizes this inertia weighted by the momentum parameter β, instead of just relying on the current point x_t. Much of the state-of-the-art performance has been achieved with SGDM (Huang et al., 2017; Howard et al., 2017; He et al., 2016).

Yet, SGDM could be hard to tune: SGDM adds another hyperparameter—momentum β—to an already sensitive stochastic procedure of SGD. As such, various works have found that such motions could aggravate the instability of SGD. For instance, Liu and Belkin (2019) and Kidambi et al. (2018) show that accelerated SGD does not in general provide any acceleration over SGD, regardless of careful tuning; further, accelerated SGD may diverge for step sizes that SGD converges. Assran and Rabbat (2020) also shows that, even with finite-sum of quadratic functions, accelerated SGD may diverge under usual choices of step size and momentum. See also Loizou and Richtárik (2020); Devolder et al. (2014); d'Aspremont (2008) for more discussions on this topic.

Stability via Proximal Updates. With respect to numerical stability, variants of SGD that utilize proximal updates have recently been proposed (Ryu and Boyd, 2017; Toulis et al., 2014; Toulis and Airoldi, 2017; Toulis et al., 2021; Asi and Duchi, 2019; Asi et al., 2020). In particular, Toulis et al. (2021) introduced stochastic errors in proximal point algorithms (SPPA) and analyzed its convergence and stability, which iterates similar to:

$$x_{t+1} = x_t - \eta \left(\nabla f(x_{t+1}) + \varepsilon_{t+1} \right). \tag{4}$$

Without stochastic errors, (4) is known as the proximal point algorithm (PPA) (Rockafellar, 1976; Güler, 1991) or the implicit gradient descent (IGD). PPA/IGD is known to converge with minimal assumptions on hyperparameter tuning, by improving the conditioning of the optimization problem; more details in Section 2. In the stochastic setting, Toulis et al. (2021) show that SPPA enjoys an exponential discount of the initial condition, regardless of the step size η and the smoothness parameter L. On the contrary, for SGD, both η and L show up within an exponential term, amplifying the initial conditions, leading to even divergence if misspecified (Moulines and Bach, 2011).

Our Focus and Contributions. Stochastic accelerated variants of PPA have received limited attention: how momentum interacts with the stability that PPA provides remains unstudied. To the best of our knowledge, *no momentum has been considered for stochastic proximal point updates that, beyond convergence, also studies the stability of the acceleration motions.* This is the aim of this work. Our contributions are summarized as:

- We introduce stochastic PPA with momentum (SPPAM), and study its convergence and stability behavior. SPPAM directly incorporates the momentum term akin to (3) into (4):

$$x_{t+1} = x_t - \eta \left(\nabla f(x_{t+1}) + \varepsilon_{t+1}\right) + \beta(x_t - x_{t-1}). \tag{5}$$

  We study whether adding momentum β results in faster convergence akin to SGDM, while preserving the numerical stability, inherited by utilizing proximal updates akin to SPPA.
- We show that SPPAM enjoys linear convergence to a neighborhood (Theorem 9) with a better contraction factor than SPPA (Lemma 6). We further characterize the conditions on η and β that result in acceleration (Corollay 7). Finally, we characterize the condition that leads to the exponential discount of initial conditions for SPPAM (Theorem 10), which is significantly easier to satisfy compared to SGDM.
- Empirically, we confirm our theory with experiments on generalized linear models (GLM), including linear and Poisson regressions with different condition numbers. As expected, SGD and SGDM converge only for specific choices of η and β, while SPPA converges for a much wider range of η. SPPAM enjoys the advantages of both acceleration from the momentum and stability from the proximal step: it converges for the range of η that SPPA converges but with faster rate, which improves or matches that of SGDM, when the latter converges.

2. PRELIMINARIES

Proximal Point Algorithm (PPA). The proximal point algorithm (PPA) (Rockafellar, 1976; Güler, 1991) obtains the next iterate for minimizing $f(\cdot)$ by solving the following optimization problem:

$$x_{t+1} = \arg\min_{x \in \mathbb{R}^p} \left\{ f(x) + \tfrac{1}{2\eta}\|x - x_t\|_2^2 \right\}, \tag{6}$$

which is equivalent to implicit gradient descent (IGD) by the first-order optimality condition:

$$x_{t+1} = x_t - \eta \nabla f(x_{t+1}). \tag{7}$$

In words, instead of minimizing $f(\cdot)$ directly, PPA minimizes $f(\cdot)$ with an additional quadratic term. This small change brings a major advantage that PPA enjoys: if $f(\cdot)$ is convex, the added quadratic term can make the problem strongly convex; if $f(\cdot)$ is non-convex, PPA can make it convex (Ahn, 2020). Due to this better conditioning of the problem, PPA exhibits different behavior compared to GD in the deterministic setting. Güler (1991) proved that for a convex function $f(\cdot)$, PPA satisfies:

$$f(x_T) - f(x^\star) \leq O\left(\frac{1}{\sum_{t=1}^T \eta_t}\right), \tag{8}$$

after T iterations. By setting the step size η_t to be large, PPA can converge "arbitrarily" fast.

PPA was soon considered in the stochastic setting. In Ryu and Boyd (2017), a stochastic version of PPA, dubbed as stochastic proximal iterations (SPI), was analyzed, where an approximation of $f(\cdot)$ using a single data $f_i(\cdot)$ was considered. The same algorithm was (statistically) analyzed under the name of implicit stochastic gradient descent (ISGD) (Toulis et al., 2014; Toulis and Airoldi, 2017), and was extended to the Robbins-Monro procedure in Toulis et al. (2021). Similar algorithms were analyzed recently in Asi and Duchi (2019); Asi et al. (2020) where each $f_i(\cdot)$ was further approximated by simpler surrogate functions. These works generally indicate that, in the asymptotic regime, SGD and SPI/ISGD have the same convergence behavior, but in the non-asymptotic regime, SPI/ISGD outperforms SGD due to numerical stability provided by utilizing proximal updates.

Table 1: Comparison of different algorithms in Section 2. $(\cdot)^{\alpha}$ is Rockafellar (1976); Güler (1991); $(\cdot)^{\beta}$ is Güler (1992); Lin et al. (2015, 2018); $(\cdot)^{\gamma}$ is Polyak (1964, 1987); $(\cdot)^{\delta}$ is Toulis et al. (2014); Toulis and Airoldi (2017); Ryu and Boyd (2017); $(\cdot)^{\epsilon}$ is Asi and Duchi (2019); Asi et al. (2020); $(\cdot)^{\zeta}$ is Kulunchakov and Mairal (2019); $(\cdot)^{\eta}$ is Chadha et al. (2021). We highlight with color the algorithms that include momentum motions.

Method	Deterministic
PPA/IGD$^{\alpha}$	$x_{t+1} = \arg\min_x \left\{ f(x) + \frac{1}{2\eta_t}\|x - x_t\|_2^2 \right\}$ $\Leftrightarrow x_{t+1} = x_t - \eta_t \nabla f(x_{t+1})$
Acc. PPA/Catalyst$^{\beta}$	$x_{t+1} \approx \arg\min_x \left\{ f(x) + \frac{\kappa}{2}\|x - y_t\|_2^2 \right\}$ $y_t = x_t + \beta_t(x_t - x_{t-1})$ where $\alpha_t^2 = (1-\alpha_t)\alpha_{t-1}^2 + \frac{\mu}{\mu+\kappa}\alpha_t, \quad \beta_t = \frac{\alpha_{t-1}(1-\alpha_{t-1})}{\alpha_{t-1}^2+\alpha_t}$
	Stochastic
SGDM$^{\gamma}$	$x_{t+1} = x_t - \eta \nabla f_{i_t}(x_t) + \beta(x_t - x_{t-1})$
SPI/ISGD$^{\delta}$	$x_{t+1} = \arg\min_x \left\{ f_{i_t}(x) + \frac{1}{2\eta_t}\|x - x_t\|_2^2 \right\}$ $\Leftrightarrow x_{t+1} = x_t - \eta_t \nabla f_{i_t}(x_{t+1})$
APROX$^{\epsilon}$	Set $f_{i_t}(x) := \max\{f_{i_t}(x_t) + \langle \nabla f_{i_t}(x_t), x - x_t\rangle, \inf_z f_{i_t}(z)\}$ from SPI
Stochastic Catalyst$^{\zeta}$	Set $f(x) := f(y_t) + \langle g_t, x - y_t\rangle + \frac{\kappa+\mu}{2}\|x - y_t\|_2^2$ from Catalyst
Acc. APROX$^{\eta}$	$y_t = (1-\beta_t)x_t + \beta_t z_t$ $z_t = \arg\min_x \left\{ f_{i_t}(x) + \frac{1}{\eta_t}\|x - z_t\|_2^2 \right\}$ $x_{t+1} = (1-\beta_t)x_t + \beta_t z_{t+1}$ where $f_{i_t}(x) := \max\{f_{i_t}(x) + \langle \nabla f_{i_t}(x), y - x\rangle, \inf_z f_{i_t}(z)\}$
SPPAM (this work)	$x_{t+1} = x_t - \eta\left(\nabla f(x_{t+1}) + \varepsilon_{t+1}\right) + \beta(x_t - x_{t-1})$

Accelerated PPA. Under a deterministic setting, accelerated PPA was first proposed in Güler (1992), where Nesterov's acceleration was applied *after* solving the proximal step in (6). This yields the convergence rate of the form:

$$f(x_T) - f(x^\star) \leq O\left(\frac{1}{\left(\sum_{t=1}^T \sqrt{\eta_t}\right)^2}\right), \tag{9}$$

which is faster than the rate in (8). This bound is based on Nesterov's momentum schedules, but does not study the effect in stability different tuning pairs (η, β) might have. Moreover, as can be seen in (8), we can already achieve arbitrarily fast convergence, given PPA is implemented exactly.

Following works focus on studying the conditions under which the proximal step in (6) can be computed inexactly, while still exhibiting some acceleration (Lin et al., 2015, 2018). This was later extended to the stochastic setting in Kulunchakov and Mairal (2019). Chadha et al. (2021) also considered accelerated stochastic PPA. Both of these works apply a convoluted 2- or 3-step Nesterov's procedure after the proximal step, where $f_i(\cdot)$ was further approximated with auxiliary functions. Yet, stability arguments via proximal updates are less apparent due to the auxiliary functions, requiring specific step size and momentum schedules, which might involve an additional one-dimensional optimization per iteration; see also Theorem 10. A summary of the above is provided in Table 1.

Intuition of SPPAM in (5). In contrast to the aforementioned works, we include Polyak's momentum (Polyak, 1964) directly to SPPA, yielding (5). Apart from the similarity between SPPAM in (5) and SGDM in (3), SPPAM shares the same geometric intuition as Polyak's momentum for SGDM. Disregarding the stochastic error, the update in (5) follows from the solution of: [1]

$$\arg\min_{x\in\mathbb{R}^p}\left\{f(x)+\tfrac{1}{2\eta}\|x-x_t\|_2^2-\tfrac{\beta}{\eta}\langle x_t-x_{t-1},x\rangle\right\}.$$

We can get a sense of the behavior of SPPAM from this expression. First, for large η, the algorithm is minimizing the original $f(x)$. For small η, the algorithm not only tries to stay local by minimizing the quadratic term, but also tries to minimize $-\frac{\beta}{\eta}\langle x_t - x_{t-1}, x\rangle$. By the definition of inner product, this means that x, on top of minimizing $f(x)$ and staying close to x_t, also tries to move along the direction from x_{t-1} to x_t. This intuition aligns with that of Polyak's momentum.

3. THE QUADRATIC MODEL CASE

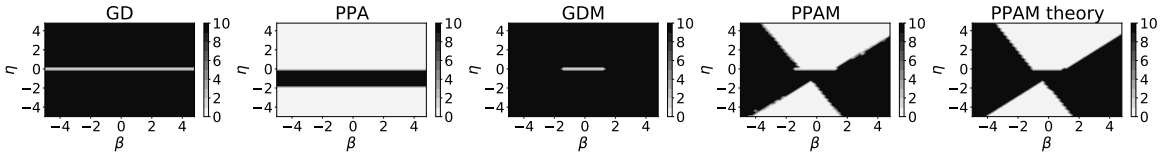

Figure 1: We generate $A \in \mathbb{R}^{p\times p}$ and $b, x^\star \in \mathbb{R}^p$ from $\mathcal{N}(0, I)$, where $p = 100$ and the condition number of A is 10. We sweep η and β from -5 to 5, with 0.2 interval. We plot the accuracy $\|x_t - x^\star\|_2^2$ after 100 iterations, with the maximum replaced by 10.

For simplicity, we first consider the convex quadratic optimization problem under the deterministic setting. Specifically, we consider the objective function:

$$f(x) = \frac{1}{2}x^\top Ax - b^\top x, \tag{10}$$

where $A \in \mathbb{R}^{p\times p}$ is positive semi-definite with eigenvalues $[\lambda_1, \ldots, \lambda_p]$. Under this scenario, we can study how the step size η and momentum β affect each other, by deriving exact conditions that lead to convergence for each algorithm. The comparison lists includes gradient descent (GD), gradient descent with momentum (GDM), the PPA, and PPA with momentum (PPAM). Propositions 1 and 3 for GD and GDM are from Goh (2017), and included for completeness. Proofs for PPA and PPAM in Propositions 2 and 4 can be found in the extended version of this work (Kim et al., 2021).

Proposition 1 (GD (Goh, 2017)) *To minimize (10) with gradient descent, the step size η needs to satisfy $0 < \eta < \frac{2}{\lambda_i}$, $\forall i$, where λ_i is the i-th eigenvalue of A.*

Proposition 2 (PPA) *To minimize (10) with PPA, the step size η needs to satisfy $\left|\frac{1}{1+\eta\lambda_i}\right| < 1$, $\forall i$.*

Proposition 3 (GDM (Goh, 2017)) *To minimize (10) with gradient descent with momentum, the step size η needs to satisfy $0 < \eta\lambda_i < 2 + 2\beta$, $\forall i$ and $0 \le \beta \le 1$.*

1. While preparing this manuscript, we became aware of a recent parallel work (Deng and Gao, 2021) that considers similar extension of proximal operators. However, their proposed algorithm with momentum involves two-step procedures, further with approximated auxiliary functions similarly to Chadha et al. (2021).

Proposition 4 (PPAM) *Let* $\delta_i = \left(\frac{\beta+1}{1+\eta\lambda_i}\right)^2 - \frac{4\beta}{1+\eta\lambda_i}$. *To minimize* (10) *with PPAM, the step size* η *and momentum* β *need to satisfy* $\forall i$*:*

- $\eta > \frac{\beta-1}{\lambda_i}$, *if* $\delta_i \leq 0$*;*
- $\frac{\beta+1}{1+\eta\lambda_i} + \sqrt{\delta_i} < 2$, *if* $\delta_i > 0$ *and* $\frac{\beta+1}{1+\eta\lambda_i} \geq 0$*;*
- $\frac{\beta+1}{1+\eta\lambda_i} - \sqrt{\delta_i} > -2$, *otherwise.*

Given the above propositions, we can study the stability with respect to the step size η and the momentum β for the considered algorithms. Numerical simulations support the above propositions and are illustrated in Figure 1, matching the theoretical conditions exhibited above. In particular, for GD (1st), only a small range of step sizes η leads to convergence (small white band); this "white band" corresponds to the restriction that η has to satisfy $\eta < \frac{2}{\lambda_i}$ for all i. On the other hand, PPA/IGD (2nd) converges in much wider choices of η; this is apparent from Proposition 2, since $\left|\frac{1}{1+\eta\lambda_i}\right|$ can be arbitrarily small for larger values of η. GDM (3rd) requires both η and β to be in a small region to converge, following Proposition 3. Finally, PPAM (4th) converges in much wider choices of η and β; e.g., the conditions in Proposition 4 define different regions of the pair (η, β) that lead to convergence, some of which set both η and β being negative. Note that the empirical convergent region for PPAM almost exactly matches the theoretical region that leads to convergence in Proposition 4 (5th). In the remainder of the paper, we study how such pattern translates to a general strongly convex function $f(\cdot)$, with stochasticity.

4. THEORY

In this section, we theoretically characterize the convergence and stability behavior of SPPAM.[2] We follow the stochastic errors of PPA, as set up in Toulis et al. (2021); we can thus express (5) as[3]:

$$x_{t+1}^+ = x_t - \eta\nabla f(x_{t+1}^+) + \beta(x_t - x_{t-1})$$
$$x_{t+1} = x_{t+1}^+ - \eta\varepsilon_{t+1}.$$

We further assume the following:

Assumption 1 $f(\cdot)$ *is a* μ*-strongly convex function: for some fixed* $\mu > 0$ *and for all* x *and* y*,*

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x - y\|_2^2.$$

Assumption 2 *There exists fixed* $\sigma^2 > 0$ *such that, given history* $\mathcal{F}_{t-1}$*,*

$$\mathbb{E}\left[\varepsilon_t \mid \mathcal{F}_{t-1}\right] = 0 \text{ and } \mathbb{E}\left[\|\varepsilon_t \mid \mathcal{F}_{t-1}\|^2\right] \leq \sigma^2 \text{ for all } t.$$

We now study whether SPPAM enjoys faster convergence than SPPA in (4). We start with the iteration invariant bound:

2. All proofs are provided in the extended version of this work (Kim et al., 2021).
3. This is equivalent to (5) by substituting x_{t+1}^+ in the last expression; x_{t+1}^+ is an auxiliary intermediate variable that is used for the analysis only.

Theorem 5 *For μ-strongly convex $f(\cdot)$, SPPAM satisfies the following iteration invariant bound:*

$$\mathbb{E}\left[\|x_{t+1}-x^\star\|_2^2\right] \leq \tfrac{4}{(1+\eta\mu)^2}\mathbb{E}\left[\|x_t-x^\star\|_2^2\right] + \tfrac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}\mathbb{E}\left[\|x_{t-1}-x^\star\|_2^2\right] + \eta^2\sigma^2. \quad (11)$$

Notice that all terms –except the last one– are divided by $(1+\eta\mu)^2$. Thus, large step sizes η help convergence (to a neighborhood), reminiscent of the convergence behavior of PPA in (8). Based on (11), we can write the following 2×2 system that characterizes the progress of SPPAM:

$$\begin{bmatrix}\mathbb{E}\left[\|x_{t+1}-x^\star\|_2^2\right]\\ \mathbb{E}\left[\|x_t-x^\star\|_2^2\right]\end{bmatrix} \leq \underbrace{\begin{bmatrix}\frac{4}{(1+\eta\mu)^2} & \frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}\\ 1 & 0\end{bmatrix}}_{A} \cdot \begin{bmatrix}\mathbb{E}\left[\|x_t-x^\star\|_2^2\right]\\ \mathbb{E}\left[\|x_{t-1}-x^\star\|_2^2\right]\end{bmatrix} + \begin{bmatrix}\eta^2\sigma^2\\ 0\end{bmatrix}. \quad (12)$$

It is clear that the spectrum of the contraction matrix A determines the convergence rate to a neighborhood, as in (Goh, 2017). This is summarized in the following lemma:

Lemma 6 *The maximum eigenvalue of A, which determines the convergence rate of SPPAM, is:*

$$\tfrac{2}{(1+\eta\mu)^2} + \sqrt{\tfrac{4}{(1+\eta\mu)^4} + \tfrac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}}. \quad (13)$$

Notice the one-step contraction factor in (13) is of order $O(1/\eta^2)$, exhibiting acceleration compared to that of SPPA for strongly convex objectives (Toulis et al., 2021): $1/(1+2\eta\mu) \approx O(1/\eta)$. However, due to the additional terms, it is not immediately obvious when SPPAM enjoys faster convergence than SPPA. We thus characterize this condition more precisely in the following corollary:

Corollary 7 *For μ-strongly convex $f(\cdot)$, SPPAM enjoys better contraction factor than SPPA if:*

$$\frac{4\beta^2}{4-(1+\beta)^2} < \frac{\eta^2\mu^2-6\eta\mu-3}{(1+\eta\mu)^2}.$$

In words, for a fixed step size η and given a strongly convex parameter μ, there is a range of momentum parameters β that exhibits acceleration compared to SPPA.

Remark 8 *In contrast to (stochastic) gradient method analyses in convex optimization, where acceleration is usually shown by improving the dependency on the condition number from $\kappa = \frac{L}{\mu}$ to $\sqrt{\kappa}$, such a claim can hardly be made for stochastic proximal point methods. This is also the case in deterministic setting; see (8) and (9). As shown in Theorem 5, our convergence analysis of SPPAM does not depend on L-smoothness at all. This robustness of SPPAM is also confirmed in numerical simulations in Section 5, where SPPAM exhibits the fastest convergence rate, virtually independent of the different settings considered.*

We now formalize the convergence behavior of SPPAM. In particular, we characterize the condition that leads to the exponential discount of initial conditions. By unrolling the recursion of SPPAM in (12) for T iterations, we obtain:

$$\begin{bmatrix}\mathbb{E}\left[\|x_T-x^\star\|_2^2\right]\\ \mathbb{E}\left[\|x_{T-1}-x^\star\|_2^2\right]\end{bmatrix} \leq A^T \cdot \begin{bmatrix}\|x_0-x^\star\|_2^2\\ \|x_{-1}-x^\star\|_2^2\end{bmatrix} + \left(\sum_{i=1}^{T-1} A^i\right)\begin{bmatrix}1\\0\end{bmatrix}\eta^2\sigma^2.$$

It is clear from the above that the convergence is determined by A^T and $\left(\sum_{i=1}^{T-1} A^i\right)$, where A was defined in (12). Our next theorem derives convergence to a neighborhood based on the spectrum of these quantities, akin to Assran and Rabbat (2020, Theorem 1) and Toulis et al. (2021, Theorem 3).

Theorem 9 *For μ-strongly convex $f(\cdot)$, assume SPPAM is initialized with $x_0 = x_{-1}$. Then, after T iterations, we have:*

$$\mathbb{E}\left[\|x_T - x^\star\|_2^2\right] \leq \frac{2\sigma_1^T}{\sigma_1 - \sigma_2}\left(\left(\|x_0 - x^\star\|_2^2 + \frac{\eta^2\sigma^2}{1-\theta}\right)\cdot(1+\theta)\right) + \frac{\eta^2\sigma^2}{1-\theta}, \tag{14}$$

where $\theta = \frac{4}{(1+\eta\mu)^2} + \frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}$. Here, $\sigma_{1,2}$ are the eigenvalues of A, and

$$\frac{2\sigma_1^T}{\sigma_1-\sigma_2} = \tau^{-1}\cdot\left(\frac{2}{(1+\eta\mu)^2} + \tau\right)^T \quad \textit{with} \quad \tau = \sqrt{\frac{4}{(1+\eta\mu)^4} + \frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}}. \tag{15}$$

The above theorem states that the term in (15) determines the discounting rate of the initial conditions. In particular, the condition that leads to an exponential discount of the initial conditions is characterized by the following theorem:

Theorem 10 *Let the following condition hold:*

$$\tau = \sqrt{\frac{4}{(1+\eta\mu)^4} + \frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}} < \frac{1}{2}. \tag{16}$$

Then, for μ-strongly convex $f(\cdot)$, initial conditions of SPPAM exponentially discount: i.e., in (14),

$$\frac{2\sigma_1^T}{\sigma_1-\sigma_2} = \tau^{-1}\cdot\left(\frac{2}{(1+\eta\mu)^2} + \tau\right)^T = C^T, \quad \textit{where} \quad C \in (0,1).$$

Remark 11 *The condition in (16) is much easier to satisfy than SGDM. E.g., as described below, the required condition for SGDM to converge linearly to a neighborhood in strongly convex quadratic objective relies on knowing $\eta = \frac{1}{L}$ and momentum $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ (Assran and Rabbat, 2020), where both L and κ are unknown in practice. While this is also true for SPPAM (i.e., μ is an unknown quantity), (16) suggests that one can essentially set η sufficiently large to ensure the exponential discount, even without knowing μ exactly.*

Remark 12 *Other works that study variants of accelerated stochastic PPA (Kulunchakov and Mairal, 2019; Chadha et al., 2021) still require specific choices of step size and momentum (e.g., $\eta_t = \frac{1}{L+c_0\sqrt{t+1}}$, $\beta_t = \frac{2}{t+2}$ for the latter; see Table 1 for the former), similarly to SGDM.*

To provide more context of the condition in Theorem 10, we make an "unfair" comparison of (16), which holds for general strongly convex $f(\cdot)$, to the condition that SGDM requires for strongly convex *quadratic objective* in (10). Assran and Rabbat (2020, Theorem 1) show that SGDM converges to a neighborhood at a linear rate for strongly convex quadratic objective if $\max\{\rho_\mu(\eta,\beta),\ \rho_L(\eta,\beta)\} < 1$, where $\rho_\lambda(\eta,\beta)$ for $\lambda \in \{\mu, L\}$ is defined as:

$$\rho_\lambda(\eta,\beta) = \begin{cases} \frac{|(1+\beta)(1-\eta\lambda)|}{2} + \frac{\sqrt{\Delta_\lambda}}{2} & \text{if } \Delta_\lambda \geq 0, \\ \sqrt{\beta(1-\eta\lambda)} & \text{otherwise,} \end{cases} \tag{17}$$

with $\Delta_\lambda = (1+\beta)^2(1-\eta\lambda)^2 - 4\beta(1-\eta\lambda)$. This condition for convergence can thus be divided into three cases, depending on the range of $\eta\lambda$. Define $\psi_{\beta,\eta,\lambda} = (1+\beta)(1-\eta\lambda)$. Then:

$$\begin{cases} \eta\lambda \geq 1, & \text{Converges if } -\psi_{\beta,\eta,\lambda} + \sqrt{\Delta_\lambda} < 2, \\ \frac{(1-\beta)^2}{(1+\beta)^2} \leq \eta\lambda < 1, & \text{Always converges,} \\ \eta\lambda < \frac{(1-\beta)^2}{(1+\beta)^2}, & \text{Converges if } \psi_{\beta,\eta,\lambda} + \sqrt{\Delta_\lambda} < 2. \end{cases}$$

Now, consider the standard momentum value $\beta = 0.9$. For the first case, the convergence requirement translates to $1 \leq \eta\lambda \leq \frac{24}{19}$. The second range is given by $\frac{1}{361} \leq \eta\lambda < 1$. The third condition is lower bounded by 2 for $\beta = 0.9$, leading to divergence. Combining, SGDM requires $0.0028 \approx \frac{1}{361} \leq \eta\lambda \leq \frac{24}{19} \approx 1.26$ to converge for strongly convex quadratic objectives, set aside that this bound has to satisfy for (unknown) μ or L.

Albeit an unfair comparison, for general strongly convex objective, (16) becomes $\eta\mu > 4.81$ for $\beta = 0.9$. Even though μ is unknown, one can see this condition is easy to satisfy, by using a sufficiently large step size η.

5. EXPERIMENTS

In this section, we perform numerical experiments to study the convergence behaviors of SPPAM, SPPA, SGDM, and SGD, using generalized linear models (GLM) Nelder and Wedderburn (1972).

Let $b_i \in \mathbb{R}$ be the label, $a_i \in \mathbb{R}^p$ be the features, and $x^\star \in \mathbb{R}^p$ be the model parameter of interest. GLM assumes that b_i follows an exponential family distribution: $b_i \mid a_i \sim \exp\left(\frac{\gamma b_i - c_1(\gamma)}{\omega} c_2(b_i, \omega)\right)$. Here, $\gamma = \langle a_i, x^\star \rangle$ is the linear predictor, ω is the dispersion parameter related to the variance of b_i, and $c_1(\cdot)$ and $c_2(\cdot)$ are known real-valued functions. GLM subsumes a wide family of models including linear, logistic, and Poisson regressions. Different models connects the linear predictor $\gamma = \langle a_i, x^\star \rangle$ through different *mean* functions $h(\cdot)$. We focus on linear and Poisson regression models, where mean functions are defined respectively as $h(\gamma) = \gamma$ and $h(\gamma) = e^\gamma$. The former is an "easy" case, where objective is strongly convex, satisfying Assumption 1. The latter is a "hard" case with non-Lipschitz continuous gradients, where SGD and SGDM are expected to suffer.

Toulis et al. (2014) introduced an efficient, exact implementation of SPPA for GLM. We adapt this procedure to SPPAM; see Algorithm 1. Its derivation can be found in Kim et al. (2021).

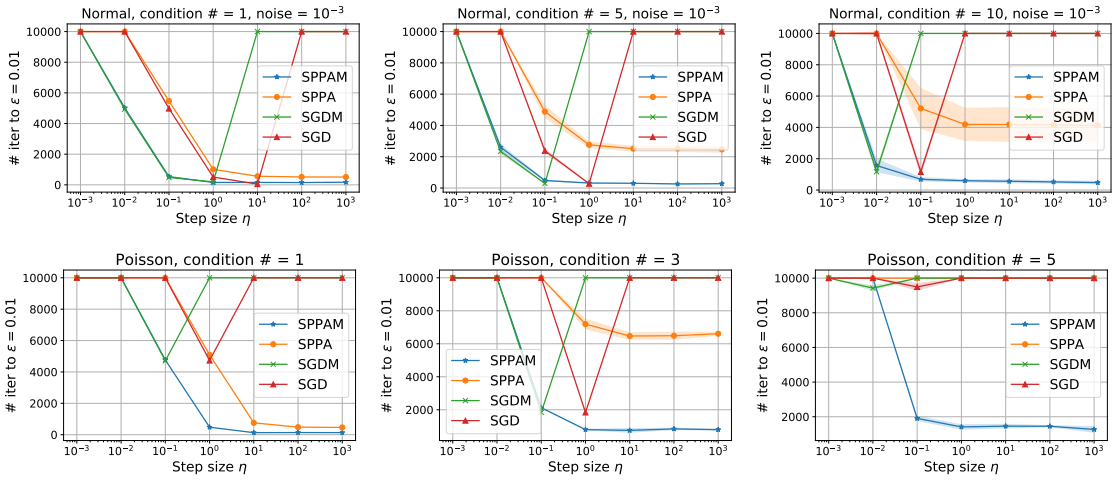

Figure 2: Top: Linear regression with condition number $\kappa \in \{1, 5, 10\}$ with gaussian noise level `1e-3`. Bottom: Poisson regression with condition number $\kappa \in \{1, 3, 5\}$. We set $p = n = 100$ in both cases. Batch size is 10 for all algorithms. The median number of iterations to reach $\varepsilon = 0.01$ is plotted. Shaded area are the standard deviations across 5 experiments.

Algorithm 1 SPPAM for GLM

for $t = 1, 2, \dots$ **do**
  Sample $i_t \sim \text{Unif}(1, n)$
  $r_t \leftarrow \eta(b_{i_t} - h(\langle a_{i_t}, x_{t-1}\rangle)$
  $B_t \leftarrow [0, r_t]$
  if $r_t \leq 0$ **then**
    $B_t \leftarrow [r_t, 0]$
  end
  $\xi_t = \eta\big[b_{i_t} - h((1+\beta)\langle a_{i_t}, x_{t-1}\rangle - \beta\langle a_{i_t}, x_{t-2}\rangle + \xi_t \cdot \|a_{i_t}\|_2^2)\big],\ \xi_t \in B_t$
  $x_t \leftarrow x_{t-1} + \xi_t \cdot a_{i_t} + \beta(x_{t-1} - x_{t-2})$
end

We generate the data as follows. $A \in \mathbb{R}^{p\times n}$ and $x^\star \in \mathbb{R}^p$ are drawn from $\mathcal{N}(0, I)$. For the normal case, we generate $b_i = \langle a_i, x^\star\rangle$, and for the Poisson case, we generate $b_i \sim \text{Poisson}(e^{\langle a_i, x^\star\rangle})$ for $i = 1, \dots, n$. For each experimental setup, we run SPPAM (blue), SPPA (orange), SGDM (green), and SGD (red) for 10^4 iterations. We repeat each experiment for 5 independent trials, and plot the median number of iterations to reach precision $\varepsilon \leq 10^{-2}$, along with the standard deviation. We measure the precision $\varepsilon = \frac{\|b-\hat{b}\|_2^2}{\|b\|_2^2}$, where b is the true label and $\hat{b}$ is the predicted label.

In Figure 2 (Top), we present the results for the linear regression with different condition numbers, with gaussian noise level `1e-3`. We run each algorithm constant step size η varying from 10^{-3} to 10^3 with $10\times$ increment, and with $\beta = 0.9$. As expected, SGD and SGDM only converge for specific step size η, while SPPA and SPPAM converge for much wider ranges. In terms of convergence rate, SPPAM converges faster than SPPA in all scenarios, which improves or matches the rate of SGDM, when it converges. As κ increases, the range of η that leads to convergence for SGD and SGDM shrinks; notice the sharper " $\vee$ " shape for SGD and SGDM for $\kappa = 10$ (3rd), compared to $\kappa = 5$ (2nd) or $\kappa = 1$ (1st). SPPA also slightly slows down as κ increases, while SPPAM converges essentially in the same manner for all scenarios.

Such trend is much more pronounced for the Poisson regression case presented in Figure 2 (Bottom). Due to the exponential mean function $h(\cdot)$ for Poisson model, the outcomes are extremely sensitive, and its likelihood does not satisfy standard assumptions like L-smoothness. As such, SGD and SGDM struggles with slow convergence even when $\kappa = 1$ (1st), while also exhibiting instability—each method converges only for a single choice of η considered. Similar trend is shown when $\kappa = 3$ (2nd) where SPPA starts slowing down. For $\kappa = 5$ (3rd), all methods except for SPPAM did not make much progress in 10^4 iterations, for the entire range of η and β considered. Quite remarkably, SPPAM still converges in the same manner without sacrificing both the convergence rate and the range of hyperparameters that lead to convergence.

6. CONCLUSION

We propose the stochastic proximal point algorithm with momentum (SPPAM), which directly incorporates Polyak's momentum inside the proximal step. We show that SPPAM converges to a neighborhood at a faster rate than stochastic proximal point algorithm (SPPA), and characterize the conditions that result in acceleration. Further, we prove linear convergence of SPPAM to a neighborhood, and provide conditions that lead to an exponential discount of the initial conditions, akin to SPPA. We confirm our theory with numerical simulations on linear and Poisson regression models; SPPAM converges for all the step sizes that SPPA converges, with a faster rate that matches or improves SGDM.

References

Kwangjun Ahn. From Proximal Point Method to Nesterov's Acceleration. *arXiv:2005.08304 [cs, math]*, June 2020. URL http://arxiv.org/abs/2005.08304. arXiv: 2005.08304.

Z. Allen-Zhu and L. Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

Hilal Asi and John C. Duchi. Stochastic (Approximate) Proximal Point Methods: Convergence, Optimality, and Adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, January 2019. ISSN 1052-6234, 1095-7189. doi: 10.1137/18M1230323. URL http://arxiv.org/abs/1810.05633. arXiv: 1810.05633.

Hilal Asi, Karan Chadha, and Gary Cheng. Minibatch Stochastic Approximate Proximal Point Methods. *34th Conference on Neural Information Processing Systems*, page 11, 2020.

Mahmoud Assran and Michael Rabbat. On the Convergence of Nesterov's Accelerated Gradient Method in Stochastic Settings. *Proceedings of the 37 th International Conference on Machine Learning*, page 11, 2020. URL https://proceedings.icml.cc/static/paper_files/icml/2020/5529-Paper.pdf.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 1*, pages 773–781, 2013.

Léon Bottou and Olivier Bousquet. 13 the tradeoffs of large-scale learning. *Optimization for machine learning*, page 351, 2011.

Léon Bottou. Stochastic Gradient Descent Tricks. In *Neural Networks: Tricks of the Trade*, volume 7700, pages 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35288-1 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_25. URL http://link.springer.com/10.1007/978-3-642-35289-8_25. Series Title: Lecture Notes in Computer Science.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2):223–311, January 2018. ISSN 0036-1445, 1095-7200. doi: 10.1137/16M1080173. URL https://epubs.siam.org/doi/10.1137/16M1080173.

S. Bubeck, Y.-T. Lee, and M. Singh. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

Karan Chadha, Gary Cheng, and John C. Duchi. Accelerated, Optimal, and Parallel: Some Results on Model-Based Stochastic Optimization. *arXiv:2101.02696 [cs, math, stat]*, January 2021. URL http://arxiv.org/abs/2101.02696. arXiv: 2101.02696.

Alexandre d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.

A. Defazio. On the curved geometry of accelerated optimization. In *Advances in Neural Information Processing Systems*, pages 1764–1773, 2019.

Qi Deng and Wenzhi Gao. Minibatch and Momentum Model-based Methods for Stochastic Weakly Convex Optimization. *arXiv:2106.03034 [cs, math]*, November 2021. URL http://arxiv.org/abs/2106.03034. arXiv: 2106.03034.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, 2014.

Gabriel Goh. Why momentum really works. *Distill*, 2017. doi: 10.23915/distill.00006. URL http://distill.pub/2017/momentum.

Robert M Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtarik. SGD: General Analysis and Improved Rates. *Proceedings of the 36 th International Conference on Machine Learning*, page 10, 2019.

Osman Güler. On the Convergence of the Proximal Point Algorithm for Convex Minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419, March 1991. ISSN 0363-0129. doi: 10.1137/0329022. URL https://epubs.siam.org/doi/10.1137/0329022. Publisher: Society for Industrial and Applied Mathematics.

Osman Güler. New Proximal Point Algorithms for Convex Minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992. ISSN 1052-6234. doi: 10.1137/0802032. URL https://epubs.siam.org/doi/abs/10.1137/0802032. Publisher: Society for Industrial and Applied Mathematics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

B. Hu and L. Lessard. Dissipativity theory for Nesterov's accelerated method. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1549–1557. JMLR. org, 2017.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham M. Kakade. On the insufficiency of existing momentum schemes for Stochastic Optimization. February 2018. URL https://openreview.net/forum?id=rJTutzbA-.

Junhyung Lyle Kim, Panos Toulis, and Anastasios Kyrillidis. Convergence and stability of the stochastic proximal point algorithm with momentum. *Author website*, 2021. URL https://jlylekim.github.io/files/SPPAM.pdf.

Andrei Kulunchakov and Julien Mairal. A Generic Acceleration Framework for Stochastic Composite Optimization. *Advances in Neural Information Processing Systems*, 32, October 2019. arXiv: 1906.01164.

M. Laborde and A. Oberman. A Lyapunov analysis for accelerated gradient methods: From deterministic to stochastic case. *arXiv preprint arXiv:1908.07861*, 2019.

L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

H. Lin, Julien Mairal, and Zaid Harchaoui. Catalyst Acceleration for First-order Convex Optimization: from Theory to Practice. *Journal of Machine Learning Research*, 18:1–54, 2018.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A Universal Catalyst for First-Order Optimization. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3384–3392. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5928-a-universal-catalyst-for-first-order-optimization.pdf.

Chaoyue Liu and Mikhail Belkin. Accelerating SGD with momentum for over-parameterized learning. September 2019. URL https://openreview.net/forum?id=r1gixp4FPH.

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

Eric Moulines and Francis R. Bach. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 451–459. Curran Associates, Inc., 2011.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, 19(4):1574–1609, January 2009. ISSN 1052-6234, 1095-7189. doi: 10.1137/070704277. URL http://epubs.siam.org/doi/10.1137/070704277.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer Optimization and Its Applications. Springer International Publishing, 2 edition, 2018. ISBN 978-3-319-91577-7. doi: 10.1007/978-3-319-91578-4. URL https://www.springer.com/gp/book/9783319915777.

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, January 1964. ISSN 0041-5553. doi: 10.1016/0041-5553(64)90137-5. URL http://www.sciencedirect.com/science/article/pii/0041555364901375.

Boris T. Polyak. Introduction to optimization. *Inc., Publications Division, New York*, 1, 1987.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. ISSN 0003-4851. URL https://www.jstor.org/stable/2236626. Publisher: Institute of Mathematical Statistics.

R. Tyrrell Rockafellar. Monotone Operators and the Proximal Point Algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, August 1976. ISSN 0363-0129. doi: 10.1137/0314056. URL https://epubs.siam.org/doi/abs/10.1137/0314056. Publisher: Society for Industrial and Applied Mathematics.

Ernest K Ryu and Stephen Boyd. Stochastic Proximal Iteration: A Non-Asymptotic Improvement Upon Stochastic Gradient Descent. *Author website*, page 42, 2017.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical programming*, 127(1):3–30, 2011.

W. Su, S. Boyd, and E. Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

Panos Toulis and Edoardo M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, August 2017. ISSN 0090-5364, 2168-8966. doi: 10.1214/16-AOS1506. URL https://projecteuclid.org/journals/annals-of-statistics/volume-45/issue-4/Asymptotic-and-finite-sample-properties-of-estimators-based-on-stochastic/10.1214/16-AOS1506.full. Publisher: Institute of Mathematical Statistics.

Panos Toulis, Jason Rennie, and Edoardo M Airoldi. Statistical analysis of stochastic gradient methods for generalized linear models. *International Conference on Machine Learning*, pages 667–675, 2014. URL http://proceedings.mlr.press/v32/toulis14.html.

Panos Toulis, Thibaut Horel, and Edoardo M. Airoldi. The proximal Robbins–Monro method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(1):188–212, 2021. ISSN 1467-9868. doi: 10.1111/rssb.12405. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12405.

A. Wibisono, A. Wilson, and M. Jordan. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

K. Williams. The n-th power of a $2\times$ 2 matrix. *Mathematics Magazine*, 65(5):336–336, 1992.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116, 2004.

Appendix A. Proofs for Section 3

In this section, we provide proofs for the propositions in Section 3. Proofs below utilize "classical momentum" form, which iterates:

$$\begin{aligned} y_{t+1} &= \beta y_t + \nabla f(x_t) \\ x_{t+1} &= x_t - \eta y_{t+1}. \end{aligned}$$

This is equivalent to Polyak's momentum in the sense that, plugging in the first equation to the second, we get

$$\begin{aligned} x_{t+1} = x_t - \eta y_{t+1} &= x_t - \eta \nabla f(x_t) - \eta \beta y_t \\ &= x_t - \eta \nabla f(x_t) + \beta (x_t - x_{t-1}), \end{aligned}$$

where the last equality is from the second equation of classical momentum.

Proof of Proposition 2

Proof Recall PPA/IGD recursion. For quadratic problem in (10), the gradient can be computed in closed form:

$$x_{t+1} = x_t - \eta \nabla f(x_{t+1}) = x_t - \eta (A x_{t+1} - b).$$

Consider the eigenvalue decomposition of $A = QDQ^\top$ and the change of basis $z_t = Q^\top (x_t - x^\star)$. Then, $Qz_t = x_t - x^\star$ and $Qz_t + x^\star = x_t$. Also using $AQ = QD$ and $Ax^\star = b$, we can write down the recursion above as:

$$\begin{aligned} Qz_{t+1} + x^\star &= Qz_t + x^\star - \eta \left(A(Qz_{t+1} + x^\star) - b \right) \\ &= Qz_t + x^\star - \eta Q D z_{t+1} \end{aligned}$$

Multiplying $Q^\top$ on both sides,

$$\begin{aligned} &z_{t+1} + Q^\top x^\star = z_t + Q^\top x^\star - \eta D z_{t+1} \Rightarrow \\ &(1 + \eta D) z_{t+1} = z_t. \end{aligned}$$

Writing the above component-wise, we get

$$z_{t+1}^i = \left(\frac{1}{1 + \eta \lambda_i} \right) \cdot z_t^i = \left(\frac{1}{1 + \eta \lambda_i} \right)^{t+1} \cdot z_0^i$$

Going back to the change of basis, $z_t = Q^\top (x_t - x^\star)$, we have the relation $x_t - x^\star = Qz_t$. Therefore, using the component-wise relation for z_{t+1}^i above,

$$x_t - x^\star = Qz_t = \sum_{i=1}^n z_0^i \left(\frac{1}{1 + \eta \lambda_i} \right)^t q^i.$$

Thus, in order for IGD to converge, one needs to satisfy $\left| \frac{1}{1+\eta\lambda_i} \right| < 1$.

■

Proof of Proposition 4

Proof Consider the "classical momentum" form of PPAM:

$$\begin{aligned} y^{k+1} &= \beta y^k + \nabla f(x^{k+1}) \\ x^{k+1} &= x^k - \eta y^{k+1}. \end{aligned} \tag{18}$$

We perform change of basis: $z^k = Q^\top (x^k - x^\star)$ and $\phi^k = Q^\top y^k$.

For the first line of (18), we have:

$$\begin{aligned} y^{k+1} &= \beta y^k + \nabla f(x^{k+1}) \\ Q^\top y^{k+1} &= \beta Q^\top y^k + Q^\top (Ax^{k+1} - b) \\ &= \beta Q^\top y^k + Q^\top (A(Qz^{k+1} + x^\star) - b) \\ &= \beta Q^\top y^k + Q^\top (AQz^{k+1} + Ax^\star - b) \\ &= \beta Q^\top y^k + Q^\top AQz^{k+1} \\ &= \beta Q^\top y^k + DQ^\top Qz^{k+1} \\ &= \beta Q^\top y^k + Dz^{k+1}. \end{aligned}$$

Change of basis and writing component-wise, we get:

$$\begin{aligned} \phi_i^{k+1} &= \beta \phi_i^k + \lambda_i z_i^{k+1} \\ &= \beta \phi_i^k + \lambda_i (z_i^k - \eta \phi_i^{k+1}) \\ &= \beta \phi_i^k + \lambda_i z_i^k - \eta \lambda_i \phi_i^{k+1} \\ (1 + \eta \lambda_i) \phi_i^{k+1} &= \beta \phi_i^k + \lambda_i z_i^k \\ \phi_i^{k+1} &= \frac{\beta}{1 + \eta \lambda_i} \phi_i^k + \frac{\lambda_i}{1 + \eta \lambda_i} z_i^k. \end{aligned}$$

For the second line of (18), we have:

$$\begin{aligned} x^{k+1} &= x^k - \eta y^{k+1} \\ Qz^{k+1} + x^\star &= Qz^k + x^\star - \eta Q \phi^{k+1} \\ Q^\top Qz^{k+1} &= Q^\top Qz^k - \eta Q^\top Q \phi^{k+1}. \end{aligned}$$

Again, change of basis and writing component-wise, we get:

$$z_i^{k+1} = z_i^k - \eta \phi_i^{k+1}.$$

Therefore, (18) can be written as, component-wise,

$$\begin{aligned} \phi_i^{k+1} &= \frac{\beta}{1 + \eta \lambda_i} \phi_i^k + \frac{\lambda_i}{1 + \eta \lambda_i} z_i^k \\ z_i^{k+1} &= z_i^k - \eta \phi_i^{k+1}. \end{aligned}$$

We can write above in matrix form:

$$\begin{bmatrix} 1 & 0 \\ \eta & 1 \end{bmatrix} \cdot \begin{bmatrix} \phi_i^{k+1} \\ z_i^{k+1} \end{bmatrix} = \begin{bmatrix} \frac{\beta}{1+\eta\lambda_i} & \frac{\lambda_i}{1+\eta\lambda_i} \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \phi_i^k \\ z_i^k \end{bmatrix}.$$

Multiplying the inverse of the first matrix, i.e., $\begin{bmatrix} 1 & 0 \\ \eta & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ -\eta & 1 \end{bmatrix}$ on both sides,

$$\begin{aligned} \begin{bmatrix} \phi_i^{k+1} \\ z_i^{k+1} \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ -\eta & 1 \end{bmatrix} \cdot \begin{bmatrix} \frac{\beta}{1+\eta\lambda_i} & \frac{\lambda_i}{1+\eta\lambda_i} \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \phi_i^k \\ z_i^k \end{bmatrix} \\ &= \begin{bmatrix} \frac{\beta}{1+\eta\lambda_i} & \frac{\lambda_i}{1+\eta\lambda_i} \\ \frac{-\eta\beta}{1+\eta\lambda_i} & \frac{1}{1+\eta\lambda_i} \end{bmatrix} \cdot \begin{bmatrix} \phi_i^k \\ z_i^k \end{bmatrix}. \end{aligned}$$

Therefore, we can write the above as

$$\begin{bmatrix} \phi_i^k \\ z_i^k \end{bmatrix} = R^k \cdot \begin{bmatrix} \phi_i^0 \\ z_i^0 \end{bmatrix}, \quad R = \begin{bmatrix} \frac{\beta}{1+\eta\lambda_i} & \frac{\lambda_i}{1+\eta\lambda_i} \\ \frac{-\eta\beta}{1+\eta\lambda_i} & \frac{1}{1+\eta\lambda_i} \end{bmatrix}.$$

To compute R^k, we use the method presented in Williams (1992). Then, denoting σ_1 and σ_2 as the eigenvalues of R, we have:

$$R^k = \begin{cases} \sigma_1^k R_1 - \sigma_2^k R_2 & \sigma_1 \neq \sigma_2 \\ \sigma_1^k \left(k\frac{R}{\sigma_1} - (k-1)I \right) & \sigma_1 = \sigma_2 \end{cases}, \quad R_j = \frac{R - \sigma_j I}{\sigma_1 - \sigma_2}.$$

To get the convergence condition, we compute the eigenvalues of R explicitly:

$$\sigma_{1,2} = \frac{1}{2} \left(\frac{\beta+1}{1+\eta\lambda_i} \pm \sqrt{\left(\frac{\beta+1}{1+\eta\lambda_i} \right)^2 - \frac{4\beta}{1+\eta\lambda_i}} \right).$$

Now, to get the convergence criterion, we need to examine the conditions that lead to $\max\{|\sigma_1|, |\sigma_2|\} < 1$. If the eigenvalues computed above are complex, i.e.,

$$\begin{aligned} \left(\frac{\beta+1}{1+\eta\lambda_i} \right)^2 - \frac{4\beta}{1+\eta\lambda_i} < 0 \implies |\sigma_1| = |\sigma_2| &= \sqrt{\frac{1}{4}\left(\frac{\beta+1}{1+\eta\lambda_i} \right)^2 + \left| \frac{1}{4}\left(\frac{\beta+1}{1+\eta\lambda_i} \right)^2 - \frac{\beta}{1+\eta\lambda_i} \right|} \\ &= \sqrt{\frac{\beta}{1+\eta\lambda_i}}. \end{aligned}$$

We need the above quantity to be less than 1 to converge, so we need

$$\sqrt{\frac{\beta}{1+\eta\lambda_i}} < 1 \iff \beta - 1 < \eta\lambda_i \iff \eta > \frac{\beta-1}{\lambda_i}.$$

Now, if the eigenvalues are real, i.e.,

$$\left(\frac{\beta+1}{1+\eta\lambda_i} \right)^2 - \frac{4\beta}{1+\eta\lambda_i} \geq 0 \implies \max\{|\sigma_1|, |\sigma_2|\} = \frac{1}{2} \max \left\{ \left| \frac{\beta+1}{1+\eta\lambda_i} \pm \sqrt{\left(\frac{\beta+1}{1+\eta\lambda_i} \right)^2 - \frac{4\beta}{1+\eta\lambda_i}} \right| \right\}.$$

Cases can be further divided into two. In the first case, when we have $\frac{\beta+1}{1+\eta\lambda_i} > 0$, we have $\sigma_1 \geq \sigma_2 \geq 0$, because the square-root term is non-negative. Therefore, to have $\max\{|\sigma_1|, |\sigma_2|\} < 1$, we need

$$\sigma_1 = \frac{1}{2}\left(\frac{\beta+1}{1+\eta\lambda_i} + \sqrt{\left(\frac{\beta+1}{1+\eta\lambda_i}\right)^2 - \frac{4\beta}{1+\eta\lambda_i}}\right) < 1.$$

In the second case, when we have $\frac{\beta+1}{1+\eta\lambda_i} < 0$, we have $|\sigma_2| \geq |\sigma_1|$. Therefore, to have $\max\{|\sigma_1|, |\sigma_2|\} < 1$, we need

$$\sigma_2 = \frac{1}{2}\left(\frac{\beta+1}{1+\eta\lambda_i} - \sqrt{\left(\frac{\beta+1}{1+\eta\lambda_i}\right)^2 - \frac{4\beta}{1+\eta\lambda_i}}\right) > -1.$$

■

Appendix B. Proofs for Section 4

Recall the recursion of SPPAM:

$$x_{t+1}^+ = x_t - \eta\nabla f(x_{t+1}^+) + \beta(x_t - x_{t-1}) \tag{19}$$
$$x_{t+1} = x_{t+1}^+ - \eta\varepsilon_{t+1}. \tag{20}$$

We will refer to (19) as PPAM (without stochastic errors).

Lemma 13 *For μ-strongly convex $f(\cdot)$, it holds that for all x,*

$$\|\nabla f(x)\|_2^2 \geq \mu^2\|x - x^\star\|_2^2.$$

Proof By strong convexity, we have for all x and y,

$$f(y) \geq f(x) + \nabla f(x)^\top(y - x) + \tfrac{\mu}{2}\|y - x\|_2^2.$$

Since minimization retains inequality, we can minimize each side. On the left hand side, we have $\min_y\{f(y)\} = f(x^\star)$. On the right hand side, we take the derivative with respect to y and set it to zero to obtain:

$$\nabla f(x) + \mu(y - x) = 0 \Rightarrow y = x - \tfrac{1}{\mu}\nabla f(x).$$

Plugging back, we get:

$$\begin{aligned} f(x^\star) &\geq f(x) + \nabla f(x)^\top(x - \tfrac{1}{\mu}\nabla f(x) - x) + \tfrac{\mu}{2}\|x - x + \tfrac{1}{\mu}\nabla f(x)\|_2^2 \\ &= f(x) - \tfrac{1}{2\mu}\|\nabla f(x)\|_2^2. \end{aligned}$$

Rearraging, we have

$$\begin{aligned} \|\nabla f(x)\|_2^2 &\geq 2\mu(f(x) - f(x^\star)) \\ &\geq \mu^2\|x - x^\star\|_2^2, \end{aligned}$$

where last inequality uses strong convexity with $y = x$ and $x = x^\star$. ■

Proof of Theorem 5

Proof From PPAM in (19), subtract $x^\star$ on both sides and take the norm squared:

$$\begin{aligned} x_{t+1}^+ - x^\star &= x_t - x^\star - \eta\nabla f(x_{t+1}^+) + \beta(x_t - x_{t-1}) \Rightarrow \\ \|x_{t+1}^+ - x^\star\|_2^2 &= \|x_t - x^\star\|_2^2 + \eta^2\|\nabla f(x_{t+1}^+)\|_2^2 + \beta^2\|x_t - x_{t-1}\|_2^2 \\ &\quad - 2\eta(x_t - x^\star)^\top\nabla f(x_{t+1}^+) \\ &\quad + 2\beta(x_t - x^\star)^\top(x_t - x_{t-1}) \\ &\quad - 2\beta\eta(x_t - x_{t-1})^\top\nabla f(x_{t+1}^+). \end{aligned}$$

For the fourth term, observe that

$$\begin{aligned} (x_t - x^\star)^\top\nabla f(x_{t+1}^+) &= (x_{t+1}^+ - x^\star + \eta\nabla f(x_{t+1}^+) - \beta(x_t - x_{t-1}))^\top\nabla f(x_{t+1}^+) \\ &= (x_{t+1}^+ - x^\star)^\top\nabla f(x_{t+1}^+) + \eta\|\nabla f(x_{t+1}^+)\|_2^2 - \beta(x_t - x_{t-1})^\top\nabla f(x_{t+1}^+) \Rightarrow \\ -2\eta(x_t - x^\star)^\top\nabla f(x_{t+1}^+) &= -2\eta(x_{t+1}^+ - x^\star)^\top\nabla f(x_{t+1}^+) - 2\eta^2\|\nabla f(x_{t+1}^+)\|_2^2 + 2\beta\eta(x_t - x_{t-1})^\top\nabla f(x_{t+1}^+) \\ &\leq -2\eta\mu\|x_{t+1}^+ - x^\star\|_2^2 - 2\eta^2\|\nabla f(x_{t+1}^+)\|_2^2 + 2\beta\eta(x_t - x_{t-1})^\top\nabla f(x_{t+1}^+), \end{aligned}$$

where the last inequality uses strong convexity of $f(\cdot)$.

For the third and fifth term, observe that

$$\begin{aligned} &\beta^2\|x_t - x_{t-1}\|_2^2 + 2\beta(x_t - x^\star)^\top(x_t - x_{t-1}) \\ &\qquad = \|\beta(x_t - x_{t-1}) + (x_t - x^\star)\|_2^2 - \|x_t - x^\star\|_2^2 \\ &\qquad = \|(1+\beta)x_t - \beta x_{t-1} - x^\star\|_2^2 - \|x_t - x^\star\|_2^2 \\ &\qquad = \|(1+\beta)(x_t - x^\star) - \beta(x_{t-1} - x^\star)\|_2^2 - \|x_t - x^\star\|_2^2 \\ &\qquad \leq (1+\beta)^2(1+\zeta)\|x_t - x^\star\|_2^2 + \beta^2(1+\tfrac{1}{\zeta})\|x_{t-1} - x^\star\|_2^2 - \|x_t - x^\star\|_2^2, \end{aligned}$$

where last inequality uses Young's inequality: $\|a+b\|_2^2 \leq (1+\zeta)\|a\|_2^2 + (1+\frac{1}{\zeta})\|b\|_2^2$ for $\zeta > 0$.

Combining terms, we have

$$\begin{aligned} \|x_{t+1}^+ - x^\star\|_2^2 &\leq \|x_t - x^\star\|_2^2 + \eta^2\|\nabla f(x_{t+1}^+)\|_2^2 \\ &\quad - 2\eta\mu\|x_{t+1}^+ - x^\star\|_2^2 - 2\eta^2\|\nabla f(x_{t+1}^+)\|_2^2 + \cancel{2\beta\eta(x_t - x_{t-1})^\top\nabla f(x_{t+1}^+)} \\ &\quad + (1+\beta)^2(1+\zeta)\|x_t - x^\star\|_2^2 + \beta^2(1+\tfrac{1}{\zeta})\|x_{t-1} - x^\star\|_2^2 - \|x_t - x^\star\|_2^2 \\ &\quad \cancel{- 2\beta\eta(x_t - x_{t-1})^\top\nabla f(x_{t+1}^+)} \\ &= -2\eta\mu\|x_{t+1}^+ - x^\star\|_2^2 - \eta^2\|\nabla f(x_{t+1}^+)\|_2^2 + (1+\beta)^2(1+\zeta)\|x_t - x^\star\|_2^2 \\ &\quad + \beta^2(1+\tfrac{1}{\zeta})\|x_{t-1} - x^\star\|_2^2 \\ &\leq -(2\eta\mu + \eta^2\mu^2)\|x_{t+1}^+ - x^\star\|_2^2 + (1+\beta)^2(1+\zeta)\|x_t - x^\star\|_2^2 \\ &\quad + \beta^2(1+\tfrac{1}{\zeta})\|x_{t-1} - x^\star\|_2^2, \end{aligned}$$

where the last inequality is by Lemma 13. Grouping the same terms, we get:

$$\begin{aligned} &(1+\eta\mu)^2\|x_{t+1}^+ - x^\star\|_2^2 \leq (1+\beta)^2(1+\zeta)\|x_t - x^\star\|_2^2 + \beta^2(1+\tfrac{1}{\zeta})\|x_{t-1} - x^\star\|_2^2 \Rightarrow \\ &\qquad \|x_{t+1}^+ - x^\star\|_2^2 \leq \frac{(1+\beta)^2(1+\zeta)}{(1+\eta\mu)^2}\|x_t - x^\star\|_2^2 + \frac{\beta^2(1+\frac{1}{\zeta})}{(1+\eta\mu)^2}\|x_{t-1} - x^\star\|_2^2. \end{aligned} \tag{21}$$

Now, choose $\zeta = \frac{4}{(1+\beta)^2} - 1$, which is positive for $0 < \beta < 1$. Then, each coefficient in the two terms on the RHS above reduces to:

$$\frac{(1+\beta)^2(1+\zeta)}{(1+\eta\mu)^2} = \frac{4}{(1+\eta\mu)^2}, \quad \text{and} \quad \frac{\beta^2(1+\frac{1}{\zeta})}{(1+\eta\mu)^2} = \frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}.$$

Therefore, our original recursion in (21) reduces to

$$\|x_{t+1}^+ - x^\star\|_2^2 \leq \frac{4}{(1+\eta\mu)^2}\|x_t - x^\star\|_2^2 + \frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}\|x_{t-1} - x^\star\|_2^2. \tag{22}$$

Note that from (19), we have

$$x_{t+1}^+ + \eta\nabla f(x_{t+1}^+) = x_t + \beta(x_t - x_{t-1}).$$

Thus, x_{t+1}^+ is deterministic given x_t and x_{t-1}. Therefore, going back to SPPAM in (20) and taking expectations, we have:

$$\begin{aligned}
\mathbb{E}\left[\|x_{t+1} - x^\star\|_2^2\right] &= \mathbb{E}\left[\|x_{t+1}^+ - x^\star\|_2^2\right] - 2\eta\mathbb{E}\left[\langle x_{t+1}^+ - x^\star, \varepsilon_{t+1}\rangle\right] + \eta^2\mathbb{E}\left[\|\varepsilon_{t+1}\|_2^2\right] \\
&= \mathbb{E}\left[\|x_{t+1}^+ - x^\star\|_2^2\right] + \eta^2\mathbb{E}\left[\|\varepsilon_{t+1}\|_2^2\right] \\
&\leq \tfrac{4}{(1+\eta\mu)^2}\mathbb{E}\left[\|x_t - x^\star\|_2^2\right] + \tfrac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}\mathbb{E}\left[\|x_{t-1} - x^\star\|_2^2\right] + \eta^2\sigma^2,
\end{aligned}$$

where the last inequality follows from (22) and Assumption 2.

■

Proof of Lemma 6

Proof (11) in the main text leads to the following 2×2 recursion:

$$\begin{bmatrix} \mathbb{E}\left[\|x_{t+1} - x^\star\|_2^2\right] \\ \mathbb{E}\left[\|x_t - x^\star\|_2^2\right] \end{bmatrix} \leq \underbrace{\begin{bmatrix} \frac{4}{(1+\eta\mu)^2} & \frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)} \\ 1 & 0 \end{bmatrix}}_{:=A} \cdot \begin{bmatrix} \mathbb{E}\left[\|x_t - x^\star\|_2^2\right] \\ \mathbb{E}\left[\|x_{t-1} - x^\star\|_2^2\right] \end{bmatrix} + \begin{bmatrix} \eta^2\sigma^2 \\ 0 \end{bmatrix}. \tag{23}$$

Eigenvalues of a 2×2 matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is given by $\frac{(a+d) \pm \sqrt{(a+d)^2 - 4(ad-bc)}}{2}$. Thus, eigenvalues of the contraction matrix A is given by

$$\begin{aligned}
\sigma_{1,2} &= \frac{4}{2(1+\eta\mu)^2} \pm \frac{1}{2} \cdot \sqrt{\left(\frac{4}{(1+\eta\mu)^2}\right)^2 - 4\left(-\frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}\right)} \\
&= \frac{2}{(1+\eta\mu)^2} \pm \sqrt{\frac{4}{(1+\eta\mu)^4} + \frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}}.
\end{aligned}$$

Note that all terms are positive. Thus, the maximum eigenvalue is determined by

$$\sigma_1 = \frac{2}{(1+\eta\mu)^2} + \sqrt{\frac{4}{(1+\eta\mu)^4} + \frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}}.$$

■

Proof of Corollary 7

Proof We want to see under what condition the following holds:

$$\frac{1}{1+2\eta\mu} > \frac{2}{(1+\eta\mu)^2} + \sqrt{\frac{4}{(1+\eta\mu)^4} + \frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}} \Rightarrow$$
$$\frac{1}{1+2\eta\mu} - \frac{2}{(1+\eta\mu)^2} > \sqrt{\frac{4}{(1+\eta\mu)^4} + \frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}}$$

Squaring both sides [4] and grouping the same terms, we get

$$\left(\frac{1}{1+2\eta\mu}\right)^2 - \frac{4}{(1+\eta\mu)^2(1+2\eta\mu)} > \frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)} \Rightarrow$$
$$\frac{\eta^2\mu^2 - 6\eta\mu - 3}{(1+2\eta\mu)^2} > \frac{4\beta^2}{4-(1+\beta)^2}.$$

■

Proof of Theorem 9

Proof Unrolling the recursion (23) for T iterations, we get

$$\begin{aligned}
\begin{bmatrix} \mathbb{E}\left[\|x_T - x^\star\|_2^2\right] \\ \mathbb{E}\left[\|x_{T-1} - x^\star\|_2^2\right] \end{bmatrix} &\leq A^T \cdot \begin{bmatrix} \|x_0 - x^\star\|_2^2 \\ \|x_{-1} - x^\star\|_2^2 \end{bmatrix} + \left(\sum_{i=1}^{T-1} A^i\right) \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta^2\sigma^2 \\
&= \left(\frac{\sigma_1^T - \sigma_2^T}{\sigma_1 - \sigma_2} A - \sigma_1\sigma_2 \frac{\sigma_1^{T-1} - \sigma_2^{T-1}}{\sigma_1 - \sigma_2} I\right) \cdot \begin{bmatrix} \|x_0 - x^\star\|_2^2 \\ \|x_{-1} - x^\star\|_2^2 \end{bmatrix} + \left(\sum_{i=1}^{T-1} A^i\right) \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta^2\sigma^2 \\
&\leq \frac{2\sigma_1^T}{\sigma_1 - \sigma_2}(A + I) \cdot \begin{bmatrix} \|x_0 - x^\star\|_2^2 \\ \|x_{-1} - x^\star\|_2^2 \end{bmatrix} + \left(\sum_{i=1}^{T-1} A^i\right) \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta^2\sigma^2, \qquad (24)
\end{aligned}$$

where the last equality is using the formula in Williams (1992), and the last inequality is due to

$$\frac{\sigma_1^T - \sigma_2^T}{\sigma_1 - \sigma_2} \leq \frac{|\sigma_1|^T + |\sigma_2|^T}{\sigma_1 - \sigma_2} \leq \frac{2|\sigma_1|^T}{\sigma_1 - \sigma 2} = \frac{2\sigma_1^T}{\sigma_1 - \sigma 2},$$

and

$$\begin{aligned}
-\sigma_1\sigma_2 \frac{\sigma_1^{T-1} - \sigma_2^{T-1}}{\sigma_1 - \sigma_2} &\leq |\sigma_1\sigma_2| \frac{|\sigma_1|^{T-1} + |\sigma_2|^{T-1}}{\sigma_1 - \sigma_2} \\
&= \frac{|\sigma_2| \cdot |\sigma_1|^T - |\sigma_1| \cdot |\sigma_2|^T}{\sigma_1 - \sigma_2} \leq \frac{|\sigma_1|^T + |\sigma_1|^T}{\sigma_1 - \sigma_2} \leq \frac{2\sigma_1^T}{\sigma_1 - \sigma_2},
\end{aligned}$$

under the assumption that $|\sigma_{1,2}| < 1$, which we justified in Theorem 10.

Now, focusing on the error term, $\sum_{i=1}^{T-1} A^i$ converge to:

$$\sum_{i=1}^{T-1} A^i = (I - A)^{-1}(I - A^T) := B(I - A^\top).$$

4. Here, to square both sides and maintain the inequality, we assume $\eta\mu > 1$, which holds by the condition (16).

Then,

$$\begin{aligned}\left(\sum_{i=1}^{T-1} A^i\right)\begin{bmatrix}1\\0\end{bmatrix}\eta^2\sigma^2 &= (I-A)^{-1}(I-A^T)\begin{bmatrix}1\\0\end{bmatrix}\eta^2\sigma^2\\ &= -BA^T\begin{bmatrix}1\\0\end{bmatrix}\eta^2\sigma^2 + B\begin{bmatrix}1\\0\end{bmatrix}\eta^2\sigma^2\\ &\leq B\left(-\frac{\sigma_1^T-\sigma_2^T}{\sigma_1-\sigma_2}A + \sigma_1\sigma_2\frac{\sigma_1^{T-1}-\sigma_2^{T-1}}{\sigma_1-\sigma_2}I\right) + B\begin{bmatrix}1\\0\end{bmatrix}\eta^2\sigma^2\\ &\leq B\left(\frac{2\sigma_1^T}{\sigma_1-\sigma_2}A + \frac{2\sigma_1^T}{\sigma_1-\sigma_2}I\right) + B\begin{bmatrix}1\\0\end{bmatrix}\eta^2\sigma^2\\ &= \frac{2\sigma_1^T}{\sigma_1-\sigma_2}B(A+I)\begin{bmatrix}1\\0\end{bmatrix}\eta^2\sigma^2 + B\begin{bmatrix}1\\0\end{bmatrix}\eta^2\sigma^2. \end{aligned} \tag{25}$$

Computing $(I-A)^{-1} := B$ term first, we get

$$\begin{aligned}(I-A)^{-1} &= \left(1-\frac{4}{(1+\eta\mu)^2}-\frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}\right)^{-1}\begin{bmatrix}1 & \frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}\\ 1 & 1-\frac{4}{(1+\eta\mu)^2}\end{bmatrix}\\ &:= \frac{1}{p-q}\cdot\begin{bmatrix}1 & q\\ 1 & p\end{bmatrix}\end{aligned}$$

Then,

$$B(A+I)\begin{bmatrix}1\\0\end{bmatrix} = \frac{1}{p-q}\begin{bmatrix}2-p+q\\2\end{bmatrix}, \quad \text{and} \quad B\begin{bmatrix}1\\0\end{bmatrix} = \frac{1}{p-q}\begin{bmatrix}1\\1\end{bmatrix}. \tag{26}$$

Combining (24), (25), and (26), we have

$$\begin{aligned}\begin{bmatrix}\mathbb{E}\left[\|x_T-x^\star\|_2^2\right]\\ \mathbb{E}\left[\|x_{T-1}-x^\star\|_2^2\right]\end{bmatrix} &\leq \frac{2\sigma_1^T}{\sigma_1-\sigma_2}(A+I)\cdot\begin{bmatrix}\|x_0-x^\star\|_2^2\\ \|x_{-1}-x^\star\|_2^2\end{bmatrix} + \left(\sum_{i=1}^{T-1}A^i\right)\begin{bmatrix}1\\0\end{bmatrix}\eta^2\sigma^2\\ &\leq \frac{2\sigma_1^T}{\sigma_1-\sigma_2}\left((A+I)\cdot\begin{bmatrix}\|x_0-x^\star\|_2^2\\ \|x_{-1}-x^\star\|_2^2\end{bmatrix} + \frac{1}{p-q}\begin{bmatrix}2-p+q\\2\end{bmatrix}\eta^2\sigma^2\right) + \frac{1}{p-q}\begin{bmatrix}1\\1\end{bmatrix}\eta^2\sigma^2.\end{aligned}$$

Since we assume $x_0 = x_{-1}$, we have $\|x_0-x^\star\|_2^2 = \|x_{-1}-x^\star\|_2^2$. Using this and computing $(A+I)$ explicitly, the top row results in:

$$\mathbb{E}\left[\|x_T-x^\star\|_2^2\right] \leq \frac{2\sigma_1^T}{\sigma_1-\sigma_2}\left((2-p+q)\cdot\left(\|x_0-x^\star\|_2^2 + \frac{\eta^2\sigma^2}{p-q}\right)\right) + \frac{\eta^2\sigma^2}{p-q}.$$

Observe that

$$\begin{aligned}p-q &= 1-\frac{4}{(1+\eta\mu)^2}-\frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)} := 1-\theta \Rightarrow\\ 2-p+q &= 1+\theta,\end{aligned}$$

so we can write the above recursion as:

$$\mathbb{E}\left[\|x_T-x^\star\|_2^2\right] \leq \frac{2\sigma_1^T}{\sigma_1-\sigma_2}\left((1+\theta)\cdot\left(\|x_0-x^\star\|_2^2 + \frac{\eta^2\sigma^2}{1-\theta}\right)\right) + \frac{\eta^2\sigma^2}{1-\theta}.$$

Thus, we see that the initial condition is discounted by the factor

$$\frac{2\sigma_1^T}{\sigma_1-\sigma_2} = \tau^{-1}\cdot\left(\frac{2}{(1+\eta\mu)^2}+\tau\right)^T$$

where $\tau = \sqrt{\frac{4}{(1+\eta\mu)^4}+\frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}}$, up to a region that depends on $O(\eta^2\sigma^2)$. ■

Proof of Theorem 10

Proof We want to analyze the term

$$\frac{2\sigma_1^T}{\sigma_1-\sigma_2} = \frac{\left(\frac{2}{(1+\eta\mu)^2}+\sqrt{\frac{4}{(1+\eta\mu)^4}+\frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}}\right)^T}{\sqrt{\frac{4}{(1+\eta\mu)^4}+\frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}}}.$$

First, notice that $\frac{2}{(1+\eta\mu)^2} \le \sqrt{\frac{4}{(1+\eta\mu)^4}+\frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}}$. Thus,

$$\begin{aligned}\frac{2\sigma_1^T}{\sigma_1-\sigma_2} &\le \frac{\left(2\sqrt{\frac{4}{(1+\eta\mu)^4}+\frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}}\right)^T}{\sqrt{\frac{4}{(1+\eta\mu)^4}+\frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}}} \\ &= 2\cdot\left(2\sqrt{\frac{4}{(1+\eta\mu)^4}+\frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}}\right)^{T-1}.\end{aligned}$$

Therefore, if $2\sqrt{\frac{4}{(1+\eta\mu)^4}+\frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}} < 1$, we have exponential discount of the initial conditions. This condition leads to the desired result immediately. Also note that this condition justifies $|\sigma_{1,2}| < 1$, which we required in the proof of Theorem 9. ■

Appendix C. Derivation of Algorithm 1

In this section, we present the derivation of the procedure in Algorithm 1. Note that the following derivation is based on implicit SGD procedure presented in Toulis et al. (2014), extended to SPPAM.

We have

$$\begin{aligned}x_t &= x_{t-1}+\eta\nabla f_i(x_t)+\beta(x_{t-1}-x_{t-2}) \\ &= x_{t-1}+\eta\left(b_{i_t}-h(x_t^\top a_{i_t})\right)a_{i_t}+\beta(x_{t-1}-x_{t-2}),\end{aligned}$$

where in the last equality we substituted the gradient for GLM for a (uniformly sampled) single data point (a_{i_t}, b_{i_t}), with $h(\cdot)$ being the *mean* function from the main text.

First, multiply both sides by a_{i_t}. Then,

$$x_t^\top a_{i_t} = x_{t-1}^\top a_{i_t}+\eta\left(b_{i_t}-h(x_t^\top a_{i_t})\right)a_{i_t}^\top a_{i_t}+\beta(x_{t-1}-x_{t-2})^\top a_{i_t}.$$

Now let $\xi_t := \eta\left(b_{i_t} - h(x_t^\top a_{i_t})\right) \in \mathbb{R}$. Then, we have:

$$x_t^\top a_{i_t} = \xi_t \|a_{i_t}\|_2^2 + (1+\beta)x_{t-1}^\top a_{i_t} - \beta x_{t-2}^\top a_{i_t}.$$

We now apply the transfer function $h(\cdot)$ on both sides to get:

$$h(x_t^\top a_{i_t}) = h\left(\xi_t \|a_{i_t}\|_2^2 + (1+\beta)x_{t-1}^\top a_{i_t} - \beta x_{t-2}^\top a_{i_t}\right). \tag{27}$$

But from $\xi_t = \eta\left(b_{i_t} - h(x_t^\top a_{i_t})\right)$, we can re-arrange to get:

$$h(x_t^\top a_{i_t}) = b_{i_t} - \frac{\xi_t}{\eta}.$$

Plugging this back into the left-hand side of (27) and solving for ξ_t, we have:

$$\begin{aligned}
\xi_t &= \eta\left(y - h(\xi_t \|a_{i_t}\|_2^2 + (1+\beta)x_{t-1}^\top a_{i_t} - \beta x_{t-2}^\top a_{i_t})\right) \\
x_t &= (1+\beta)x_{t-1} + \xi_t a_{i_t} - \beta x_{t-2} = x_{t-1} + \xi_t a_{i_t} + \beta(x_{t-1} - x_{t-2}),
\end{aligned}$$

arriving at Algorithm 1. Note that this derivation is assuming a single data point is sampled, but the derivation for mini-batch version is straightforward.