Acceleration and Stability of the Stochastic Proximal Point Algorithm

Junhyung Lyle Kim (Rice CS), Panos Toulis (UChicago Booth), Anastasios Kyrillidis (Rice CS)

[Empirical risk minimization and SGD]

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

- SGD and SGD with momentum (SGDM) became the de facto algorithms. BUT:
  - SGD can take long to converge with smell step size/ diverge easily if step size is misspecified

GD: $O(1/t)$ VS. SGD: $O(1/\sqrt{t})$

SGD : $\mathbb{E}\|x_t - x^\star\|_2^2 \leq 2\exp(4L^2\eta_1^2\log(t))\|x_0 - x^\star\|_2^2 \cdots$

  - SGDM can be more unstable than SGD due to the gradient noise accumulation

E.g., Liu and Belkin (2019), Assran and Rabbat (2020).

[Why Proximal Point Algorithm?]

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

- PPA changes the conditioning of the problem by adding a quadratic term to the objective function
- Equivalent to implicit gradient descent (IGD) by the first-order optimality condition
- Stochastic setting:

SPPA : $\mathbb{E}\|x_t - x^\star\|_2^2 \leq \exp(-\log(1+2\eta_1\mu)\log(t))\|x_0 - x^\star\|_2^2 \cdots$

[Intuition about SPPAM]

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

- Disregarding the stochastic error for simplicity, above can be written as the solution to:

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

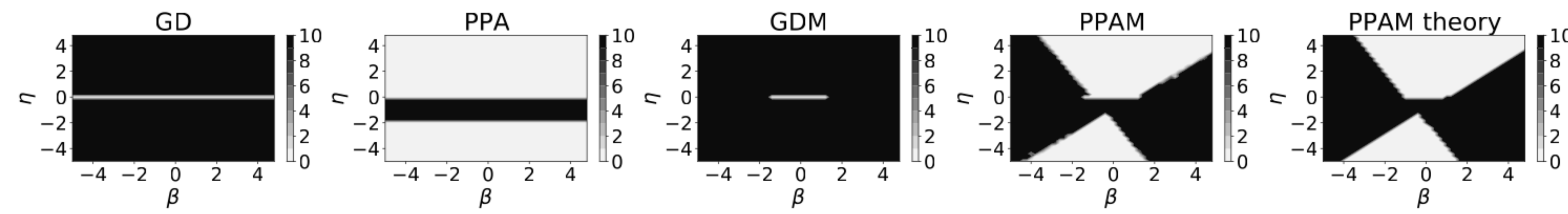
- On top of minimizing $f(x)$ and staying close to x_t, the algorithm also tries to move along the direction from x_{t-1} to x_t
- This intuition exactly aligns with that of Polyak's momentum applied to e.g., SGD

[Our Contribution]

- We show that SPPAM enjoys linear convergence with a better contraction factor than SPPA, and characterize the conditions on η and β that result in acceleration.
- We also characterize the condition that leads to the exponential discount of initial conditions for SPPAM, which is significantly easier to satisfy compared to SGDM.
- Empirically SPPAM enjoys the both advantages: 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.

[The Quadratic Model Case]

- Conditions on η and β for different algorithms to solve: $f(x) = \frac{1}{2}x^\top Ax - b^\top x$



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

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

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:*

- $\eta > \frac{\beta-1}{\lambda_i}$, *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.*

[Acceleration]

- Main assumptions:

Assumption 1. *$f(\cdot)$ is a μ-strongly convex function, satisfying:*

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

for some fixed $\mu > 0$ and for all x and y.

Assumption 2. *There exists fixed $\sigma^2 > 0$ such that:*

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

- Iteration invariant bound of SPPAM:

Theorem 1. *For μ-strongly convex $f(\cdot)$, SPPAM in (5) satisfies the following iteration invariant bound:*

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

- Above can be written in 2 x 2 system where the contraction matrix A determines convergence rate

$$\begin{bmatrix}\mathbb{E}\left[\|x_{t+1} - x^\star\|_2^2\right]\\ \mathbb{E}\left[\|x_t - x^\star\|_2^2\right]\end{bmatrix} \leq A\begin{bmatrix}\mathbb{E}\left[\|x_t - x^\star\|_2^2\right]\\ \mathbb{E}\left[\|x_{t-1} - x^\star\|_2^2\right]\end{bmatrix} + \begin{bmatrix}\eta^2\sigma^2\\ 0\end{bmatrix}$$

- Condition on η and β that lead to faster discount factor than SPPA:

Corollary 1. *For μ-strongly convex $f(\cdot)$, SPPAM in (5) converges faster than stochastic PPA in (4) if:*

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

[Stability]

- Convergence (to a neighborhood):

Theorem 3. *For μ-strongly convex $f(\cdot)$, assume SPPAM in (5) is initialized with $x_0 = x_{-1}$. Then, after T iterations, we have:*

$$\mathbb{E}\left[\|x_T - x^\star\|_2^2\right] \leq \frac{2\sigma_1^T}{\sigma_1 - \sigma_2}\left(\left(\|x_0 - x^\star\|_2^2 + \frac{\eta^2\sigma^2}{1-\theta}\right)\cdot(1+\theta)\right) + \frac{\eta^2\sigma^2}{1-\theta} \tag{16}$$

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

$$\frac{2\sigma_1^T}{\sigma_1 - \sigma_2} = \tau^{-1}\cdot\left(\frac{2}{(1+\eta\mu)^2} + \tau\right)^T \tag{17}$$

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

- Condition on η and β that lead to exponential discount of initial conditions:

Theorem 4. *Let the following condition hold:*

$$\tau = \sqrt{\frac{4}{(1+\eta\mu)^4} + \frac{4\beta^2}{(1+\eta\mu)^2(4-(1+\beta)^2)}} < \frac{1}{2}. \tag{18}$$

Then, for μ-strongly convex $f(\cdot)$, the initial conditions of SPPAM exponentially discount: i.e., in (16),

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

where $C \in (0, 1)$.

[Unfair Comparison]

- Assran and Rabbit (2020): for Nesterov's accelerated SGD to converge for *strongly convex quadratic* $f(\cdot)$:

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

↓ $\beta = 0.9$

Nesterov's accelerated SGD (strongly convex *quadratic*):

$0.0028 \approx \frac{1}{361} \leq \eta\lambda \leq \frac{24}{19} \approx 1.26$ for $\lambda \in \{\mu, L\}$

VS.

SPPAM (strongly convex, Theorem 4):

$\eta\mu > 4.81$ with $\beta = 0.9$

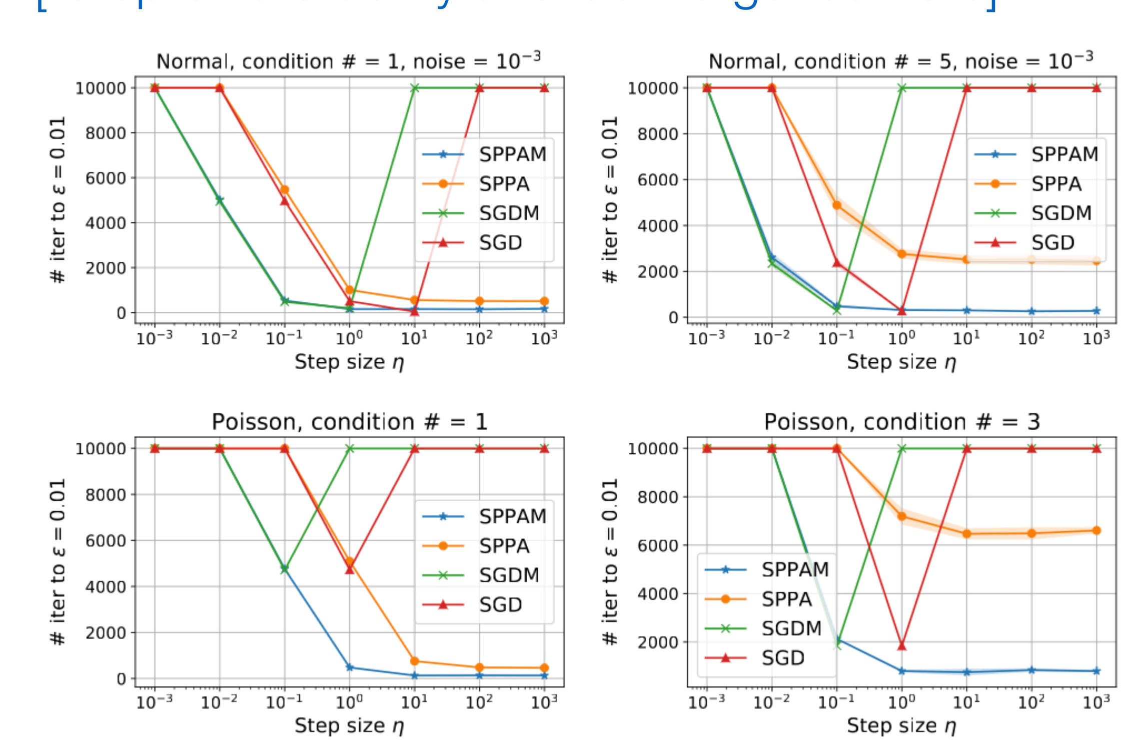
[Experiments]

- Generalized Linear Model (GLM) :
  - Labels: $b_i \in \mathbb{R}$
  - Features: $a_i \in \mathbb{R}^p$
  - True model parameter: $x^\star \in \mathbb{R}^p$

$$b_i \mid a_i \sim \exp\left(\frac{\gamma b_i - c_1(\gamma)}{\omega}c_2(b_i, \omega)\right)$$

- Linear predictor $\gamma = \langle a_i, x^\star\rangle$ with mean functions $h(\cdot)$:
  - Normal: $h(\gamma) = \gamma$
  - Logistic: $h(\gamma) = e^\gamma(1+e^\gamma)^{-1}$
  - Poisson: $h(\gamma) = e^\gamma$

[Step Size Stability and Convergence Rate]



- SGD and SGDM only converge for specific η and β
- SPPA and SPPAM converge for much wider ranges
- SPPAM converges faster than SPPA
  - Convergence rate of SPPAM matches that of SGDM when the latter converges