# Accelerated, Optimal, and Parallel: Some Results on Model-Based Stochastic Optimization

Karan Chadha$^\star$, Gary Cheng$^\star$, John Duchi

Stanford Univeristy

ICML 2022

# Motivation

**Objective**

$$\begin{aligned} \text{minimize} \;\; & f(x) = \mathbb{E}_P[F(x; S)] = \int_{\mathcal{S}} F(x; s) dP(s) \\ \text{subject to} \;\; & x \in \mathcal{X}. \end{aligned}$$

# Motivation

## Objective

$$\begin{aligned} \text{minimize} \;\; & f(x) = \mathbb{E}_P[F(x; S)] = \int_{\mathcal{S}} F(x; s) dP(s) \\ \text{subject to} \;\; & x \in \mathcal{X}. \end{aligned}$$

## Stochastic Gradient Method (SGM)

$$x_{k+1} = x_k - \alpha_k g_k, \qquad g_k \in \partial F(x_k; S_k)$$

# Motivation

## Objective

$$\begin{aligned} &\text{minimize} \;\; f(x) = \mathbb{E}_P[F(x;S)] = \int_{\mathcal{S}} F(x;s)dP(s) \\ &\text{subject to} \;\; x \in \mathcal{X}. \end{aligned}$$

## Stochastic Gradient Method (SGM)

$$x_{k+1} = x_k - \alpha_k g_k, \qquad g_k \in \partial F(x_k; S_k)$$

**Pros:**

- Efficient and easy
- Simple minibatch extensions
- Easy to analyze

**Cons:**

- Not robust to stepsize choice
- Weak stability guarantees

# aProx Update

Asi and Duchi [2019] suggest using aProx family of algorithms

**Repeat:**

- Sample $S_k$
- Construct a model $F_{x_k}(\cdot; S_k)$ of sample function $F(\cdot; S_k)$ at $x_k$
- Update:

$$x_{k+1} := \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ F_{x_k}(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\},$$

# aProx Update

$$x_{k+1} := \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ F_{x_k}(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\},$$

**Model Conditions:**

1. (Convex)

$$y \mapsto F_x(y; s) \quad \text{is convex}$$

2. (Lower Bound)

$$F_x(y; s) \le F(y; s) \quad \text{for all } y.$$

3. (Local Accuracy)

$$F_x(x; s) = F(x; s).$$

# aProx Update

$$x_{k+1} := \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ F_{x_k}(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\},$$

**Examples:**

- *Stochastic gradient methods:*

$$F_x(y; s) := F(x; s) + \langle F'(x; s), y - x \rangle.$$

- *Proximal point methods:*

$$F_x(y; s) := F(y; s).$$

- *Truncated methods:*

$$F_x(y; s) := \Big( F(x; s) + \langle F'(x; s), y - x \rangle \Big)_+.$$

# aProx Update

$$x_{k+1} := \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ F_{x_k}(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\},$$

**Pros:**

- Robust to stepsize choice
- Better models yield better convergence
- Efficient and easy
- Fast rates on growth + interpolation problems
- Minibatching (Asi et al. [2020])

**Cons:**

- Acceleration?
- Optimal for growth + interpolation problems?

# Results

# Accelerated aProx

$$
\begin{aligned}
y_k &= (1-\theta_k)x_k + \theta_k z_k \\
z_{k+1} &= \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ F_{y_k}(x; S_k) + \frac{1}{2\alpha_k} \|x - z_k\|_2^2 \right\} \\
x_{k+1} &= (1-\theta_k)x_k + \theta_k z_{k+1}
\end{aligned}
$$

# Non-Asymptotic Results (Smooth Functions)

Theorem (Chadha, C. & Duchi 22)

*For a smooth population function $f$, under bounded variance of the gradient, accelerated* aProx *with minibatch $m$ has error rate*

$$\mathbb{E}\left[f\left(x_{k+1}\right)-f(x^{\star})\right] \lesssim \frac{1}{k^2}+\frac{1}{\sqrt{km}}.$$

**(matches accelerated SGM)**

# Interpolation Problems

**Definition (Interpolation Problems)**

if for all $s \in \mathcal{S}$ we have, $\inf_{x \in \mathcal{X}} F(x; s) = F(x^\star; s)$.

# Interpolation Problems

Definition (Interpolation Problems)

if for all $s \in \mathcal{S}$ we have, $\inf_{x \in \mathcal{X}} F(x; s) = F(x^\star; s)$.

Assumption (Quadratic-Growth)

$$\mathbb{E}\left[\frac{(F(x; S) - F(x^\star, S))^2}{\|F'(x; S)\|_2^2}\right] \geq \lambda_1 \mathrm{dist}(x, \mathcal{X}^\star)^2.$$

# Fast and Optimal Convergence

Theorem (Chadha, C. & Duchi 22)

*For Quadratic-growth interpolation problems,* aProx *with truncated models has error rates*

$$\mathbb{E}[\mathrm{dist}(x_{k+1}, \mathcal{X}^\star)^2] \lesssim \exp(-\lambda_1 k)\, \mathrm{dist}(x_1, \mathcal{X}^\star)^2.$$

*for step sizes* $\alpha_k \propto k^{-\beta}$ **for any** $\beta \in [0, 1]$

# Fast and Optimal Convergence

Theorem (Chadha, C. & Duchi 22)

*For Quadratic-growth interpolation problems,* aProx *with truncated models has error rates*

$$\mathbb{E}[\mathrm{dist}(x_{k+1}, \mathcal{X}^\star)^2] \lesssim \exp(-\lambda_1 k)\,\mathrm{dist}(x_1, \mathcal{X}^\star)^2.$$

*for step sizes* $\alpha_k \propto k^{-\beta}$ **for any** $\beta \in [0, 1]$

Theorem (Chadha, C. & Duchi 22)

*For Quadratic-growth interpolation problems,* aProx *with truncated models attains the minimax optimal convergence rate.*

Visit poster #605 tonight to learn more!

# Takeaways

Previously, we had known aProx

- is robust to stepsize choice
- has fast rates on interpolation problems with growth
- has linear minibatch speedup

**Now** we know aProx

- can be accelerated, and we know how to do it
- has fast rates on interpolation problems with growth **for a wide variety of step size schedules**
- is optimal for interpolation problems with growth