Adaptive Sampling Stochastic Sequential Quadratic Programming

Raghu Bollapragada

Joint work with Albert S. Berahas (Univ. of Michigan) and Baoyu Zhou (Chicago Booth)

The University of Texas at Austin

January 2023

12^{th} US - Mexico Workshop on Optimization and its Applications

In the honor of Steve Wright's 60th birthday

Huatulco, Mexico

Optimization Problem

$$\min_{x \in \mathbb{R}^n} f(x) = \mathbb{E}_\zeta[F(x, \zeta)]$$
$$\text{s.t.} \quad c(x) = 0$$

- **Assumptions:**
  - $F : \mathbb{R}^n \times \Omega \to \mathbb{R}, c : \mathbb{R}^n \to \mathbb{R}^m$ - smooth differentiable functions
  - Only equality constraints
  - Constraint qualifications hold
- **Applications:** Constrained machine learning, optimal power flow, portfolio optimization, PDE constrained optimization,...

Stochastic Gradient (SG)

$$\boxed{\min_{x\in\mathbb{R}^n} f(x) = \mathbb{E}_\zeta[F(x,\zeta)]}$$

- SG and its variants are quite popular

$$x_{k+1} = x_k - \alpha_k \bar{g}_k, \quad \bar{g}_k = \frac{1}{|S_k|}\sum_{\zeta_i \in S_k} \nabla F(x_k, \zeta_i)$$

where $\mathbb{E}_k[\bar{g}_k] = g_k = \nabla f(x_k)$

- Used extensively in machine learning
- Recently SG methods are developed for constrained stochastic optimization based on the SQP paradigm

[Berahas et al., 2021], [Na et al., 2022]

Stochastic SQP - Main Idea

$$\min_{x \in \mathbb{R}^n} f(x) = \mathbb{E}_\zeta[F(x, \zeta)]$$
$$\text{s.t.} \quad c(x) = 0$$

Stochastic SQP iterate update:

$$x_{k+1} = x_k + \bar{\alpha}_k \bar{d}_k$$

where $\bar{d}_k$ is an approximate solution of

$$\min_{d \in \mathbb{R}^n} f(x_k) + \bar{g}_k{}^T d + \frac{d}{2}^T H_k d \qquad \begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} \bar{d}_k \\ \bar{\delta}_k \end{bmatrix} \approx - \begin{bmatrix} \bar{g}_k + J_k^T y_k \\ c(x_k) \end{bmatrix}$$
$$s.t. \quad c(x_k) + J_k^T d = 0$$

$J_k = \nabla c(x_k)$; H_k assumed to be positive definite on $Null(\nabla c)$; y_k - Lagrangian multiplier.

Stochastic SQP - Details

- **Merit Function:** guides algorithm, $\tau > 0$ (merit parameter)

$$\phi(x, \tau) = \tau f(x) + \|c(x)\|_1$$

- **Model of Merit Function:**

$$l(x, \tau, g, d) = \tau(f(x) + g^T d) + \|c(x) + \nabla c(x) d\|_1$$

- Given $(\bar{g}_k, \bar{d}_k)$, update $\bar{\tau}_k$ to ensure reduction in the model

$$\begin{aligned} \Delta l(x_k, \bar{\tau}_k, \bar{g}_k, \bar{d}_k) &= l(x_k, \bar{\tau}_k, \bar{g}_k, 0) - l(x_k, \bar{\tau}_k, \bar{g}_k, \bar{d}_k) \\ &= -\bar{\tau}_k \bar{g}_k^T \bar{d}_k + \|c(x_k)\|_1 - \|c(x_k) + \nabla c(x_k) \bar{d}_k\|_1 \gg 0 \end{aligned}$$

- Choose $\bar{\alpha}_k$ sufficiently small based on Lipschitz constants that reduces an upper bound on the decrease in stochastic merit function

SQP - Details

Theorem (Berahas et al., 2021)

If $\{\bar{\tau}_k\}$ eventually remains fixed at sufficient small $\tau_{\min}$, the for large k

$$\bar{\alpha}_k = \mathcal{O}(1): \quad \mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\left(\|\nabla f(x_k) + \nabla c(x_k)^T y_k\|_2^2 + \|c(x_k)\|_2\right)\right] \leq \mathcal{O}(M)$$

$$\bar{\alpha}_k = \mathcal{O}\left(\frac{1}{k}\right): \quad \liminf_{k\to\infty} \mathbb{E}\left[\left(\|\nabla f(x_k) + \nabla c(x_k)^T y_k\|_2^2 + \|c(x_k)\|_2\right)\right] = 0$$

- Diminishing stepsizes, slow convergence, difficult to parallelize

SQP - Details

Theorem (Berahas et al., 2021)

If $\{\bar{\tau}_k\}$ *eventually remains fixed at sufficient small* $\tau_{\min}$*, the for large* k

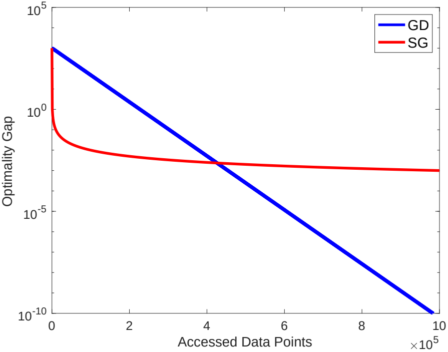
$$\bar{\alpha}_k = \mathcal{O}(1): \quad \mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\left(\|\nabla f(x_k) + \nabla c(x_k)^T y_k\|_2^2 + \|c(x_k)\|_2\right)\right] \leq \mathcal{O}(M)$$

$$\bar{\alpha}_k = \mathcal{O}\left(\frac{1}{k}\right): \quad \liminf_{k\to\infty} \mathbb{E}\left[\left(\|\nabla f(x_k) + \nabla c(x_k)^T y_k\|_2^2 + \|c(x_k)\|_2\right)\right] = 0$$

- Diminishing stepsizes, slow convergence, difficult to parallelize
- Is it possible to obtain results similar to deterministic case without diminishing stepsizes?

Unconstrained Settings

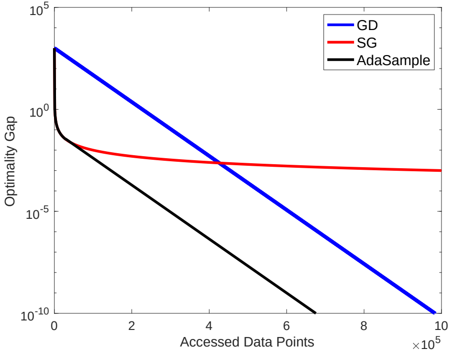
$$\min_{x \in \mathbb{R}^n} f(x) = \mathbb{E}_\zeta[F(x, \zeta)]$$



GD - Gradient Descent SG - Stochastic Gradient

Unconstrained Settings

$$\min_{x \in \mathbb{R}^n} f(x) = \mathbb{E}_\zeta[F(x,\zeta)]$$



GD - Gradient Descent SG - Stochastic Gradient

Adaptive Sampling Methods

$$\bar{g}_k = \frac{1}{|S_k|} \sum_{\zeta_i \in S_k} \nabla F(x_k, \zeta_i)$$

- Gradually increase sample sizes $|S_k|$
- To increase accuracy in gradient estimation
- Optimal theoretical sampling rates and practical adaptive sampling rules have been established in the literature

  [Friedlander & Schmidt 2012], [Pasupathy et al., 2015], [Byrd et al., 2015], [Bollapragada et al., 2019]
- Recently, adapted to projected gradient methods

  [Xie et al., 2021]
- Overcomes the limitations of SG methods

Goal: Develop an adaptive sampling method based on the SQP paradigm

AdaSQP Framework

Ada SQP

Input: x_0 (initial iterate); $\tau_{-1} > 0$ (initial penalty parameter)

for $k = 0, 1, 2, \ldots$ **do**
    Choose a set S_k consisting of random realizations of ζ
    Compute the stochastic gradient approximation $\bar{g}_k$
    Solve Newton-SQP system to compute $(\bar{d}_k, \bar{\delta}_k)$
    Update $\bar{\tau}_k > 0$ to ensure $\Delta l(x_k, \tau_k, \bar{g}_k \bar{d}_k) \gg 0$
    Compute stepsize $\bar{\alpha}_k$ based on estimates of Lipschitz constants
    Update $x_{k+1} \leftarrow x_k + \bar{\alpha}_k \bar{d}_k$; $y_{k+1} \leftarrow y_k + \bar{\alpha}_k \bar{\delta}_k$;
end for

Key Questions: How to choose sample size S_k, merit parameter τ_k, step size $\bar{\alpha}_k$?

Note: In this talk - linear systems are solved exactly

Sample Size Selection

- A popular test in unconstrained settings - *norm test*

$$\mathbb{E}_k[\|\bar{g}_k - \nabla f(x_k)\|^2] \leq \theta \|\nabla f(x_k)\|^2, \quad \theta > 0$$

- Control variance relative to the norm of the gradient
- Not readily applicable to constrained settings
- **Note:** $g_k = \nabla f(x_k) \not\to 0$ as we approach optimal solution
- Need a different optimality measure on the right-hand side

Sample Size Selection

- A popular test in unconstrained settings - *norm test*

$$\mathbb{E}_k[\|\bar{g}_k - \nabla f(x_k)\|^2] \leq \theta \|\nabla f(x_k)\|^2, \quad \theta > 0$$

- Control variance relative to the norm of the gradient
- Not readily applicable to constrained settings
- **Note:** $g_k = \nabla f(x_k) \not\to 0$ as we approach optimal solution
- Need a different optimality measure on the right-hand side
- **Observation:** Linear reduction in the model $\Delta l(x, \tau, g, d) \to 0$ as we approach optimal solution

$$\Delta l(x_k, \tau_k, g_k, d_k) = -\tau_k g_k^T d_k + \|c(x_k)\|_1$$

Modified Norm Test

- Modified Norm test for SQP settings:

$$\mathbb{E}_k \left[\|\bar{g}_k - \nabla f(x_k)\|_2^2\right] \leq \theta_1 \beta^\sigma \Delta l(x_k, \tau_k, g_k, d_k),$$

where $\theta_1 > 0$, $\beta \in (0, 1)$, $\sigma \in [2, 4]$.

- Boils down to *norm test* when there are no constraints where $g_k = -d_k$

$$\begin{aligned} \Delta l(x_k, \tau_k, g_k, d_k) &= -\tau_k g_k^T d_k + \|c(x_k)\|_1 \\ &= \|g_k\|^2 = \|\nabla f(x_k)\|^2 \end{aligned}$$

Practical Implementation

- Can be approximated as

$$\boxed{\frac{\mathbb{E}_k\left[\|\nabla F(x_k,\xi) - \nabla f(x_k)\|_2^2\right]}{|S_k|} \leq \theta_1 \beta^\sigma \Delta l(x_k, \tau_k, g_k, d_k)}$$

- Requires knowledge of unknown population quantities
- **Practical Settings:** Approximate population quantities with sample quantities

$$\mathbb{E}_k\left[\|\nabla F(x_k,\xi) - \nabla f(x_k)\|_2^2\right] \approx \frac{1}{|S_k|-1} \sum_{\zeta_i \in S_k} \|\nabla F(x_k, \xi_i) - \bar{g}_k\|^2$$

$$\Delta l(x_k, \tau_k, g_k, d_k) \approx \Delta l(x_k, \bar{\tau}_k, \bar{g}_k, \bar{d}_k)$$

Merit Parameter Update

- Ensures sufficient reduction in

$$\Delta l(x_k, \bar{\tau}_k, \bar{g}_k, \bar{d}_k) \geq \bar{\tau}_k \omega_1 \max\{\bar{d}_k{}^T H_k \bar{d}_k, \epsilon_d \|\bar{d}_k\|_2^2\} + \omega_1 \|c_k\|_1$$

$$\bar{\tau}_k^{trial} \leftarrow \begin{cases} \infty & \text{if } (\bar{g}_k{}^T \bar{d}_k + \max\{\bar{d}_k{}^T H_k \bar{d}_k, \epsilon_d \|\bar{d}_k\|_2^2\}) \leq 0 \\ \frac{(1-\omega_1)\|c_k\|_1}{(\bar{g}_k{}^T \bar{d}_k + \max\{\bar{d}_k{}^T H_k \bar{d}_k, \epsilon_d \|\bar{d}_k\|_2^2\}} & \text{otherwise,} \end{cases}$$

$$\bar{\tau}_k \leftarrow \begin{cases} \bar{\tau}_{k-1} & \text{if } \bar{\tau}_{k-1} \leq (1-\epsilon_\tau)\bar{\tau}_k^{trial} \\ (1-\epsilon_\tau)\bar{\tau}_k^{trial} & \text{otherwise,} \end{cases}$$

- Results in a *discontinuous* update formulae which involves multiple cases depending on

$$(\bar{g}_k{}^T \bar{d}_k + \max\{\bar{d}_k{}^T H_k \bar{d}_k, \epsilon_d \|\bar{d}_k\|_2^2\}) \leq \text{ or } > 0$$

Merit Parameter Update

- Need to analyze the difference between τ_k and $\bar{\tau}_k$ to establish strong non-asymptotic results
- Could differ a lot due to discontinuous update formula
- Need additional assumption that avoids difficult scenarios

Assumption

$$|(\bar{g}_k{}^T \bar{d}_k + \max\{\bar{d}_k{}^T H_k \bar{d}_k, \epsilon_d \|\bar{d}_k\|_2^2\}) - (g_k{}^T d_k + \max\{d_k{}^T H_k d_k, \epsilon_d \|d_k\|_2^2\})| \\ \leq \theta_3 \beta^{\sigma/2} |g_k{}^T d_k + \max\{d_k{}^T H_k d_k, \epsilon_d \|d_k\|_2^2\}|$$

- Trivially satisfied in the unconstrained settings ($\bar{d}_k = -\bar{g}_k$)
- Ensures $\bar{\tau}_k \geq \bar{\tau}_{\min}$

Merit Parameter Analysis

Theorem

Under the standard assumptions and if the previously mentioned assumption is satisfied then,

$$|(\bar{\tau}_k - \tau_k){g_k}^T d_k| \leq c\beta^{\sigma/2}\Delta l(x_k, \tau_k, g_k, d_k)$$

for some $c > 0$.

- Non-asymptotic merit parameter analysis
- Additional assumption is not necessary for asymptotic analysis

Stepsize selection

- Choose $\bar{\alpha}_k$ sufficiently small based on Lipschitz constants that reduces an upper bound on the decrease in stochastic merit function
- The formula for $\bar{\alpha}_k$ satisfies

$$\underline{\alpha}\beta \leq \bar{\alpha}_k \leq \alpha_u \beta^{(2-\sigma/2)}$$

- $\sigma \in [2, 4]$ balances gradient accuracy with stepsize

$$\boxed{\frac{\mathbb{E}_k\left[\|\nabla F(x_k,\xi) - \nabla f(x_k)\|_2^2\right]}{|S_k|} \leq \theta_1 \beta^\sigma \Delta l(x_k, \tau_k, g_k, d_k)}$$

- Higher σ results in higher gradient accuracy (or larger sample sizes) and potentially lead to the acceptance of large stepsizes (higher upper bound)

Convergence Results

Theorem

Under standard assumptions and if the previously mentioned conditions are satisfied with sufficiently small β*, then,*

$$\lim_{k\to\infty} \mathbb{E}\left[\Delta l(x_k, \tau_k, g_k, d_k)\right] = 0$$

Convergence Results

Theorem

Under standard assumptions and if the previously mentioned conditions are satisfied with sufficiently small β, then,

$$\lim_{k\to\infty} \mathbb{E}\left[\Delta l(x_k, \tau_k, g_k, d_k)\right] = 0$$

Consequently,

$$\lim_{k\to\infty} \mathbb{E}\left[\|d_k\|_2^2\right] = 0, \quad \lim_{k\to\infty} \mathbb{E}\left[\|c_k\|_2\right] = 0, \quad \lim_{k\to\infty} \mathbb{E}\left[\|g_k + J_k^T(y_k + \delta_k)\|_2\right] = 0.$$

- Results are similar to deterministic case

Iteration Complexity Results

- Consider iteration complexity to achieve

$$\mathbb{E}[\|g_k + J_k^T(y_k + \delta_k)\|_2] \leq \epsilon_L, \quad \text{and} \quad \mathbb{E}[\|c_k\|_1] \leq \epsilon_c$$

Theorem

Under the previously mentioned conditions, algorithm generates iterates $\{(x_k, y_k)\}$ that satisfies the above condition in at most

$$K_\epsilon = \mathcal{O}\left(\max\left\{\epsilon_L^{-2}, \epsilon_c^{-1}\right\}\right)$$

iterations. Moreover, if $\epsilon_L = \epsilon$ and $\epsilon_c = \epsilon^2$, then $K_\epsilon = \mathcal{O}(\epsilon^{-2})$.

- Matches with complexity results of the deterministic case
- Significant computational savings in number of linear system solves compared to Stochastic SQP ($\mathcal{O}(\epsilon^{-4})$)
- **Note:** More samples required as iterations progress
- Need to analyze sample complexity

Predetermined Sampling

- Instead of adaptively controlling the sample sizes - use predetermined sublinear sampling schemes
- That is, consider,

$$\frac{\mathbb{E}_k\left[\|\nabla F(x_k,\xi)-\nabla f(x_k)\|_2^2\right]}{|S_k|} \leq \frac{\theta_1 \beta^\sigma}{(k+1)^\nu}, \quad \nu > 1$$

- **Note:** Right hand side is monotonically decreasing
- Not efficient in practice but provides guidance on sampling complexity
- Similar theoretical convergence and iteration complexity results

Sample Complexity

$$\mathbb{E}[\|g_k + J_k^T(y_k + \delta_k)\|_2] \le \epsilon_L, \quad \text{and} \quad \mathbb{E}[\|c_k\|_1] \le \epsilon_c$$

Theorem

Under the previously mentioned conditions, algorithm generates iterates $\{(x_k, y_k)\}$ that satisfies the above condition in at most

Sample Complexity

$$\mathbb{E}[\|g_k + J_k^T(y_k + \delta_k)\|_2] \leq \epsilon_L, \quad \text{and} \quad \mathbb{E}[\|c_k\|_1] \leq \epsilon_c$$

Theorem

Under the previously mentioned conditions, algorithm generates iterates $\{(x_k, y_k)\}$ *that satisfies the above condition in at most*

$$W_\epsilon = \mathcal{O}\left(\left(\max\left\{\epsilon_L^{-2}, \epsilon_c^{-1}\right\}\right)^{(\nu+1)}\right), \quad \nu > 1$$

stochastic gradient evaluations.

Sample Complexity

$$\mathbb{E}[\|g_k + J_k^T(y_k + \delta_k)\|_2] \leq \epsilon_L, \quad \text{and} \quad \mathbb{E}[\|c_k\|_1] \leq \epsilon_c$$

Theorem

Under the previously mentioned conditions, algorithm generates iterates $\{(x_k, y_k)\}$ that satisfies the above condition in at most

$$W_\epsilon = \mathcal{O}\left(\left(\max\left\{\epsilon_L^{-2}, \epsilon_c^{-1}\right\}\right)^{(\nu+1)}\right), \quad \nu > 1$$

stochastic gradient evaluations. Moreover, if $\epsilon_L = \epsilon$ and $\epsilon_c = \epsilon^2$, then $W_\epsilon = \mathcal{O}(\epsilon^{-2(\nu+1)}) \approx \mathcal{O}(\epsilon^{-4})$.

- W_ϵ arbitrarily close to the typical expected work complexity for stochastic SQP

Inexact Linear Systems

- So far, the linear systems are solved exactly - Expensive

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} \bar{d}_k \\ \bar{\delta}_k \end{bmatrix} = - \begin{bmatrix} \bar{g}_k + J_k^T y_k \\ c(x_k) \end{bmatrix}$$

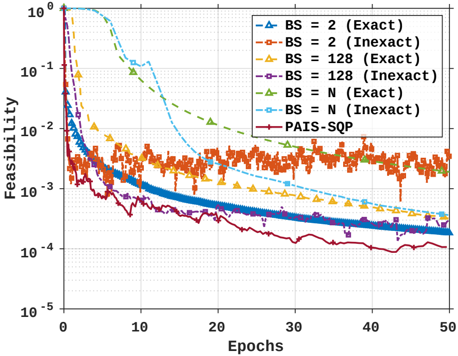
- Instead, solve the linear system inexactly to get $(\tilde{d}_k, \tilde{\delta}_k)$ that satisfies

$$\left\| \begin{bmatrix} \tilde{d}_k \\ \tilde{\delta}_k \end{bmatrix} - \begin{bmatrix} \bar{d}_k \\ \bar{\delta}_k \end{bmatrix} \right\|_2^2 \leq \theta_2 \beta^\sigma \Delta l(x_k, \tau_k, g_k, d_k),$$

  where $\theta_2 > 0$, $\beta \in (0, 1)$, $\sigma \in [2, 4]$.
- **Practical Settings:** Use residuals on the left hand side and approximate population quantities with sample quantities
- Similar theoretical convergence, iteration complexity and sample complexity results as in the exact case

Numerical Experiments - Constrained Logistic Regression

$$\min_{x \in \mathbb{R}n} \; f(x) = \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + e^{-y_i(X_i^T x)}\right) \quad \text{s.t} \quad Ax = b_1, \quad \|x\|_2^2 = b_2,$$

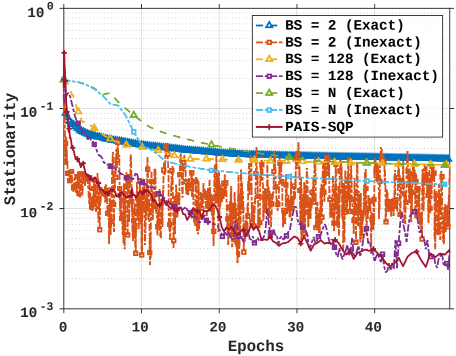


Australian Dataset

- **PAIS-SQP**: Our Method

Numerical Experiments - Constrained Logistic Regression

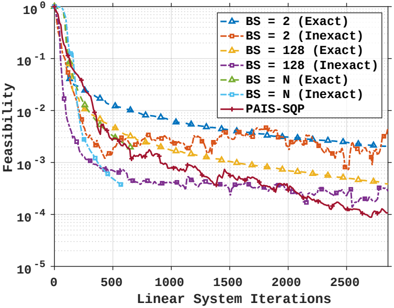
$$\min_{x \in \mathbb{R}n} \; f(x) = \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + e^{-y_i (X_i^T x)}\right) \quad \text{s.t} \quad Ax = b_1, \quad \|x\|_2^2 = b_2,$$



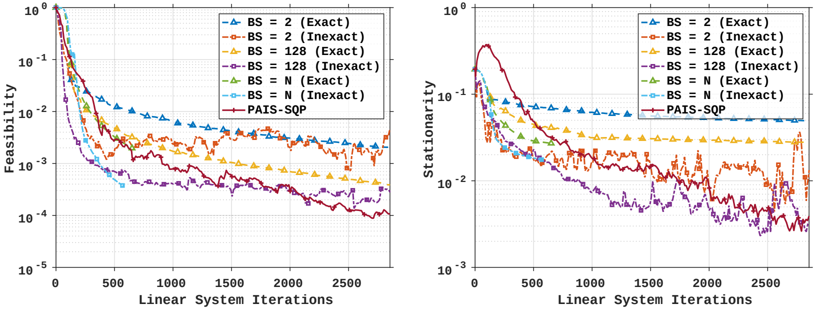
Australian Dataset

- **PAIS-SQP**: Our Method

Numerical Experiments - Constrained Logistic Regression

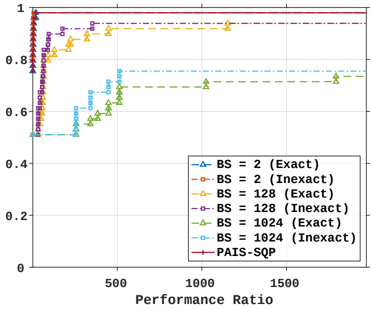


Numerical Experiments - Constrained Logistic Regression

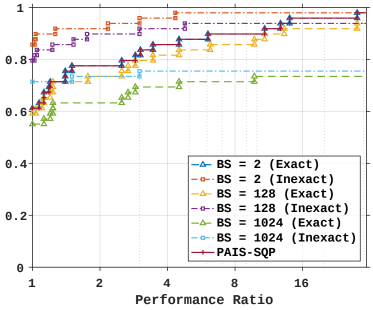


Australian Dataset

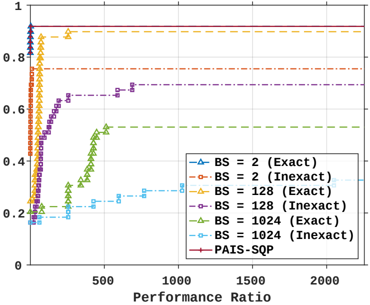
Numerical Experiments - CUTE problems



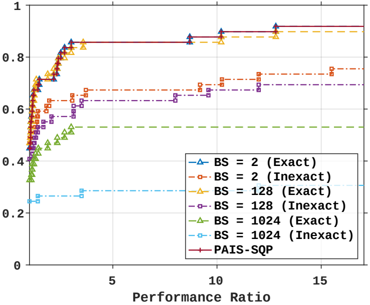
(a) Feasibility vs. Epochs



(b) Feasibility vs. Linear Sys. Iter.



(a) Stationarity vs. Epochs



(b) Stationarity. vs. Linear Sys. Iter.

Summary & Future Work

- Developed adaptive sampling framework for SQP Paradigm
- Non-asymptotic convergence and iteration complexity results - similar to deterministic SQP
- Sample complexity results - comparable to stochastic SQP
- Promising numerical results

Summary & Future Work

- Developed adaptive sampling framework for SQP Paradigm
- Non-asymptotic convergence and iteration complexity results - similar to deterministic SQP
- Sample complexity results - comparable to stochastic SQP
- Promising numerical results

Future Research Directions

- Quasi-Newton variants
- Subsampled Hessians
- Inequality constraints

https://arxiv.org/abs/2206.00712

Thank you

https://arxiv.org/abs/2206.00712

Questions?

https://arxiv.org/abs/2206.00712