Retrospective Approximation for Stochastic Constrained Problems Using Sequential Quadratic Programming

Shagun Gupta

Raghu Bollapragada and Albert S. Berahas

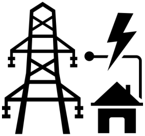
Constrained Stochastic Optimization

$$\min_{x \in \mathbb{R}^n} \; {\color{red} f(x)} = \mathbb{E}[{\color{blue} F(x, \xi)}] \qquad (f : \mathbb{R}^n \to \mathbb{R})$$

$$\text{s.t.} \quad c_E(x) = 0 \qquad (c_E : \mathbb{R}^n \to \mathbb{R}^{m_E})$$

$$c_I(x) \leq 0 \qquad (c_I : \mathbb{R}^n \to \mathbb{R}^{m_I})$$

where $F : \mathbb{R}^n \times \Omega \to \mathbb{R}$, ξ has a probability space $(\xi, \mathcal{F}, P)$, $\mathbb{E}[\cdot]$ with respect to P



(a) Optimal Power Flow

Adobe Stock

(b) Machine Learning

Tom Taulli, Forbes 2019

(c) Logistics

Adobe Stock

Empirical Risk Minimization

$$\min_{x\in\mathbb{R}^n} \quad f_S(x) = \frac{1}{|S|}\sum_{i\in\mathcal{S}} F(x,\xi_i)$$
$$\text{s.t.} \quad c(x) = 0$$

where $\mathcal{S} = \{\xi_1, \xi_2, \dots\}$ are a set of i.i.d observations.

Deterministic Methods [Nocedal and Wright 1999]

1. Penalty Methods
2. Projection Methods
3. Sequential Quadratic Programming
4. Interior Point Methods

Stochastic Optimization

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

where $\mathcal{S} = \{\xi_1, \xi_2, \ldots\}$ are a set of i.i.d observations.

Stochastic Methods

1. Penalty Methods

   Ravi et al. 2019; Nandwani, Pathak, and Singla 2019, ...

2. Projection Methods

   Lan 2020; Ghadimi, Lan, and Zhang 2016, ...

3. Sequential Quadratic Programming (SQP)

   Berahas, et al., 2021; Berahas, et al., 2022; Na, et al, 2023; O'Neill, Michael J., 2024; ...

4. Interior Point Methods

   Curtis, et al., 2023; Curtis, et al., 2024

Stochastic Optimization

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

where $\mathcal{S} = \{\xi_1, \xi_2, \dots\}$ are a set of i.i.d observations.

Stochastic Methods

1. Penalty Methods

   Ravi et al. 2019; Nandwani, Pathak, and Singla 2019, ...

2. Projection Methods

   Lan 2020; Ghadimi, Lan, and Zhang 2016, ...

3. **Sequential Quadratic Programming (SQP)**

   Berahas, et al., 2021; Berahas, et al., 2022; Na, et al, 2023; O'Neill, Michael J., 2024; ...

4. Interior Point Methods

   Curtis, et al., 2023; Curtis, et al., 2024

Key Takeaways

1. We introduce a Retrospective Approximation framework for efficient stochastic constrained optimization.

2. The framework leverages deterministic SQP solvers to solve stochastic problems.
   - Eliminates need to tune parameters for step size and merit parameter.

3. The framework achieves:
   3.1 Optimal Sample Gradient Complexity - $\mathcal{O}(\epsilon^{-4})$
   3.2 Optimal Linear System Solve Complexity - $\mathcal{O}(\epsilon^{-2})$

Constrained Optimization

$$\min_{x \in \mathbb{R}^n} \quad f(x)$$
$$\text{s.t.} \quad c(x) = 0$$

Let $L(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ denote the Lagrangian of the problem:

$$L(x, \lambda) = f(x) + c(x)^T \lambda$$

where $x \in \mathbb{R}^n$ is the primal and $\lambda \in \mathbb{R}^m$ is the dual variable
Necessary conditions at any local solution

$$\nabla_x L(x, \lambda) = \nabla f(x) + \nabla c(x)^T \lambda = 0$$
$$c(x) = 0$$

Sequential Quadratic Programming (SQP)

$$\min_{x \in \mathbb{R}^n} \quad f(x)$$
$$\text{s.t.} \quad c(x) = 0$$

Interpretation: @ x_k

$$\min_{d \in \mathbb{R}^n} \quad f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T H_k d$$
$$\text{s.t.} \quad c(x_k) + \nabla c(x_k)^T d = 0$$

Sequential Quadratic Programming (SQP)

$$\min_{x \in \mathbb{R}^n} \; f(x)$$
$$\text{s.t.} \; c(x) = 0$$

Interpretation: @ x_k

$$\min_{d \in \mathbb{R}^n} \; f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T H_k d$$
$$\text{s.t.} \; c(x_k) + \nabla c(x_k)^T d = 0$$

Step computation "Newton-SQP system":

- $\nabla c(x_k)^T$ has linearly independent rows **(LICQ)**
- H_k is positive definite over $Null\left(\nabla c(x_k)^T\right)$

$$\begin{bmatrix} H_k & \nabla c(x_k) \\ \nabla c(x_k)^T & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \lambda_k \end{bmatrix} = - \begin{bmatrix} \nabla f(x_k) \\ c(x_k) \end{bmatrix}$$

Sequential Quadratic Programming (SQP)

$$\min_{x \in \mathbb{R}^n} \quad f(x)$$
$$\text{s.t.} \quad c(x) = 0$$

Interpretation: @ x_k

$$\min_{d \in \mathbb{R}^n} \quad f(x_k) + \nabla f(x_k)^T d + \tfrac{1}{2} d^T H_k d$$
$$\text{s.t.} \quad c(x_k) + \nabla c(x_k)^T d = 0$$

Step computation "Newton-SQP system":

$$\begin{bmatrix} H_k & \nabla c(x_k) \\ \nabla c(x_k)^T & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \lambda_k \end{bmatrix} = - \begin{bmatrix} \nabla f(x_k) \\ c(x_k) \end{bmatrix}$$

Update merit parameter: $\tau_k > 0$ to ensure $\phi'(x_k, \tau_k, d_k) \ll 0$

$$\phi(x_k, \tau_k) = \tau_k f(x) + \|c(x)\|_1$$
$$\phi'(x_k, \tau_k, d) = \tau_k \nabla f(x)^T d - \|c(x)\|_1$$

Sequential Quadratic Programming (SQP)

$$\min_{x \in \mathbb{R}^n} \; f(x) \quad \text{s.t.} \; c(x) = 0$$

Interpretation: @ x_k

$$\min_{d \in \mathbb{R}^n} \; f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T H_k d \quad \text{s.t.} \; c(x_k) + \nabla c(x_k)^T d = 0$$

Step computation "Newton-SQP system":

$$\begin{bmatrix} H_k & \nabla c(x_k) \\ \nabla c(x_k)^T & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \lambda_k \end{bmatrix} = - \begin{bmatrix} \nabla f(x_k) \\ c(x_k) \end{bmatrix}$$

Update merit parameter: $\tau_k > 0$ to ensure $\phi'(x_k, \tau_k, d_k) \ll 0$

Update iterate (Line Search): $x_{k+1} \leftarrow x_k + \alpha_k d_k$

Stochastic SQP

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

Interpretation: @ x_k

$$\min_{\tilde{d} \in \mathbb{R}^n} \; \tilde{f}(x_k) + \nabla \tilde{f}(x_k)^T d + \tfrac{1}{2} \tilde{d}^T H_k \tilde{d}$$
$$\text{s.t.} \; c(x_k) + \nabla c(x_k)^T \tilde{d} = 0$$

Step computation "Newton-SQP system":

$$\begin{bmatrix} H_k & \nabla c(x_k) \\ \nabla c(x_k)^T & 0 \end{bmatrix} \begin{bmatrix} \tilde{d}_k \\ \tilde{\lambda}_k \end{bmatrix} = - \begin{bmatrix} \nabla \tilde{f}(x_k) \\ c(x_k) \end{bmatrix}$$

Update merit parameter: $\tilde{\tau}_k > 0$ to ensure $\phi'(x, \tilde{\tau}_k, \tilde{d}_k) \ll 0$
Update iterate: $x_{k+1} \leftarrow x_k + \tilde{\alpha}_k \tilde{d}_k$

Stochastic SQP

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

Require additional control over the stochastic parameters.

Theorem (Informal)

If $\{\tilde{\tau}_k\}$ remains eventually fixed at a sufficiently small τ_{min} and the step size $\tilde{\alpha}_k = \mathcal{O}(\beta_k)$

$$\beta_k = \beta = \mathcal{O}(1): \quad \mathbb{E}\left[\|\nabla L(x_k,\lambda_k)\|_2^2 + \|c(x_k)\|_2\right] \leq \mathcal{O}(M)$$
$$\beta_k = \mathcal{O}\left(\frac{1}{k}\right): \quad \liminf_{k\to\infty} \mathbb{E}\left[\|\nabla L(x_k,\lambda_k)\|_2^2 + \|c(x_k)\|_2\right] = 0$$

[Berahas et al. 2021]

Retrospective Approximation

Newton 2023; Chen and Schmeiser 2001; Jalilzadeh and Shanbhag 2016; Deng and Ferris 2009; Pasupathy 2010; Royset 2013, ...

1. Outer Iteration $k \in \mathbb{N}$
   Construct sub-sampled problem

$$\min_{x \in \mathbb{R}^n} \; f_{S_k}(x) = \frac{1}{|S_k|} \sum_{i \in \mathcal{S}_k} F(x, \xi_i)$$
$$\text{s.t.} \;\; c(x) = 0$$

where $\mathcal{S}_k \subset \mathcal{S}$

Retrospective Approximation

Newton 2023; Chen and Schmeiser 2001; Jalilzadeh and Shanbhag 2016; Deng and Ferris 2009; Pasupathy 2010; Royset 2013, ...

1. Outer Iteration $k \in \mathbb{N}$
   Construct sub-sampled problem

$$\min_{x \in \mathbb{R}^n} \; f_{S_k}(x) = \frac{1}{|S_k|} \sum_{i \in \mathcal{S}_k} F(x, \xi_i)$$
$$\text{s.t.} \; c(x) = 0$$

   where $\mathcal{S}_k \subset \mathcal{S}$

2. Inner Iterations
   Solve the sub-sampled problem to an accuracy
      - NOT to Optimality

Retrospective Approximation

Algorithm Retrospective Approximation Framework

Input : x_0, $\{S_k\}$, $\{\mathcal{T}_{S_k}\}$

for k = 1, 2, 3, ... **do**
    Obtain batch $S_k = \{\xi_1, \xi_2, ..., \xi_{|S_k|}\}$
    Construct sub-sampled problem
    Solve initialized at x_{k-1} until $\mathcal{T}_{S_k}$ is satisfied for x_k
end for

Retrospective Approximation

Algorithm Retrospective Approximation Framework

Input : x_0, $\{S_k\}$, $\{\mathcal{T}_{S_k}\}$

1: **for** k = 1, 2, 3, ... **do**
2: Obtain batch $S_k = \{\xi_1, \xi_2, ..., \xi_{|S_k|}\}$
3: Construct sub-sampled problem
4: Solve initialized at x_{k-1} until $\mathcal{T}_{S_k}$ is satisfied for x_k
5: **end for**

Questions to be addressed:

- Deterministic Solver
- Termination Criterion sequence $\{\mathcal{T}_{S_k}\}$
- Sample Batch sequence $\{S_k\}$

Deterministic Solver - Line Search SQP

$$\min_{x\in\mathbb{R}^n} \quad f_{S_k}(x) = \frac{1}{b_k}\sum_{i\in\mathcal{S}_k} F(x,\xi_i)$$

$$\text{s.t.} \quad c(x) = 0$$

Algorithm Line Search-SQP Nocedal and Wright 1999

Input: x_0, $\tau_{-1} > 0$

for $k = 0, 1, 2, \ldots$ **do**

Compute Newton step d

Update merit parameter: $\tau_k > 0$ to ensure $\phi'(x_k, \tau_k, d_k) \ll 0$

Line Search: find α_k such that $x_{k+1} \leftarrow x_k + \alpha_k d_k$ yields

$$\phi(x_k + \alpha_k d_k, \tau_k) \leq \phi(x_k, \tau_k) - \eta\alpha_k \Delta q(x_k, \tau_k, \nabla f(x), H_k, d)$$

Update iterate : $x_{k+1} \leftarrow x_k + \alpha_k d_k$

end for

Retrospective Approximation - SQP

- Iterates $\{x_{k,j}\}, \{\lambda_{k,j}\}$
- N_k iterations performed on sub-sampled problem S_k
- Lagrangian Hessian Approximations $\{H_{k,j}\}$
- Search direction $\{d_{k,j}\}$

Retrospective Approximation - SQP

Algorithm RA-SQP

Input : $x_{0,0}$, $\{S_k\}$, $\{\gamma_k\}$, $\{\epsilon_k\}$

1: **for** k = 1, 2, 3, ... **do**

2: Construct sub-sampled constraint problem $f_{S_k}(x)$

3: Solve $f_{S_k}(x)$ using **line search SQP**

- Initialized at $x_{k,0}$ and τ_{-1}
- Terminate at N_k if

$$\left\| \begin{matrix} \nabla L_{S_k}(x_{k,N_k}, \lambda_{k,N_k}) \\ c(x_{k,N_k}) \end{matrix} \right\| \leq \gamma_k \left\| \begin{matrix} \nabla L_{S_k}(x_{k,0}, \lambda_{k,0}) \\ c(x_{k,0}) \end{matrix} \right\| + \epsilon_k$$

4: $x_{k+1,0} = x_{k,N_k}$

5: **end for**

Termination Criterion

Theorem

The termination criterion for $\gamma < 1$

$$\left\| \begin{array}{c} \nabla L_S(x_j, \lambda_j) \\ c(x_j) \end{array} \right\| \leq \gamma \left\| \begin{array}{c} \nabla L_S(x_0, \lambda_0) \\ c(x_0) \end{array} \right\| + \epsilon$$

is satisfied when the following conditions are satisfied:

- *Step norm condition with sufficiently small* γ_d *and* ϵ_d.

$$\|d_j\| \leq \gamma_d \|d_0\| + \epsilon_d$$

- *Merit function model condition with sufficiently small* γ_q *and* ϵ_q.

$$\begin{aligned} &\Delta q(x_j, d_j, H_j, \nabla f_S(x_j), \tau_j) \\ &\qquad \leq \gamma_q \Delta q(x_0, d_0, H_0, \nabla f_S(x_0), \tau_0) + \epsilon_q \end{aligned}$$

Convergence Analysis

Assumption

We define the error metric $G_S = \max_{x \in \chi}\left[\frac{\|\nabla f(x) - \nabla f_S(x)\|_2}{u + \|\nabla f(x)\|_2}\right]$ *where* $u > 0$ *and as* $\{|S_k|\} \to \infty$, $E[G_{S_k}^2] \to 0$.

Convergence Analysis

Assumption

We define the error metric $G_S = \max_{x \in \chi} \left[\frac{\|\nabla f(x) - \nabla f_S(x)\|_2}{u + \|\nabla f(x)\|_2} \right]$ *where* $u > 0$ *and as* $\{|S_k|\} \to \infty$, $E[G_{S_k}^2] \to 0$.

Theorem (Convergence)

Under reasonable assumptions, when $\{|S_k|\} \to \infty$, *if*

- $E[\epsilon_k] \to 0$ *and* $\{\gamma_k\} < \gamma < 1$,
- $E[\|H_{k+1,0} - H_{k,N_k}\|_2^2] \to 0$

then,

$$\mathbb{E}\left[\left\| \begin{matrix} \nabla L(x_{k,N_k}, \lambda_{k,N_k}) \\ c(x_{k,N_k}) \end{matrix} \right\|_2 \right] \to 0.$$

Convergence Analysis

Theorem (Hessian Approximation)

The Hessian approximation condition, i.e., $E[\|H_{k+1,0} - H_{k,N_k}\|^2] \to 0$*, is satisfied when employing:*

1. *A constant Hessian, i.e.* $\{H_{k,j}\} = I_n$.

Convergence Analysis

Theorem (Hessian Approximation)

The Hessian approximation condition, i.e., $E[\|H_{k+1,0} - H_{k,N_k}\|^2] \to 0$, *is satisfied when employing:*

1. *A constant Hessian, i.e.* $\{H_{k,j}\} = I_n$.
2. *A deterministic BFGS or LBFGS approximation that is carried over to the next outer iteration, i.e.* $H_{k+1,0} = H_{k,N_k}$.

Convergence Analysis

Theorem (Hessian Approximation)

The Hessian approximation condition, i.e., $E[\|H_{k+1,0} - H_{k,N_k}\|^2] \to 0$, *is satisfied when employing:*

1. *A constant Hessian, i.e.* $\{H_{k,j}\} = I_n$.
2. *A deterministic BFGS or LBFGS approximation that is carried over to the next outer iteration, i.e.* $H_{k+1,0} = H_{k,N_k}$.
3. *A deterministic subsampled Hessian, i.e.*

$$H_{k,j} = \nabla^2 f_{S_k}(x_{k,j}) + \sum_{i=1}^{m} [\lambda_{k,j-1}]_i \nabla^2 c_i(x_{k,j}), \quad j > 0$$

$$H_{k+1,0} = \nabla^2 f_{S_{k+1}}(x_{k+1,0}) + \sum_{i=1}^{m} [\lambda_{k,N_k-1}]_i \nabla^2 c_i(x_{k+1,0})$$

under reasonable assumptions on the subsampled Hessian.

Complexity Analysis

Condition (Batch Size)

At each iteration k, S_k is selected such that,

$$\mathbb{E}[\|\nabla f_{S_{k+1}}(x_{k,N_k}) - \nabla f(x_{k,N_k})\|] \leq \theta \left\| \begin{matrix} \nabla L(x_{k,N_k}, \lambda_{k,N_k}) \\ c(x_{k,N_k}) \end{matrix} \right\| + a\beta^k.$$

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

- $a = 0$, norm condition for constrained optimization
- $\theta = 0$, geometric increase in batch size

Complexity Analysis

Condition (Batch Size)

At each iteration k, S_k is selected such that,

$$\mathbb{E}[\|\nabla f_{S_{k+1}}(x_{k,N_k}) - \nabla f(x_{k,N_k})\|] \leq \theta \left\| \begin{matrix} \nabla L(x_{k,N_k}, \lambda_{k,N_k}) \\ c(x_{k,N_k}) \end{matrix} \right\| + a\beta^k.$$

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

Assumption (CLT scaling)

The stochastic gradient error G_S satisfies CLT scaling, i.e.,

$$\mathbb{E}[G_S^2] \leq \frac{\sigma^2}{|S|} \quad , \textit{where } \sigma < \infty.$$

Thus if x and S are independent,

$$\mathbb{E}[\|\nabla f_S(x) - \nabla f(x)\|^2 | x] \leq \frac{M^2}{|S|} \quad , \textit{where } M < \infty.$$

Complexity Analysis

Theorem

Under the stated assumptions, if $\epsilon_k = \frac{\delta}{\sqrt{|S_k|}}$ *where* $0 < \delta < \infty$, $\gamma = 0$, $A = \delta + (u + \kappa_g)\sigma$ *and* $\theta < \frac{M}{A}$

$$\mathbb{E}\left[\left\| \begin{matrix} \nabla L(x_{k,N_k}, \lambda_{k,N_k}) \\ c(x_{k,N_k}) \end{matrix} \right\|\right] \leq \left(\max\left\{\beta, \frac{\theta A}{M}\right\}\right)^k \left[\frac{A}{\sqrt{|S_0|}} + \frac{2a\theta A^2}{M|\beta M - \theta A|}\right]$$

If $0 < a < \infty$*, the work complexity for:*

1. *The number of linear system solves is* $\mathcal{O}\left(\epsilon^{-2}\right)$
2. *The number of sample gradients is* $\mathcal{O}\left(\epsilon^{-4}\right)$

Complexity of Stochastic SQP methods

SQP Methods	Linear Solves	Sample Gradients
Deterministic	$\mathcal{O}(\epsilon^{-2})$	-
Stochastic	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-4})$
Adaptive Sampling	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2(v+1)}), v > 1$
Retrospective Approximation	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-4})$

Numerical Experiments

- Termination Criterion
  - $\epsilon_k = 0$ and $\gamma_k = \gamma$
  - Merit function model condition, $\gamma = 0.1$
  - Step size norm condition, $\gamma = 0.5$
- Batch Size Conditions
  - Geometric Batch Size

$$|S_k| = |S_0|\omega^k, \quad \omega > 1$$

  - Norm Test Condition

$$|S_{k+1}| \geq \frac{Var(\nabla F_{\tilde{S}_k}(x_{k,N_k}))}{\theta \Delta q(x_{k,N_k}, g_{\tilde{S}_k}(x_{k,N_k}), \tilde{d}_k, H_{k,N_k}, \tau_{trial}, c_{k,N_k})}$$

where $|S_k| = |\tilde{S}_k|$ and are independent samples.

Logistic Regression

$$\min_{x \in \mathbb{R}^n} F(x) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-v_i^T Q_i x})$$

$$\text{s.t.} \quad Ax = b_1,$$

$$\|x\|_2^2 = b_2,$$

with a9a dataset.

Logistic Regression

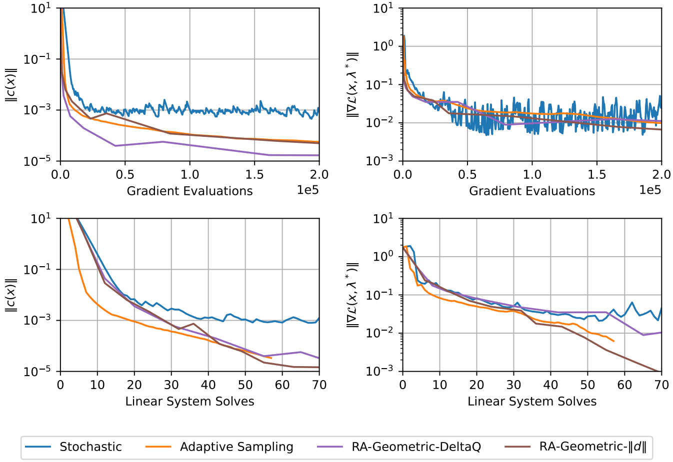


Figure: Comparison of SQP methods

Logistic Regression

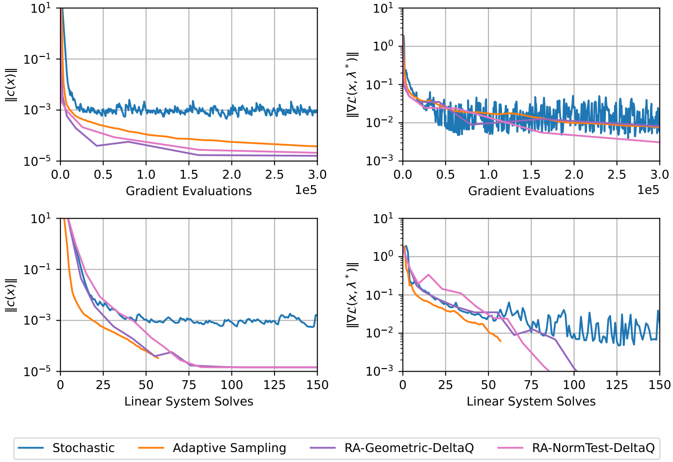


Figure: Norm Condition in RA

Logistic Regression

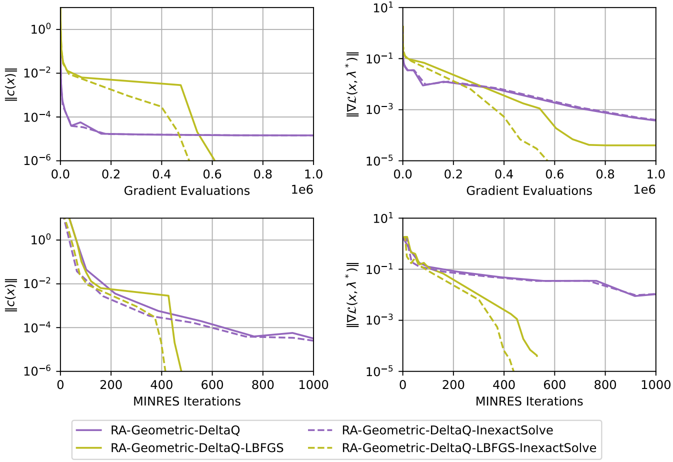


Figure: LBFGS and Inexact SQP system in RA

CUTEst

$$\min_{x \in \mathbb{R}^n} F(x) = f(x) + \xi \|x - x_0 - 1\|^2$$
$$\text{s.t.} \quad c(x) = 0,$$

where $\xi \sim U[-1, 1]$. [Gratton and Toint 2024]

- Creates adversarial noise in the objective function.
- Requires sharp increases in batch sizes.

CUTEst

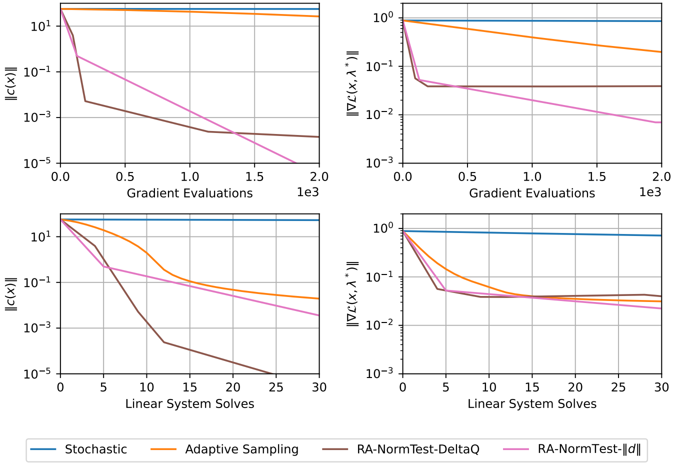


Figure: Comparison of SQP methods - BT6

Final Remarks

1. We introduce Retrospective Approximation Framework for efficient stochastic constrained optimization
2. The framework can achieve optimal complexity for linear system solves and sample gradients.
3. The framework allows utilizing advanced techniques from deterministic solvers to improve the performance of stochastic solvers.

Ongoing Work

1. Extending the framework to solve general non-linear inequality constraints.

Thank You!

Questions?

Backup Slides

Algorithm Line Search-SQP Nocedal and Wright 1999

Input: x_0, $\tau_{-1} > 0$

for $k = 0, 1, 2, \ldots$ **do**

  Compute Newton step d

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

$$\phi'(x, \tau, d) = \tau \nabla f(x)^T d - \|c(x)\|_1$$

  Update merit parameter: $\tau_k > 0$ to ensure $\phi'(x, \tau_k, d_k) \ll 0$

  Line Search: find α_k such that $x_{k+1} \leftarrow x_k + \alpha_k d_k$ yields

$$\phi(x_k + \alpha_k d_k, \tau_k) \leq \phi(x_k, \tau_k) + \frac{1}{2}\alpha_k \phi'(x, \tau_k, d_k)$$

end for