Adaptive, Fast, and Scalable Algorithms for Nonlinear Stochastic Optimization

Raghu Bollapragada

Operations Research and Industrial Engineering
The University of Texas at Austin

East Coast Optimization Meeting 2025
George Mason University
Arlington, Virginia

April 2025

College of Science
CENTER FOR MATHEMATICS AND ARTIFICIAL INTELLIGENCE
George Mason University

Sandia National Laboratories

NSF

siam
Society for Industrial and Applied Mathematics

Collaborators

Shagun Gupta
UT Austin

Thomas O'Leary-Roseberry
UT Austin

This Talk

1. **B** and Gupta (2025). *On the Convergence and Complexity of Proximal and Accelerated Proximal Gradient Methods under Adaptive Sampling Strategies. In preparation.*
2. O'Leary-Roseberry and **B** (2024). *Fast Unconstrained Optimization via Hessian Averaging and Adaptive Gradient Sampling Methods. arXiv preprint arXiv:2408.07268v1. Under review.*

Outline

Outline

Optimization Problem

$$\boxed{\min_{w \in \mathbb{R}^d} \; f(w)}$$

- $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, bounded below, potentially nonconvex

Expectation Problem

$$f(w) := \mathbb{E}_\zeta[F(w,\zeta)]$$

- $F : \mathbb{R}^d \times \Omega \to \mathbb{R}$, ζ has a probability space $(\Omega, \mathcal{F}, P)$.
- $\mathbb{E}_\zeta[\cdot]$ with respect to P

Finite-Sum Problem

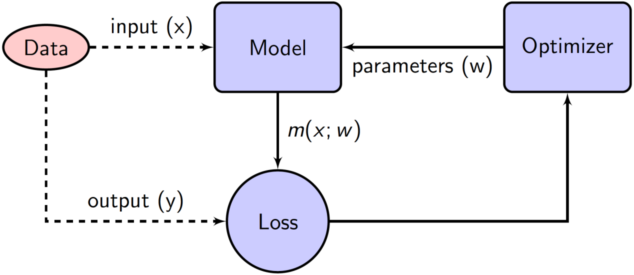
$$f(w) := \frac{1}{n} \sum_{i=1}^{n} F_i(w)$$

- $F_i : \mathbb{R}^d \to \mathbb{R}$, stochastic realizations of $F(w, \zeta_i)$.
- n: Number of samples

Key Challenges

- Expensive stochastic function evaluations or large n
- Severe nonlinearity (ill-conditioned problems)
- High dimensional settings (large d)

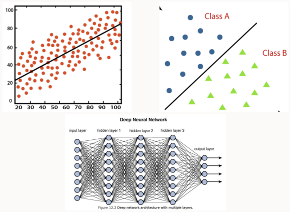
Applications: Supervised Machine Learning



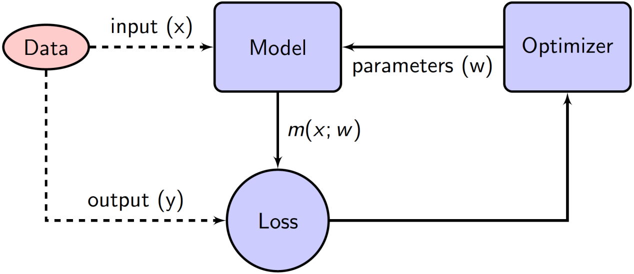
- Learn a parametric model $(m(x; w))$
  - e.g., linear models, neural networks
- Optimize using a loss function
  - e.g., squared loss, logistic loss

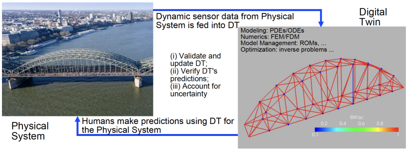
Empirical Risk: $$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} l(m(x_i; w), y_i)$$

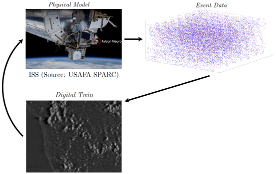
Expected Risk: $$\min_{w \in \mathbb{R}^d} \mathbb{E}_{(x,y)}[l(m(x; w), y)]$$



Applications: Digital Twins







Antil 2024

- Model calibration, stochastic inverse problems, neural operator training, surrogate modeling, control optimization, etc.

Deterministic Algorithms

Gradient Descent

$$w_{k+1} = w_k - \alpha_k \nabla f(w_k)$$

- Requires only gradient oracles
- Relies on small step size (α_k)
- Exhibits slow local linear convergence
- Sensitive to ill-conditioning

Newton's Method

$$w_{k+1} = w_k + \alpha_k p_k, \quad \nabla^2 f(w_k) p_k = -\nabla f(w_k)$$

- Requires gradient and Hessian oracles
- Allows large (often unit) step sizes (α_k)
- Achieves fast local quadratic convergence
- Robust to ill-conditioning

- ∇f and $\nabla^2 f$ are expensive or unavailable

$$\nabla f : \mathcal{O}(nd); \quad \nabla^2 f : \mathcal{O}(nd^2)$$

- Not suitable for stochastic or large-scale settings

Stochastic Gradient

$$w_{k+1} = w_k - \alpha_k \nabla F_{S_k^g}(w_k), \quad \nabla F_{S_k}^g(w_k) := \frac{1}{|S_k^g|} \sum_{i \in S_k^g} \nabla F_i(w_k)$$

Choose a subset $S_k^g \subset \{1, 2, \ldots\}$ of data at random.

- $|S_k^g|$ very small (128, 256)
- Low cost per iteration ($\mathcal{O}(d)$)
- Simple and easy to implement
- Widely used in machine learning

- α_k is heuristic, requires tuning
- Slower sublinear convergence
- Sensitive to ill-conditioning
- Hours of computing time

Stochastic Gradient

$$w_{k+1} = w_k - \alpha_k \nabla F_{S_k^g}(w_k), \quad \nabla F_{S_k}^g(w_k) := \frac{1}{|S_k^g|} \sum_{i \in S_k^g} \nabla F_i(w_k)$$

Choose a subset $S_k^g \subset \{1, 2, \ldots\}$ of data at random.

- $|S_k^g|$ very small (128, 256)
- Low cost per iteration ($\mathcal{O}(d)$)
- Simple and easy to implement
- Widely used in machine learning

- α_k is heuristic, requires tuning
- Slower sublinear convergence
- Sensitive to ill-conditioning
- Hours of computing time

Our Goal

- Design efficient optimization algorithms with fast convergence and low computational cost

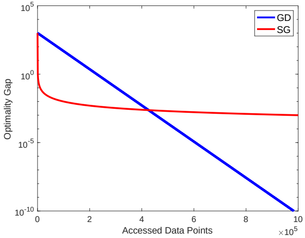
Outline

Adaptive Sampling - Motivation

$$w_{k+1} = w_k - \alpha_k \nabla F_{S_k^g}(w_k), \quad \nabla F_{S_k^g}(w_k) := \frac{1}{|S_k^g|} \sum_{i \in S_k^g} \nabla F_i(w_k)$$

Choose a subset $S_k^g \subset \{1, 2, \ldots\}$ of data at random.

- Gradually increase sample size $|S_k^g|$
- Improves accuracy of gradient estimation
- Inaccurate gradients suffice far from the solution
- Accuracy increases as iterates approach the solution
- Gradient computation can be parallelized



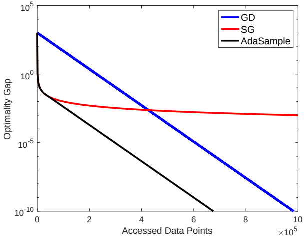
- How should $|S_k^g|$ be chosen for optimal theoretical and practical performance?

Adaptive Sampling - Motivation

$$w_{k+1} = w_k - \alpha_k \nabla F_{S_k^g}(w_k), \quad \nabla F_{S_k^g}(w_k) := \frac{1}{|S_k^g|} \sum_{i \in S_k^g} \nabla F_i(w_k)$$

Choose a subset $S_k^g \subset \{1, 2, \ldots\}$ of data at random.

- Gradually increase sample size $|S_k^g|$
- Improves accuracy of gradient estimation
- Inaccurate gradients suffice far from the solution
- Accuracy increases as iterates approach the solution
- Gradient computation can be parallelized



- How should $|S_k^g|$ be chosen for optimal theoretical and practical performance?

Adaptive Sampling Condition

For any $\theta_k > 0$, $\iota_k > 0$,

$$\mathbb{E}_{S_k^g}[\|\nabla F_{S_k^g}(w_k) - \nabla f(w_k)\|^2 | w_k] \leq \theta_k \|\nabla f(w_k)\|^2 + \iota_k$$

- Choose $|S_k^g|$ to satisfy *relaxed norm condition*
- Controls variance relative to the gradient norm
- Ensures quality in the search direction

Adaptive Sampling Condition

For any $\theta_k > 0$, $\iota_k > 0$,

$$\mathbb{E}_{S_k^g}[\|\nabla F_{S_k^g}(w_k) - \nabla f(w_k)\|^2 | w_k] \leq \theta_k \|\nabla f(w_k)\|^2 + \iota_k$$

- Choose $|S_k^g|$ to satisfy *relaxed norm condition*
- Controls variance relative to the gradient norm
- Ensures quality in the search direction

- If the population gradient variance is bounded, i.e., $\mathbb{E}[\|\nabla F(w, \zeta) - \nabla f(w)\|^2 | w] = \sigma^2 < \infty$, then relaxed norm condition is satisfied if

$$|S_k^g| \geq \frac{\sigma^2}{\theta_k \|\nabla f(w_k)\|^2 + \iota_k}$$

- In practice, estimate population quantities using samples

Theoretical Results

$$\mathbb{E}_{S_k^g}[\|\nabla F_{S_k^g}(w_k) - \nabla f(w_k)\|^2 | w_k] \leq \theta_k \|\nabla f(w_k)\|^2 + \iota_k$$

- f is bounded below, ∇f is Lipschitz continuous, and the gradient variance is bounded

Theorem [**B** & Gupta 2025]

Under stated assumptions, with f^* as optimal value and α_k sufficiently small, we have:

Setting	**Decay Condition**	**Global Convergence Rate**	**Type**	**Gradient Complexity**
Strongly Convex:	$\iota_k = \rho^k$	$\mathbb{E}[f(w_k) - f^*] = \mathcal{O}(\rho^k)$	Linear	$\mathcal{O}(\epsilon^{-1})$

- Retains global convergence rate of gradient descent
- Matches gradient complexity of stochastic gradient
- No fast local convergence

Theoretical Results

$$\mathbb{E}_{S_k^g}[\|\nabla F_{S_k^g}(w_k) - \nabla f(w_k)\|^2 | w_k] \leq \theta_k \|\nabla f(w_k)\|^2 + \iota_k$$

- f is bounded below, ∇f is Lipschitz continuous, and the gradient variance is bounded

Theorem [**B** & Gupta 2025]

Under stated assumptions, with f^* as optimal value and α_k sufficiently small, we have:

Setting	Decay Condition	Global Convergence Rate	Type	Gradient Complexity
Strongly Convex:	$\iota_k = \rho^k$	$\mathbb{E}[f(w_k) - f^*] = \mathcal{O}(\rho^k)$	Linear	$\mathcal{O}(\epsilon^{-1})$
General Convex:	$\sum \iota_k < \infty$	$\mathbb{E}[f(w_k) - f^*] = \mathcal{O}(\frac{1}{k})$	Sublinear	$\mathcal{O}(\epsilon^{-2})$
Nonconvex:	$\sum \iota_k < \infty$	$\min_{i=0,\ldots,k-1} \mathbb{E}[\|\nabla f(w_i)\|^2] = \mathcal{O}(\frac{1}{k})$	Sublinear	$\mathcal{O}(\epsilon^{-2})$

- Retains global convergence rate of gradient descent
- Matches gradient complexity of stochastic gradient
- No fast local convergence

Outline

Subsampled Newton Methods

$$w_{k+1} = w_k + \alpha_k p_k$$

$$\nabla^2 F_{S_k^h}(w_k) p_k = -\nabla f(w_k), \quad \nabla^2 F_{S_k^h}(w_k) := \frac{1}{|S_k^h|} \sum_{i \in S_k^h} \nabla^2 F_i(w_k)$$

Choose a subset $S_k^h \subset \{1, 2, \dots\}$ of data at random.

- Use $\nabla F_{S_k^g}(w_k)$ via adaptive sampling instead of exact $\nabla f(w_k)$
- $\nabla f(w_k)$ shown for simplicity in presentation and discussion
- Reduces the Hessian cost:

$$\nabla^2 f : \mathcal{O}(nd^2) \text{ vs. } \nabla^2 F_{S_k^h} : \mathcal{O}(|S_k^h| d^2)$$

- $\nabla^2 F_{S_k^h}$ may not retain positive definiteness of $\nabla^2 f$
- $|S_k^h|$ must grow for fast local convergence:

Expectation problem: $|S_k^h| \to \infty$

Finite-Sum Problem: $|S_k^h| \to n$

[**B**, Byrd & Nocedal 2019; Roosta-Khorasani & Mahoney 2019]

Subsampled Newton Methods

$$w_{k+1} = w_k + \alpha_k p_k$$

$$\nabla^2 F_{S_k^h}(w_k) p_k = -\nabla f(w_k), \quad \nabla^2 F_{S_k^h}(w_k) := \frac{1}{|S_k^h|} \sum_{i \in S_k^h} \nabla^2 F_i(w_k)$$

Choose a subset $S_k^h \subset \{1, 2, \dots\}$ of data at random.

- Use $\nabla F_{S_k^g}(w_k)$ via adaptive sampling instead of exact $\nabla f(w_k)$
- $\nabla f(w_k)$ shown for simplicity in presentation and discussion
- Reduces the Hessian cost:

  $$\nabla^2 f : \mathcal{O}(nd^2) \text{ vs. } \nabla^2 F_{S_k^h} : \mathcal{O}(|S_k^h| d^2)$$

- $\nabla^2 F_{S_k^h}$ may not retain positive definiteness of $\nabla^2 f$
- $|S_k^h|$ must grow for fast local convergence:

  Expectation problem: $|S_k^h| \to \infty$

  Finite-Sum Problem: $|S_k^h| \to n$

[**B**, Byrd & Nocedal 2019; Roosta-Khorasani & Mahoney 2019]

- Can we relax the requirements on $|S_k^h|$?

Hessian Averaging - Motivation

$$\widehat{H}_k := \frac{1}{k+1}\sum_{i=0}^{k} \nabla^2 F_{S_i^h}(w_i)$$

- Reduce the error in Hessian approximation via average of previous Hessians

$$\widehat{H}_k - \nabla^2 f(w_k) = \underbrace{\frac{1}{k+1}\sum_{i=0}^{k}\left(\nabla^2 F_{S_i^h}(w_i) - \nabla^2 F_{S_i^h}(w_k)\right)}_{\text{Hessian memory error}} + \underbrace{\frac{1}{k+1}\sum_{i=0}^{k}\nabla^2 F_{S_i^h}(w_k) - \nabla^2 f(w_k)}_{\text{sampling error}}$$

- sampling error goes to 0 as k increases
- Hessian memory error goes to zero as iterates converge $(w_i, w_k \to w^*)$
  - **Key Observation:** Convergence driven by gradients, not Hessians
- $|S_k^h|$ can be remained fixed

Cyclic Sampling

Focus on finite-sum problems for the rest of the talk

$$\underbrace{\frac{1}{k+1}\sum_{i=0}^{k} \nabla^2 F_{S_i^h}(w_k) - \nabla^2 f(w_k)}_{\text{sampling error}}$$

- $S_k^h \subset \{1, 2, \ldots, n\}$ of data drawn at random in Hessian averaging [Na, Dereziňski & Mahoney 2023]
- Need $k \to \infty$ for sampling error$\to 0$
- Instead, use cyclic sampling: $|S_i^h| = m$, $n = pm$

$$\underbrace{1, \ldots, m}_{S_0^h}, \underbrace{m+1, \ldots, 2m}_{S_1^h}, \ldots, \underbrace{\ldots, n}_{S_{p-1}^h}$$

- sampling error$= 0$ after each full cycle
- Yields better convergence

Algorithm

$$\widetilde{H}_k = \begin{cases} |\widehat{H}_k| & \text{if } \lambda_{\min}(|\widehat{H}_k|) \geq \tilde{\mu} \\ |\widehat{H}_k| + \left(\tilde{\mu} - \lambda_{\min}(|\widehat{H}_k|)\right) I & \text{otherwise,} \end{cases}$$

- $\widehat{H}_k$ may not be positive-definite
- Earlier works skipped the update – expensive

  [Na, Dereziňski & Mahoney 2023]
- Modified $\widetilde{H}_k \succeq \tilde{\mu} I$

Algorithm

$$\widetilde{H}_k = \begin{cases} |\widehat{H}_k| & \text{if } \lambda_{\min}(|\widehat{H}_k|) \geq \tilde{\mu} \\ |\widehat{H}_k| + \left(\tilde{\mu} - \lambda_{\min}(|\widehat{H}_k|)\right) I & \text{otherwise,} \end{cases}$$

- $\widehat{H}_k$ may not be positive-definite
- Earlier works skipped the update – expensive

  [Na, Dereziňski & Mahoney 2023]
- Modified $\widetilde{H}_k \succeq \tilde{\mu} I$

Hessian Averaging Algorithm

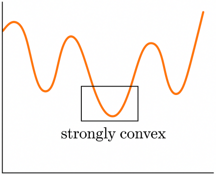
Input: $x_0 \in \mathbb{R}^d$; Hessian sample size m; $\{\alpha_k\} > 0$

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: Compute $\nabla f(w_k)$
3: Choose S_k^h samples in a cyclic manner with $|S_k^h| = m$
4: Compute sample Hessian: $\nabla^2 F_{S_k^h}(w_k)$
5: Compute average Hessian $\widehat{H}_k = \frac{1}{k+1}\left(k\widehat{H}_{k-1} + \nabla^2 F_{S_k^h}(w_k)\right)$
6: Modify $\widehat{H}_k$ to get $\widetilde{H}_k$
7: Solve the linear system: $\widetilde{H}_k p_k = -\nabla f(w_k)$
8: Update: $w_{k+1} = w_k + \alpha_k p_k$
9: **end for**

- $\mathcal{O}(d^2)$ memory required
- Step sizes via line search

Global Convergence

- f is bounded below, **globally nonconvex**, and **locally strongly convex**
- $\nabla^2 F_{S_k^h}$ are bounded above and are Lipschitz continuous



Theorem [O'Leary-Roseberry & **B** 2024]

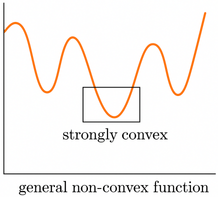
Under stated assumptions, with f^* as optimal value and α_k sufficiently small, we have:

Setting	Convergence Rate	Type
Global Nonconvex:	$\min_{i=0,\ldots,k-1} \mathbb{E}[\|\nabla f(w_i)\|^2] = \mathcal{O}(\frac{1}{k})$	Sublinear
Local Strongly Convex:	$\mathbb{E}[f(w_k) - f^*] = \mathcal{O}(\rho^k)$	Linear

- Retains global convergence rate of gradient descent

Global Convergence

- f is bounded below, **globally nonconvex**, and **locally strongly convex**
- $\nabla^2 F_{S_k^h}$ are bounded above and are Lipschitz continuous



Lemma [O'Leary-Roseberry & **B** 2024]

After $k \geq \frac{cn}{m}$ iterations, the following holds under stated assumptions:

$$\widetilde{H}_k = \widehat{H}_k$$

- Retains positive-definiteness of Hessian

Local Superlinear Convergence

Theorem [O'Leary-Roseberry & **B** 2024]

Under the stated assumptions, with w^* being a local minimizer and $\alpha_k = 1$ for all $k \geq k_{sup}$, we have:

$$\frac{\|w_{k+1} - w^*\|}{\|w_k - w^*\|} = \mathcal{O}\left(\frac{1}{k}\right) \quad \forall\, k \geq k_{sup} = c \max\left\{\frac{n}{m}, \kappa^2\right\}$$

κ : **Condition Number**

- **Deterministic** Q-superlinear rate: $\mathcal{O}\left(\frac{1}{k}\right)$
- Unit step size is naturally accepted via line search
- k_{sup} — Iteration index marking **transition** to local superlinear rate

Comparison of Superlinear Results

κ : **Condition Number**; k_{sup} : **Transition Iterations**; $|S_k^h| = m$

Sampling	$\lvert S_k^h \rvert$	Assumptions $\nabla^2 F_{S_k^h}$	Rate	Type	k_{sup}	Cite
Random	$\uparrow$	Strongly Convex	-	Expectation	-	[1]
Random	$\uparrow$	-	-	Probability	-	[2]

[1]: **B**, Byrd & Nocedal 2019 [2]: Roosta-Khorasani & Mahoney 2019

Comparison of Superlinear Results

κ : **Condition Number**; k_{sup} : **Transition Iterations**; $|S_k^h| = m$

Sampling	$\lvert S_k^h \rvert$	Assumptions $\nabla^2 F_{S_k^h}$	Rate	Type	k_{sup}	Cite
Random	$\uparrow$	Strongly Convex	-	Expectation	-	[1]
Random	$\uparrow$	-	-	Probability	-	[2]
Random	m	Subexp. Errors	$\mathcal{O}\left(\sqrt{\frac{\log(k)}{k}}\right)$	Probability	$\mathcal{O}\left(\kappa^6\right)$	[3]

[1]: **B**, Byrd & Nocedal 2019

[2]: Roosta-Khorasani & Mahoney 2019

[3]: Na, Dereziňski & Mahoney 2023

[3] has improved transition iterations $k_{sup} = \mathcal{O}(\kappa^2)$ for nonuniform averaging

Comparison of Superlinear Results

κ : **Condition Number**; k_{sup} : **Transition Iterations**; $|S_k^h| = m$

Sampling	$\lvert S_k^h \rvert$	Assumptions $\nabla^2 F_{S_k^h}$	Rate	Type	k_{sup}	Cite
Random	$\uparrow$	Strongly Convex	-	Expectation	-	[1]
Random	$\uparrow$	-	-	Probability	-	[2]
Random	m	Subexp. Errors	$\mathcal{O}\left(\sqrt{\frac{\log(k)}{k}}\right)$	Probability	$\mathcal{O}\left(\kappa^6\right)$	[3]
Cyclic	m	Lipschitz	$\mathcal{O}\left(\frac{1}{k}\right)$	**Deterministic**	$\mathcal{O}\left(\max\{\frac{n}{m}, \kappa^2\}\right)$	[4]

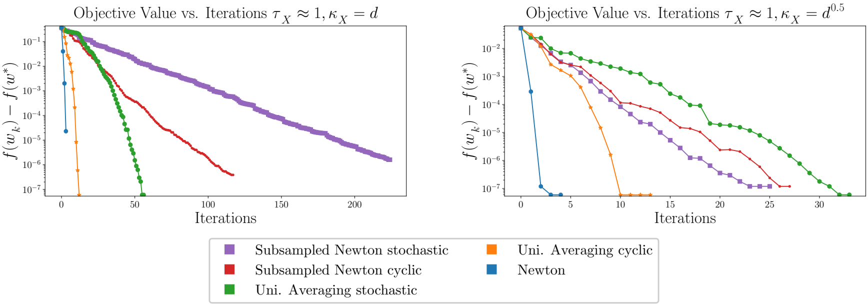
[1]: **B**, Byrd & Nocedal 2019
[2]: Roosta-Khorasani & Mahoney 2019
[3]: Na, Dereziňski & Mahoney 2023
[4]: **This Work**

- Improved superlinear rate deterministically with better or comparable transition iterations

[3] has improved transition iterations $k_{sup} = \mathcal{O}(\kappa^2)$ for nonuniform averaging

Hessian Averaging and Exact Gradients

$$f(w) = \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i(w^T x_i))) + \frac{1}{2n}\|w\|^2$$

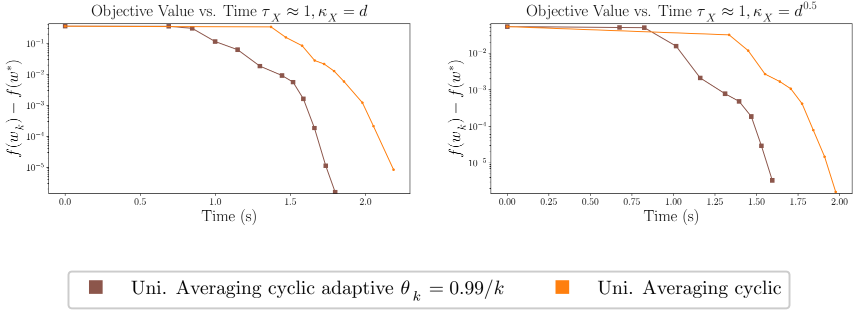


$n = 1000$; $d = 100$; κ_X : condition number

- All methods except Newton's method have same per iteration cost

Additional Experiments

Hessian Averaging and Adaptive Gradient Sampling



$n = 1000$; $d = 100$; κ_X : condition number

Remarks

Outline

Large-Scale Problems

- When d is very large, forming $(\mathcal{O}(d^2))$ and inverting Hessian $(\mathcal{O}(d^3))$ is infeasible
- Need efficient and scalable Hessian approximations
- **Idea:** Use diagonal approximations of Hessians
  [Yao et al., 2021]
- Efficient diagonal estimate via Hutchinson's randomized estimator:

$$D_k = \mathrm{diag}(\nabla^2 F_{S_k^h}(w_k)) \approx \mathbb{E}_z \left[z \nabla^2 F_{S_k^h}(w_k) z^T \right]$$

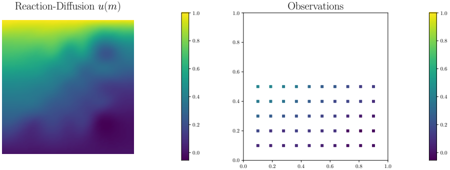
- Diagonal averaged Hessian:

$$\widehat{D}_k = \frac{1}{k+1} \sum_{i=0}^{k} D_i$$

- Note: Different averaging scheme than AdaHessian

Diagonally Averaged Newton (DAN): $w_{k+1} = w_k - \alpha_k \widetilde{D}^{-1} \nabla F_{S_k}^{g}(w_k)$

Numerical Results: Neural Operator Training



- Derivative informed Neural operator (DINO) for reaction diffusion PDE problem

$$\min_{w} \mathbb{E}_\pi \left[\|y - y_w\|^2_{\mathcal{Y}} + \|D_x y - D_x y_w\|^2_{HS(\mathcal{X},\mathcal{Y})} \right]$$

	SGD	ADAM	AdaHessian	DAN
y_r rel error $\searrow$	0.042	0.008	0.010	**0.007**
$\nabla_{x_r} y_r$ rel error $\searrow$	0.310	**0.255**	0.258	0.256

$n = 4500$; $d = 742,050$;

Outline

Final Remarks

$$\boxed{\min_{w \in \mathbb{R}^d} \quad f(w) = \mathbb{E}_\zeta[F(w,\zeta)] \quad \text{or} \quad \min_{w \in \mathbb{R}^d} \quad f(w) = \frac{1}{n}\sum_{i=1}^{n} F_i(w)}$$

- Proposed efficient adaptive gradient sampling with Hessian averaging
- Adaptive sampling: Retains gradient descent behavior with optimal stochastic gradient complexity
- Cyclic Hessian averaging: Enables fast local superlinear convergence — $(\mathcal{O}(\frac{1}{k}))$ for finite-sum problems
- Introduced scalable diagonal variants for high-dimensional problems
- Demonstrated efficient and robust performance in practice

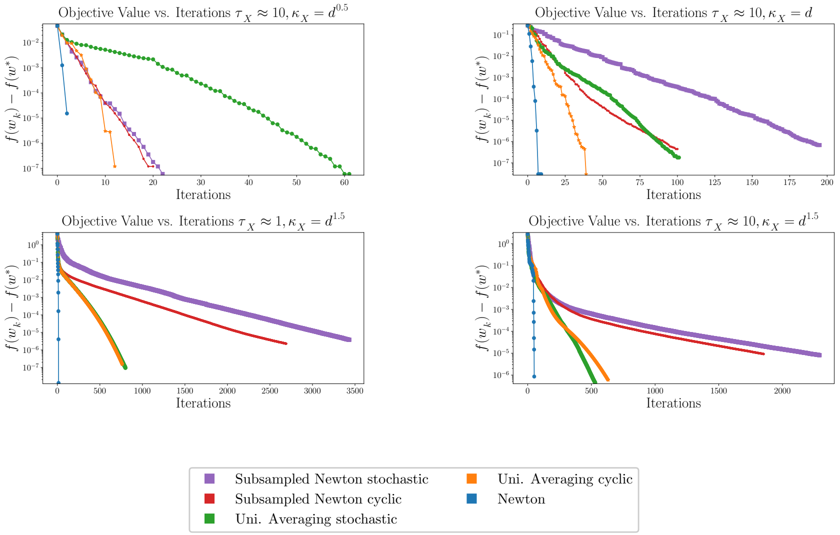
Extensions

- Several interesting research questions
- Inexact functions, simple constraints (proximal Newton methods), general nonlinear constraints (SQP methods), distributed settings, $\cdots$

Thank You!

Questions?

Hessian Averaging - Logistic Regression



Main Presentation

Reaction Diffusion Problem

$$
\begin{aligned}
-\nabla \cdot (e^m \nabla u) + u^3 &= s \text{ in } \Omega \\
u &= 1 \text{ on } \Gamma_{\text{top}} \\
e^m \nabla u \cdot n &= 0 \text{ on } \Gamma_{\text{sides}} \\
u &= 0 \text{ on } \Gamma_{\text{bottom}}
\end{aligned}
$$

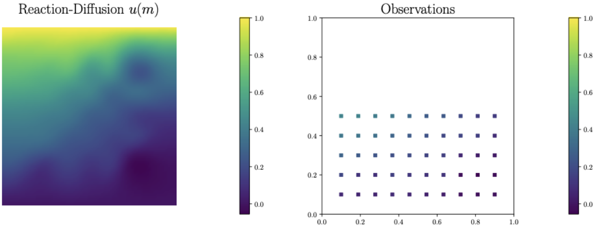


Figure 1: An instance of reaction-diffusion state and observables

- lognormal diffusion coefficient field $m \sim \mathcal{N}(0, C)$, s, is a sum of 25 smoothed point sources