

# A Stochastic Newton-type Method for Non-smooth Optimization

Titus Pinta

July 1, 2025  
EUROPT 2025, Southampton

# Problem Formulation

Given  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , **non convex, non smooth**

# Problem Formulation

Given  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , **non convex, non smooth** with **generalized 2nd order information** (i.e. semi-smooth gradient)

# Problem Formulation

Given  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , **non convex, non smooth** with **generalized 2nd order information** (i.e. semi-smooth gradient)

## Problem

Find

$$\operatorname{argmin}_{x \in \mathbb{R}^n} f(x) \quad (\mathcal{P})$$

# Problem Formulation

Given  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , **non convex, non smooth** with **generalized 2nd order information** (i.e. semi-smooth gradient)

## Problem

Find

$$\operatorname{argmin}_{x \in \mathbb{R}^n} f(x) \quad (\mathcal{P})$$

The tricky part

# Problem Formulation

Given  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , **non convex**, **non smooth** with **generalized 2nd order information** (i.e. semi-smooth gradient)

## Problem

Find

$$\operatorname{argmin}_{x \in \mathbb{R}^n} f(x) \quad (\mathcal{P})$$

The tricky part:  $\mathcal{H}f$  (2nd order information) only available from a **BAD stochastic oracle**  $\mathbb{H}$

# Problem Formulation

Given  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , **non convex, non smooth** with **generalized 2nd order information** (i.e. semi-smooth gradient)

## Problem

Find

$$\operatorname{argmin}_{x \in \mathbb{R}^n} f(x) \quad (\mathcal{P})$$

The tricky part:  $\mathcal{H}f$  (2nd order information) only available from a **BAD stochastic oracle**  $\mathbb{H}$

► biased:  $\mathbb{E}(\mathbb{H}) \neq \mathcal{H}f$

# Problem Formulation

Given  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , **non convex, non smooth** with **generalized 2nd order information** (i.e. semi-smooth gradient)

## Problem

Find

$$\operatorname{argmin}_{x \in \mathbb{R}^n} f(x) \quad (\mathcal{P})$$

The tricky part:  $\mathcal{H}f$  (2nd order information) only available from a **BAD stochastic oracle**  $\mathbb{H}$

- ▶ biased:  $\mathbb{E}(\mathbb{H}) \neq \mathcal{H}f$
- ▶ heavy tails

# Problem Formulation

Given  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , **non convex, non smooth** with **generalized 2nd order information** (i.e. semi-smooth gradient)

## Problem

Find

$$\operatorname{argmin}_{x \in \mathbb{R}^n} f(x) \quad (\mathcal{P})$$

The tricky part:  $\mathcal{H}f$  (2nd order information) only available from a **BAD stochastic oracle**  $\mathbb{H}$

- ▶ biased:  $\mathbb{E}(\mathbb{H}) \neq \mathcal{H}f$
- ▶ heavy tails
- ▶ unbounded moments:  $\mathbb{E}((\mathbb{H})^2) = \infty$

# Outline

Stochastic Analysis

Error Bound Theory

Regularity

Algorithm

Examples

Noisy Newton

Sketching

XFEL Imaging

# Probability Estimates

$X$  a random variable,

Theorem (Markov)

$$X \geq 0 \implies \mathbb{P}(X \geq \alpha) \leq \frac{\mathbb{E}(X)}{\alpha}$$

# Probability Estimates

$X$  a random variable,

Theorem (Markov)

$$X \geq 0 \implies \mathbb{P}(X \geq \alpha) \leq \frac{\mathbb{E}(X)}{\alpha}$$

Theorem (Cernoff)

$\{T^k\}_{k \in \mathbb{N}}$  i.i.d. Bernoulli,  $\mathbb{P}(T^k = 1) \geq (1 - \delta)$  Then  $\gamma \in (0, 1)$ ,  
 $\forall n \in \mathbb{N}$

$$\mathbb{P} \left( \sum_{k=0}^{n-1} T^k \leq (1 - \gamma) \mathbb{E} \left( \sum_{k=0}^{n-1} T^k \right) \right) \leq e^{-n(1-\delta) \frac{\gamma^2}{2}}$$

# Stochastic Processes

$\{T^k\}_{k \in \mathbb{N}}$  a stochastic process

# Stochastic Processes

$\{T^k\}_{k \in \mathbb{N}}$  a stochastic process,  $S^n = \sum_{k=0}^{n-1} T^k$

# Stochastic Processes

$\{T^k\}_{k \in \mathbb{N}}$  a stochastic process,  $S^n = \sum_{k=0}^{n-1} T^k$

## Definition (Stopping Times)

$K$  is a **stopping time** if  $I\{K > k\}$  is independent from any  $T^n$  with  $n \geq k$

# Stochastic Processes

$\{T^k\}_{k \in \mathbb{N}}$  a stochastic process,  $S^n = \sum_{k=0}^{n-1} T^k$

## Definition (Stopping Times)

$K$  is a **stopping time** if  $I\{K > k\}$  is independent from any  $T^n$  with  $n \geq k$

Example:  $K = \min\{n \in \mathbb{N} \mid T^n \geq \alpha\}$

# Stochastic Processes

$\{T^k\}_{k \in \mathbb{N}}$  a stochastic process,  $S^n = \sum_{k=0}^{n-1} T^k$

## Definition (Stopping Times)

$K$  is a **stopping time** if  $I\{K > k\}$  is independent from any  $T^n$  with  $n \geq k$

Example:  $K = \min\{n \in \mathbb{N} \mid T^n \geq \alpha\}$

## Theorem (Expectation)

If  $\mathbb{E}(T^k) \geq \alpha \forall k$ ,  $K$  a **stopping time** for  $\{S^k\}_{k \in \mathbb{N}}$  and  $\mathbb{E}(K) < \infty$ , then

$$\mathbb{E}(S^K) \geq \mathbb{E}(K)\alpha.$$

## Theorem (Young<sup>1</sup>)

$\{T^k\}_{k \in \mathbb{N}}$  i.i.d. Bernoulli,  $S^n = \sum_{k=0}^{n-1} T^k$ ,  $\mathbb{P}(T^k = 1) \geq (1 - \delta)$

---

<sup>1</sup>N. Young *Chernoff-type bounds for a stopped sum of independent random variables*, unpublished

## Theorem (Young<sup>1</sup>)

$\{T^k\}_{k \in \mathbb{N}}$  i.i.d. Bernoulli,  $S^n = \sum_{k=0}^{n-1} T^k$ ,  $\mathbb{P}(T^k = 1) \geq (1 - \delta)$

Then  $\gamma \in (0, 1)$ ,  $K$  a stopping time with finite expectation

$$\mathbb{P}((1 - \gamma)K \geq 2S^K + \gamma(1 - \gamma^2)\mathbb{E}(K)) \leq e^{-\gamma^2\mathbb{E}(K)}$$

---

<sup>1</sup>N. Young *Chernoff-type bounds for a stopped sum of independent random variables*, unpublished

## Theorem (Young<sup>1</sup>)

$\{T^k\}_{k \in \mathbb{N}}$  i.i.d. Bernoulli,  $S^n = \sum_{k=0}^{n-1} T^k$ ,  $\mathbb{P}(T^k = 1) \geq (1 - \delta)$

Then  $\gamma \in (0, 1)$ ,  $K$  a stopping time with finite expectation

$$\mathbb{P}((1 - \gamma)K \geq 2S^K + \gamma(1 - \gamma^2)\mathbb{E}(K)) \leq e^{-\gamma^2\mathbb{E}(K)}$$

## Proof (sketch).

Define

$$\Phi^N = (1 + \gamma)^{\sum_{k=0}^{N-1} (1 - T^k)} (1 - \gamma)^{\sum_{k=0}^{N-1} T^k} e^{-\gamma^2\mathbb{E}(K)}$$

---

<sup>1</sup>N. Young *Chernoff-type bounds for a stopped sum of independent random variables*, unpublished

## Theorem (Young<sup>1</sup>)

$\{T^k\}_{k \in \mathbb{N}}$  i.i.d. Bernoulli,  $S^n = \sum_{k=0}^{n-1} T^k$ ,  $\mathbb{P}(T^k = 1) \geq (1 - \delta)$

Then  $\gamma \in (0, 1)$ ,  $K$  a stopping time with finite expectation

$$\mathbb{P}((1 - \gamma)K \geq 2S^K + \gamma(1 - \gamma^2)\mathbb{E}(K)) \leq e^{-\gamma^2\mathbb{E}(K)}$$

## Proof (sketch).

Define

$$\Phi^N = (1 + \gamma)^{\sum_{k=0}^{N-1} (1 - T^k)} (1 - \gamma)^{\sum_{k=0}^{N-1} T^k} e^{-\gamma^2\mathbb{E}(K)}$$

Turns out  $\Phi^N$  is a martingale

$$\mathbb{E}(\Phi^K) \leq \mathbb{E}(\Phi^0)$$

---

<sup>1</sup>N. Young *Chernoff-type bounds for a stopped sum of independent random variables*, unpublished

## Theorem (Young<sup>1</sup>)

$\{T^k\}_{k \in \mathbb{N}}$  i.i.d. Bernoulli,  $S^n = \sum_{k=0}^{n-1} T^k$ ,  $\mathbb{P}(T^k = 1) \geq (1 - \delta)$

Then  $\gamma \in (0, 1)$ ,  $K$  a stopping time with finite expectation

$$\mathbb{P}((1 - \gamma)K \geq 2S^K + \gamma(1 - \gamma^2)\mathbb{E}(K)) \leq e^{-\gamma^2\mathbb{E}(K)}$$

## Proof (sketch).

Define

$$\Phi^N = (1 + \gamma)^{\sum_{k=0}^{N-1} (1 - T^k)} (1 - \gamma)^{\sum_{k=0}^{N-1} T^k} e^{-\gamma^2\mathbb{E}(K)}$$

Turns out  $\Phi^N$  is a martingale

$$\mathbb{E}(\Phi^K) \leq \mathbb{E}(\Phi^0)$$

## Markov and convex inequalities



<sup>1</sup>N. Young *Chernoff-type bounds for a stopped sum of independent random variables*, unpublished

## Gradient Lower Bound and Sufficient Decrease

$f : \mathbb{R}^n \rightarrow \mathbb{R}$  Lipschitz and a sequence  $\{x^k\}_{k \in \mathbb{N}}$

## Gradient Lower Bound and Sufficient Decrease

$f : \mathbb{R}^n \rightarrow \mathbb{R}$  Lipschitz and a sequence  $\{x^k\}_{k \in \mathbb{N}}$

Definition (Gradient Lower Bound)

$$\exists \tau, \quad \|\nabla f(x^k)\| \leq \tau \|x^{k+1} - x^k\|$$

## Gradient Lower Bound and Sufficient Decrease

$f : \mathbb{R}^n \rightarrow \mathbb{R}$  Lipschitz and a sequence  $\{x^k\}_{k \in \mathbb{N}}$

Definition (Gradient Lower Bound)

$$\exists \tau, \quad \|\nabla f(x^k)\| \leq \tau \|x^{k+1} - x^k\|$$

Definition (Sufficient Decrease)

$$\exists \rho, \quad f(x^k) - f(x^{k+1}) \geq \rho \|x^{k+1} - x^k\|^2$$

# Gradient Lower Bound and Sufficient Decrease

$f : \mathbb{R}^n \rightarrow \mathbb{R}$  Lipschitz and a sequence  $\{x^k\}_{k \in \mathbb{N}}$

Definition (Gradient Lower Bound)

$$\exists \tau, \quad \|\nabla f(x^k)\| \leq \tau \|x^{k+1} - x^k\|$$

Definition (Sufficient Decrease)

$$\exists \rho, \quad f(x^k) - f(x^{k+1}) \geq \rho \|x^{k+1} - x^k\|^2$$

Theorem

$$\min_{n \leq k} \|\nabla f(x^n)\| < \varepsilon$$

for all

$$k > \mathcal{O}(\varepsilon^{-2})$$

# Newton Differentiability

## Definition

$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is *pointwise (weakly) Newton differentiable at  $\bar{x}$*  if there are  $c \in \mathbb{R} (c \neq 0)$  and  $\mathcal{H}F : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n \times n}$

# Newton Differentiability

## Definition

$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is *pointwise (weakly) Newton differentiable at  $\bar{x}$*  if there are  $c \in \mathbb{R} (c \neq 0)$  and  $\mathcal{H}F : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n \times n}$  with

$$\lim_{x \rightarrow \bar{x}} \sup_{H \in \mathcal{H}F(x)} \frac{\|F(x) - F(\bar{x}) - H(x - \bar{x})\|}{\|x - \bar{x}\|} = c$$

# Newton Differentiability

## Definition

$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is *pointwise (weakly) Newton differentiable at  $\bar{x}$*  if there are  $c \in \mathbb{R} (c \neq 0)$  and  $\mathcal{H}F : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n \times n}$  with

$$\lim_{x \rightarrow \bar{x}} \sup_{H \in \mathcal{H}F(x)} \frac{\|F(x) - F(\bar{x}) - H(x - \bar{x})\|}{\|x - \bar{x}\|} = c$$

## Definition

$F : \mathbb{R}^m \rightarrow \mathbb{R}^n$  is *uniformly (weakly) Newton differentiable at  $\bar{x}$*  if there are  $c \in \mathbb{R} (c \neq 0)$ ,  $\mathcal{H}F : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n \times n}$  such that  $\forall \varepsilon > 0 \exists \delta$ ,  $\forall x \in \mathbb{R}^m$ ,  $\forall y \in V$  with  $\|x - y\| \leq \delta$ ,

# Newton Differentiability

## Definition

$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is *pointwise (weakly) Newton differentiable at  $\bar{x}$*  if there are  $c \in \mathbb{R} (c \neq 0)$  and  $\mathcal{H}F : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n \times n}$  with

$$\lim_{x \rightarrow \bar{x}} \sup_{H \in \mathcal{H}F(x)} \frac{\|F(x) - F(\bar{x}) - H(x - \bar{x})\|}{\|x - \bar{x}\|} = c$$

## Definition

$F : \mathbb{R}^m \rightarrow \mathbb{R}^n$  is *uniformly (weakly) Newton differentiable at  $\bar{x}$*  if there are  $c \in \mathbb{R} (c \neq 0)$ ,  $\mathcal{H}F : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n \times n}$  such that  $\forall \varepsilon > 0 \exists \delta$ ,  $\forall x \in \mathbb{R}^m$ ,  $\forall y \in V$  with  $\|x - y\| \leq \delta$ ,

$$\sup_{H \in \mathcal{H}F(x)} \frac{\|F(x) - F(\bar{x}) - H(x - \bar{x})\|}{\|x - \bar{x}\|} \leq c + \varepsilon$$

# Examples

## Proposition

$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $K$  compact,  $F \in C^1(K)$ , then  $F$  is uniformly Newton differentiable on  $K$

# Examples

## Proposition

$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $K$  compact,  $F \in \mathcal{C}^1(K)$ , then  $F$  is uniformly Newton differentiable on  $K$  with a Newton differential  $\mathcal{H}(x) := \{\nabla F(x)^T\}$

# Examples

## Proposition

$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $K$  compact,  $F \in \mathcal{C}^1(K)$ , then  $F$  is uniformly Newton differentiable on  $K$  with a Newton differential  $\mathcal{H}(x) := \{\nabla F(x)^T\}$

## Proposition

$f : \mathbb{R}^n \rightarrow \mathbb{R}$   $\mu$ -strongly convex  $\mathcal{C}^1$  and  $L$ -smooth,  $K$  compact, then  $\nabla f$  is uniformly weakly Newton differentiable on  $K$

# Examples

## Proposition

$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $K$  compact,  $F \in \mathcal{C}^1(K)$ , then  $F$  is uniformly Newton differentiable on  $K$  with a Newton differential  $\mathcal{H}(x) := \{\nabla F(x)^T\}$

## Proposition

$f : \mathbb{R}^n \rightarrow \mathbb{R}$   $\mu$ -strongly convex  $\mathcal{C}^1$  and  $L$ -smooth,  $K$  compact, then  $\nabla f$  is uniformly weakly Newton differentiable on  $K$  with a Newton differential  $\mathcal{H}(x) := \{\alpha \mathbf{I} \mid \frac{\sqrt{L^2 + \alpha^2} - 2\alpha\mu}{\alpha} \leq 1\}$

# Examples

## Definition

$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is Lipschitz *uniformly semi-smooth\** on  $V$  if  $\forall \varepsilon > 0$   
 $\exists \delta$  such that  $\forall x, \forall y \in V$  with  $\|x - y\| \leq \delta$

# Examples

## Definition

$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is Lipschitz *uniformly semi-smooth\** on  $V$  if  $\forall \varepsilon > 0$   
 $\exists \delta$  such that  $\forall x, \forall y \in V$  with  $\|x - y\| \leq \delta$ ,

$$\frac{\|F'(x; x - y) - F'(y; x - y)\|}{\|x - y\|} \leq \varepsilon$$

# Examples

## Definition

$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is Lipschitz *uniformly semi-smooth\** on  $V$  if  $\forall \varepsilon > 0$   
 $\exists \delta$  such that  $\forall x, \forall y \in V$  with  $\|x - y\| \leq \delta$ ,

$$\frac{\|F'(x; x - y) - F'(y; x - y)\|}{\|x - y\|} \leq \varepsilon$$

## Proposition

$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $F$  *uniformly semi-smooth\** on  $V$ , then  $F$  is  
uniformly Newton differentiable on  $V$

# Examples

## Definition

$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is Lipschitz *uniformly semi-smooth\** on  $V$  if  $\forall \varepsilon > 0$   
 $\exists \delta$  such that  $\forall x, \forall y \in V$  with  $\|x - y\| \leq \delta$ ,

$$\frac{\|F'(x; x - y) - F'(y; x - y)\|}{\|x - y\|} \leq \varepsilon$$

## Proposition

$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $F$  *uniformly semi-smooth\** on  $V$ , then  $F$  is  
*uniformly Newton differentiable* on  $V$  with a Newton differential  
 $\mathcal{H}(x) := \overline{\text{conv}} \left\{ H \in \mathbb{R}^{n \times m} \mid \exists \{x^k\}_{k \in \mathbb{N}}, \lim_{k \rightarrow \infty} \nabla F(x^k)^T = H \right\}$

# Convergence Result

## Theorem

Let  $f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$  be *Lipschitz* with a *local minimum* at  $\bar{x}$  and with *gradient uniformly weakly Newton differentiable* ( $\mathcal{H}f$ ) around  $\bar{x}$ ,  $c$  from Newton differentiable *small enough* for all  $x \in U$  and  $H \in \mathcal{H}f$ ,  $H$  is *positive semi-definite* and  $\|H^{-1}\| \leq \Omega$

# Convergence Result

## Theorem

Let  $f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$  be *Lipschitz* with a *local minimum* at  $\bar{x}$  and with *gradient uniformly weakly Newton differentiable* ( $\mathcal{H}f$ ) around  $\bar{x}$ ,  $c$  from Newton differentiable *small enough* for all  $x \in U$  and  $H \in \mathcal{H}f$ ,  $H$  is *positive semi-definite* and  $\|H^{-1}\| \leq \Omega$

$$x^{k+1} \in x^k - \mathcal{H}f(x^k)^{-1} \nabla f(x^k)$$

# Convergence Result

## Theorem

Let  $f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$  be *Lipschitz* with a *local minimum* at  $\bar{x}$  and with *gradient uniformly weakly Newton differentiable* ( $\mathcal{H}f$ ) around  $\bar{x}$ ,  $c$  from Newton differentiable *small enough* for all  $x \in U$  and  $H \in \mathcal{H}f$ ,  $H$  is *positive semi-definite* and  $\|H^{-1}\| \leq \Omega$

$$x^{k+1} \in x^k - \mathcal{H}f(x^k)^{-1} \nabla f(x^k)$$

then

$$\begin{aligned} \exists \tau, \quad \|\nabla f(x^k)\| &\leq \tau \|x^{k+1} - x^k\| \\ \exists \rho, \quad f(x^k) - f(x^{k+1}) &\geq \rho \|x^{k+1} - x^k\|^2 \end{aligned}$$

# Convergence Result

## Corollary

*With the same assumptions,*

$$\min_{n \leq k} \|\nabla f(x^n)\| < \varepsilon$$

*for all*

$$k > \mathcal{O}(\varepsilon^{-2})$$

# Convergence Result

## Corollary

*With the same assumptions,*

$$\min_{n \leq k} \|\nabla f(x^n)\| < \varepsilon$$

*for all*

$$k > \mathcal{O}(\varepsilon^{-2})$$

NOT a good result

# Convergence Result

## Corollary

*With the same assumptions,*

$$\min_{n \leq k} \|\nabla f(x^n)\| < \varepsilon$$

*for all*

$$k > \mathcal{O}(\varepsilon^{-2})$$

**NOT** a good result:

## Theorem

*With the same assumptions*

$$\|x^{k+1} - \bar{x}\| \leq (c\Omega)^k \|x^k - \bar{x}\|$$

# Setup

Deterministic algorithm

$$x^+ \in x - \mathcal{H}f(x)^{-1} \nabla f(x)$$

# Setup

Deterministic algorithm

$$x^+ \in x - \mathcal{H}f(x)^{-1} \nabla f(x)$$

We don't have access to  $\mathcal{H}f$

# Setup

Deterministic algorithm

$$x^+ \in x - \mathcal{H}f(x)^{-1} \nabla f(x)$$

We don't have access to  $\mathcal{H}f$

Only to a stochastic oracle  $\mathbb{H}$

# Setup

Deterministic algorithm

$$x^+ \in x - \mathcal{H}f(x)^{-1}\nabla f(x)$$

We don't have access to  $\mathcal{H}f$

Only to a stochastic oracle  $\mathbb{H}$

$$\mathbb{P}(\mathbb{H}(x) \in \mathcal{H}f(x)) \geq 1 - \delta$$

# Setup

Deterministic algorithm

$$x^+ \in x - \mathcal{H}f(x)^{-1}\nabla f(x)$$

We don't have access to  $\mathcal{H}f$

Only to a stochastic oracle  $\mathbb{H}$

$$\mathbb{P}(\mathbb{H}(x) \in \mathcal{H}f(x)) \geq 1 - \delta$$

We know

# Setup

## Deterministic algorithm

$$x^+ \in x - \mathcal{H}f(x)^{-1}\nabla f(x)$$

We don't have access to  $\mathcal{H}f$

Only to a stochastic oracle  $\mathbb{H}$

$$\mathbb{P}(\mathbb{H}(x) \in \mathcal{H}f(x)) \geq 1 - \delta$$

We know: if  $\mathbb{H}(x) \in \mathcal{H}f(x)$ , with  $x^+ \sim x - \mathbb{H}(x)^{-1}\nabla f(x)$

$$\begin{aligned} \exists \tau, \quad & \|\nabla f(x)\| \leq \tau \|x^+ - x\| \\ \exists \rho, \quad & f(x) - f(x^+) \geq \rho \|x^+ - x\|^2 \end{aligned} \tag{1}$$

# Setup

## Deterministic algorithm

$$x^+ \in x - \mathcal{H}f(x)^{-1}\nabla f(x)$$

We don't have access to  $\mathcal{H}f$

Only to a stochastic oracle  $\mathbb{H}$

$$\mathbb{P}(\mathbb{H}(x) \in \mathcal{H}f(x)) \geq 1 - \delta$$

We know: if  $\mathbb{H}(x) \in \mathcal{H}f(x)$ , with  $x^+ \sim x - \mathbb{H}(x)^{-1}\nabla f(x)$

$$\begin{aligned} \exists \tau, \quad \|\nabla f(x)\| &\leq \tau \|x^+ - x\| \\ \exists \rho, \quad f(x) - f(x^+) &\geq \rho \|x^+ - x\|^2 \end{aligned} \tag{1}$$

Idea

# Setup

## Deterministic algorithm

$$x^+ \in x - \mathcal{H}f(x)^{-1}\nabla f(x)$$

We don't have access to  $\mathcal{H}f$

Only to a stochastic oracle  $\mathbb{H}$

$$\mathbb{P}(\mathbb{H}(x) \in \mathcal{H}f(x)) \geq 1 - \delta$$

We know: if  $\mathbb{H}(x) \in \mathcal{H}f(x)$ , with  $x^+ \sim x - \mathbb{H}(x)^{-1}\nabla f(x)$

$$\begin{aligned} \exists \tau, \quad \|\nabla f(x)\| &\leq \tau \|x^+ - x\| \\ \exists \rho, \quad f(x) - f(x^+) &\geq \rho \|x^+ - x\|^2 \end{aligned} \tag{1}$$

Idea: use (1) as a test for  $H(x) \in \mathcal{H}f(x)$  via backtracking

## Algorithm

**Data:**  $f, \nabla f, x^0, c_0 \in [0, \infty), \alpha \in (0, 1), \varepsilon \in (0, 1)$ ;

- 1  $k \leftarrow 0$ ;
- 2 **while**  $\|\nabla f(x^k)\| \geq \varepsilon$  **do**
  - 3 **Sample:**  $B^k$ ;
  - 4  $y \leftarrow x^k - B^k \partial_1 f(x^k)$ ;
  - 5 **if**  $f(x^k) - f(y) \geq c_k \|y - x^k\|^2$  **and**  
 $\|\partial_1 f(x^k)\| \leq c_k^{-1} \|y - x^k\|$  **then**
    - 6  $x^{k+1} \leftarrow y$ ;
    - 7  $c_{k+1} \leftarrow c_k$ ;
  - 8 **else**
    - 9  $x^{k+1} \leftarrow x^k$ ;
    - 10  $c_{k+1} \leftarrow \alpha c_k$ ;
  - 11 **end**
- 12  $k \leftarrow k + 1$ ;
- 13 **end**
- 14 **return**  $x^k$ ;

# Main Result

## Theorem

Let  $f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$  be *Lipschitz* with a *local minimum* at  $\bar{x}$  and with *gradient uniformly weakly Newton differentiable* ( $\mathcal{H}f$ ) around  $\bar{x}$ ,  $c$  from Newton differentiable *small enough* for all  $x \in U$  and  $H \in \mathcal{H}f$ ,  $H$  is *positive semi-definite* and  $\|H^{-1}\| \leq \Omega$

# Main Result

## Theorem

Let  $f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$  be *Lipschitz* with a *local minimum* at  $\bar{x}$  and with *gradient uniformly weakly Newton differentiable* ( $\mathcal{H}f$ ) around  $\bar{x}$ ,  $c$  from Newton differentiable *small enough* for all  $x \in U$  and  $H \in \mathcal{H}f$ ,  *$H$  is positive semi-definite and  $\|H^{-1}\| \leq \Omega$*

Let  $x^k$  be the sequence produced by the algorithm and  $K \in \mathbb{N} \cup \{\infty\}$  the total number of iterations

# Main Result

## Theorem

Let  $f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$  be *Lipschitz* with a *local minimum* at  $\bar{x}$  and with *gradient uniformly weakly Newton differentiable* ( $\mathcal{H}f$ ) around  $\bar{x}$ ,  $c$  from Newton differentiable *small enough* for all  $x \in U$  and  $H \in \mathcal{H}f$ ,  $H$  is *positive semi-definite* and  $\|H^{-1}\| \leq \Omega$

Let  $x^k$  be the sequence produced by the algorithm and  $K \in \mathbb{N} \cup \{\infty\}$  the total number of iterations

Then

1.

$$\mathbb{E}(K) \leq (1 - \delta)^{-1} \mathcal{O}(\varepsilon^{-2})$$

# Main Result

## Theorem

Let  $f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$  be *Lipschitz* with a *local minimum* at  $\bar{x}$  and with *gradient uniformly weakly Newton differentiable* ( $\mathcal{H}f$ ) around  $\bar{x}$ ,  $c$  from Newton differentiable *small enough* for all  $x \in U$  and  $H \in \mathcal{H}f$ ,  $H$  is *positive semi-definite* and  $\|H^{-1}\| \leq \Omega$

Let  $x^k$  be the sequence produced by the algorithm and  $K \in \mathbb{N} \cup \{\infty\}$  the total number of iterations

Then

1.

$$\mathbb{E}(K) \leq (1 - \delta)^{-1} \mathcal{O}(\varepsilon^{-2})$$

2.

$$\mathbb{P}(K \geq \gamma^2 \mathbb{E}(K)) \leq \mathcal{O}(e^{-\gamma^2(1-\delta)^{-1}\varepsilon^{-2}})$$

# Proof Idea

1. Any **successful**<sup>2</sup> iteration makes progress

---

$$\begin{aligned} &^2 x^{k+1} \neq x^k \\ &^3 x^k \in \mathcal{H}f(x^k) \end{aligned}$$

# Proof Idea

1. Any **successful**<sup>2</sup> iteration makes progress
2. From **error bounds**, we need **at most**  $\mathcal{O}(\varepsilon^{-2})$  **true**<sup>3</sup> iterations

---

<sup>2</sup> $x^{k+1} \neq x^k$   
<sup>3</sup> $x^k \in \mathcal{H}f(x^k)$

# Proof Idea

1. Any **successful**<sup>2</sup> iteration makes progress
2. From **error bounds**, we need at most  $\mathcal{O}(\varepsilon^{-2})$  **true**<sup>3</sup> iterations
3.  $K$  is bounded by **sum of Bernoulli** (iteration is true)

---

$$\begin{aligned} &^2 x^{k+1} \neq x^k \\ &^3 x^k \in \mathcal{H}f(x^k) \end{aligned}$$

# Proof Idea

1. Any **successful**<sup>2</sup> iteration makes progress
2. From **error bounds**, we need at most  $\mathcal{O}(\varepsilon^{-2})$  **true**<sup>3</sup> iterations
3.  $K$  is bounded by **sum of Bernoulli** (iteration is true)
4. **Expectation of sums** gives  $\mathbb{E}(K) \leq (1 - \delta)^{-1} \mathcal{O}(\varepsilon^{-2})$

---

<sup>2</sup>  $x^{k+1} \neq x^k$   
<sup>3</sup>  $x^k \in \mathcal{H}f(x^k)$

# Proof Idea

1. Any **successful**<sup>2</sup> iteration makes progress
2. From **error bounds**, we need at most  $\mathcal{O}(\varepsilon^{-2})$  true<sup>3</sup> iterations
3.  $K$  is bounded by **sum of Bernoulli** (iteration is true)
4. **Expectation of sums** gives  $\mathbb{E}(K) \leq (1 - \delta)^{-1} \mathcal{O}(\varepsilon^{-2})$
5. **Chernoff** gives  $\mathbb{P}(K \geq \gamma^2 \mathbb{E}(K)) \leq \mathcal{O}(e^{-\gamma^2(1-\delta)^{-1} \varepsilon^{-2}})$

---

<sup>2</sup>  $x^{k+1} \neq x^k$   
<sup>3</sup>  $x^k \in \mathcal{H}f(x^k)$

# Noisy Newton

$f \in \mathcal{C}^2$  convex

# Noisy Newton

$f \in \mathcal{C}^2$  convex, and we sample  $B(x) \sim \nabla^2 f(x) + \mathcal{N}(0, \Sigma)$

# Noisy Newton

$f \in \mathcal{C}^2$  convex, and we sample  $B(x) \sim \nabla^2 f(x) + \mathcal{N}(0, \Sigma)$

**Problem**

# Noisy Newton

$f \in \mathcal{C}^2$  convex, and we sample  $B(x) \sim \nabla^2 f(x) + \mathcal{N}(0, \Sigma)$

**Problem:**  $\mathbb{P}(B(x) \in \{\nabla^2 f(x)\}) = 0$

# Noisy Newton

$f \in \mathcal{C}^2$  convex, and we sample  $B(x) \sim \nabla^2 f(x) + \mathcal{N}(0, \Sigma)$

**Problem:**  $\mathbb{P}(B(x) \in \{\nabla^2 f(x)\}) = 0$

## Proposition

If  $F$  is **uniformly Newton differentiable** with  $\mathcal{H}F$ , then it is weakly Newton differentiable with  $\mathcal{H}F + \mathbb{B}_\varepsilon[0]$

# Noisy Newton

$f \in \mathcal{C}^2$  convex, and we sample  $B(x) \sim \nabla^2 f(x) + \mathcal{N}(0, \Sigma)$

**Problem:**  $\mathbb{P}(B(x) \in \{\nabla^2 f(x)\}) = 0$

## Proposition

If  $F$  is **uniformly Newton differentiable** with  $\mathcal{H}F$ , then it is weakly Newton differentiable with  $\mathcal{H}F + \mathbb{B}_\epsilon[0]$

## Proposition

$$\mathbb{P}(B(x) \in \nabla^2 f(x) + \mathbb{B}_{n^{3/2}\sqrt{\text{tr}\Sigma}}[0]) \geq 1 - n^{-1}$$

## Lemma (Johnson-Lindenstrauss)

For any  $n \in \mathbb{N}$ ,  $\varepsilon > 0$ ,  $\delta < 1/2$  and  $d \in \mathcal{O}(-\varepsilon^{-2} \log(\delta))$ , we sample  $A \sim \mathcal{D}$ ,  $A \in \mathbb{R}^{d \times n} \forall x \in \mathbb{R}^n$  with  $\|x\| = 1$

## Lemma (Johnson-Lindenstrauss)

For any  $n \in \mathbb{N}$ ,  $\varepsilon > 0$ ,  $\delta < 1/2$  and  $d \in \mathcal{O}(-\varepsilon^{-2} \log(\delta))$ , we sample  $A \sim \mathcal{D}$ ,  $A \in \mathbb{R}^{d \times n} \forall x \in \mathbb{R}^n$  with  $\|x\| = 1$

$$\mathbb{P}(|\|Ax\|^2 - 1| \leq \varepsilon) \leq \delta.$$

## Algorithm

$f \in \mathcal{C}^2$  convex

## Lemma (Johnson-Lindenstrauss)

For any  $n \in \mathbb{N}$ ,  $\varepsilon > 0$ ,  $\delta < 1/2$  and  $d \in \mathcal{O}(-\varepsilon^{-2} \log(\delta))$ , we sample  $A \sim \mathcal{D}$ ,  $A \in \mathbb{R}^{d \times n} \forall x \in \mathbb{R}^n$  with  $\|x\| = 1$

$$\mathbb{P}(|\|Ax\|^2 - 1| \leq \varepsilon) \leq \delta.$$

## Algorithm

$f \in \mathcal{C}^2$  convex, and we sample  $B(x)^{-1} \sim \mathcal{D}^T (\mathcal{D} \nabla^2 f(x) \mathcal{D}^T)^{-1} \mathcal{D}$

## Lemma (Johnson-Lindenstrauss)

For any  $n \in \mathbb{N}$ ,  $\varepsilon > 0$ ,  $\delta < 1/2$  and  $d \in \mathcal{O}(-\varepsilon^{-2} \log(\delta))$ , we sample  $A \sim \mathcal{D}$ ,  $A \in \mathbb{R}^{d \times n} \forall x \in \mathbb{R}^n$  with  $\|x\| = 1$

$$\mathbb{P}(|\|Ax\|^2 - 1| \leq \varepsilon) \leq \delta.$$

## Algorithm

$f \in \mathcal{C}^2$  convex, and we sample  $B(x)^{-1} \sim \mathcal{D}^T (\mathcal{D} \nabla^2 f(x) \mathcal{D}^T)^{-1} \mathcal{D}$

## Proposition

If  $F$  is **uniformly Newton differentiable** with  $\mathcal{H}F$ , then it is weakly Newton differentiable with  $\mathcal{G}F$  if

$$\|x - y\| \leq \delta \implies \frac{\|((\mathcal{H}F(x) - \mathcal{G}F(x))(x - y))\|}{\|x - y\|} \leq \varepsilon$$

# Image Denoising

## Problem

$$n : \mathbb{R}^{m_1 \times m_2} \rightarrow \mathbb{R} \text{ and } d : \mathbb{R}^{m_1 \times m_2} \rightarrow \mathbb{R}$$

# Image Denoising

## Problem

$n : \mathbb{R}^{m_1 \times m_2} \rightarrow \mathbb{R}$  and  $d : \mathbb{R}^{m_1 \times m_2} \rightarrow \mathbb{R}$

$$n(x) = \|\nabla x\|^2 := \sum_{i=2}^{m_1-1} \sum_{j=2}^{m_1-1} (x_{i,j+1} - x_{i,j-1})^2 + (x_{i+1,j} - x_{i-1,j})^2$$
$$d(x) = \|x - o\|_F^2$$

# Image Denoising

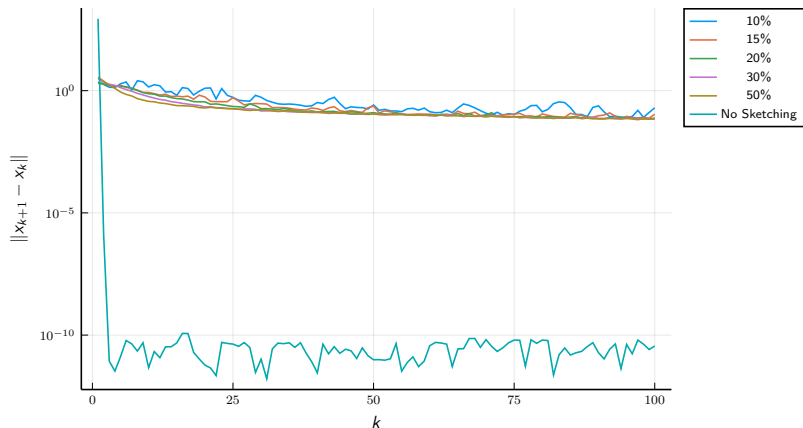
## Problem

$n : \mathbb{R}^{m_1 \times m_2} \rightarrow \mathbb{R}$  and  $d : \mathbb{R}^{m_1 \times m_2} \rightarrow \mathbb{R}$

$$n(x) = \|\nabla x\|^2 := \sum_{i=2}^{m_1-1} \sum_{j=2}^{m_1-1} (x_{i,j+1} - x_{i,j-1})^2 + (x_{i+1,j} - x_{i-1,j})^2$$
$$d(x) = \|x - o\|_F^2$$

$$\min_{x \in \mathbb{R}^{m_1 \times m_2}} n(x) + \alpha d(x)$$

# Image Denoising



**Figure:** The probability distribution of the embedding matrices is  $\text{UnifII}(n, d)$

# Image Denoising

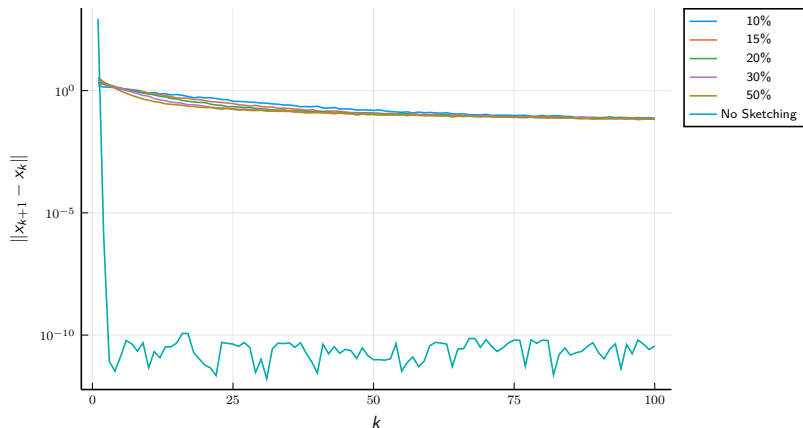


Figure: The probability distribution of the embedding matrices is  $\mathcal{N}(0, I)$

# Problem

1. the object we want  $u \in \mathcal{E} \subseteq \mathcal{L}^2(\mathbb{R}^3, \mathbb{C})$

## Problem

1. the object we want  $u \in \mathcal{E} \subseteq \mathcal{L}^2(\mathbb{R}^3, \mathbb{C})$
2. unknown random noise  $\rho = \text{UnifSO}(3)$

# Problem

1. the object we want  $u \in \mathcal{E} \subseteq \mathcal{L}^2(\mathbb{R}^3, \mathbb{C})$
2. unknown random noise  $\rho = \text{UnifSO}(3)$
3. the function  $|\mathcal{F}| : \mathcal{E} \rightarrow \Pi(\mathbb{R}^2)$

# Problem

1. the object we want  $u \in \mathcal{E} \subseteq \mathcal{L}^2(\mathbb{R}^3, \mathbb{C})$
2. unknown random noise  $\rho = \text{UnifSO}(3)$
3. the function  $|\mathcal{F}| : \mathcal{E} \rightarrow \Pi(\mathbb{R}^2)$
4. the amplitude of the electron intensity on the detector plane

$$\delta_u(A) = \int_{\text{SO}(3)} |\mathcal{F}|(u \circ \rho)(A) \, d\rho$$

# Problem

1. the object we want  $u \in \mathcal{E} \subseteq \mathcal{L}^2(\mathbb{R}^3, \mathbb{C})$
2. unknown random noise  $\rho = \text{UnifSO}(3)$
3. the function  $|\mathcal{F}| : \mathcal{E} \rightarrow \Pi(\mathbb{R}^2)$
4. the amplitude of the electron intensity on the detector plane

$$\delta_u(A) = \int_{\text{SO}(3)} |\mathcal{F}|(u \circ \rho)(A) d\rho$$

5.  $\delta_u(A)$  is the expectation number of electrons to detect in  $A \subseteq \mathbb{R}^2$

# Problem

1. the object we want  $u \in \mathcal{E} \subseteq \mathcal{L}^2(\mathbb{R}^3, \mathbb{C})$
2. unknown random noise  $\rho = \text{UnifSO}(3)$
3. the function  $|\mathcal{F}| : \mathcal{E} \rightarrow \Pi(\mathbb{R}^2)$
4. the amplitude of the electron intensity on the detector plane

$$\delta_u(A) = \int_{\text{SO}(3)} |\mathcal{F}|(u \circ \rho)(A) d\rho$$

5.  $\delta_u(A)$  is the expectation number of electrons to detect in  $A \subseteq \mathbb{R}^2$
6. we can measure this!

# Problem

1. the object we want  $u \in \mathcal{E} \subseteq \mathcal{L}^2(\mathbb{R}^3, \mathbb{C})$
2. unknown random noise  $\rho = \text{UnifSO}(3)$
3. the function  $|\mathcal{F}| : \mathcal{E} \rightarrow \Pi(\mathbb{R}^2)$
4. the amplitude of the electron intensity on the detector plane

$$\delta_u(A) = \int_{\text{SO}(3)} |\mathcal{F}|(u \circ \rho)(A) d\rho$$

5.  $\delta_u(A)$  is the expectation number of electrons to detect in  $A \subseteq \mathbb{R}^2$
6. we can measure this!
- 7.

$$\mathcal{U}_k(x_1, \dots, x_N) = \operatorname{argmin}_{u \in \mathcal{E}} \sum_{i=1}^N -\log \frac{\partial \delta_u}{\partial \lambda}(u, x_i)$$

# Algorithm

$$\mathcal{U}_N(x_1, \dots, x_N) = \operatorname{argmin}_{u \in \mathcal{E}} \sum_{i=1}^N -\log \frac{\partial \delta_u}{\partial \lambda}(u, x_i)$$

# Algorithm

$$\mathcal{U}_N(x_1, \dots, x_N) = \operatorname{argmin}_{u \in \mathcal{E}} \sum_{i=1}^N -\log \frac{\partial \delta_u}{\partial \lambda}(u, x_i)$$

*N* is very large!

# Algorithm

$$\mathcal{U}_N(x_1, \dots, x_N) = \operatorname{argmin}_{u \in \mathcal{E}} \sum_{i=1}^N -\log \frac{\partial \delta_u}{\partial \lambda}(u, x_i)$$

$N$  is very large!

We use batch size  $B$

# Algorithm

$$\mathcal{U}_N(x_1, \dots, x_N) = \operatorname{argmin}_{u \in \mathcal{E}} \sum_{i=1}^N -\log \frac{\partial \delta_u}{\partial \lambda}(u, x_i)$$

*N* is very large!

We use batch size *B*

$$g_k(u) = \nabla \sum_{i=1}^B -\log f(u, X_{\sigma_k(i)})$$

# Algorithm

$$\mathcal{U}_N(x_1, \dots, x_N) = \operatorname{argmin}_{u \in \mathcal{E}} \sum_{i=1}^N -\log \frac{\partial \delta_u}{\partial \lambda}(u, x_i)$$

$N$  is very large!

We use batch size  $B$

$$\mathbf{g}_k(u) = \nabla \sum_{i=1}^B -\log f(u, X_{\sigma_k(i)})$$

$$H^{k+1} = \begin{cases} \alpha I & \text{if } M \text{ divides } k, \\ \mathcal{S}(u^k, \mathbf{g}_{\lfloor k/M \rfloor}(u^k), H^k) & \text{else,} \end{cases}$$

# XFEL Imaging

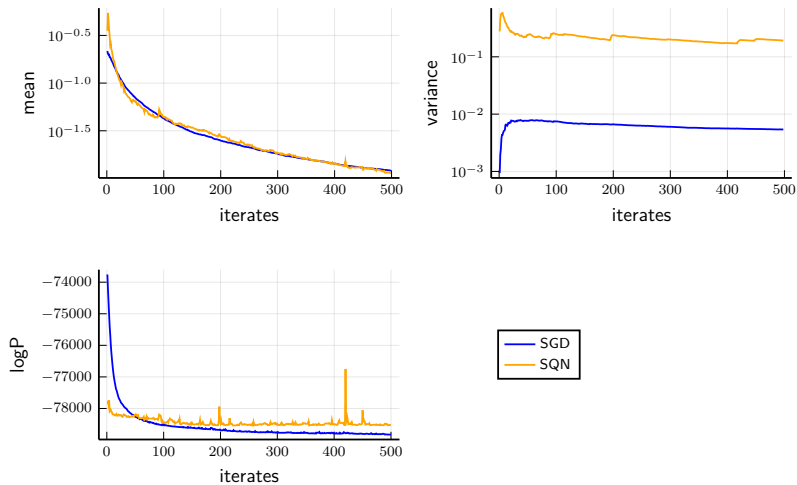


Figure:  $M = 10$

# XFEL Imaging

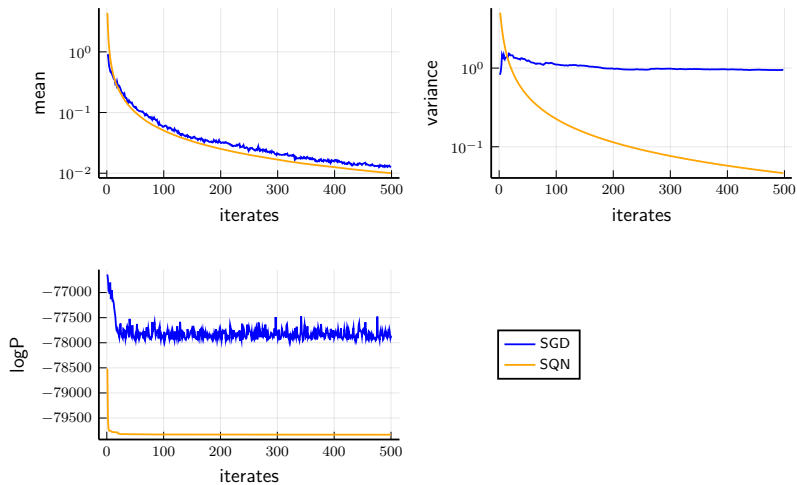


Figure:  $M = 100$

T. Pinça, *A Stochastic Newton-type Method for Non-smooth Optimization*. Math. Program. (under review).

