Numerical frugality in optimization: Newton's methods in mixed precision

PEQUAN Team presentation

G. Carrino, N. Brisebarre, T. Mary, E. Riccietti | Monday 13th April, 2026

Problem statement

Optimization problem

Solving the following optimization problem:

$$\min_{x \in \mathbb{R}^n} f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function.

Optimizers

Two popular optimizers:

Optimizers

Two popular optimizers:

- **Gradient descent**: $x_{i+1} = x_i - \underbrace{\alpha}_{\text{Step size}} g(x_i)$

Optimizers

Two popular optimizers:

- **Gradient descent**: $x_{i+1} = x_i - \underbrace{\alpha}_{\text{Step size}} \; g(x_i)$
- **Newton's method**: $x_{i+1} = x_i - \alpha \; \underbrace{H(x_i)^{-1}}_{\text{Curvature information}} \; g(x_i)$

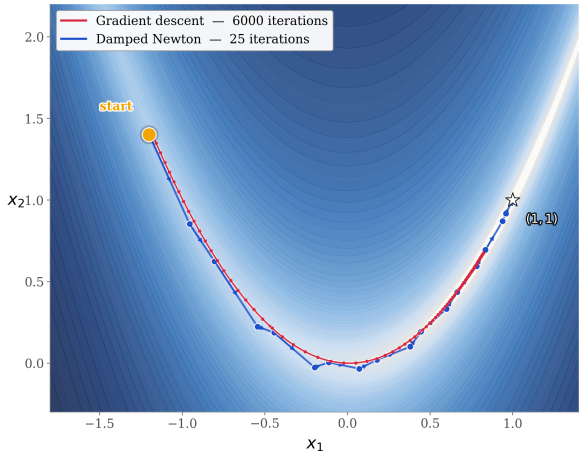
Optimizers

Two popular optimizers:

- **Gradient descent**: $x_{i+1} = x_i - \underbrace{\alpha}_{\text{Step size}} \quad g(x_i)$
- **Newton's method**: $x_{i+1} = x_i - \alpha \quad \underbrace{H(x_i)^{-1}}_{\text{Curvature information}} \quad g(x_i)$

Newton's method convergence is faster, but more expensive per iteration.

GD vs. Newton


Gradient descent — 6000 iterations
Damped Newton — 25 iterations
start
(1, 1)
2.0
1.5
1.0
0.5
0.0
x_2
−1.5
−1.0
−0.5
0.0
0.5
1.0
x_1

Newton's cost

Newton's method comes with two main sources of cost...

Computing the Hessian

$H(x_i)$

Solving the linear system

$H(x_i)\, p_i = -g(x_i)$

Newton's cost

Newton's method comes with two main sources of cost…

Computing the Hessian

$H(x_i)$

$\downarrow$

quasi-Newton methods

$B(x) \approx H(x)$

Solving the linear system

$H(x_i)\, p_i = -g(x_i)$

$\downarrow$

inexact Newton methods

$H(x_i)p_i + g(x_i) \leq \eta \|g(x_i)\|$

…but many possible approximate variants!

Quasi-Newton: an example

When solving least-squares problems,

Quasi-Newton: an example

When solving least-squares problems,

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|R(x)\|^2, \quad R : \mathbb{R}^n \to \mathbb{R}^m,$$

Quasi-Newton: an example

When solving least-squares problems,

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|R(x)\|^2, \quad R : \mathbb{R}^n \to \mathbb{R}^m,$$

the Hessian is given by

Quasi-Newton: an example

When solving least-squares problems,

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|R(x)\|^2, \quad R : \mathbb{R}^n \to \mathbb{R}^m,$$

the Hessian is given by

$$H(x) = J_R(x)^T J_R(x) + S(x), \quad S(x) = \sum_{i=1}^{m} R_i(x) \nabla^2 R_i(x)$$

Quasi-Newton: an example

When solving least-squares problems,

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|R(x)\|^2, \quad R : \mathbb{R}^n \to \mathbb{R}^m,$$

the Hessian is given by

$$H(x) = J_R(x)^T J_R(x) + S(x), \quad S(x) = \sum_{i=1}^{m} R_i(x) \nabla^2 R_i(x)$$

Gauss-Newton discards this!

Reducing the cost

- ***Issue***: these variants can also be expensive, especially for large-scale problems.

Reducing the cost

- ***Issue***: these variants can also be expensive, especially for large-scale problems.
- ***Idea***: use **low precision floating-point arithmetic** to reduce the overall cost.

Floating-point precision

$$\mathbf{fl}(x \circ y) = (x \circ y)(1 + \delta), \quad |\delta| \leq u, \quad \circ \in \{+, -, *, /\}$$

Floating-point arithmetic model.

Floating-point precision

$$\mathrm{fl}(x \circ y) = (x \circ y)(1+\delta), \quad |\delta| \leq u, \quad \circ \in \{+, -, *, /\}$$

Floating-point arithmetic model.

Type	Size	Unit roundoff (u)	Approx. range
bfloat16	16 bits	4×10^{-3}	$10^{\pm 38}$
fp16	16 bits	5×10^{-4}	$10^{\pm 4}$
fp32	32 bits	6×10^{-8}	$10^{\pm 38}$
fp64	64 bits	1×10^{-16}	$10^{\pm 308}$

$\approx 30\times$

$\approx 4\times$

Floating point formats.

Reducing the cost

- ***Issue***: also these variants can be expensive, especially for large-scale problems.
- ***Idea***: use **low floating-point precision** to reduce the overall cost.
- ***Challenge***: how to do it without sacrificing convergence?

Convergence analysis

Error model

For each iteration i, the mixed precision Newton's step satisfies

$$\widehat{d}_i \coloneqq -\left(H(\widehat{x}_i) + E_i^H\right)^{-1}\left(g(\widehat{x}_i) + e_i^g\right),$$
$$\widehat{x}_{i+1} = \widehat{x}_i + \widehat{d}_i + e_i^+ .$$

Error model

For each iteration i, the mixed precision Newton's step satisfies

$$\widehat{d}_i \coloneqq -\left(H(\widehat{x}_i) + E_i^H\right)^{-1}\left(g(\widehat{x}_i) + e_i^g\right),$$
$$\widehat{x}_{i+1} = \widehat{x}_i + \widehat{d}_i + e_i^+.$$

- E_i^H: Hessian approximations and/or inexact linear solvers.

Error model

For each iteration i, the mixed precision Newton's step satisfies

$$\widehat{d}_i \coloneqq -\left(H(\widehat{x}_i) + E_i^H\right)^{-1}\left(g(\widehat{x}_i) + e_i^g\right),$$
$$\widehat{x}_{i+1} = \widehat{x}_i + \widehat{d}_i + e_i^+ .$$

- E_i^H: Hessian approximations and/or inexact linear solvers.
- e_i^g: inexact gradient computations.

Error model

For each iteration i, the mixed precision Newton's step satisfies

$$\widehat{d}_i := -\left(H(\widehat{x}_i) + E_i^H\right)^{-1}\left(g(\widehat{x}_i) + e_i^g\right),$$
$$\widehat{x}_{i+1} = \widehat{x}_i + \widehat{d}_i + e_i^+.$$

- E_i^H: Hessian approximations and/or inexact linear solvers.
- e_i^g: inexact gradient computations.
- e_i^+: errors in the update step.

Convergence analysis

$$\|\widehat{x}_{i+1} - x^*\| \leq \alpha_i \|\widehat{x}_i - x^*\|^2 + \beta_i \|\widehat{x}_i - x^*\| + \gamma_i \,,$$

Mixed precision Newton's method convergence rate.

Convergence analysis

$$\|\widehat{x}_{i+1} - x^*\| \leq \boxed{\alpha_i} \|\widehat{x}_i - x^*\|^2 + \beta_i \|\widehat{x}_i - x^*\| + \gamma_i \,,$$

Mixed precision Newton's method convergence rate.

- α_i: standard Newton's quadratic convergence.

Convergence analysis

$$\|\widehat{x}_{i+1} - x^*\| \leq \boxed{\alpha_i} \|\widehat{x}_i - x^*\|^2 + \boxed{\beta_i} \|\widehat{x}_i - x^*\| + \gamma_i \,,$$

Mixed precision Newton's method convergence rate.

- α_i: standard Newton's quadratic convergence.
- β_i: linear convergence influenced by Hessian errors.

Convergence analysis

$$\|\widehat{x}_{i+1} - x^*\| \leq \boxed{\alpha_i} \|\widehat{x}_i - x^*\|^2 + \boxed{\beta_i} \|\widehat{x}_i - x^*\| + \boxed{\gamma_i},$$

Mixed precision Newton's method convergence rate.

- α_i: standard Newton's quadratic convergence.
- β_i: linear convergence influenced by Hessian errors.
- γ_i: **limiting accuracy** mainly related to gradient errors.

Error model appplications

Mixed precision Newton algorithm

Algorithm: Mixed precision Newton

Input: initial guess x_0, Hessian H, gradient g
Output: an approximation x_{i+1} to the minimizer x^*

1: **for** $i = 0, 1, \ldots$ until convergence **do**
2: Compute $g_i = g(x_i)$ in precision with roundoff u_g
3: Solve $H(x_i)d_i = -g_i$ in precision with roundoff u_H
4: Update $x_{i+1} = x_i + d_i$ in precision with roundoff u
5: **end for**
6: **return** x_{i+1}

Finite precision and error model

Roundoffs can be easily mapped to the error model:

Finite precision and error model

Roundoffs can be easily mapped to the error model:

- $\|e_i^+\| = u$, by standard floating-point arithmetic;

Finite precision and error model

Roundoffs can be easily mapped to the error model:

- $\|e_i^+\| = u$, by standard floating-point arithmetic;
- $\|E_i^H\| \approx O(u_H)$, if linear system is solved with a backward stable method;

Finite precision and error model

Roundoffs can be easily mapped to the error model:

- $\|e_i^+\| = u$, by standard floating-point arithmetic;
- $\|E_i^H\| \approx O(u_H)$, if linear system is solved with a backward stable method;
- $\|e_i^g\| \approx O(u_g)$.

Mixed precision setting

The setting of interest is

$$u_g \leq u \leq u_H$$

Mixed precision setting

The setting of interest is

$$\boxed{u_g} \leq u \leq u_H$$

- **Accurate gradient** (u_g small): better limiting accuracy γ_i.

Mixed precision setting

The setting of interest is

$$\boxed{u_g} \leq u \leq \boxed{u_H}$$

- **Accurate gradient** (u_g small): better limiting accuracy γ_i.
- **Reduced precision linear solver** (u_H large): reduced cost per iteration.

Mixed precision setting

The setting of interest is

$$\boxed{u_g} \leq \boxed{u} \leq \boxed{u_H}$$

- **Accurate gradient** (u_g small): better limiting accuracy γ_i.
- **Reduced precision linear solver** (u_H large): reduced cost per iteration.
- **Target precision**: the accuracy we aim to achieve.

Mixed precision setting

The setting of interest is

$$\boxed{u_g} \leq \boxed{u} \leq \boxed{u_H}$$

- **Accurate gradient** (u_g small): better limiting accuracy γ_i.
- **Reduced precision linear solver** (u_H large): reduced cost per iteration.
- **Target precision**: the accuracy we aim to achieve.
- **Trade-off**: potentially slower convergence.

Error model: extensions

Two key instances of E_i^H from the error model:

Inexact Newton

linear solver stopped early at tolerance η

Gauss-Newton

Hessian approximation discarding $S(\widehat{x}_i)$

Error model: extensions

Two key instances of E_i^H from the error model:

Inexact Newton

linear solver stopped early at tolerance η

$$\|E_i^H\| \leq \frac{\|g(\widehat{x}_i)\|}{\|H(\widehat{x}_i)\|\|\widehat{d}_i\|}\eta$$

Gauss-Newton

Hessian approximation discarding $S(\widehat{x}_i)$

Error model: extensions

Two key instances of E_i^H from the error model:

Inexact Newton

linear solver stopped early at tolerance η

$$\|E_i^H\| \leq \frac{\|g(\widehat{x}_i)\|}{\|H(\widehat{x}_i)\|\|\widehat{d}_i\|}\eta$$

Gauss-Newton

Hessian approximation discarding $S(\widehat{x}_i)$

$$\|E_i^H\| \leq \frac{\|S(\widehat{x}_i)\|}{\|H(\widehat{x}_i)\|}$$

Inexact Newton and Gauss-Newton both fit the E_i^H framework.

Numerical experiments

Numerical experiments

$$f(x) = 3 + \sum_{i=0}^{n-2} (x_i^2 + x_{i+1}^2)^2 - 4x_i.$$

ENGVAL1 function (from CUTEst dataset).

Numerical experiments

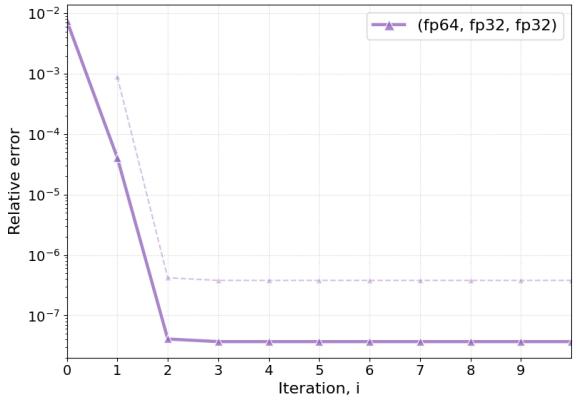
$$f(x) = 3 + \sum_{i=0}^{n-2} (x_i^2 + x_{i+1}^2)^2 - 4x_i.$$

ENGVAL1 function (from CUTEst dataset).

$$R(x) = x_0 z + \sum_{k=1}^{\lceil \frac{n-1}{2} \rceil} x_{2k-1} z^{k+1} + \sum_{k=1}^{\lfloor \frac{n-1}{2} \rfloor} x_{2k} \sin(x_{2k} z),$$

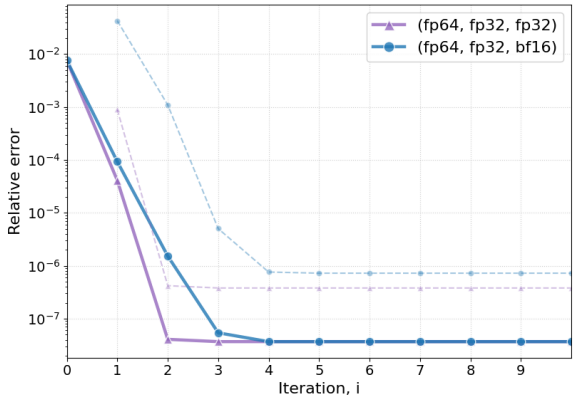
SINREG least squares residual, with z fixed data.

Charts - Mixed precision Newton



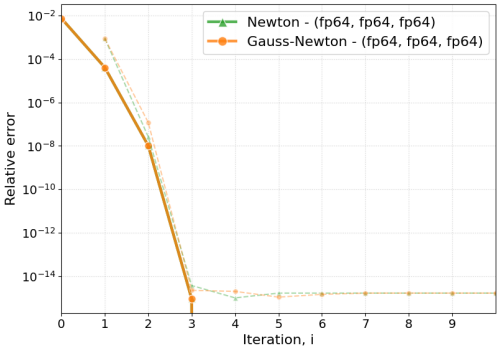
SINREG. Mixed precision Newton relative error and its predicted rate.

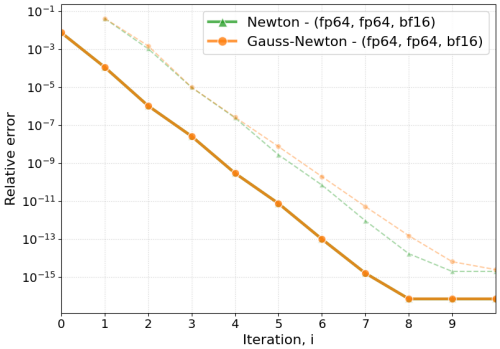
Charts - Two mixed precision settings



SINREG. Different mixed precision settings relative error.

Charts - Approximated Hessian



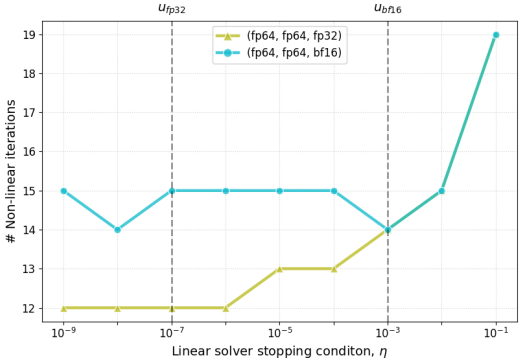


SINREG. Gauss-Newton vs. Newton's method.

Charts - Inexact linear solver

Inexact Newton **Hessian error**

$$\epsilon_i^H = c_{\text{solver}} u_H + \frac{\|g(\widehat{x}_i)\|}{\|H(\widehat{x}_i)\|\|\widehat{d}_i\|}\eta$$



ENGVAL1. Inexact Newton: iterations vs. η.

Conclusions

Our contribution

- A comprehensive **error model** for mixed precision Newton's methods.

Our contribution

- A comprehensive **error model** for mixed precision Newton's methods.
- A rigorous **convergence analysis**, highlighting the interplay between *precision* and *convergence rates*.

Our contribution

- A comprehensive **error model** for mixed precision Newton's methods.
- A rigorous **convergence analysis**, highlighting the interplay between *precision* and *convergence rates*.
- **Numerical experiments** validated theoretical findings, demonstrating the practical benefits of mixed precision strategies.

Future directions

Global convergence

Numerical errors in trust regions and line search strategies.

Future directions

Global convergence

Numerical errors in trust regions and line search strategies.

Sub-sampled Newton

Stochastic gradients and Hessians under mixed precision.

Future directions

Global convergence

Numerical errors in trust regions and line search strategies.

Sub-sampled Newton

Stochastic gradients and Hessians under mixed precision.

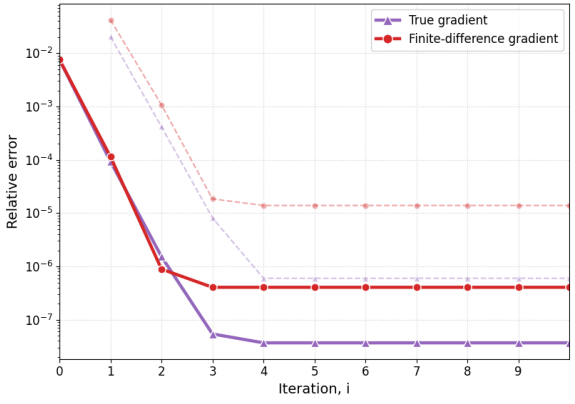
HPC benchmarking

Practical implementations for large-scale optimization.

Thank you!

Questions?

Charts - Approximated gradient



Finite differences with (fp64, fp32, bf16).