# Accelerated Stochastic Matrix Inversion:

## General Theory and Speeding up BFGS Rules for Faster Second-Order Optimization

Robert M. Gower[1] Filip Hanzely[2] Peter Richtárik[2, 3, 4] Sebastian U. Stich[5]

[1]Télécom ParisTech, [2]KAUST, [3]University of Edinburgh [4]Moscow Institute of Physics and Technology [5]EPFL

## Linear Systems in Euclidean Space

$$\mathcal{A}x = b,$$

For $\mathcal{X}$ and $\mathcal{Y}$ finite dimensional Euclidean spaces, $\mathcal{A}: \mathcal{X} \mapsto \mathcal{Y}$ a linear operator.

**Optimization Problem**: For $x_0 \in \mathcal{X}$:

$$x^* \overset{\text{def}}{=} \arg\min_{x \in \mathcal{X}} \tfrac{1}{2}\|x - x_0\|^2 \quad \text{subject to} \quad \mathcal{A}x = b.$$

### Motivation: Matrix Inversion

$$A^{-1} = \arg\min_X \|X\|^2_{F(A)} = \|A^{1/2}XA^{1/2}\|_F$$
$$\text{s.t.} \quad AX = I, \; X = X^\top$$

for symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$. Adaptive sketch-and-project is competitive with state of the art [1].

## Sketch-and-Project Updates

**Sketch and project iteration:**

$$x_{k+1} = \operatorname{argmin}_x \|x_k - x\|^2_B \quad \text{s.t.} \quad S_k^\top A x = S_k^\top b,$$

where $\|x\|_B^2 = \langle Bx, x\rangle$ for some $B \succ 0$ and $S_k$ is a random sketching matrix.

Classical rate:

$$\mathbf{E}\left[\|x_k - x^*\|_B^2\right] \le (1-\mu)^k \|x_0 - x^*\|_B^2.$$

$$\mu \overset{\text{def}}{=} \lambda^+_{\min} \mathbf{E}\left[B^{-\frac{1}{2}} A^\top S_k (S_k^\top A B^{-1} A^\top S_k)^\dagger S_k^\top A B^{-\frac{1}{2}}\right]$$

**Extending [2], we analyze accelerated sketch-and-project algorithms in Euclidean spaces.**

## Main Contributions

- **Accelerated Sketch and Project in Euclidean Spaces.**
- **Faster Algorithms for Matrix Inversion.**
- **Randomized Accelerated Quasi-Newton.**
- **Accelerated Quasi-Newton.**

## Algorithm

**Algorithm 1** Accelerated Sketch-and-Project

1: **Parameters:** $\mu, \nu > 0$, $\mathcal{D}$ = distribution over random linear operators $\mathcal{S}$
2: Choose $x_0 \in \mathcal{X}$ and set $v_0 = x_0$, $\beta = 1 - \sqrt{\frac{\mu}{\nu}}$, $\gamma = \sqrt{\frac{1}{\mu\nu}}$, $\alpha = \frac{1}{1+\gamma\nu}$.
3: **for** $k = 0, 1, \dots$ **do**
4: $\quad y_k = \alpha v_k + (1-\alpha)x_k$
5: $\quad$ Sample an independent copy $\mathcal{S}_k \sim \mathcal{D}$
6: $\quad g_k = \mathcal{A}^*\mathcal{S}_k^*(\mathcal{S}_k\mathcal{A}\mathcal{A}^*\mathcal{S}_k^*)^\dagger \mathcal{S}_k(\mathcal{A}y_k - b)$
7: $\quad x_{k+1} = y_k - g_k$
8: $\quad v_{k+1} = \beta v_k + (1-\beta)y_k - \gamma g_k$
9: **end for**

$$\mu \overset{\text{def}}{=} \inf_{x \in \mathbf{Range}(\mathcal{A}^*)} \frac{\langle \mathbf{E}[Z]x, x\rangle}{\langle x, x\rangle} \quad \text{(strong convexity)}$$

$$\nu \overset{\text{def}}{=} \sup_{x \in \mathbf{Range}(\mathcal{A}^*)} \frac{\langle \mathbf{E}\left[Z\mathbf{E}[Z]^\dagger Z\right]x, x\rangle}{\langle \mathbf{E}[Z]x, x\rangle} \quad \text{(new parameter)}$$

$$Z \overset{\text{def}}{=} \mathcal{A}^*\mathcal{S}_k^*(\mathcal{S}_k\mathcal{A}\mathcal{A}^*\mathcal{S}_k^*)^\dagger \mathcal{S}_k\mathcal{A}$$

**Lemma:**

$$1 \le \nu \le \tfrac{1}{\mu} = \|\mathbf{E}[Z]^\dagger\|$$

and if $\mathbf{Range}(\mathcal{A}^*) = \mathcal{X}$, then $\frac{\mathbf{Rank}(\mathcal{A}^*)}{\mathbf{E}[\mathbf{Rank}(Z)]} \le \nu$.

**Example:**(Linear systems in $\mathbb{R}^n$)

Choose $B = A$ and $S = e_i$ with probability proportional to $A_{i,i}$. Then

$$\mu = \frac{\lambda_{\min}(A)}{\mathbf{Tr}(A)} \quad \text{and} \quad \nu = \frac{\mathbf{Tr}(A)}{\min_i A_{i,i}}.$$

### Theorem

If $\mathbf{Null}(\mathcal{A}) = \mathbf{Null}(\mathbf{E}[Z])$, (exactness)

$$\mathbf{E}\left[\|v_k - x_*\|^2_{\mathbf{E}[Z]^\dagger} + \tfrac{1}{\mu}\|x_k - x_*\|^2\right] \le \left(1 - \sqrt{\tfrac{\mu}{\nu}}\right)^k \mathbf{E}\left[\|v_0 - x_*\|^2_{\mathbf{E}[Z]^\dagger} + \tfrac{1}{\mu}\|x_0 - x_*\|^2\right]$$

## References

[1] Robert M Gower and Peter Richtárik.
Randomized quasi-Newton updates are linearly convergent matrix inversion algorithms.
*SIAM Journal on Matrix Analysis and Applications*, 38(4):1380–1409, 2017.

[2] Peter Richtárik and Martin Takáč.
Stochastic reformulations of linear systems: accelerated method.
*Manuscript, October 2017*, 2017.

## Accelerated BFGS updates

**Optimization Problem:**

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

for $f: \mathbb{R}^d \to \mathbb{R}$ convex and sufficiently smooth.

**Quasi-Newton Methods:**

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

where $X_k \approx (\nabla^2 f(w_k))^{-1}$.

**Quasi-Newton update:** (Secant equation)

$$X_k(\nabla f(w_k) - \nabla f(w_{k-1})) = w_k - w_{k-1}, \quad X_k = X_k^\top.$$

This can also be written as

$$X_{k+1} = \operatorname{argmin}_X \|X - X_k\|^2_{F(A)}$$
$$\text{s.t.} \quad X(w_{k+1} - w_k) = \nabla f(w_{k+1}) - \nabla f(w_k)$$
$$X = X^\top.$$

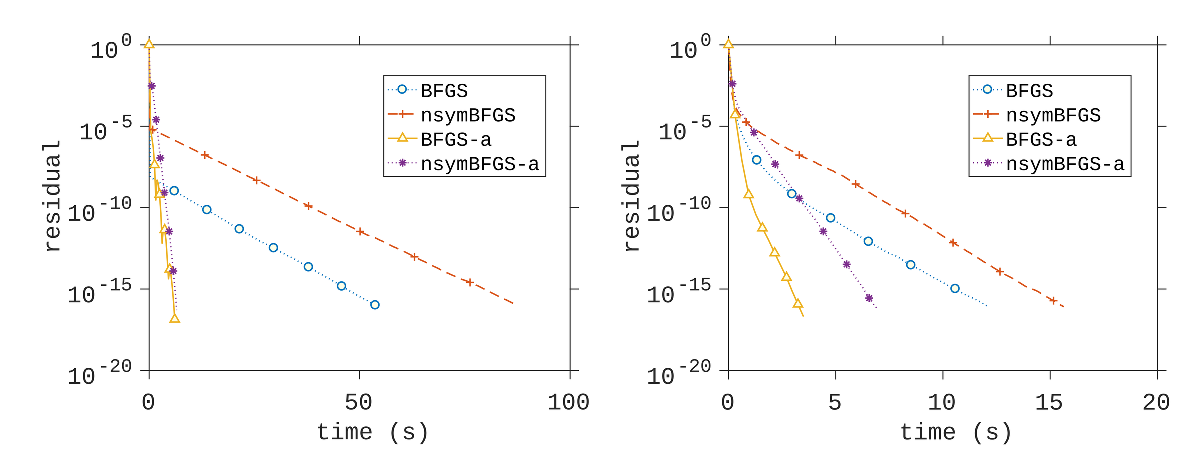
**Algorithm 2** BFGS method with accelerated BFGS update

1: **Parameters:** $\mu, \nu > 0$, stepsize $\eta$.
2: Choose $X_0 \in \mathcal{X}$, $w_0$ and set $V_0 = X_0$, $\beta = 1 - \sqrt{\frac{\mu}{\nu}}$, $\gamma = \sqrt{\frac{1}{\mu\nu}}$, $\alpha = \frac{1}{1+\gamma\nu}$.
3: **for** $k = 0, 1, \dots$ **do**
4: $\quad w_{k+1} = w_k - \eta X_k \nabla f(w_k)$
5: $\quad s_k = w_{k+1} - w_k, \quad \zeta_k = \nabla f(w_{k+1}) - \nabla f(w_k)$
6: $\quad Y_k = \alpha V_k + (1-\alpha)X_k$
7: $\quad X_{k+1} = \frac{\delta_k\delta_k^\top}{\delta_k^\top\zeta_k} + \left(I - \frac{\delta_k\zeta_k^\top}{\delta_k^\top\zeta_k}\right) Y_k \left(I - \frac{\zeta_k\delta_k^\top}{\delta_k^\top\zeta_k}\right)$
8: $\quad V_{k+1} = \beta V_k + (1-\beta)Y_k - \gamma(Y_k - X_{k+1})$
9: **end for**

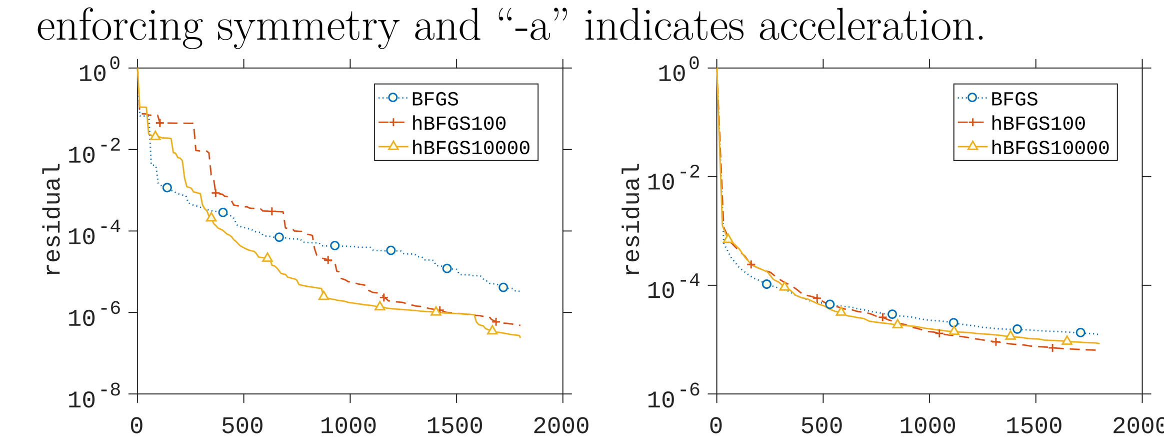
**Remark:** Here the Sketch-and-Project update is deterministic, the theory does not apply.

## Experiments

### Accelerated Matrix Inversion

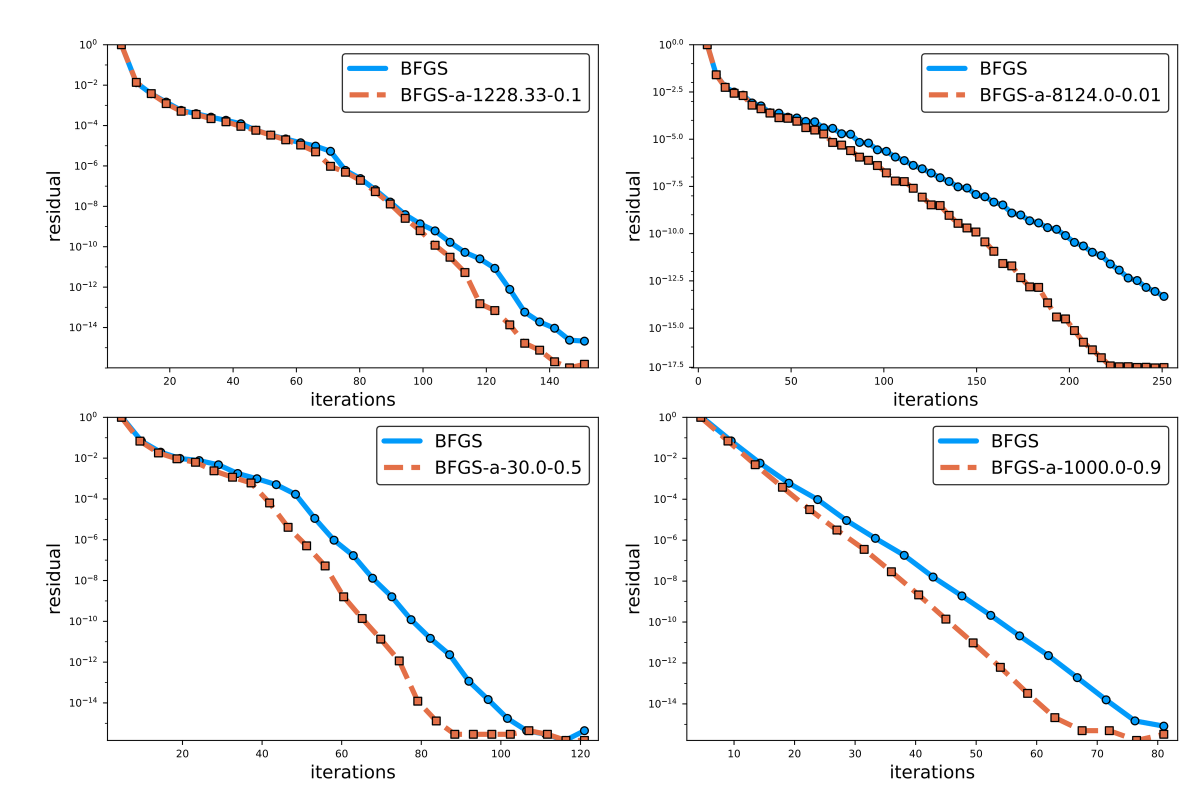



Left: Eigenvalues of $A \in \mathbb{R}^{100\times 100}$ are $1, 10^3, 10^3, \dots, 10^3$ and coordinate sketches with probabilities proportional to $\operatorname{diag}(A)$ are used. Right: Eigenvalues of $A \in \mathbb{R}^{100\times 100}$ are $1, 2, \dots, n$ and Gaussian sketches are used. Label "nsym" indicates non-enforcing symmetry and "-a" indicates acceleration.




Left: Epsilon dataset ($n = 2000$), uniform coordinate sketches. Right: SVHN ($n = 3072$), coordinate sketches with probabilities proportiaonal to $\operatorname{diag}(A)$. We choose $\mu = \frac{1}{100\nu}$ or $\mu = \frac{1}{10000\nu}$.

### BFGS with accelerated update




Algorithm 2 vs standard BFGS. From left to right: `phishing`, `mushrooms`, `australian` and `splice` dataset. Acceleration parameters chosen via grid search.

## Challenges

- Limited memory updates
- Convergence guarantees for Algorithm 2
- Adaptive sketches