# Stochastic Block BFGS: Squeezing More Curvature out of Data

Robert M. Gower† Donald Goldfarb* Peter Richtárik†

† University of Edinburgh *Columbia University

## 1. Problem

Find an approximate minima of

$$\min_{x\in\mathbb{R}^d} f(x) \overset{\text{def}}{=} \tfrac{1}{n}\textstyle\sum_{i=1}^n f_i(x), \qquad (1)$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex and twice differentiable, $d$ is large and **$n$ is very large**.

## 2. Variable Metric Methods

Given $x_0 \in \mathbb{R}^d$, many successful methods for solving (1) fit the format

$$x_{t+1} = x_t - \eta H_t g_t,$$

where $\mathbf{E}[g_t] = \nabla f(x_t)$, $H_t \approx \nabla^2 f(x_t)^{-1}$, and $\eta > 0$ is a stepsize. To update $g_t$ and $H_t$, effective methods use only the *subsampled gradient* and *subsampled Hessian*

$$\nabla f_S(x) \overset{\text{def}}{=} \frac{1}{|S|}\sum_{i\in S}\nabla f_i(x), \quad \nabla^2 f_T(x) \overset{\text{def}}{=} \frac{1}{|T|}\sum_{i\in T}\nabla^2 f_i(x)$$

where $S, T \subseteq [n] \overset{\text{def}}{=} \{1,2,\ldots,n\}$ selected uniformly at random.

**Challenge:** Update $H_t$ using subsampled Hessians.

**Novelty:** We develop a new stochastic Block BFGS method for updating/maintaining $H_t$ based on sketching. We also present a new limited memory variant.

## 5. Block L-BFGS update

Let $V_t = I - D_t\Delta_t Y_t^T$. Expanding $M$ block BFGS updates applied to $H_{t-M}$ gives

$$\begin{aligned} H_t &= V_t H_{t-1} V_t^T + D_t\Delta_t D_t^T \\ &= V_t\cdots V_{t+1-M} H_{t-M} V_{t+1-M}^T\cdots V_t^T \\ &+ \sum_{i=t}^{t+1-M} V_t\cdots V_{i+1} D_i\Delta_i D_i^T V_{i+1}^T\cdots V_t^T. \end{aligned}$$

Therefore $H_t$ is a function of $H_{t-M}$ and the triples

$$(D_{t+1-M}, Y_{t+1-M}, \Delta_{t+1-M}), \ldots, (D_t, Y_t, \Delta_t). \quad (5)$$

Set $H_{t-M} = I$ and **only store the triples in (5).**

**Algorithm 1** Block L-BFGS Update (Two-loop Recursion)

**inputs:** $g_t \in \mathbb{R}^d, D_i, Y_i \in \mathbb{R}^{d\times q}$ and $\Delta_i \in \mathbb{R}^{q\times q}$ for $i \in \{t+1-M,\ldots,t\}$.
**initiate:** $v \leftarrow g_t$
**for** $i = t,\ldots,t-M+1$ **do**
  $\alpha_i \leftarrow \Delta_i D_i^T v, \quad v \leftarrow v - Y_i\alpha_i$
**end for**
**for** $i = t-M+1,\ldots,t$ **do**
  $\beta_i \leftarrow \Delta_i Y_i^T v, \quad v \leftarrow v + D_i(\alpha_i - \beta_i)$
**end for**
**output** $H_t g_t \leftarrow v$

## 6. Algorithm

**Algorithm 2** Stochastic Block BFGS Method

**inputs:** $w_0 \in \mathbb{R}^d$, stepsize $\eta > 0$, $q$ = sample action size, and length of inner loop $m$.
**initiate:** $H_{-1} = I$
**for** $k = 0,1,2,\ldots$ **do**
  Compute the full gradient $\mu = \nabla f(w_k)$
  Set $x_0 = w_k$
  **for** $t = 0,\ldots,m-1$ **do**
    Sample $S_t, T_t \subseteq [n]$, independently
    Compute variance-reduced stochastic gradient
      $g_t = \nabla f_{S_t}(x_t) - \nabla f_{S_t}(w_k) + \mu$
    Form $D_t \in \mathbb{R}^{d\times q}$ so that $\text{rank}(D_t) = q$
    Compute sketch $Y_t = \nabla^2 f_{T_t}(x_t) D_t$
    Compute $d_t = -H_t g_t$ via Algorithm 1
    Set $x_{t+1} = x_t + \eta d_t$
  **end for**
  **Option I:** Set $w_{k+1} = x_m$
  **Option II:** Set $w_{k+1} = x_i$, where $i$ is selected uniformly at random from $[m] = \{1,2,\ldots,m\}$
**end for**
**output** $w_{k+1}$

## 3. Hessian Sketching

**Fact:** Evaluating Hessian-vector products is cheap

$$\nabla^2 f_T(x_t)v = \frac{d}{d\alpha}\nabla f_T(x_t + \alpha v)\Big|_{\alpha=0} \qquad (2)$$

We would like $H_t$ to satisfy the *inverse equation*

$$H_t\nabla^2 f_T(x_t) = I,$$

but calculating the inverse of $d\times d$ matrix is expensive.

**Solution:** finding $H_t$ that satisfies a *sketched* version of inverse equation

$$H_t\nabla^2 f_T(x_t) D_t = D_t, \qquad (3)$$

is cheap (2), where $D_t \in \mathbb{R}^{d\times q}$ and $q \ll \min\{d,n\}$.

We employ three different sketching strategies:

**1) gauss.** $D_t$ has standard Gaussian entries sampled i.i.d at each iteration.

**2) prev.** Let $d_t = -H_t g_t$. Store search directions $D_t = [d_{t+1-q},\ldots,d_t]$ and update $H_t$ once every $q$ iterations.

**3) fact.** Sample $C_t \subseteq \{1,\ldots,d\}$ uniformly at random and set $D_t = L_{t-1}I_{:C_t}$, where $L_{t-1}L_{t-1}^T = H_{t-1}$ and $I_{:C_t}$ denotes the concatenation of the columns of the identity matrix indexed by a set $C_t \subset \{1,\ldots,d\}$.

## 4. Block BFGS Update

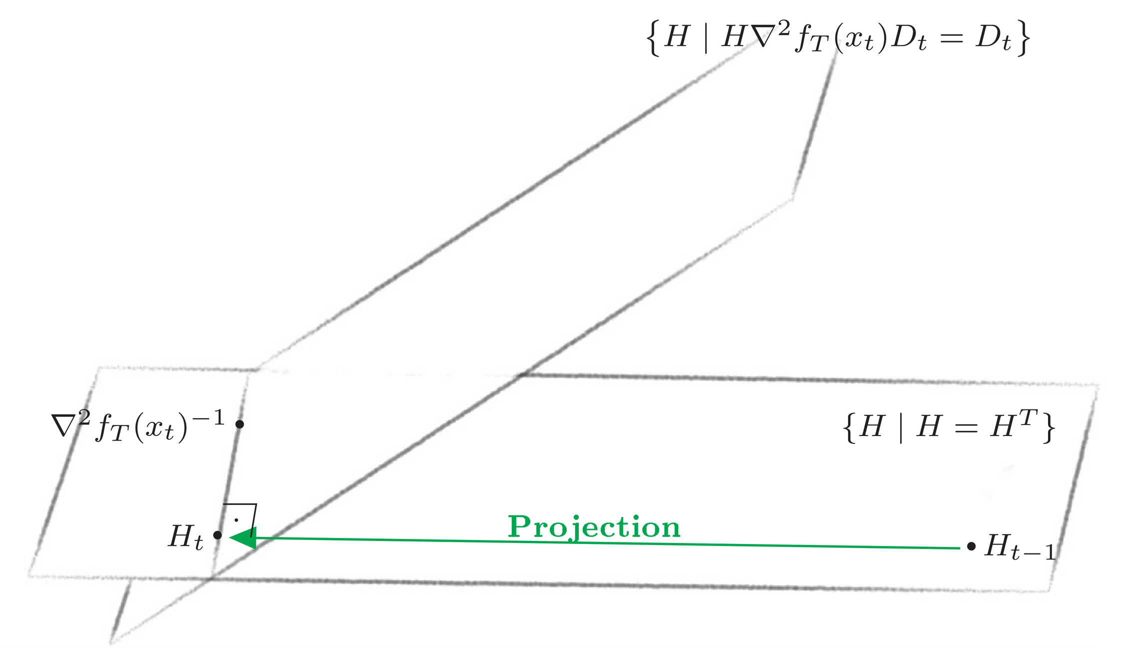
The sketched equation (3) is not enough to determine $H_t$ uniquely. So we make use of the following projection

$$H_t = \arg\min_{H\in\mathbb{R}^{d\times d}} \|H - H_{t-1}\|_t^2$$
$$\text{subject to } H\nabla^2 f_T(x_t)D_t = D_t, \ H = H^T, \qquad (4)$$

where $\|H\|_t^2 \overset{\text{def}}{=} \mathbf{Tr}\left(H\nabla^2 f_T(x_t)H^T\nabla^2 f_T(x_t)\right)$. The closed form solution of (4) is

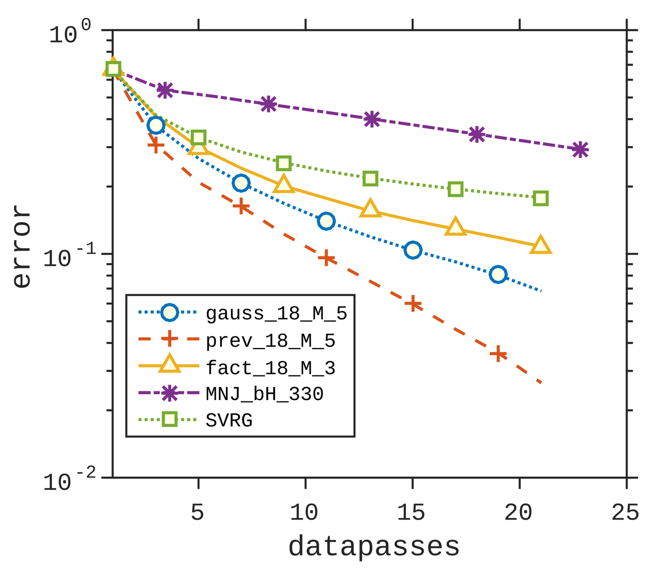
$$H_t = D_t\Delta_t D_t^T + \left(I - D_t\Delta_t Y_t^T\right)H_{t-1}(I - Y_t\Delta_t D_t),$$

where $\Delta_t = (D_t^T Y_t)^{-1}$ and $Y_t = \nabla^2 f_T(x_t)D_t$.

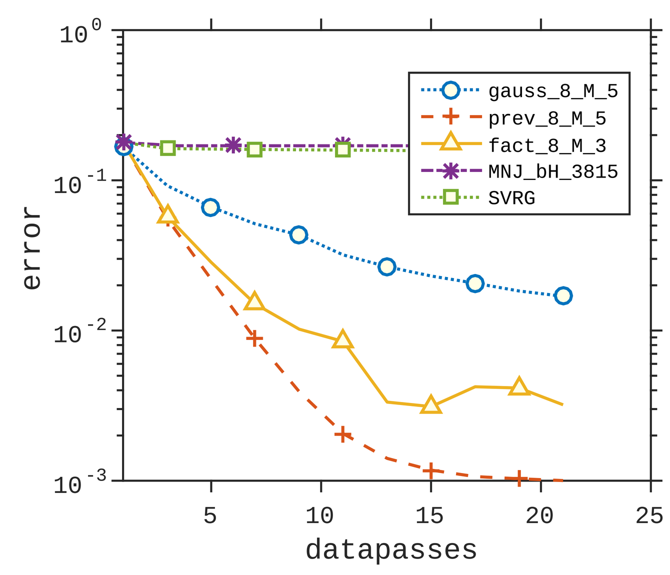



## 7. Tests on logistic loss with $L2$ regularizer

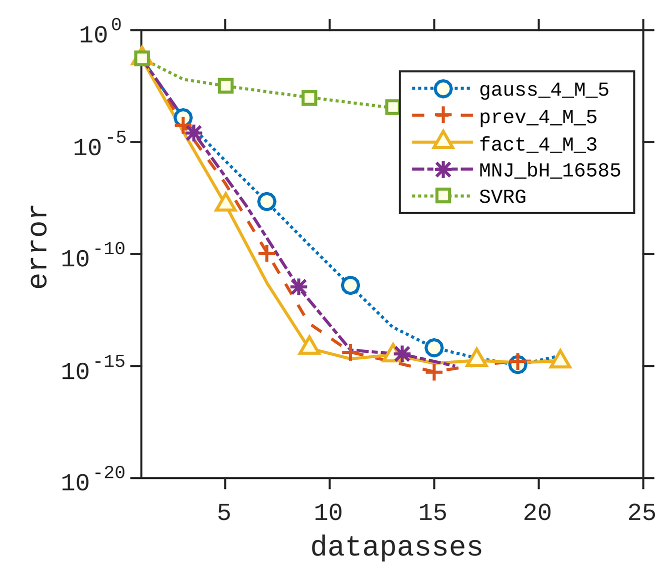
gisette $(n;d) = (6,000; 5,000)$



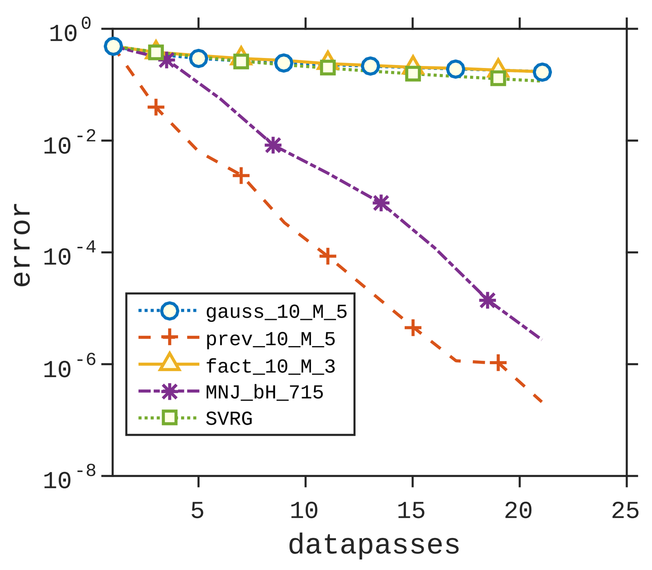

covtype $(n;d) = (581,012; 54)$



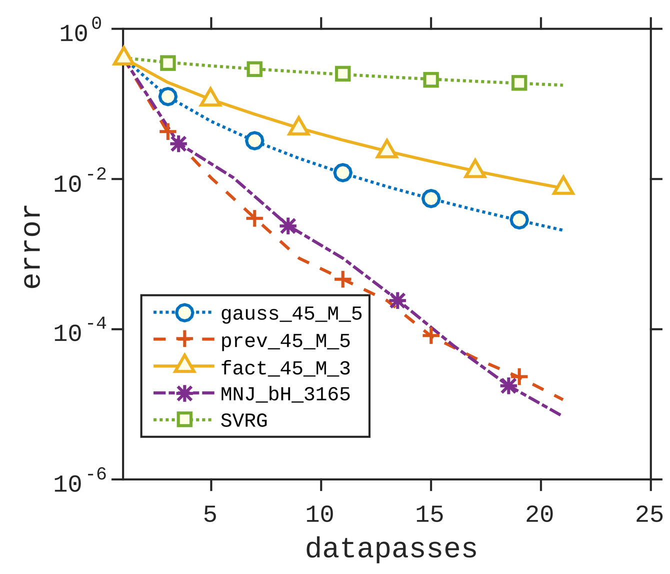

HIGGS $(n;d) = (11,000,000; 28)$



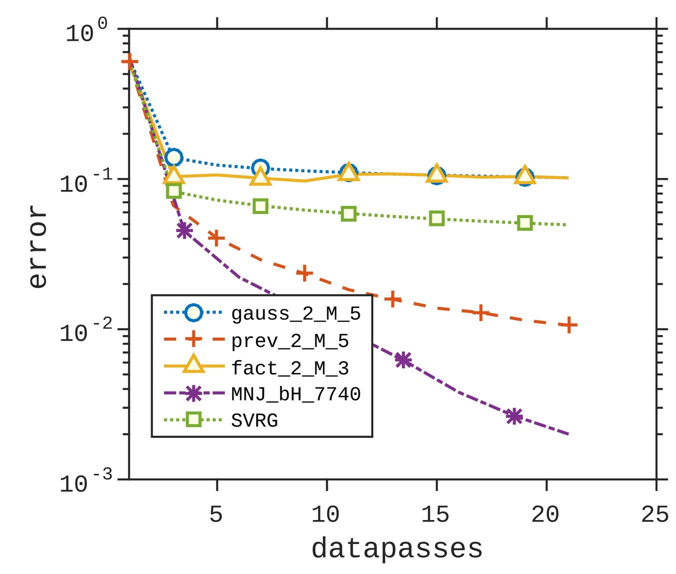

rcv1 $(n;d) = (20,242; 47,236)$




epsilon $(n;d) = (400,000; 2,000)$




url_comb $(n;d) \approx (2\times 10^6; 3\times 10^6)$




## 8. Convergence

**Assumption 1.** *There exist constants $0 < \lambda \le \Lambda$ such that*

$$\lambda I \preceq \nabla^2 f_T(x) \preceq \Lambda I \qquad (6)$$

*for all $x \in \mathbb{R}^d$ and all $T \subseteq [n]$.*

**Lemma 1.** *There exists $\Gamma \ge \gamma > 0$ such that*

$$\gamma I \preceq H_t \preceq \Gamma I \qquad \forall t, \qquad (7)$$

*where*

$$\frac{1}{1+M\Lambda} \le \gamma \le \Gamma \le (1+\sqrt{\kappa})^{2M}\left(1+\frac{1}{\lambda(2\sqrt{\kappa}+\kappa)}\right)$$

*and $\kappa \overset{def}{=} \Lambda/\lambda$.*

**Theorem 1.** *If we select parameters $m, \eta$ such that*

$$m \ge \frac{1}{2\eta(\gamma\lambda - \eta\Gamma^2\Lambda(2\Lambda - \lambda))}, \qquad \eta < \gamma\lambda/(2\Gamma^2\Lambda^2)$$

*then Algorithm 2 with Option II gives*

$$\mathbf{E}[f(w_k) - f(w_*)] \le \rho^k\mathbf{E}[f(w_0) - f(w_*)], \quad k \ge 0$$

*where the convergence rate is given by*

$$\rho = \frac{1/2m\eta + \eta\Gamma^2\Lambda(\Lambda-\lambda)}{\gamma\lambda - \eta\Gamma^2\Lambda^2} < 1.$$

## 9. Summary

We proposed a novel limited-memory stochastic block BFGS update for incorporating enriched curvature information in stochastic approximation methods. In our method, the estimate of the inverse Hessian matrix is updated at each iteration using a sketch of the Hessian. We presented three sketching strategies, a new quasi-Newton method that uses stochastic block BFGS updates combined with the variance reduction approach SVRG to compute batch stochastic gradients, and proved linear convergence of the resulting method.

## References

[1] R.M. Gower and P. Richtárik (2016), Randomized Quasi-Newton Updates are Linearly Convergent Matrix Inversion Algorithms, *arXiv:1602.01768*.

[2] R. Johnson and T. Zhang (2013). Accelerating stochastic gradient descent using predictive variance reduction, *NIPS*.

[3] P. Moritz, R. Nishihara and M. I. Jordan (2016). A Linearly-Convergent Stochastic L-BFGS Algorithm, *AISTATS*

[4] R.M. Gower and Jacek Gondzio Action constrained quasi-Newton methods, *1412.8045* 2014.