1. Problem
Find an approximate minima of
\[ \min_{x \in \mathbb{R}^d} f(x) \defeq \frac{1}{n} \sum_{i=1}^{n} f_i(x), \]  
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
\[ x_{i+1} = x_i - \eta_i H_i g_i, \]
where \( E[g_i] = \nabla f(x_i) \), \( H_i \approx \nabla^2 f(x_i) \), and \( \eta_i > 0 \) is a stepsize. To update \( g_i \) and \( H_i \), effective methods use only the subsampled gradient and subsampled Hessian
\[ \nabla f_i(x) \defeq \frac{1}{m} \sum_{j \in S} \nabla f_j(x), \quad \nabla^2 f_i(x) \defeq \frac{1}{m} \sum_{j \in S} \nabla^2 f_j(x) \]
where \( S,T \subseteq [n] \defeq \{1,2,\ldots,n\} \) selected uniformly at random.

**Challenge:** Use \( H_i \) using subsampled Hessians.

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

3. Hessian Sketching
**Fact:** Evaluating Hessian-vector products is cheap
\[ \nabla^2 f(x)w = \frac{d}{dx} \nabla f(x + \alpha w)|_{\alpha = 0} \]  
We would like \( H_i \) to satisfy the inverse equation
\[ H_i \nabla^2 f_i(x) = I, \]
but calculating the inverse of \( d \times d \) matrix is expensive.

**Solution:** Finding \( H_i \) that satisfies a sketched version of inverse equation
\[ H_i \nabla^2 f_i(x)D_i = D_i, \]
is cheap (2), where \( D_i \in \mathbb{R}^{d \times d} \) and \( q \ll \min(d,n) \).

We employ three different sketching strategies:

1. **gauss:** \( D_i \) has standard Gaussian entries sampled i.i.d. at each iteration.
2. **prev:** Let \( D_{i+1} = H_i g_i \). Store search directions \( D_{i+1} = \{d_{i+1}, \ldots, d_i\} \) and update \( H_i \) once every \( q \) iterations.
3. **fact:** Sample \( C_i \subseteq \{1, d \} \) uniformly at random and set \( D_{i+1} = L_{i+1} F_{i+1} \) where \( L_{i+1} \) and \( F_{i+1} \) denotes the concatenation of the columns of the identity matrix indexed by a set \( C_i \subseteq \{1, \ldots, d\} \).

4. Block BFGS Update
The sketched equation (3) is not enough to determine \( H_i \) uniquely. So we make use of the following projection
\[ H_i \rightarrow \arg \min_{H \in \mathbb{R}^{d \times d}} \|H - H_i \|_F^2 \]
such that \( H \nabla^2 f_i(x)D_i = D_i, \quad H = H^T \), (4)
where \( \|H\|_F \defeq \text{tr}(H \nabla^2 f_i(x)H^T \nabla^2 f_i(x)) \). The closed form solution of (4) is
\[ H_i = D_i L_i \Delta_i^{-1} + (I - D_i \Delta_i^{-1} L_i^T)H_{i-1}(I - Y_i \Delta_i^{-1} D_i), \]
where \( \Delta_i \defeq (D_i L_i)^{-1} \) and \( Y_i \defeq \nabla^2 f_i(x)D_i \).

5. Block L-BFGS update
Let \( V_1 = I - D_1 \Delta_1 Y_1^T \). Expanding \( M \) block BFGS updates applied to \( H_{i-M} \) gives
\[ H_i = V_{1} H_{i-1} V_{1}^T + D_i \Delta_i D_i^T \]
\[ = V_{1} \cdots V_{i-M} H_{i-M} V_{i-M+1} \cdots V_{1} H_{1-M+1} \cdots V_{1} \]
\[ + \sum_{t=1}^{i-M} V_1 \cdots V_{t+1} H_{t+1-M+1} \cdots V_{1} \]
Therefore \( H_i \) is a function of \( H_{i-M} \) and the triples \((D_{1-M}, Y_{1-M}, \Delta_{1-M}), \ldots, (D_{1}, Y_{1}, \Delta_{1})\). (5)
Set \( H_{i-M} = I \) and only store the triples in (5).

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

**inputs:** \( g_i, D_i, Y_i \in \mathbb{R}^{d \times d} \), \( H_i \in \mathbb{R}^{d \times d} \) for \( i \in \{1, \ldots, i-M\} \).

**initiate:** \( \eta \leftarrow g_0 \)

for \( i = 1, \ldots, i-M+1 \) do
\( \alpha_i \leftarrow \Delta_i D_i^T v_i \)
\( v_i \leftarrow v_i - Y_i \alpha_i \)
end for

for \( i = M, \ldots, i \) do
\( \beta_i \leftarrow \Delta_i^{-1} Y_i^T v_i \)
\( v_i \leftarrow v_i + D_i (\alpha_i - \beta_i) \)
end for

**output** \( H_i \leftarrow v \)

6. Algorithm
**Algorithm 2** Stochastic Block BFGS Method

**inputs:** \( w_0 \in \mathbb{R}^{d} \), stepsize \( \eta > 0 \), \( q \) = sample action size, and inner loop of size \( m \).

**initiate:** \( H_1 = I \)

for \( k = 0, 1, 2, \ldots \) do
Compute the full gradient \( \mu \defeq \nabla f(w_k) \)
Set \( x_0 = w_k \)
for \( t = 0, \ldots, m-1 \) do
Sample \( S_t \subseteq \{n\} \) independently
Compute variance-reduced stochastic gradient \( g_t = \nabla f_i(x_t) - \nabla f_{j}(w_{k}) + \mu \)
Form \( D_1 \in \mathbb{R}^{d \times d} \) so that \( \text{rank}(D_1) = q \)
Compute sketch \( Y_t \defeq \nabla^2 f_i(x_t)D_t \)
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 \( w_{k+1} = x_t \), where \( t \) is selected uniformly at random from \( \{1,2,\ldots,m\} \)

end for

**output** \( w_{k+1} \)

7. Tests on logistic loss with L2 regularizer

gisette \((n;d) = (6,000; 5,000)\)
covtype \((n;d) = (581,012; 54)\)
HIGGS \((n;d) = (11,000,000; 28)\)

8. Convergence

**Assumption 1.** There exist constants \( 0 < \lambda \leq \Lambda \)

such that
\[ M \leq \nabla^2 f(x) \leq M \]
for all \( x \in \mathbb{R}^d \) and all \( T \subseteq [n] \).

**Lemma 1.** There exists \( \gamma \geq 0 \) such that
\[ \gamma I \leq H_i \leq \gamma I \quad \forall i, \]
where
\[ \frac{1}{1 - \gamma M} \leq \gamma \leq \frac{1}{1 + \lambda^2} \]
and \( \kappa \defeq \lambda / \gamma \).

**Theorem 1.** If we set parameters \( m, \eta \) such that
\[ m \geq 2q \gamma \left( \frac{\lambda}{\lambda - \gamma} \right)^2 \]
then Algorithm 2 with Option II gives
\[ \frac{1}{2m \eta} \leq \nabla^2 f(w_{k}) \leq \frac{1}{2m \eta}, \quad k \geq 0 \]
where the convergence rate is given by \( 2m \eta \).

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


