1. Problem

Find an approximate minima of
\[ \min_{\mathbf{x} \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x), \]
where \( f_i : \mathbb{R}^d \rightarrow \mathbb{R} \) 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
\[ x_{t+1} = x_t - \eta H_t \nabla f(x_t), \]
where \( \mathcal{E}_{[t]} = \nabla f(x_t), H_t \approx \nabla^2 f(x_t) \), and \( \eta > 0 \) is a stepsize. To update \( g_t \) and \( H_t \), effective methods use only the subsampled gradient and subsampled Hessian
\[ \nabla f(x_t) \approx \frac{1}{|S|} \sum_{i \in S} \nabla f(x_i), \quad \nabla^2 f(x_t) \approx \frac{1}{|T|} \sum_{i \in T} \nabla^2 f_i(x), \]
where \( S,T \subseteq [n] \) and \( S \cap T = \emptyset \) uniformly at random.

Challenge: Choose \( H_t \) 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 \( V_t = I - \Delta_t \Delta_t^T \). Expanding \( M \) block BFGS updates applied to \( H_{t-M} \) gives
\[ H_t = V_t H_{t-M} V_t^T + \Delta_t \Delta_t^T = V_{t-1} \cdots V_{t-M} H_{t-M} V_{t-M+1}^T \cdots V_t^T + \Delta_t \Delta_t^T, \]
Therefore \( H_t \) is a function of \( H_{t-M} \) and the triples \((D_{t-M}, V_{t-M+1} \cdots V_t, \Delta_t)\). (5)
Set \( H_{t-M} = I \) and only store the triples in (5).

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

inputs: \( g_t \in \mathbb{R}^d, D_t, Y_t \in \mathbb{R}^{d \times d} \) and \( \Delta_t \in \mathbb{R}^{d \times d} \) for \( t \in \{t+1 \ldots T, t\}. \)
initiate: \( v_t = -g_t \)
for \( i = t \ldots t-M + 1 \) do
\[ v_i = \Delta_t D_t v_i, \quad v_t = -Y_t v_t, \]
end for
for \( i = t-M+1 \ldots t \) do
\[ \beta_i = \Delta_t^T Y_t v_t, \quad v_t = -D_t (v_t - \beta_i), \]
end for
output: \( H_t g_t = v_t \)

6. Algorithm

Algorithm 2 Stochastic Block BFGS Method

inputs: \( w_0 \in \mathbb{R}^d \), stepsize \( \eta > 0 \), \( q \) is sample action size, and length of inner loop \( m \).
initiate: \( H_0 = \frac{1}{m} \sum_{i=1}^{m} \nabla^2 f(w_i) \)
for \( k = 0, 1, 2 \ldots \) do
Compute the full gradient \( \mu = \nabla f(w_k) \)
Set \( x_0 = w_k \)
for \( r = 0 \ldots m-1 \) independently
Sample \( S_r, T_r \subseteq [n] \)
Compute variance-reduced stochastic gradient \( g_{0} = \nabla f_S(x_0) + \mu \)
Form \( D_r \in \mathbb{R}^{d \times d} \) so that \( \mathcal{E}_{[r]} = \nabla f \)
Compute sketch \( Y_r = \frac{1}{|T_r|} \sum_{i \in T_r} \nabla^2 f_i(x_0)D_r \)
Compute \( d_r = H_0 g_r \) via Algorithm 1
Set \( x_{r+1} = x_0 + \eta d_r \)
end for
Option I: Set \( w_{k+1} = x_m \)
Option II: Set \( w_{k+1} = x_r \), where \( r \) is selected uniformly at random from \([m] = \{1, 2, \ldots, m\}\)
end for
output: \( w_{k+1} \)

8. Convergence

Assumption 1. There exist constants \( 0 < \lambda \leq \Lambda \) such that
\[ \mathcal{M} \leq \nabla^2 f(x) \leq \mathcal{M} \]
for all \( x \in \mathbb{R}^d \) and \( \mathcal{T} \subseteq [n] \).
Lemma 1. There exists \( \lambda > \gamma > 0 \) such that
\[ \gamma I \leq H_t \leq \lambda I \quad \forall t, \]
where
\[ \frac{1}{1 + \mathcal{M} A} \leq \gamma \leq \Gamma \leq (1 + \sqrt{2})^2 \left( 1 + \frac{1}{2 \sqrt{\pi} + \epsilon} \right), \]
and \( \kappa \triangleq \lambda / \gamma \).

Theorem 1. If we select parameters \( m, \eta \) such that
\[ m \geq 4 \gamma^{-1} \left( \lambda - \sqrt{2} \mathcal{M} (2 \lambda - \lambda) \right), \quad \eta < \gamma / (2 \sqrt{\mathcal{M} A}), \]
then Algorithm 2 with Option II has
\[ E \left[ f(w_{k+1}) - f(w_k) \right] \leq \rho E \left[ f(w_k) - f(w_0) \right], \quad \rho \geq 0 \]
where the convergence rate is given by
\[ \rho = \frac{1}{2m} \eta - \eta^2 \lambda^2 (\Lambda - \lambda) - \eta^2 \lambda^{-2}. \]

4. Block BFGS Update

The sketched equation (3) is not enough to determine \( H_t \) uniquely. So we make use of the following projection
\[ H_t = \arg \min_{H \in \mathbb{R}^{d \times d}} \| H - H_{t-1} \|_F^2 \]
such that \( H^2 = H_{t-1}^2 + H - H_{t-1} \) where \( H_{t-1} = \Delta_t \Delta_t^T \). The closed form solution of (4) is
\[ H_t = \Delta_t \Delta_t^T + \left( I - \Delta_t \Delta_t^T \right) H_{t-1} \left( I - \Delta_t \Delta_t^T \right) \]
where \( \Delta_t = \left( \Delta_t^2 \right)^{1-t} \) and \( Y_t = \nabla^2 f(x_t) \Delta_t \).

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