Sharpened Lazy Quasi-Newton Method

Aakash Lahoti*, Spandan Senapati*, Ketan Rajawat, Alec Koppel

Overview

- SLIQN is the **first** incremental Quasi-Newton method with an **explicit superlinear rate,** an $O(d^2)$ cost and a **superior empirical performance** over other methods.
- Past works either have (a) an asymptotic convergence rate, or (b) $O(d^3)$ cost, which is prohibitively large for high dimensional problems.

Quasi-Newton Methods

Quasi-Newton methods are Newton-like algorithms that use an easy-to-invert Hessian approximation to take descent steps. This reduces the cost for $O(d^3)$ to $O(d^2)$.

$$x^{t+1} = x^t - (B^t)^{-1}\nabla f(x^t)$$

x^t is the current iterate, B^t is the Hessian approximation. Let $K^t = \int_0^1 \nabla^2 f(x^t + (x^{t+1} - x^t)\lambda)\, d\lambda$ and the descent direction $u = x^{t+1} - x^t$ then,

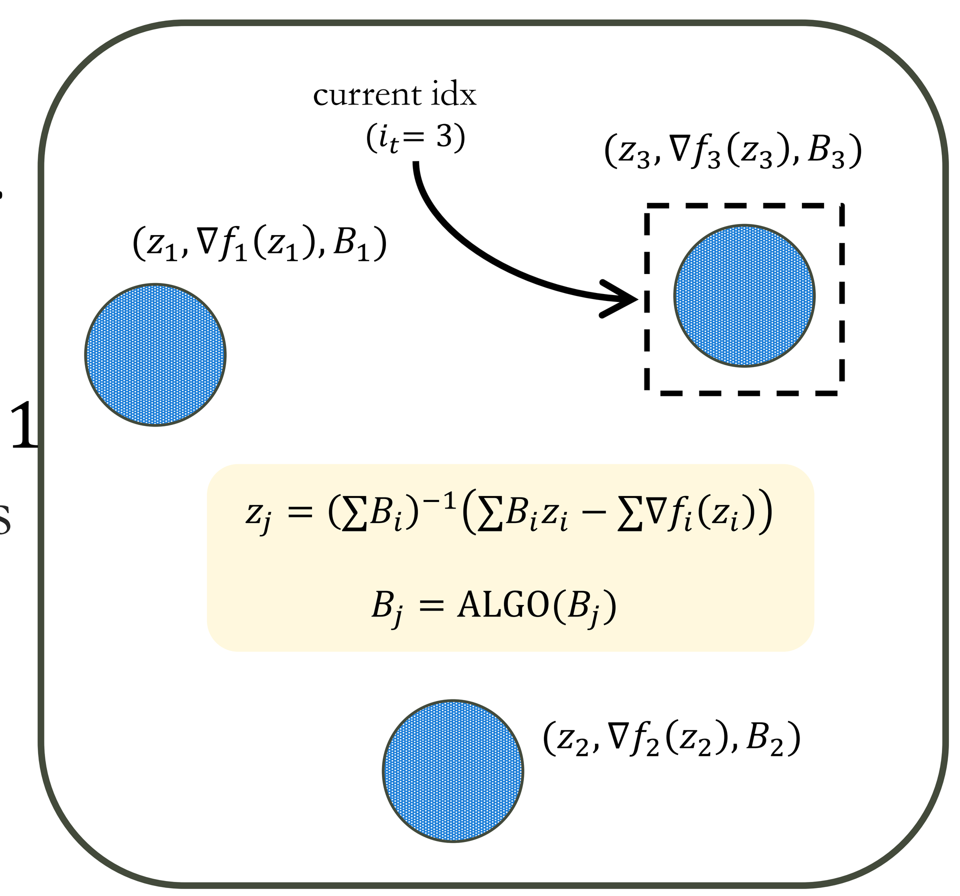
$$B^{t+1} \coloneqq \text{BFGS}(B^t, K^t, u) = B^t - \frac{B^t u u^T B^t}{u^T B^t u,} + \frac{K^t u u^T K^t}{u^T K^t u}$$

IQN Framework

An incremental approach to Quasi Newton methods:

1. Each sample 'i' stores the tuple $(z_i, \nabla f_i(z_1), B_i)$.

2. Iteration t procedure:
2a) Set $j = (t-1)\%n + 1$
2b) Compute $z_j, \nabla f(z_j)$ as shown to the right
2c) Compute B_j using custom ALGO module



Main Result

Under the assumptions of smoothness, strong convexity of the functions and the Lipschitz continuity of the Hessian, and if the initial iterate x^0, and the initial Hessian approximation B_i^0 are close enough to x^*, and $\nabla^2 f_i(x^0)$ respectively, then

$$\| x^t - x^\star \| \leq \zeta^{\lceil \frac{t-1}{n} \rceil} \qquad \zeta^k \leq \left(1 - \frac{\mu}{dL}\right)^{\frac{(k+1)(k+2)}{2}}$$

SLIQN's ALGO Module

SLIQN applies a scaled classic BFGS update followed by a Greedy BFGS update to obtain the Hessian approximation

$$Q^t = \text{BFGS}\left((1+\beta_t)^2 B_{i_t}^{t-1}, (1+\beta_t)K^t, z_{i_t}^t - z_{i_t}^{t-1}\right)$$

$$B_{i_t}^t = \text{BFGS}\left(Q^t, \nabla^2 f_{i_t}(z_{i_t}^t), \bar{u}(Q^t, \nabla^2 f_{i_t}(z_{i_t}^t))\right)$$

where, $\bar{u}(B,K) = \max_j \frac{e_j^T B e_j}{e_j^T K e_j}$ is greedy vector, and β_t is scaling factor.

Proof Sketch

Let $\xi(Z_i^{t-1}) \coloneqq \|z_i^{t-1} - x^\star\|$, $\xi(B_i^{t-1}) \coloneqq \|B_i^{t-1} - \nabla^2 f(z_i^{t-1})\|$. Then $\xi(Z_{i_t}^t)$ can be bounded by the previous n residuals as,

One-Step Inequality (Simplified)

$$\xi(Z_{i_t}^t) = \mathcal{O}\left(\sum_{i=1}^{n} \xi^2(Z_i^{t-1}) + \xi(Z_i^{t-1}) \cdot \xi(B_i^{t-1})\right)$$

Linear Convergence (Informal)

The iterates of SIQN are locally convergent and satisfy:
$\xi(z_{i_t}) = \mathcal{O}\left(\rho^{\frac{t}{n}}\right)$ and $\sigma\left(B_{i_t}^t, \nabla^2 f(z_{i_t}^t)\right) = \mathcal{O}\left(\rho^{\frac{t}{n}}\right)$, where $\sigma(\cdot)$ is some metric on the space of matrices, and $\rho \in (0,1)$.

The above result is derived using an induction on t, wherein the closeness conditions are used to bound the relevant terms in the the one-step inequality

Mean Superlinear Convergence (Informal)

$$\|x^t - x^\star\| = \mathcal{O}\left(\left(1 - \frac{\mu}{dL}\right)^{\frac{t}{n}} \cdot \frac{1}{n}\sum_{i=1}^{n}\|x^{t-i} - x^\star\|\right)$$

The proof follows by substituting the linear convergence result back into the one-step inequality.

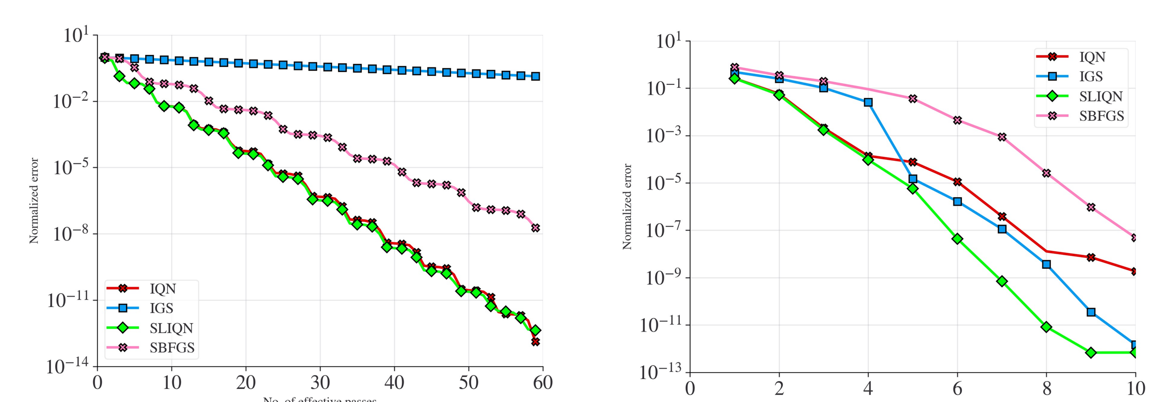
Superlinear Convergence (Informal)

$$\|x^t - x^\star\| = \mathcal{O}\left(\left(1 - \frac{\mu}{dL}\right)^{\frac{t^2}{n^2}}\right)$$

The proof follows from the mean superlinear convergence result. Specifically, we show that $\|x^t - x^\star\| \leq \zeta^t$, where ζ_t is a sequence which is defined by the recursion $\zeta_t \leq \left(1 - \frac{\mu}{dL}\right)^{t+1} \zeta_{t-1}$.

Experiments

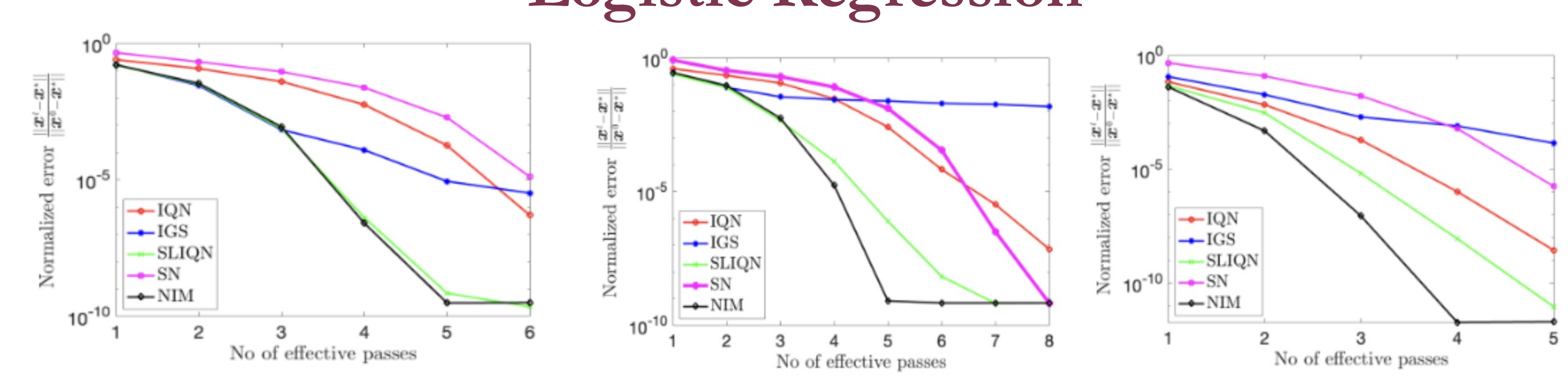
Quadratic Minimization



$n = 20, d = 500$ $\qquad$ $n = 50000, d = 10$

IQN descends fast in earlier epochs, while IGS descends fast in the later ones; SLIQN combines the best of both worlds!

Logistic Regression



a9a $\qquad$ w8a $\qquad$ Protein