1. High Dimensional Optimization

Consider the optimization problem

$$x_{\ast} = \arg \min_{x \in \mathbb{R}^d} f(x),$$

where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is $L$-smooth and $d$ is very big. This arises in training ML models with a very large number of parameters, or when data is high dimensional and acquiring data is expensive/hard.

Example: genomics, seismoology, neurology and high resolution sensors in medicine.

Notation:
- Gradient & Hessian: $g(x) = \nabla f(x)$, $H(x) = \nabla^2 f(x)$
- Level set: $\mathcal{Q} = \{x \in \mathbb{R}^d : f(x) \leq f(x_0)\}$
- Hessian inner product: $\langle u, v \rangle_{H(x)} = \langle H(x)u, v \rangle$

2. Assumptions (New)

Assumption 1: Gradient invariance: $g(x) \in \text{Range}(H(x))$ for all $x \in \mathbb{R}^d$.

Assumption 2: $f$ is $L$-smooth and $\mu$-convex relative to its Hessian. That is, there exist $L \geq \mu > 0$ such that for all $x, y \in \mathbb{Q}$,

$$f(x) \leq f(y) + \langle g(y), x - y \rangle + \frac{L}{2} \|x - y\|_H^2,$$

$$f(x) \geq f(y) + \langle g(y), x - y \rangle + \frac{\mu}{2} \|x - y\|_H^2.$$  

This is a weak assumption since:
- $L$-smoothness $\Rightarrow$ $\mu$-stability [1] $\Rightarrow$ $L$-smoothness $\mu$-convexity.

Example: Both assumptions hold for smooth generalized linear models with $L_2$ regularization.

3. Newton’s Method

Newton’s method applied to problem (1) has the form

$$x_{k+1} = x_k - H(x_k)^{-1} \nabla f(x_k).$$

where
- $\gamma > 0$ is the stepsize
- $H(x_k)$ is the Moore-Penrose pseudoinverse of $H(x_k)$

Pros: Can handle curvature, invariant to coordinate transformations

Cons: Cost of each iteration is very high: $O(d^3)$

4. Sketching and Dimension Reduction

Let $S \in \mathbb{R}^{d \times r}$ be a random matrix drawn from $S \sim \mathcal{D}$.

$$S^\top \mathbb{E} f(x) S = \mathbb{E} f(x),$$

where

$$S^\top \mathbb{E} f(x) = \mathbb{E} f(x).$$

Assumption 3: With probability 1, the sketching matrix $S$ satisfies

$$\text{Null}(S^\top H(x) S) = \text{Null}(S), \quad \forall x \in \mathcal{Q}.$$  

5. Randomized Subspace Newton

Algorithm 1: Randomized Subspace Newton

1. input: $x_0 \in \mathbb{R}^d$
2. parameters: $D$ = distribution over random matrices
3. for $k = 0, 1, 2, \ldots$ do
4. Sample a fresh sketching matrix $S_k \sim \mathcal{D}$
5. $x_{k+1} = x_k - S_k^\top H(x_k) S_k S_k^\top g(x_k)$
6. end for

Output: last iterate $x_k$

Computation of sketched Newton direction:

$$S_k^\top H(x_k) S_k S_k^\top g(x_k) \rightarrow S_k^\top H(x_k) S_k S_k^\top g(x_k) \rightarrow S_k^\top H(x_k) S_k S_k^\top g(x_k)$$

Can be computed with directional derivatives:

$$\frac{d}{d \lambda} f(S_k^\top g(x_k) S_k^\top \lambda) \bigg|_{\lambda = 0} = S_k^\top H(x_k) S_k S_k^\top g(x_k).$$

Advantages of RSN:
- Uses second-order information & hence enjoys better dependence on condition number
- Enjoy global convergence theory
- Is a descent method: $f(x_{k+1}) \leq f(x_k)$
- Is a feasible method: $x_k \in \mathcal{Q}$ for all $k \geq 0$
- Applicable for very large $d$

Example: Single Column Sketches

Let $0 \leq U \in \mathbb{R}^{d \times d}$ be a symmetric positive definite matrix such that $H(x) \preceq U$, $\forall x \in \mathcal{R}$. Let $M = [m_1, \ldots, m_d] \in \mathbb{R}^{d \times d}$ be an invertible matrix such that $m_i H(x), m_i \neq 0$ for all $x \in \mathcal{Q}$ and $i = 1, \ldots, d$. If we sample according to $m_i U_m$, the $i$-th column of $U_m$ then the update on line 5 of Algorithm 1 is given by

$$x_{k+1} = x_k - \frac{m_i g(x_k)}{\langle m_i g(x_k), m_i \rangle} \frac{1}{\text{Trace}(M U_m)}$$

with probability $p_i$, (6) costs $O(d)$ and has linear iteration complexity (10) given by

$$k \geq \max_{x \in \mathcal{Q}} \frac{1}{\text{Trace}(M U_m) \text{Min} \{H(x)^{\top} M H(x)^{\top} \} \| \log_{\beta} \frac{1}{\beta} \|}$$

Sufficient Condition for $p > 0$

If (5) holds and $R(K(x)) \subset R(S_k H(x_k) S_k^\top)$, then $\rho > 0$, and

$$\rho = \max_{d} \left\{ \frac{\text{Trace}(M U_m)}{\text{Min} \{H(x)^{\top} M H(x)^{\top} \}} \right\}$$

Example: Generalized Linear Models

Let $0 \leq u \leq f$. Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be a twice differentiable function such that $u = \phi(\beta)$ for $i = 1, \ldots, n$. Let $a \in \mathbb{R}^d$ for $i = 1, \ldots, n$ and $A = [a_1, \ldots, a_n] \in \mathbb{R}^{d \times n}$. We say that $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a generalized linear model when $f(x) = \frac{1}{2} \sum_{i=1}^{n} \phi(a_i^\top x) + \frac{1}{2} \|x\|_2^2$.

$\phi$ is $L$-smooth and $\mu$-convex relative to its Hessian with

$$L = \text{trace}(A^\top A) + \mu I$$

and

$$\mu = \text{trace}(A^\top A) + \lambda I$$

RSN has iteration complexity (10) given by

$$k \geq \frac{1}{\mu} \left( \text{trace}(A^\top A) + \lambda I \right)^2 \log \frac{1}{\epsilon}.$$  

8. Experiments

We compare RSN to Gradient descent (GD), accelerated gradient descent (AGD) [2] and full Newton method. For RSN we use coordinate sketches defined by $S_k \in \{0,1\}^{d \times d}$, with exactly one non-zero entry per row and per column of $S_k$.

References

Global linear convergence of Newton’s method without strong-convexity or Lipschitz gradients.
Introductory Lectures on Convex Optimization: A Basic Course.