日本オペレーションズ・リサーチ学会
2022年 秋季研究発表会

Randomized subspace regularized Newton method for unconstrained non-convex optimization

Tokyo University — Terunari Fuji
RIKEN-AIP — Pierre-Louis Poirion*
01308490 Tokyo University/RIKEN-AIP — Akiko Takeda

1. Introduction

In this paper, we develop a Newton-type iterative method with random projections for the following unconstrained optimization problem:

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is a non-convex twice differentiable function. In our method, at each iteration, we restrict the function f to a random subspace and compute the next iterate by choosing a descent direction on this random subspace. When f is convex, Randomized Subspace Newton (RSN) is introduced in [1]. At each iteration, it computes the descent direction d_k^{RSN} and the next iterate as

$$d_k^{\mathrm{RSN}} = -P_k^\top (P_k\nabla^2 f(x_k)P_k^\top)^{-1}P_k\nabla f(x_k),$$
$$x_{k+1} = x_k + \frac{1}{\hat{L}}d_k^{\mathrm{RSN}},$$

where $P_k\in\mathbb{R}^{s\times n}$ is a random matrix and $\hat{L}$ is some fixed constant. RSN is expected to be highly computationally efficient, with respect to the original Newton method since it does not require computation of the inverse of the full Hessian. If the objective function f is not convex, the Hessian is not always positive semidefinite and d_k^{RSN} is not guaranteed to be a descent direction so that we need to use a modified Hessian. Based on the regularized Newton method (RNM) for the unconstrained non-convex optimization [2, 3], we propose the randomized subspace regularized Newton method (RS-RNM):

$$d_k = -P_k^\top (P_k\nabla^2 f(x_k)P_k^\top + \eta_k I_s)^{-1}P_k\nabla f(x_k),$$
$$x_{k+1} = x_k + t_k d_k,$$

where η_k is defined to ensure that search direction d_k is a descent direction and the step size t_k is chosen so that it satisfies Armijo's rule. As with RSN, this algorithm is expected to be computationally efficient since we use projections onto lower-dimensional spaces. In this paper, we show that RS-RNM has global convergence under appropriate assumptions, more precisely, we have $\|\nabla f(x_k)\|\le\varepsilon$ after at most $O(\varepsilon^{-2})$ iterations with some probability. We will then prove that under additional assumptions, we can obtain a linear convergence rate locally. In particular, the conditions we obtain are, to the best of our knowledge, the weakest conditions until now. From the weakest conditions, we can derive a random-projection version of the PL inequality. We will then prove that linear convergence is the best rate we can hope for this method.

2. Randomized subspace regularized Newton method

Let $\mathcal{D}$ be a distribution over random projection matrices of size $s\times n$. With a Gaussian random matrix P_k from $\mathcal{D}$, the regularized sketched Hessian:

$$M_k = P_k\nabla^2 f(x_k)P_k^\top + \eta_k I_s \in \mathbb{R}^{s\times s},$$

where $\eta_k := c_1\Lambda_k + c_2\|\nabla f(x_k)\|^\gamma$ and where $\Lambda_k := \max(0, -\lambda_{\min}(P_k\nabla^2 f(x_k)P_k^\top))$, is computed. We then compute the search direction:

$$d_k = -P_k^\top M_k^{-1}P_k\nabla f(x_k). \tag{2}$$

The costly part of Newton-based methods, the inverse computation of a (approximate) Hessian matrix, is done in the subspace of size s. We note that d_k defined by (2) is a descent direction for f at x_k, since it turns out that M_k is positive definite from the definition of Λ_k, and therefore $P_k^\top M_k^{-1}P_k$ is also positive definite.

The backtracking line search with Armijo's rule finds the smallest integer $l_k\ge 0$ such that

$$f(x_k) - f(x_k+\beta^{l_k}d_k) \ge -\alpha\beta^{l_k}g_k^\top d_k. \tag{3}$$

Starting with $l_k=0$, l_k is increased by $l_k\leftarrow l_k+1$ until the condition (3) holds.

3. Global convergence

The following theorem asserts that when the Hessian is Lipschitz continuous and the level sets of f are bounded, then we have that with high probability, $\|\nabla f(x_k)\| \leq \varepsilon$ after at most $O(\varepsilon^{-2})$ iterations.

Theorem 1 *Assume that the level set of f at the initial point x_0 is bounded and that the Hessian is Lipschitz continuous. Then, with probability at least $1 - 2m\left(\exp(-\frac{\mathcal{C}_0}{4}s) - \exp(-s)\right)$, we have*

$$\sqrt{\frac{f(x_0) - f^*}{mp}} \geq \min_{k=0,1,\ldots,m-1} \|\nabla f(x_k)\|,$$

where p is constant that depends on $\frac{n}{s}$ and on some parameters of the function f.

4. Local convergence

In this section, we investigate local convergence properties of the sequence $\{x_k\}$ assuming that it converges to a strict local minimal $\bar{x}$. First we will show that the sequence converges locally linearly to the strict local minima. Then we will prove that when f is strongly convex, we cannot aim at local super-linear convergence using random subspace. We make the following assumptions

Assumption 1 *We have that $s = o(n)$*

Assumption 2 *We assume that*

(i) There exists $\sigma > 0$ such that $r = rank(\nabla^2 f(\bar{x})) \geq \sigma n$

(ii) There exists $0 < \rho < 3$ and $\tilde{C}$ such that in a neighborhood of $\bar{x}$, $f(x_k) - f(\bar{x}) \geq \tilde{C}\|x_k - \bar{x}\|^\rho$ holds.

Assumption 3 *We assumed that*

$$(\mathcal{C} + 2)^2 s < n.$$

The next Theorem proves that under some assumptions, the sequence $\{f(x_k) - f(\bar{x})\}$ converge linearly locally to 0.

Theorem 2 *Assume that Assumptions 1, 2, 3 hold, that the Hessian is Lipschitz continuous and that the level set of f at the initial point x_0 is bounded. There exists $0 < \kappa < 1$ and $k_0 \in \mathbb{N}$ such that if $k \geq k_0$, then*

$$f(x_{k+1}) - f(\bar{x}) \leq \kappa(f(x_k) - f(\bar{x})).$$

holds with probability at least $1 - 6(\exp(-s) + \exp(-\frac{\mathcal{C}_0}{4}s))$.

We now restrict to the case where f is locally strictly convex.

Theorem 3 *Assume that Assumption 3 holds and that the level set of f at the initial point x_0 is bounded. There exists a constant $c > 0$ such that for k large enough,*

$$\|x_{k+1} - x^*\| \geq c\|x_k - x^*\|,$$

holds with probability at least $1 - 2\exp(-\frac{\mathcal{C}_0}{4}) - 2\exp(-s)$. Furthermore this implies the existence of a constant $c' > 0$ such that

$$f(x_{k+1}) - f(\bar{x}) \geq c'(f(x_k) - f(\bar{x})).$$

[1] R. Gower, D. Kovalev, F. Lieder, and P. Richtárik. RSN: Randomized Subspace Newton. *Adv. Neural Inf. Process. Syst.*, 32:616–625, 2019.

[2] K. Ueda and N. Yamashita. Convergence properties of the regularized newton method for the unconstrained nonconvex optimization. *Appl. Math. Optim.*, 62(1):27–46, 2010.

[3] K. Ueda and N. Yamashita. A regularized newton method without line search for unconstrained optimization. *Comput. Optim. Appl.*, 59(1-2):321–351, 2014.