Randomized Subspace Newton Method for Unconstrained Non-Convex Optimization

PRAIRIE -RIKEN Workshop

Pierre-Louis Poirion (RIKEN-AIP)
joint work with Terunari Fuji and Akiko Takeda

March 19, 2023

Overview

The gist

Non-convex unconstrained minimization

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable

The gist

Non-convex unconstrained minimization

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable

Subspace optimization

$$\min_{u \in \mathbb{R}^s} f(x + P^\top u),$$

where $P \in \mathbb{R}^{s \times n}$ is a random matrix.

The gist

Non-convex unconstrained minimization

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable

Subspace optimization

$$\min_{u \in \mathbb{R}^s} f(x + P^\top u),$$

where $P \in \mathbb{R}^{s \times n}$ is a random matrix.

- Can we speed up the computation time?
- Global and local convergence properties?

Previous works

Random Subspace Newton (RSN) [Gower et al., 2019](f is convex)

By computing the Newton direction on the function $u \mapsto f(x_k + P_k^\top u_k)$, we obtain $u_k = -(P_k \nabla^2 f(x_k) P_k^\top)^{-1} P_k \nabla f(x_k)$, hence

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

They prove global sub-linear convergence and local linear convergence if f is strongly convex.

Previous works

Random Subspace Newton (RSN) [Gower et al., 2019](f is convex)

By computing the Newton direction on the function $u \mapsto f(x_k + P_k^\top u_k)$, we obtain $u_k = -(P_k \nabla^2 f(x_k) P_k^\top)^{-1} P_k \nabla f(x_k)$, hence

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

They prove global sub-linear convergence and local linear convergence if f is strongly convex.

- [Hanzely et al., 2020]: Cubically-regularized subspace Newton method.
- [Kovalev et al., 2020]: random subspace version of the BFGS method.
- [Roberts and Royer, 2022]: probabilistic direct-search method in reduced random spaces (non-convex problems). The authors prove sub-linear convergence.

Our work

Based on regularized Newton method (RNM) for the unconstrained non-convex optimization [Ueda and Yamashita, 2010], we propose the randomized subspace regularized Newton method (RS-RNM):

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

where η_k is defined to ensure $P_k \nabla^2 f(x_k) P_k^\top + \eta_k I_s \succ 0$ and t_k satisfies Armijo's rule.

Our work

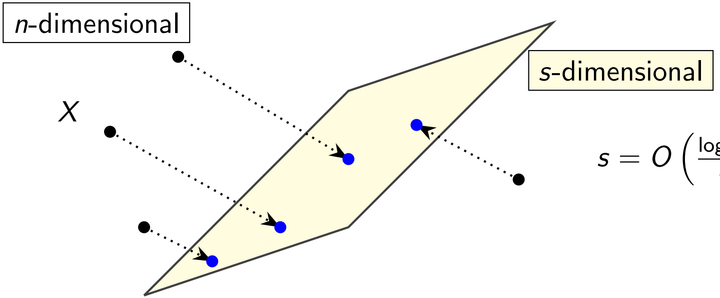
Based on regularized Newton method (RNM) for the unconstrained non-convex optimization [Ueda and Yamashita, 2010], we propose the randomized subspace regularized Newton method (RS-RNM):

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

where η_k is defined to ensure $P_k \nabla^2 f(x_k) P_k^\top + \eta_k I_s \succ 0$ and t_k satisfies Armijo's rule.

- In [Ueda and Yamashita, 2010] the authors prove global sub-linear convergence and local quadratic convergence under local-error bound condition.
- Can we extend these results to the random subspace setting ?

What is Random Projection



$$s = O\left(\frac{\log|X|}{\varepsilon^2}\right)$$

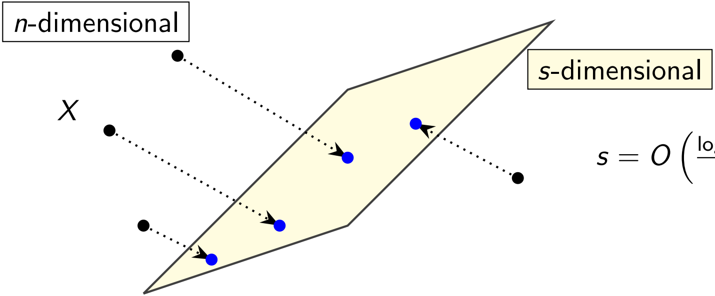
Random Projection

Lemma JLL

Let $P \in \mathbb{R}^{d \times n}$, $P_{ij} \sim \mathrm{N}(0, 1/s)$, i.i.d..
Then for any $x \in \mathbb{R}^n$ and $\varepsilon \in (0, 1)$, we have

$$\mathrm{Prob}\ [(1-\varepsilon)\|x\|_2^2 \le \|Px\|_2^2 \le (1+\varepsilon)\|x\|_2^2] \ge 1 - 2\exp(-\mathcal{C}\varepsilon^2 s),$$

where $\mathcal{C}$ is an absolute constant.



$$s = O\left(\frac{\log|X|}{\varepsilon^2}\right)$$

Concentration inequality for random matrices

$$P \in \mathbb{R}^{s \times n}$$

Proposition

There exists a constant $\mathcal{C}_1 > 0$s such that:

$$\left\| \frac{1}{n} PP^\top - I_s \right\| \leq \mathcal{C}_1 \frac{s}{n},$$

holds with probability at least $1 - 2\exp(-s)$.

Why is it useful ?

Remember that

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

Why is it useful ?

Remember that

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

Therefore, with high probability,

$$d_k = 0 \quad \iff \quad \nabla f(x_k) = 0.$$

Algorithm 1 Randomized subspace regularized Newton method (RS-RNM)

input: $x_0 \in \mathbb{R}^n$, $\gamma \geq 0, c_1 > 1, c_2 > 0, \alpha, \beta \in (0, 1)$

$k \leftarrow 0$

repeat

  sample a random matrix: $P_k \sim$ Gaussian matrix $\mathcal{N}(0, 1/s)^{s \times n}$

  compute the regularized sketched hessian:
  $M_k = P_k \nabla^2 f(x_k) P_k^\top + c_1 \Lambda_k I_s + c_2 \|\nabla f(x_k)\|^\gamma I_s$, where $\Lambda_k = \max(0, -\lambda_{\min}(P_k \nabla^2 f(x_k) P_k^\top))$

  compute the search direction: $d_k = -P_k^\top M_k^{-1} P_k \nabla f(x_k)$

  apply the backtracking line search with Armijo's to compute $l_k \geq 0$ such that (1) holds. Set $t_k = \beta^{l_k}$, $x_{k+1} = x_k + t_k d_k$ and $k \leftarrow k + 1$

until the stopping criteria is satisfied

return the last iterate x_k

$$f(x_k) - f(x_k + \beta^{l_k} d_k) \geq -\alpha \beta^{l_k} g_k^\top d_k. \quad (1)$$

Global convergence

Assumption (1)

The level set of f *at the initial point* x_0 *is compact, i.e.,*
$\Omega := \{\mathbb{R}^n : f(x) \leq f(x_0)\}$ *is compact.*

Global convergence

Assumption (1)

The level set of f *at the initial point* x_0 *is compact, i.e.,*
$\Omega := \{\mathbb{R}^n : f(x) \leq f(x_0)\}$ *is compact.*

Assumption (2)

1. $\gamma \leq 1/2$,
2. $\alpha \leq 1/2$,
3. *There exists* $L_H > 0$ *such that*

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L_H \|x - y\|, \quad {}^\forall x, y \in \Omega + B(0, r_1),$$

where $r_1 := \dfrac{\mathcal{C} U_g^{1-\gamma} n}{c_2 s}$, *and* $\|\nabla f(x_k)\| \leq U_g$.

Global convergence

Let

$$t_{\min} = \min\left(1, \frac{\beta c_2^2 s^2}{\mathcal{C}^2 L_H U_g^{1-2\gamma} n^2}\right) \quad p = \frac{\alpha t_{\min}}{2\mathcal{C}(1+c_1)\frac{n}{s} U_H + 2c_2 U_g^{\gamma}}.$$

Theorem

Suppose that Assumptions (1) and (2) hold. Let

$$m = \left\lfloor \frac{f(x_0) - f^*}{p\varepsilon^2} \right\rfloor + 1$$

Then, with probability at least $1 - 2m\left(\exp(-\frac{\mathcal{C}_0}{4}s) - \exp(-s)\right)$, we have

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

$O(\varepsilon^{-2})$ complexity: same that [Ueda and Yamashita, 2010].

Success probability

We want $1 - 2m\left(\exp(-\frac{\mathcal{C}_0}{4}s) - \exp(-s)\right)$ as close to one as possible.

Success probability

We want $1 - 2m\left(\exp(-\frac{\mathcal{C}_0}{4}s) - \exp(-s)\right)$ as close to one as possible.

Assume that $\|x_0 - x^*\| \leq \bar{C}\sqrt{n}$ for some constant $\bar{C} > 0$, then for some constant $\hat{C} > 0$,

$$m \leq \hat{C}\frac{n^{9/2}}{\varepsilon}.$$

Success probability

We want $1 - 2m\left(\exp(-\frac{\mathcal{C}_0}{4}s) - \exp(-s)\right)$ as close to one as possible.
Assume that $\|x_0 - x^*\| \leq \bar{C}\sqrt{n}$ for some constant $\bar{C} > 0$, then for some constant $\hat{C} > 0$,

$$m \leq \hat{C}\frac{n^{9/2}}{\varepsilon}.$$

By taking $s = D\log(n)$, for $D > 9/2$ ensure $1 - 2m\left(\exp(-\frac{\mathcal{C}_0}{4}s) - \exp(-s)\right)$ tends to 1.

Local convergence

Assume that $\{x_k\}$ converge to a strict local minima $\bar{x}$. We show that

- the sequence $\{f(x_k)\}$ converges locally linearly to $f(\bar{x})$
- when f is strongly convex, we cannot aim at local super-linear convergence using random subspace.

Local convergence: assumptions

Assumption (2')

In a neighborhood of $\bar{x}$, we have

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L_H \|x - y\|.$$

Assumption (3)

We have that $s = o(n)$, that is, $\lim_{n \to +\infty} \frac{s}{n} = 0$.

Local convergence: assumptions

Assumption (2')

In a neighborhood of $\bar{x}$, we have

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L_H \|x - y\|.$$

Assumption (3)

We have that $s = o(n)$, that is, $\lim_{n \to +\infty} \frac{s}{n} = 0$.

Assumption (4)

We assume that

1. *There exists $\sigma \in (0,1)$ such that $r = rank(\nabla^2 f(\bar{x})) \geq \sigma n$*
2. *There exists $\rho \in (0,3)$ 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.*

Proposition 1

Let $0 < \varepsilon_0 < 1$. Then under Assumptions (3) and (4.1) there exists $n_0 \in \mathbb{N}$ (which depends only on ε_0 and σ) and a neighborhood $B^* \subseteq \bar{B}$ such that if $n \geq n_0$, for any $x \in B^*$,

$$P\nabla^2 f(x)P^\top \succeq \frac{(1-\varepsilon_0)^2 n}{2s}\sigma\bar{\lambda}I_s \quad \bar{\lambda} = \lambda_r(\bar{x})/2$$

holds with probability at least $1 - 6\exp(-s)$.

Proposition 1

Let $0 < \varepsilon_0 < 1$. Then under Assumptions (3) and (4.1) there exists $n_0 \in \mathbb{N}$ (which depends only on ε_0 and σ) and a neighborhood $B^* \subseteq \bar{B}$ such that if $n \geq n_0$, for any $x \in B^*$,

$$P\nabla^2 f(x)P^\top \succeq \frac{(1-\varepsilon_0)^2 n}{2s}\sigma\bar{\lambda} I_s \quad \bar{\lambda} = \lambda_r(\bar{x})/2$$

holds with probability at least $1 - 6\exp(-s)$.

PL inequality

There exists $n_0 \in \mathbb{N}$ (which depends only on ε_0 and σ) and neighborhoods $\hat{B} \subset B^*$ and B_0 (a neighborhood of $0 \in \mathbb{R}^s$) such that if $n \geq n_0$, for any $x \in \hat{B}$,

$$\|P\nabla f(x)\|^2 \geq \frac{(1-\varepsilon_0)^2 n}{s}\sigma\bar{\lambda}\left(f(x) - \min_{u \in B_0} f(x + P^\top u)\right)$$

holds with probability at least $1 - 6\exp(-s)$.

Proposition 2

Under Assumptions (1),(2') and (4). there exists $0 < \kappa < 1$, $k_0 \in \mathbb{N}$, $n_0 \in \mathbb{N}$, and $\bar{C} > 0$ such that if $n \geq n_0$, $k \geq k_0$, we have with probability $1 - 6(\exp(-s) + \exp(-\frac{\mathcal{C}_0}{4}s))$:

$$f(x_k) - \min_{u \in B_0} f(x_k + P_k^\top u) \geq \bar{C}(f(x_k) - f(\bar{x})).$$

Local convergence: Theorem 1

Theorem

Under Assumptions (1),(2'),(3) and (4), there exists $0 < \kappa < 1$, $k_0 \in \mathbb{N}$, and $n_0 \in \mathbb{N}$ such that if $n \geq n_0$, $k \geq k_0$, then

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

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

Local convergence: Theorem 1

Theorem

Under Assumptions (1),(2'),(3) and (4), there exists $0 < \kappa < 1$, $k_0 \in \mathbb{N}$, and $n_0 \in \mathbb{N}$ such that if $n \geq n_0$, $k \geq k_0$, then

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

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

Theorem

Under Assumptions (1),(2'),(3) and (4), there exists $0 < \kappa' < 1$, $s_0 \in \mathbb{N}$, $k_0 \in \mathbb{N}$, and $n_0 \in \mathbb{N}$ such that if $n \geq n_0$, $k \geq k_0$, then

$$\mathbb{E}[f(x_{k+1}) - f(\bar{x})] \leq \kappa' \mathbb{E}[f(x_k) - f(\bar{x})].$$

holds if $s \geq s_0$.

Super-linear convergence?

Assumption (5)

We assume that

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

Super-linear convergence?

Assumption (5)

We assume that

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

Theorem

Under Assumptions (2') and (5), if f is locally strongly convex around $\bar{x}$. There exists a constant $c > 0$ such that for k large enough,

$$\|x_{k+1} - \bar{x}\| \geq c\|x_k - \bar{x}\|$$

holds with probability at least $1 - 2\exp(-\frac{\mathcal{C}_0}{4}) - 2\exp(-s)$.

Super-linear convergence?

Assumption (5)

We assume that

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

Theorem

Under Assumptions (2') and (5), if f is locally strongly convex around $\bar{x}$. There exists a constant $c > 0$ such that for k large enough,

$$\|x_{k+1} - \bar{x}\| \geq c\|x_k - \bar{x}\|$$

holds with probability at least $1 - 2\exp(-\frac{\mathcal{C}_0}{4}) - 2\exp(-s)$.

We deduce from the theorem and the assumptions that there exists a constant c' such that

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

with probability at least $1 - 2\exp(-\frac{\mathcal{C}_0}{4}) - 2\exp(-s)$.

Theorem

Under Assumptions (2') and (5), if f is locally strongly convex around $\bar{x}$. There exists a constant $c' > 0$ such that for k large enough, and s greater than some constant,

$$\mathbb{E}[\|x_{k+1} - \bar{x}\|] \geq c' \mathbb{E}[\|x_k - \bar{x}\|].$$

Numerical experiments: Support vector regression

Data: $\forall i \leq m,\ (x_i, y_i) \in \mathbb{R}^n \times \{0, 1\}$, we aim minimizing sum of a loss function and a regularizer

$$f(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(y_i - x_i^\top w) + \lambda \|w\|^2.$$

- Internet advertisements dataset from UCI repository[Dua and Graff, 2017] processed so that the number of instances is $m = 600$ and and $n = 1500$.
- Comparison with Gradient Descent (GD) and Regularized Newton Method (RNM)
- Step sizes are all determined by Armijo backtracking line search
- The parameters are fixed as follows:

$$c_1 = 2, c_2 = 1, \gamma = 0.5, \alpha = 0.3, \beta = 0.5, s \in \{100, 200, 400\}.$$

Loss function

$$\ell(t) = \frac{2t^2}{t^2 + 4}$$

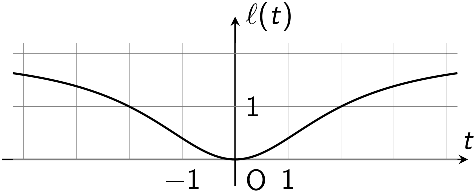


Figure: The robust loss functions.

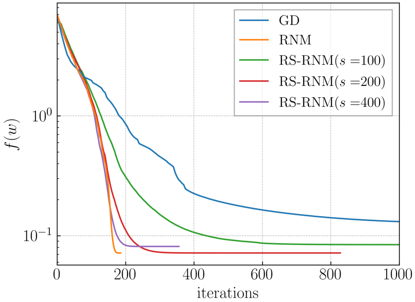


Figure: iterations versus $f(w)$ ($\log_{10}$-scale)

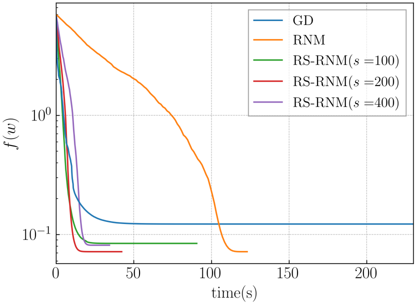


Figure: time versus $f(w)$ ($\log_{10}$-scale)).

Future work

Can we find a second order subspace algorithm with local superlinear convergence ? Full paper: "T. Fuji, P.L. Poirion, A. Takeda, **Randomized subspace regularized Newton method for unconstrained non-convex optimization**. arXiv:2209.04170, (2022)"

Reference I

Dua, D. and Graff, C. (2017).
UCI machine learning repository.

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

Hanzely, F., Doikov, N., Richtárik, P., and Nesterov, Y. (2020).
Stochastic subspace cubic Newton method.
In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4027–4038. PMLR.

Kovalev, D., Gower, R. M., Richtárik, P., and Rogozin, A. (2020).
Fast linear convergence of randomized bfgs.
arXiv preprint arXiv:2002.11337.

Roberts, L. and Royer, C. W. (2022).
Direct search based on probabilistic descent in reduced spaces.
arXiv preprint arXiv:2204.01275.

Reference II

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