Cubic Regularization Literature Review

Yufeng Yang

ABSTRACT: Overview of main achievements in the area of Cubic Regularization. Will be updated weekly

KEYWORDS: Non-convex Optimization, Cubic Regularization, Distributed Optimization, Machine Learning

1 Main text for Cubic Regularization for Newton's method and its global performance

suppose we need to solve the optimization problem.

$$\min_{x \in Q} f(x) \tag{1}$$

Where Q is a closed convex set. Then we can choose the next point x_{k+1} in our sequence to be:

$$\min_{y \in Q} \xi_{1,x_k}(y) . \xi_{1,x_k} = f(x_k) + \langle f'(x_k), y - x_k \rangle + \frac{1}{2} D \left\| y - x \right\|^2 \tag{2}$$

Convergence of this scheme follows from the fact that $\xi_{1,x_k}(y)$ is an upper first-order approximation of the objective function, that is $\xi_{1,x_k}(y) \geq f(y)$. for $Q = R^n$, then the rule results in a natural gradient scheme:

$$x_{k+1} = x_k - \frac{1}{D} f'(x_k) \tag{3}$$

Assume the Lipschitz hessian:

$$\left\| f''(x) - f''(y) \right\| \leq L \left\| x - y \right\| \tag{4}$$

Then, we can do cubic approximation about the original objective function:

$$\xi_{2,x}(y) = f(x) + \langle f'(x), y - x \rangle + \frac{1}{2} \langle f''(x)(y-x), (y-x) \rangle + \frac{L}{6} \|y - x\|^3 \tag{5}$$

In this way, we can propose that the next point will be $x_{k+1} \in argmin_y \xi x_k(y)$ We call it Cubic Regularization of Newton's method. The problem here is actually non-convex and it have many local minima. However, with suitable deduction, it can be reduced into a convex optimization with one variable.

***Assumption*1** The Hessian of function f is Lipschitz continuous on F:

$$\|f''(x) - f''(y)\| \leq L \|x - y\| \tag{6}$$

for some $L \geq 0$.

Then we have:

Lemma 1. *For any x and y from F we have:*

$$\|f'(y) - f'(x) - f''(x)(y-x)\| \leq \frac{L}{2} \|y - x\|^2 \tag{7}$$

Let M be a positive parameter. Define a modified Newton step using the following cubic regularization of quadratic approximation of function f(x).:

$$T_M(x) \in Arg\min \left[\langle f'(x), y - x \rangle + \frac{1}{2} \langle f''(x)(y-x), y - x \rangle + \frac{M}{6} \|y - x\|^3 \right] \tag{8}$$

Algorithm 1 Cubic Regularization Algorithm by Yuri Nesterov

```
Require: : initialize $x_0 \in R^n$
Ensure: $M_k \in [L_0, 2L]$ such that $f(T_{M_k}(x_k)) \leq \bar{f}_{M_k}(x_k)$
  $x_{k+1} = T_{M_k}(x)$
```

Next, Yuri Nesterov talks about the implementing issues of above algorithm. Remember the optimization problem we want to solve is (5). We can solve it by constructing a "primal-duality" scheme. Define

$$v_u(h) = \langle g, h \rangle + \frac{1}{2} \langle Hh, h \rangle + \frac{M}{6} \|h\|^3 \tag{9}$$

and

$$v_l(r) = -\frac{1}{2}\left\langle (H + \frac{Mr}{2}I)^{-1}g, g \right\rangle - \frac{Mr^3}{12} \tag{10}$$

The theorem is:

theoremor any M¿0 we have the following relation:

$$\min_{h \in R^n} v_u(h) = \sup_{r \in D} v_l(r) \tag{11}$$

For any $r \in D$, direction $h(r) = -(H + \frac{M}{2})^{-1}g$ satisfies the equation:

$$0 \leq v_u(h(r)) - v_l(r) = \frac{M}{12}(r + 2\left\|h(r)\right\|)(\left\|h(r) - r\right\|)^2 \tag{12}$$

When achieves optimal value, r satisfy the following constraint:

$$r = \left\| \left(H + \frac{Mr}{2}I \right)^{-1} tg \right\|, r \geq \frac{2}{M}(-\lambda_n(H))_+ \tag{13}$$

To be more precise, the above formula can be transformed into:

$$r^2 = \sum_1^n \frac{\tilde{g_i}^2}{(\lambda_i + \frac{M}{2}r)^2}, r \geq \frac{2}{M}\left(-\lambda_n\right)_+ \tag{14}$$

In this article, Yuri Nesterov also argues that Algorithm 1 actually works. Recall the optimality condition for a local minimum should be:

$$f'(x) = 0, f''(x) \geq 0 \tag{15}$$

Therefore, it is reasonable to define a measure of local optimality:

$$\mu_M(x) = \max \left\{ \sqrt{\frac{2}{L+M}\left\|f'(x)\right\|}, \frac{-2}{2L+M}\lambda_n(f''(x)) \right\} \tag{16}$$

After defining above measure, in this article, Nesterov also proves the following theorem:

theoremet the sequence x_i be generated by method(3.3). Assume that the objective function f(x) is bounded below: $f(x) \geq f^* \; \forall x \in F$ Then $\sum_{i=0}^{\infty} r_{M_i}^3(x_i) \leq \frac{12}{L_0}(f(x_0) - f^*)$.

Moreover: $\lim_{x\to\infty} \mu_L(x_i) = 0$ and for any $k \geq 1$ we have:

$$\min_{1\leq i\leq k} \leq \frac{8}{3}\left(\frac{3(f(x_0) - f^*)}{2kL_0}\right)^{\frac{1}{3}} \tag{17}$$

The next theorem states the existence of local minimum satisfy the optimality condition:

theoremet sequence x_i be generated by algorithm1. For some $i \geq 0$, assume the set $L(f(x_i))$ be bounded. Then there exists a limit $\lim_{i\to\infty} f(x_i) = f^*$. The set X^* of the limit points of this sequence is non-empty. Moreover, this is a connected set, such that for any $x^* \in X^*$, we have:

$$f(x^*) = f^*, f'(x^*) = 0, f''(x^*) \geq 0 \tag{18}$$

Denote $\delta_k = \frac{L\|f'(x_k)\|}{\lambda_n^2(f''(x_k))}$ The following theorem indicates the convergence behavior for Algorithm 1.

theoremet $f''(x_0) \geq 0$ and $\delta_0 \leq \frac{1}{4}$. let the points x_k be generated by algorithm 1. Then:

1. For all $k \geq 0$ and δ_k are well defined and they converge quadratically to zero:

$$\delta_{k+1} \leq \frac{3}{2}\left(\frac{\delta_k}{1-\delta_k}\right)^2 \leq \frac{8}{3}\delta_k^2 \leq \frac{2}{3}\delta_k \tag{19}$$

2. Minimal eigenvalue of all Hessians $f''(x_k)$ lie within the following bounds:

$$e^{-1}\lambda_n(f''(x_0)) \leq \lambda(f''(x_k)) \leq e^{3/4}\lambda_n(f''(x_0)) \tag{20}$$

3. The whole sequence of x_i converges quadratically to a point x^*, which is a non-degenerate local minimum of function f(x). In particular, for any $k \geq 1$ we have:

$$\|f'(x_k)\| \leq \lambda_n^2(f''(x_k))\frac{9e^{3/2}}{16L}(\frac{1}{2}^{2^k}) \tag{21}$$

2 Trust Region Methods for Solving Cubic Regularization Problem

2.1 Notations used for identifying trust region problem

The problem we consider in this section is defined as:

$$m_k(x_k + s) = f(x_k) + \langle g_k, s \rangle + \frac{1}{2} \langle s, H_k s \rangle \tag{22}$$

g_k is the gradient of $f(x_k)$ and H_k is a symmetric bounded approximation of $\nabla_{kk} f(x_k)$. The model minimizer is defined as x_k^M. For this optimization problem, we need to add an constraint on the norm of step s, which is called trust region. Typical choice of norm would be l_1,l_2,and l_∞. For simplicity of derivation, we use l_2 norm in the following deduction.

2.2 Trust Region Problem Formulation

The optimization problem for l_2 norm model minimizer is:

$$\begin{aligned} \min_{s \in R^n} \quad & q(s) = \langle g, s \rangle + \tfrac{1}{2} \langle s, Hs \rangle \\ \text{s.t.} \quad & \|s\|_2 \leq \triangle \end{aligned} \tag{23}$$

if the solution is interior to the trust region, the trust region bound may as well not have been there, and therefore s^M is the model minimizer of $q(s)$. But this can only happen when $q(s)$ is convex (i.e the Hessian H is P.S.D). In the nonconvex case, a solution must lie on the boundary of the trust region, while in the convex case a solution may or may not do so. Thus, if the model is unbounded from below, or if the unconstrained minimizer lies outside the trust region, then the model minimizer must occur on the boundary and thus can be found as the global minimizer of $q(s)$ s.t $\|s\|_2 = \triangle$.

theoremny global minimizer of $q(s)$ s.t $\|s\|_2 = \triangle$ satisfies the equation:

$$H(\lambda^M) s^M = -g. \tag{24}$$

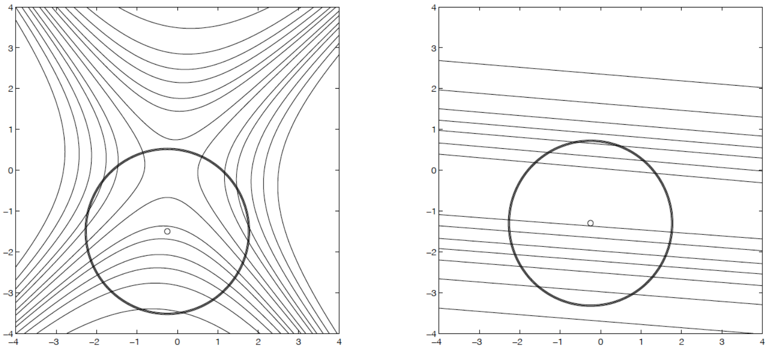
where $H(\lambda^M) = H + \lambda^M I$. if $H(\lambda^M)$ is positive definite, s^M is unique. (Pf skipped, use taylor expansion)

Corollary 1.1. *Any global minimizer of $q(s)$ s.t $\|s\|_2 = \triangle$ satisfies the equation:*

$$H(\lambda^M)s^M = -g. \tag{25}$$

where $H(\lambda^M) = H + \lambda^M I$. if $H(\lambda^M)$ is positive definite, $\lambda \geq 0$, and $\lambda^M(\|s^M\| - \triangle) = 0$, if $H(\lambda^M)$ is positive definite, s^M is unique.(Pf Skipped)

It is also possible to understand above corollary in the ”regularization way”. Recall Lagrange multiplier method, using trust region is equal as adding quadractic term for original objective function, which makes it to be convex. In this way, we can guarantee that there is unique s^M satisfy the requirement.



(a)

(b)

Figure 7.2.3: (a) The contours of a model with two solutions and the trust-region boundary, and (b) the contours of the singular (convex) modified model.

2.3 Finding the solution s^M

Using the obvious equations in corollary has a lot of drawbacks. Instead, we prefer to use newton method to find s^M.

Suppose that H has the following eigen-decomposition:$H = U^T \Lambda U$, where Λ is the diagonal matrix of eigenvalues $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ Thus, $H(\lambda) = U^T(\Lambda + \lambda I)U$. From above corollary, we want to seek the value of λ satisfy $\lambda_M \geq -\lambda_1$ (Also, if the strict inequality holds, the solution s^M is unique). If that is the case, the solution s^M can be expressed as:

$$s(\lambda) = -H(\lambda)^{-1}g = -U^T(\Lambda + \lambda I)Ug \tag{26}$$

Besides, the solution should also satisfy the nonlinear inequality constraint: $\|s(\lambda)\| \leq \triangle$. In order to make detailed analysis, we define

$$\psi(\lambda) = \|s(\lambda)\|^2 = \left\|U^T(\Lambda + \lambda I)^{-1}Ug\right\|^2 = \sum_{i=1}^{n} \frac{\gamma_i^2}{(\lambda_i + \lambda)^2} \tag{27}$$

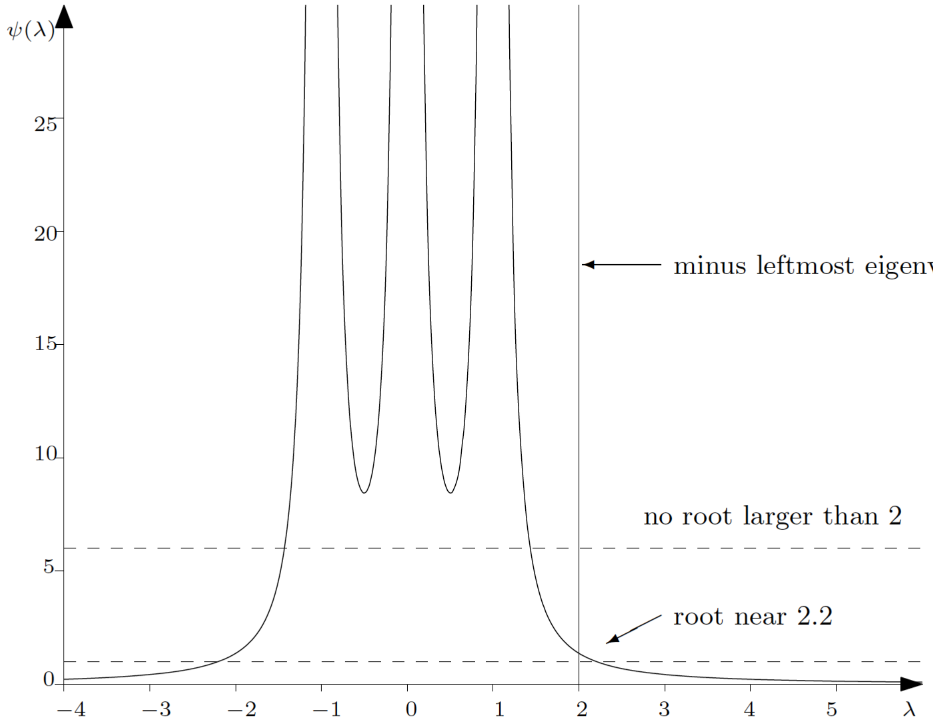
Where γ_i is $[Ug]_i$,the ith component of Ug.The author then make several plots for different case to show hard case we need to consider.

2.3.1 convex case

When the function is convex, then all eigenvalue of them should be non-negative. Thus, All the poles are less than equal to 0. For any threshold $\triangle$, we can always find a corresponding λ, which is always larger than $-\lambda_1$.

2.3.2 A harder case

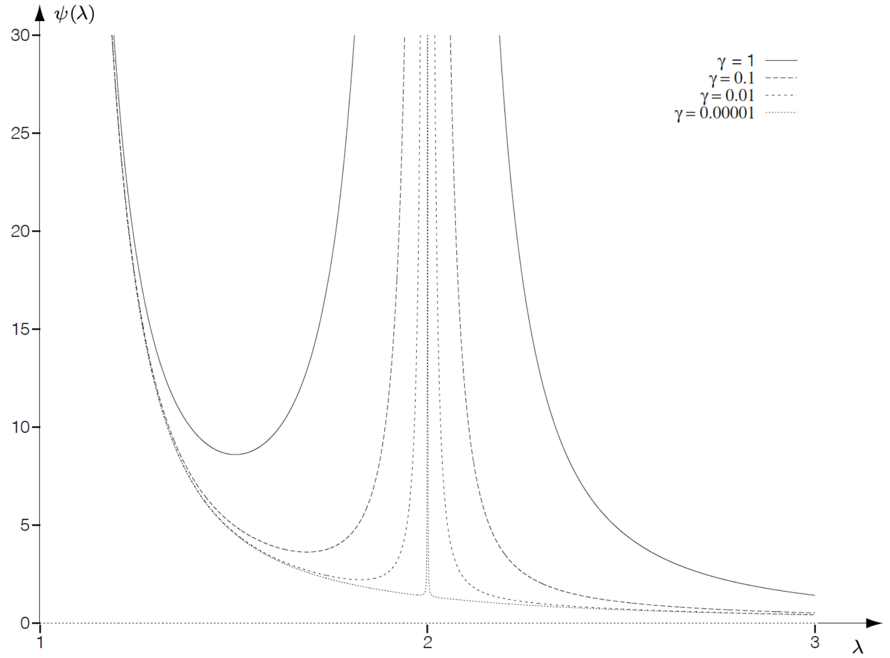
The following graph is the plot of $\psi(\lambda)$ for a non-convex function:



As we can see from above, we there exists negative eigenvalues, there exists poles at positive value.By corollary, we need to require $\lambda \geq 2$ (because the smallest hessian diagonal element is -2). Then, there is no solution for $\triangle$ larger than 1.2. However, if the first element of g is 0, then $\gamma 1$ equals to 0 and thus the most negative eigenvalue vanishes. If this case happens, we got into trouble. Not only we need to solve a singular system, but also we need to compute a partial eigensolution of H.

2.4 Use newton method to find the solution s^M

In general, ψ has poles but possibly no zeros.



Thus, it is reasonable to transform the solution finding problem from $\|s(\lambda) - \triangle\| = 0$ into $\Phi(\lambda) = \frac{1}{\|s(\lambda)\|_2} - \frac{1}{\triangle} = 0$

Lemma 2. *Suppose $g \neq 0.Thenthefunction\phi$ is strictly increasing, when $\lambda \geq -\lambda_1$, and concave. Its first derivative are :*

$$\phi'(x) = -\frac{\langle s(\lambda), \bigtriangledown_\lambda s(\lambda)\rangle}{\|s(\lambda)\|_2^3} \tag{28}$$

and

$$\phi''(\lambda) = \frac{3\,\langle s(\lambda), \bigtriangledown_\lambda s(\lambda)\rangle^2 - \|s(\lambda)\|_2^2\,\|\bigtriangledown_\lambda s(\lambda)\|_2^2}{\|s(\lambda)\|_2^5} \tag{29}$$

where

$$\bigtriangledown_\lambda\, s(\lambda) = -H(\lambda)^{-1}s(\lambda) \tag{30}$$

(Pf skipped)

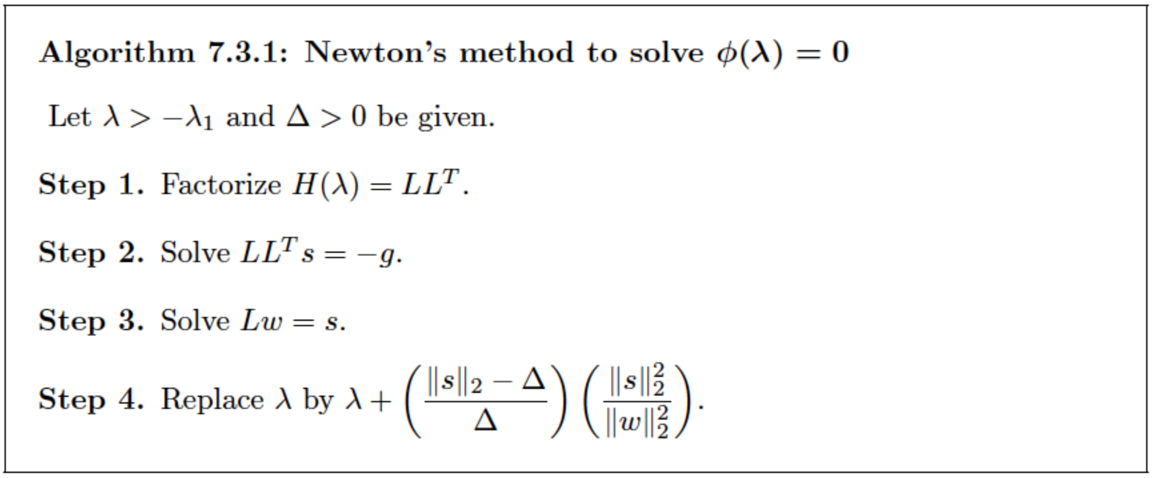
The general form of Newton's algorithm:

$$\lambda^{\lambda} - \phi(\lambda)/\phi'(\lambda) \tag{31}$$

If $H(\lambda)$ is positive definite, thus we may use cholesky factors $H(\lambda) = L(\lambda)L^T(\lambda)$ and also there is a relation:

$$\langle s, H(\lambda)^{-1}s\rangle = \langle s, L^{-T}L^{-1}s\rangle = \langle L^{-1}s, L^{-1}s\rangle = \|w\|_2^2 \tag{32}$$

Use above facts, we have the algorithm:



Algorithm 7.3.1: Newton's method to solve $\phi(\lambda) = 0$

Let $\lambda > -\lambda_1$ and $\Delta > 0$ be given.

Step 1. Factorize $H(\lambda) = LL^T$.

Step 2. Solve $LL^T s = -g$.

Step 3. Solve $Lw = s$.

Step 4. Replace λ by $\lambda + \left(\frac{\|s\|_2 - \Delta}{\Delta}\right)\left(\frac{\|s\|_2^2}{\|w\|_2^2}\right)$.

2.4.1 Safe Guarantee for Newton Algorithm

The most important thing we need to guarantee for newton's algorithm is its convergence.

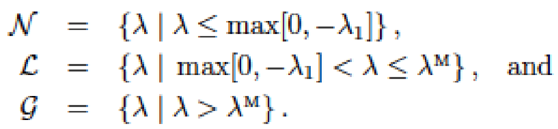
Lemma 3. *Suppose $\lambda \geq -\lambda_1$ and $\phi(\lambda) \leq 0$. Then the method starting from λ will inherit these properties and converges monotonically towards the required root, λ^M. The convergence is globally Q-linear with factor at least:*

$$\gamma_\lambda = 1 - \phi(\lambda^M)/\phi'(\lambda) \leq 1 \tag{33}$$

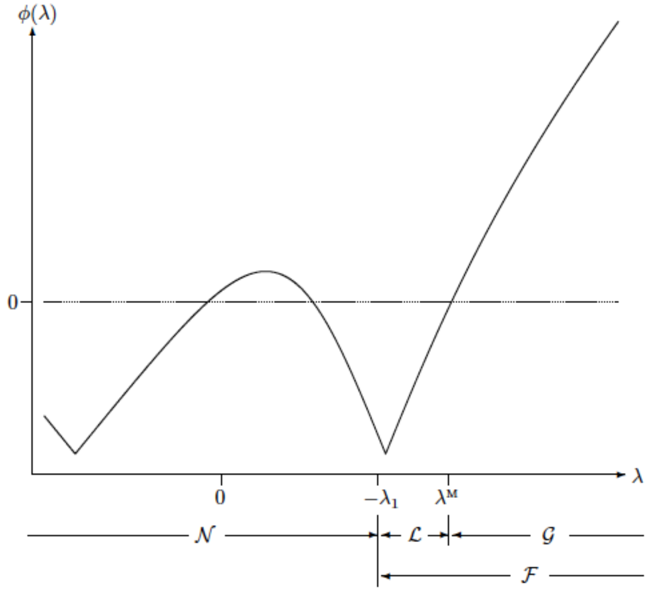
and is ultimately Q-quadratic.

Lemma 4. *Suppose $\lambda \geq -\lambda_1$ and $\phi(\lambda) \geq 0$. Then the next iteration satisfies that $\lambda^+ \leq \lambda^M$ and will additionally satisfy either $\lambda^+ \geq -\lambda_1$ and $\phi(\lambda^+) \leq 0$ or $\lambda \leq -\lambda_1$*

But there is hard case that we cannot find the solution $\phi(\lambda^+) \leq 0$ and $\lambda^+ \geq -\lambda_1$. We need to be more cautious for the choice of λ. We shall seperate three possible values of



$$\begin{aligned} \mathcal{N} &= \{\lambda \mid \lambda \leq \max[0, -\lambda_1]\}, \\ \mathcal{L} &= \{\lambda \mid \max[0, -\lambda_1] < \lambda \leq \lambda^{\mathrm{M}}\}, \quad \text{and} \\ \mathcal{G} &= \{\lambda \mid \lambda > \lambda^{\mathrm{M}}\}. \end{aligned}$$



λ into disjoint sets:

And also $F = L \cup G$. The following Graph is an illustration:

2.4.2 Update λ

We need an estimation interval $\left[\lambda^L, \lambda^U\right]$ to ensure that newton's method will not diverge(Algorithm 7.3.2/7.3.3).
The estimation of λ is geometric mean $\lambda = \sqrt{\lambda^L \lambda^U}$

3 Solving Cubic Regularization Problem Distributedly

Distributed Optimization is motivated by finite sum minimization problem, which has wide applications in communication networks or machine learning problems. The optimization is formulated as:

$$\min_{x\in Q}\left\{f(x)=\sum_{i=1}^{n}f^i(x)\right\} \tag{34}$$

Where $f^i(x)$ is the local emprirical risk of a subset of data points stored locally by an agent i. Each agent can only access f^i only. Also, we assume that each units are connected over a network that allows for sparse communication between them. Thus, the proposed solution needs to be executed locally at each agent, using local information only and achieve convergence rate as if they has access to the completed dataset.

3.0.1 Notations and Problem reformulation

Consider a network of m agents, modeled as a fixed, connected and undirected graph $G = (V, E)$,where $V = (1, 2...m)$ and $E \subseteq V \times V$ is a set of edges. It satisfy that $(i, j) \in E$ if and only if agent j connected to agent i. Agents try to jointly solve(34), but agents can only access to the information of $f^i(x)$. But they are allowed to exchange information by each other. Assume that each $f^i : Q \to R$ is convex with Lipschitz continuous gradient and hessian. defined in a nonempty, convex and compact set $Q \subseteq R^n$. Also, assume that the minimizer x^* can be found in the interior of set Q.

Recall Graph Laplacian, it is defined as :

$$\left\{\begin{matrix}[W_G]_{i,j}=-1,\forall\,(i,j)\in E\\ [W_G]_{i,j}=deg(i),\forall i=j\\ [W_G]_{i,j}=0\end{matrix}\right\} \tag{35}$$

The matrix W_G is symmetric and positive semi-definite.It has two important properties.

Lemma 5. *When all the elements for vector x are equal, i.e: $x_1 = x_2 = ... = x_m$, $W_G x = 0$.*

Lemma 6. *The corresponding eigenvector for 0 eigenvalue of W_G is $\{\mathbf{1}_m\}$*

Define $\boldsymbol{W} = W_G \otimes \boldsymbol{I_n}$ The finite sum minimization problem can be reformulated as:

$$\min_{x\in Q^n,\sqrt{(}W)x=0} F(x)=\sum_{i=1}^{m}f^i(x^i) \tag{36}$$

Where $Q^m = \left\{x^{nm}|x^T = [(x^1)^T...(x^m)^T, x^i \in Q\forall i \in V\right\}$

3.1 Definitions and Assumptions

Definition 3.1. *A point is called a $(\epsilon, \hat{\epsilon})$ solution if $F(\hat{x}) - F^* \leq \epsilon$*

Definition 3.2. *Define $\hat{x} \approx_\delta argmin_{x \in Q} f(x)$ as a point in X such that $f(\hat{x}) - f^* \leq \delta$ where f^* is the minimum value of function $f(x)$ over the set X*

The following are the main assumptions to guarantee the convergence result of algorithm.

Assumption 1. *(Lipstichiz gradient)Each function $f^i(x)$ has $M^i - Lipschitz$ continous gradient over the set Q satisfy $\|\nabla f^i(x) - \nabla f^i(y)\| \leq M_1^i \|x - y\|$*

Assumption 2. *(Lipschitz Hessian) Each function is twice-differentiable and M_2^i Lipschtiz continuous hessian over the set Q satisfy $\|\nabla^2 f^i(x) - \nabla^2 f^i(y)\| \leq M_1^i \|x - y\|$*

Assumption 3. *The diameter of the compact set Q is upper bound by a constant D_Q*

With the above definitions and assumptions, we are enough to introduce the main algorithm and ways for solving it.
Notice the cubic regularization function is an upper bound for $F(x)$ at each point. Instead of solving it, we solve the cubic regularization problem:

$$\widehat{F}(x, z) = F(z) + \langle \nabla F(z), x - z \rangle \frac{1}{2} \left\langle \nabla^2 F(z)(x - z), x - z \right\rangle + \frac{N}{6} \|x - z\|^3 \tag{37}$$

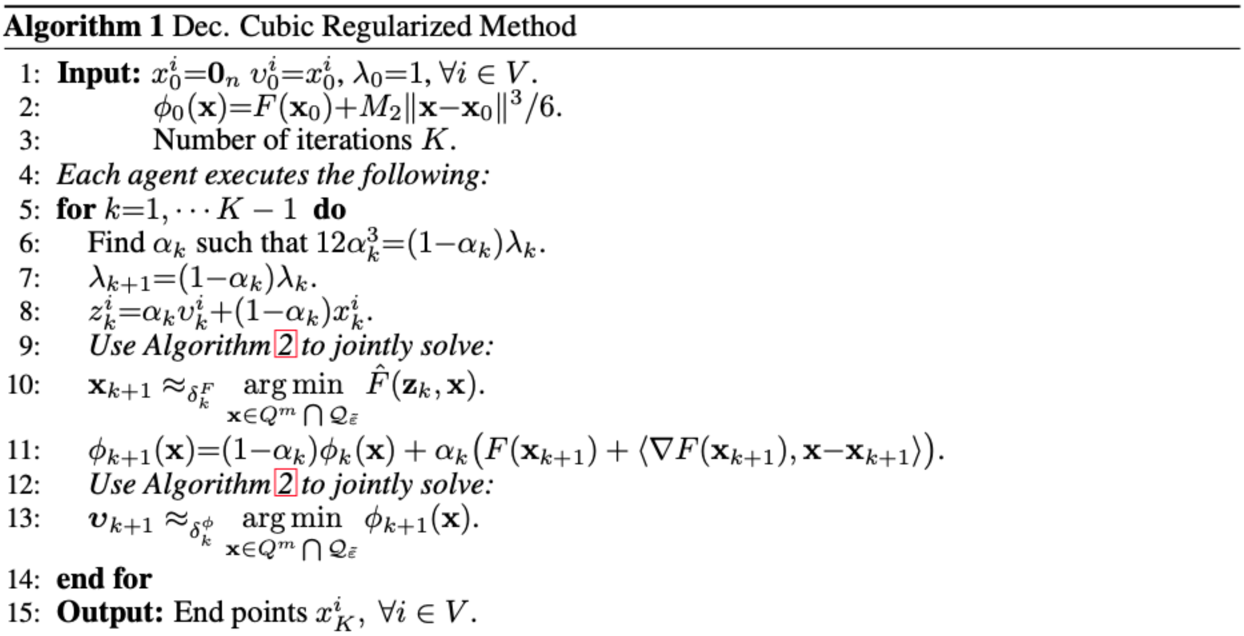
Using the following algorithm, we can make the approximated solution getting from Cubic function close the true solution with controllable error. And the algorithm complexity is $O(k^{-3})$The algorithm presents as following:

Algorithm 2 is one way for finding solution of cubic function provided in algorithm1.

3.1.1 Finding solution for Distributed Cubic function

For convenience, replacing the $\sqrt{W}$ matrix by a generic matrix whose null space is consensus subspace.Denote $g = \left[g_1^T, g_2^T ......g_m^T\right]$ Where g_i is denoted the gradient of agents i at some point in agents i, and $H = (H_1, ...H_m)$ where $H_i = \nabla^2 f^i(z_k^i)$ $h = x - z$. Then, the problem can be rewritten as:

$$\min_{h \in H \subset R^{n,m} Ah=0} \left\{ \Phi(x) = \langle g, h \rangle + \frac{1}{2} \langle Hh, h \rangle + \frac{N}{6} \|h\|^3 \right\} \tag{38}$$



Algorithm 1 Dec. Cubic Regularized Method

```
Input: $x_0^i = \mathbf{0}_n$ $v_0^i = x_0^i$, $\lambda_0 = 1$, $\forall i \in V$.
        $\phi_0(\mathbf{x}) = F(\mathbf{x}_0) + M_2\|\mathbf{x} - \mathbf{x}_0\|^3/6$.
        Number of iterations $K$.
Each agent executes the following:
for $k = 1, \cdots K-1$ do
    Find $\alpha_k$ such that $12\alpha_k^3 = (1-\alpha_k)\lambda_k$.
    $\lambda_{k+1} = (1-\alpha_k)\lambda_k$.
    $z_k^i = \alpha_k v_k^i + (1-\alpha_k)x_k^i$.
    Use Algorithm 2 to jointly solve:
    $\mathbf{x}_{k+1} \approx_{\delta_k^F} \arg\min_{\mathbf{x} \in Q^m \bigcap \mathcal{Q}_{\bar{\varepsilon}}} \hat{F}(\mathbf{z}_k, \mathbf{x})$.
    $\phi_{k+1}(\mathbf{x}) = (1-\alpha_k)\phi_k(\mathbf{x}) + \alpha_k\big(F(\mathbf{x}_{k+1}) + \langle \nabla F(\mathbf{x}_{k+1}), \mathbf{x} - \mathbf{x}_{k+1}\rangle\big)$.
    Use Algorithm 2 to jointly solve:
    $\boldsymbol{v}_{k+1} \approx_{\delta_k^\phi} \arg\min_{\mathbf{x} \in Q^m \bigcap \mathcal{Q}_{\bar{\varepsilon}}} \phi_{k+1}(\mathbf{x})$.
end for
Output: End points $x_K^i$, $\forall i \in V$.
```

Figure 1: Decentralized Cubic

By definition of g and H. we can write them distributedly. i.e $\langle g, h\rangle + \frac{1}{2}\langle Hh, h\rangle = \sum_{i=1}^m g_i^T h_i + \frac{1}{2}h_i^T H_i h_i$. However, the cubic term cannot be directly written in a distributed manner. The following lemma formulates cubic term as the optimal solution of an optimization problem.

Proposition 1.

$$\max_{\tau_i \geq 0, \tau_i = \tau_j} \quad \sum_{i=1}^m \|x_i\|^2 \tau_i - \frac{4}{3m}\sum_{i=1}^m \tau_i^3 = \frac{1}{3}\|x\|^3 \tag{39}$$

The proof is using the optimality condition for above optimziation with respect to ${}_i$. And plug ${}_i^*$ into term and get the desired result.

Using this proposition, the original optimization can be formulated as an minmax optimization problem:

$$\min_{h \in H \subset R^{nm}, Ah = 0} \max_{\tau_i \geq 0, B\tau = 0} \left\{ \langle g, h\rangle + \frac{1}{2}\langle Hh, h\rangle + \frac{N}{2}\sum_{i=1}^m \|h_i\|^2 \tau_i - \frac{2}{3}\sum_{i=1}^m \tau_i^3 \right\} \tag{40}$$

Where B is also a generic matrix satisfy that $B\tau = 0$ when $\tau_1 = \tau_2 = \ldots \tau_m$

Using lemma 3, it can be shown that the above optimization problem is equivalient as:

$$\min_{h\in H\subset R^{n,m},Ah=0} \max_{\tau_i\geq 0,} \left\{ \langle g,h\rangle + \frac{1}{2}\langle Hh,h\rangle + \frac{N}{2}\sum_{i=1}^{m} \|h_i\|^2 \tau_i - \frac{2}{3}\sum_{i=1}^{m}\tau_i^3 \right\} \tag{41}$$

Proof. The KKT optimality condition for original minmax optimization problem is:

$$\begin{aligned} Ah &= 0 \\ B\tau &= 0 \\ g + (H+NT)h - A^Ty &= 0 \\ \frac{N}{4}\|h_i\|^2 - \tau_i^2 + B^T\eta &= 0 \forall i \in 1...m \end{aligned} \tag{42}$$

When $Ah = 0$ and $B\tau = 0$, by the property of generic matrix A and B, it indicates that all the entries in h and τ are equal. Thus, if we view the last inequality, it indicates that $B^T\eta = \alpha I$ for some α. We can argue that this α here can only be 0.
If $\alpha = 0$, it indicates that $B^T\tau = 0$, which indicates all the elements are equal. Assume $\alpha \neq 0$, we prove by contradiction.
Use eigen-decomposition to represent $B = V^T$. Thus it indicates $V\Lambda V^T = \alpha I$ then $\Lambda V^T\eta = \alpha V1$ Using the property for matrix A and B, the corresponding eigenvector for eigenvalue $\lambda = 0$ is $\mathbf{1}$.Thus, the right-hand-side system exists an element which is equal to 0, which indicates α can only be 0. □

In this way, we can eliminate the effect of set B. The optimization problem may be much easier to solve.Define:

$$\varphi(x) = \min_{h\in H}\max_{\tau_i} \left\{ \langle g,h\rangle + \frac{1}{2}\langle (H+NT)h,h\rangle - \frac{2N}{3m}\sum_{i=1}^{n}\tau_i^3 - \langle y,Ah\rangle \right\} \tag{43}$$

Using Demyanov-Danskin's theorem, it follows that $\nabla\varphi(y) = Ah^*(A^Ty)$. Denote $h^*(A^Ty)$ as the unique solution of the inner maximization problem:

$$h^*(A^Ty) = arg\min_{h\in H}\max_{\tau_i} \left\{ \langle g,h\rangle + \frac{1}{2}\langle (H+NT)h,h\rangle - \frac{2N}{3m}\sum_{i=1}^{n}\tau_i^3 - \langle y,Ah\rangle \right\} \tag{44}$$

The solution of above optimization problem can be computed locally at each node, namely:

$$h_i^*([A^T y]_i) = arg \min_{h \in H} \max_{\tau_i} \left\{ \left\langle g_i, \hat{h} \right\rangle + \frac{1}{2} \left\langle (H_i + N\tau_i I_n)\hat{h}, \hat{h} \right\rangle - \frac{2N}{3m}\tau_i^3 - \left\langle y_i, [A\hat{h}]_i \right\rangle \right\} \tag{45}$$

Using the optimality condition, and taking derivative with respect to h_i and τ_i for each agent, we have the following system of equations:

$$\begin{aligned} (H + N\tau_i I_n)h_i &= ([A^T y]_i - g_i) \\ \frac{m}{4} \|h_i\|^2 - \tau_i^2 &= 0 \end{aligned} \tag{46}$$

Recall Corollary1.1, we can use that lemma to solve it by trust region method.
Define $h_i = s \forall i \in 1...m$, and $H(\tau) = H_i + N\tau_i I_n$ and $\hat{g} = [A^T y]_i - g_i$ and $\Delta^2 = \frac{4}{m}\tau_i^2$
Use the normal way, we first do eigen-decomposition $H = U^T \Lambda U$ Plug them into the second inequality, we have:

$$\begin{aligned} \frac{m}{4} \left\|(\Lambda + N\tau_i I_n)^{-1}([A^T y]_i - g_i)\right\|^2 &= \tau_i^2 \\ \frac{m}{4} \sum_{j=1}^{d} \frac{\gamma_j^2}{(\lambda_j^2 + N\tau_i)} &= \tau_i^2 \end{aligned} \tag{47}$$

These two relations leads to the algorithm 2 for finding solution of $h^*(A^T y)$

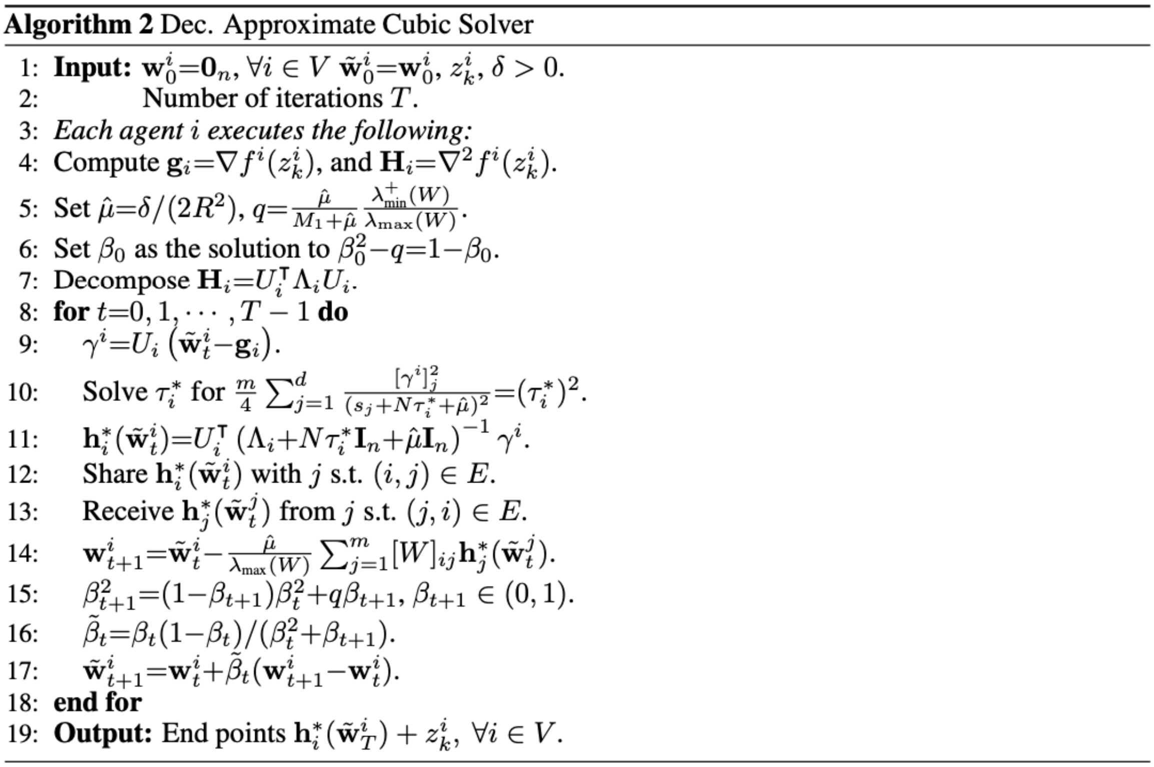
3.2 $O(\frac{1}{k^3})$ proof of algorithm 1

3.3 Modify original decentralized algorithm by Yuri Nesterov's inexact second order method

4 Citations

Quotations must follow the ABNT NBR 10520 guidelines. In indirect citations (paraphrases), the author should be cited in parentheses by last name, in all capital letters, sepa-rated by a comma from the year of publication (SWERTS, 1997). If the name of the author is cited in the text, only the year should be indicated in parentheses: According to Oliveira Jr (2000), [. . .].

Whenever the indication of the page is necessary, as in a direct quotation, the number



```
Algorithm 2 Dec. Approximate Cubic Solver
Input: $\mathbf{w}_0^i = \mathbf{0}_n, \forall i \in V$ $\tilde{\mathbf{w}}_0^i = \mathbf{w}_0^i$, $z_k^i$, $\delta > 0$.
        Number of iterations $T$.
Each agent $i$ executes the following:
Compute $\mathbf{g}_i = \nabla f^i(z_k^i)$, and $\mathbf{H}_i = \nabla^2 f^i(z_k^i)$.
Set $\hat{\mu} = \delta/(2R^2)$, $q = \frac{\hat{\mu}}{M_1+\hat{\mu}} \frac{\lambda_{\min}^+(W)}{\lambda_{\max}(W)}$.
Set $\beta_0$ as the solution to $\beta_0^2 - q = 1 - \beta_0$.
Decompose $\mathbf{H}_i = U_i^\intercal \Lambda_i U_i$.
for $t = 0, 1, \cdots, T-1$ do
    $\gamma^i = U_i \left(\tilde{\mathbf{w}}_t^i - \mathbf{g}_i\right)$.
    Solve $\tau_i^*$ for $\frac{m}{4} \sum_{j=1}^d \frac{[\gamma^i]_j^2}{(s_j + N\tau_i^* + \hat{\mu})^2} = (\tau_i^*)^2$.
    $\mathbf{h}_i^*(\tilde{\mathbf{w}}_t^i) = U_i^\intercal \left(\Lambda_i + N\tau_i^* \mathbf{I}_n + \hat{\mu}\mathbf{I}_n\right)^{-1} \gamma^i$.
    Share $\mathbf{h}_i^*(\tilde{\mathbf{w}}_t^i)$ with $j$ s.t. $(i,j) \in E$.
    Receive $\mathbf{h}_j^*(\tilde{\mathbf{w}}_t^j)$ from $j$ s.t. $(j,i) \in E$.
    $\mathbf{w}_{t+1}^i = \tilde{\mathbf{w}}_t^i - \frac{\hat{\mu}}{\lambda_{\max}(W)} \sum_{j=1}^m [W]_{ij} \mathbf{h}_j^*(\tilde{\mathbf{w}}_t^j)$.
    $\beta_{t+1}^2 = (1-\beta_{t+1})\beta_t^2 + q\beta_{t+1}$, $\beta_{t+1} \in (0,1)$.
    $\tilde{\beta}_t = \beta_t(1-\beta_t)/(\beta_t^2 + \beta_{t+1})$.
    $\tilde{\mathbf{w}}_{t+1}^i = \mathbf{w}_t^i + \tilde{\beta}_t(\mathbf{w}_{t+1}^i - \mathbf{w}_t^i)$.
end for
Output: End points $\mathbf{h}_i^*(\tilde{\mathbf{w}}_T^i) + z_k^i$, $\forall i \in V$.
```

Figure 2: Solve Cubic Sub-problem

of the page must be placed right after the year, separated from it by a comma and preceded by "p.", for example, (OLIVEIRA JR, 2000, p. 95). Direct quotations of up to three lines are made inside the text in double quotation marks, while direct quotations of more than three lines should be separated from the text with a 4cm indentation to the left, font size 10, with-out quotation marks and without indentation of the first line.

The citations of several works by the same author published in the same year should be distinguished by lower case letters after the date, without spacing (FERREIRA, 2007a).

When the work has two or three authors, all may be indicated, separated by semicolons (HUETTIG; ROMMERS; MEYER, 2011); when there are more than three authors, the first name is indicated, followed by **et al.** (ALMEIDA **et al.**, 2013). In this case, it is advisa-ble to indicate all authors in the References.

The References at the end of the text must also comply with ABNT norms. The con-tent of the articles and the accuracy of the references are the sole responsibility of the au-thors. The Editors and the Brazilian Association of Linguistics do not assume any responsi-bility for the opinions or statements of the authors. All and only the works of authors cited in the text should appear in the References, which are typed with single spacing between lines, separated by a simple space, and listed in alphabetical order by the surname of the first author. Where available, the authors should include the URLs and DOIs of the refer-ences used.

Acknowledgements

Acknowledgments and mention of research funding (scholarships, financial aid, and the respective funding agencies) that enabled the research should not be included in a footnote but in an Acknowledgments section at the end of the article, before the References.

References

References: (according to ABNT NBR 6023 (2018), the DOI must be referenced when-ever possible).

BRAYNER, A. R. A.; MEDEIROS, C. B. Incorporação do tempo em SGBD orientado a objetos. *In*: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 9., 1994, São Paulo. Anais [...]. São Paulo: USP, 1994. p. 16-29. (**Example of a work published in Proceedings**)

DANTAS, José Alves *et al.* Regulação da auditoria em sistemas bancários: análise do cenário internacional e fatores determinantes. **Revista Contabilidade Finanças**, São Paulo, v. 25, n. 64, p. 7-18, jan./abr. 2014. DOI http://dx.doi.org/10.1590/S1519-70772014000100002. Acesso em: 20 maio 2014 **(Example of an article published in a journal with a DOI**).

LUCK, Heloisa. **Liderança em gestão escolar.** 4. ed. Petrópolis: Vozes, 2010. (**Example of a book**)

MORAES, João Antônio de; RILLIARD, Albert. "Prosody and Emotion in Brazilian Portu-guese". *In*: ARMSTRONG, Meghan E.; HENRIKSEN, Nicholas; VANRELL, Maria del Mar. **Intonational Grammar in Ibero-Romance**: Approaches across linguistic subfields. Amsterdam: John Benjamins, 2016, p. 135-152. (Example of a book chapter)

RODRIGUES, Ana Lúcia Aquilas. **Impacto de um programa de exercícios no local de trabalho sobre o nível de atividade física e o estágio de prontidão para a mudança de comportamento**. 2009. Dissertação (Mestrado em Fisiopatologia Experimental) – Faculdade de Medicina, Universidade de São Paulo, São Paulo, 2009. (**Example of an aca-demic work – thesis, dissertation, etc.**)

SEKEFF, Gisela. O emprego dos sonhos. **Domingo**, Rio de Janeiro, ano 26, n. 1344, p. 30-36, 3 fev. 2002. (**Example of a journal article without URL or DOI**)