The University of Texas at Austin
Department of Electrical and Computer Engineering

ECE 381V: Large-Scale Optimization II — Spring 2022

Lecture 24

Caramanis & Mokhtari Wednesay, April 20, 2022

Goal: In this lecture, we study the Cubic Regularization of Newton's method (CRN) and its convergence rate for nonconvex problems.

1 Setting the stage

Suppose we aim to solve the following unconstrained problem

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

Consider an open convex set $\mathcal{F}$ such that $\mathcal{F} \subseteq \mathbb{R}^n$.
We first formally define the only assumption that we need to define the CRN method.

Assumption 1. *f is twice differentiable and its Hessian is L_2-Lipschitz continuous on $\mathcal{F}$, i.e.,*

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\hat{\mathbf{x}})\|_2 \leq L_2\|\mathbf{x}-\hat{\mathbf{x}}\|, \qquad \text{for all} \quad \mathbf{x}, \hat{\mathbf{x}} \in \mathcal{F}$$

Further, suppose we are given an arbitrary initial point $\mathbf{x}_0 \in \mathcal{F}$ with function value $f(\mathbf{x}_0)$. We define it corresponding sublevel set as

$$\mathcal{S}(f(\mathbf{x}_0)) := \{\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$$

All we need that the convex set $\mathcal{F}$ is large enough that contains the sublevel set $\mathcal{S}(f(\mathbf{x}_0))$.

2 Cubic update

The main update of the CRN method requires solving the following cubic problem

$$\min_{\mathbf{y}} \phi_M(\mathbf{y};\mathbf{x}_k) = \min_{\mathbf{y}} \ \nabla f(\mathbf{x}_k)^\top(\mathbf{y}-\mathbf{x}_k) + \frac{1}{2}(\mathbf{y}-\mathbf{x}_k)^\top\nabla^2 f(\mathbf{x}_k)(\mathbf{y}-\mathbf{x}_k) + \frac{M}{6}\|\mathbf{y}-\mathbf{x}_k\|^3$$

Denote $T_M(\mathbf{x}_k)$ as an optimal solution of this problem for the case that cubic parameter is M. The parameter M should be selected based on the constant L_2. If L_2 is known we simply set $M := L_2$ and our next iterate can be defined as $\mathbf{x}_{k+1} = T_{L_2}(\mathbf{x}_k)$. However, we often don't know the exact value of L_2. In that case, we run a backtracking line-search on M to find a valid choice.
To define our method let us define the following function

$$f_M(\mathbf{x}) := \min_{\mathbf{y}} \ \left\{ f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^\top\nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{M}{6}\|\mathbf{y}-\mathbf{x}\|^3 \right\}$$

The formal version of the CRN method is the following: (for the case that L_2 is unknown)

Initialization step: Select $\mathbf{x}_0$ and L_0 that is smaller than L_2 and set $M_0 = L_0$

At step k:

- Set $M = M_k$
- Solve the cubic problem with M and find $T_M(\mathbf{x}_k)$
- If $f(T_M(\mathbf{x}_k)) \leq f_M(\mathbf{x}_k)$ then set $\mathbf{x}_{k+1} = T_M(\mathbf{x}_k)$ and go to step $k+1$
- If $f(T_M(\mathbf{x}_k)) > f_M(\mathbf{x}_k)$ then set $M_k = 2M_k$ and go back to the second step

We will show that M_k won't be larger than $2L_2$

2.1 Important Properties

Note that the optimality condition of the cubic problem implies that

$$\nabla f(\mathbf{x}_k) + \nabla^2 f(\mathbf{x}_k)(T_M(\mathbf{x}_k) - \mathbf{x}_k) + \frac{M}{2}\|T_M(\mathbf{x}_k) - \mathbf{x}_k\|(T_M(\mathbf{x}_k) - \mathbf{x}_k) = \mathbf{0} \tag{1}$$

which leads to the following expression if we multiply both sides by $T_M(\mathbf{x}_k) - \mathbf{x}_k$

$$\nabla f(\mathbf{x}_k)^\top (T_M(\mathbf{x}_k) - \mathbf{x}_k) + (T_M(\mathbf{x}_k) - \mathbf{x}_k)^\top \nabla^2 f(\mathbf{x}_k)(T_M(\mathbf{x}_k) - \mathbf{x}_k) + \frac{M}{2}\|\mathbf{x}_k - T_M(\mathbf{x}_k)\|^3 = 0 \tag{2}$$

One can further show that the the solution of the cubic problem satisfies the following condition

Lemma 1.

$$\nabla^2 f(\mathbf{x}_k) + \frac{M}{2}\|\mathbf{x}_k - T_M(\mathbf{x}_k)\|\mathbf{I} \ \succeq\ \mathbf{0}$$

Using the above result, one can show that the CRN update leads to a descent direction.

Lemma 2.

$$\nabla f(\mathbf{x}_k)^\top (T_M(\mathbf{x}_k) - \mathbf{x}_k) \leq 0$$

Proof. Simply multiply both sides of the result in Lemma 1 by $\mathbf{x}_{k+1} - \mathbf{x}_k$ from left and right to obtain

$$(T_M(\mathbf{x}_k) - \mathbf{x}_k)^\top \nabla^2 f(\mathbf{x}_k)(T_M(\mathbf{x}_k) - \mathbf{x}_k) + \frac{M}{2}\|\mathbf{x}_k - T_M(\mathbf{x}_k)\|^3 \geq 0$$

now the claim follows simply from (2) □

An important property of minimizing the cubic loss function is $f(T_M(\mathbf{x}_k)) \leq f_M(\mathbf{x}_k)$ for $M \geq L_2$. This simply follows from the fact that

$$\begin{aligned}
f(T_M(\mathbf{x}_k)) &\leq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top (T_M(\mathbf{x}_k) - \mathbf{x}_k) + \frac{1}{2}(T_M(\mathbf{x}_k) - \mathbf{x}_k)^\top \nabla^2 f(\mathbf{x}_k)(T_M(\mathbf{x}_k) - \mathbf{x}_k) + \frac{L_2}{6}\|T_M(\mathbf{x}_k) - \mathbf{x}_k\|^3 \\
&\leq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top (T_M(\mathbf{x}_k) - \mathbf{x}_k) + \frac{1}{2}(T_M(\mathbf{x}_k) - \mathbf{x}_k)^\top \nabla^2 f(\mathbf{x}_k)(T_M(\mathbf{x}_k) - \mathbf{x}_k) + \frac{M}{6}\|T_M(\mathbf{x}_k) - \mathbf{x}_k\|^3 \\
&= f_M(\mathbf{x}_k)
\end{aligned}$$

Hence, we have

$$\boxed{f(T_M(\mathbf{x}_k)) \leq f_M(\mathbf{x}_k), \qquad \text{for } M \geq L_2}$$

We further know that

$$\begin{aligned} f(\mathbf{x}_k) - f_M(\mathbf{x}_k) &= -\nabla f(\mathbf{x}_k)^\top (T_M(\mathbf{x}_k) - \mathbf{x}_k) - \frac{1}{2}(T_M(\mathbf{x}_k) - \mathbf{x}_k)^\top \nabla^2 f(\mathbf{x}_k)(T_M(\mathbf{x}_k) - \mathbf{x}_k) - \frac{M}{6}\|T_M(\mathbf{x}_k) - \mathbf{x}_k\|^3 \\ &= -\frac{1}{2}\nabla f(\mathbf{x}_k)^\top (T_M(\mathbf{x}_k) - \mathbf{x}_k) + \frac{M}{12}\|T_M(\mathbf{x}_k) - \mathbf{x}_k\|^3 \\ &\geq \frac{M}{12}\|T_M(\mathbf{x}_k) - \mathbf{x}_k\|^3 \end{aligned}$$

where the second equality follows from (2). Hence, we have

$$\boxed{f(\mathbf{x}_k) - f_M(\mathbf{x}_k) \geq \frac{M}{12}\|T_M(\mathbf{x}_k) - \mathbf{x}_k\|^3}$$

By combining these results we have if $M \geq L_2$ then

$$\boxed{f(T_M(\mathbf{x}_k)) \leq f(\mathbf{x}_k) - \frac{M}{12}\|T_M(\mathbf{x}_k) - \mathbf{x}_k\|^3 \qquad \text{for } M \geq L_2}$$

Remark 1. *Note that for $M \geq L_2$ we know that f is monotonically decreasing when we follows the CRN update. Hence, the M_k selected by the line-search method would be at most $2L_2$.*

Lemma 3. *The total number of additional times that we will solve the cubic subproblem after N steps is bounded above by*

$$\sum_{k=0}^{N} i_k = \sum_{k=0}^{N} \log_2 \frac{M_{k+1}}{M_k} = \log \frac{M_{N+1}}{M_0} \leq \log \frac{2L_2}{L_0} = 1 + \log \frac{L_2}{L_0}$$

Hence, after N steps, overall we solve the cubic subproblem at most

$$N + 2 + \log \frac{L_2}{L_0}$$

3 Convergence in the nonconvex setting

Note that in nonconvex setting, a second-order stationary point is defined as

$$\nabla f(\hat{\mathbf{x}}) = \mathbf{0}, \qquad \nabla^2 f(\hat{\mathbf{x}}) \succeq \mathbf{0}$$

This is a necessary condition for a local minimum of the problem.
One can find an approximate version of this by finding an (ϵ, δ)-second-order stationary point

$$\|\nabla f(\hat{\mathbf{x}})\| \leq \epsilon, \qquad \nabla^2 f(\hat{\mathbf{x}}) \succeq -\delta \mathbf{I}$$

which can be also written as

$$\|\nabla f(\hat{\mathbf{x}})\| \leq \epsilon, \qquad -\lambda_{min}(\nabla^2 f(\hat{\mathbf{x}})) \leq \delta \mathbf{I}$$

We show that CRN can find such a point efficiently!
To do so we need to establish upper bounds on norm of gradient and the negative of the minimum eigenvalue of the Hessian in terms of $\|T_M(\mathbf{x}_k) - \mathbf{x}_k\|$.

Lemma 4.

$$\|\nabla f(T_M(\mathbf{x}_k))\| \leq \frac{L_2+M}{2}\|T_M(\mathbf{x}_k)-\mathbf{x}_k\|^2$$

Proof. Note that

$$\|\nabla f(T_M(\mathbf{x}_k)) - \nabla f(\mathbf{x}_k) - \nabla^2 f(\mathbf{x}_k)(T_M(\mathbf{x}_k)-\mathbf{x}_k)\| \leq \frac{L}{2}\|T_M(\mathbf{x}_k)-\mathbf{x}_k\|^2$$

Hence

$$\|\nabla f(T_M(\mathbf{x}_k))\| \leq \|\nabla f(\mathbf{x}_k)+\nabla^2 f(\mathbf{x}_k)(T_M(\mathbf{x}_k)-\mathbf{x}_k)\|+\frac{L}{2}\|T_M(\mathbf{x}_k)-\mathbf{x}_k\|^2 = \frac{M}{2}\|T_M(\mathbf{x}_k)-\mathbf{x}_k\|^2+\frac{L}{2}\|T_M(\mathbf{x}_k)-\mathbf{x}_k\|^2$$

where the last inequality follows from (1). □

This result implies that

$$\boxed{\sqrt{\frac{2}{L_2+M}}\sqrt{\|\nabla f(T_M(\mathbf{x}_k))\|} \leq \|T_M(\mathbf{x}_k)-\mathbf{x}_k\|}$$

Lemma 5.

$$\nabla^2 f(T_M(\mathbf{x}_k)) \;\succeq\; \nabla^2 f(\mathbf{x}_k) - L_2\|T_M(\mathbf{x}_k)-\mathbf{x}_k\|\mathbf{I} \;\succeq\; -\left(\frac{M}{2}+L_2\right)\|T_M(\mathbf{x}_k)-\mathbf{x}_k\|\mathbf{I}$$

Proof. The first inequality simply follows from L_2-Lipschitz continuity of the Hessian, and the second one follows from Lemma 1. □

The above result implies an important result that

$$\boxed{-\frac{2}{2L_2+M}\;\lambda_{min}(\nabla^2 f(T_M(\mathbf{x}_k))) \leq \|T_M(\mathbf{x}_k)-\mathbf{x}_k\|}$$

Theorem 1. *If we define e_k as*

$$e_k = \max\left\{\frac{2}{3L_2}\sqrt{\|\nabla f(\mathbf{x}_k)\|}, -\frac{1}{2L_2}\;\lambda_{min}(\nabla^2 f(\mathbf{x}_k))\right\}$$

then for the iterates of RCN we have

$$\min_{i=0,\ldots,k-1} e_k \leq \left(\frac{12(f(\mathbf{x}_0)-f^*)}{kL_0}\right)^{1/3}$$

Proof. Note that using the fact that

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{M_k}{12}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^3$$

we can show that

$$\sum_{i=0}^{k-1}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^3 \leq \frac{12}{L_0}(f(\mathbf{x}_0)-f(\mathbf{x}_k)) \leq \frac{12}{L_0}(f(\mathbf{x}_0)-f^*)$$

This result implies that

$$\min_{i=0,\ldots,k-1}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^3 \leq \frac{12(f(\mathbf{x}_0)-f^*)}{kL_0}$$

which implies that

$$\min_{i=0,\dots,k-1} \|\mathbf{x}_{i+1} - \mathbf{x}_i\| \leq \left(\frac{12(f(\mathbf{x}_0) - f^*)}{kL_0}\right)^{1/3}$$

Now note that

$$\max\left\{\frac{2}{3L_2}\sqrt{\|\nabla f(\mathbf{x}_{i+1})\|}, -\frac{1}{2L_2}\ \lambda_{min}(\nabla^2 f(\mathbf{x}_{i+1}))\right\} \leq \|\mathbf{x}_{i+1} - \mathbf{x}_i\|$$

Hence,

$$\min_{i=0,\dots,k-1} \max\left\{\sqrt{\frac{2}{3L_2}\|\nabla f(\mathbf{x}_{i+1})\|}, -\frac{1}{2L_2}\ \lambda_{min}(\nabla^2 f(\mathbf{x}_{i+1}))\right\} \leq \left(\frac{12(f(\mathbf{x}_0) - f^*)}{kL_0}\right)^{1/3}$$

□

This result shows that to find an (ϵ, δ)-SOSP the CRN requires

$$k = \mathcal{O}\left(\frac{1}{\epsilon^{3/2}} + \frac{1}{\delta^3}\right)$$

Note that for nonconvex functions GD requires

$$k = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$$

to find an (ϵ)-FOSP