EE C227C Convex Optimization and Approximation Lecture 22 - 04/07/2022

Lecture 22: Convergence of Cubic Regularization

Lecturer: Jiantao Jiao *Scribe: Kamyar Salahi and Licong Lin*

1 Motivation

In the previous lecture, we discussed approaches to mitigate the divergence of Newton's Method due to poor initialization. Most of these approaches were heuristics with few convergence guarantees. In the last lecture, we introduced cubic regularization as a method for second-order function optimization with global theoretical guarantees. However, it turns out that for certain function families, we can show that convergence with cubic regularization is considerably faster than our global convergence results from the last lecture.

Recall from the previous lecture, the formulation of cubic regularization:

$$T_M(x) = \arg\min[\langle \nabla f(x), (y-x)\rangle + \frac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle + \frac{M}{6}\|x-y\|^2]$$

For every iteration, we define $x_{k+1} = T_L(x_k)$.

We will now consider three function families and show convergence results for cubic regularization in each.

2 Star-Convex Functions

We define f to be a star-convex function if its set of global minima X^* is non-empty and for any $x^* \in X^*$ and any $x \in \mathbb{R}^n$ we have $f(\alpha x^* + (1-\alpha)x) \leq \alpha f(x^*) + (1-\alpha)f(x)$ for all $\alpha \in [0,1]$.

Theorem 1. *[Theorem 4.1.4 in [1]] Assume that the objective function f is star-convex and continuously twice-differentiable. Further, assume that $x \in \mathcal{F}$ where $\mathcal{F}$ is an open convex set, the sublevel set $\{x \in \mathbb{R}^n \parallel f(x) \leq f(x_0)\} \subseteq \mathcal{F}$, and F is bounded such that $diam(\mathcal{F}) = D$. Then, applying cubic regularization, we have*

1. *If the initial value of the objective function is large enough:*

$$f(x_0) - f(x^*) \geq \frac{3}{2}LD^3 \tag{1}$$

then after one step the algorithm converges as follows:

$$f(x_1) - f^* \leq \frac{1}{2}LD^3 \tag{2}$$

where L is the Lipschitz constant of the Hessian.

2. *If the initial value of the objective function is small:*

$$f(x_0) - f(x^*) \leq \frac{3}{2}LD^3 \tag{3}$$

then the rate of convergence of the process is fast:

$$f(x_k) - f(x^*) \leq \frac{3LD^3}{2(1+\frac{k}{3})^2} \tag{4}$$

Note:

$$\operatorname{diam}(\mathcal{F}) = \sup_{x,y\in\mathcal{F}} \|x - y\|$$

From 1, our first stage of convergence shows that if the function value at the initial point is very large for a star-convex function f, then a single step will get you very close to $f(x^*)$. In the second stage of convergence, we are no longer converging as quickly.

3 Global Non-Degenerate Functions

We define the optimal set X^* of a function f to be globally non-degenerate if $\exists \mu > 0$ such that for any $x \in \mathcal{F}$, $f(x) - f(x^*) \geq \frac{\mu}{2}\rho^2(x, X^*)$ where X^* is the set of global minima and $\rho(x, X^*) = \inf_{y\in X^*} \|x - y\|$.

Global non-degeneracy holds for strongly convex functions where the function's growth is similarly lower bounded by the square of the distance. In the case of a convex function, X^* will be a singleton. However, global non-degeneracy can hold for non-convex functions as well:

Example 2. Consider $f(x) = (\|x\|^2 - 1)^2$. The set of minima is defined as $X^* = \{x \mid \|x\| = 1\}$. This function is not convex but still globally non-degenerate.

Theorem 3. *[Theorem 4.1.5 in [1]] Assume that the objective function f is star-convex and has a globally non-degenerate optimal set. Then, applying cubic regularization, we have*

1. *If $f(x_0) - f(x^*) \geq \frac{4}{9}\bar{\omega}$, then at the first phase of the process we have the following rate of convergence:*

$$f(x_k) - f(x^*) \leq \left[(f(x_0) - f(x^*))^{\frac{1}{4}} - \frac{k}{6}\sqrt{\frac{2}{3}}\bar{\omega}^{\frac{1}{4}} \right]^4 . \tag{5}$$

   This phase is ends as soon as $f(x_{k_0} - f(x^*) \leq \frac{4}{9}\bar{\omega})$ for some $k_0 \geq 0$.

2. *For $k \geq k_0$, the sequence converges super linearly:*

$$f(x_{k+1}) - f(x^*) \leq \frac{1}{2}(f(x_k) - f(x^*))\sqrt{\frac{f(x_k) - f(x^*)}{\bar{\omega}}} \tag{6}$$

where $\bar{\omega} := \frac{1}{L^2}\left(\frac{\mu}{2}\right)^3$

4 Gradient-Dominated Functions

We define a function f to be gradient dominated of degree $p \in [1, 2]$ if it attains a global minimum at some point x^* and for all $x \in \mathcal{F}$ we have

$$f(x) - f(x^*) \leq \tau_f \|\nabla f(x)\|^p , \tag{7}$$

where τ_f is a positive constant and p is the degree of domination.

The definition of global dominance is useful since there exist cases where the gradient grows very slowly, meaning that we would like our bounds to be in terms of the $\nabla f(x)$ rather than x. For example, convex functions that are not strongly convex will satisfy gradient dominance but will not satisfy global non-degeneracy.

Example 4. Let f be convex on $\mathbb{R}^n$ with its minimum at x^*. By the definition of convexity,

$$f(x) - f(x^*) \leq \langle \nabla f(x), x - x^* \rangle \leq \|\nabla f(x)\| \, \|x - x^*\|$$

where the second inequality follows by Cauchy-Schwarz. Since we have a bounded domain, we can further utilize $\|x - x^*\| \leq D$ to state that $f(x) - f(x^*) \leq \|\nabla f(x)\| D$. In this case, we have $p = 1$ and $\tau_f = D$.

Example 5. Let f be differentiable and strongly convex on $\mathbb{R}^n$ with its minimum at x^*. By the definition of strong convexity, we have that

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

By minimizing both sides of this inequality over y, we find that

$$f(x) - f(x^*) \leq \frac{1}{2\mu} \|\nabla f(x)\|^2, \ \forall x \in \mathbb{R}^n.$$

In this case, we have $p = 2$ with $\tau_f = \frac{1}{2\mu}$.

For gradient dominated functions, we have the following result.

Theorem 6. *[Theorem 4.1.7 in [1]] Assume f has L-Lipschitz continuous Hessian $\nabla^2 f(x)$ and is gradient dominated with parameter τ_f and of degree $p = 2$. Applying cubic regularization to minimize f, we have*

1. *If the initial value of the objective function is large enough:*

$$f(x_0) - f(x^*) \geq \tilde{\omega} := \frac{1}{324 L^2 \tau_f^3} \tag{8}$$

then at its first phase the process converges as follows:

$$f(x_k) - f(x^*) \leq (f(x_0) - f(x^*)) \cdot e^{-k \cdot \sigma} \tag{9}$$

where $\sigma = \frac{\tilde{\omega}^{1/4}}{\tilde{\omega}^{1/4} + (f(x_0) - f(x^*))^{1/4}}$. This phase ends at the first iteration k_0 for which (8) does not hold.

2. *For $k \geq k_0$, the rate of convergence is super-linear:*

$$f(x_{k+1}) - f(x^*) \leq \tilde{\omega} \cdot \left(\frac{f(x_k) - f(x^*)}{\tilde{\omega}} \right)^{4/3} \tag{10}$$

From Theorem 6, we see in the first stage the gradient dominated functions converge at a linear rate. This is faster than the rate in the first stage for the globally nondegenerate functions. Moreover, cubic regularization has an even faster super-linear convergence rate in the second stage, when the difference $f(x_k) - f^*$ is small enough.

So far we have talked about several assumptions (star convexity, globally nondegeneracy, gradient dominance) on f such that the cubic regularization can achieve better convergence rate. In the next section we will introduce how to solve the cubic regularization problem.

5 Implementation of cubic regularization

To apply cubic regularization problem, we need to solve the following regularization problem

$$\min_{h \in \mathbb{R}^n} \left[v(h) := \langle g, h \rangle + \frac{1}{2} \langle Hh, h \rangle + \frac{M}{6} \|h\|^3, \right] \tag{11}$$

where H is the Hessian. When H is definite the problem is convex and hence can be solved efficiently. However, when H is indefinite, this problem is non-convex. Nevertheless, we can solve the problem through Lagrangian duality.

Note that (11) can be expressed as

$$v(h) = \min_{\tau \in \mathbb{R}} \left\{ \tilde{v}(h,\tau) \stackrel{\text{def}}{=} \langle g, h \rangle + \frac{1}{2}\langle Hh, h \rangle + \frac{M}{6}|\tau|^{3/2} : \|h\|^2 \leq \tau \right\}$$

Thus $T_M(x)$ solves the following problem

$$\min_{h \in \mathbb{R}^n, \tau \in \mathbb{R}} \left[\tilde{v}(h,\tau) : f(h,\tau) \stackrel{\text{def}}{=} \frac{1}{2}\|h\|^2 - \frac{1}{2}\tau \leq 0 \right]$$

Since this is already a constrained minimization problem, we can form for it a Lagrangian dual problem. Indeed, define the Lagrangian $\mathcal{L}(h,\tau,\lambda) = \tilde{v}(h,\tau) + \lambda\left[\frac{1}{2}\|h\|^2 - \frac{1}{2}\tau\right]$ with $h \in \mathbb{R}^n$ and $\tau, \lambda \in \mathbb{R}$. Then the dual function is

$$\psi(\lambda) = \inf_{h \in \mathbb{R}^n, \tau \in \mathbb{R}} \left\{ \langle g, h \rangle + \frac{1}{2}\langle Hh, h \rangle + \frac{M}{6}|\tau|^{3/2} + \lambda \left[\frac{1}{2}\|h\|^2 - \frac{1}{2}\tau \right] \right\}$$

The optimal value of τ can be found from the equation $\frac{M}{4}|\tau|^{1/2}\operatorname{sign}(\tau) = \frac{1}{2}\lambda$. Therefore, $\tau(\lambda) = \frac{4\lambda|\lambda|}{M^2}$, and we have

$$\psi(\lambda) = \inf_{h \in \mathbb{R}^n} \left\{ \langle g, h \rangle + \frac{1}{2}\left\langle \left(H + \lambda I_n\right) h, h \right\rangle - \frac{2}{3M^2}|\lambda|^3 \right\},$$

$$\operatorname{dom}\, \psi = \left\{ \lambda \in \mathbb{R} : \inf_{h \in \mathbb{R}^n} \left[q_\lambda(h) \stackrel{\text{def}}{=} \langle g, h \rangle + \frac{1}{2}\left\langle \left(H + \lambda I_n\right) h, h \right\rangle \right] > -\infty \right\}.$$

Without loss of generality, assume that $H = \operatorname{diag}\{H_1, ..., H_n\}$ is a diagonal matrix and define $H_{\min} = \min_{1 \leq i \leq n} H_i$.

If $\lambda > -H_{\min}$, then $\lambda \in \operatorname{dom}\, \psi$ since $q_\lambda(h)$ is convex. If $\lambda < -H_{\min}$, then $\lambda \notin \operatorname{dom}\, \psi$. Thus, only the status of the point $\lambda = -H_{\min}$ can be different. Define

$$G^2 = \sum_{i \in I^*} \left(g^{(i)}\right)^2, \quad I^* = \{i : H_i = H_{\min}\}.$$

There are three possibilities.

1. $G^2 > 0$. Then $\operatorname{dom}\, \psi = \{\lambda \in \mathbb{R} : \lambda > -H_{\min}\}$. For any λ in this domain we have

$$\psi(\lambda) = -\frac{1}{2}\frac{G^2}{H_{\min} + \lambda} - \frac{1}{2}\sum_{i \notin I^*} \frac{\left(g^{(i)}\right)^2}{H_i + \lambda} - \frac{2}{3M^2}|\lambda|^3.$$

   Meanwhile, the optimal vector for the function $q_\lambda(\cdot)$ has the form

$$h(\lambda) = -\left(H + \lambda I_n\right)^{-1} g.$$

   Since this vector and value $\tau(\lambda)$ are uniquely defined and continuous on $\operatorname{dom}\, \psi$], it follows from Theorem 1.3.2 in [1] that

$$\min_{h \in \mathbb{R}^n} v(h) = \max_{\lambda \in \operatorname{dom}\, \psi \cap \mathbb{R}_+} \psi(\lambda), \tag{12}$$

   and a global minimal point of (11) is $h(\lambda^*)$.

2. $G^2 = 0$ and $\lambda^* > H_{\min}$. In this case, for any $\lambda > -H_{\min}$, the optimal vector is uniquely defined as follows:

$$h^{(i)}(\lambda) = \begin{cases} \frac{g^{(i)}}{H_i + \lambda}, & \text{if } i \notin I^*, \\ 0, & \text{otherwise} \end{cases} \qquad i = 1, \ldots, n$$

This vector is continuous on dom ψ. Therefore, if

$$\lambda^* := \arg \max_{\lambda \in \operatorname{dom} \psi \bigcap \mathbb{R}_+} \psi(\lambda) > -H_{\min}$$

then the conditions of Theorem 1.3.2 in [1] are satisfied. Hence, in this case (12) is also valid.

3. $G^2 = 0$ and $\lambda^* = H_{\min}$.

For the third cases, similar arguments can be applied to solve the optimization problem. We refer to page 264 of [1] for more details.

References

[1] Y. Nesterov *et al.*, *Lectures on convex optimization*. Springer, 2018, vol. 137.