Basics of optimization theory

Cong Ma

University of Chicago, Autumn 2021

Unconstrained optimization

Consider an unconstrained optimization problem

$$\text{minimize}_{\boldsymbol{x}} \qquad f(\boldsymbol{x})$$

- For simplicity, we assume $f(\boldsymbol{x})$ is twice differentiable
- We assume the minimizer $\boldsymbol{x}_{\text{opt}}$ exists, i.e.,

$$\boldsymbol{x}_{\text{opt}} := \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

(Local) strong convexity and smoothness

Definition 7.1

A twice differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be α-strongly convex in a set $\mathcal{B}$ if for all $\boldsymbol{x} \in \mathcal{B}$

$$\nabla^2 f(\boldsymbol{x}) \succeq \alpha \boldsymbol{I}_n.$$

Definition 7.2

A twice differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be β-smooth in a set $\mathcal{B}$ if for all $\boldsymbol{x} \in \mathcal{B}$

$$\|\nabla^2 f(\boldsymbol{x})\| \leq \beta.$$

Gradient descent theory revisited

Gradient descent method with step size $\eta > 0$

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)$$

Lemma 7.3

Suppose f is α-strongly convex and β-smooth in the local ball $\mathcal{B}_\delta(\boldsymbol{x}_{\mathsf{opt}}) := \{\boldsymbol{x} \mid \|\boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}}\|_2 \leq \delta\}$. Running gradient descent from $\boldsymbol{x}^0 \in \mathcal{B}_\delta(\boldsymbol{x}_{\mathsf{opt}})$ with $\eta = 1/\beta$ achieves linear convergence

$$\|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2 \leq \left(1 - \frac{\alpha}{\beta}\right)^t \|\boldsymbol{x}^0 - \boldsymbol{x}_{\mathsf{opt}}\|_2, \quad t = 0, 1, 2, \ldots$$

Implications

- Condition number β/α determines rate of convergence
- Attains ε-accuracy (i.e., $\|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2 \leq \varepsilon \|\boldsymbol{x}_{\mathsf{opt}}\|_2$) within

  $$O\left(\frac{\beta}{\alpha}\log\frac{1}{\varepsilon}\right)$$

  iterations
- Needs initialization $\boldsymbol{x}^0 \in \mathcal{B}_\delta(\boldsymbol{x}_{\mathsf{opt}})$

Proof of Lemma 7.3

Since $\nabla f(\boldsymbol{x}_{\mathsf{opt}}) = \boldsymbol{0}$, we can rewrite GD as

$$\begin{aligned}\boldsymbol{x}^{t+1} - \boldsymbol{x}_{\mathsf{opt}} &= \boldsymbol{x}^{t} - \eta \nabla f(\boldsymbol{x}^{t}) - [\boldsymbol{x}_{\mathsf{opt}} - \eta \nabla f(\boldsymbol{x}_{\mathsf{opt}})] \\ &= \left[\boldsymbol{I}_n - \eta \int_0^1 \nabla^2 f(\boldsymbol{x}(\tau)) \mathsf{d}\tau \right] (\boldsymbol{x}^{t} - \boldsymbol{x}_{\mathsf{opt}}),\end{aligned}$$

where $\boldsymbol{x}(\tau) := \boldsymbol{x}_{\mathsf{opt}} + \tau(\boldsymbol{x}^{t} - \boldsymbol{x}_{\mathsf{opt}})$. By local strong convexity and smoothness, one has

$$\alpha \boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}(\tau)) \preceq \beta \boldsymbol{I}_n, \qquad \text{for all } 0 \le \tau \le 1$$

Therefore $\eta = 1/\beta$ yields

$$\boldsymbol{0} \preceq \boldsymbol{I}_n - \eta \int_0^1 \nabla^2 f(\boldsymbol{x}(\tau)) \mathsf{d}\tau \preceq (1 - \frac{\alpha}{\beta}) \boldsymbol{I}_n,$$

which further implies

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_{\mathsf{opt}}\|_2 \le \left(1 - \frac{\alpha}{\beta}\right) \|\boldsymbol{x}^{t} - \boldsymbol{x}_{\mathsf{opt}}\|_2$$

Regularity condition

More generally, for update rule

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^{t} - \eta \boldsymbol{g}(\boldsymbol{x}^{t}),$$

where $g(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}^n$

Definition 7.4

$\boldsymbol{g}(\cdot)$ is said to obey RC(μ, λ, δ) for some $\mu, \lambda, \delta > 0$ if

$$2\langle \boldsymbol{g}(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}} \rangle \geq \mu \|\boldsymbol{g}(\boldsymbol{x})\|_2^2 + \lambda \, \|\boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 \quad \forall \boldsymbol{x} \in \mathcal{B}_{\delta}(\boldsymbol{x}_{\mathsf{opt}})$$

- Negative search direction $\boldsymbol{g}$ is positively correlated with error $\boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}} \Longrightarrow$ one-step improvement
- $\mu\lambda \leq 1$ by Cauchy-Schwarz

RC = one-point strong convexity + smoothness

- One-point α-strong convexity:

$$f(\boldsymbol{x}_{\mathsf{opt}}) - f(\boldsymbol{x}) \geq \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}_{\mathsf{opt}} - \boldsymbol{x} \rangle + \frac{\alpha}{2} \|\boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 \qquad (7.1)$$

- β-smoothness:

$$\begin{aligned} f(\boldsymbol{x}_{\mathsf{opt}}) - f(\boldsymbol{x}) &\leq f\Big(\boldsymbol{x} - \frac{1}{\beta}\nabla f(\boldsymbol{x})\Big) - f(\boldsymbol{x}) \\ &\leq \Big\langle \nabla f(\boldsymbol{x}), -\frac{1}{\beta}\nabla f(\boldsymbol{x}) \Big\rangle + \frac{\beta}{2}\Big\|\frac{1}{\beta}\nabla f(\boldsymbol{x})\Big\|_2^2 \\ &= -\frac{1}{2\beta}\|\nabla f(\boldsymbol{x})\|_2^2 \qquad (7.2) \end{aligned}$$

RC = one-point strong convexity + smoothness

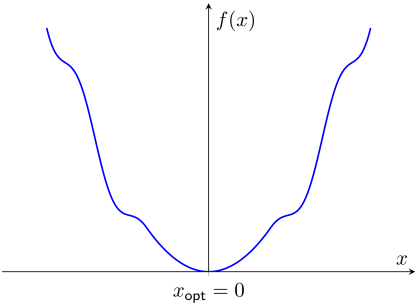
Combining (7.1) and (7.2) yields

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_{\text{opt}} \rangle \geq \frac{\alpha}{2} \|\boldsymbol{x} - \boldsymbol{x}_{\text{opt}}\|_2^2 + \frac{1}{2\beta} \|\nabla f(\boldsymbol{x})\|_2^2$$

— *RC holds with $\mu = 1/\beta$ and $\lambda = \alpha$*

Extension of convex functions

When $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$, f is not necessarily convex



$$f(x) = \begin{cases} x^2, & |x| \leq 6, \\ x^2 + 1.5|x|(cos(|x| - 6) - 1), & |x| > 6 \end{cases}$$

Convergence under RC

Lemma 7.5

Suppose $\boldsymbol{g}(\cdot)$ *obeys* RC(μ, λ, δ). *The update rule* $(\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \boldsymbol{g}(\boldsymbol{x}^t))$ *with* $\eta = \mu$ *and* $\boldsymbol{x}^0 \in \mathcal{B}_\delta(\boldsymbol{x}_{\mathsf{opt}})$ *obeys*

$$\|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 \leq (1 - \mu\lambda)^t \, \|\boldsymbol{x}^0 - \boldsymbol{x}_{\mathsf{opt}}\|_2^2$$

- $\boldsymbol{g}(\cdot)$: more general search directions
  - example: in vanilla GD, $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$
- The product $\mu\lambda$ determines the rate of convergence
- Attains ε-accuracy within $O(\frac{1}{\mu\lambda} \log \frac{1}{\varepsilon})$ iterations

Proof of Lemma 7.5

By definition, one has

$$
\begin{aligned}
\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 &= \|\boldsymbol{x}^t - \eta \boldsymbol{g}(\boldsymbol{x}^t) - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 \\
&= \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 + \eta^2 \|\boldsymbol{g}(\boldsymbol{x}^t)\|_2^2 - 2\eta \left\langle \boldsymbol{g}(\boldsymbol{x}^t), \boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}} \right\rangle \\
&\le \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 + \eta^2 \|\boldsymbol{g}(\boldsymbol{x}^t)\|_2^2 - \eta \left(\lambda \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 + \mu \|\boldsymbol{g}(\boldsymbol{x}^t)\|_2^2 \right) \\
&= (1 - \eta\lambda) \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 + \eta(\eta - \mu) \|\boldsymbol{g}(\boldsymbol{x}^t)\|_2^2 \\
&\le (1 - \eta\mu) \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2^2
\end{aligned}
$$