Unconstrained minimization

Topics

- gradient descent method
- Newton's method
- convergence rate analysis
- self-concordant functions

Unconstrained minimization

Problem:

$$\min_{x \in \mathrm{dom}\, f} \quad f(x)$$

Assumptions:

- f is convex, twice continuously differentiable (means $\mathrm{dom}\, f$ is open)
- optimal value $p^* = \inf_x f(x)$ is attained and finite

Iterative methods:

- produce sequence of points $(x^{(0)}, x^{(1)}, \cdots, x^{(k)}, \cdots) \in \mathrm{dom}\, f$ with

$$f(x^{(k)}) \to p^*$$

- initial point $x^{(0)}$ and sublevel set
  - $x^{(0)} \in \mathrm{dom}\, f$
  - sublevel set $S = \{x \mid f(x) \leq f(x^{(0)}\}$ is closed

Descent methods

---

given a starting point $x \in \text{dom } f$
repeat

1. Determine a descent direction Δx
2. Line search: choose a step size $t > 0$
3. Update: $x := x + t\Delta x$

until stopping criterion is satisfied

Notation:

$$x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)}$$

- $\Delta x^{(k)}$ descent direction: $f(x^{(k+1)}) < f(x^{(k)})$
- $t > 0$ is the step size
- f convex $\implies \nabla f(x^{(k)})^T \Delta x^{(k)} < 0$

Linear search methods

1. **Exact line search:**

$$t = \text{argmin}_{t>0} \; f(x + t\Delta x)$$

2. **Backtracking line search:**

---

given a descent direction Δx, $\alpha \in (0, 1/2)$, $\beta \in (0, 1)$
$t := 1$
while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x, \quad t := \beta t$

Gradient descent methods

descent method with $\Delta x = -\nabla f(x)$

---

given a starting point $x \in \mathbf{dom}\, f$
repeat

1. $\Delta x := -\nabla f(x)$
2. Line search: choose a step size $t > 0$
3. Update: $x := x + t\Delta x$

until stopping criterion is satisfied

- stopping criterion: $\|\nabla f(x)\|_2 \leq \epsilon$
- convergence result: for strongly convex f,

$$f(x^{(k)}) - p^* \leq c^k (f(x^{(0)}) - p^*)$$

where $c \in (0, 1)$, called linear convergence

Newton's method

Newton step:

$$\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

interpretations:

- $x + \Delta x_{nt}$ minimizes the second order approximation of f

$$\hat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$

- $x + \Delta x_{nt}$ solves linearized optimality condition

$$\nabla f(x+v) \approx \nabla \hat{f}(x+v) = \nabla f(x) + \nabla^2 f(x) v = 0$$

Newton decrement

$$\lambda(x) = (\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x))^{1/2}$$

a measure of proximity of x to x^*

perperties

- provides an estimate of $f(x) - p^*$, using quadratic approximation $\hat{f}$:

$$f(x) - \inf_y \hat{f}(y) = \frac{1}{2}\lambda(x)^2$$

- directional derivative in the Newton's direction

$$\nabla f(x)^T \Delta x_{nt} = -\lambda(x)^2$$

- norm of Newton step in the Hessian norm

$$\lambda(x) = (\Delta x_{nt}^T \nabla^2 f(x) \Delta x_{nt})^{1/2}$$

- affine invariant

Newton's method

---

given a starting point $x \in \operatorname{dom} f$, tolerance $\epsilon > 0$
repeat

1. Compute the Newton step and decrement

$$\Delta x_{nt} := -\nabla^2 f(x)^{-1} \nabla f(x) \quad \lambda(x)^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$$

2. Stopping criterion: quit if $\lambda^2/2 \leq \epsilon$
3. Line search: choose a step size $t > 0$ by backtracking line search
4. Update: $x := x + t\Delta x$

Rate of convergence

Let $z_1, z_2, \cdots, z_n, \cdots, \to z$ be a convergent sequence. The **rate of convergence** of this sequence is r^* if

$$r^* = \sup \left\{ r \mid \limsup_{k\to\infty} \frac{|z_{k+1} - z|}{|z_k - z|^r} < \infty \right\}$$

Types of convergence: Let

$$\beta = \limsup_{k\to\infty} \frac{|z_{k+1} - z|}{|z_k - z|^{r^*}}$$

1. $p^* = 1, 0 < \beta < 1$: **linear** (geometric) rate of convergence
2. $p^* = 1, \beta = 0$: **super-linear** rate of convergence
3. $p^* = 1, \beta = 1$: **sub-linear** rate of convergence
4. $p^* = 2$: **quadratic** rate of convergence

Rate of convergence: examples

1. $z_k = a^k$, $a \in (0,1)$: $\beta = a$, converge linearly to 0
2. $z_k = a^{2^k}$, $a \in (0,1)$: converge quadratically to 0
3. $z_k = a^{k^2}$, $a \in (0,1)$: converge super-linearly to 0
4. $z_k = a^{\sqrt{k}}$, $a \in (0,1)$: converge sub-linearly to 0
5. $z_k = 1/k$: converge super-linearly to 0

Strong convexity

Definition (strong convexity)

f is strongly convex on S if $\exists m > 0$ such that

$$\nabla^2 f(x) \succeq m\mathrm{I} \quad \forall x \in S$$

If f is strongly convex, then

- for all $x, y \in S$,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}\|x - y\|_2^2$$

hence S is bounded. (assume $\inf f$ is attained and finite)

Because S is bounded and $\lambda_{\max}(\nabla^2 f(x))$ is a continuous function of x, so $\lambda_{\max}(\nabla^2 f(x))$ is bounded above in S. That is, $\exists M > 0$ such that

$$\nabla^2 f(x) \preceq M\mathrm{I} \quad \forall x \in S$$

Strong convexity: property

If $p^* > -\infty$, then for all $x \in S$,

- lower and upper bound on Hessian

$$mI \preceq \nabla^2 f(x) \preceq MI$$

- upper bound on difference

$$f(x) - p^* \leq \frac{1}{2m} \|\nabla f(x)\|_2^2$$

Convergence analysis: gradient descent

$$f(x^{(k)}) - p^* \le c^k (f(x^{(0)}) - p^*)$$

where $c < 1$

1. Exact line search $c = 1 - m/M$
2. Backtracking line search: $c = 1 - \min\{2m\alpha, 2\alpha\beta m/M\}$

Remark:

- Linear rate of convergence. The number of iterations is bounded above by
$$\frac{\log((f(x^{(0)}) - p^*)/\epsilon)}{\log(1/c)}$$
- Conditional number of Hessian: $k(\nabla^2 f(x)) = \lambda_{\max}/\lambda_{\min} \le M/m$
- Slow when $k(\nabla^2 f(x))$ is large.

Convergence analysis: Newton's method

Assumptions

- f is strongly convex on S: $\nabla^2 f(x) \succeq mI$
- $\nabla^2 f$ is Lipschitz continuous on S, with constant $L > 0$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L\|x - y\|_2$$

Convergence analysis: there exist constants $\eta \in (0, m^2/L), \gamma > 0$ such that

- if $\|\nabla f(x^{(k)})\|_2 \geq \eta$, then $f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma$
- if $\|\nabla f(x^{(k)})\|_2 < \eta$, then

$$\frac{L}{2m^2}\|\nabla f(x^{(k+1)}\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x^{(k)}\|_2\right)^2$$

Convergence analysis

damped Newton phase: $\|\nabla f(x)\|_2 \geq \eta$

- mostly requires backtracking steps
- requires $f(x^{(0)} - p^*)/\gamma$ iterations

quadratically convergent phase: $\|\nabla f(x)\|_2 < \eta$

- no need for backtracking: $t = 1$ always
- quadratical convergence:

$$\frac{L}{2m^2}\|\nabla f(x^{(l)}\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x^{(k+1)}\|_2\right)^{2^{l-k}} \leq \left(\frac{1}{2}\right)^{2^{l-k}}$$

The total number of iterations is bounded above by

$$\frac{f(x^{(0)}) - p^*}{\gamma} + \log_2 \log_2(\epsilon_0/\epsilon)$$

where $\epsilon_0 = 2m^3/L^2$.

Self-concordant functions

Definition (Self-concordant function in R)

A convex function $f : R \to R$ is **self-concordant** if

$$|f'''(x)| \leq 2f''(x)^{3/2}, \quad \forall x \in \text{dom } f$$

Definition (Self-concordant function in R^n)

A convex function $f : R^n \to R$ is **self-concordant** if it is self-concordant along every line in its domain, i.e., if $\tilde{f}(t) = f(x + tv)$ is a self-concordant function of t for all $x \in \text{dom}\, f$ and all v.

Self-concordant functions: examples

- affine or quadratic functions
- negative log: $f(x) = -\log x$
- log barrier for linear inequalities: $f(x) = -\sum_{i=1}^{m} \log(b_i - a_i^T x)$
- log-determinant: $f(X) = -\log \det X$

Complexity of Newton's method for self-concordant functions

If the objective function is self-concordant, then the number of Newton iterations is bounded above by

$$\frac{f(x^{(0)}) - p^*}{\gamma} + \log_2 \log_2(1/\epsilon)$$

with

$$\gamma = \frac{\alpha\beta(1-2\alpha)^2}{20-8\alpha}$$

Example: with $\alpha = 0.1, \beta = 0.8$, and tolerance $\epsilon = 10^{-10}$, the bound is

$$375(f(x^{(0)}) - p^*) + 6$$