Gradient Descent Algorithm

Yuqian Zhang

Rutgers University

yqz.zhang@rutgers.edu

February 22, 2021

Overview

1. Choice of step size
2. Convergence of gradient descent
   - Strongly convex
   - Lipschitz gradient condition

General descent algorithm

Choose a starting point $x^{(0)}$
Do

- determine a descent direction $d^{(k)}$
- choose a step size $t \geq 0$
- update $x^{(k)}$ as $x^{(k-1)} + td^{(k)}$
- check convergence criteria

until convergence

General descent algorithm

- **Gradient Descent**

$$d^{(k)} = -\nabla f(x^{(k-1)})$$

  - This is the direction of steepest descent in ℓ^2
  - Gradient descent iterations are cheap, but typically many iterations are required for convergence.

- **Newton's method**

$$d^{(k)} = -(\nabla^2 f(x^{(k-1)}))^{-1}\nabla f(x^{(k-1)})$$

  - tend to be expensive (as they require a system solve), but they typically converge in far fewer iterations than gradient descent

Line search

- Exact step size
  - Solve the 1D optimization problem

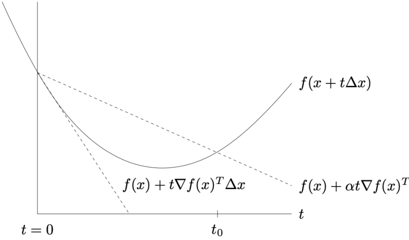
$$\text{minimize} \quad f(\boldsymbol{x}^{(k-1)} + t\boldsymbol{d}^{(k)})$$

  - heavy computation
  - unless there exist analytical solution
- Fixed step size
  - works well when the step size is small enough
  - too many iterations

Backtracking line search

Start with a step size of $t = 1$, then decrease by a factor of β until the update is below a certain line.

- Fix $\alpha \in (0, 0.5)$ and $\beta \in (0, 1)$
- Given a starting point $\boldsymbol{x}$ and direction $\boldsymbol{d}$
- $t = 1$
- Repeat
    - if $f(\boldsymbol{x} + t\boldsymbol{d}) < f(\boldsymbol{x}) + \alpha t \langle \boldsymbol{d}, \nabla f(\boldsymbol{x}) \rangle$, converged
    - else $t = \beta t$
- until convergence



Strong convexity

f is twice differentiable

$$m\boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq M\boldsymbol{I}$$

- the eigenvalues of the Hessian are bounded between $m > 0$ and $M < \infty$.
- Lower bounds implies strict convexity $\nabla^2 f(\boldsymbol{x}) > \boldsymbol{0}$.

Basic inequalities

- By convexity

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

- By Taylor's Theorem

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle + \frac{1}{2} (\boldsymbol{y} - \boldsymbol{x})^T \nabla^2 f(\boldsymbol{z}) (\boldsymbol{y} - \boldsymbol{x})$$

then

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle + \frac{m}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2$$
$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle + \frac{M}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2$$

Basic inequalities

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle + \frac{m}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2$$

Minimizing the right hand side over $\boldsymbol{y}$, the optimal solution is

$$\tilde{\boldsymbol{y}} = \boldsymbol{x} - m^{-1} \nabla f(\boldsymbol{x})$$

plugging $\tilde{\boldsymbol{y}}$ into the right hand side yields

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) - \frac{1}{2m} \|\nabla f(\boldsymbol{x})\|_2^2$$

Hence the optimal value satisfies

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

Basic inequalities

$$
\begin{aligned}
p^* = f(\boldsymbol{x}^*) &\geq f(\boldsymbol{x}) + \langle \boldsymbol{x}^* - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle + \frac{m}{2} \|\boldsymbol{x}^* - \boldsymbol{x}\|_2^2 \\
&\geq f(\boldsymbol{x}) - \|\boldsymbol{x}^* - \boldsymbol{x}\|_2 \|\nabla f(\boldsymbol{x})\|_2 + \frac{m}{2} \|\boldsymbol{x}^* - \boldsymbol{x}\|_2^2
\end{aligned}
$$

since $p^* \leq f(\boldsymbol{x})$,

$$
- \|\boldsymbol{x}^* - \boldsymbol{x}\|_2 \|\nabla f(\boldsymbol{x})\|_2 + \frac{m}{2} \|\boldsymbol{x}^* - \boldsymbol{x}\|_2^2 \leq 0
$$

and so

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

Convergence of gradient descent

- **exact line search**
- at each iteration, the gap $f(x^{(k)}) - p^*$ gets cut down by a fixed factor.
- use x to denote the current point, and $x^+ = x - t_{exact}\nabla f(x)$ to denote the result of the gradient step.
- choose t_{exact} by minimizing the following function:

$$\tilde{f}(t) = f(x - t\nabla f(x))$$

Convergence of gradient descent

- By strong convexity

$$\tilde{f}(t) \leq f(x) - t \|\nabla f(x)\|_2^2 + \frac{Mt^2}{2} \|\nabla f(x)\|_2^2$$

- By definition of t_{exact}, we know

$$f(x^+) = \tilde{f}(t_{exact}) \leq \tilde{f}(1/M) \leq f(x) - \frac{1}{2M} \|\nabla f(x)\|_2^2$$

- since $p^* \geq f(x) - \frac{1}{2m} \|\nabla f(x)\|_2^2$, then

$$\|\nabla f(x)\|_2^2 \geq 2m(f(x) - p^*)$$

Convergence of gradient descent

- therefore

$$f(x^+) - p^* \leq f(x) - p^* - \frac{m}{M}(f(x) - p^*)$$

- which means

$$\frac{f(x^+) - p^*}{f(x) - p^*} \leq \left(1 - \frac{m}{M}\right)$$

- the gap between the current functional evaluation and the optimal value has been cut down by a factor of $1 - m/M < 1$

Convergence of gradient descent

- Applying this inequality recursively

$$\frac{f(x^{(k)}) - p^*}{f(x^{(0)}) - p^*} \leq \left(1 - \frac{m}{M}\right)^k$$

- Another way to say this is that we can achieve accuracy

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

by taking steps

$$k \geq \frac{\log(E_0/\epsilon)}{\log(1 - m/M)}, \quad E_0 = f(x^{(0)}) - p^*$$

Lipschitz gradient condition

- Similar results for gradient descent on strongly convex functions using backtracking – with the same linear convergence but with constants that depend on α and β along with m and M.
- We can also get (much weaker) convergence results when f is not strongly convex (or even necessarily twice differentiable), but has a **Lipschitz gradient**

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2 , \quad L > 0$$

- Upper bound

$$f(y) \leq f(x) + \langle y - x, \nabla f(x) \rangle + \frac{L}{2} \|y - x\|_2^2$$

Lipschitz gradient condition

From fundamental theorem of calculus

$$f(y) - f(x) = \int_0^1 \langle y - x, \nabla f((1-t)x + ty) \rangle \, dt$$

then

$$\begin{aligned} f(y) - f(x) - \langle y - x, \nabla f(x) \rangle &= \int_0^1 \langle y - x, \nabla f((1-t)x + ty) - \nabla f(x) \rangle \, dt \\ &\le \|y - x\|_2 \int_0^1 \|\nabla f((1-t)x + ty) - \nabla f(x)\|_2 \\ &\le L \|y - x\|_2^2 \int_0^1 t \; dt \\ &\le \frac{L}{2} \|y - x\|_2^2 \end{aligned}$$

Convergence of gradient descent

- **fixed step size** $t \leq 1/L$

$$
\begin{aligned}
f(x^+) &\leq f(x) - \left(1 - \frac{Lt}{2}\right) t \left\|\nabla f(x)\right\|_2^2 \\
&\leq f(x) - \frac{t}{2} \left\|\nabla f(x)\right\|_2^2
\end{aligned}
$$

- By convexity

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

Convergence of gradient descent

Therefore

$$f(x^+) \leq f(x^*) + \langle x - x^*, \nabla f(x) \rangle - \frac{t}{2} \|\nabla f(x)\|_2^2$$

Substituting $\nabla f(x) = (x - x^+) / t$ yields

$$\begin{aligned} f(x^+) - f(x^*) &\leq \frac{1}{t} \left\langle x - x^*, x - x^+ \right\rangle - \frac{1}{2t} \left\| x - x^+ \right\|_2^2 \\ &= \frac{1}{2t} \left(\left\langle x - x^*, x - x^+ \right\rangle - \left\langle x^* - x^+, x - x^+ \right\rangle \right) \\ &= \frac{1}{2t} \left(\left\| x - x^* \right\|_2^2 - \left\| x^+ - x^* \right\|_2^2 \right) \end{aligned}$$

Convergence of gradient descent

Summing over k iterations:

$$\begin{aligned}
\sum_{i=1}^{k} f(x^{(i)}) - f(x^*) &\leq \frac{1}{2t}\left(\sum_{i=1}^{k} \left\|x^{(i-1)} - x^*\right\|_2^2 - \left\|x^{(i)} - x^*\right\|_2^2\right) \\
&= \frac{1}{2t}\left(\left\|x^{(0)} - x^*\right\|_2^2 - \left\|x^{(k)} - x^*\right\|_2^2\right) \\
&\leq \frac{1}{2t}\left\|x^{(0)} - x^*\right\|_2^2
\end{aligned}$$

and the k-th term is smaller than average, then

$$\begin{aligned}
f(x^{(k)}) - f(x^*) &\leq \frac{1}{k}\sum_{i=1}^{k} f(x^{(i)}) - f(x^*) \\
&\leq \frac{1}{2tk}\left\|x^{(0)} - x^*\right\|_2^2
\end{aligned}$$

Convergence of gradient descent

- Strongly convex

$$f(x^{(k)}) - p^* \leq \left(1 - \frac{m}{M}\right)^k \left(f(x^{(0)}) - p^*\right)$$

- Lipschitz gradient

$$f(x^{(k)}) - f(x^*) \leq \frac{1}{2tk} \left\| x^{(0)} - x^* \right\|_2^2$$

General descent algorithm

- **Gradient Descent**

$$d^{(k)} = -\nabla f(x^{(k-1)})$$

  - This is the direction of steepest descent
  - Gradient descent iterations are cheap, but typically many iterations are required for convergence.

- **Newton's method**

$$d^{(k)} = -(\nabla^2 f(x^{(k-1)}))^{-1} \nabla f(x^{(k-1)})$$

  - tend to be expensive (as they require a system solve), but they typically converge in far fewer iterations than gradient descent