DS-GA 1003 Machine Learning
Lecture 2

Feb 9, 2021

1 Gradient Descent

- Gradient is the steepest ascent direction.
  - Derivative tells us how much the function value $f(x)$ changes if we move x a tiny bit.
  - For multivariable functions, we need directional derivatives to know how fast $f(x)$ changes along u.
  - The fastest ascent direction is given by

    $$\arg\max_{\|u\|_2=1} \nabla f(x) \cdot u = \frac{\nabla f(x)}{\|\nabla f(x)\|_2}$$

    * Show by Cauchy-Schwarz.
    * (draw) Geometric explanation: $a \cdot b = \|a\|_2 \|b\|_2 \cos\theta$.
- Where does gradient descent converge?
  - Stationary/Critical points: x where $\nabla f(x) = 0$.
  - (draw) Local/global minimum/maximum, flat region of critical points
  - (draw) Are all critical points local minima/maxima? [no, saddle points.]
  - In general, GD converges to stationary points. With certain conditions (e.g. f is convex, gradient cannot change arbitrarily fast, small step size), we can reach global minimum.
- What is the true "step size"?
  - $\eta\|\nabla f(x)\|_2$. Step is smaller as we move towards the extremum.
- Line search methods
  - Exact line search: find the optimize step size along a descent direction

    $$\arg\min_{\eta \geq 0} f(x - \eta \nabla f(x))$$

    Ususally we cannot minimize it exactly.
  - Back-tracking line search: find the step size so that we get the expected amount of decrease in $f(x)$

    * Start with $\eta = 1$, repeat $\eta \leftarrow \beta\eta$ until

$$f(x^k - \eta\nabla f(x^k)) \le f(x^k) - \alpha\eta\nabla^T f(x^k)\nabla f(x^k) = f(x^k) - \alpha\eta\|\nabla f(x^k)\|_2^2$$

    * (draw function of the step size)
    * Can prevent step sizes that are too large

2 Case study: Least Square Regression

- Closed form solution:

$$(X^T X)w = Xy$$

  - $X^T X$: $O(nd^2)$
  - Xy: $O(nd)$
  - Solving $d \times d$ linear system: $O(d^3)$

- Gradient descent:

$$f(w) = \frac{1}{2}\|Xw - y\|_2^2 \tag{1}$$

$$\nabla_w f(w) = X^T(Xw - y) \tag{2}$$

$$w^{t+1} = w^t - \eta^t X^T(Xw - y) \tag{3}$$

  - Compute the gradient: $O(nd)$
  - Gradient descent: $O(ndt)$

- GD can be faster if d is very large.