Recitation #13

Irina Espejo (iem244@nyu.edu)

Center for Data Science

DS-GA 1014: Optimization and Computational Linear Algebra for Data Science

recall

Gradient Descent

Let $f : \mathbb{R}^n \to \mathbb{R}$. Gradient descent is an algorithm to find minimize a convex and twice differentiable function. The update algorithm for gradient descent is:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_n)$$

for $\alpha_t \in \mathbb{R}$ is the step size at time t

Convergence

Convergence is ensured if the function is convex and twice differentiable. The speed of convergence follows the inequality:

$$f(x_t) - f(x^\star) \leq \frac{2L\|x_0 - x^\star\|^2}{t+4}$$

Alert! Many times in Data Science we have non-convex functions, generally we still apply methods derived from gradient descent but convergence is not ensured.

recall

Newton's method

This is basically adjusting the step α_t to the "optimal step" under "nice" conditions of f:

$$x_{t+1} = x_t - H f(x_t)^{-1} \nabla f(x_t)$$

If we take the step $\alpha_t = F f(x_t)$ and the function f is nice then convergence is super super fast $\|x_t - x^\star\|^2 \leq Ce^{-a2^t}$

Accelerating the descent

gradient descent with momentum

The upgrade algorithm for an accelerated gradient descent is

$$x_{t+1} = x_t + v_t$$

where the velocity $v_t = -\alpha_t \nabla f(x_t) + \beta_t v_{t-1}$ adds momentum to the descent trajectory to get to the minimum faster (under "nice" conditions of the function f)

practice: gradient descent

Exercise 1 (ex 0.14 2019 review)

Assume we use gradient descent when minimizing the least-square cost $f(x) = \|Ax - y\|^2$. assume that the columns of A are linearly dependent, meaning that $ker(A) \neq \{0\}$.
Write the gradient descent step update for this problem. //

Exercise 1

practice: gradient descent

Exercise 1 (ex 0.14 2019 review)

If now we use a Newton update, which is the gradient descent update?

Exercise 1

practice: gradient descent

Exercise 1 (ex 0.14 2019 review)

What happens to the speed of gradient descent for linear regression when we first perform some dimensionality reduction on the features?

Exercise 1

practice: gradient descent

Exercise 1 (ex 0.14 2019 review)

Think about possible stopping criteria for the gradient descent algorithm.

recall

Stochastic Gradient Descent

we use this variant of gradient descent when the function to optimize is stochastic. Plus, instead of calculating the full gradient we calculate a cheaper but noisy gradient. Turns out under suitable conditions, this algorithm will converge to a local minimum.

review final exam

Exercise