Convex Optimization

Part 2: Gradient descent (1/2)

Namhoon Lee

POSTECH

19 Sep 2022

Admin

Assignment 1 is due by midnight on Friday 30 September.

- Please keep in mind the course policies on late submission and cheating/plagiarism.

Unconstrained optimization

Let us consider the following unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \ .$$

- There is no constraint on x.
- We may assume that f is convex and differentiable.
- The goal is to find a minimum value f^*.

Example 1: linear regression

Consider linear prediction

$$\hat{y} = \beta^\top x$$

- Given data $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, the goal is to find a linear relationship between x and y.

This problem can be casted as a minimization problem:

$$\min_\beta \sum_{i=1}^{n} (\beta^\top x_i - y_i)^2 \quad \text{or} \quad \min_\beta \|X\beta - y\|^2$$

where the goal is to find β^* that minimizes the squared loss (hence least squares).

- We can also put some regularization term (*e.g.*, ridge regression).

Solving the least squares problem is quite simple.

- Take the derivative and set it equal to zero

- The solution

Some questions:

- How can this procedure be justified?
- Does the solution always exist? Is the solution unique?
- How expensive is it to compute the solution?

Gradient descent

For the following unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \ ,$$

consider applying the Gradient Descent (GD) algorithm.

Gradient Descent

Start with some initial point x_1, repeat the following update step iteratively

$$x_{t+1} = x_t - \eta \nabla f(x_t) \ ,$$

and stop at some point. Here η is a step size.

Interpreting gradient descent

GD by function approximation:

$$f(x) \approx f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{1}{2\eta}\|x - x_0\|^2$$

i.e., given x_0 approximate f as a linear function plus a quadratic penalty term.

- Alternatively as a second-order Taylor expansion with the Hessian replaced with identity.

Then choosing the next point as the minimum of the approximation gives

$$x^+ = x - \eta \nabla f(x) \ ,$$

the iterative update rule that is essentially GD.

Gradient descent for least squares

Solving the least squares with GD

How does it compare to the analytic solution?

Example 2: simple quadratic

Consider the following problem:

$$\min_x f(x) = 3x^2 + 4x - 2$$

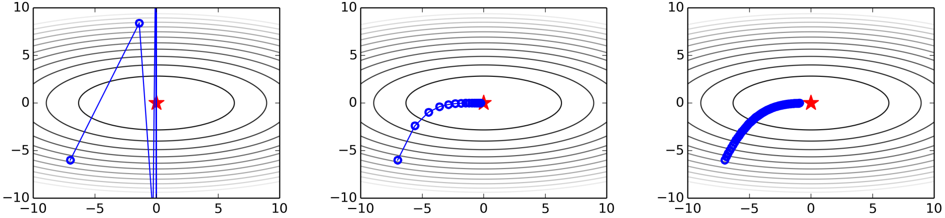
- The solution is achieved at $x^* = -2/3$.

Apply GD to the above?

- The same solution is achieved as $t \to \infty$ with a step size chosen appropriately.

Step size

GD with different step sizes:



- Too large step size can overshoot.
- Too small step size can take too long to converge (if it does).

Line search

The step size can be adjusted adaptively at each iteration.

- The idea is to impose on η so that it leads to *some* reduction in f.

Backtracking line search:

- The reduction in f should be proportional to both the step size and the directional derivative.
- At each iteration t, start with some large step size η and decrease it to be $\alpha\eta$ with $\alpha \in (0, 1)$ until it satisfies the Armijo condition:

$$f(x_t - \eta \nabla f(w_t)) \leq f(x_t) - \gamma\eta \|\nabla f(x_t)\|^2$$

  where $\gamma \in (0, 1)$.
- More conditions can be added.

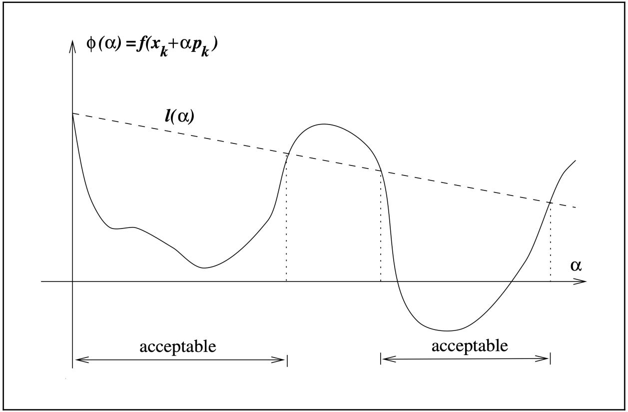


Figure: Sufficient decrease condition (from the NW book).

Smoothness

Let us consider the case where f is differentiable and ∇f is Lipschitz continuous. We often call in this case the function is *smooth*.

Definition (Smoothness)

A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ *is called* β*-smooth when* ∇f *is Lipschitz continuous with Lipschitz constant* $\beta > 0$, *i.e., if there exists some constant* β *such that the following is satisfied:*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2 \qquad \forall \{x, y\} \ .$$

- This ensures that gradients do not change arbitrarily quickly.
- This also means $\nabla^2 f(x) \preceq \beta I$ if f is twice differentiable.

A consequence of β-smoothness:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|y - x\|^2 \qquad \forall \{x, y\} \ .$$

i.e., a quadratic upper bound on f.

Proof.

Recall from the fundamental theorem of calculus that $\int_0^1 f'(t)dt = f(1) - f(0)$. Then we can write

$$
\begin{aligned}
f(y) &= f(x) + \int_0^1 \nabla f((1-t)x + ty)^\top (y-x)dt \\
&= f(x) + \nabla f(x)^\top (y-x) + \int_0^1 \big(\nabla f((1-t)x+ty) - \nabla f(x)\big)^\top (y-x)dt \\
&\leq f(x) + \nabla f(x)^\top (y-x) + \int_0^1 \|\nabla f((1-t)x+ty) - \nabla f(x)\| \|(y-x)\| dt \\
&\leq f(x) + \nabla f(x)^\top (y-x) + \int_0^1 t\beta \|(y-x)\|^2 dt \\
&= f(x) + \nabla f(x)^\top (y-x) + \frac{\beta}{2}\|(y-x)\|^2
\end{aligned}
$$

□

Also, consider running GD with $\eta = 1/\beta$ for smooth f, *i.e.*

$$x_{t+1} = x_t - \frac{1}{\beta} \nabla f(x_t) \ .$$

By substituting variables in the smoothness upper bound we can write

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{\beta}{2} \|x_{t+1} - x_t\|^2 \\ &= f(x_t) + \langle \nabla f(x_t), -\frac{1}{\beta} \nabla f(x_t) \rangle + \frac{\beta}{2} \| - \frac{1}{\beta} \nabla f(x_t) \|^2 \\ &= f(x_t) - \frac{1}{2\beta} \|\nabla f(x_t)\|^2 \end{aligned}$$

- This implies that GD guarantees to decrease f (a.k.a. progress bound).

Any questions?