Convex Optimization

Part 2: Gradient descent (2/2)

Namhoon Lee

POSTECH

21 Sep 2022

Consequence of quadratic upper bound

Bound on suboptimality

If f is β-smooth, then

$$\frac{1}{2\beta}\|\nabla f(x)\|_2^2 \leq f(x) - f(x^*) \leq \frac{\beta}{2}\|x - x^*\|^2 \qquad \forall x$$

Proof.

- (right) it follows from the quadratic upper bound set with $y = x, x = x^*$.
- (left) it follows from minimizing the bound w.r.t. y, plugging it in, and lower bounding with $f(x^*)$.

□

Co-coercivity of gradient

Co-coercivity

If f is convex and β-smooth, then

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\|_2^2 \qquad \forall x, y$$

- Notice, this in turn implies the smoothness (by Cauchy-Schwarz).
- Thus, smoothness $\Rightarrow$ upper bound $\Rightarrow$ co-coercivity $\Rightarrow$ smoothness, meaning that they are equivalent.

Proof.

Define two convex functions f_x, f_y

$$f_x(z) = f(z) - \langle \nabla f(x), z \rangle \qquad \text{and} \qquad f_y(z) = f(z) - \langle \nabla f(y), z \rangle$$

Notice that $z = x$ minimizes $f_x(z)$, and similiarly, $z = y$ minimizes $f_y(z)$. Now write

$$\begin{aligned} f(y) - (f(x) + \langle \nabla f(x), y - x \rangle) &= f(y) - \langle \nabla f(x), y \rangle - (f(x) - \langle \nabla f(x), x \rangle) \\ &= f_x(y) - f_x(x) \\ &\geq \frac{1}{2\beta} \|\nabla f_x(y)\|_2^2 \qquad \text{(from suboptimality bound)} \\ &= \frac{1}{2\beta} \|\nabla f(y) - \nabla f(x)\|_2^2 \end{aligned}$$

Similarly,

$$f(x) - (f(y) + \langle \nabla f(y), x - y \rangle) \geq \frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|_2^2$$

Adding these will give co-coercivity. □

Equivalence to smoothness

For f being β-smooth is equivalent to the following:

$$\frac{\beta}{2}\|x\|_2^2 - f(x) \quad \text{is a convex function.}$$

Proof.

By Cauchy-Schwarz on smoothness, we can write

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \beta \|x - y\|_2^2 \ .$$

This is monotonicity of $\beta x - \nabla f(x)$ (*i.e.*, prove immediately by definition). This further leads to the desired result, *i.e.*, $\frac{\beta}{2}\|x\|_2^2 - f(x)$, because of the equivalence between monotonicity of gradient and convexity. □

- Notice this can be used to show the smoothness characterization for twice differentiable f, *i.e.*, $\nabla^2 f(x) \preceq \beta I$.

Convergence analysis

Does gradient descent ever converge? How fast does it converge when it does?

- We need to analyse its convergence properties or convergence rate.

Convergence of smooth functions

Theorem

For β-smooth functions, gradient descent with the step size $\eta = 1/\beta$ after T iterations satisfies

$$\min_{t=\{1,\ldots,T\}} \|\nabla f(x_t)\|^2 \leq \frac{2\beta R}{T}$$

where $R = f(x_1) - f^*$.

Proof.

The proof is straightforward from the progress bound and noting that $f(x_t) \geq f^*$. □

Notes

- After T iterations we find at least one t with $\|\nabla f(x_t)\|^2 = \mathcal{O}(1/t)$; *i.e.*, the suboptimality gap or error ϵ decreases proportionally to $1/t$ rate.
- The number of iterations required to achive ϵ-accuracy is proportional to $1/\epsilon$.
- This result does not mean that it is the last t that minimizes f or the minimum found is a global minimum.

Convergence of smooth convex functions

Theorem

For β-smooth convex functions, gradient descent with the step size $\eta = 1/\beta$ after T iterations satisfies

$$f(\frac{1}{T}\sum_{t=1}^{T} x_t) - f^* \leq \frac{\beta R^2}{2T}$$

where $R = \|x_1 - x^*\|$.

Proof.

The proof is straightforward from the convexity and progress bound (see next). □

To complete the proof, we can write

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

Rearranging terms gives

$$
f(x_{t+1}) - f(x^*) \leq \frac{\beta}{2}(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2)
$$

By taking the sum over T iterations (and additional steps) we get the desired result.

Any questions?