6.S979 Topics in Deployable ML – Fall 2019 September 5, 2019

Lecture 1: Continuous Optimization Fundamentals

Lecturer: Aleksander Mądry *Scribes: Scott Foster, Luke Kulik, Kevin Li*
(Revised by Andrew Ilyas and Dimitris Tsipras)

1 Introduction

Continuous Optimization is an important toolkit for deep learning. It aims to solve the canonical problem is solving the unconstrained minimization problem:

$$\underset{x}{\text{minimize}}\ f(x), \qquad \text{where } f \text{ is continuous and smooth.} \tag{1}$$

(For us, smooth means that derivatives of all orders exist.) This problem is (of course) intractable in general, so we will make additional assumptions in order to be able to design algorithms. It turns out that all of our algorithms will be iterative. So that we will need to specify a primitive that, given the current solution x^t determines in what direction and how far we should go to obtain an even better solution x^{t+1}.

2 Gradient Descent Method

The most fundamental tool for solving continuous optimization problems is gradient descent method. The basic idea here relies on the Taylor expansion of our function around our current point x (in order to find the best step Δ to take to arrive to the new solution). This expansion states that

$$f(x+\Delta) = \underbrace{f(x) + \nabla f(x)^\intercal \Delta}_{\varphi_x(\Delta)} + \underbrace{\frac{1}{2}\Delta^\intercal \nabla^2 f(x)\Delta + \ldots}_{\varrho_x(\Delta)}, \tag{2}$$

where $\varphi_x(\Delta) = f(x) + \nabla f(x)^\intercal \Delta$ can be viewed as the linear approximation of our function at the point x and $\varrho_x(\Delta)$ is the "tail error" of this approximation. Note that

$$\|\varphi_x(\Delta)\| \sim \|\Delta\| \quad \text{and} \quad \|\varrho_x(\Delta)\| \leq O(\|\Delta\|^2). \tag{3}$$

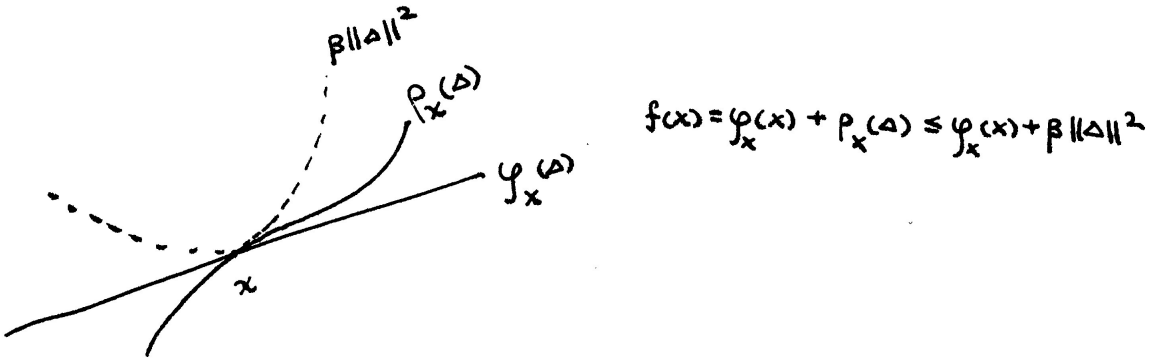
for sufficiently small Δ. So, in principle, by taking Δ small enough we can always ensure that the error of our linear approximation is smaller than the benefit from moving in the direction that minimizes $\varphi_x(\Delta)$.

2.1 Smoothness Assumptions

The second condition in (3) can be seen as a form a of smoothness condition. Specifically, we define the notion of β-smoothness.

Definition 1 *A function f is β-smooth iff*

$$\varrho_x(\Delta) \leq \frac{1}{2}\beta\|\Delta\|^2, \quad \textit{for all } x \textit{ and } \Delta. \tag{4}$$



In particular, this implies that f is dominated by a quadratic function $\varphi_x(\Delta) + \frac{1}{2}\beta\|\Delta\|^2$. That is,

$$f(x+\Delta) \leq \varphi_x(\Delta) + \frac{1}{2}\beta\|\Delta\|^2,$$

for all x and Δ.

Thus we can consider minimizing the proxy $\varphi_x(\Delta) + \frac{1}{2}\beta\|\Delta\|^2$ as a function of Δ in lieu of minimizing $f(x+\Delta)$ directly. Since the minimizer is $\Delta^* = -\frac{1}{\beta}\Delta f(x)$, we arrive at the gradient descent algorithm.

- Pick an initial point, say, $x^0 = 0$.
- For $t = 0, 1, \ldots, T$, set

$$x^{t+1} = x^t - \frac{1}{\beta}\nabla f(x^t). \tag{5}$$

This basic algorithm has the following guarantee on the amount of progress made in each step

$$f(x^t) - f(x^{t+1}) \geq \frac{1}{2\beta}\|\nabla f(x^t)\|. \tag{6}$$

As a result, it eventually is bound to a point $\widehat{x}$ such that

$$\|\nabla f(\hat{x})\| \approx 0. \tag{7}$$

This means that $\hat{x}$ will be a critical point. However, it might be a saddle point and not necessarily a local extremum. So, further assumptions are needed to guarantee convergence to optimality.

2.2 Convexity

If we assume that f is convex, then $\hat{x}$ is indeed guaranteed to be a global minimum. Convexity is equivalent to having the bound

$$0 \leq \varrho_x(\Delta) \leq \frac{1}{2}\beta\|\Delta\|^2, \tag{8}$$

for all x and Δ.

It can then be shown that after $T = O(\beta R^2 \varepsilon^{-1})$ steps, where $R = \|x^0 - x^*\|$, we have that $f(x^T) - f(x^*) \leq \varepsilon$. However, this linear dependence on ε^{-1} and R can be quite inconvenient. So, we need a stronger assumption to get a better bound.

2.3 Strong-α Convexity Assumption

Definition 2 *The function f is said to be strong-α convex iff*

$$\frac{1}{2}\alpha\|\Delta\|^2 \leq \varrho_x(\Delta) \quad \textit{for all } x \textit{ and } \Delta \textit{ and } \alpha > 0. \tag{9}$$

So, if we assume that f is strong-α convex and β-smooth, this means that the error $\varrho_x(\Delta)$ is "sandwiched" by two quadratics, i.e.,

$$\frac{1}{2}\alpha\|\Delta\|^2 \leq \varrho_x(\Delta) \leq \frac{1}{2}\beta\|\Delta\|^2,$$

for all x and Δ. Consequently, the function $f(x+\Delta)$ is "sandwiched" as well by corresponding quadratics too. That is, we have

$$\varphi_x(\Delta) + \frac{1}{2}\alpha\|\Delta\|^2 \leq \varrho_x(\Delta) \leq \frac{1}{2}\beta\|\Delta\|^2,$$

for all x and Δ. (See the Figure 1.)

Once this assumption is in place, we can attain $f(x^T) - f(x^*) \leq \varepsilon$ using only

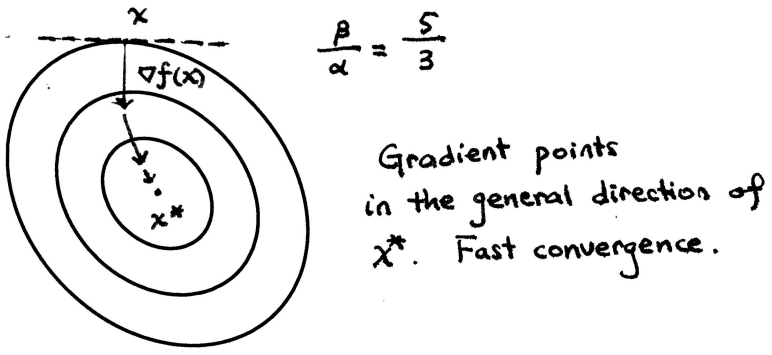
$$T = O\left(\frac{\beta}{\alpha}\log\frac{R}{\varepsilon}\right) \text{ steps. That is, we have a logarithmic dependency in } \varepsilon^{-1}. \tag{10}$$

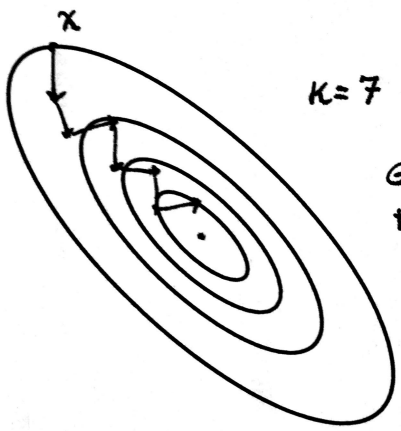
The quantity $\kappa := \beta/\alpha$ is called the condition number of f. It can be viewed as a reflection of the "badness" of the (Euclidean) geometry of f.

In particular, if f is twice differentiable everywhere (which we assumed here), the values of α and β correspond to the bounds on the smallest and largest eigenvalues of the Hessian. That is,

$$\alpha = \inf_x \lambda_{\min}(\nabla^2 f(x)) \quad \text{and} \quad \beta = \sup_x \lambda_{\max}(\nabla^2 f(x)) \tag{11}$$

Now, to get some intuition regarding why this condition number is important one should note that when $\kappa = 1$ (i.e., when f is quadratic), the gradient point directly in the direction of the minimizer. Conversely, when κ is large, the gradient direction does not correlate well with the direction towards minimum. As a result, the optimization path tends to slowly zig-zag toward the minimizer. See the figures below.





Gradient descent zig-zags "around" the minimum; slow convergence.