CSE6243: Advanced Machine Learning **Fall 2023**

Lecture 1: Convex Optimization I

Lecturer: Bo Dai *Scribes: Aditya Sasanur & Aryaman Jha*

Note: *LaTeX template courtesy of UC Berkeley EECS Department.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1.1 Recap

1. Course policies which are available in the files section on Canvas;
2. Course content covering background, generative models, differentiable programming, and reinforcement learning;
3. The broad paradigms of ML tasks.

1.2 New Content

1.2.1 Broad classification of ML problems

Supervised Learning:

TL;DR - $\mathcal{D} = \{x_i, y_i\}|_{i=0}^n, Alg(\mathcal{D}) \Rightarrow f(\cdot) : X \to Y$

We have a labelled dataset $\mathcal{D}$:

$$\mathcal{D} = \{x_i, y_i\}_{i=1}^N \tag{1.1}$$

Supervised learning algorithms provide a map f from an input set X to an output set Y

$$Alg(\mathcal{D}) \implies f : X \to Y \tag{1.2}$$

Regression:

$$\mathcal{D} = \{x_i, y_i\}_{i=0}^n \quad f_w(x) = w^T x \quad y \in \mathbb{R} \tag{1.3}$$

$$\min_w \sum_{i=1}^n (y_i - t_w(x_i))^2 = \sum_{i=1}^n \left(y_i - w^\top x_i\right)^2 \tag{1.4}$$

Classification / Logistic Regression:

$$f_\omega(x) = \frac{1}{1 + \exp(-\omega^T x)}$$
$$\min_\omega \sum_{i=1} y_i \log f_\omega(x_i) + (1 - y_i) \log(1 + f_\omega(x_i))$$

- E.g., 2-layer net: $f_x(x) = v_2^T \sigma\left(v_1^t x + b_2\right) + b_2$

Unsupervised Learning:

TL;DR - $\mathcal{D} = \{x_i\}|_{i=0}^n, Alg(\mathcal{D}) \Rightarrow f(\cdot) : X \to Z$

We have an unlabelled dataset X such that the unsupervised learning algorithms map the input set X to the set Z

$$Alg(X) \implies f : X \to Z \tag{1.5}$$

$$D = \{x_i\}_{i=1}^n = X \tag{1.6}$$

$$\min_{u,v} \|x - uv^T\|^2 \tag{1.7}$$

Reinforced Learning:

TL;DR - *Given environment*, $Alg(Env) \Rightarrow \pi(\cdot) \in \Delta(\mathcal{A})$

We have an agent interacting with an environment. The agent operates with a policy Π over a set of actions $\mathcal{A}$. Reinforcement learning algorithms learn an optimal policy Π over the action set $\mathcal{A}$:

$$Alg(\cdot) \implies \Pi(\cdot) \subseteq \Delta(\mathcal{A}) \tag{1.8}$$

- E.g., multi-arm bandit:

$$\max_{\pi(a)} \mathbb{E}_{\pi(a)}[R(a)] \tag{1.9}$$

1.2.2 Convex Optimization

Convex optimization is a special class of optimization problems. Not all optimization problems of interest to us in ML are convex optimization but studying it is of interest to us for the following reasons:

- It has been solved well theoretically for most cases, and it possesses a global optimum
- It has many rich results which can serve as a starting point for other optimization problems. e.g., ADAM is an optimization technique, and in its documentation, the theoretical results have mainly been discussed with convex problems.

1.2.3 Definitions

An optimization problem is convex if

- the optimization is over a convex set;
- we are optimizing for a convex function.

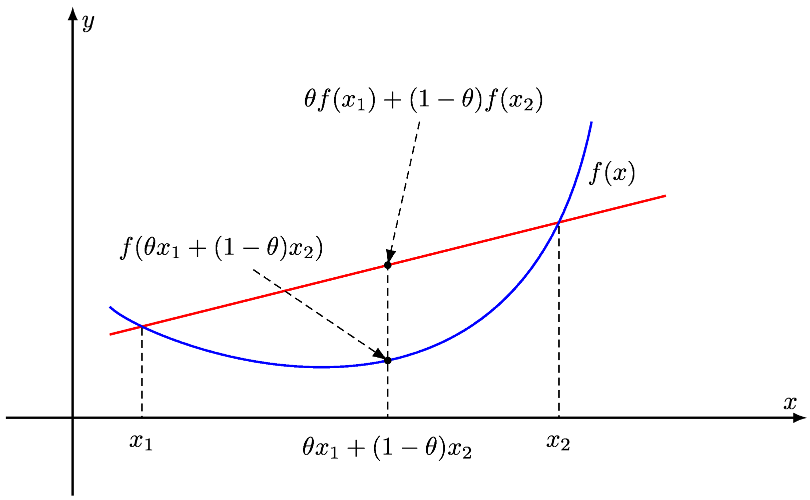

Figure 1.1: Illustration of convex function

These concepts are defined as follows:

Convex Set

A set Ω is a convex set if for all $u, v \in \Omega$ and t $\in$ [0,1] then x $= ut + v(1-t) \ \in \ \Omega$

Convex Function

A function l is a convex function iff the domain of l is a convex set Ω and for $x, y \in \ \Omega$, t $\in$ [0,1], we have : $l(xt + y(1-t)) < l(x)$ and $l(xt + y(1-t)) < l(y)$

1.2.4 Results

Theorem 1 Any local minimum is also a global minimum in convex optimization, i.e. if $w^* \in \Omega$, $\|w^* - u\| < \rho$ then, $l(u) \geq l(w^*) \implies \forall \mu \ \in \ \Omega, \ l(w^*) < l(\mu)$

Proof Assume w_0 where $l\left(w_0\right) \leq l\left(w^*\right)$

$$\begin{aligned} l\left(tw_0 + (1-t)w^*\right) &\leq tl\left(w_0\right) + (1-t)l\left(w^*\right) \\ &\leq t\ell\left(w^*\right) + (1-t)l\left(w^*\right) \\ &= l\left(w^*\right) \end{aligned}$$

This is a contradiction so this w_0 can't exist unless it is w^*

Theorem 2 First-order characteristic condition for global optimum of convex problem

$$w^* = \arg\min_{w \in \Omega} \text{ iff } \nabla l(u)^T (u - w^*) \geq 0, \quad \forall u \in \Omega \tag{1.10}$$

Proof A visual proof may be obtained for lower dimensional spaces by drawing the two vectors $\nabla l(u)$ and $(u - w^*)$ and observing the acuteness of the dot product, implying a positive value.

Least Squares Linear Regression

The solution to the unconstrained optimization of the least squares linear regression is defined as:

$$f(x) = w^T x, \quad \min_w ||Y - w^T X|| = l(w) \tag{1.11}$$

is given by $w^* = (XX^T)^{-1} XY^T$.

Proof

At the global minimum,

$$\nabla l(w^*) = 0 \tag{1.12}$$

$$\implies (Y - w^T X) X^T = 0 \tag{1.13}$$

$$\implies (YX^T)^T = (w^T XX^T)^T \tag{1.14}$$

$$\implies w^* = (XX^T)^{-1} XY^T \tag{1.15}$$