ONLINE OPTIMIZATION

SIDDHANT CHAUDHARY

These are my course notes for the **Online Optimization** course I took in CMI. The reference book used can be found on arXiv: Introduction to Online Convex Optimisation. *Online Learning* can also be described as *learning from experience.*

Contents

Date: 21 September, 2021.

1. The Online Learning Model

1.1. Model Description and Restrictions. In many kinds of ML problems, getting the training set is costly and often not possible. In the usual ML setting, one collects the training data beforehand, and trains a model based on the data. *Online learning*, on the other hand, helps us in the situation when data can't be collected beforehand (called *offline learning*). In such situations, the data is fed continuously to the learner, and the learner *learns through experience*.

Let us now describe the model. In online convex optimisation, there is an *online learner* who interacts with the environment. At each time step t, the learner has to make decisions without knowing the outcome beforehand. After a decision is made, the consequences of the decision are revealed. These consequences are given in terms of *losses* incurred for making the decision. Also, the losses incurred are unknown beforehand. The losses can be adversarially chosen, and they depend on the action taken by the learner.

To make this situation useful, we make the following restrictions.

(1) The losses incurred cannot be unbounded.
(2) The *decision set* (i.e the set of possible choices that the learner can make) must be somehow structured/bounded. It doesn't have to be finite. Usually, we work with convex subsets of the Euclidean space.
(3) Often, the loss functions are bounded convex functions on the decision set.

Let us now set up some notation, and see how the online learning model works.

(1) The *decision set* will be denoted by $\mathcal{K} \subseteq \mathbf{R}^n$, and this set will be a convex region. As mentioned above, the loss functions will be functions $f : \mathcal{K} \to \mathbf{R}^n$.
(2) At each time step, the learner chooses a point $\boldsymbol{x}_t \in \mathcal{K}$.
(3) After the learner has made the choice, a convex loss function $f_t \in \mathcal{F} : \mathcal{K} \to \mathbf{R}$ is revealed. Here $\mathcal{F}$ is the set of bounded convex loss functions available to the adversary.
(4) The loss incurred by the learner at time step t is $f_t(\boldsymbol{x}_t)$. In many situations, only the value $f_t(\boldsymbol{x}_t)$ is revealed, and not the whole function f_t.
(5) T will denote the total number of time steps the learner plays the game. Often, T is unknown beforehand.
(6) The total loss incurred by the learner is

$$\sum_{i=1}^{T} f_t(\boldsymbol{x}_t)$$

1.2. Regret of an algorithm. In this section, we will see how we measure the *goodness* of an OCO algorithm that makes the decisions. This is defined in terms of the *regret* of the algorithm.

Definition 1.1. Let $\mathcal{A}$ be an algorithm for OCO, which maps a certain history to a decision in the decision set. The *regret* of $\mathcal{A}$ is defined to be

$$\text{regret}_T(\mathcal{A}) = \sup_{\{f_1,\ldots,f_T\}\subseteq\mathcal{F}} \left\{ \sum_{t=1}^{T} f_t(\boldsymbol{x}_t) - \min_{\boldsymbol{x}\in\mathcal{K}} \sum_{t=1}^{T} f_t(\boldsymbol{x}) \right\}$$

Let us disambiguate the above definition. Note that $\boldsymbol{x}_t$ is dependent upon $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}\}$, $\{f_1, \ldots, f_{t-1}\}$ and the algorithm $\mathcal{A}$. Ofcourse, it is assumed that $\mathcal{A}$ is a deterministic algorithm, and hence the choice $\boldsymbol{x}_1$ is fixed if the decision set $\mathcal{K}$ is fixed.

Remark 1.0.1. Note that, in the above definition, $\boldsymbol{x}$ is fixed in the second summation, i.e the same choice is made at each step. We can also have the slightly better definition given below.

$$\text{regret}_T(\mathcal{A}) = \sup_{\{f_1,\ldots,f_T\}\subseteq\mathcal{F}} \left\{ \sum_{t=1}^{T} f_t(\boldsymbol{x}_t) - \sum_{t=1}^{T} \min_{\boldsymbol{x}_i\in\mathcal{K}} f_t(\boldsymbol{x}_t) \right\}$$

It turns out that this problem is more difficult to solve. In our setting, the regret is calculated as the difference between the loss incurred by the learner's choices and that of the best *fixed* decision in hindsight.

1.3. **Expert Advice: An example.** Let's look at an example where OCO can be used. The learner has to choose among the advice of n given experts. After making their choice, the learner incurs a loss between 0 and 1. This process is repeated. The goal of the learner is to do as well as the best expert in hindsight.

In this scenario, the decision set is the n-dimensional simplex Δ_n.

$$\mathcal{K} = \Delta_n := \left\{ \boldsymbol{x} \in \mathbf{R}^n \; : \; \sum_i x_i = 1, x_i \geq 0 \right\}$$

Each point in the decision set is interpreted as a probability distribution over n elements, which the ith coordinate represents the probability of choosing expert i. Suppose the loss incurred by the ith expert at time step i is $g_t(i)$. Let $\boldsymbol{g}_t$ be the loss vector of all n experts. The cost function is then the expected loss incurred at time step i, i.e

$$f_t(\boldsymbol{x}) = \boldsymbol{g}_t^T \boldsymbol{x}$$

Clearly, the cost functions are linear (hence convex) and bounded.

1.4. **Learning from Expert Advice.** It turns out that the expert advice example in the previous section is an important class of problems in OCO.

Again, recall the expert advice problem. Suppose the time steps are $t = 1, 2, ...T$. At each time step, the learner faces two choices, namely choice A and choice B (eg. buy or sell a certain stock). The learner can take advice from N experts. As usual, after a choice has been made, the loss function at time t is revealed, which describes how much loss the learner incurred at time t. For simplicity, we will assume that the loss functions are $0-1$ loss functions, i.e the correct choice has 0 loss and the wrong choice has 1 loss.

Suppose the learner, at every time step, chooses choices A, B uniformly at random. We ask the following question, a question about *relative performance*: can the learner make as few mistakes as the best expert in hindsight? The following theorem shows that this is not possible for deterministic learners.

Theorem 1.1. *Suppose the best expert in hindsight makes $L \leq \frac{T}{2}$ mistakes. Then there does not exist a deterministic algorithm that can guarantee less than $2L$ mistakes. So, all deterministic algorithms can be forced to make $\geq 2L$ mistakes.*

Proof. See **Theorem 1.1** in the reference book; the proof provides a counterexample, this proving the theorem. ■

1.5. **The Weighted Majority Algorithm.** We will now see our first algorithm to solve the expert advice problem.

Let us first set the problem up (we did it before, but we'll do it again). We have N experts, from whom we are taking advice. Each expert at each time step gives advice either A or B, where A means buy the stock, and B means sell the stock. At time t, expert i incurs a loss of $M_t(i)$; for simplicity, we assume that the loss is zero for a correct decision, and the loss is 1 for an incorrect decision. With this assumption, $M_t(i)$ is the number of mistakes made by the ith expert till time t. Also, we assume that at each time step, the distribution associated with the N experts is a vertex of the N-simplex Δ_N, i.e exactly one coordinate of the distribution is 1, and all others are 0. In simple words, at each time step, we go with exactly one expert with probability 1.

The algorithm we use to solve this problem is called the *weighted majority algorithm*. The procedure is as follows.

- Each expert i is assigned a weight $W_t(i)$ at every time step t. Initially, at time $t = 1$, $W_1(i) = 1$ for all i. This means at $t = 1$, we keep all the experts at the same level.
- For $t \in [T]$, let $S_t(A), S_t(B) \subseteq [N]$ be the sets of all experts with choice A, B respectively at time step t. Then, we consider the following two sums.

$$W_t(A) := \sum_{i \in S_t(A)} W_t(i) \qquad W_t(B) := \sum_{i \in S_t(B)} W_t(i)$$

  We then compare these two sums; if $W_t(A) \geq W_t(B)$, we make the prediction A at time step t, otherwise we make the prediction B.
- After the prediction is made, we update the weights of the experts as follows.

$$W_{t+1}(i) = \begin{cases} W_t(i) & \text{if expert } i \text{ was correct} \\ W_t(i)(1-\epsilon) & \text{if expert } i \text{ was wrong} \end{cases}$$

  Here $\epsilon \in (0,1)$ is a parameter which we can fix. In simple words, we penalise those experts which gave us a wrong prediction.

A bound on the number of mistakes made by this algorithm is given in **Lemma 1.3** of the reference book. We won't cover that here; instead, we will analyze a more general version of this algorithm in the next section.

1.6. **Randomized Weighted Majority Algorithm.** We now present a randomized version of the weighted majority algorithm. The algorithm is very similar to the deterministic one, with a few differences. In this analysis, we will assume that we the loss functions are f_t^i, i.e the loss of the ith expert at time step t is f_t^i, where $f_t^i \in [-1, 1]$. So, the loss vector function f_t is a function experts $\to [-1, 1]^N$.

- Weights are again assigned to each expert at each time step, and all the initial weights are again 1. In addition to the weights $W_t(i)$, we assign a probability $p_t(i)$ to the ith expert at the ith time step; this is computed as follows.

$$p_t(i) = \frac{W_t(i)}{\sum_{i=1}^N W_t(i)}$$

  The probability $p_t(i)$ is interpreted as the probability of choosing the ith expert at time t.

- There is a minor change in the updating rule of the weights as well; let $\epsilon > 0$ be a parameter. We update the weights as follows.
$$W_{t+1}(i) = W_t(i)(1 - \epsilon f_t^i)$$
Observe that if f_t^i is positive, i.e the ith expert incurs a positive loss at time t, we are penalising him by reducing his weight. Similarly, if f_t^i is negative, we are increasing the weight of the expert.

This sort of update rule is known as the *multiplicative weight update* rule. We now prove an important bound.

Theorem 1.2. *For all $t \in [T]$, assume that $||f_t||_\infty \leq 1$, and let $0 < \epsilon \leq \frac{1}{2}$. Let p_t be the distribution at time step t, and as usual let f_t be the loss vector at time t. Then*
$$\sum_{i=1}^{T} \langle p_t, f_t \rangle - \inf_p \sum_{i=1}^{T} \langle p, f_t \rangle \leq \frac{\ln N}{\epsilon} + \epsilon T$$
The first sum is the total expected loss incurred by the algorithm, and the second term (over which the infimum is taken) is the least average loss in hindsight.

Corollary 1.2.1. *If $\epsilon T = \frac{\ln N}{\epsilon}$ i.e $T = \frac{\ln N}{\epsilon^2}$, the error is atmost $2\epsilon T$, and hence the average error is atmost*
$$\frac{2\epsilon T}{T} = 2\epsilon$$

Proof. Define Φ_t to be the sum of the weights at time t, i.e
$$\Phi_t = \sum_{i=1}^{N} W_t(i)$$
So, we have that $\Phi_1 = N$. Observe the following.
$$\begin{aligned}
\Phi_{t+1} &= \sum_{i=1}^{N} W_{t+1}(i) \\
&= \sum_{i=1}^{N} W_t(i)(1 - \epsilon f_t^i) \\
&= \sum_{i=1}^{N} W_t(i) - \epsilon \sum_{i=1}^{N} W_t(i) f_t^i \\
&= \sum_{i=1}^{N} W_t(i) - \epsilon \Phi_t \sum_{i=1}^{N} \frac{W_t(i) f_t^i}{\Phi_t} \\
&= \Phi_t - \epsilon \Phi_t \sum_{i=1}^{N} p_t(i) f_t^i \\
&= \Phi_t - \epsilon \Phi_t \langle p_t, f_t \rangle \\
&= \Phi_t (1 - \epsilon \langle p_t, f_t \rangle) \\
&\leq \Phi_t e^{-\epsilon \langle p_t, f_t \rangle}
\end{aligned}$$

where in the last step we have used the inequality $1 - x \leq e^{-x}$. Writing this out for each t inductively, we get

$$(*) \qquad \Phi_{T+1} \leq \Phi_1 e^{-\epsilon \sum_{t=1}^{T} \langle p_t, f_t \rangle} = N e^{-\epsilon \sum_{t=1}^{T} \langle p_t, f_t \rangle}$$

■

This gives us an upper bound on Φ_{T+1}. Now we will try to find a lower bound. Note that trivially, we have $\Phi_{T+1} \geq W_{T+1}(i)$ for each $i \in [N]$. This gives us the following for each $i \in [N]$.

$$\begin{aligned}
\Phi_{T+1} &\geq W_{T+1}(i) \\
&= W_T(i)(1 - \epsilon f_T^i) \\
&= W_{T-1}(i)(1 - \epsilon f_{T-1}^i)(1 - \epsilon f_T^i) \\
&\quad \vdots \\
&= W_1(i) \prod_{t=1}^{T} (1 - \epsilon f_t^i) \\
&= \prod_{t=1}^{T} (1 - \epsilon f_t^i) \geq e^{\sum_{t=1}^{T} -\epsilon f_t^i - \epsilon^2 (f_t^i)^2}
\end{aligned}$$

where in the last step we have used the inequality $e^{-x-x^2} \leq 1 - x$. So, we get

$$(\dagger) \qquad \Phi_{T+1} \geq e^{-\epsilon \sum_{t=1}^{T} f_t^i - \epsilon^2 \sum_{t=1}^{T} (f_t^i)^2}$$

Inequalities $(*)$ and $(\dagger)$ give us the following inequality.

$$e^{-\epsilon \sum_{t=1}^{T} f_t^i - \epsilon^2 \sum_{t=1}^{T} (f_t^i)^2} \leq N e^{-\epsilon \sum_{t=1}^{T} \langle p_t, f_t \rangle}$$

Taking the logarithm of both sides, we get the following.

$$-\epsilon \sum_{t=1}^{T} f_t^i - \epsilon^2 \sum_{t=1}^{T} (f_t^i)^2 \leq \ln N - \epsilon \sum_{t=1}^{T} \langle p_t, f_t \rangle$$

Rearranging the above, we get the following.

$$\epsilon \sum_{t=1}^{T} \langle p_t, f_t \rangle - \epsilon \sum_{t=1}^{T} f_t^i \leq \ln N + \epsilon^2 T$$

where above we have crucially used the fact that $||f_t||_\infty \leq 1$. Note that the above inequality is true for each $i \in [N]$. Now, suppose $p = (p_1, ..., p_N)$ is any element of the N-simplex Δ_N. Above, we multiply the ith inequality by p_i (possible because each $p_i \geq 0$), and add all the obtained inequalities. Since $p_1 + \cdots + p_N = 1$, we get the following.

$$\epsilon \sum_{t=1}^{T} \langle p_t, f_t \rangle - \epsilon \sum_{t=1}^{T} \langle p, f_t \rangle \leq \ln N + \epsilon^2 T$$

Since $p \in \Delta_N$ was arbitrary, the claim follows.

2. Introduction to Convexity

2.1. **Basic Definitions.** Through, we will use the notation $\mathcal{K}$ for the convex set in question.

Definition 2.1. A subset $\mathcal{K} \subseteq \mathbf{R}^d$ is said to be *convex* if for all points $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{K}$ and $\alpha \in [0,1]$, it is true that

$$\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y} \in \mathcal{K}$$

This means that the line segment between the points lies in the set.

Example 2.1 (LP). Convex sets are really useful in *linear programming.* Suppose we are given a matrix inequality of the form

$$A\boldsymbol{x} \leq \boldsymbol{b}$$

where the variables are the coordinates of $\boldsymbol{x}$, A is some $m \times n$ matrix, and the vector $\boldsymbol{b}$ represents constraints. This inequality is equivalent to m linear inequalities in the n variables $\boldsymbol{x} = (x_1, .., x_n)$. It can be very easily shown that the solution space of this system of inequalities is a convex set.

Example 2.2 (Max Flow). Consider the maximum flow problem. With some work, this problem can be stated as an optimisation problem, where we are trying to maximise some linear function subject to some linear constraints.

Definition 2.2. Let $\mathcal{K}$ be a convex set. A function $f : \mathcal{K} \to \mathbf{R}$ is said to be *convex* if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{K}$ and all $\alpha \in [0,1]$, it is true that

$$f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y})$$

Theorem 2.1. *Let $\mathcal{K}$ be an open convex set, and let $f : \mathcal{K} \to \mathbf{R}$ be a differentiable function on $\mathcal{K}$. Then, f is convex iff. and only if*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

Proof. First, suppose f is convex, and let $\lambda \in (0,1)$. Then, we have the following.

$$f((1-\lambda)\boldsymbol{x} + \lambda\boldsymbol{y}) \leq (1-\lambda) f(\boldsymbol{x}) + \lambda f(\boldsymbol{y})$$

This implies the following.

$$\lambda f(\boldsymbol{y}) \geq \lambda f(\boldsymbol{x}) + f(\lambda\boldsymbol{y} + (1-\lambda)\boldsymbol{x}) - f(\boldsymbol{x})$$

Dividing throughout by λ, we see that

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \frac{f(\boldsymbol{x} + \lambda(\boldsymbol{y} - \boldsymbol{x})) - f(\boldsymbol{x})}{\lambda}$$

Letting $\lambda \to 0$, the second term becomes the directional derivative in the direction $\boldsymbol{y} - \boldsymbol{x}$. Hence,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

The converse is easy to prove. Suppose

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

Let $\boldsymbol{z} = \lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}$. Then, $\boldsymbol{z} \in \mathcal{K}$. So, the following inequalities hold.

$$f(\boldsymbol{x}) \geq f(\boldsymbol{z}) + \langle \nabla f(\boldsymbol{z}), \boldsymbol{x} - \boldsymbol{z} \rangle$$
$$f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + \langle \nabla f(\boldsymbol{z}), \boldsymbol{y} - \boldsymbol{z} \rangle$$

Multiply the first inequality by λ and the second by $1 - \lambda$ and add. Doing this, we get

$$\lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) \geq f(\boldsymbol{z})$$

which shows that f is indeed convex. ■

2.2. **Subgradients.** Often in optimization, we encounter convex functions which are differentiable at many points except a few. In those cases, we ofcourse can't use the notion of *gradients*. For example, consider the graph of $|x|$. It is not differentiable at 0. Still, the techniques of convex optimisation can be applied to these functions by making a few modifications.

Definition 2.3. Let $\mathcal{K} \subseteq \mathbf{R}^n$ be a convex set, and let $f : \mathcal{K} \to \mathbf{R}$ be a convex function. Let $\boldsymbol{x} \in \mathcal{K}$. We say that $\boldsymbol{v} \in \mathbf{R}^n$ is a *subgradient* of f at $\boldsymbol{x}$ if for all $\boldsymbol{y} \in \mathcal{K}$, it is true that

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{x} \rangle$$

Note that this is the inequalitty of **Theorem** 2.1. The set of subgradients is denoted by $\partial f(\boldsymbol{x})$.

Definition 2.4. The quantity

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

is called the *Bregman divergence.*

2.3. **Alternative Characterisations of Convexity.** In this section, we will prove some alternate characterisations of convexity, which often make it much simpler to prove that a function is convex.

Lemma 2.2. *Let $f : \mathcal{K} \to \boldsymbol{R}$ be a twice differentiable function, where $\mathcal{K}$ is a convex open set. Then, f is convex iff. for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{K}$, we have*

$$\langle \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \geq 0$$

Proof. First, suppose f is convex. Then, by **Theorem** 2.1, we know that the following hold for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{K}$.

$$f(\boldsymbol{x}) \geq f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle$$
$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

Adding the above two inequalities, we get

$$0 \geq \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

The above inequality is the same as what we wanted to show, by rearranging.

Conversely, suppose the given inequality holds. We want to show that f is convex. Consider the following one variable function.

$$g(t) = f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))$$

Here our interval of interest is $[0, 1]$. Clearly, g is differentiable on $[0, 1]$, and the derivative is given by the following.

$$g'(t) = \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

Infact, g is second order differentiable (provided f is), and we have the following.

$$g''(t) = \left\langle \nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle \tag{†}$$

Here, $\nabla^2 f$ is the *Hessian.* Let us now prove this. First,

$$g'(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \cdot (y_i - x_i)$$

Now, since f is twice differentiable, each of the partial derivatives $\dfrac{\partial f}{\partial x_i}$ are differentiable too. So, g' is differentiable, and

$$\begin{aligned} g''(t) &= \sum_{i=1}^{n} \left\langle \nabla \frac{\partial f}{\partial x_i}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x} \right\rangle \cdot (y_i - x_i) \\ &= \sum_{i=1}^{n} \left[\sum_{j=1}^{n} \frac{\partial f}{\partial x_j \partial x_i}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \cdot (y_j - x_j) \right] \cdot (y_i - x_i) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} (y_i - x_i) \cdot \frac{\partial f}{\partial x_i \partial x_j}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \cdot (y_j - x_j) \end{aligned}$$

In the last step, we have used the equality of mixed partials. It is now straightforward to check that the last sum is equal to the sum given in equation (†).

Now, using the fundamental theorem of calculus, we get the following.

$$\langle \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle = g'(1) - g'(0) = \int_0^1 g''(t) \ \mathrm{d}t$$

Also, by the same theorem, we have the following.

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \int_0^1 g'(t) \ \mathrm{d}t$$

Let $\boldsymbol{x}_t = \boldsymbol{x} + t(\mathbf{y} - \boldsymbol{x})$. The above equation implies the following.

$$\begin{aligned} f(\boldsymbol{y}) = f(\boldsymbol{x}) &+ \int_0^1 \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x} \rangle \ \mathrm{d}t \\ &= f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \int_0^1 \langle \nabla f(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \ \mathrm{d}t \\ &= f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \int_0^1 \frac{1}{t} \langle \nabla f(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}), \boldsymbol{x}_t - \boldsymbol{x} \rangle \ \mathrm{d}t \end{aligned}$$

By our assumption, the last integral is a positive quantity. So,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

and hence by **Theorem** 2.1, f is convex. ■

Theorem 2.3. *Let $\mathcal{K}$ be convex, and let $f : \mathcal{K} \to \mathbf{R}$ be a twice differentiable function. Then f is convex iff. the Hessian $\nabla^2 f(\boldsymbol{x})$ is positive semi-definite for each $\boldsymbol{x} \in \mathcal{K}$.*

Proof. First, suppose f is convex, and let $\boldsymbol{x} \in \mathcal{K}$. We pick a small neighborhood of $\boldsymbol{x}$ inside $\mathcal{K}$ (ofcourse, we assume that $\mathcal{K}$ is open, since we are talking about differentiability). Let $\boldsymbol{s} \in \mathbf{R}^n$ be some direction given to us. We will assume that, for small enough t, $\boldsymbol{x}_t = \boldsymbol{x} + t\boldsymbol{s}$ is inside $\mathcal{K}$. Since f is convex, we can apply **Lemma** 2.2 to get

$$\langle \nabla f(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}), \boldsymbol{x}_t - \boldsymbol{x} \rangle \geq 0$$

Also, assume that $t \geq 0$. So, we get

$$\begin{aligned} 0 &\leq \frac{1}{t^2} \langle \nabla f(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}), \boldsymbol{x}_t - \boldsymbol{x} \rangle \\ &= \frac{1}{t} \langle \nabla f(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}), \boldsymbol{s} \rangle \\ &= \frac{1}{t} \int_0^t \left\langle \nabla^2 f(\boldsymbol{x}_t)\boldsymbol{s}, \boldsymbol{s} \right\rangle \, \mathrm{d}t \end{aligned}$$

Now, take the limit as $t \to 0$, and use the fundamental theorem of calculus. We get

$$0 \leq \left\langle \nabla^2 f(\boldsymbol{x})\boldsymbol{s}, \boldsymbol{s} \right\rangle$$

and hence $\nabla^2 f(\boldsymbol{x})$ is positive semi-definite.

Conversely, suppose $\nabla^2 f(\boldsymbol{x})$ is positive semi-definite. Then,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

This is true by Taylor's Theorem, where we do a second order approximation. Hence, f is convex. ■

Definition 2.5. A function f is said to be *α-strongly convex* if

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\alpha}{2} ||\boldsymbol{y} - \boldsymbol{x}||$$

Moreover, f is said to be *β-smooth* if

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

Proposition 2.4. *Let $\mathcal{K}$ be a convex set.*

1. *Suppose f_1 is α_1-strongly convex on $\mathcal{K}$, and suppose f_2 is α_2-strongly convex on $\mathcal{K}$. Then $f_1 + f_2$ is $\alpha_1 + \alpha_2$-strongly convex on $\mathcal{K}$.*
2. *Suppose f_1 is β_1-smooth on $\mathcal{K}$, and suppose f_2 is β_2-smooth on $\mathcal{K}$. Then $f_1 + f_2$ is $\beta_1 + \beta_2$-smooth on $\mathcal{K}$.*

Proof. The proofs are straightforward. Let us consider (1) first. Suppose $\boldsymbol{x}, \boldsymbol{y}$ are arbitrary points in $\mathcal{K}$. Then, we have the following two inequalities by strong convexity.

$$\begin{aligned} f_1(\boldsymbol{x}) &\geq f_1(\boldsymbol{y}) + \langle \nabla f_1(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{\alpha_1}{2} ||\boldsymbol{x} - \boldsymbol{y}||^2 \\ f_2(\boldsymbol{x}) &\geq f_2(\boldsymbol{y}) + \langle \nabla f_2(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{\alpha_2}{2} ||\boldsymbol{x} - \boldsymbol{y}||^2 \end{aligned}$$

Adding the above two inequalities, the claim follows.

(2) is similarly proven. This completes the proof. ■

2.4. Convex bodies and Hyperplanes.

In this section, we will prove a fundamental and important fact about convex sets, namely that convex sets can be separated from points outside them by hyperplanes.

Definition 2.6. Let H be some hyperplane in $\mathbf{R}^n$, and let $\mathcal{K}$ be some set. Then H is said to be *supporting* at $\boldsymbol{x} \in \mathcal{K}$ if $H \cap \mathcal{K} = \{\boldsymbol{x}\}$.

Theorem 2.5. *Let $\mathcal{K}$ be a convex set in $\mathbf{R}^n$, and in addition suppose that $\mathcal{K}$ is closed. Let $y \in \mathbf{R}^n - \mathcal{K}$. Then, there is a hyperplane that separates y from $\mathcal{K}$; formally, there is some $\boldsymbol{h} \in \mathbf{R}^n$ and $c \in \mathbf{R}$ such that*

$$\boldsymbol{h}^t \boldsymbol{y} \geq c$$

and

$$\boldsymbol{h}^t \boldsymbol{x} < c$$

for all $x \in \mathcal{K}$.

Proof. The proof is motivated by geometric intuition. First, we assume that $\mathcal{K}$ is compact; the non-compact case will be dealt with later. For $\boldsymbol{x} \in \mathcal{K}$, consider the quantity

$$||\boldsymbol{x} - \boldsymbol{y}||$$

and let

$$\boldsymbol{x}^* = \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{K}} ||\boldsymbol{x} - \boldsymbol{y}||$$

We claim that such an $\boldsymbol{x}^*$ exists; to see this, note that the quantity $||\boldsymbol{x} - \boldsymbol{y}||$ is always *positive*, and hence

$$\inf_{\boldsymbol{x} \in \mathcal{K}} ||\boldsymbol{x} - \boldsymbol{y}|| \geq 0$$

Moreover, since $||\cdot||$ is a continuous function and since $\mathcal{K}$ is a *compact* set, this infimum is attained at some point. Also, since $\boldsymbol{y} \notin \mathcal{K}$, this infimum is *positive.*

Next, we claim that for all $\boldsymbol{x} \in \mathcal{K}$, it is true that

$$\langle \boldsymbol{x}^* - \boldsymbol{x}, \boldsymbol{x}^* - \boldsymbol{y} \rangle \leq 0$$

The proof is quite simple. Suppose $\boldsymbol{x} \in \mathcal{K}$, and put

$$\boldsymbol{x}_t = t\boldsymbol{x} + (1-t)\boldsymbol{x}^*$$

Clearly, since $\mathcal{K}$ is convex, $\boldsymbol{x}_t \in \mathcal{K}$ for all $t \in [0,1]$. By the definition of $\boldsymbol{x}^*$, we know that

$$||\boldsymbol{x}_t - \boldsymbol{y}||^2 \geq ||\boldsymbol{x}^* - \boldsymbol{y}||^2$$

But, note that

$$\begin{aligned} ||\boldsymbol{x}_t - \boldsymbol{y}||^2 &= ||t\boldsymbol{x} + (1-t)\boldsymbol{x}^* - \boldsymbol{y}||^2 \\ &= ||\boldsymbol{x}^* - \boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{x}^*)||^2 \\ &= ||\boldsymbol{x}^* - \boldsymbol{y}||^2 + 2t \langle \boldsymbol{x^*} - y, \boldsymbol{x} - \boldsymbol{x}^* \rangle + t^2 ||\boldsymbol{x} - \boldsymbol{x}^*||^2 \end{aligned}$$

So, it follows that

$$||\boldsymbol{x}^* - \boldsymbol{y}||^2 + 2t \langle \boldsymbol{x^*} - y, \boldsymbol{x} - \boldsymbol{x}^* \rangle + t^2 ||\boldsymbol{x} - \boldsymbol{x}^*||^2 \geq ||\boldsymbol{x}^* - \boldsymbol{y}||^2$$

and this means that

$$2t \langle \boldsymbol{x^*} - y, \boldsymbol{x} - \boldsymbol{x}^* \rangle + t^2 ||\boldsymbol{x} - \boldsymbol{x}^*||^2 \geq 0$$

Dividing throughout by t, we get

$$2 \langle \boldsymbol{x^*} - y, \boldsymbol{x} - \boldsymbol{x}^* \rangle + t ||\boldsymbol{x} - \boldsymbol{x}^*||^2 \geq 0$$

Since $t \in [0,1]$ is arbitrary, this is only possible if

$$\langle \boldsymbol{x^*} - y, \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0$$

which means

$$\langle \boldsymbol{x}^* - \boldsymbol{x}, \boldsymbol{x^*} - y \rangle \leq 0$$

and this proves our claim.

Now, let $\boldsymbol{h} = \boldsymbol{y} - \boldsymbol{x}^*$. Clearly, $\boldsymbol{h} \neq 0$. Moreover, the above inequality implies

$$\langle \boldsymbol{y} - \boldsymbol{x}^*, \boldsymbol{x} - \boldsymbol{x}^* \rangle \leq 0$$

which means

$$\langle \boldsymbol{y} - \boldsymbol{x}^*, \boldsymbol{x} \rangle \leq \langle \boldsymbol{y} - \boldsymbol{x}^*, \boldsymbol{x}^* \rangle < \langle \boldsymbol{y} - \boldsymbol{x}^*, \boldsymbol{y} \rangle$$

where the last inequality is strict because $||h|| > 0$. This proves the theorem in the case when $\mathcal{K}$ is compact.

Now, suppose $\mathcal{K}$ is closed, but not compact (i.e it is unbounded). Let $\boldsymbol{x}' \in \mathcal{K}$ be some fixed point. Consider the set

$$S := \{\boldsymbol{x} \in \mathcal{K} \mid \ ||\boldsymbol{x} - \boldsymbol{y}|| \leq ||\boldsymbol{x}' - \boldsymbol{y}||\}$$

Clearly, S is a bounded subset of $\mathcal{K}$, and it is non-empty. We will show that S is also closed (and hence compact). Then, by applying the same process as above, the theorem will be proved in the general case as well.

To prove that S is closed, let $\boldsymbol{x}_0$ be a limit point of $\boldsymbol{S}$. So, there is some sequence $\boldsymbol{x}_n$ of points of S converging to $\boldsymbol{x}_0$. We need to show that

$$||\boldsymbol{x}_0 - y|| \leq ||\boldsymbol{x}' - \boldsymbol{y}||$$

But this is easy to see as follows: because $\boldsymbol{x}_n \to \boldsymbol{x}_0$ and the inequality holds for all $\boldsymbol{x}_n$, the inequality must hold for $\boldsymbol{x}_0$ too since $||\cdot||$ is a continuous function. Hence, $\boldsymbol{x}_0 \in S$, and the proof is complete. ■

Remark 2.5.1. The point $\boldsymbol{x}^*$ constructed above for a fixed $\boldsymbol{y}$ is called the projection of $\boldsymbol{y}$ onto $\mathcal{K}$. We will revisit these in a later section.

2.5. **Local Minima are Global Minima.** In this section, we will prove an important fact about convex functions, namely: if a convex function has a local minima at some point, then it is infact a global minima.

Theorem 2.6. *Let $f : \mathcal{K} \to \mathbf{R}$ be a convex function, where $\mathcal{K}$ is a convex set. Suppose f attains a local minima at some point $\boldsymbol{x} \in \mathcal{K}$. Then, f attains a global minima at the point $\boldsymbol{x}$.*

Proof. The proof is straightforward. Suppose f attains a local minima at some $\boldsymbol{x} \in \mathcal{K}$. This means that there is some $\delta > 0$ such that

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x})$$

for all $\boldsymbol{y} \in B_{\boldsymbol{x},\delta}$. For the sake of contradiction, suppose there is some point $\boldsymbol{x}' \in \mathcal{K}$ such that

$$f(\boldsymbol{x}') < f(\boldsymbol{x})$$

For $t \in [0, 1]$, put

$$\boldsymbol{x}_t = t\boldsymbol{x} + (1 - t)\boldsymbol{x}'$$

Since f is convex, we know that for any $t \in [0, 1]$,

$$f(\boldsymbol{x}_t) \leq tf(\boldsymbol{x}) + (1 - t)f(\boldsymbol{x}') < tf(\boldsymbol{x}) + (1 - t)f(\boldsymbol{x}) = f(\boldsymbol{x})$$

which is clearly a contradiction for small enough t. So, $\boldsymbol{x}$ is a point of global minima. ■

Another fact of the same type is the following.

Lemma 2.7. *Let $f : \mathcal{K} \to \mathbf{R}$ be a convex function, where $\mathcal{K}$ is a closed convex set. Suppose $\boldsymbol{x}^*$ is the minimizer of f on $\mathcal{K}$. Then for all $\boldsymbol{y} \in \mathcal{K}$, we have*

$$\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^* \rangle \geq 0$$

Proof. For the sake of contradiction, suppose the claim is false, i.e there is some $\boldsymbol{y} \in \mathcal{K}$ such that

$$\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^* \rangle < 0$$

Next, define the function $g : [0,1] \to \mathbf{R}$ by the following.

$$g(t) = f((1-t)\boldsymbol{x}^* + t\boldsymbol{y})$$

Then, observe that

$$g(0) = f(\boldsymbol{x}^*)$$

Also, we have the following by the chain rule, for all $t \in [0,1]$.

$$g'(t) = \langle \nabla f((1-t)\boldsymbol{x}^* + t\boldsymbol{y}), \boldsymbol{y} - \boldsymbol{x}^* \rangle$$

Above, $g'(0)$ should be interpreted as a one-sided limit, where $t \to 0^+$. We immediately see that

$$g'(0) = \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^* \rangle < 0$$

Now, by the definition of the derivative,

$$g'(0) = \lim_{t \to 0^+} \frac{g(t) - g(0)}{t} < 0$$

This means that, for sufficiently small $t \in (0,1)$,

$$\frac{g(t) - g(0)}{t} < 0$$

which implies that

$$g(t) - g(0) < 0$$

and hence

$$g(t) < g(0)$$

But, this means that

$$f((1-t)\boldsymbol{x}^* + t\boldsymbol{y}) < f(\boldsymbol{x}^*)$$

which contradicts the fact that $\boldsymbol{x}^*$ is the minimizer of f. ■

2.6. **Convex Projections.** Let $\mathcal{K} \subseteq \mathbf{R}^d$ be a convex body, and let $\boldsymbol{y} \in \mathbf{R}^d$. The *projection* of $\boldsymbol{y}$ onto $\mathcal{K}$ is defined as follows.

$$\Pi_{\mathcal{K}}(\boldsymbol{y}) := \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{K}} ||\boldsymbol{x} - \boldsymbol{y}||$$

In **Theorem** 2.5, it was shown that if $\mathcal{K}$ is a closed convex set, then projections exist.

Projections have a crucial property that we'll use a lot, known as the *Pythagorean Theorem.*

Theorem 2.8 (**Pythagoras**). *Let $\mathcal{K} \subseteq \mathbf{R}^d$ be a closed convex set, $\boldsymbol{y} \in \mathbf{R}^d$ and let $\boldsymbol{x} = \Pi_{\mathcal{K}}(\boldsymbol{y})$. Then for all $\boldsymbol{z} \in \mathcal{K}$ it is true that*

$$||\boldsymbol{y} - \boldsymbol{z}|| \geq ||\boldsymbol{x} - \boldsymbol{z}||$$

Proof. Consider the function

$$\boldsymbol{q} \mapsto ||\boldsymbol{q} - \boldsymbol{y}||^2$$

on the convex set $\mathcal{K}$. Clearly, this is a differentiable function, and by hypothesis, it's minimizer over $\mathcal{K}$ is $\boldsymbol{x}$. We then appeal to **Lemma** 2.7 (ofcourse using the fact that norm squared is a convex function); using that lemma, we see that for any $\boldsymbol{z} \in \mathcal{K}$, it is true that

$$\langle 2(\boldsymbol{x} - \boldsymbol{y}), \boldsymbol{z} - \boldsymbol{x} \rangle = 2 \langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{z} - \boldsymbol{x} \rangle \geq 0$$

which implies that

$$2 \langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{x} - \boldsymbol{z} \rangle \geq 0$$

Now, for any $\boldsymbol{z} \in \mathcal{K}$, we have the following.

$$\begin{aligned}||\boldsymbol{y}-\boldsymbol{z}||^2 &= ||\boldsymbol{y}-\boldsymbol{x}+\boldsymbol{x}-\boldsymbol{z}||^2 \\ &= ||\boldsymbol{y}-\boldsymbol{x}||^2 + ||\boldsymbol{x}-\boldsymbol{z}||^2 + 2\langle \boldsymbol{y}-\boldsymbol{x}, \boldsymbol{x}-\boldsymbol{z}\rangle \\ &\geq ||\boldsymbol{x}-\boldsymbol{z}||^2\end{aligned}$$

and this proves the claim by taking square roots. ■

3. First Order Convex Optimization

3.1. **Gradient Descent.** The first convex optimisation algorithm that we will look at is called *gradient descent*. Before describing the algorithm, we will prove some properties of α-strongly convex and β-smooth functions.

Definition 3.1. Let $f : \mathcal{K} \to \mathbb{R}$ be an α-strongly convex and β-smooth function. Then f is said to be γ-*well conditioned*, where

$$\gamma = \frac{\alpha}{\beta}$$

Proposition 3.1. *Let f be a γ-well conditioned function on a convex domain $\mathcal{K}$. Let T be any positive integer, called the time horizon. Let $\boldsymbol{x}_0$ be a point in $\mathcal{K}$, and for each $0 \leq t \leq T-2$, let*

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \nabla_t$$

where

$$\nabla_t = \nabla f(\boldsymbol{x}_t)$$

and each η_t is a scalar. Also, let

$$\begin{aligned}h_t &= f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \\ d_t &= ||\boldsymbol{x}_t - \boldsymbol{x}^*||\end{aligned}$$

where $\boldsymbol{x}^*$ is the minimizer of f over $\mathcal{K}$. Then the following inequalities hold, provided $\boldsymbol{x}_t \in \mathcal{K}$ for all t.

(1) $h_t \geq \frac{\alpha}{2} d_t^2$

(2) $h_t \leq \frac{\beta}{2} d_t^2$, *provided that $\mathcal{K}$ contains some open ball around $\boldsymbol{x}^*$ (which happens if $\mathcal{K} = \mathbf{R}^d$, the case of unconstrained optimization).*

(3) $h_t \geq \frac{1}{2\beta} ||\nabla_t||^2$ *if $\eta_t = \frac{1}{\beta}$ for all t.*

(4) $h_t \leq \frac{||\nabla_t||^2}{2\alpha}$ *provided $\mathcal{K} = \mathbf{R}^d$, the case of unconstrained optimization.*

Proof. First, let us prove (1). By the convexity of f and the fact that f is α-strongly convex, we know the following.

$$\begin{aligned}f(\boldsymbol{x}_t) &\geq f(\boldsymbol{x}^*) + \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x}_t - \boldsymbol{x}^*\rangle + \frac{\alpha}{2}||\boldsymbol{x}_t - \boldsymbol{x}^*||^2 \\ &\geq f(\boldsymbol{x}^*) + \frac{\alpha}{2}||\boldsymbol{x}_t - \boldsymbol{x}^*||^2 \\ &= f(\boldsymbol{x}^*) + \frac{\alpha}{2} d_t^2\end{aligned}$$

where in the second step above we used **Lemma** 2.7 to prove the non-negativity of the second term. This clearly implies that

$$h_t \geq \frac{\alpha}{2} d_t^2$$

and proves part (1). Note that this inequality still holds even if the function f is not β-smooth.

Now consider inequality (2). By the fact that f is β-smooth, we have the following.

$$f(\boldsymbol{x}_t) \leq f(\boldsymbol{x}^*) + \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x}_t - \boldsymbol{x}^* \rangle + \frac{\beta}{2} \left|\left|\boldsymbol{x}_t - \boldsymbol{x}^*\right|\right|^2$$

By our assumption that some ball around $\boldsymbol{x}^*$ is contained in $\mathcal{K}$ (which is the case for *unconstrained optimization*), we know that $\nabla f(\boldsymbol{x}^*) = 0$ (because $\boldsymbol{x}^*$ is the minimizer). This means

$$h_t \leq \frac{\beta}{2} d_t^2$$

which is exactly what we wanted to show.

Let us now prove (3). So suppose $\eta_t = \eta$ for some η (we will eventually let $\eta = \frac{1}{\beta}$). By β-smoothness, we have the following.

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \rangle + \frac{\beta}{2} \left|\left|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\right|\right|^2$$

which can implies the following.

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1}) \geq \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\beta}{2} \left|\left|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\right|\right|^2$$

So, we have the following.

$$\begin{aligned} h_t &= f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \\ &\geq f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1}) \\ &\geq \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\beta}{2} \left|\left|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\right|\right|^2 \\ &= \langle \nabla f(\boldsymbol{x}_t), \eta \nabla_t \rangle - \frac{\beta}{2} \eta^2 \left|\left|\nabla_t\right|\right|^2 \\ &= \eta \left|\left|\nabla_t\right|\right|^2 - \frac{\beta}{2} \eta^2 \left|\left|\nabla_t\right|\right|^2 \end{aligned}$$

Note that the above inequality is true for any η. So, put $\eta = \dfrac{1}{\beta}$. Doing so, we obtain

$$h_t \geq \frac{1}{2\beta} \left|\left|\nabla_t\right|\right|^2$$

which is exactly what we wanted to prove.

Let us prove (4) now. Let $\boldsymbol{x} \in \mathcal{K} = \mathbf{R}^d$ be fixed, and consider the following expression.

$$f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{x} \rangle + \frac{\alpha}{2} \left|\left|\boldsymbol{z} - \boldsymbol{x}\right|\right|^2$$

Since $\mathcal{K} = \mathbf{R}^d$, this expression attains a global minimum at the point where the gradient (w.r.t $\boldsymbol{z}$) is zero, i.e when

$$\boldsymbol{z} = \boldsymbol{x} - \frac{\nabla f(\boldsymbol{x})}{\alpha}$$

So, for any $\boldsymbol{y}, \boldsymbol{x} \in \mathcal{K} = \mathbf{R}^d$, we have the following, where we substitute the above value of $\boldsymbol{z}$ in the given expression.

$$\begin{aligned} f(\boldsymbol{y}) &\geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\alpha}{2} \left|\left|\boldsymbol{y} - \boldsymbol{x}\right|\right|^2 \\ &\geq f(\boldsymbol{x}) - \frac{\left|\left|\nabla f(\boldsymbol{x})\right|\right|}{2\alpha} \end{aligned}$$

Now, let $\boldsymbol{y} = \boldsymbol{x}^*$ and let $\boldsymbol{x} = \boldsymbol{x}_t$. This gives us

$$h_t \leq \frac{\left|\left|\nabla_t\right|\right|^2}{2\alpha}$$

which is what we wanted to prove. ■

3.2. **The Polyak Step Size.** Let h_t, d_t, ∇_t and η_t have the same meanings as in the previous section. The *Polyak Step Size* is defined as follows.

$$\eta_t = \frac{h_t}{\left|\left|\nabla_t\right|\right|^2}$$

Lemma 3.2. *Suppose f is γ-well conditioned and $\boldsymbol{x}_0, ..., \boldsymbol{x}_{T-1}$ are such that*

$$d_{t+1}^2 \leq d_t^2 - \frac{h_t^2}{\left|\left|\nabla_t\right|\right|^2}$$

Put

$$\overline{\boldsymbol{x}} = \underset{\boldsymbol{x}_t}{\operatorname{argmin}} \left\{f(\boldsymbol{x}_t)\right\}$$

Then

$$f(\overline{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \leq \frac{1}{T}\sum_{t=0}^{T-1} h_t \leq B_T$$

where

$$B_T = \min\left\{\frac{Gd_0}{\sqrt{T}}, \frac{2\beta d_0^2}{T}, \frac{4G^2}{\alpha T}, \beta d_0^2\left(1 - \frac{\gamma}{4}\right)^T\right\}$$

where for all t,

$$\left|\left|\nabla_t\right|\right| \leq G$$

The above condition is the same as Lipschitz continuity.

Proof. First, consider the differences $f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)$ for $t = 0$ to $T - 1$. By the fact that the average is $\geq$ the minimum, we immediately see that

$$f(\overline{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \leq \frac{1}{T}\sum_{t=0}^{T} h_t$$

This establishes the first half of the inequality. We now prove the second half, which is the trickier one.

Now, first consider the case when $\left|\left|\nabla_t\right|\right| \leq G$, the Lipschitz continuity case of f. We know the following.

$$d_{t+1}^2 - d_t^2 \leq \frac{-h_t^2}{\left|\left|\nabla_t\right|\right|^2} \leq \frac{-h_t^2}{G^2}$$

The above implies that

$$d_t^2 - d_{t+1}^2 \geq \frac{1}{G^2}h_t^2$$

Summing the above inequalities from $t = 0$ to $t = T - 1$, we get the following.

$$d_0^2 - d_T^2 \geq \frac{1}{G^2} \sum_{t=0}^{T-1} h_t^2$$

The above inequality implies that $G^2 d_0^2$ is an upper bound of $\sum_{t=0}^{T-1} h_t^2$. Combining this fact with the Cauchy-Schwarz inequality, we get that

$$\frac{h_0 + h_1 + ... + h_{T-1}}{T} = (h_0, ..., h_{T-1}) \cdot \left(\frac{1}{T}, ..., \frac{1}{T}\right) \leq \frac{\sqrt{\sum_{t=0}^{T-1} h_t^2}}{\sqrt{T}} \leq \frac{\sqrt{G^2 d_0^2}}{\sqrt{T}} = \frac{G d_0}{\sqrt{T}}$$

Next, suppose that f is β-smooth, and that $\eta_t = \dfrac{1}{\beta}$ for all t. Then, by part (3) of **Proposition** 3.1, we know the following.

$$d_{t+1}^2 - d_t^2 \leq \frac{h_t^2}{||\nabla_t||^2} \leq \frac{-h_t}{2\beta}$$

Again, summing the above inequalities for each t, we get

$$d_0^2 - d_T^2 \geq \frac{1}{2\beta} \sum_{t=0}^{T-1} h_t$$

which implies

$$\frac{2\beta d_0^2}{T} \geq \frac{1}{T} \sum_{t=0}^{T-1} h_t$$

Next, suppose f is γ-well conditioned, and again suppose that $\eta_t = \dfrac{1}{\beta}$. By part (3) **Proposition** 3.1, we know that

$$d_{t+1}^2 - d_t^2 \leq \frac{-h_t^2}{||\nabla_t||^2} \leq \frac{-h_t}{2\beta}$$

Also, by part (1) of the same proposition, we know that

$$h_t \geq \frac{\alpha}{2} d_t^2$$

Combining the above two inequalities, we see that

$$d_{t+1}^2 - d_t^2 \leq \frac{-h_t}{2\beta} \leq -\frac{\alpha}{4\beta} d_t^2$$

which implies that

$$d_{t+1}^2 \leq \left(1 - \frac{\gamma}{4}\right) d_t^2$$

By induction, this implies that

$$d_T^2 \leq \left(1 - \frac{\gamma}{4}\right)^T d_0^2$$

Finally, using the inequality $h_T \leq \beta d_T^2$ (follows from the fact that $\boldsymbol{x}^*$ is the minimizer), we see that

$$f(\overline{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \leq h_T \leq \beta d_T^2 \leq d_0^2 \beta \left(1 - \frac{\gamma}{4}\right)^T$$

and this proves the claim.

Finally, suppose f is an α-strongly convex function (and ofcourse we are assuming that the gradients are bounded by G). Now, by our assumption, we know that

$$d_{t+1}^2 - d_t^2 \leq \frac{-h_t^2}{||\nabla_t||^2} \leq \frac{-h_t^2}{G^2}$$

By α-strong convexity of f, we have the following.

$$\frac{-h_t^2}{G^2} \leq \frac{-\alpha^2 d_t^4}{4G^2}$$

The above two inequalities combined together give us the following.

$$d_{t+1}^2 \leq d_t^2 \left(1 - \frac{\alpha^2 d_t^2}{4G^2}\right)$$

Multiplying both sides by $\alpha^2/4G^2$, we get the following.

$$\frac{\alpha^2 d_{t+1}^2}{4G^2} \leq \frac{\alpha^2 d_t^2}{4G^2}\left(1 - \frac{\alpha^2 d_t^2}{4G^2}\right)$$

Now let $a_t = \frac{\alpha^2 d_t^2}{4G^2}$. The above inequality can be written as

$$a_{t+1} \leq a_t(1 - a_t)$$

and also

$$a_0 = \frac{\alpha^2 d_0^2}{4G^2}$$

Observe that

$$f(\boldsymbol{x}_0) \geq f(\boldsymbol{x}^*) + \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle + \frac{\alpha}{2} d_0^2 \geq f(\boldsymbol{x}^*) + \frac{\alpha}{2} d_0^2$$

where in the last inequality, we have used the fact that the above inner product is non-negative, as $\boldsymbol{x}^*$ is the minimizer of f. So, we see that

$$G d_0 \geq f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*) \geq \frac{\alpha}{2} d_0^2$$

where above we have used the Lipschitz condition. Clearly, this implies that

$$2G \geq \alpha d_0$$

which means that

$$a_0 \leq 1$$

By induction, and using the fact that $a_{t+1} \leq a_t(1 - a_t)$, we can show that

$$a_t \leq \frac{1}{t+1}$$

Next, we know that

$$\frac{h_t^2}{||\nabla_t||^2} \leq d_t^2 - d_{t+1}^2$$

which implies that

$$h_t^2 \leq G^2(d_t^2 - d_{t+1}^2)$$

Summing the above quantity from $T = T/2$ to $T - 1$, we see that

$$\frac{1}{T/2} \sum_{t=T/2}^{T-1} h_t^2 \leq \frac{2G^2}{T}[d_{T/2}^2 - d_T^2] = \frac{8G^2}{\alpha^2 T}[a_{T/2} - a_T] \leq \frac{16G^4}{\alpha^2 T^2}$$

by using the fact that $a_{T/2} \leq 1/(T/2+1)$. So, there is some t between $T/2$ and $T-1$ such that

$$h_t^2 \leq \frac{16G^4}{\alpha^2 T^2}$$

which implies that

$$h_t \leq \frac{4G^2}{\alpha T}$$

and hence we've shown all the four inequalities. ∎

Proposition 3.3. *The Polyak step size satisfies*

$$d_{t+1}^2 - d_t^2 \leq \frac{-h_t^2}{||\nabla_t||^2}$$

Infact, this holds even in a constrained optimization problem (i.e $\mathcal{K}$ need not be $\mathbf{R}^d$) and even if we replace the gradient by a subgradient.

Proof. We will prove this for a general constrained optimization problem. In that case, the update rules are as follows.

$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_t - \eta_t \nabla_t \quad , \quad \boldsymbol{x}_{t+1} = \Pi_{\mathcal{K}} \boldsymbol{y}_{t+1}$$

where Π represents the projection. First, observe that

$$||\boldsymbol{x}_{t+1} - \boldsymbol{x}^*||^2 \leq ||\boldsymbol{y}_{t+1} - \boldsymbol{x}^*||^2$$

which is true by **Theorem** 2.8 (**Pythagoras Theorem**). The above inequality gives us the following.

$$\begin{aligned} ||\boldsymbol{x}_{t+1} - \boldsymbol{x}^*||^2 &\leq ||\boldsymbol{x}_t - \eta_t \nabla_t - \boldsymbol{x}^*||^2 \\ &= ||\boldsymbol{x}_t - \boldsymbol{x}^*||^2 + \eta_t^2 ||\nabla_t||^2 - 2\eta_t \langle \nabla_t, \boldsymbol{x}_t - \boldsymbol{x}^* \rangle \end{aligned}$$

The above inequality can be written as follows.

$$d_{t+1}^2 \leq d_t^2 + \eta_t^2 ||\nabla_t||^2 - 2\eta_t \langle \nabla_t, \boldsymbol{x}_t - \boldsymbol{x}^* \rangle$$

Now, by convexity of f, we know the following.

$$f(\boldsymbol{x}^*) \geq f(\boldsymbol{x}_t) + \langle \nabla_t, \boldsymbol{x}^* - \boldsymbol{x}_t \rangle$$

Combining the above inequality with the previous inequality, we get the following.

$$\begin{aligned} d_{t+1}^2 &\leq d_t^2 + \eta_t^2 ||\nabla_t||^2 - 2\eta_t (f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) \\ &= d_t^2 + \eta_t^2 ||\nabla_t||^2 - 2\eta_t h_t \end{aligned}$$

Now, we use the fact that η_t is actually the Polyak step size, i.e

$$\eta_t = \frac{h_t}{||\nabla_t||^2}$$

Doing so, we get the following.

$$d_{t+1}^2 \leq d_t^2 - \frac{h_t^2}{||\nabla_t||^2}$$

and this proves the claim. ∎

3.3. **Exponential Convergence for Projected GD in Unconstrained Optimization.** In this section, we will prove an exponential convergence bound for the general case of unconstrained optimization, i.e $\mathcal{K}$ need not be $\mathbf{R}^d$. We have the following theorem.

Theorem 3.4. *Let f be a $\gamma = \alpha/\beta$-well conditioned convex function on $\mathcal{K}$. Let $\boldsymbol{x}^*$ be the minimizer of f on $\mathcal{K}$. Then, projected gradient descent with $\eta_t = 1/\beta$ satisfies the following.*

$$h_{t+1} \leq h_1 e^{-\frac{\gamma t}{4}}$$

Proof. This is just **Theorem 2.4** of Hazan's book. We will just complete the missing details here.

The only missing detail in the proof is showing the following equality (which is also problem 6. of the same chapter).

$$\underset{\boldsymbol{x}\in\mathcal{K}}{\operatorname{argmin}} \left\{ \langle \nabla_t, \boldsymbol{x} - \boldsymbol{x}_t \rangle + \frac{1}{2\eta_t} \left|\left| \boldsymbol{x} - \boldsymbol{x}_t \right|\right|^2 \right\} = \underset{\boldsymbol{x}\in\mathcal{K}}{\operatorname{argmin}} \left\{ \left|\left| \boldsymbol{x} - (\boldsymbol{x}_t - \eta_t \nabla_t) \right|\right|^2 \right\}$$

This is easy to see, because we have the following.

$$\begin{aligned} \left|\left| \boldsymbol{x} - (\boldsymbol{x}_t - \eta_t \nabla_t) \right|\right|^2 &= \left|\left| \boldsymbol{x} - \boldsymbol{x}_t + \eta_t \nabla_t \right|\right|^2 \\ &= \left|\left| \boldsymbol{x} - \boldsymbol{x}_t \right|\right|^2 + \eta_t^2 \left|\left| \nabla_t \right|\right|^2 + 2\eta_t \langle \nabla_t, \boldsymbol{x} - \boldsymbol{x}_t \rangle \end{aligned}$$

Observe that the quantity $\eta_t^2 \left|\left| \nabla_t \right|\right|^2$ is independent of the choice of $\boldsymbol{x}$. So, minimizing

$$\left|\left| \boldsymbol{x} - (\boldsymbol{x}_t - \eta_t \nabla_t) \right|\right|^2$$

Is the same as minimizing

$$\left|\left| \boldsymbol{x} - \boldsymbol{x}_t \right|\right|^2 + 2\eta_t \langle \nabla_t, \boldsymbol{x} - \boldsymbol{x}_t \rangle$$

which in turn is the same as minimizing

$$\langle \nabla_t, \boldsymbol{x} - \boldsymbol{x}_t \rangle + \frac{1}{2\eta_t} \left|\left| \boldsymbol{x} - \boldsymbol{x}_t \right|\right|^2$$

and this proves the equality. ■

3.4. **Online Gradient Descent.** In this section, we will explore the online version of the usual gradient descent algorithm. As usual, our input will be a convex body $\mathcal{K}$, an initial point $\boldsymbol{x}_1 \in \mathcal{K}$, a time horizon T and step sizes η_t. At each time t, the loss function f_t is revealed, and we want to minimize the *regret* (as we defined before).

$$\text{regret}_T = \sum_{t=1}^{T} f_t(\boldsymbol{x}_t) - \min_{\boldsymbol{x}^*\in\mathcal{K}} \sum_{t=1}^{T} f_t(\boldsymbol{x}^*)$$

Algorithm 1 Online Gradient Descent

```
Input: $\mathcal{K}$, $\boldsymbol{x}_1 \in \mathcal{K}$, $T$, $\eta_t$
for $t = 1$ to $T$ do
    Play $\boldsymbol{x}_t$ and get the cost $f_t(\boldsymbol{x}_t)$.
    $\boldsymbol{y}_{t+1} = \boldsymbol{x}_t - \eta_t \nabla f_t(\boldsymbol{x}_t)$
    $\boldsymbol{x}_{t+1} = \Pi_{\mathcal{K}}(\boldsymbol{y}_{t+1})$
end for
return $\boldsymbol{x}_{T+1}$
```

Now we will show that this online version of the gradient descent algorithm achieves *sublinear* regret.

Theorem 3.5. *Let the setup be as above. Then, online gradient descent with step sizes* $\eta_t = \frac{D}{G\sqrt{t}}$ *for* $t \in [T]$ *achieves the following for all* $T \geq 1$*.*

$$\text{regret}_T \leq \frac{3}{2} GD\sqrt{T}$$

where G *is an upper bound on the gradients, and* D *is the diameter of the convex body* $\mathcal{K}$*.*

Proof. First, let

$$\boldsymbol{x}^* = \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{K}} \sum_{t=1}^{T} f_t(\boldsymbol{x})$$

Now, by **Theorem** 2.8 (**Pythagoras Theorem**), we know the following.

$$||\boldsymbol{x}_{t+1} - \boldsymbol{x}^*||^2 \leq ||\boldsymbol{y}_{t+1} - \boldsymbol{x}^*||^2 = ||\boldsymbol{x}_t - \eta_t \nabla_t - \boldsymbol{x}^*||^2$$

So, we get the following.

$$\begin{aligned} ||\boldsymbol{x}_{t+1} - \boldsymbol{x}^*||^2 &\leq ||\boldsymbol{x}_t - \eta_t \nabla_t - \boldsymbol{x}^*||^2 \\ &= ||\boldsymbol{x}_t - \boldsymbol{x}^*||^2 + \eta_t^2 ||\nabla_t||^2 - 2\eta_t \langle \nabla_t, \boldsymbol{x}_t - \boldsymbol{x}^* \rangle \end{aligned}$$

Rearranging the above inequality, we get the following.

$$\begin{aligned} 2 \langle \nabla_t, \boldsymbol{x}_t - \boldsymbol{x}^* \rangle &\leq \frac{||\boldsymbol{x}_t - \boldsymbol{x}^*||^2 - ||\boldsymbol{x}_{t+1} - \boldsymbol{x}^*||^2}{\eta_t} + \eta_t ||\nabla_t||^2 \\ &\leq \frac{||\boldsymbol{x}_t - \boldsymbol{x}^*||^2 - ||\boldsymbol{x}_{t+1} - \boldsymbol{x}^*||^2}{\eta_t} + \eta_t G^2 \end{aligned}$$

Moreover, by convexity of f_t, we know the following.

$$f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}^*) \leq \langle \nabla_t, \boldsymbol{x}_t - \boldsymbol{x}^* \rangle$$

Combining the last two inequalities, we get the following.

$$2(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}^*)) \leq \frac{||\boldsymbol{x}_t - \boldsymbol{x}^*||^2 - ||\boldsymbol{x}_{t+1} - \boldsymbol{x}^*||^2}{\eta_t} + \eta_t G^2$$

Note that the above inequality is true for all $t \in [T]$. So, summing the above inequality for $t \in [T]$, we get the following, where we are using the fact that $||\boldsymbol{x}_{T+1} - \boldsymbol{x}^*|| \geq 0$.

$$\begin{aligned} 2 \sum_{t=1}^{T} f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}^*) &\leq \sum_{t=1}^{T} \frac{||\boldsymbol{x}_t - \boldsymbol{x}^*||^2 - ||\boldsymbol{x}_{t+1} - \boldsymbol{x}^*||^2}{\eta_t} + G^2 \sum_{t=1}^{T} \eta_t \\ &= \sum_{t=1}^{T} ||\boldsymbol{x}_t - \boldsymbol{x}^*||^2 \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) + G^2 \sum_{t=1}^{T} \eta_t \\ &\leq \sum_{t=1}^{T} D^2 \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) + G \sum_{t=1}^{T} \eta_t \\ &= D^2 \frac{1}{\eta_T} + G^2 \sum_{t=1}^{T} \eta_t \end{aligned}$$

where above we are using the convention that $\frac{1}{\eta_0} = 0$. Finally, by our choice of η_t, we have the following.

$$\begin{aligned} D^2 \frac{1}{\eta_T} + G^2 \sum_{t=1}^{T} \eta_t &= D^2 \frac{G\sqrt{T}}{D} + G^2 \sum_{t=1}^{T} \frac{D}{G\sqrt{t}} \\ &= DG\sqrt{T} + DG \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \\ &\leq 3DG\sqrt{T} \end{aligned}$$

and this proves the claim. ■

3.5. **OGD for Strongly Convex Functions.** Next, we will show that for strongly convex functions, a better step size selection can lead to a better regret bound.

Theorem 3.6. *Let $\mathcal{K}$ be a convex domain, $\boldsymbol{x}_1 \in \mathcal{K}$ be an initial point, and T be a time horizon. Let f_t be the revealed cost functions. Suppose each f_t is α-strongly convex. Then, doing OGD with step sizes $\eta_t = \frac{1}{\alpha t}$ gives the following regret bound.*

$$\text{regret}_T \leq \frac{G^2}{2\alpha}(1 + \log T)$$

where G is an upper bound on the gradients.

Proof. The proof is very similar to that of **Theorem** 3.5. As usual, let

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}\in\mathcal{K}}{\text{argmin}} \sum_{t=1}^{T} f_t(\boldsymbol{x})$$

Now, by α-strong convexity, we know that

$$f_t(\boldsymbol{x}^*) \geq f_t(\boldsymbol{x}_t) + \langle \nabla_t, \boldsymbol{x}^* - \boldsymbol{x}_t \rangle + \frac{\alpha}{2} ||\boldsymbol{x}^* - \boldsymbol{x}_t||^2$$

The above inequality implies the following inequality.

$$2(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}^*)) \leq 2 \langle \nabla_t, \boldsymbol{x}_t - \boldsymbol{x}^* \rangle) - \alpha ||\boldsymbol{x}_t - \boldsymbol{x}^*||^2 \tag{3.1}$$

Also, just like in the proof of **Theorem** 3.5, we have the following inequality.

$$2 \langle \nabla_t, \boldsymbol{x}_t - \boldsymbol{x}^* \rangle \leq \frac{||\boldsymbol{x}_t - \boldsymbol{x}^*||^2 - ||\boldsymbol{x}_{t+1} - \boldsymbol{x}^*||^2}{\eta_t} + \eta_t G^2$$

So, from the above inequality and inequality (3.1), we get the following.

$$2(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}^*)) \leq \frac{||\boldsymbol{x}_t - \boldsymbol{x}^*||^2 - ||\boldsymbol{x}_{t+1} - \boldsymbol{x}^*||^2}{\eta_t} + \eta_t G^2 - \alpha ||\boldsymbol{x}_t - \boldsymbol{x}^*||^2$$

Finally, summing the above inequalities over all t, we get the following, where our convention is $1/\eta_0 = 0$, and we are using the fact that $||\boldsymbol{x}_{T+1} - \boldsymbol{x}^*|| \geq 0$.

$$\begin{aligned} 2\,\text{regret}_T &\leq \sum_{t=1}^{T} ||\boldsymbol{x}_t - \boldsymbol{x}^*||^2 \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \alpha \right) + G^2 \sum_{t=1}^{T} \eta_t \\ &= 0 + \frac{G^2}{\alpha} \sum_{t=1}^{T} \frac{1}{t} \\ &\leq \frac{G^2}{\alpha}(1 + \log T) \end{aligned}$$

and this proves the claim. ■

3.6. **OGD without knowing D and G.** In this section, we will devise an online GD algorithm which gives the same regret bounds as the one we studied in the previous sections (i.e $O(\sqrt{T})$ regret), but the catch here will be: we don't actually know what D and G are. We only know that they exist. Using only this information, we can prove the following theorem.

Theorem 3.7. *Let the notation be as before. Suppose f is a convex function on $\mathcal{K}$, such that $||\nabla f(\boldsymbol{x})|| \leq G$ for all $\boldsymbol{x} \in \mathcal{K}$ and the diameter of $\mathcal{K}$ is D. Further, assume that the actual values of G, D are not known. For each $t \in [T]$, define D_t as follows.*

$$D_1 = 1$$
$$D_t = \begin{cases} D_{t-1} & , \quad \text{if } ||\boldsymbol{x}_t - \boldsymbol{x}_1|| \leq D_{t-1} \\ 2D_{t-1} & , \quad \text{otherwise} \end{cases}$$

Similarly, for each $t \in [T]$, define G_t as follows.

$$G_1 = ||\nabla_1||$$
$$G_t = \max(G_{t-1}, ||\nabla_t||)$$

Then, for step sizes $\eta_t = \frac{D_t}{G_t\sqrt{t}}$, OGD gives the following guarantee on regret.

$$\text{regret}_T \leq O(\sqrt{T})$$

Proof. As usual, let

$$\boldsymbol{x}^* = \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{K}} \sum_{t=1}^{T} f_t(\boldsymbol{x})$$

First, observe that $D_1 \leq D_2 \leq \cdots \leq D_T$ and similarly $G_1 \leq G_2 \leq \cdots \leq G_T$. This is easy to see from the definitions of these sequences.

Now, just as in the proof of **Theorem** 3.5, we will get the following inequality.

$$2(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}^*)) \leq \frac{||\boldsymbol{x}_t - \boldsymbol{x}^*||^2 - ||\boldsymbol{x}_{t+1} - \boldsymbol{x}^*||^2}{\eta_t} + \eta_t\, ||\nabla_t||^2$$

Note that the above inequality is true for all $t \in [T]$. So, summing over all t, we get the following, where again the convention is $1/\eta_0 = 0$ and we are using the fact that

$||\boldsymbol{x}_{T+1} - \boldsymbol{x}^*|| \geq 0$.

$$\begin{aligned}
2 \cdot \text{regret}_T &\leq \sum_{t=1}^{T} ||\boldsymbol{x}_t - \boldsymbol{x}^*||^2 \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \sum_{t=1}^{T} \eta_t \, ||\nabla_t||^2 \\
&\leq \sum_{t=1}^{T} D^2 \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \sum_{t=1}^{T} \eta_t \, ||\nabla_t||^2 \\
&\leq \frac{D^2}{\eta_T} + \sum_{t=1}^{T} \frac{D_t}{G_t\sqrt{t}} G_t^2 \\
&= \frac{D^2 G_T \sqrt{T}}{D_T} + \sum_{t=1}^{T} \frac{D_t G_t}{\sqrt{t}} \\
&\leq \frac{D^2 G_T \sqrt{T}}{D_T} + D_T G_T \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \\
&\leq \frac{D^2 G_T \sqrt{T}}{D_T} + 2 D_T G_T \sqrt{T}
\end{aligned}$$

Above, we have used the facts that D_t, G_t are non-decreasing sequences. Now, observe that $G_T \leq G$ (because G is an upper bound on the gradients, and G_T is the maximum norm of a gradient seen till time T). So, we get that

$$2 \cdot \text{regret}_T \leq \frac{D^2 G \sqrt{T}}{D_T} + 2 D_T G \sqrt{T}$$

Now, we consider two cases.

(1) In the first case, we have $D_T \leq D$. Also, we know that $1 \leq D_T$. So, in this case we see that

$$\frac{D^2 G \sqrt{T}}{D_T} + 2 D_T G \sqrt{T} \leq D^2 G \sqrt{T} + 2 D G \sqrt{T} = O(\sqrt{T})$$

and hence we have an $O(\sqrt{T})$ regret bound.

(2) In the second case, we have $D < D_T$. Suppose $t_0 + 1 \leq T$ is the *last time step* when the sequence D_t was updated, i.e

$$D_{t_0+1} = 2D_{t_0}$$

Clearly, we see that $D_T = D_{t_0+1} = 2D_{t_0}$. Also, by our definition, this update happened only because

$$D_{t_0} < ||\boldsymbol{x}_{t_0+1} - \boldsymbol{x}_1|| \leq D$$

So, we have that

$$D_{t_0} < D < D_T$$

which is the same as the inequality

$$\frac{D_T}{2} < D < D_T$$

In this case, we have that

$$\frac{D^2 G \sqrt{T}}{D_T} + 2 D_T G \sqrt{T} \leq D G \sqrt{T} + 4 D G \sqrt{T} = O(\sqrt{T})$$

and hence in this case as well, we have an $O(\sqrt{T})$ regret bound.

So, in all cases the given regret bound follows, and this completes the proof of the claim. ■

Lemma 3.8. *If the sequence $\{D_t\}$ is defined as in the previous theorem, then for any $\tau \leq T$, it is true that for all $t \leq \tau$,*

$$||\boldsymbol{x}_t - \boldsymbol{x}_1|| \leq D_\tau$$

Proof. We will prove this by induction. For the base case, suppose $\tau = 1$, and we know that $D_1 = 1$. Now, note that

$$||\boldsymbol{x}_1 - \boldsymbol{x}_1|| = 0 < D_1$$

and this proves the base case.

For the inductive case, suppose the statement is true for some τ. We will prove it for $\tau + 1$. Now, if $t \leq \tau$, then by the inductive hypothesis, combined with the fact that the sequence $\{D_t\}$ is monotonic, we have that

$$||\boldsymbol{x}_t - \boldsymbol{x}_1|| \leq D_\tau \leq D_{\tau+1}$$

Now, consider the time step $\tau + 1$. If

$$||\boldsymbol{x}_{\tau+1} - \boldsymbol{x}_1|| \leq D_\tau$$

then by definition we know that $D_{\tau+1} = D_\tau$, and hence there is nothing to prove. So, suppose $||\boldsymbol{x}_{\tau+1} - \boldsymbol{x}_1|| > D_\tau$. In that case, we know that $D_{\tau+1} = 2D_\tau$. Now, observe the following.

$$\begin{aligned} ||\boldsymbol{x}_{\tau+1} - \boldsymbol{x}_1|| &\leq ||\boldsymbol{x}_{\tau+1} - \boldsymbol{x}_\tau|| + ||\boldsymbol{x}_\tau - \boldsymbol{x}_1|| \\ &\leq ||\boldsymbol{x}_{\tau+1} - \boldsymbol{x}_\tau|| + D_\tau \end{aligned}$$

Now, because $\boldsymbol{x}_{\tau+1} = \pi_{\mathcal{K}}(\boldsymbol{y}_{\tau+1})$, by **Theorem** 2.8 (**Pythagoras Theorem**) we have that

$$\begin{aligned} ||\boldsymbol{x}_{\tau+1} - \boldsymbol{x}_\tau|| &\leq ||\boldsymbol{y}_{t+1} - \boldsymbol{x}_\tau|| \\ &= ||\boldsymbol{x}_\tau - \eta_\tau \nabla_\tau - \boldsymbol{x}_\tau|| \\ &= \eta_\tau \, ||\nabla_\tau|| \\ &\leq \frac{D_\tau}{G_\tau \sqrt{\tau}} \cdot G_\tau \\ &\leq D_\tau \end{aligned}$$

So, combining this with the previous inequality, we see that

$$||\boldsymbol{x}_{\tau+1} - \boldsymbol{x}_1|| \leq D_\tau + D_\tau = 2D_\tau = D_{\tau+1}$$

and this completes the proof by induction. ■

3.7. **Stochastic Gradient Descent.** In this section, we will introduce *stochastic gradient descent* and analyse it for two cases: the first case for normal convex functions, and the second for strongly convex functions.

As usual, we are given a convex set $\mathcal{K}$, and we want to minimize some function f on $\mathcal{K}$. However, unlike offline GD, we are given access to a gradient oracle, defined as follows.

$$\mathcal{O}(\boldsymbol{x}) := \tilde{\nabla}_{\boldsymbol{x}}$$

Moreover, the oracle has the property that

$$\mathbf{E}\left[\tilde{\nabla}_{\boldsymbol{x}}\right] = \nabla f(\boldsymbol{x}) \quad , \quad \mathbf{E}\left[\left|\left|\tilde{\nabla}_{\boldsymbol{x}}\right|\right|^2\right] \leq G^2$$

In simple words, the expected value of the gradient returned by the oracle for a point $\boldsymbol{x}$ is the true gradient $\nabla f(\boldsymbol{x})$.

Example 3.1. A version of SGD works as follows: we are given a data set S of n points, and we want to optimize some parameter $\boldsymbol{\theta}$. The loss function in many scenarios is of the following form.

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} L_{\boldsymbol{x}_i, y_i}(\boldsymbol{\theta})$$

where $(\boldsymbol{x}_i, y_i)$ is the ith data point. So, to compute the gradient $\nabla L(\boldsymbol{\theta})$, we need to iterate over all the data points. which might be costly. Instead, we use SGD, by randomly sampling an index j (uniformly at random), and computing the gradient $\nabla L_{\boldsymbol{x}_j, y_j}(\boldsymbol{\theta})$. This almost fits in with our oracle description, because the expected value of the gradient is $\frac{\nabla L(\boldsymbol{\theta})}{n}$.

Algorithm 2 Stochastic Gradient Descent

```
Input: $f$, $\mathcal{K}$, $\boldsymbol{x}_1 \in \mathcal{K}$, step sizes $\eta_t$.
for $t = 1$ to $T$ do
    Let $\tilde{\nabla}_t = \mathcal{O}(\boldsymbol{x}_t)$ and define $f_t(\boldsymbol{x}) := \left\langle \tilde{\nabla}_t, \boldsymbol{x} \right\rangle$.
    Update $\boldsymbol{y}_{t+1} \leftarrow \boldsymbol{x}_t - \eta_t \tilde{\nabla}_t$
    Project $\boldsymbol{x}_{t+1} = \pi_{\mathcal{K}}(\boldsymbol{y}_{t+1})$
end for
return $\overline{\boldsymbol{x}_T} := \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t$
```

Theorem 3.9. *Let the notation be as above. Then SGD with step sizes* $\eta_t = \frac{D}{G\sqrt{t}}$ *guarantees the following.*

$$\mathbf{E}\left[f(\overline{\boldsymbol{x}_T})\right] \leq \min_{\boldsymbol{x}^* \in \mathcal{K}} f(\boldsymbol{x}^*) + \frac{3GD}{2\sqrt{T}}$$

Proof. We will use the regret guarantee of the OGD algorithm as we proved in **Theorem** 3.5. We have the following. (In the second step below, we use Jensen's Inequality;

in the third step, we use the convexity of f.

$$\begin{aligned}
&\mathbf{E}\left[f(\overline{\boldsymbol{x}_T})\right] - f(\boldsymbol{x}^*) \\
&\le \frac{1}{T}\mathbf{E}\left[\sum_t f(\boldsymbol{x}_t)\right] - f(\boldsymbol{x}^*) \\
&= \frac{1}{T}\mathbf{E}\left[\sum_t [f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)]\right] \\
&\le \frac{1}{T}\mathbf{E}\left[\sum_t \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^* \rangle\right] \\
&= \frac{1}{T}\mathbf{E}\left[\sum_t \left\langle \tilde{\nabla}_t, \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle\right] \\
&= \frac{1}{T}\mathbf{E}\left[\sum_t f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}^*)\right] \\
&\le \frac{\text{regret}_T}{T} \\
&\le \frac{3GD}{2\sqrt{T}}
\end{aligned}$$

Ofcourse, here we are heavily using the fact that **Theorem** 3.5 works for any choice of f_t, i.e the theorem holds against even an adaptive adversary. ■

As before, we can make the convergence guarantee of SGD stronger for strongly-convex functions as we did for OGD. We will now prove this.

Theorem 3.10 (**SGD for Strongly Convex Functions**). *Let the notation be as above, and in addition suppose f is α-strongly convex. Then, with step sizes $\eta_t = \frac{1}{\alpha t}$, SGD has the following convergence guarantee.*

$$\mathbf{E}\left[f(\overline{\boldsymbol{x}_T})\right] \le \min_{\boldsymbol{x}^* \in \mathcal{K}} f(\boldsymbol{x}^*) + \frac{G^2}{2\alpha}\frac{(1 + \log T)}{T}$$

Proof. The proof is very similar to that of **Theorem** 3.9. For each t, we define the following function.

$$g_t(\boldsymbol{x}) = \left\langle \tilde{\nabla}_t, \boldsymbol{x} \right\rangle + \frac{\alpha}{2}\left|\left|\boldsymbol{x} - \boldsymbol{x}_1\right|\right|^2$$

It is clear that g_t is an α-strongly convex function for each t. Next, we have the following.

$$\begin{aligned}
&\mathbf{E}\left[f(\overline{\boldsymbol{x}_T})\right] - f(\boldsymbol{x}^*)\\
&\leq \frac{1}{T}\mathbf{E}\left[\sum_t f(\boldsymbol{x}_t)\right] - f(\boldsymbol{x}^*)\\
&= \frac{1}{T}\mathbf{E}\left[\sum_t [f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)]\right]\\
&\leq \frac{1}{T}\mathbf{E}\left[\sum_t \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^*\rangle - \frac{\alpha}{2}\left|\left|\boldsymbol{x}_t - \boldsymbol{x}^*\right|\right|^2\right]\\
&= \frac{1}{T}\mathbf{E}\left[\sum_t \left\langle \tilde{\nabla}_t, \boldsymbol{x}_t - \boldsymbol{x}^*\right\rangle - \frac{\alpha}{2}\left|\left|\boldsymbol{x}_t - \boldsymbol{x}^*\right|\right|^2\right]
\end{aligned}$$

Now, using the trivial inequality

$$-\frac{\alpha}{2}\left|\left|\boldsymbol{x}_t - \boldsymbol{x}^*\right|\right|^2 \leq \frac{\alpha}{2}\left|\left|\boldsymbol{x}_t - \boldsymbol{x}_1\right|\right|^2 - \frac{\alpha}{2}\left|\left|\boldsymbol{x}^* - \boldsymbol{x}_1\right|\right|^2$$

we get the following.

$$\begin{aligned}
&\frac{1}{T}\mathbf{E}\left[\sum_t \left\langle \tilde{\nabla}_t, \boldsymbol{x}_t - \boldsymbol{x}^*\right\rangle - \frac{\alpha}{2}\left|\left|\boldsymbol{x}_t - \boldsymbol{x}^*\right|\right|^2\right]\\
&\leq \frac{1}{T}\mathbf{E}\left[\sum_t \left\langle \tilde{\nabla}_t, \boldsymbol{x}_t - \boldsymbol{x}^*\right\rangle + \frac{\alpha}{2}\left|\left|\boldsymbol{x}_t - \boldsymbol{x}_1\right|\right|^2 - \frac{\alpha}{2}\left|\left|\boldsymbol{x}^* - \boldsymbol{x}_1\right|\right|^2\right]\\
&= \frac{1}{T}\mathbf{E}\left[\sum_t g_t(\boldsymbol{x}_t) - g_t(\boldsymbol{x}^*)\right]\\
&\leq \frac{\text{regret}_T}{T}\\
&\leq \frac{G^2}{2\alpha}\frac{(1 + \log T)}{T}
\end{aligned}$$

where in the second last step we have used **Theorem** 3.6, the convergence bound of OGD on strongly convex functions. This completes the proof. ■

Remark 3.10.1. Another way of saying the above theorem is that for strongly convex functions, the convergence bound is $\tilde{O}\left(\frac{1}{T}\right)$; this notation hides logarithmic factors.

4. Regularization

4.1. **Follow The Leader (FTL).** First, let us introduced the so called *Follow The Leader Strategy.* Recall that in the OCO framework, our goal is to optimize the *regret* of the algorithm. This motivates the following naive strategy: at time step $t+1$, choose the *best decision* at the time, i.e choose

$$\boldsymbol{x}_{t+1} = \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{K}} \sum_{\tau=1}^{t} f_\tau(\boldsymbol{x})$$

We will now show an example where this strategy fails miserably.

Example 4.1. Let $\mathcal{K} = [-1, 1]$ and let $f_1(x) = \frac{x}{2}$. Then, we have

$$f_t(x) = \begin{cases} -x & , \quad t \text{ is even} \\ x & , \quad \text{otherwise} \end{cases}$$

With these loss functions, it is clear that $\sum_{\tau=1}^{t} f_\tau(x) = \frac{x}{2}$ if t is odd, and $\sum_{\tau=1}^{t} f_\tau(x) = \frac{-x}{2}$ otherwise. So, the *Follow The Leader* strategy will fluctuate between the choice of -1 and 1. Thus at each time step, we incur a loss of $\frac{1}{2}$, which is *linear regret.* So, this strategy is obviously not the best strategy.

To prevent the algorithm from fluctuating as it did in the above example, we use the technique of *regularization*, i.e instead of juts minimizing $\sum_{\tau=1}^{t} f_t(\boldsymbol{x})$, we add a *regularizer* R to this sum and minimize the resulting expression. We will now make this formal.

Definition 4.1. Let $R : \mathcal{K} \to \mathbf{R}$ be a strongly convex function. Most of the time we will assume that R is a twice differentiable function, and strong convexity implies that $\nabla^2 R(\boldsymbol{x}) \succ 0$ for all $\boldsymbol{x} \in \mathcal{K}$, i.e the Hessian of R at each point is positive definite. The *diameter* of $\mathcal{K}$ with respect to R is defined as

$$D_R := \sqrt{\max_{\boldsymbol{x},\boldsymbol{y} \in \mathcal{K}} \{R(\boldsymbol{x}) - R(\boldsymbol{y})\}}$$

Definition 4.2. For any norm $||\cdot||$, define the *dual norm* $||\cdot||^*$ as follows.

$$||\boldsymbol{y}||^* := \max_{||\boldsymbol{x}|| \leq 1} \langle \boldsymbol{x}, \boldsymbol{y} \rangle$$

Proposition 4.1 (Generalised Cauchy-Schwarz Inequality). *For all $\boldsymbol{x}, \boldsymbol{y} \in V$, where V is some vector space with positive definit inner product $\langle \cdot, \cdot \rangle$,*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq ||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||^*$$

Proof. This is really trivial and follows from the definition of the dual norm: if $\boldsymbol{x} = \mathbf{0}$, then there is nothing to prove. So, suppose $\boldsymbol{x} \neq \mathbf{0}$, and hence $\frac{\boldsymbol{x}}{||\boldsymbol{x}||}$ is a unit vector. Now, from the definition of the dual norm, we see that

$$||\boldsymbol{y}||^* \geq \left\langle \frac{\boldsymbol{x}}{||\boldsymbol{x}||}, \boldsymbol{y} \right\rangle$$

The claim follows from here. ■

Definition 4.3. Given a positive definite matrix A, the *matrix norm* $||\cdot||_A$ is defined as follows.

$$||\boldsymbol{x}||_A := \sqrt{\boldsymbol{x}^T A \boldsymbol{x}}$$

Proposition 4.2 (Dual of Matrix Norm). *Let A be a positive definite matrix, and let $\langle \cdot, \cdot \rangle$ be defined as above. Then, for all $\boldsymbol{x} \in \mathbf{R}^n$,*

$$||\boldsymbol{x}||_A^* = ||\boldsymbol{x}||_{A^{-1}}$$

In simple words, the dual norm of the matrix norm induced by A is the matrix norm induced by A^{-1}.

Proof. To be completed. Do this proof! ■

4.2. **Bregman Divergence.** We have seen this quantity before. In this section, we will define it more generally.

Definition 4.4. The *Bregman Divergence* $B_R(\boldsymbol{x}||\boldsymbol{y})$ with respect to a function R is defined as follows.

$$B_R(\boldsymbol{x}||\boldsymbol{y}) = R(\boldsymbol{x}) - R(\boldsymbol{y}) - \langle \nabla R(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle$$

For twice differentiable functions R, **Taylor's Theorem** gives us the following expression for the Bregman Divergence between two points.

$$B_R(\boldsymbol{x}||\boldsymbol{y}) = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{y})^T\nabla^2 R(\boldsymbol{z})(\boldsymbol{x}-\boldsymbol{y}) =: \frac{1}{2}||\boldsymbol{x}-\boldsymbol{y}||_{\boldsymbol{z}}^2$$

Above, $\boldsymbol{z}$ is some point on the line segment between $\boldsymbol{x}$ and $\boldsymbol{y}$. Note that $\nabla^2 R(\boldsymbol{z})$ is an $n\times n$ matrix, and hence the quantity is nothing but the square of the matrix norm we defined before. Sometimes we will use the notation

$$\frac{1}{2}||\boldsymbol{x}-\boldsymbol{y}||_{\boldsymbol{x},\boldsymbol{y}}^2$$

for the above quantity.

In the OCO framework, as usual we choose points $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T$. Consider consecutive points $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$. We will simply use the notation $||\cdot||_t^2$ to denote the Bregman Divergence $||\boldsymbol{x}_t - \boldsymbol{x}_{t+1}||_{\boldsymbol{x}_t,\boldsymbol{x}_{t+1}}^2$, i.e we will use the notation

$$B_R(\boldsymbol{x}_t||\boldsymbol{x}_{t+1}) = \frac{1}{2}||\boldsymbol{x}_t - \boldsymbol{x}_{t+1}||_{\boldsymbol{x}_t,\boldsymbol{x}_{t+1}}^2 =: \frac{1}{2}||\boldsymbol{x}_t - \boldsymbol{x}_{t+1}||_t^2$$

4.3. **Regularized Follow The Leader (RFTL).** In this section, we will see a modification to the FTL algorithm that actually works very nicely. Let R be strongly convex, smooth and twice differentiable. Consider the following algorithm.

Algorithm 3 Regularized Follow The Leader

```
Input: $\eta$, $R$, $\mathcal{K}$.
Let $\boldsymbol{x}_1 \leftarrow \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{K}} R(\boldsymbol{x})$
for $t = 1$ to $T$ do
    Reveal $\boldsymbol{x}_t$.
    Observe $f_t$ and let $\nabla_t = \nabla f_t(\boldsymbol{x}_t)$.
    Update: $\boldsymbol{x}_{t+1} \leftarrow \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{K}} \left\{\eta\sum_{s=1}^t \langle\nabla_s, \boldsymbol{x}\rangle + R(\boldsymbol{x})\right\}$
end for
```

Note that RFTL is very similar to the usual FTL; the only difference is the addition of the regularizer, and the introduction of the hyperparameter η. We will now prove some regret bounds for this algorithm.

Lemma 4.3. *Consider the RFTL algorithm. For any T, the following is true.*

$$\text{regret}_T \le \sum_{t=1}^T \langle\nabla_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1}\rangle + \frac{1}{\eta}D_R^2$$

Proof. By convexity of the f_ts, we have the following inequality.

$$\sum_{t=1}^t f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}^*) \le \sum_{t=1}^T \langle\nabla_t, \boldsymbol{x}_t - \boldsymbol{x}^*\rangle \tag{4.1}$$

Define $g_0(\boldsymbol{x}) = \frac{R(\boldsymbol{x})}{\eta}$, and for all $t \geq 1$, define $g_t(\boldsymbol{x}) = \langle \nabla_t, \boldsymbol{x} \rangle$.

We claim that for all $u \in \mathcal{K}$,

$$\sum_{t=0}^{T} g_t(\boldsymbol{u}) \geq \sum_{t=0}^{T} g_t(\boldsymbol{x}_{t+1}) \tag{4.2}$$

We will prove this by induction. For the base case, note that

$$g_0(\boldsymbol{u}) \geq g_0(\boldsymbol{x}_1)$$

by the definition of g_0 and the choice of $\boldsymbol{x}_1$. Now suppose the claim is true for some τ, i.e suppose it is true that

$$\sum_{t=0}^{\tau} g_t(\boldsymbol{u}') \geq \sum_{t=0}^{\tau} g_t(\boldsymbol{x}_{t+1})$$

for all $\boldsymbol{u}' \in \mathcal{K}$ and we will prove it for $\tau + 1$. We want to show that

$$\sum_{t=0}^{\tau+1} g_t(\boldsymbol{u}) \geq \sum_{t=0}^{\tau+1} g_t(\boldsymbol{x}_{t+1})$$

Now by our choice of $\boldsymbol{x}_{\tau+2}$, we know that

$$\begin{aligned}
\sum_{t=0}^{\tau+1} g_t(\boldsymbol{u}) &\geq \sum_{t=0}^{\tau+1} g_t(\boldsymbol{x}_{\tau+2}) \\
&= g_{\tau+1}(\boldsymbol{x}_{\tau+2}) + \sum_{t=0}^{\tau} g_t(\boldsymbol{x}_{\tau+2}) \\
&\geq g_{\tau+1}(\boldsymbol{x}_{\tau+2}) + \sum_{t=0}^{\tau} g_t(\boldsymbol{x}_{t+1}) \\
&= \sum_{t=0}^{\tau+1} g_t(\boldsymbol{x}_{t+1})
\end{aligned}$$

and hence this proves our claim (4.2) by induction.

Now, (4.2) implies that for all $\boldsymbol{u} \in \mathcal{K}$,

$$g_0(\boldsymbol{u}) + \sum_{t=1}^{T} g_t(\boldsymbol{u}) \geq g_0(\boldsymbol{x}_1) + \sum_{t=1}^{T} g_t(\boldsymbol{x}_{t+1})$$

which implies that

$$-\sum_{t=1}^{T} g_t(\boldsymbol{u}) \leq -g_0(\boldsymbol{x}_1) - \sum_{t=1}^{T} g_t(\boldsymbol{x}_{t+1}) + g_0(\boldsymbol{u})$$

So, it follows that

$$\begin{aligned}\sum_{t=1}^{T} g_t(\boldsymbol{x}_t) - g_t(\boldsymbol{u}) &\leq \sum_{t=1}^{T} g_t(\boldsymbol{x}_t) - g_t(\boldsymbol{x}_{t+1}) + g_0(\boldsymbol{u}) - g_0(\boldsymbol{x}_1)\\ &= \sum_{t=1}^{T} g_t(\boldsymbol{x}_t) - g_t(\boldsymbol{x}_{t+1}) + \frac{R(\boldsymbol{u}) - R(\boldsymbol{x}_1)}{\eta}\\ &\leq \sum_{t=1}^{T} g_t(\boldsymbol{x}_t) - g_t(\boldsymbol{x}_{t+1}) + \frac{D_R^2}{\eta}\end{aligned}$$

and combining the above inequality with inequality (4.1), the lemma is proven. ■

Theorem 4.4. *For all $\boldsymbol{u} \in \mathcal{K}$, the following is true.*

$$\text{regret}_T \leq 2\eta \sum_{t=1}^{T} ||\nabla_t||_t^{*2} + \frac{R(\boldsymbol{u}) - R(\boldsymbol{x}_1)}{\eta}$$

Proof. We will use **Lemma** 4.3 (the previous lemma) to prove this; infact, this inequality is just a consequence of the previous lemma. For ease of notation, define Φ_t as follows.

$$\Phi_t(\boldsymbol{x}) = \eta \sum_{s=1}^{t} \langle \nabla_s, \boldsymbol{x} \rangle + R(\boldsymbol{x})$$

Then observe that in the RFTL algorithm, $\boldsymbol{x}_{t+1}$ is picked to minimize Φ_t. Now, by the definition of Bregman Divergence, we can write

$$\begin{aligned}\Phi_t(\boldsymbol{x}_t) &= \Phi_t(\boldsymbol{x}_{t+1}) + \langle \nabla\Phi_t(\boldsymbol{x}_{t+1}), \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle + B_{\Phi_t}(\boldsymbol{x}_t || \boldsymbol{x}_{t+1})\\ &\geq \Phi_t(\boldsymbol{x}_{t+1}) + B_{\Phi_t}(\boldsymbol{x}_t || \boldsymbol{x}_{t+1})\end{aligned}$$

Above, we have used the fact that $\langle \nabla\Phi_t(\boldsymbol{x}_{t+1}), \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle \geq 0$ (recall **Lemma** 2.7). Now, also note that the function Φ_t is a sum of linear functions and the regularizer R. Clearly, at any point in it's domain, the Hessian of Φ_t will be equal to the Hessian of R, since the Hessians of the linear terms will vanish. So, we see that

$$\Phi_t(\boldsymbol{x}_{t+1}) + B_{\Phi_t}(\boldsymbol{x}_t || \boldsymbol{x}_{t+1}) = \Phi_t(\boldsymbol{x}_{t+1}) + B_R(\boldsymbol{x}_t || \boldsymbol{x}_{t+1})$$

Combining the last two inequalities, we get the following.

$$\begin{aligned}B_R(\boldsymbol{x}_t || \boldsymbol{x}_{t+1}) &\leq \Phi_t(\boldsymbol{x}_t) - \Phi_t(\boldsymbol{x}_{t+1})\\ &= \Phi_{t-1}(\boldsymbol{x}_t) - \Phi_{t-1}(\boldsymbol{x}_{t+1}) + \eta \langle \nabla_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle\\ &\leq \eta \langle \nabla_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle\end{aligned}$$

where in the last step we have used the fact that $\Phi_{t-1}(\boldsymbol{x}_t) - \Phi_{t-1}(\boldsymbol{x}_{t+1}) \leq 0$. Now, by the generalised Cauchy Schwarz Inequality along with the above inequality, we have

$$\begin{aligned}\eta \langle \nabla_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle &\leq \eta \, ||\boldsymbol{x}_t - \boldsymbol{x}_{t+1}||_{\boldsymbol{x}_t, \boldsymbol{x}_{t+1}} \cdot ||\nabla_t||_t^*\\ &= \eta \sqrt{2B_R(\boldsymbol{x}_t || \boldsymbol{x}_{t+1})} \cdot ||\nabla_t||_t^*\\ &\leq \eta \sqrt{2\eta \langle \nabla_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle} \cdot ||\nabla_t||_t^*\end{aligned}$$

Squaring and rearranging the above inequality, we get the following.

$$\langle \nabla_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle \leq 2\eta \, ||\nabla_t||_t^*$$

The above inequality combined with **Lemma** 4.3 proves the theorem. ■

Remark 4.4.1. So, if there is an upper bound on $||\nabla_t||_t^*$, say G, then we can take step size $\eta_t = \frac{D_R}{G\sqrt{t}}$ to get $O(\sqrt{T})$ regret bounds.

4.4. **Online Mirrored Descent.** Didn't get time to typeset notes for this; but this is a very important idea. check out the book for this.

5. Appendix

5.1. **Singular Value Decomposition.** In this section, we will explore the *singular value decomposition* of a matrix.

Let A be an $n \times d$ matrix. Consider the rows of A as n data points living in $\mathbf{R}^d$. The *singular value decomposition* of A finds the *best* linear subspace of $\mathbf{R}^d$ that contains these points. There are various notions of *best*, but in our case, *best* will mean the following: we want to minimize the sum of the squares of the (perpendicular) distances of the points from the subspace. By the Pythagorean Theorem, this is equivalent to maximising the sum of the squares of the lengths of the (orthogonal) projections of the points onto the subspace. We will now make all of this formal.

5.1.1. *Singular Vectors.* As above, let A be an $n \times d$ matrix, where the rows of A are interpreted as n data points. Suppose these rows are $\boldsymbol{a}_1, ..., \boldsymbol{a}_n \in \mathbf{R}^d$. Consider the best fit line through the origin, and let $\boldsymbol{v}$ be a unit vector along this line. The length of the projection of $\boldsymbol{a}_i$ onto $\boldsymbol{v}$ is clearly $|\boldsymbol{a}_i \cdot \boldsymbol{v}|$, and hence the squared projection length is $|\boldsymbol{a}_i \cdot \boldsymbol{v}|^2$. So, it follows that the sum of the squares of the projection of the points onto this line is

$$\sum_{i=1}^{n} |\boldsymbol{a}_i \cdot \boldsymbol{v}|^2 = ||A\boldsymbol{v}||^2$$

With this observation, the *first singular vector* $\boldsymbol{v}_1$ of A is defined to be

$$\boldsymbol{v}_1 = \operatorname*{argmax}_{||\boldsymbol{v}||=1} ||A\boldsymbol{v}||$$

(Clearly the maximum exists because the given function is continuous on a compact domain). Also, note that this argmax need not be unique. The *first singular value* is defined as

$$\sigma_1(A) = ||A\boldsymbol{v}_1||$$

So it is clear that σ_1^2 is the sum of the squared lengths of the projections of the points onto the best fit line.

Now, inductively, successive singular values can be defined as follows: suppose singular vectors $\boldsymbol{v}_1, ..., \boldsymbol{v}_{i-1}$ are defined, and let the corresponding singular values be $\sigma_1, ..., \sigma_{i-1}$. We define the *ith singular vector* as follows.

$$\boldsymbol{v}_i = \operatorname*{argmax}_{\boldsymbol{v} \perp \langle \boldsymbol{v}_1, ..., \boldsymbol{v}_{i-1} \rangle, ||\boldsymbol{v}||=1} ||A\boldsymbol{v}||$$

So, the vector $\boldsymbol{v}_i$ is defined to be the minimizer of the same quantity as before, with the condition that $\boldsymbol{v}_i$ must lie in the space orthogonal to the first $i-1$ singular vectors. The corresponding *singular value* is defined as follows.

$$\sigma_i(A) = ||A\boldsymbol{v}_i||$$

The process stops at some $r \leq n$, where

$$\operatorname*{argmax}_{\boldsymbol{v} \in \langle \boldsymbol{v}_1, ..., \boldsymbol{v}_r \rangle, ||\boldsymbol{v}||=1} = 0$$

Theorem 5.1. *Let A be an $n \times d$ matrix with singular vectors $\boldsymbol{v}_1, ..., \boldsymbol{v}_r$. For $1 \leq k \leq r$, let V_k be the subspace spanned by $\boldsymbol{v}_1, ..., \boldsymbol{v}_k$. For each k, V_k is the best fit k-dimensional subspace for points in A.*

Proof. For a proof, look at **Theorem 3.1** of Kannan's book on *Foundations of Data Science.* This just shows that the greedy algorithm to find the best k-dimensional subspace works. ■

Proposition 5.2. *Let r be the integer where the process stops. Then,*

$$\sum_{i=1}^{n} ||\boldsymbol{a}_i||^2 = \sum_{i=1}^{r} \sigma_i^2(A) = ||A||_F$$

where $||\cdot||_F$ is the Frobenius norm of a matrix.

Proof. This is actually very easy to see; just look at the discussion in the book. ■

Definition 5.1. The vectors $\boldsymbol{v}_r, ..., \boldsymbol{v}_r$ constructed above are called the *right singular vectors* of A. By definition, these vectors are orthonormal. For each $1 \leq i \leq r$, the unit vector

$$\boldsymbol{u}_i = \frac{A\boldsymbol{v}_i}{\sigma_i(A)}$$

is called the ith *left singular vector* of A.

5.1.2. *Singular Value Decomposition.* As above, let A be an $n \times d$ matrix with right singular vectors $\boldsymbol{v}_1, ..., \boldsymbol{v}_r$ and singular values $\sigma_1, ..., \sigma_r$. Let the left singular vectors be $\boldsymbol{u}_1, ..., \boldsymbol{u}_r$. Observe that the matrix $\sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$ is a rank one matrix. The next theorem shows that the matrix A can be written as a sum of rank one matrices of the above form.

Theorem 5.3. *Let the notation be as above. Then,*

$$A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$$

Proof. The proof of this is rather straightforward. Note that two real matrices A and B of the same dimensions are equal if and only if $A\boldsymbol{v} = B\boldsymbol{v}$ for all vectors $\boldsymbol{v}$. We will use this fact. Let

$$M = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$$

We will show that for all i, $A\boldsymbol{v}_i = M\boldsymbol{v}_i$. We claim that this is enough to show that $A = M$. To see this, observe that the $\boldsymbol{v}_i$s are orthonormal; hence, any vector $\boldsymbol{v}$ can be written as a linear combination of $\boldsymbol{v}_i$s and some vector orthogonal to all the $\boldsymbol{v}_i$s, i.e some vector in the kernel of A. So, if $A\boldsymbol{v}_i = M\boldsymbol{v}_i$ for each i, it follows that $A\boldsymbol{v} = M\boldsymbol{v}$ for all vectors $\boldsymbol{v}$, and that will prove that $A = M$.

Showing that $A\boldsymbol{v}_i = M\boldsymbol{v}_i$ is straightforward.

$$M\boldsymbol{v}_i = \sum_{j=1}^{r} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^T \boldsymbol{v}_i = \sigma_i \boldsymbol{u}_i = A\boldsymbol{v}_i$$

and this proves the claim. ■

5.1.3. *SVD as a matrix product.* The SVD of a matrix A can also be written in a more convenient form, which we will now see.

Proposition 5.4. *Let r be any positive integer. Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_r$ and $\boldsymbol{y}_1, ..., \boldsymbol{y}_r$ be d-dimensional vectors. Let $c_1, ..., c_r$ be real numbers. Let X be the $d \times r$ matrix whose columns are $\boldsymbol{x}_i$s, Y be the $d \times r$ matrix whose columns are $\boldsymbol{y}_i$s, and let C be the $r \times r$ diagonal matrix which has $c_1, ..., c_r$ as it's diagonal entries. Then*

$$\sum_{i=1}^{r} c_i \boldsymbol{x}_i \boldsymbol{y}_i^T = XCY^T$$

Proof. The proof of this follows from the distributive law of matrix multiplication. Write X as a bunch of matrices with all but one non-zero columns. Write C as a sum of matrices with all but one diagonal non-zero. Similarly, write Y^T as a sum of matrices with all but one row non-zero. Then apply the distributive law. ■

Corollary 5.4.1. *If A is an $n \times d$ matrix, then the SVD of A is*

$$A = U\sigma V^T$$

where U is the matrix whose columns are the left singular vectors of A, σ is the diagonal matrix whose diagonals are the singular values of A, and V is the matrix whose columns are the right singular vectors of A.

Proof. This is immediate from the previous proposition and the SVD of a matrix. ■

5.2. **The Moore Penrose Pseudo Inverse.** Suppose A is an $n \times n$ square matrix. Suppose the SVD of A is

$$A = U\sigma V^T = \sum_{i=1}^{r} \sigma_i(A) \boldsymbol{u}_i \boldsymbol{v}_i^T$$

where as usual $\boldsymbol{u}_i$, $\boldsymbol{v}_i$ are the left and right singular vectors, and U, V have the same meaning as before. Note that the left and right singular vectors have the same dimensions, because A is a square matrix. Also, by definition, note that $\sigma_i(A) > 0$ for all $1 \leq i \leq r$ (after all, we ignore zero singular values).

Consider the matrix B defined by the following.

$$B = \sum_{i=1}^{r} \frac{1}{\sigma_i(A)} \boldsymbol{v}_i \boldsymbol{u}_i^T = V\sigma^+ U^T$$

where σ^+ is the $r \times r$ diagonal matrix whose diagonal entries are the inverses of the diagonal entries of σ. We will show that

$$BA\boldsymbol{x} = \boldsymbol{x}$$

for all $\boldsymbol{x}$ in the span of the right singular vectors of A. For this reason, the matrix B is called the *Moore-Penrose Pseudo Inverse* of A, as it acts as an inverse.

To see this, first observe the following.

$$\begin{aligned} BA &= \left(\sum_{i=1}^{r} \sigma_i(A) \boldsymbol{v}_i \boldsymbol{u}_i^T\right) \left(\sum_{j=1}^{r} \frac{1}{\sigma_j(A)} \boldsymbol{u}_i \boldsymbol{v}_i^T\right) \\ &= \sum_{i=1}^{r} \boldsymbol{v}_i \boldsymbol{v}_i^T \end{aligned}$$

The above equation is true simply because the left singular vectors are orthonormal, and so are the right singular vectors. Now, suppose $\boldsymbol{x}$ is in the span of $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_r\}$, i.e suppose

$$\boldsymbol{x} = a_1\boldsymbol{v}_1 + \cdots + a_r\boldsymbol{v}_r$$

Again, using the fact that the $\boldsymbol{v}_i$'s form an orthonormal system, we have the following.

$$\begin{aligned}\left(\sum_{i=1}^{r}\boldsymbol{v}_i\boldsymbol{v}_i^T\right)(a_1\boldsymbol{v}_1 + \cdots + a_r\boldsymbol{v}_r) &= \sum_{1\leq i,j\leq r} a_j\boldsymbol{v}_i\boldsymbol{v}_i^T\boldsymbol{v}_j \\ &= \sum_{i=1}^{r} a_j\boldsymbol{v}_j \\ &= \boldsymbol{x}\end{aligned}$$

and we've just shown that

$$BA\boldsymbol{x} = \boldsymbol{x}$$

and this proves the claim.

5.3. **Matrix Differentials.** Let $F : M_n(\mathbb{R}) \to M_n(\mathbb{R})$ be a differentiable map. To be completed.

5.4. **Fenchel Conjugates and Fenchel Duality.** In this section, we will see how *strong convexity* is really *smoothness* looked at from a different perspective. This phenomenon is called *Fenchel's Duality.*

Definition 5.2. Let f be a function defined on a suitable domain. Then the *Fenchel conjugate* f^* is defined as follows.

$$f^*(\boldsymbol{\theta}) = \max_{\boldsymbol{w}} \langle \boldsymbol{w}, \boldsymbol{\theta}\rangle - f(\boldsymbol{w})$$

The *Fenchel-Young inequality* immediately follows from this: for all $\boldsymbol{u}$,

$$f^*(\boldsymbol{\theta}) \geq \langle \boldsymbol{u}, \boldsymbol{\theta}\rangle - f(\boldsymbol{u})$$

Definition 5.3. A convex function f is said to be *closed* if for all $\alpha \in \mathbb{R}$, the *sublevel set* $\{\boldsymbol{x} \in \text{dom}f \mid f(\boldsymbol{x}) \leq \alpha\}$ is a closed set. Section to be completed.

5.5. **A simple fact about projection onto simplex.** Throughout this section, let Δ_n denote the n-simplex. First, observe that Δ_n is contained in a hyperplane; to see this, it is enough to observe that if $\boldsymbol{\theta} \in \Delta_n$, then we know that

$$\sum_{i=1}^{n}\theta_i = 1$$

So, $\boldsymbol{\theta}$ lies in the hyperplane described by the equation

$$\boldsymbol{w}^T\boldsymbol{x} - 1 = 0$$

where $\boldsymbol{w} = (1, 1, ..., 1)$. Also, this shows us that $\boldsymbol{w}$ is normal to the hyperplane containing Δ_n.

Now, consider the following definition.

$$S := \{\boldsymbol{\theta} + t\boldsymbol{w} \mid \boldsymbol{\theta} \in \Delta_n, t \in \mathbb{R}\}$$

Geometrically, S is the set obtained by *swiping* the set Δ_n along the axis parallel to $\boldsymbol{w}$; for example, in $\mathbf{R}^3$, the set S will be an infinite length prism, because Δ_3 is a triangle. It is also clear that S is a convex set.

We claim that projecting a point in S to Δ_n is easy; for any point $\boldsymbol{y}$ of the form $\boldsymbol{y} = \boldsymbol{\theta} + t\boldsymbol{w}$ where $\boldsymbol{\theta} \in \Delta_n$ and $t \in \mathbb{R}$, we have

$$\Pi_{\Delta_n}(\boldsymbol{y}) = \boldsymbol{\theta} \tag{5.1}$$

Geometrically, all we are doing is dropping a perpendicular from $\boldsymbol{y}$ onto the set Δ_n. To prove this, first note that if $\boldsymbol{\theta}' \in \Delta_n$, then the vector $\boldsymbol{\theta}' - \boldsymbol{\theta}$ is orthogonal to the vector $t\boldsymbol{w}$ (because $\boldsymbol{w}$ is a normal to the plane). So, this gives us the following for any $\boldsymbol{\theta}' \in \Delta_n$.

$$\begin{aligned} ||\boldsymbol{\theta}' - \boldsymbol{y}||^2 &= ||\boldsymbol{\theta}' - \boldsymbol{\theta} - t\boldsymbol{w}||^2 \\ &= ||\boldsymbol{\theta}' - \boldsymbol{\theta}||^2 + t^2 ||\boldsymbol{w}||^2 \end{aligned}$$

Above, we simply used the Pythagoras Theorem. So, it follows that $||\boldsymbol{\theta}' - \boldsymbol{y}||^2$ is minimised at the point $\boldsymbol{\theta}' = \boldsymbol{\theta}$, and this proves equation (5.1).

Now, let $\boldsymbol{y} \in \mathbf{R}^n$ be *any* point. We will now prove the following equation.

$$\Pi_{\Delta_n}(\boldsymbol{y}) = \Pi_{\Delta_n}(\Pi_S(\boldsymbol{y})) \tag{5.2}$$

In simple words, to find the projection of $\boldsymbol{y}$ onto Δ_n, it is enough to find the projection of $\boldsymbol{y}$ onto S, and then project the result onto Δ_n. Let us now prove this. Now, suppose $\Pi_S(\boldsymbol{y}) = \boldsymbol{\theta} + t\boldsymbol{w}$ for some $t \in \mathbb{R}$ and some $\boldsymbol{\theta} \in \Delta_n$. We claim that the vectors $\boldsymbol{w}$ and $\boldsymbol{\theta} + t\boldsymbol{w} - \boldsymbol{y}$ are orthogonal. To see this, define the following function on $\mathbb{R}$.

$$h(x) = ||\boldsymbol{\theta} + x\boldsymbol{w} - \boldsymbol{y}||^2$$

Clearly, h is a differentiable function on $\mathbb{R}$, and h attains a minimum at the point $x = t$. So, it follows that $h'(t) = 0$. But, note that

$$h'(t) = 2\langle \boldsymbol{w}, \boldsymbol{\theta} + t\boldsymbol{w} - \boldsymbol{y}\rangle = 0$$

and so it follows that the vectors $\boldsymbol{w}$ and $\boldsymbol{\theta} + \boldsymbol{tw} - \boldsymbol{y}$ are orthogonal.

Next, suppose $\boldsymbol{\theta}' \in \Delta_n$ is an arbitrary point. We claim that the vectors $\boldsymbol{tw}$ and $\boldsymbol{\theta}' + t\boldsymbol{w} - \boldsymbol{y}$ are orthogonal. To see this, note that

$$\begin{aligned} \langle t\boldsymbol{w}, \boldsymbol{\theta}' + t\boldsymbol{w} - \boldsymbol{y}\rangle &= \langle t\boldsymbol{w}, \boldsymbol{\theta}' - \boldsymbol{\theta} + \boldsymbol{\theta} + t\boldsymbol{w} - \boldsymbol{y}\rangle \\ &= \langle t\boldsymbol{w}, \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle + \langle t\boldsymbol{w}, \boldsymbol{\theta} + t\boldsymbol{w} - \boldsymbol{y}\rangle \\ &= 0 + 0 \end{aligned}$$

where the first quantity is 0 because $\boldsymbol{w}$ is a normal vector to the plane containing Δ_n, and the second quantity is zero which was just shown above.

So, we have the following, where we are just using the Pythagorean Theorem.

$$\begin{aligned} ||\boldsymbol{\theta}' - \boldsymbol{y}||^2 &= ||\boldsymbol{\theta}' + t\boldsymbol{w} - t\boldsymbol{w} - \boldsymbol{y}||^2 \\ &= t^2 ||\boldsymbol{w}||^2 + ||\boldsymbol{\theta}' + t\boldsymbol{w} - \boldsymbol{y}||^2 \end{aligned}$$

So, minimizing $||\boldsymbol{\theta}' - \boldsymbol{y}||^2$ is the same as minimizing $||\boldsymbol{\theta}' + t\boldsymbol{w} - \boldsymbol{y}||^2$. But, we know that this minimum is achieved for $\boldsymbol{\theta}' = \boldsymbol{\theta}$, because $\Pi_S(\boldsymbol{y}) = \boldsymbol{\theta} + t\boldsymbol{w}$. So, it follows that

$$\Pi_{\Delta_n}(\boldsymbol{y}) = \boldsymbol{\theta} = \Pi_{\Delta_n}(\Pi_S(\boldsymbol{y}))$$

and this proves equation (5.2).