Online Learning and Online Convex Optimization

August 9, 2021

Sadie Zhao, Denizalp Goktas, Amy Greenwald

In this set of notes, we provide a modern overview of online learning. We will give readers a sense of some of interesting ideas in online learning and underscore the centrality of convexity in deriving efficient online learning algorithms.

Contents

In this set of notes, we provide a modern overview of online learning. We will give readers a sense of some of interesting ideas in online learning and underscore the centrality of convexity in deriving efficient online learning algorithms.

1 Introduction

Online learning is the process of answering a sequence of questions given (maybe partial) knowledge of the correct answers to previous questions and possibly additional available information.

Algorithm 1 `Online Learning`

```
for t=1,2,... do
    receive question $\mathbf{x_t} \in \mathcal{X}$
    predict $p_t \in \mathcal{D}$
    receive true answer $y_t \in \mathcal{Y}$
    suffer loss $l(p_t, y_t)$
end for
```

The learner's ultimate goal is to minimize the cumulative loss suffered along its run.

Note that we make no assumptions regarding the origin of the sequence of examples. Thus, the sequence can be deterministic, stochastic, or even adversarially adaptive to the learner's own behavior. An adversary can make the cumulative loss arbitrarily large, so we need to restrict the problem. We consider two natural restrictions:

1. We assume that all the answers are generated by some target best mapping, $h^* : \mathcal{X} \to \mathcal{Y}$. Furthermore, h^* is taken from a fixed set, called a hypothesis class and denoted by $\mathcal{H}$. For an online learning algorithm, A, we denote by $M_A(\mathcal{H})$ the maximal number of mistakes A might make on a sequence of examples which is labeled by some $h^* \in \mathcal{H}$. A bound on $M_A(\mathcal{H})$ is called a *mistake-bound* and we will try to minimize $M_A(\mathcal{H})$.

2. We no longer assume that all answers are generated by some $h \in \mathcal{H}$. We define the *regret* of the algorithm relative to h when running on a sequence of T examples as:

$$\text{Regret}_T(h) = \sum_{t=1}^{T} l(p_t, y_t) - \sum_{t=1}^{T} l(h(x_t), y_t) \tag{1}$$

   and the regret of the algorithm relative to a hypothesis class $\mathcal{H}$ is

$$\text{Regret}_T(\mathcal{H}) = \max_{h \in \mathcal{H}} \text{Regret}_T(h) \tag{2}$$

   Now, the learner's new goal is to minimize regret relative to $\mathcal{H}$. We will sometimes be satisfied with "low regret" algorithms, by which we mean that $\text{Regret}_T(\mathcal{H})$ grows sub-linearly.

1.1 Examples

In this section, we will list some online prediction problems and their possible hypothesis classes.

Online Regression.

In regression problem, $\mathcal{X} = \mathbb{R}^d$ which corresponds to a set of d features, and $\mathcal{Y} = \mathcal{D} = \mathbb{R}$. Common loss functions are the squared loss, $l(p, y) = (p - y)^2$, and the absolute loss, $l(p, y) = |p - y|$. The simplest hypothesis class for regression is the class of linear predictors, $\mathcal{H} = \{\mathbf{x} \mapsto \sum_{i=1}^{d} w[i]x[i] : \forall i, w[i] \in \mathbb{R}\}$. This is the hypothesis class for *online linear regression problem*.

Prediction with Expert Advice

On each round t, the learner has to choose from the advice of d given expert. Therefore, $\mathbf{x_t} \in \mathcal{X} \subset \mathbb{R}^d$, where $x_t[i]$ is the advice of the ith expert, and $p_t \in \mathcal{D} = \{1, 2, ..., d\}$. The true answer is a vector $y_t \in \mathcal{Y} = [0, 1]^d$, where $y_t[i]$ is the cost of following the advice of ith expert. Thus, $l(p_t, y_t) = y_t[p_t]$. A common hypothesis class is the set of constant predictors, $\mathcal{H} = \{h_1, ..., h_d\}$, where $h_i(\mathbf{x}) = i$ for all $\mathbf{x} \in \mathcal{X}$.

1.2 A Gentle Start

In this section, we will start with discussing online classification problem. In this problem, $\mathcal{Y} = \mathcal{D} = \{0, 1\}$, and $l(p, y) = |p - y|$. Moreover, we assume that we have a finite hypothesis class, that is, $|\mathcal{H}| < \infty$.

Recall that the regret relative to the hypothesis class is defined as

$$\text{Regret}_T(\mathcal{H}) = \max_{h \in \mathcal{H}} \left(\sum_{t=1}^{T} |p_t - y_t| - \sum_{t=1}^{T} |h(x_t) - y_t| \right) \tag{3}$$

as $l(p, y) = |p - y|$ in this case.

We have Cover's impossibility results: no algorithm can obtain a sublinear regret bound even if $|\mathcal{H}| = 2$. We can sidestep this result by further restricting the power of the adversarial environment as we described above.

1.2.1 Realizability Assumption

We make one additional assumption: assume that all target labels are generated by some perfect $h^* \in \mathcal{H}$, namely, $y_t = h^*(\mathbf{x}_t)\ \forall 1 \leq t \leq T$. Our goal is to minimize the mistake bound, $M_A(\mathcal{H})$.

Our intuition is to design an algorithm that at any round, use any hypothesis which is consistent with all past examples.

Algorithm 2 `Consistent`(Online Classification)

```
input: A finite hypothesis class $\mathcal{H}$
Initialize: $V_1 = \mathcal{H}$
for t=1,2,... do
    receive $\mathbf{x}_t$
    choose any $h \in V_t$
    predict $p_t = h(\mathbf{x}_t)$
    receive true answer $y_t = h^*(\mathbf{x}_t)$
    Update $V_{t+1} = \{h \in V_t : h(\mathbf{x}_t) = y(t)\}$
end for
```

Obviously, whenever `Consistent` makes a prediction mistake, at least one hypothesis is removed from V_t. Therefore, after making M mistakes we have $|V_t| \leq |\mathcal{H}| - M$. Since by the realizability assumption, $h^* \in \mathcal{H}$ (so V_t is always nonempty), we have $1 \leq |V_t| \leq |\mathcal{H}| - M$. That is,

Theorem 1.1. *Let $\mathcal{H}$ be a finite hypothesis class. The `Consistent` algorithm enjoys the mistake bound*

$$M_{\texttt{Consistent}}(\mathcal{H}) \leq |\mathcal{H}| - 1$$

Then, we present a better algorithm which is guaranteed to make exponentially fewer mistakes. The idea is to predict according to the majority of hypothesis in V_t rather than according to some arbitrary $h \in V_t$.

Algorithm 3 `Halving`(Online Classification)

```
input: A finite hypothesis class $\mathcal{H}$
Initialize: $V_1 = \mathcal{H}$
for t=1,2,... do
    receive $\mathbf{x}_t$
    predict $p_t = \arg\max_{r \in \{0,1\}} |\{h \in V_t : h(\mathbf{x}_t) = r\}|$    (in case of a tie, predict $p_t = 1$)
    receive true answer $y_t$
    Update $V_{t+1} = \{h \in V_t : h(\mathbf{x}_t) = y(t)\}$
end for
```

In this way, whenever we make mistake, we are guaranteed to remove at least half of the hypotheses from the version space V_t. That is, whenever the algorithm errs, we have $|V_{t+1}| \leq |V_t|/2$. Therefore, if M is the total number of mistakes, we have

$$1 \leq |V_{t+1}| \leq |\mathcal{H}| 2^{-M}.$$

Rearranging the above inequality, we can conclude that:

Theorem 1.2. *Let $\mathcal{H}$ be a finite hypothesis class. The `Halving` algorithm enjoys the mistake bound*

$$M_{\texttt{Halving}}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$$

1.2.2 Randomization

Instead of having the predictions domain being $D = \{0, 1\}$, we allow it to be $D = [0, 1]$, and interpret $p_t \in D$ as the probability to predict the label 1 on round t. The loss function is still $l(p_t, y_t) = |p_t - y_t|$.

With this assumption, it is possible to derive a low regret algorithm as stated in the following theorem.

Theorem 1.3. *Let $\mathcal{H}$ be a finite hypothesis class. There exists an algorithm for online classification, whose predictions come from $D = [0, 1]$, and enjoys the regret bound*

$$\sum_{t=1}^{T} |p_t - y_t| - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} |h(x_t) - y_t| \leq \sqrt{0.5 \ln(|\mathcal{H}|) T}$$

We will provide a constructive proof of this theorem in the next section.

1.3 Notation and Basic Definitions

We denote **scalars** with lower case letters (e.g., x), and **vectors** with bold ace letters (e.g., $\mathbf{x}$). $\mathbf{x}[i]$ denotes the ith element of vector $\mathbf{x}$. Since online learning is performed in a sequence of rounds, we denote by $\mathbf{x}_t$ the tth vector in a sequence of vectors $\mathbf{x}_1$,$\mathbf{x}_2$,...,$\mathbf{x}_T$.

The **inner product** between vectors $\mathbf{x}$ and $\mathbf{w}$ is denoted by $\langle \mathbf{x}, \mathbf{w}\rangle$.Whenever we do not specify the vector space, we assume that it is the d-dimensional Euclidean space and then $\langle \mathbf{x}, \mathbf{w}\rangle = \sum_{i=1}^{d} x[i]w[i]$. The Euclidean (or ℓ_2) **norm** of a vector $\mathbf{w}$ is $\|\mathbf{w}\|_2 = \sqrt{\langle \mathbf{w}, \mathbf{w}\rangle}$. We also use other ℓ_p norms, $\|\mathbf{w}\|_p = (\sum_i |w[i]|^p)^{1/p}$. In particular, $\|\mathbf{w}\|_1 = \sum_i |w[i]|$ and $\|\mathbf{w}\|_\infty = \max_i |w[i]|$. A generic norm of a vector $\mathbf{w}$ is denoted by $\|\mathbf{w}\|$ and its **dual norm** is defined as

$$\|\mathbf{x}\|_* = \max\{\langle \mathbf{w}, \mathbf{x}\rangle : \|\mathbf{w}\| \leq 1\}.$$

The definition of the dual norm immediately implies the inequality

$$\langle \mathbf{w}, \mathbf{z}\rangle \leq \|\mathbf{w}\| \, \|\mathbf{z}\|_* \,. \tag{4}$$

Given a predicate π, we use the notation $\mathbf{1}_{[\pi]}$to denote the **indicator function** that outputs 1 if π holds and 0 if otherwise.

A function f is called L-**Lipschitz** over a set S with respect to a norm $\|\cdot\|$ if for all $\mathbf{u}$,$\mathbf{w} \in S$, we have

$$|f(\mathbf{u}) - f(\mathbf{w})| \leq L \, \|\mathbf{u} - \mathbf{w}\| \,.$$

The **gradient** of a differentiable function f is denoted by ∇f and the **Hessian** is denoted by $\nabla^2 f$.

A set S is **convex** if for all $\mathbf{w}, \mathbf{v} \in S$ and $\alpha \in [0, 1]$, we have that $\alpha\mathbf{w} + (1 - \alpha)\mathbf{v} \in S$ as well. Similarly, a function $f : S \to \mathbb{R}$ is convex if for all $\mathbf{w}, \mathbf{v}$ and $\alpha \in [0, 1]$, we have $f(\alpha\mathbf{w} + (1 - \alpha)\mathbf{v}) \leq \alpha f(\mathbf{w}) + (1 - \alpha)f(\mathbf{v})$.

2 Online Convex Optimization

Algorithm 4 Online Convex Optimization (OCO)

```
input: A convex set $S$
for t=1,2,... do
    predict a vector $\mathbf{w}_t \in S$
    receive a convex loss function $f_t : S \to \mathbb{R}$
    suffer loss $f_t(\mathbf{w}_t)$
end for
```

In this section, we describe algorithms for online convex optimization and analyze their regret. Recall that the regret of an online algorithm with respect a computing hypothesis (which here will be come vector $\mathbf{u}$) is defined as

$$\text{Regret}_T(\mathbf{u}) = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \sum_{t=1}^{T} f_t(\mathbf{u})$$

Then, the regret of the algorithm relative to a set of competing vectors, U, is defined as

$$\text{Regret}_T(U) = \max_{\mathbf{u}\in U} \text{Regret}_T(\mathbf{u})$$

Remark 2.1. *Note that the predictions of the learner should come from the set S, while we analyze the regret respect to the set U. While in some situations it makes sense to set $U = S$, this is not always true. If we do not specify the value of U, we use the default value $U = S$, and our default setting for S is $S = \mathbb{R}^d$.*

2.1 Convexification

Some online prediction problems can be seamlessly cast in the online convex optimization framework.

Example 2.2 (Online linear regression)**.** *Recall the online regression problem described in Section 1.1. On each online round, the learner first receives a vector of features, $\mathbf{x}_t \in A \subset \mathbf{R}^d$, and then predict a scalar, p_t. Next, the learner receive the true answer $y_t \in \mathbf{R}$, and suffer the loss $l(p_t, y_t) = |p_t - y_t|$. The learner should be competitive with the set of linear predictors of the from $\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$. Assume that the predictions of the learner are also based on linear functions, then we can easily cast this online prediction problem in the online convex optimization framework as follows:*

The learner should decide on a vector $\mathbf{w}_t$ which yields the prediction $p_t = \langle \mathbf{w}_t, \mathbf{x}_t \rangle$. The loss function becomes $|p_t - y_t| = |\langle \mathbf{w}_t, \mathbf{x}_t \rangle - y_t|$. Therefore, consider $f_t(\mathbf{w}) = |\langle \mathbf{w}, \mathbf{x}_t \rangle - y_t|$, which is indeed a convex function, we obtain that $f_t(\mathbf{w}_t) = l(p_t, y_t)$.

Note that other online prediction problems do not seem to fit into the online convex optimization framework. For example, in the online classification problem, the predictions domain S or the loss functions are not convex. We will then describe two techniques that allow us to utilize the online convex optimization framework in additional scenarios.

2.1.1 Convexification by Randomization

Consider the problem of prediction with expert advice, where on each online round, the learner has to choose from the advice of d given experts. Denote by $p_t \in \{1, ..., d\}$ the chosen expert. Then, the learner receives a vector $\mathbf{y}_t \in [0,1]^d$, where $y_t[i]$ is the cost of following the advice of the ith expert. Finally, the learner suffers the loss $l(p_t) = y_t[p_t]$.

By allowing the learner to randomize his predictions, we can cast the problem in the online convex optimization framework, and therefore can obtain low regret algorithm for this problem

Formally, let $S = \{\mathbf{w} \in \mathbb{R}_+^d : \|\mathbf{w}\|_1 = 1\}$ be the probability simplex, which forms a convex set. At round t, the learner chooses $\mathbf{w}_t \in S$ and based on $\mathbf{w}_t$ picks an expert at random according to $\mathbb{P}[P_t = i] = w_t[i]$ where P_t a random categorical variable, i.e., $P_t \sim \mathrm{Cat}(d, \mathbf{w}_t)$. Then, the cost vector $\mathbf{y}_t$ is revealed and the learner pays for his expected cost

$$\mathbb{E}[l(P_t))] = \mathbb{E}[y_t[P_t])] = \sum_{i=1}^{d} \mathbb{P}[p_t = i] y_t[i] = \langle \mathbf{w}_t, \mathbf{y}_t \rangle$$

Now we can cast the problem as online convex optimization since S is a convex set and the loss function $f_t(\mathbf{w}) = \langle \mathbf{w}, \mathbf{y}_t \rangle$ is a linear function (hence convex).

2.1.2 Convexification by Surrogate Loss function

We again start with the specific problem of online classification with a finite hypothesis class. Recall the realizability assumption we used to sideup Cover's impossibility result. That is, we assumed that there exists a perfect $h^* \in \mathcal{H}$ such that $y_t = h^*(\mathbf{x}_t)$ for all t. With this assumption, we described the `Halving` algorithm with at most $\log_2(|\mathcal{H}|)$ prediction mistakes.

We now try derive a similar guarantee using the language of online convex optimization. Let $\mathcal{H} = \{h_1, ..., h_d\}$ and let $S = \{\mathbf{w} \in \mathbb{R}_+^d : ||w||_1 = 1\}$ be the probability simplex. For each online round, define $\mathbf{v}_t = (h_1(\mathbf{x}_t), ..., h_d(\mathbf{x}_t)) \in \{0,1\}^d$. Our algorithm will maintain $\mathbf{w}_t \in S$ and will predict the label according to

$$p_t = \begin{cases} 1 & \text{if } \langle \mathbf{w}_t, \mathbf{v}_t \rangle \geq 1/2 \\ 0 & \text{if } \langle \mathbf{w}_t, \mathbf{v}_t \rangle < 1/2 \end{cases}$$

Let $\mathcal{M} = \{t : p_t \neq y_t\}$ be the rounds on which our algorithm makes a prediction mistake. We define the loss function

$$\boldsymbol{f}_t(\mathbf{w}) = \begin{cases} 2|\langle \mathbf{w}_t, \mathbf{v}_t \rangle - y_t| & \text{if } t \in \mathcal{M} \\ 0 & \text{if } t \notin \mathcal{M} \end{cases}$$

f_t has two key properties:

- f_t is a convex function
- $f_t(\mathbf{w}_t) \geq |p_t - y_t|$, namely, the convex loss upper bounds the original non-convex loss. (Note that when $t \in \mathcal{M}$, $f_t(\mathbf{w}) \geq 1$.)

Hence, we name it **surrogate convex loss**. Since S is a convex set and f_t is a convex function for all t, we have obtained an online convex optimization problem.

In the next section, we will derive algorithms for online convex optimization problems. In particular, one of these algorithms enjoys the regret bound

$$\forall \mathbf{u} \in S, \ \ \sum_{t=1}^{T} f_t(\mathbf{w}_t) \le \sum_{t=1}^{T} f_t(\mathbf{u}) + \frac{\log(d)}{\eta} + 2\eta \sum_{t=1}^{T} L_t$$

where η is a parameter, which we will set here to be $\eta = 1/4$, and L_t is a Lipschitz parameter of the function f_t. In our case, $L_t = 1$ (consider f_t function for both $p_t = 0$ and $p_t = 1$), if $t \in \mathcal{M}$ and $L_t = 0$ if $t \notin \mathcal{M}$. Hence,

$$\forall \mathbf{u} \in S, \ \ \sum_{t=1}^{T} f_t(\mathbf{w}_t) \le \sum_{t=1}^{T} f_t(\mathbf{u}) + 4\log(d) + \frac{1}{2}|\mathcal{M}|$$

By the surrogate property of f_t, we can lower bound the left hand side by $|\mathcal{M}|$. Rearranging, we obtain:

$$|\mathcal{M}| \le 2\sum_{t=1}^{T} f_t(\mathbf{u}) + 8\log(d)$$

(For all $t \in \mathcal{M}$, $f_t(\mathbf{w}_t) \ge 1$, so $\sum_{t=1}^{T} f_t(\mathbf{w}_t) \ge |\mathcal{M}|$.)

This type of bound, where the number of mistakes is upper bounded by the convex surrogate loss of a competing hypothesis, is often called a **relative loss bound**.

In the realizable case, there exists a true (perfect) hypothesis h^*. Then, consider the vector $\mathbf{u} = (0, ..., 0, 1, 0, ..., 0) \in S$, where the 1 is placed in the coordinate corresponding to the true hypothesis h^*. Then, by our construction, $f_t(\mathbf{u}) = 0$ for all t, which yields

$$|\mathcal{M}| \le 8\log(d)$$

Remark 2.3. *Here is a general process of this surrogate loss function technique:*

1. *Reparameterize of the problem such that the decision space becomes convex (instead of maintaining the set V_t in `Halving`, we now maintain the vector $\mathbf{w}_t \in S$).*

2. *Construct a function f_t of the predicted parameter that satisfied two requirements: it is convex and it should upper bound the original loss function.*

3. *Construct a convex surrogate for which there exists some $\mathbf{u} \in S$ that attains a low cumulative loss. Otherwise, the resulting bound will be meaningless. Typically, this is done by assuming more on the problem. For example, in the above the realizability assumption enable us to construct a surrogate for which there was $\mathbf{u} \in S$ such that $f_t(\mathbf{u}) = 0$ for all t.*

2.2 Follow-the-leader

In this section, we try to derive some algorithms for online convex optimization.

The most natural learning rule is, at any online round, using the vector which has minimal loss on all past rounds. This is the same spirit of the `Consistent` algorithm, and in the context of online convex optimization, it is usually referred to as `Follow-The-Leader`.

Algorithm 5 `Follow-the-leader (FTL)`

```
input: A convex set $S$
for t=1,2,... do
    $$\mathbf{w}_t \in \arg\min_{\mathbf{w}\in S} \sum_{i=1}^{t-1} f_i(\mathbf{w})$$
end for
```

To analyze FTL, we will first show that the regret of FTL is upper bounded by the cumulative difference between the loss of $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$.

Lemma 2.4. *Let $\mathbf{w}_1$, $\mathbf{w}_2$,... be the sequence of vectors produced by FLT. Then, for all $\mathbf{u} \in S$, we have*

$$\mathit{Regret}_T(\mathbf{u}) = \sum_{t=1}^{T}(f_t(\mathbf{w}_t) - f_t(\mathbf{u})) \leq \sum_{t=1}^{T}(f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}))$$

Proof. Subtracting $\sum_t f_t(\mathbf{w}_t)$ from both sides and rearranging, the desired inequality can be rewritten as

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) \leq \sum_{t=1}^{T} f_t(\mathbf{u}) \;\; \forall \mathbf{u} \in S.$$

We want to prove this inequality by induction. The base case is $T = 1$ follows directly from the definition of $\mathbf{w}_{t+1}$. Assume the inequality holds for $T - 1$, then for all $\mathbf{u} \in S$ we have

$$\sum_{t=1}^{T-1} f_t(\mathbf{w}_{t+1}) \leq \sum_{t=1}^{T-1} f_t(\mathbf{u}).$$

Adding $f_T(\mathbf{w}_{T+1})$ to both sides, we get

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) \leq f_T(\mathbf{w}_{T+1}) + \sum_{t=1}^{T-1} f_t(\mathbf{u}).$$

Since this holds for all $\mathbf{u} \in S$, we can pick $\mathbf{u} = \mathbf{w}_{T+1}$. Thus,

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) \leq \sum_{t=1}^{T} f_t(\mathbf{w}_{T+1}).$$

By definition of $\mathbf{w}_{T+1} \in \arg\min_{\mathbf{w}\in S} \sum_{t=1}^{T} f_t(\mathbf{w})$, we know that for all $\mathbf{u} \in S$,

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{T+1}) \leq \sum_{t=1}^{T} f_t(\mathbf{u}).$$

Therefore,

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) \leq \sum_{t=1}^{T} f_t(\mathbf{u}) \quad \forall \mathbf{u} \in S$$

□

Definition 2.5 (Online Quadratic Optimization)**.** *This is an online convex optimization problem where at each round* $f_t(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{w} - \mathbf{z}_t\right\|_2^2$ *for some vector* $\mathbf{z}_t$*.*

Next, we can use lemma 2.4 to derive a regret bound for the following sub-family of online convex optimization.

Corollary 2.6. *Consider running FTL on an Online Quadratic Optimization problem with* $S = \mathbb{R}^d$ *and let* $L = \max_t \left\|\mathbf{z}_t\right\|_2$*. Then, the regret of FTL with respect to all vectors* $\mathbf{u} \in \mathbb{R}^d$ *is at most* $4L^2(\log(T) + 1)$*.*

Proof. We further assume that $S = \mathbb{R}^d$. For this case, we can verify that the FTL becomes

$$\mathbf{w}_t \in \underset{\mathbf{w}\in S}{\arg\min} \sum_{i=1}^{t-1} f_i(\mathbf{w}) = \underset{\mathbf{w}\in S}{\arg\min} \sum_{i=1}^{t-1} \frac{1}{2}\left\|\mathbf{w} - \mathbf{z}_i\right\|_2^2 = \frac{1}{t-1}\sum_{i=1}^{t-1} \mathbf{z}_i,$$

namely, $\mathbf{w}_t$ is the average of $\mathbf{z}_1, \cdots, \mathbf{z}_{t-1}$.
($\mathbf{w}_t \in \arg\min_{\mathbf{w}\in S} \sum_{i=1}^{t-1} \frac{1}{2}\left\|\mathbf{w} - \mathbf{z}_i\right\|_2^2$ implies $\sum_{i=1}^{t-1} \mathbf{w}_t - \mathbf{z}_i = 0$ by taking derivative, so $(t-1)\mathbf{w}_t = \sum_{i=1}^{t-1} \mathbf{z}_i$, and finally $\mathbf{w}_t = \frac{1}{t-1}\sum_{i=1}^{t-1} \mathbf{z}_i$.)

Thus,

$$\mathbf{w}_{t+1} = \left(\frac{t-1}{t}\right)\mathbf{w}_t + \left(\frac{1}{t}\right)\mathbf{z}_t = \left(1 - \frac{1}{t}\right)\mathbf{w}_t + \left(\frac{1}{t}\right)\mathbf{z}_t$$

which yields

$$\mathbf{w}_{t+1} - \mathbf{z}_t = \left(1 - \frac{1}{t}\right)\mathbf{w}_t + \left(\frac{1}{t} - 1\right)\mathbf{z}_t = \left(1 - \frac{1}{t}\right)(\mathbf{w}_t - \mathbf{z}_t).$$

Therefore,

$$\begin{aligned}
f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) &= \frac{1}{2}\left\|\mathbf{w}_t - \mathbf{z}_t\right\|^2 - \frac{1}{2}\left\|\mathbf{w}_{t+1} - \mathbf{z}_t\right\|^2 \\
&= \frac{1}{2}\left\|\mathbf{w}_t - \mathbf{z}_t\right\|^2 - \frac{1}{2}\left\|\left(1 - \frac{1}{t}\right)(\mathbf{w}_t - \mathbf{z}_t)\right\|^2 \\
&= \frac{1}{2}\left(1 - \left(1 - \frac{1}{t}\right)^2\right)\left\|\mathbf{w}_t - \mathbf{z}_t\right\|^2 \\
&= \frac{1}{2}\left(\frac{2}{t} - \frac{1}{t^2}\right)\left\|\mathbf{w}_t - \mathbf{z}_t\right\|^2 \\
&= \frac{1}{t}\left(1 - \frac{1}{2t}\right)\left\|\mathbf{w}_t - \mathbf{z}_t\right\|^2 \\
&\leq \frac{1}{t}\left\|\mathbf{w}_t - \mathbf{z}_t\right\|^2.
\end{aligned}$$

Let $L = \max_t \left\|\mathbf{z}_t\right\|$. Since $\mathbf{w}_t$ is the average of $\mathbf{z}_1, \cdots, \mathbf{z}_{t-1}$, we have that $\left\|\mathbf{w}_t\right\| \leq L$ and therefore by triangle inequality, $\left\|\mathbf{w}_t - \mathbf{z}_t\right\| \leq 2L$. We have therefore obtained:

$$\sum_{t=1}^{T}(f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})) \leq (2L)^2 \sum_{t=1}^{T} \frac{1}{t}.$$

Combining the above with lemma 2.4 and using the inequality $\sum_{t=1}^{T} \frac{1}{t} \leq \log(T) + 1$ we can conclude that the regret of FTL with respect to all vectors $\mathbf{u} \in \mathbb{R}^d$ is at most $4L^2(\log(T) + 1)$. □

Definition 2.7 (Online Linear Optimization)**.** *This is an online convex optimization problem where at each round* $f_t(\mathbf{w}) = \langle \mathbf{w}, \mathbf{z}_t \rangle$ *for some vector* $\mathbf{z}_t$*.*

While the above results about Online Quadratic Optimization problem seems promising, we will next show that the FTL rule does not guarantee low regret for another important sub-family.

Example 2.8 (Failure of FTL)**.** *Let* $S = [-1, 1] \subset \mathbb{R}$ *and consider the sequence of linear functions such that* $f_t(w) = z_t w$ *where*

$$z_t = \begin{cases} -0.5 & \textit{if } t = 1 \\ 1 & \textit{if } t \textit{ is even} \\ -1 & \textit{if } t > 1 \textit{ and } t \textit{ is odd} \end{cases}$$

Then, if t *is odd,*

$$w_t \in \operatorname*{arg\,min}_{w \in S} \sum_{i=1}^{t-1} f_i(w) = \operatorname*{arg\,min}_{w \in S} \sum_{i=1}^{t-1} z_i w = \operatorname*{arg\,min}_{w \in S} 0.5w = -1,$$

Similarly, if t *is even,*

$$w_t \in \operatorname*{arg\,min}_{w \in S} \sum_{i=1}^{t-1} f_i(w) = \operatorname*{arg\,min}_{w \in S} \sum_{i=1}^{t-1} z_i w = \operatorname*{arg\,min}_{w \in S} -0.5w = 1.$$

The cumulative loss of the FTL algorithm will therefore be $\sum_{t=1}^{T} f_t(w_t) = \sum_{t=1}^{T} z_t w_t = T$*. Moreover, the cumulative loss of the fixed solution* $u = 0 \in S = 0$ *(let* $U = S$ *here), so Regret*$_T(u) = T$*. Thus, the regret of FTL is at least T.*

Intuitively, FTL fails in the Example 2.8 because its predictions are *not stable*–w_t shifts drastically from round to round. In contrast, FTL works fine for the quadratic game since $\mathbf{w}_{t+1}$ is "close" to $\mathbf{w}_t$. One way to stabilize FTL is by adding regularization, which is the topic of next section.

2.3 Follow-the-Regularized-Leader

Follow-the-Regularized-Leader is a natural modification of the basic FTL algorithm in which we minimize the loss on all past rounds plus a regularization term. The goal of the regularization term is to stabilize the solution. Formally, for a regularization function $R : S \to \mathbb{R}$ we define

Algorithm 6 `Follow-the-Regularized-leader(FoReL)`

```
input: A convex set $S$
for t=1,2,... do
    $\mathbf{w}_t \in \operatorname*{arg\,min}_{\mathbf{w} \in S} \left( \sum_{i=1}^{t-1} f_i(\mathbf{w}) + R(\mathbf{w}) \right)$
end for
```

Naturally, different regularization functions will yield different algorithms with different regret bounds. But, first, let us specify FoReL for the case of linear functions and squared-ℓ_2-norm regularization, which we often call the Euclidean regularization case.

Example 2.9. *Consider again the Online linear Optimization problem where* $f_t(\mathbf{w}) = \langle \mathbf{w}, \mathbf{z}_t \rangle$ *and let* $S = \mathbb{R}^d$*. Suppose we run FoReL with the regularization function* $R(\mathbf{w}) = \frac{1}{2\eta} \|\mathbf{w}\|_2^2$ *for some positive scalar* η*. Then, it is easy to verify that*

$$\mathbf{w}_{t+1} = -\eta \sum_{i=1}^{t} \mathbf{z}_i = \mathbf{w}_t - \eta \mathbf{z}_t \tag{5}$$

Proof.

$$\begin{aligned}
\mathbf{w}_t &= \arg\min_{\mathbf{w} \in S} \sum_{i=1}^{t-1} f_i(\mathbf{w}) + R(\mathbf{w}) \\
&= \arg\min_{\mathbf{w} \in S} \sum_{i=1}^{t-1} \langle \mathbf{w}, \mathbf{z}_i \rangle + \frac{1}{2\eta} \|\mathbf{w}\|_2^2 \text{, which implies} \sum_{i=1}^{t-1} z_i + \frac{1}{\eta}\mathbf{w}_t = 0, \\
\text{so } \mathbf{w}_t &= -\eta \sum_{i=1}^{t-1} \mathbf{z}_i, \mathbf{w}_{t+1} = -\eta \sum_{i=1}^{t} \mathbf{z}_i = \mathbf{w}_t - \eta \mathbf{z}_t
\end{aligned}$$

□

Note that $\mathbf{z}_t$ *is the gradient of* f_t *at* $\mathbf{w}_t$ *(in fact, at any point). Therefore, the recursive rule,* $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{z}_t$ *can be rewritten as* $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla f_t(\mathbf{w}_t)$*. Hence, this rule is often called the Online Gradient Descent.*

Then, we turn to analysis of FoReL. As with the analysis of FTL, we first relate the regret of FoReL with the cumulative difference between the loss of $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$.

Lemma 2.10. *Let* $\mathbf{w}_1$,$\mathbf{w}_2$,*... be the sequence of vectors produced by FoReL. Then, for all* $\mathbf{u} \in S$ *we have*

$$\sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{u})) \leq R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})).$$

Proof. Observing that running FoReL on $f_1, ..., f_T$ is equivalent to running FTL on $f_0, f_1, ..., f_T$ where $f_0 = R$. Using Lemma 2.4, we obtain that

$$\sum_{t=0}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{u})) \leq \sum_{t=0}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})).$$

That is,

$$\sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{u})) + R(\mathbf{w}_0) - R(\mathbf{u}) \leq \sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})) + R(\mathbf{w}_0) - R(\mathbf{w}_1).$$

Therefore, by rearranging this, we can conclude that

$$\sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{u})) \leq R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})).$$

□

Based on the above lemma, we can easily derive a regret bound for online linear optimization with the regularizer $R(\mathbf{w}) = \frac{1}{2\eta} \|\mathbf{w}\|_2^2$.

Theorem 2.11. *Consider running FoReL on a sequence of linear functions, $f_t(\mathbf{w}) = \langle \mathbf{w}, \mathbf{z}_t \rangle$ for all t, with $S = \mathbb{R}$, and with the regularizer $R(\mathbf{w}) = \frac{1}{2\eta} \|\mathbf{w}\|_2^2$, which yields the predictions given in Equation (5). Then, for all $\mathbf{u}$ we have*

$$\mathit{Regret}_T(\mathbf{u}) \leq \frac{1}{2\eta} \|\mathbf{u}\|_2^2 + \eta \sum_{t=1}^{T} \|\mathbf{z}_t\|_2^2 .$$

In particular, consider the set $U = \{\mathbf{u} : \|\mathbf{u}\| \leq B\}$ and let L be such that $\frac{1}{T} \sum_{t=1}^{T} \|\mathbf{z}_t\|_2^2 \leq L^2$, then by setting $\eta = \frac{B}{L\sqrt{2T}}$ we obtain

$$\mathit{Regret}_T(\mathbf{u}) \leq BL\sqrt{2T}.$$

Proof. Using Lemma 2.10 and Equation (5),

$$\begin{aligned}
\mathrm{Regret}_T(\mathbf{u}) &\leq R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})) && \text{by lemma 2.10} \\
&= \frac{1}{2\eta} \|\mathbf{u}\|_2^2 + \sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{w}_{t+1}, \mathbf{z}_t \rangle && \text{Note that } \mathbf{w}_1 = 0 \\
&= \frac{1}{2\eta} \|\mathbf{u}\|_2^2 + \sum_{t=1}^{T} \langle \eta \mathbf{z}_t, \mathbf{z}_t \rangle && \text{as } \mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{z}_t \text{ by Equation (5)} \\
&= \frac{1}{2\eta} \|\mathbf{u}\|_2^2 + \eta \sum_{t=1}^{T} \|\mathbf{z}_t\|_2^2
\end{aligned}$$

□

Remark 2.12. *The parameter η in the above theorem depends on the time horizon T. It is possible to derive a similar result without using the time horizon. In the next section, we show a generic way to do so.*

We see that the Euclidean regularization function guarantees low regret for linear functions with bounded gradient using the ℓ_2-norm, because it stabilizes the predictions. We shall later generalize the above results in two aspects:

- *First, we allow any sequence of Lipschitz functions rather than linear functions with bounded gradient.*
- *Second, we consider other regularization functions which guarantee stability in other scenarios.*

2.3.1 The Doubling Trick

Sadie: add own explanation

Consider an algorithm that enjoys a regret bound of the form $\alpha\sqrt{T}$, but the parameters require the knowledge of T. The doubling trick, described below, enables us to convert such an algorithm into an algorithm that does not need to know the time horizon. The idea is to divide the time into periods of increasing size and run the original algorithm on each period. The regret of A on each period of 2^m rounds is at most $\alpha\sqrt{2^m}$. Therefore, the total regret is at most

Algorithm 7 The Doubling Trick

```
input: algorithm $A$ whose parameters depend on the time horizon $T$
for $m = 0, 1, 2, \ldots$ do
    run $A$ on the $2^m$ rounds: $t = 2^m, ..., 2^{m+1} - 1$
end for
```

$$\begin{aligned}
\sum_{m=1}^{\lceil \log_2(T) \rceil} \alpha\sqrt{2^m} &= \alpha \sum_{m=1}^{\lceil \log_2(T) \rceil} \left(\sqrt{2}\right)^m \\
&= \alpha \left(\frac{1 - (\sqrt{2})^{\lceil \log_2(T) \rceil + 1}}{1 - \sqrt{2}} \right) \\
&\leq \alpha \left(\frac{1 - \sqrt{2T}}{1 - \sqrt{2}} \right) \\
&\leq \frac{\sqrt{2}}{\sqrt{2} - 1} \left(\alpha\sqrt{T} \right)
\end{aligned}$$

That is, we obtain that the regret is worse by a constant multiplicative factor.

2.4 Online Gradient Descent: Linearization of Convex Functions

In the previous section we introduced the FoReL approach and analyzed it for the case of linear functions, $S = \mathbb{R}^d$, and Euclidean regularization. We now generalize this results by deriving a simple reduction from convex functions to linear function.

Definition 2.13. *Let S be a convex set. A function $f : S \to \mathbb{R}$ is convex iff for all $\mathbf{w} \in S$, there exists $\mathbf{z}$ such that*

$$\forall \mathbf{u} \in S, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \mathbf{z} \rangle. \tag{6}$$

In words, convexity is characterized by the existence of tangents $\mathbf{z}$ that lie below the function.

Definition 2.14 (sub-gradient)**.** *A vector $\mathbf{z}$ that satisfies eq.* (6) *is called a **sub-gradient** of f at $\mathbf{w}$. The set of sub-gradients of f at $\mathbf{w}$ is denoted $\partial f(\mathbf{w})$. Furthermore, if f is differentiable at $\mathbf{w}$ then $\partial f(\mathbf{w})$ contains single element – the gradient of f at $\mathbf{w}$, $\nabla f(\mathbf{w})$.*

Getting back to online convex optimization, by definition 2.13, for each round t, there exists $\mathbf{z}_t$ such that

$$f_t(\mathbf{w}_t) - f_t(\mathbf{u}) \leq \langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t \rangle$$

It follows that for any sequence of convex functions $f_1, ..., f_T$ and vectors $\mathbf{w}_1, ..., \mathbf{w}_T$, if for all t, $\mathbf{z}_t \in \partial f_t(\mathbf{w}_t)$ (namely, it is a sub-gradient), then

$$\sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{u})) \leq \sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t \rangle = \sum_{t=1}^{T} (\langle \mathbf{w}_t, \mathbf{z}_t \rangle - \langle \mathbf{u}, \mathbf{z}_t \rangle) \tag{7}$$

Combining the above observation with the FoReL procedure with Euclidean regularization (eq. (5)) yields the Online Gradient Descent algorithm:

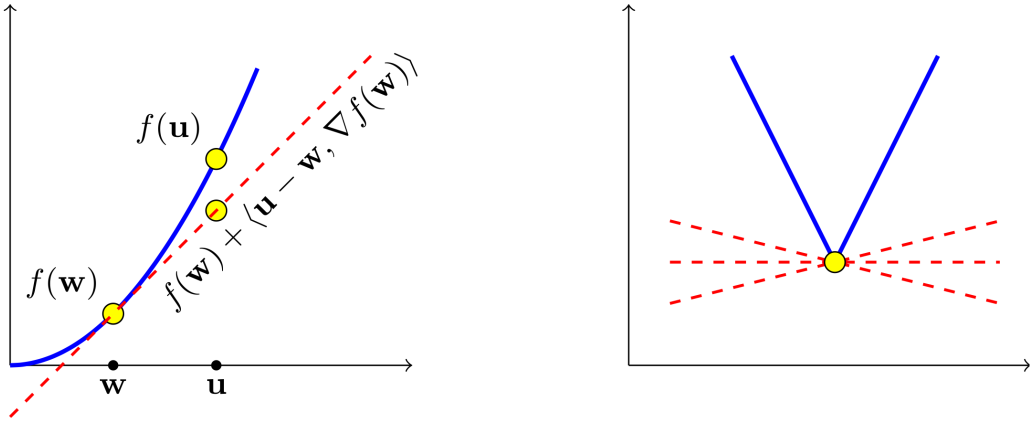
$f(\mathbf{u})$
$f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla f(\mathbf{w})\rangle$
$f(\mathbf{w})$
$\mathbf{w}$
$\mathbf{u}$

Figure 1: Left: Illustration of Equation (6). Right: Illustration of several sub-gradients of a non-differentiable convex function

Algorithm 8 `Online Gradient Descent (OGD)`

```
parameter: $\eta > 0$
initialize: $\mathbf{w}_1 = 0$
for t=2,3,...,T do
    $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{z}_t$ where $\mathbf{z}_t \in \partial f_t(\mathbf{w}_t)$
end for
```

To analyze OGD, we combine Equation (7) with the analysis for linear functions given in theorem 2.11, to get that

$$\begin{aligned}
\mathrm{Regret}_T(\mathbf{u}) &\le R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})) && \text{by lemma 2.10} \\
&\le \frac{1}{2\eta} \|\mathbf{u}\|_2^2 + \sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{w}_{t+1}, \mathbf{z}_t \rangle && \text{by eq. (7)} \\
&= \frac{1}{2\eta} \|\mathbf{u}\|_2^2 + \sum_{t=1}^{T} \langle \eta \mathbf{z}_t, \mathbf{z}_t \rangle && \text{as } \mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{z}_t \\
&= \frac{1}{2\eta} \|\mathbf{u}\|_2^2 + \eta \sum_{t=1}^{T} \|\mathbf{z}_t\|_2^2 && (8)
\end{aligned}$$

This regret bound depends on the norms of the sub-gradients of the vectors produced by the algorithm, and is therefore not satisfactory. To derive a more concrete bound, we must assure that the norms of sub-gradients will not be excessively large. One way to do this is by assuming that the functions are Lipschitz.

But before relating norms of sub-gradients to Lipschitzness of f_t, we first prove a useful corollary.

Corollary 2.15. *(Cauchy-Schwarz inequality for dual norm)*
For any vector $\mathbf{w}$, $\mathbf{z}$, $\langle \mathbf{w}, \mathbf{z} \rangle \le \|\mathbf{w}\| \, \|\mathbf{z}\|_*$ *for some norm* $\|\cdot\|$ *and its dual* $\|\cdot\|_*$.

Proof. Consider $\mathbf{u} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$. We have $\langle \mathbf{u}, \mathbf{z} \rangle = \frac{1}{\|\mathbf{w}\|} \langle \mathbf{w}, \mathbf{z} \rangle$. By the definition of dual norm, we have

$$\|\mathbf{z}\|_* = \max_{\|\mathbf{v}\| \le 1} \langle \mathbf{v}, \mathbf{z} \rangle$$

Since $\|\mathbf{u}\| = \left\| \frac{\mathbf{w}}{\|\mathbf{w}\|} \right\| = 1$, $\langle \mathbf{u}, \mathbf{z} \rangle \leq \|\mathbf{z}\|_*$. Therefore,

$$\begin{aligned} \langle \mathbf{w}, \mathbf{z} \rangle &= \|\mathbf{w}\| \langle \mathbf{u}, \mathbf{z} \rangle \\ &\leq \|\mathbf{w}\| \|\mathbf{z}\|_* \end{aligned}$$

□

Lemma 2.16. *Let $f : S \to \mathbb{R}$ be a convex function. Then, f is L-Lipschitz over S with respect to a norm $\|\cdot\|$ iff for all $\mathbf{w} \in S$ and $\mathbf{z} \in \partial f(\mathbf{w})$ we have $\|\mathbf{z}\|_* \leq L$, where $\|\mathbf{z}\|_*$ is the dual norm ($\|\mathbf{z}\|_* = \max\{\langle \mathbf{w}, \mathbf{z} \rangle : \|\mathbf{w}\| \leq 1\}$).*

Proof. Assume that f is Lipschitz. Choose some $\mathbf{w} \in S$, $\mathbf{z} \in \partial f(\mathbf{w})$. Let $\mathbf{u}$ be such that $\mathbf{u} - \mathbf{w} = \arg\max_{\mathbf{v}:\|\mathbf{v}\| \leq 1} \langle \mathbf{v}, \mathbf{z} \rangle$, $\|\mathbf{u} - \mathbf{w}\| \leq 1$. Thus, $\langle \mathbf{u} - \mathbf{w}, \mathbf{z} \rangle = \|\mathbf{z}\|_*$.

By the convexity of f, and from the definition of sub-gradient,

$$f(\mathbf{u}) - f(\mathbf{w}) \geq \langle \mathbf{u} - \mathbf{w}, \mathbf{z} \rangle$$

On the other hand, since f is Lipschitz, we have

$$L = L \cdot 1 \geq L \|\mathbf{u} - \mathbf{w}\| \geq f(\mathbf{u}) - f(\mathbf{w})$$

Combining the above two inequalities, we conclude that $\|\mathbf{z}\|_* \leq L$.

For the other direction, assume that $\|\mathbf{z}\|_* \leq L$. Since f is convex and $\mathbf{z} \in \partial f(\mathbf{w})$, we also have

$$f(\mathbf{w}) - f(\mathbf{u}) \leq \langle \mathbf{w} - \mathbf{u}, \mathbf{z} \rangle.$$

Combining the above with corollary 2.15: $\langle \mathbf{w}, \mathbf{z} \rangle \leq \|\mathbf{w}\| \|\mathbf{z}\|_*$, we obtain

$$f(\mathbf{w}) - f(\mathbf{u}) \leq \langle \mathbf{w} - \mathbf{u}, \mathbf{z} \rangle \leq \|\mathbf{w} - \mathbf{u}\| \|\mathbf{z}\|_* \leq L \|\mathbf{w} - \mathbf{u}\|.$$

Hence, f is L-Lipschitz. □

Since the dual of ℓ_2 norm is ℓ_2 norm, $\|\mathbf{z}\|_* = \|\mathbf{z}\|_2$ with respect to norm ℓ_2. Therefore, in eq. (8), as f_t is L_t-Lipschitz,

$$\sum_{t=1}^{T} \|\mathbf{z}\|_2^2 = \sum_{t=1}^{T} \|\mathbf{z}\|_*^2 \leq \sum_{t=1}^{T} L_t^2,$$

We conclude:

Corollary 2.17. *Assume that OGD is running on a sequence $f_1, ..., f_T$ of convex functions. Then, for all $\mathbf{u}$, we have*

$$\textit{Regret}_T(\mathbf{u}) \leq \frac{1}{2\eta} \|\mathbf{u}\|_2^2 + \eta \sum_{t=1}^{T} \|\mathbf{z}_t\|_2^2.$$

If we further assume that each f_t is L_t-Lipschitz with respect to $\|\cdot\|_2$, and let L be such that $\frac{1}{T}\sum_{t=1}^{T} L_t^2 \leq L^2$. Then, for all $\mathbf{u}$, the regret of OGD satisfies

$$\textit{Regret}_T(\mathbf{u}) \leq \frac{1}{2\eta} \|\mathbf{u}\|_2^2 + \eta T L^2.$$

In particular, if $U = \{\mathbf{u} : \|\mathbf{u}\|_2 \leq B\}$ *and* $\eta = \frac{B}{L\sqrt{2T}}$ *then*

$$Regret_T(U) \leq BL\sqrt{2T}$$

Let us discuss some consequences of Corollary 2.16 starting with the online linear regression problem (Example 2.2). Recall that for that example, $f_t(\mathbf{w}) = |\langle \mathbf{w}, \mathbf{x}_t\rangle - t_y|$, where $\mathbf{x}_t$ comes from a set A. If the set A is contained in a ball of radius of L (with respect to ℓ_2 norm), then f_t is $L-$Lipschitz. We therefore obtain a regret bound of $BL\sqrt{2T}$ which holds for all competing vectors $\mathbf{u}$ with $\|\mathbf{u}\|_2 \leq B$.

2.5 Strong Convex Regularizers

So far we applied FoReL with the Euclidean regularization function. However, this regularization cannot be used for learning with expert advice problem as it does not guarantee that $\mathbf{w}_t$ will always be in the probability simplex. In this section, we consider other regularization functions and underscore strong convexity as an important property of them.

2.5.1 Strong Convexity

Intuitively, a function is strongly convex if it grows faster than a linear function.

Definition 2.18. *A function* $f : S \to \mathbb{R}$ *is* σ*-strongly-convex over* S *with respect to a norm* $\|\cdot\|$ *if for any* $\mathbf{w} \in S$ *we have*

$$\forall \mathbf{z} \in \partial f(\mathbf{w}),\ \forall \mathbf{u} \in S,\ f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{z}, \mathbf{u} - \mathbf{w}\rangle + \frac{\sigma}{2}\|\mathbf{u} - \mathbf{w}\|^2$$

Lemma 2.19. *Let* S *be a nonempty convex set. Let* $f : S \to \mathbb{R}$ *be a* σ*-strongly-convex function over* S *with respect to a norm* $\|\cdot\|$*. Let* $\mathbf{w} = \arg\min_{\mathbf{v}\in S} f(\mathbf{v})$*. Then, for all* $\mathbf{u} \in S$

$$f(\mathbf{u}) - f(\mathbf{w}) \geq \frac{\sigma}{2}\|\mathbf{u} - \mathbf{w}\|^2.$$

Proof. If $\mathbf{u} \geq \mathbf{w}$, then $\mathbf{z} \geq 0$, so $\langle \mathbf{z}, \mathbf{u} - \mathbf{w}\rangle \geq 0$.
If $\mathbf{u} < \mathbf{w}$, then $\mathbf{z} \leq 0$, so $\langle \mathbf{z}, \mathbf{u} - \mathbf{w}\rangle \geq 0$.
Sadie: not very sure how to write the proof of these facts rigorously Thus, for all $\mathbf{u} \in S$, by strong convexity,

$$\forall \mathbf{z} \in \partial f(\mathbf{w}),\ \forall \mathbf{u} \in S,\ f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{z}, \mathbf{u} - \mathbf{w}\rangle \tag{9}$$

$$\begin{aligned} f(\mathbf{u}) - f(\mathbf{w}) &\geq \langle \mathbf{z}, \mathbf{u} - \mathbf{w}\rangle + \frac{\sigma}{2}\|\mathbf{u} - \mathbf{w}\|^2 \\ &\geq \frac{\sigma}{2}\|\mathbf{u} - \mathbf{w}\|^2 \end{aligned}$$

□

Corollary 2.20. *If R is twice differentiable, then a sufficient condition for strong convexity of R is that for all $\mathbf{w}, \mathbf{x} \in S$, $\langle \nabla^2 R(\mathbf{w})\mathbf{x}, \mathbf{x}\rangle = \mathbf{x}^T \nabla^2 R(\mathbf{w})\mathbf{x} \geq \sigma \|\mathbf{x}\|^2$, where $\nabla^2 R(\mathbf{w})$ is the Hessian matrix of R at $\mathbf{w}$, namely the matrix of second-order partial derivatives of R at $\mathbf{w}$.*

Proof. By Taylor expansion,

$$\forall \mathbf{u} \in S,\ R(\mathbf{u}) \geq R(\mathbf{w}) + \langle \nabla R(\mathbf{w}), \mathbf{u} - \mathbf{w}\rangle + \frac{1}{2}\langle \nabla^2 R(\mathbf{w})(\mathbf{u} - \mathbf{w}), \mathbf{u} - \mathbf{w}\rangle$$
$$\forall \mathbf{u} \in S,\ R(\mathbf{u}) \geq R(\mathbf{w}) + \langle \nabla R(\mathbf{w}), \mathbf{u} - \mathbf{w}\rangle + \frac{1}{2}(\mathbf{u} - \mathbf{w})^T \nabla^2 R(\mathbf{w})(\mathbf{u} - \mathbf{w})$$

We also know that R is σ-strongly-convex if

$$\forall \mathbf{z} \in \partial R(\mathbf{w}),\ \forall \mathbf{u} \in S,\ R(\mathbf{u}) \geq R(\mathbf{w}) + \langle \mathbf{z}, \mathbf{u} - \mathbf{w}\rangle + \frac{\sigma}{2}\|\mathbf{u} - \mathbf{w}\|^2$$

That is, R is σ-strongly-convex if

$$\forall \mathbf{u} \in S,\ \frac{1}{2}(\mathbf{u} - \mathbf{w})^T \nabla^2 R(\mathbf{w})(\mathbf{u} - \mathbf{w}) \geq \frac{\sigma}{2}\|\mathbf{u} - \mathbf{w}\|^2$$
$$\forall \mathbf{w}, \mathbf{x} \in S, \mathbf{x}^T \nabla^2 R(\mathbf{w})\mathbf{x} \geq \sigma \|\mathbf{x}\|^2$$

□

Example 2.21 (Euclidean regularization)**.** *The function $R(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$ is 1-strongly-convex with respect to the ℓ_2 norm over $\mathbb{R}^d$. To see this, simply note that the Hessian of R at any $\mathbf{w}$ is the identity matrix.*

Example 2.22 (Entropic regularization)**.** *The function $R(\mathbf{w}) = \sum_{i=1}^d w[i]\log(w[i])$ is $\frac{1}{B}$-strongly-convex with respect to the ℓ_1 norm over the set $S = \{\mathbf{w} \in \mathbb{R}^d : \mathbf{w} > 0 \wedge \|\mathbf{w}\|_1 \leq B\}$. In particular, R is 1-strongly-convex over the probability simplex, which is the positive vectors whose elements sum to 1, since $B = 1$ in that case. To see this, note that*

$$\begin{aligned}
\langle \nabla^2 R(\mathbf{w})\mathbf{x}, \mathbf{x}\rangle &= \sum_i \frac{x[i]^2}{w[i]} \\
&= \frac{1}{\|\mathbf{w}\|_1}\left(\sum_i w[i]\right)\left(\frac{x[i]^2}{w[i]}\right) && \|\mathbf{w}\|_1 = \sum_i w[i] \\
&\geq \frac{1}{\|\mathbf{w}\|_1}\left(\sum_i \sqrt{w[i]}\frac{|x[i]|}{\sqrt{w[i]}}\right)^2 && \textit{by Cauchy-Schwartz inequality} \\
&= \frac{1}{\|\mathbf{w}\|_1}\|\mathbf{x}\|_1^2 && (10)
\end{aligned}$$

Again, recall that if R is twice differentiable, then a sufficient condition for strong convexity of R is that for all $\mathbf{w}, \mathbf{x}$, $\langle \nabla^2 R(\mathbf{w})\mathbf{x}, \mathbf{x}\rangle \geq \sigma \|\mathbf{x}\|^2$ (Corollary 2.20). In this case, choose $\sigma = \frac{1}{B}$, we know R is $\frac{1}{B}$-strongly-convex.

Additional useful properties are given in the following lemmas, whose proof follows directly from the definition of strong convexity.

Lemma 2.23. *Given $\theta \in \mathbb{R}$, if R is 1-strongly-convex over S with respect to some norm, then θR is θ-strongly-convex over S with respect to the same norm. In addition, if S' is a convex subset of S, then R is 1-strongly convex over S' as well.*

Lemma 2.24. *The addition of a convex function to a strongly convex function keeps the strong convexity property.*

Proof. Say that f is convex, then by definition, for all $\mathbf{w} \in S$, there exists $\mathbf{z}$ such that

$$\forall \mathbf{u} \in S,\ f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \mathbf{z} \rangle.$$

We denote the set of such $\mathbf{z}$ as $\partial f(\mathbf{w})$.

Say that g is σ-strongly convex with respect to norm $\|\cdot\|$, then by definition, for all $\mathbf{w} \in S$, we have

$$\forall \mathbf{z} \in \partial g(\mathbf{w}),\ \forall \mathbf{u} \in S,\ g(\mathbf{u}) \geq g(\mathbf{w}) + \langle \mathbf{z}, \mathbf{u} - \mathbf{w} \rangle + \frac{\sigma}{2} \|\mathbf{u} - \mathbf{w}\|^2 .$$

Then, if $h = f + g$, we have

$$\partial h(\mathbf{x}) = \partial f(\mathbf{x}) + \partial g(\mathbf{x}) = \{a + b \mid a \in \partial f(\mathbf{x}), b \in \partial g(\mathbf{x})\}$$

Then, $\forall \mathbf{z} = z_1 + z_2 \in \partial h(\mathbf{w}),, \forall \mathbf{u} \in S$,

$$\begin{aligned} f(\mathbf{u}) + g(\mathbf{u}) &\geq f(\mathbf{w}) + \langle \mathbf{z}_1, \mathbf{u} - \mathbf{w} \rangle + g(\mathbf{w}) + \langle \mathbf{z}_2, \mathbf{u} - \mathbf{w} \rangle + \frac{\sigma}{2} \|\mathbf{u} - \mathbf{w}\|^2 \\ h(\mathbf{u}) &\geq h(\mathbf{w}) + \langle \mathbf{z}_1 + \mathbf{z}_2, \mathbf{u} - \mathbf{w} \rangle + \frac{\sigma}{2} \|\mathbf{u} - \mathbf{w}\|^2 \\ &= h(\mathbf{w}) + \langle \mathbf{z}, \mathbf{u} - \mathbf{w} \rangle + \frac{\sigma}{2} \|\mathbf{u} - \mathbf{w}\|^2 \end{aligned}$$

Therefore,

$$\forall \mathbf{z} \in \partial h(\mathbf{w}),\ \forall \mathbf{u} \in S,\ h(\mathbf{u}) \geq h(\mathbf{w}) + \langle \mathbf{z}, \mathbf{u} - \mathbf{w} \rangle + \frac{\sigma}{2} \|\mathbf{u} - \mathbf{w}\|^2 .$$

□

2.5.2 Analyzing FoReL with Strongly Convex Regularizers

We now analyze FoReL with strongly convex regularizers. Recall the regret bound given in Lemma 2.10:

$$\sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{u})) \leq R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})).$$

If f_t is L-Lipschitz with respect to a norm $\|\cdot\|$ then

$$f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) \leq L \|\mathbf{w}_t - \mathbf{w}_{t+1}\| .$$

Therefore, we need to ensure that $\|\mathbf{w}_t - \mathbf{w}_{t+1}\|$ is small. The following lemma shows that if the regularization function $R(\mathbf{w})$ is strongly convex with respect to the same norm, then $\mathbf{w}_t$ will be closed to $\mathbf{w}_{t+1}$.

Lemma 2.25. *Let $R : S \to \mathbb{R}$ be a σ-strongly-convex function over S with respect to a norm $\|\cdot\|$. Let $\mathbf{w}_1, \mathbf{w}_2, ...$ be the predictions of the FoReL algorithm. Then, for all t, if f_t is L_t-Lipschitz with respect to $\|\cdot\|$, then*

$$f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) \leq L_t \|\mathbf{w}_t - \mathbf{w}_{t+1}\| \leq \frac{L_t^2}{\sigma}.$$

Proof. For all t, let $F_t(\mathbf{w}) = \sum_{i=1}^{t-1} f_i(\mathbf{w}) + R(\mathbf{w})$ and note that the FoReL rule is $\mathbf{w}_t = \arg\min_{\mathbf{w}\in S} F_t(\mathbf{w})$. Note also that F_t is σ-strongly-convex since the addition of a convex function to a strongly convex function keeps the strong convexity property(Lemma 2.24). Therefore, Lemma 2.19 implies that:

$$F_t(\mathbf{w}_{t+1}) \geq F_t(\mathbf{w}_t) + \frac{\sigma}{2} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 .$$

Repeating the same argument for F_{t+1} and its minimizer w_{t+1} we get

$$F_{t+1}(\mathbf{w}_t) \geq F_{t+1}(\mathbf{w}_{t+1}) + \frac{\sigma}{2} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 .$$

Summing the above two inequalities and rearranging then, we obtain

$$\begin{aligned} \sigma \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 &\leq F_{t+1}(\mathbf{w}_t) - F_t(\mathbf{w}_t) + F_t(\mathbf{w}_{t+1}) - F_{t+1}(\mathbf{w}_{t+1}) \\ &= f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}). \end{aligned} \tag{11}$$

Next, using the Lipschitzness of f_t we get that

$$f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) \leq L_t \|\mathbf{w}_t - \mathbf{w}_{t+1}\| .$$

If $\mathbf{w}_t - \mathbf{w}_{t+1} = 0$, we have $f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) = 0 = L_t \|\mathbf{w}_t - \mathbf{w}_{t+1}\| \leq \frac{L_t^2}{\sigma}$.
If $\mathbf{w}_t - \mathbf{w}_{t+1} \neq 0$, combining this with Equation (11) and rearranging, we get that $\|\mathbf{w}_t - \mathbf{w}_{t+1}\| \leq \frac{L_t}{\sigma}$, therefore,

$$f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) \leq L_t \|\mathbf{w}_t - \mathbf{w}_{t+1}\| \leq \frac{L_t^2}{\sigma}.$$

□

Combining the above Lemma with Lemma 2.10, we obtain

Theorem 2.26. *Let $f_1, ..., f_T$ be a sequence of convex functions such that f_t is L_t-Lipschitz with respect to some norm $\|\cdot\|$. Let L be such that $\frac{1}{T}\sum_{t=1}^{T} L_t^2 \leq L^2$. Assume that FoReL is run on the sequence with the regularization function which is σ-strongly-convex with respect to the same norm. Then, for all $\mathbf{u} \in S$,*

$$\mathit{Regret}_T(\mathbf{u}) \leq R(\mathbf{u}) - R(\mathbf{w}_1) + \frac{TL^2}{\sigma}.$$

given $\mathbf{w}_1, \mathbf{w}_2, \ldots$ for FoReL, with $\mathbf{w}_1 = \min_{\mathbf{v}\in S} R(\mathbf{v})$ as the initial regret is 0.

2.5.3 Derived Bounds

We now derive concrete bounds from Theorem 2.23. We start with the simplest case of Euclidean regularization, which is 1-strongly-convex over $\mathbb{R}^d$, hence the following corollary follows.

Corollary 2.27. *Let $f_1, ..., f_T$ be a sequence of convex functions such that f_t is L_t-Lipschitz with respect to $\|\cdot\|_2$. Let L be such that $\frac{1}{T}\sum_{t=1}^{T} L_t^2 \leq L^2$. Assume that FoReL is run on the sequence with the regularization function $R(\mathbf{w}) = \frac{1}{2\eta}\|\mathbf{w}\|_2^2$. Then, for all $\mathbf{u}$*

$$\textit{Regret}_T(\mathbf{u}) \leq \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \eta T L^2$$

since $\frac{1}{2}\|\mathbf{w}\|_2^2$ is 1-strongly-convexity, so $\frac{1}{2\eta}\|\mathbf{w}\|_2^2$ is $\frac{1}{\eta}$-strongly-convexity.

In particular, if $U = \{\mathbf{u} : \|\mathbf{u}\|_2 \leq B\}$ and $\eta = \frac{B}{L\sqrt{2T}}$ then

$$\textit{Regret}_T(U) \leq BL\sqrt{2T}.$$

Observe that the bound we obtained is identical to the bound of Online-Gradient-Descent given in Corollary 2.16.

Then, we consider the problem of prediction with expert advice. As mentioned previously, the Euclidean regularization cannot be applied into this problem since it does not enforce $\mathbf{w}_t$ to be in the probability simplex. A simple solution to enforce the constraint $\mathbf{w}_t \in S$ by setting $R(\mathbf{w}) = \infty$ whenever $\mathbf{w} \notin S$. The resulting regularization function is still strongly convex by Lemma 2.24. (The indicator function $I(\mathbf{w}) = \begin{cases} 0 & \text{if } \mathbf{w} \in S \\ \infty & \text{if } \mathbf{w} \notin S \end{cases}$ is convex when the set is convex, and sum of convex function and strongly convex function is strongly convex.)

Corollary 2.28. *Let $f_1, ..., f_T$ be a sequence of convex functions such that f_t is L_t-Lipschitz with respect to $\|\cdot\|_2$. Let L be such that $\frac{1}{T}\sum_{t=1}^{T} L_t^2 \leq L^2$. Let S be a convex set and assume that FoReL is run on the sequence with the regularization function*

$$R(\mathbf{w}) = \begin{cases} \frac{1}{2\eta}\|\mathbf{w}\|_2^2 & \textit{if } \mathbf{w} \in S \\ \infty & \textit{if } \mathbf{w} \notin S \end{cases}$$

Then, for all $\mathbf{u} \in S$,

$$\textit{Regret}_T(\mathbf{u}) \leq \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \eta T L^2.$$

In particular, if $B \geq \max_{\mathbf{u}\in S}\|\mathbf{u}\|_2$ and $\eta = \frac{B}{L\sqrt{2T}}$ then

$$\textit{Regret}_T(S) \leq BL\sqrt{2T}.$$

Then, we can apply the regularization function given in the above corollary to the problem of prediction with expert advice: S is the probability simplex, $\mathbf{x}_t \in [0,1]^d$. Hence, we can set $B = 1$ and $L = \max_{\mathbf{x}\in S}\|\nabla f_t(\mathbf{x})\|_2 = \sqrt{d}$ which leads to the regret bound of $\sqrt{2dT}$.

We next show another regularization function which leads to a regret bound of $\sqrt{2\log(d)T}$. That is the Entropic regularization introduced in Example 2.20.

Corollary 2.29. *Let $f_1, ..., f_T$ be a sequence of convex functions such that f_t is L_t-Lipschitz with respect to $\|\cdot\|_1$. Let L be such that $\frac{1}{T}\sum_{t=1}^{T} L_t^2 \leq L^2$. Let S be a convex set and assume that FoReL is run on the sequence with the regularization function $R(\mathbf{w}) = \frac{1}{\eta}\sum_{i=1}^{d} w[i]\log(w[i])$ and with the set $S = \{\mathbf{w} : \|\mathbf{w}\|_1 = B \wedge \mathbf{w} > 0\} \subset \mathbb{R}^d$. Then,*

$$\textit{Regret}_T(S) \leq \frac{B\log(d)}{\eta} + \eta BTL^2.$$

In particular, setting $\eta = \frac{\sqrt{\log(d)}}{L\sqrt{2T}}$ *yields*

$$\textit{Regret}_T(S) \leq BL\sqrt{2\log(d)T}$$

Proof. By theorem 2.26, say R is σ-strongly-convex, we have for any $\mathbf{u} \in S$,

$$\begin{aligned}
\text{Regret}_T(\mathbf{u}) &\leq R(\mathbf{u}) - R(\mathbf{w}_1) + \frac{TL^2}{\sigma} \\
&\leq \frac{B\log(d)}{\eta} + \frac{TL^2}{\sigma} \\
&\leq \frac{B\log(d)}{\eta} + \eta BTL^2 && R \text{ is } \frac{1}{\eta}\cdot\frac{1}{B}\text{-strongly-convex}
\end{aligned}$$

□

Note that the Entropic regularization used in the above corollary is strongly convex with respect to the ℓ_1 norm, and therefore the Lipschitzeness requirement of the loss functions is also with respect to the ℓ_1 norm.
When applying this to the problem of prediction with expert advice (see section 2.1.1), it's clear that $B = 1$, but we need to be careful with L. In this problem, the loss function $f_t(\mathbf{w}) = \langle \mathbf{w}, \mathbf{y}_t \rangle$ is a linear function. We have by Corollary 2.15 that,

$$|f_t(\mathbf{w}) - f_t(\mathbf{u})| = |\langle \mathbf{w} - \mathbf{u}, \mathbf{y}_t \rangle| \leq \|\mathbf{w} - \mathbf{u}\|_1 \|\mathbf{y}_t\|_\infty .$$

Since $\mathbf{y}_t \in [0,1]^d$, so $\|\mathbf{y}_t\|_\infty \leq 1$. Thus, we can set $L = 1$, and obtain the regret bound of $\sqrt{2\log(d)T}$.

Remark 2.30. *It's interesting to compare the two bounds given in Corollary 2.25 and Corollary 2.26. In Corollary 2.25, the parameter B imposes an ℓ_2 constraint on $\mathbf{u}$ and the parameter L captures Lipschitzness of the loss functions with respect to the ℓ_2 norm. In contrast, in Corollary 2.26, the parameter B imposes an ℓ_1 constraint on $\mathbf{u}$ and the parameter L captures Lipschitzness of the loss functions with respect to the ℓ_1 norm. Therefore, the choice of the regularization function should depend on the Lipschitzness of the loss functions and on prior assumptions on the set of competing vectors (does a competing vector have a smaller ℓ_1 norm or only a small ℓ_2 norm). In the prediction with expert advice, the competing vector would be a singleton so both the ℓ_1 and ℓ_2 norms are 1. On the other hand, the gap between Lipschitzeness with respect to ℓ_2 norm ($\sqrt{d}$) and Lipschitzness with respect to ℓ_1 norm (1) is large.*

Amy: TODO: let's come up with an example. let's choose loss functions, and plot regret bounds, to show that for different choices, sometimes the L_2 regularizer is better, but perhaps more often the L_1 regularizer is better.

2.6 Online Mirror Descent

A possible disadvantage of the FoReL approach is that it requires solving an optimization problem at each online round. In this section, we will derive and analyze the family of Online Mirror Descent algorithms from the FoReL framework. We will show that Online Mirror Descent achieves the same regret bound as FoReL but the update step is much simpler.

We start with applying FoReL on a sequence of linear functions in which $f_t(\mathbf{w}) = \langle \mathbf{w}, \mathbf{z}_t \rangle$ with some regularization function $R(\mathbf{w})$. Thoughout, we assume that $R(\mathbf{w}) = \infty \ \forall \mathbf{w} \notin S$. We use the notation $\mathbf{z}_{1:t} = \sum_{i=1}^{t} \mathbf{z}_i$. Then, we can rewrite the prediction of FoReL as:

$$\begin{aligned}\mathbf{w}_{t+1} &= \arg\min_{\mathbf{w}} R(\mathbf{w}) + \sum_{i=1}^{t} \langle \mathbf{w}, \mathbf{z}_i \rangle \\ &= \arg\min_{\mathbf{w}} R(\mathbf{w}) + \langle \mathbf{w}, \mathbf{z}_{1:t} \rangle \\ &= \arg\max_{\mathbf{w}} \langle \mathbf{w}, -\mathbf{z}_{1:t} \rangle - R(\mathbf{w}).\end{aligned}$$

Letting

$$g(\boldsymbol{\theta}) = \arg\max_{\mathbf{w}} \langle \mathbf{w}, \boldsymbol{\theta} \rangle - R(\mathbf{w}) \tag{12}$$

We can now rewrite the FoReL prediction based on the following recursive update rule:

1. $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{z}_t$
   (Let $\boldsymbol{\theta}_t = -\mathbf{z}_{1:t-1}$, then $\boldsymbol{\theta}_{t+1} = -\mathbf{z}_{1:t} = -\mathbf{z}_{1:t-1} - \mathbf{z}_t$.)

2. $\mathbf{w}_{t+1} = g(\boldsymbol{\theta}_{t+1})$
   ($\mathbf{w}_{t+1} = \arg\max_{\mathbf{w}} \langle \mathbf{w}, -\mathbf{z}_{1:t} \rangle - R(\mathbf{w}) = g(\boldsymbol{\theta}_{t+1})$.)

Now, if f_t is convex but not linear, we can use the same technique we used for deriving the Online Gradient Descent algorithm and use sub-gradients of f_t at $\mathbf{w}_t$ to linearize the problem. That is, letting $\mathbf{z}_t$ be a sub-gradient of f_t at $\mathbf{w}_t$, we have that for all $\mathbf{u} \in U$

$$\begin{aligned}f_t(\mathbf{u}) &\geq f_t(\mathbf{w}_t) + \langle \mathbf{u} - \mathbf{w}_t, \mathbf{z}_t \rangle \\ f_t(\mathbf{w}_t) - f_t(\mathbf{u}) &\leq \langle \mathbf{w}_t, \mathbf{z}_t \rangle - \langle \mathbf{u}, \mathbf{z}_t \rangle \\ \sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{u}) &\leq \sum_{t=1}^{T} \langle \mathbf{w}_t, \mathbf{z}_t \rangle - \langle \mathbf{u}, \mathbf{z}_t \rangle.\end{aligned}$$

Summing over t we obtain that the regret with respect to the nonlinear loss functions ($f_t(\mathbf{w}_t)$) is upper bounded by the regret with respect to the linear functions ($\langle \mathbf{w}_t, \mathbf{z} \rangle$). This yields the Online Mirror Descent framework.

Algorithm 9 `Online Mirror Descent (OMD)`

```
parameter: a link function $g : \mathbb{R}^d \to S$, defined by Equation (12)
initialize: $\boldsymbol{\theta}_1 = 0$
for t=1,2,... do
    predict $\mathbf{w}_t = g(\boldsymbol{\theta}_t)$
    update $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{z}_t$ where $\mathbf{z}_t \in \partial f_t(\mathbf{w}_t)$
end for
```

$$\begin{aligned}\mathbf{w}_t &= g(\boldsymbol{\theta}_t) \\ \mathbf{w}_{t+1} &= g(\boldsymbol{\theta}_{t+1}) \\ \mathbf{w}_{t+1} - \mathbf{w}_t &= -\eta \mathbf{z}_t \\ \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t &= -\mathbf{z}_t\end{aligned}$$

Algorithm 10 Deni: how about this? `Online Mirror Descent (OMD)`

```
parameter: a link function $g : \mathbb{R}^d \to S$, defined by Equation (12)
initialize: $\boldsymbol{\theta}_1 = 0$
for t=1,2,... do
    predict $\mathbf{w}_t = g(\boldsymbol{\theta}_t)$
    update $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \mathbf{z}_t$ where $\mathbf{z}_t \in \partial f_t(\mathbf{w}_t)$
end for
```

Online Gradient Descent is a special case of Online Mirror Descent where $S = \mathbb{R}^d$ and $g(\boldsymbol{\theta}) = \eta\boldsymbol{\theta}$, for some $\eta > 0$. When g is nonlinear, we obtain that the vector $\boldsymbol{\theta}$ is updated by subtracting the gradient out of it, but the actual prediction is "mirrored" or "linked" to the set S via the function g. This is why g is often referred to as a link function.

Now, we will first give a generic bound for the OMD family based on our analysis of FoReL rule.

Theorem 2.31. *Let R be a $\frac{1}{\eta}$-strongly-convex function over S with respect to a norm $\|\cdot\|$. Assume that OMD is run on the sequence with a link function*

$$g(\boldsymbol{\theta}) = \arg\max_{\mathbf{w} \in S} (\langle \mathbf{w}, \boldsymbol{\theta} \rangle - R(\mathbf{w})). \tag{13}$$

Then, for all $\mathbf{u} \in S$,

$$\mathit{Regret}_T(\mathbf{u}) \leq R(\mathbf{u}) - \min_{\mathbf{v} \in S} R(\mathbf{v}) + \eta \sum_{t=1}^{T} \|\mathbf{z}_t\|_*^2 ,$$

where $\|\cdot\|_*$ is the dual norm. Furthermore, if f_t is L_t-Lipschitz with respect to $\|\cdot\|$, then we can further upper bound $\|\mathbf{z}\|_* \leq L_t$.

Proof. As we have shown previously,

$$\sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{u})) \leq \sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t \rangle,$$

and this OMD algorithm is equivalent to running FoReL on the sequence of linear functions with the regularization $R(\mathbf{w})$. Recall Theorem 2.26 and Lemma 2.16, the theorem follows directly form them. □

2.6.1 Derived Algorithms

We now derive additional algorithms from the OMD framework.

The first algorithm we derive is often called normalized Exponential Gradient. In this algorithm, $S = \{\mathbf{w} : \|\mathbf{w}\|_1 = 1 \wedge \mathbf{w} \geq 0\}$ is the probability simplex and $g : \mathbb{R}^d \to S$ is the vector-valued function whose ith component is the function

$$g_i(\boldsymbol{\theta}) = \frac{e^{\eta\theta[i]}}{\sum_j e^{\eta\theta[j]}} \tag{14}$$

Therefore,

$$\begin{aligned} w_{t+1}[i] &= \frac{e^{\eta\theta_{t+1}[i]}}{\sum_j e^{\eta\theta_{t+1}[j]}} \\ &= \frac{e^{\eta\theta_{t+1}[i]}}{\sum_j e^{\eta\theta_{t+1}[j]}} \cdot \frac{\sum_k e^{\eta\theta_t[k]}}{\sum_k e^{\eta\theta_t[k]}} \\ &= \frac{e^{\eta\theta_t[i]}e^{-\eta z_t[i]}}{\sum_j e^{\eta\theta_t[j]}e^{-\eta z_t[j]}} \cdot \frac{\sum_k e^{\eta\theta_t[k]}}{\sum_k e^{\eta\theta_t[k]}} && \text{Since } \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{z}_t \\ &= \frac{w_t[i]e^{-\eta z_t[j]}}{\sum_j w_t[j]e^{-\eta z_t[j]}} \end{aligned}$$

Algorithm 11 `Normalized Exponential Gradient (normalized-EG)`

```
parameter: $\eta > 0$
initialize: $\mathbf{w}_1 = (\frac{1}{d}, ..., \frac{1}{d})$
update rule:
$\forall i,\ w_{t+1}[i] = \frac{w_t[i]e^{-\eta z_t[j]}}{\sum_j w_t[j]e^{-\eta z_t[j]}}$ where $\mathbf{z}_t \in \partial f_t(\mathbf{w}_t)$
```

Corollary 2.32. *The normalized-EG algoritm enjoys the regret bound given in Corollary 2.29:*

$$Regret_T(S) \le \frac{\log(d)}{\eta} + \eta T L^2$$

with $B = 1$.

Proof. To analyze the normalized-EG, we rely on Theorem 2.31. Let $R(\mathbf{w}) = \frac{1}{\eta}\sum_i w[i]\log(w[i])$ be the Entropic regularization, let S be the probability simplex, and recall that R is $\frac{1}{\eta}$-strongly-convex over S with respect to the ℓ_1 norm.

Using the technique of Lagrange multipliers, it's easy Amy: FOR DENI! to verify that the link function in Equation (14) is the solution to the optimization problem given in Equation (13). Therefore, Theorem 2.31 yields: for all $\mathbf{u} \in S$, Sadie: added here

$$\begin{aligned} \text{Regret}_T(\mathbf{u}) &\le R(\mathbf{u}) - \min_{\mathbf{v}\in S} R(\mathbf{v}) + \eta \sum_{t=1}^{T} \|\mathbf{z}_t\|_*^2 \\ &\le R(\mathbf{u}) + \eta \sum_{t=1}^{T} \|\mathbf{z}_t\|_*^2 \\ &\le \frac{\log(d)}{\eta} + \eta T L^2 && \frac{1}{\eta}\sum_i w[i]\log(w[i]) \le \frac{\log(d)}{\eta},\ \|\mathbf{z}_t\|_* \le L_t^2 \end{aligned}$$

□

Next, we derive an algorithm which is called Online Gradient Descent with Lazy Projections. To derive the algorithm, let S be a convex set and define

$$g(\boldsymbol{\theta}) = \arg\min_{\mathbf{w}\in S} \|\mathbf{w} - \eta\boldsymbol{\theta}\|_2 \,.$$

Amy: please add verification right here That is, $g(\boldsymbol{\theta})$ returns the point in S which is closest to $\eta\boldsymbol{\theta}$.

Algorithm 12 `Online Gradient Descent with Lazy Projections`

```
parameter: $\eta > 0$ and a convex set $S$
initialize: $\boldsymbol{\theta}_1 = 0$
for t=1,2,...,T do
    $\mathbf{w}_t = g(\boldsymbol{\theta}_t) = \arg\min_{\mathbf{w}\in S} \|\mathbf{w} - \eta\boldsymbol{\theta}\|_2$
    $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{z}_t$ where $\mathbf{z}_t \in \partial f_t(\mathbf{w}_t)$
end for
```

Corollary 2.33. *Online Gradient Descent with Lazy Projections enjoys the same regret bound given in Corollary 2.28: For all* $\mathbf{u} \in S$*,*

$$\textit{Regret}_T(\mathbf{u}) \le \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \eta T L^2.$$

In particular, if $B \ge \max_{\mathbf{u}\in S}\|\mathbf{u}\|_2$ *and* $\eta = \frac{B}{L\sqrt{2T}}$ *then*

$$\textit{Regret}_T(S) \le BL\sqrt{2T}.$$

Proof. To analyze the Online Gradient Descent with Lazy Projections algorithm, we consider the Euclidean regularization function $R(\mathbf{w}) = \frac{1}{2\eta}\|\mathbf{w}\|_2^2$, which is $\frac{1}{\eta}$-strongly-convex over S with respect to the ℓ_2 norm. We have that

$$\begin{aligned}
\arg\max_{\mathbf{w}\in S}(\langle \mathbf{w}, \boldsymbol{\theta}\rangle - R(\mathbf{w})) &= \arg\max_{\mathbf{w}\in S}\left(\langle \mathbf{w}, \boldsymbol{\theta}\rangle - \frac{1}{2\eta}\|\mathbf{w}\|_2^2\right) \\
&= \arg\max_{\mathbf{w}\in S}\left(\langle \mathbf{w}, \eta\boldsymbol{\theta}\rangle - \frac{1}{2}\|\mathbf{w}\|_2^2\right) \\
&= \arg\min_{\mathbf{w}\in S}\left(\frac{1}{2}\|\mathbf{w}\|_2^2 - \langle \mathbf{w}, \eta\boldsymbol{\theta}\rangle\right) \\
&= \arg\min_{\mathbf{w}\in S}\left(\frac{1}{2}\|\mathbf{w}\|_2^2 - \langle \mathbf{w}, \eta\boldsymbol{\theta}\rangle + \frac{1}{2}\|\eta\boldsymbol{\theta}\|_2^2\right) \\
&= \arg\min_{\mathbf{w}\in S}\frac{1}{2}\|\mathbf{w} - \eta\boldsymbol{\theta}\|_2^2 \\
&= \arg\min_{\mathbf{w}\in S}\|\mathbf{w} - \eta\boldsymbol{\theta}\|_2
\end{aligned}$$

Therefore, by Theorem 2.31, we can conclude that for all $\mathbf{u} \in S$,

$$\begin{aligned}
\text{Regret}_T(\mathbf{u}) &\le R(\mathbf{u}) - \min_{\mathbf{v}\in S} R(\mathbf{v}) + \eta\sum_{t=1}^{T}\|\mathbf{z}_t\|_*^2 \\
&\le \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \eta T L^2 && \text{since } \|\mathbf{z}_t\|_* \le L_t
\end{aligned}$$

□

Finally, we derive the p-norm algorithm, in which $S = \mathbb{R}^d$ and

$$g_i(\boldsymbol{\theta}) = \eta \frac{\text{sign}(\theta[i])|\theta[i]|^{p-1}}{\|\boldsymbol{\theta}\|_p^{p-2}}$$

where $p \geq 2$ is a parameter and

$$\|\boldsymbol{\theta}\|_p = \left(\sum_{i=1}^{d} |\theta[i]|^p \right)^{\frac{1}{p}}$$

Algorithm 13 p-`norm`

```
parameter: $\eta > 0$ and $p > 2$
initialize: $\boldsymbol{\theta}_1 = 0$
for t=1,2,...,T do
    $\forall i,\ w_{t,i} = \eta \frac{\text{sign}(\theta[i])|\theta[i]|^{p-1}}{\|\boldsymbol{\theta}\|_p^{p-2}}$
    $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{z}_t$ where $\mathbf{z}_t \in \partial f_t(\mathbf{w}_t)$
end for
```

Corollary 2.34. *Let $f_1, \ldots, f_T$ be a sequence of convex functions such that f_t is L_T-Lipschitz over $\mathbb{R}^d$ with respect to $\|\cdot\|_q$. Let L be such that $\frac{1}{T}\sum_{t=1}^{T} L_t^2 \leq L^2$. Then, for all $\mathbf{u}$, the regret of the p-norm algorithm satisfies*

$$Regret_T(\mathbf{u}) \leq \frac{1}{2\eta(q-1)} \|\mathbf{w}\|_q^2 + \eta T L^2.$$

In particular, if $U = \{\mathbf{u} : \|\mathbf{u}\|_q \leq B\}$ and $\eta = \frac{B}{L\sqrt{2T/(q-1)}}$ then

$$Regret_T(U) \leq BL\sqrt{\frac{2T}{q-1}}$$

Proof. To analyze the p-norm algorithm, we consider the regularization function $R(\mathbf{w}) = \frac{1}{2\eta(q-1)} \|\mathbf{w}\|_q^2$, where $q = \frac{p}{p-1}$. If $q \in (1, 2]$, then $\frac{1}{2\eta(q-1)} \geq \frac{1}{2\eta}$, then R is $\frac{1}{\eta}$-strongly-convex over $\mathbb{R}^d$ with respect to ℓ_q norm. It is also possible to verify that $g(\boldsymbol{\theta}) = \arg\max_{\mathbf{w}} \langle \mathbf{w}, \boldsymbol{\theta} \rangle - R(\mathbf{w})$ Sadie: I still don't know how to do that, Lagrange multiplier?. Therefore, by Theorem 2.31, we can conclude that for all $\mathbf{u}$, the regret of the p-norm algorithm satisfies

$$\text{Regret}_T(\mathbf{u}) \leq \frac{1}{2\eta(q-1)} \|\mathbf{w}\|_q^2 + \eta T L^2.$$

□

Note that when $q = 2$, the link function becomes $g(\boldsymbol{\theta}) = \eta\boldsymbol{\theta}$ and the p-norm algorithm becomes the Online Gradient Descent algorithm. When q is close to 1, we can see that the p-norm algorithm behaves like the Entropic regularization. In particular, when $p = \log(d)$, we can obtain a regret bound similar to the regret bound of the EG algorithm. Intermediate values of q enables us to interpolate between the properties of the Entropic and Euclidean regularizations.

2.7 The Language of Duality

In this section, we present a different proof technique that relies on duality. Here are some reasons to consider this different approach:

1. It is easier to derive tighter bounds based on duality approach. In particular, we will tighten the regret bounds we derived for the OMD framework by a factor of $\sqrt{2}$.

2. It may become convenient for developing new algorithms.

3. Many previous papers on online learning uses the language of duality.

2.7.1 Fenchel Conjugacy

There are two equivalent representations of a convex function. Either as pairs $(x, f(x))$ or as the set of tangents of f, namely pairs of the form (slope, intersection-with-y-axis). The function f^* that related slopes of tangents to their intersection with the y-axis is called the **Fenchel conjugate** of f, and is formally defined as

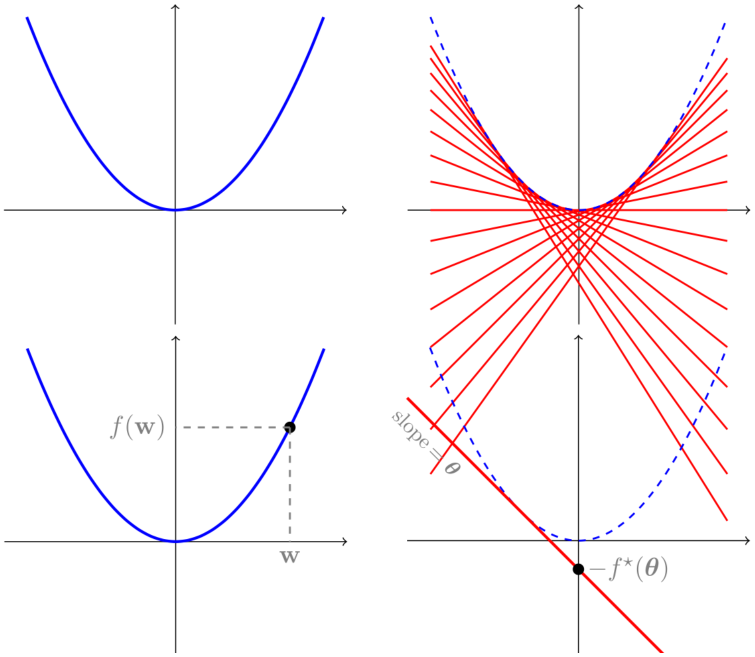

Figure 2: Left: pair $(x, f(x))$. Right: pair (slope,intersection-with-y-axis)

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

It is possible to show that $f = (f^*)^*$ if and only if f is closed and convex function Sadie: maybe need to add proof here. From now on, we always assume that our functions are closed and convex.

The definition of Fenchel conjugate immediately implies **Fenchel-Young inequality**:

$$\forall \mathbf{u},\ f^*(\boldsymbol{\theta}) \geq \langle \mathbf{u}, \boldsymbol{\theta} \rangle - f(\mathbf{u}). \tag{15}$$

It is possible to show Sadie: maybe need to add proof here, and examples that the equality holds if $\mathbf{u}$ is a sub-gradient of f^* at $\boldsymbol{\theta}$ and in particular, if f^* is differentiable, equality holds when $\mathbf{u} = \nabla f^*(\boldsymbol{\theta})$.

Table 2.1. Example of Fenchel conjugate pairs.

$f(\mathbf{w})$	$f^\star(\boldsymbol{\theta})$	Comments
$\frac{1}{2}\|\mathbf{w}\|_2^2$	$\frac{1}{2}\|\boldsymbol{\theta}\|_2^2$	
$\frac{1}{2}\|\mathbf{w}\|_q^2$	$\frac{1}{2}\|\boldsymbol{\theta}\|_p^2$	where $\frac{1}{p}+\frac{1}{q}=1$
$\sum_i w[i]\log(w[i]) + I_{\{\mathbf{v}\geq 0:\|\mathbf{v}\|_1=1\}}(\mathbf{w})$	$\log(\sum_i e^{\theta[i]})$	(normalized Entropy)
$\sum_i w[i](\log(w[i]) - 1)$	$\sum_i e^{\theta[i]}$	(un-normalized Entropy)
$\frac{1}{\eta}g(\mathbf{w})$	$\frac{1}{\eta}g^\star(\eta\boldsymbol{\theta})$	where $\eta > 0$
$g(\mathbf{w}) + \langle\mathbf{w},\mathbf{x}\rangle$	$g^\star(\boldsymbol{\theta}-\mathbf{x})$	

Here is a list of several Fenchel conjugate pairs. Recall that given a set S, we use the notation

$$I_S(\mathbf{w}) = \begin{cases} 0 & \mathbf{w} \in S \\ \infty & \mathbf{w} \notin S \end{cases}$$

2.7.2 Bregman Divergences and the Strong/Smooth Duality

A differentiable function R defines a Bergman divergence between two vectors as follows:

$$D_R(\mathbf{w}||\mathbf{u}) = R(\mathbf{w}) - (R(\mathbf{u}) + \langle\nabla R(\mathbf{u}), \mathbf{w}-\mathbf{u}\rangle). \tag{16}$$

That is, the Bregman divergence is the difference, at the point $\mathbf{w}$, between R and its linearization around $\mathbf{u}$. When R is convex, the Bregman divergence if always non-negative. However, it is not a metric measure because it is not symmetric and also does not satisfy the triangle inequality.

When $R(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$, the Bregman divergence is $D_R(\mathbf{w}||\mathbf{u}) = \frac{1}{2}\|\mathbf{w}-\mathbf{u}\|_2^2$.

When $R(\mathbf{w}) = \sum_i w[i]\log(w[i])$, the Bregman divergence between two vectors in the probability simplex becomes the Kullback-Leibler divergence, $D_R(\mathbf{w}||\mathbf{u}) = \sum_i w[i]\log\frac{w[i]}{u[i]}$.

Recall the definition of strong-convexity (definition 2.18). If R is differentiable, we can rewrite the σ-strong-convexity requirement as

$$D_R(\mathbf{w}||\mathbf{u}) \geq \frac{\sigma}{2}\|\mathbf{w}-\mathbf{u}\|^2$$

definition 2.18:

$$\forall\mathbf{z}\in\partial f(\mathbf{w}),\ \forall\mathbf{u}\in S,\ f(\mathbf{u}) \geq f(\mathbf{w}) + \langle\mathbf{z},\mathbf{u}-\mathbf{w}\rangle + \frac{\sigma}{2}\|\mathbf{u}-\mathbf{w}\|^2$$

Definition 2.35. *A function R is σ-strongly-smooth with respect to a norm $\|\cdot\|$ if it is differentiable and for all $\mathbf{u}, \mathbf{w}$ we have*

$$D_R(\mathbf{w}||\mathbf{u}) \leq \frac{\sigma}{2}\|\mathbf{w}-\mathbf{u}\|^2$$

$$\forall\mathbf{z}\in\partial f(\mathbf{w}),\ \forall\mathbf{u}\in S,\ f(\mathbf{u}) \leq f(\mathbf{w}) + \langle\mathbf{z},\mathbf{u}-\mathbf{w}\rangle + \frac{\sigma}{2}\|\mathbf{u}-\mathbf{w}\|^2$$

Not surprisingly, strong convexity and strong smoothness are dual properties.

Lemma 2.36 (Strong/Smooth Duality)**.** *Assume that R is a closed and convex function. Then R is β-strongly convex with respect to a norm $\|\cdot\|$ if and only if R^* is $\frac{1}{\beta}$-strongly smooth with respect to the dual norm $\|\cdot\|_*$.*

Proof. Sadie: make up proof later? search up online □

The above lemma implies in particular that if R is strongly convex then R^* is strongly smooth, and thus differentiable. Based on Section 2.7.1, this also implies that

$$\nabla R^*(\boldsymbol{\theta}) = \arg\max_{\mathbf{w}}(\langle \mathbf{w}, \boldsymbol{\theta}\rangle - R(\mathbf{w})) \tag{17}$$

Since we say if f^* is differentiable, $f^*(\boldsymbol{\theta}) = \langle \mathbf{u}, \boldsymbol{\theta}\rangle - f(\mathbf{u})$ holds when $\mathbf{u} = \nabla f^*(\boldsymbol{\theta})$. That is, $f^*(\boldsymbol{\theta}) = \max_{\mathbf{u}}\langle \mathbf{u}, \boldsymbol{\theta}\rangle - f(\mathbf{u}) = \langle \nabla f^*(\boldsymbol{\theta}), \boldsymbol{\theta}\rangle - f(\nabla f^*(\boldsymbol{\theta})) \implies \nabla f^*(\boldsymbol{\theta}) = \arg\max_{\mathbf{w}}(\langle \mathbf{w}, \boldsymbol{\theta}\rangle - f(\mathbf{w}))$. Amy: seems important. make a lemma! Amy: prove via Danskin's theorem in two lines

2.7.3 Analyzing OMD using Duality

Recall that the OMD rule is

$$\mathbf{w}_t = g(\boldsymbol{\theta}_t) = g(-\mathbf{z}_{1:t-1}),$$

where the link function g is

$$g(\boldsymbol{\theta}) = \arg\max_{\mathbf{w}}(\langle \mathbf{w}, \boldsymbol{\theta}\rangle - R(\mathbf{w})).$$

Based on eq. (17), we can also rewrite $g(\boldsymbol{\theta}) = \nabla R^*(\boldsymbol{\theta})$.

Lemma 2.37. *Suppose that OMD is run with a link function $g = \nabla R^*$. Then, its regret is upper bounded by*

$$\sum_{t=1}^{T}\langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t\rangle \leq R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} \boldsymbol{D}_{R^*}(-\mathbf{z}_{1:t}|| - \mathbf{z}_{1:t-1}).$$

Furthermore, equality holds for the vector $\mathbf{u}$ that minimizes $R(\mathbf{u}) + \sum_t\langle \mathbf{u}, \mathbf{z}_t\rangle$.

Proof. Amy: please insert something about the LHS of the inequality to start or the end of the proof. First, using Fenchel-Young inequality (eq. (15)), we have for all $\mathbf{u}$,

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

$$R(\mathbf{u}) + \sum_{t=1}^{T}\langle \mathbf{u}, \mathbf{z}_t\rangle = R(\mathbf{u}) - \langle \mathbf{u}, -\mathbf{z}_{1:T}\rangle \geq -R^*(-\mathbf{z}_{1:T}),$$

where equality holds for the vector $\mathbf{u}$ that maximizes $\langle \mathbf{u}, -\mathbf{z}_{1:T}\rangle - R(\mathbf{u})$, hence minimizes $R(\mathbf{u}) + \langle \mathbf{u}, \mathbf{z}_{1:T}\rangle$.

Second, using the fact that $\mathbf{w}_t = g(-\mathbf{z}_{1:t-1}) = \nabla R^*(-\mathbf{z}_{1:t-1})$ and the definition of the Bregman divergence, we can rewrite the RHS as

$$\begin{aligned} -R^*(-\mathbf{z}_{1:T}) &= -R^*(0) - \sum_{t=1}^{T}(R^*(-\mathbf{z}_{1:t}) - R^*(-\mathbf{z}_{1:t-1})) && \text{the sum term equals } R^*(-\mathbf{z}_{1:T}) - R^*(0) \\ &= -R^*(0) - \sum_{t=1}^{T}(D_{R^*}(-\mathbf{z}_{1:t}|| - \mathbf{z}_{1:t-1}) - \langle \mathbf{w}_t, \mathbf{z}_t\rangle) \\ &= -R^*(0) + \sum_{t=1}^{T}(\langle \mathbf{w}_t, \mathbf{z}_t\rangle - D_{R^*}(-\mathbf{z}_{1:t}|| - \mathbf{z}_{1:t-1})) && (18) \end{aligned}$$

(Since $D_{R^*}(-\mathbf{z}_{1:t}||-\mathbf{z}_{1:t-1}) = R^*(-\mathbf{z}_{1:t}) - R^*(-\mathbf{z}_{1:t-1}) + \langle \nabla R^*(-\mathbf{z}_{1:t-1}, -\mathbf{z}_t\rangle = R^*(-\mathbf{z}_{1:t}) - R^*(-\mathbf{z}_{1:t-1}) - \langle \mathbf{w}_t, \mathbf{z}_t\rangle$.)
Note that $R^*(0) = \max_{\mathbf{w}} \langle 0, \mathbf{w}\rangle - R(\mathbf{w}) = -\min_{\mathbf{w}} R(\mathbf{w}) = -R(\mathbf{w}_1)$. Combining all the above we conclude the proof. □

It is interesting to compare the above lemma to lemma 2.10, which for linear functions yields the regret bound

$$\sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t\rangle \leq R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{w}_{t+1}, \mathbf{z}_t\rangle.$$

We can easily derive concrete bounds from lemma 2.37 if R is strongly convex.

Theorem 2.38. *Let R be a $\frac{1}{\eta}$-strongly convex function with respect to a norm $\|\cdot\|$ and suppose the OMD algorithm is run with the link function $g = \nabla R^*$. Then,*

$$\sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t\rangle \leq R(\mathbf{u}) - R(\mathbf{w}_1) + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{z}_t\|_*^2 .$$

Proof. Since R is $\frac{1}{\eta}$-strongly convex, by lemma 2.36, we know that R^* is η-strongly smooth with respect to norm $\|\cdot\|_*$.

Thus, by definition of strongly smooth, we have

$$D_{R^*}(-\mathbf{z}_{1:t}||-\mathbf{z}_{1:t-1}) \leq \frac{\eta}{2} \|-\mathbf{z}_{1:t} - (-\mathbf{z}_{1:t-1})\|_*^2 = \frac{\eta}{2} \|\mathbf{z}_t\|_*^2 .$$

Therefore, we can conclude that

$$\sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t\rangle \leq R(\mathbf{u}) - R(\mathbf{w}_1) + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{z}_t\|_*^2 .$$

□

2.7.4 Other Proof Technique

In the previous section, we used Fenchel-Young inequality to derive bounds for OMD, which is equivalent to FoReL when the loss functions are linear. It is possible to extent this proof technique, based on Fenchel duality, and to derive a larger family of online convex optimization algorithms. Ths basic idea is to derive the Fenchel dual of the optimization problem $\min_{\mathbf{u}} R(\mathbf{u}) + \sum_t f_t(\mathbf{u})$, and to construct an online learning algorithm by incrementally solving the dual problem.

Another popular approach is to derive regret bounds by monitoring the Bregman divergence $D_R(\mathbf{w}_t||\mathbf{u})$ where $\mathbf{u}$ is the competing vector. It is important to note that the analysis using the Bregman divergence potential requires that R will be a Legendre function. In particular, Legendre functions guarantee the property that ∇R and ∇R^* are inverse mappings.

To the best of our knowledge, the two proof techniques lead to the same regret bounds.

2.8 Bounds with Local Norms

Consider running the normalized EG algorithm, namely, running FoReL on linear loss function with the normalized entropy $R(\mathbf{w}) = \frac{1}{\eta}\sum_i w[i]\log(w[i]) + I_S(\mathbf{w})$, where $S = \{\mathbf{w} \geq 0 : \|\mathbf{w}\|_1 = 1\}$. Previously, we have derived the regret bound:

$$\sum_{t=1}^{T}\langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t\rangle \leq \frac{\log(d)}{\eta} + \eta\sum_{t=1}^{T}\|\mathbf{z}_t\|_\infty^2 .$$

Note that the dual norm of $\|\cdot\|_1$ is $\|\cdot\|_\infty$.

We now derive a refined bound for the normalized EG algorithm, in which each term $\|\mathbf{z}_t\|_\infty^2$ is replaced by a term $\sum_i w_t[i]z_t[i]^2$. Since $\mathbf{w}_t$ is in the probability simplex, we clearly have that $\sum_i w_t[i]z_t[i]^2 \leq \|\mathbf{z}_t\|_\infty^2$. (Note that $\|\mathbf{z}_t\|_\infty = \max_i |z_t[i]|$.)

In fact, we can rewrite $\sum_i w_t[i]z_t[i]^2$ as a local norm $\|\mathbf{z}_t\|_t^2$ where

$$\|\mathbf{z}\|_t = \sqrt{\sum_i w_t[i]z[i]^2}.$$

Theorem 2.39. *Assume that the normalized EG algorithm is run on a sequence of linear loss functions sych that fir all* t, i, *we have* $\eta z_t[i] \geq -1$. *Then,*

$$\sum_{t=1}^{T}\langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t\rangle \leq \frac{\log(d)}{\eta} + \eta\sum_{t=1}^{T}\sum_i w_t[i]z_t[i]^2.$$

Proof. Using lemma 2.37, it suffices to show that

$$D_{R^*}(\mathbf{z}_{1:t}|| - \mathbf{z}_{1:t-1}) \leq \eta\sum_i w_t[i]z_t[i]^2,$$

where, based on section 2.7.1, the conjugate function is

$$R^*(\boldsymbol{\theta}) = \frac{1}{\eta}\log\left(\sum_i e^{\eta\theta[i]}\right).$$

Indeed,

$$\begin{aligned} D_{R^*}(\mathbf{z}_{1:t}|| - \mathbf{z}_{1:t-1}) &= R^*(-\mathbf{z}_{1:t}) - R^*(-\mathbf{z}_{1:t-1}) + \langle \mathbf{w}_t, \mathbf{z}_t\rangle && (19)\\ &= \frac{1}{\eta}\log\left(\frac{\sum_i e^{-\eta z_{1:t}[i]}}{\sum_i e^{-\eta z_{1:t-1}[i]}}\right) + \langle \mathbf{w}_t, \mathbf{z}_t\rangle && (20)\\ &= \frac{1}{\eta}\log\left(\sum_i w_t[i]e^{-{}_t[i]}\right) + \langle \mathbf{w}_t, \mathbf{z}_t\rangle. && (21) \end{aligned}$$

Using the inequality $e^{-a} \leq 1 - a + a^2$ which holds for all $a \leq -1$ (and hence holds by the assumptions of the theorem) we obtain

$$D_{R^*}(\mathbf{z}_{1:t} || - \mathbf{z}_{1:t-1}) \leq \frac{1}{\eta} \log \left(\sum_i w_t[i](1 - \eta z_t[i] + \eta^2 z_t[i]^2) \right) + \langle \mathbf{w}_t, \mathbf{z}_t \rangle.$$

Next, we use the fact that $\sum_i w_t[i] = 1$ and the inequality $log(1-a) \leq -a$, which holds for all $a \leq 1$, we obtain

$$\begin{aligned} D_{R^*}(\mathbf{z}_{1:t} || - \mathbf{z}_{1:t-1}) &\leq \frac{1}{\eta} \log \left(\sum_i w_t[i](-\eta z_t[i] + \eta^2 z_t[i]^2) \right) + \langle \mathbf{w}_t, \mathbf{z}_t \rangle \\ &= \eta \sum_i w_t[i] z_t[i]^2. \end{aligned}$$

□

2.9 Online Convex Optimization Algorithms Summary

See the table on next page.

Convex Optimization Algorithms				
Name	**Update Rule(General)**	**Conditions**	**update Rule(Specific)**	**Regret Bound**
Follow-the-Leader	$\forall t, \mathbf{w}_t =$ $\arg\min_{\mathbf{w}\in S} \sum_{i=1}^{t-1} f_i(\mathbf{w})$	**(Online Quadratic Optimization)** $S = \mathbb{R}^d$, $f_t(\mathbf{w}) = \frac{1}{2}\lVert\mathbf{w}-\mathbf{z}_t\rVert_2^2$ for some $\mathbf{z}_t$. Let $L = \max_t \lVert\mathbf{z}_t\rVert$.	None	$4L^2(\log(T)+1)$
Follow-the-Regularized-Leader	$\forall t, \mathbf{w}_t =$ $\arg\min_{\mathbf{w}\in S} \sum_{i=1}^{t-1} f_i(\mathbf{w})$ $+R(\mathbf{w})$	**(Online Linear Optimization)** $S = \mathbb{R}^d$, $f_t(\mathbf{w}) = \langle\mathbf{w},\mathbf{z}_t\rangle$ for some $\mathbf{z}_t$, $R(\mathbf{w}) = \frac{1}{2\eta}\lVert\mathbf{w}\rVert_2^2$. Let $B = \max\lVert\mathbf{u}\rVert$, and let L be such that $\frac{1}{T}\sum_{t=1}^{T}\lVert\mathbf{z}_t\rVert_2^2 \leq L^2$. Set $\eta = \frac{B}{L\sqrt{2T}}$.	None	$BL\sqrt{2T}$
		(Online Gradient Descent) f_t is convex and L_t-Lipschitz with respect to $\lVert\cdot\rVert_2$, $R(\mathbf{w}) = \frac{1}{2\eta}\lVert\mathbf{w}\rVert_2^2$. Let $B = \max\lVert\mathbf{u}\rVert_2$, and let L be such that $\frac{1}{T}\sum_{t=1}^{T} L_t^2 \leq L^2$. Set $\eta = \frac{B}{L\sqrt{2T}}$.	$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\mathbf{z}_t$ where $\eta \in \partial f_t(\mathbf{w}_t)$	$BL\sqrt{2T}$
		(Prediction with Expert Advice I) S is a convex set, f_t is convex and L_t-Lipschitz with respect to $\lVert\cdot\rVert_2$, $R(\mathbf{w}) = \begin{cases} \frac{1}{2\eta}\lVert\mathbf{w}\rVert_2^2 & \mathbf{w}\in S \\ \infty & \mathbf{w}\notin S \end{cases}.$ Let $B > \max\lVert\mathbf{u}\rVert_2$, and let L be such that $\frac{1}{T}\sum_{t=1}^{T} L_t^2 \leq L^2$. Set $\eta = \frac{B}{L\sqrt{2T}}$.	None	$BL\sqrt{2T}$. In this case, $B = 1$, $L = \sqrt{d}$, so regret bound is $\sqrt{2dT}$.
		(Prediction with Expert Advice II) $S = \{\mathbf{w} : \lVert\mathbf{w}\rVert_1 = B \wedge \mathbf{w} > 0\}$ $\subset \mathbb{R}^d$ is a convex set, f_t is convex and L_t-Lipschitz with respect to $\lVert\cdot\rVert_1$, $R(\mathbf{w}) = \frac{1}{\eta}\sum_i w[i]\log(w[i])$. Let L be such that $\frac{1}{T}\sum_{t=1}^{T} L_t^2 \leq L^2$. Set $\eta = \frac{\sqrt{\log(d)}}{L\sqrt{2T}}$.	None	$BL\sqrt{2\log(d)T}$. In this case, $B = 1$, $L = 1$, so regret bound is $\sqrt{2\log(d)T}$.

Convex Optimization Algorithms (Cont.)				
Name	**Update Rule(General)**	**Conditions**	**update Rule(Specific)**	**Regret Bound**
Online Mirror Descent	predict $\mathbf{w}_t = g(\boldsymbol{\theta}_t)$ update $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{z}_t$ where $\mathbf{z}_t \in \nabla f_t(\mathbf{w}_t)$.	(Normalized Exponentiated Gradient) $S = \{\mathbf{w} : \|\mathbf{w}\|_1 = 1 \wedge \mathbf{w} > 0\}$ is the probability simplex, $g_i(\boldsymbol{\theta}) = \frac{e^{\eta\theta[i]}}{\sum_j e^{\eta\theta[j]}}$, $R(\mathbf{w}) = \frac{1}{\eta}\sum_i w[i]\log(w[i])$. Let L be such that $\frac{1}{T}\sum_{t=1}^T L_t^2 \leq L^2$. Set $\eta = \frac{\sqrt{\log(d)}}{L\sqrt{2T}}$.	$\forall i, w_{t+1}[i] = \frac{w_t[i]e^{-\eta z_t[j]}}{\sum_j w_t[j]e^{-\eta z_t[j]}}$ where $\mathbf{z}_t \in \partial f_t(\mathbf{w}_t)$.	$\sqrt{2\log(d)T}$
		(Online Gradient Descent with Lazy Projection) S is a convex set, $g(\boldsymbol{\theta}) = \arg\min_{\mathbf{w}\in S} \|\mathbf{w} - \eta\boldsymbol{\theta}\|_2$. $R(\mathbf{w}) = \frac{1}{2\eta}\|\mathbf{w}\|_2^2$, Let $B = \max \|\mathbf{u}\|_2$, and let L be such that $\frac{1}{T}\sum_{t=1}^T L_t^2 \leq L^2$. Set $\eta = \frac{B}{L\sqrt{2T}}$.	$\mathbf{w}_t = \arg\min_{\mathbf{w}\in S}$ $\|\mathbf{w} - \eta\boldsymbol{\theta}_t\|_2$, $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{z}_t$ where $\mathbf{z}_t \in \partial f_t(\mathbf{w}_t)$.	$BL\sqrt{2T}$
		(p-norm) $S = \mathbb{R}^d$, $g_i(\boldsymbol{\theta}) = \eta\frac{\text{sign}(\theta[i])\|\theta[i]\|^{p-1}}{\|\boldsymbol{\theta}\|_p^{p-2}}$, $R(\mathbf{w}) = \frac{1}{2\eta(q-1)}\|\mathbf{w}\|_q^2$ $(q = \frac{p}{p-1})$. Let $B = \max \|\mathbf{u}\|_q$, and let L be such that $\frac{1}{T}\sum_{t=1}^T L_t^2 \leq L^2$. Set $\eta = \frac{B}{L\sqrt{2T/(q-1)}}$. Note: when $q = 2$, this becomes OGD	$\forall u$, $w_{t,i} = \eta\frac{\text{sign}(\theta[i])\|\theta[i]\|^{p-1}}{\|\boldsymbol{\theta}\|_p^{p-2}}$, $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{z}_t$ where $\mathbf{z}_t \in \partial f_t(\mathbf{w}_t)$.	$BL\sqrt{\frac{2T}{q-1}}$
		(Applying duality idea) $g(\boldsymbol{\theta}) = \nabla R^*$, $R(\mathbf{w}) = \frac{1}{2\eta}\|\mathbf{w}\|_2^2$. Let $B = \max \|\mathbf{u}\|_2$, and let L be such that $\frac{1}{T}\sum_{t=1}^T L_t^2 \leq L^2$. Set $\eta = \frac{B}{L\sqrt{T}}$.	None	$BL\sqrt{T}$