Relax and Localize: From Value to Algorithms

Alexander Rakhlin
University of Pennsylvania

Ohad Shamir
Microsoft Research

Karthik Sridharan
University of Pennsylvania

May 4, 2012

Abstract

We show a principled way of deriving online learning algorithms from a minimax analysis. Various upper bounds on the minimax value, previously thought to be non-constructive, are shown to yield algorithms. This allows us to seamlessly recover known methods and to derive new ones. Our framework also captures such "unorthodox" methods as Follow the Perturbed Leader and the R^2 forecaster. We emphasize that understanding the inherent complexity of the learning problem leads to the development of algorithms.

We define *local* sequential Rademacher complexities and associated algorithms that allow us to obtain faster rates in online learning, similarly to statistical learning theory. Based on these localized complexities we build a general adaptive method that can take advantage of the suboptimality of the observed sequence.

We present a number of new algorithms, including a family of randomized methods that use the idea of a "random playout". Several new versions of the Follow-the-Perturbed-Leader algorithms are presented, as well as methods based on the Littlestone's dimension, efficient methods for matrix completion with trace norm, and algorithms for the problems of transductive learning and prediction with static experts.

1 Introduction

This paper studies the online learning framework, where the goal of the player is to incur small regret while observing a sequence of data on which we place no distributional assumptions. Within this framework, many algorithms have been developed over the past two decades, and we refer to the book of Cesa-Bianchi and Lugosi [7] for a comprehensive treatment of the subject. More recently, a non-algorithmic minimax approach has been developed to study the *inherent complexities* of sequential problems [2, 1, 16, 21]. In particular, it was shown that a theory in parallel to Statistical Learning can be developed, with random averages, combinatorial parameters, covering numbers, and other measures of complexity. Just as the classical learning theory is concerned with the study of the supremum of empirical or Rademacher process, online learning is concerned with the study of the supremum of a martingale or a certain dyadic process. Even though complexity tools introduced in [16, 18, 17] provide ways of studying the minimax value, no algorithms have been exhibited to achieve these non-constructive bounds in general.

In this paper, we show that algorithms can, in fact, be extracted from the minimax analysis. This observation leads to a unifying view of many of the methods known in the literature, and also gives a general recipe for developing new algorithms. We show that the potential method, which has been studied in various forms, naturally arises from the study of the minimax value as a certain *relaxation*. We further show that the sequential complexity tools introduced in [16] are, in fact, relaxations and can be used for constructing algorithms that enjoy the corresponding bounds. By choosing appropriate relaxations, we recover many known methods, improved variants of some known methods, and new algorithms. One can view our framework as one for converting a non-constructive proof of an upper bound on the value of the game into an

algorithm. Surprisingly, this allows us to also study such "unorthodox" methods as Follow the Perturbed Leader [12], and the recent method of [9] under the same umbrella with others. We show that the idea of a random playout has a solid theoretical basis, and that Follow the Perturbed Leader algorithm is an example of such a method. It turns out that whenever sequential Rademacher complexity is of the same order as its i.i.d. cousin, there is a family of randomized methods that avoid certain computational hurdles. Based on these developments, we exhibit an efficient method for the trace norm matrix completion problem, novel Follow the Perturbed Leader algorithms, and efficient methods for the problems of transductive learning and prediction with static experts.

The framework of this paper gives a recipe for developing algorithms. Throughout the paper, we stress that the notion of a relaxation, introduced below, is not appearing out of thin air but rather as an upper bound on the sequential Rademacher complexity. The understanding of *inherent complexity* thus leads to the *development of algorithms*.

One unsatisfying aspect of the minimax developments so far has been the lack of a *localized* analysis. Local Rademacher averages have been shown to play a key role in Statistical Learning for obtaining fast rates. It is also well-known that fast rates are possible in online learning, on the case-by-case basis, such as for online optimization of strongly convex functions. We show that, in fact, a localized analysis can be performed at an abstract level, and it goes hand-in-hand with the idea of relaxations. Using such a localized analysis, we arrive at *local sequential Rademacher* and other local complexities. These complexities upper-bound the value of the online learning game and can lead to fast rates. What is equally important, we provide an associated generic algorithm to achieve the localized bounds. We further develop the ideas of localization, presenting a general adaptive (data-dependent) procedure that takes advantage of the actual moves of the adversary that might have been suboptimal. We illustrate the procedure on a few examples. Our study of localized complexities and adaptive methods follows from a general agenda of developing universal methods that can adapt to the actual sequence of data played by Nature, thus automatically interpolating between benign and minimax optimal sequences.

This paper is organized as follows. In Section 2 we formulate the value of the online learning problem and present the (possibly computationally inefficient) minimax algorithm. In Section 3 we develop the idea of relaxations and the meta algorithm based on relaxations, and present a few examples. Section 4 is devoted to a new formalism of localized complexities, and we present a basic localized meta algorithm. We show, in particular, that for strongly convex objectives, the regret is easily bounded through localization. Next, in Section 5, we present a fully adaptive method that constantly checks whether the sequence being played by the adversary is in fact minimax optimal. We show that, in particular, we recover some of the known adaptive results. We also demonstrate how local data-dependent norms arise as a natural adaptive method. The remaining sections present a number of new algorithms, often with superior computational properties and regret guarantees than what is known in the literature.

Notation: A set $\{x_1, \dots, x_t\}$ is often denoted by $x_{1:t}$. A t-fold product of $\mathcal{X}$ is denoted by $\mathcal{X}^t$. Expectation with respect to a random variable Z with distribution p is denoted by $\mathbb{E}_Z$ or $\mathbb{E}_{Z\sim p}$. The set $\{1, \dots, T\}$ is denoted by $[T]$, and the set of all distributions on some set $\mathcal{A}$ by $\Delta(\mathcal{A})$. The inner product between two vectors is written as $\langle a, b\rangle$ or as $a^\intercal b$. The set of all functions from $\mathcal{X}$ to $\mathcal{Y}$ is denoted by $\mathcal{Y}^\mathcal{X}$. Unless specified otherwise, ϵ denotes a vector $(\epsilon_1, \dots, \epsilon_T)$ of i.i.d. Rademacher random variables. An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is defined as a sequence $(\mathbf{x}_1, \dots, \mathbf{x}_d)$ of mappings $\mathbf{x}_t : \{\pm 1\}^{t-1} \mapsto \mathcal{X}$ (see [16]). We often write $\mathbf{x}_t(\epsilon)$ instead of $\mathbf{x}_t(\epsilon_{1:t-1})$.

2 Value and The Minimax Algorithm

Let $\mathcal{F}$ be the set of learner's moves and $\mathcal{X}$ the set of moves of Nature. The online protocol dictates that on every round $t = 1, \dots, T$ the learner and Nature simultaneously choose $f_t \in \mathcal{F}$, $x_t \in \mathcal{X}$, and observe each

other's actions. The learner aims to minimize regret

$$\mathbf{Reg}_T \triangleq \sum_{t=1}^{T} \ell(f_t, x_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f, x_t)$$

where $\ell : \mathcal{F} \times \mathcal{X} \to \mathbb{R}$ is a known loss function. Our aim is to study this online learning problem at an abstract level without assuming convexity or other properties of the loss function and the sets $\mathcal{F}$ and $\mathcal{X}$. We do assume, however, that ℓ, $\mathcal{F}$, and $\mathcal{X}$ are such that the minimax theorem in the space of distributions over $\mathcal{F}$ and $\mathcal{X}$ holds. By studying the abstract setting, we are able to develop general algorithmic and non-algorithmic ideas that are common across various application areas.

The starting point of our development is the minimax value of the associated online learning game:

$$\mathcal{V}_T(\mathcal{F}) = \inf_{q_1\in\Delta(\mathcal{F})} \sup_{x_1\in\mathcal{X}} \mathop{\mathbb{E}}_{f_1\sim q_1} \cdots \inf_{q_T\in\Delta(\mathcal{F})} \sup_{x_T\in\mathcal{X}} \mathop{\mathbb{E}}_{f_T\sim q_T} \left[\sum_{t=1}^{T} \ell(f_t, x_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f, x_t)\right] \tag{1}$$

where $\Delta(\mathcal{F})$ is the set of distributions on $\mathcal{F}$. The minimax formulation immediately gives rise to the optimal algorithm that solves the minimax expression at every round t. That is, after witnessing $x_1, \ldots, x_{t-1}$ and $f_1, \ldots, f_{t-1}$, the algorithm returns

$$\begin{aligned} &\operatorname*{argmin}_{q\in\Delta(\mathcal{F})} \left\{ \sup_{x_t} \mathop{\mathbb{E}}_{f_t\sim q} \inf_{q_{t+1}} \sup_{x_{t+1}} \mathop{\mathbb{E}}_{f_{t+1}} \cdots \inf_{q_T} \sup_{x_T} \mathop{\mathbb{E}}_{f_T} \left[\sum_{i=t}^{T} \ell(f_i, x_i) - \inf_{f\in\mathcal{F}} \sum_{i=1}^{T} \ell(f, x_i)\right]\right\} \\ &= \operatorname*{argmin}_{q\in\Delta(\mathcal{F})} \left\{ \sup_{x_t} \mathop{\mathbb{E}}_{f_t\sim q} \left[\ell(f_t, x_t) + \inf_{q_{t+1}} \sup_{x_{t+1}} \mathop{\mathbb{E}}_{f_{t+1}} \cdots \inf_{q_T} \sup_{x_T} \mathop{\mathbb{E}}_{f_T} \left[\sum_{i=t+1}^{T} \ell(f_i, x_i) - \inf_{f\in\mathcal{F}} \sum_{i=1}^{T} \ell(f, x_i)\right]\right]\right\} \end{aligned} \tag{2}$$

Henceforth, if the quantification in inf and sup is omitted, it will be understood that x_t, f_t, p_t, q_t range over $\mathcal{X}$, $\mathcal{F}$, $\Delta(\mathcal{X})$, $\Delta(\mathcal{F})$, respectively. Moreover, $\mathbb{E}_{x_t}$ is with respect to p_t while $\mathbb{E}_{f_t}$ is with respect to q_t. The first sum in (2) starts at $i = t$ since the partial loss $\sum_{i=1}^{t-1} \ell(f_i, x_i)$ has been fixed. We now notice a recursive form for defining the value of the game. Define for any $t \in [T-1]$ and any given prefix $x_1, \ldots, x_t \in \mathcal{X}$ the *conditional value*

$$\mathcal{V}_T\left(\mathcal{F}|x_1, \ldots, x_t\right) \triangleq \inf_{q\in\Delta(\mathcal{F})} \sup_{x\in\mathcal{X}} \left\{ \mathop{\mathbb{E}}_{f\sim q} \left[\ell(f, x)\right] + \mathcal{V}_T(\mathcal{F}|x_1, \ldots, x_t, x)\right\}$$

where

$$\mathcal{V}_T\left(\mathcal{F}|x_1, \ldots, x_T\right) \triangleq -\inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f, x_t) \quad \text{and} \quad \mathcal{V}_T(\mathcal{F}) = \mathcal{V}_T(\mathcal{F}|\{\}).$$

The minimax optimal algorithm specifying the mixed strategy of the player can be written succinctly

$$q_t = \operatorname*{argmin}_{q\in\Delta(\mathcal{F})} \sup_{x\in\mathcal{X}} \left\{ \mathop{\mathbb{E}}_{f\sim q} \left[\ell(f, x)\right] + \mathcal{V}_T(\mathcal{F}|x_1, \ldots, x_{t-1}, x)\right\}. \tag{3}$$

This recursive formulation has appeared in the literature, but now we have tools to study the conditional value of the game. We will show that various upper bounds on $\mathcal{V}_T(\mathcal{F}|x_1, \ldots, x_{t-1}, x)$ yield an array of algorithms, some with better computational properties than others. In this way, the non-constructive approach of [16, 17, 18] to upper bound the value of the game directly translates into algorithms.

The minimax algorithm in (3) can be interpreted as choosing the best decision that takes into account the present loss and the worst-case future. We then realize that the conditional value of the game serves as a "regularizer", and thus well-known online learning algorithms such as Exponential Weights, Mirror Descent and Follow-the-Regularized-Leader arise as relaxations rather than a "method that just works".

The first step is to appeal to the minimax theorem and perform the same manipulation as in [1, 16], but only on the value from $t+1$ onwards:

$$\mathcal{V}_T\left(\mathcal{F}|x_1, \ldots, x_t\right) = \sup_{p_{t+1}} \mathop{\mathbb{E}}_{x_{t+1}} \cdots \sup_{p_T} \mathop{\mathbb{E}}_{x_T} \left[\sum_{i=t+1}^{T} \inf_{f_i\in\mathcal{F}} \mathop{\mathbb{E}}_{x_i\sim p_i} \ell(f_i, x_i) - \inf_{f\in\mathcal{F}} \sum_{i=1}^{T} \ell(f, x_i)\right] \tag{4}$$

This expression is still unwieldy, and the idea is now to come up with more manageable, yet tight, upper bounds of the conditional value.

3 Relaxations and the Basic Meta-Algorithm

A *relaxation* $\mathbf{Rel}\,()$ is a sequence of real-valued functions $\mathbf{Rel}_T\,(\mathcal{F}|x_1,\ldots,x_t)$ for each $t \in [T]$. We shall use the notation $\mathbf{Rel}_T\,(\mathcal{F})$ for $\mathbf{Rel}_T\,(\mathcal{F}|\{\})$. A relaxation will be called *admissible* if for any $x_1,\ldots,x_T \in \mathcal{X}$,

$$\mathbf{Rel}_T\,(\mathcal{F}|x_1,\ldots,x_t) \geq \inf_{q\in\Delta(\mathcal{F})} \sup_{x\in\mathcal{X}} \left\{ \mathop{\mathbb{E}}_{f\sim q}\left[\ell(f,x)\right] + \mathbf{Rel}_T\,(\mathcal{F}|x_1,\ldots,x_t,x) \right\} \tag{5}$$

for all $t \in [T-1]$, and

$$\mathbf{Rel}_T\,(\mathcal{F}|x_1,\ldots,x_T) \geq -\inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f,x_t).$$

A strategy q that minimizes the expression in (5) defines an optimal algorithm for the relaxation $\mathbf{Rel}\,()$. This algorithm is given below under the name "Meta-Algorithm". However, minimization need not be exact: any q that satisfies the admissibility condition (5) is a valid method, and we will say that such an algorithm is *admissible with respect to the relaxation* $\mathbf{Rel}\,()$.

Algorithm 1 Meta-Algorithm **MetAlgo**

Parameters: Admissible relaxation **Rel**
for $t = 1$ to T **do**
  $q_t = \arg\min_{q\in\Delta(\mathcal{F})} \sup_{x\in\mathcal{X}} \left\{ \mathbb{E}_{f\sim q}\left[\ell(f,x)\right] + \mathbf{Rel}_T\,(\mathcal{F}|x_1,\ldots,x_{t-1},x) \right\}$
  Play $f_t \sim q_t$ and receive x_t from adversary
end for

Proposition 1. *Let* $\mathbf{Rel}\,()$ *be an admissible relaxation. For any admissible algorithm with respect to* $\mathbf{Rel}\,()$*, including the* Meta-Algorithm*, irrespective of the strategy of the adversary,*

$$\sum_{t=1}^{T} \mathbb{E}_{f_t\sim q_t} \ell(f_t,x_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f,x_t) \leq \mathbf{Rel}_T\,(\mathcal{F}) \ , \tag{6}$$

and therefore,

$$\mathbb{E}[\mathbf{Reg}_T] \leq \mathbf{Rel}_T\,(\mathcal{F}) \ .$$

We also have that

$$\mathcal{V}_T(\mathcal{F}) \leq \mathbf{Rel}_T\,(\mathcal{F}) \ .$$

If $a \leq \ell(f,x) \leq b$ *for all* $f \in \mathcal{F}, x \in \mathcal{X}$*, the Hoeffding-Azuma inequality yields, with probability at least* $1-\delta$*,*

$$\mathbf{Reg}_T = \sum_{t=1}^{T} \ell(f_t,x_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f,x_t) \leq \mathbf{Rel}_T\,(\mathcal{F}) + (b-a)\sqrt{T/2\cdot\log(2/\delta)} \ .$$

Further, if for all $t \in [T]$*, the admissible strategies* q_t *are deterministic,*

$$\mathbf{Reg}_T \leq \mathbf{Rel}_T\,(\mathcal{F}) \ .$$

The reader might recognize **Rel** as a potential function. It is known that one can derive regret bounds by coming up with a potential such that the current loss of the player is related to the difference in the potentials at successive steps, and that the loss of the best decision in hindsight can be extracted from the final potential. The origin of "good" potential functions has always been a mystery (at least to the authors). One of the conceptual contributions of this paper is to show that they naturally arise as relaxations on the conditional value. The conditional value itself can be characterized as the tightest possible relaxation.

In particular, for many problems a tight relaxation (sometimes within a factor of 2) is achieved through symmetrization applied to the expression in (4). Define the *conditional Sequential Rademacher complexity*

$$\mathfrak{R}_T(\mathcal{F}|x_1,\dots,x_t) = \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_{t+1:T}} \sup_{f\in\mathcal{F}} \left[2\sum_{s=t+1}^{T} \epsilon_s \ell(f, \mathbf{x}_{s-t}(\epsilon_{t+1:s-1})) - \sum_{s=1}^{t} \ell(f, x_s) \right] . \tag{7}$$

Here the supremum is over all $\mathcal{X}$-valued binary trees of depth $T-t$. One may view this complexity as a partially symmetrized version of the sequential Rademacher complexity

$$\mathfrak{R}_T(\mathcal{F}) \triangleq \mathfrak{R}_T(\mathcal{F} \mid \{\}) = \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_{1:T}} \sup_{f\in\mathcal{F}} \left[2\sum_{s=1}^{T} \epsilon_s \ell(f, \mathbf{x}_s(\epsilon_{1:s-1})) \right] \tag{8}$$

defined in [16]. We shall refer to the term involving the tree $\mathbf{x}$ as the "future" and the term being subtracted off – as the "past". This indeed corresponds to the fact that the quantity is conditioned on the already observed $x_1,\dots,x_t$, while for the future we have the worst possible binary tree.[1]

Proposition 2. *The conditional Sequential Rademacher complexity is admissible.*

The proof of this proposition is given in the Appendix and it corresponds to one step of the sequential symmetrization proof in [16]. We note that the factor 2 appearing in (7) is not necessary in certain cases (e.g. binary prediction with absolute loss).

We now show that several well-known methods arise as further relaxations on the conditional sequential Rademacher complexity $\mathfrak{R}_T$.

Exponential Weights Suppose $\mathcal{F}$ is a finite class and $|\ell(f,x)| \le 1$. In this case, a (tight) upper bound on sequential Rademacher complexity leads to the following relaxation:

$$\mathbf{Rel}_T(\mathcal{F}|x_1,\dots,x_t) = \inf_{\lambda>0} \left\{ \frac{1}{\lambda} \log\left(\sum_{f\in\mathcal{F}} \exp\left(-\lambda \sum_{i=1}^{t} \ell(f,x_i) \right) \right) + 2\lambda(T-t) \right\} \tag{9}$$

Proposition 3. *The relaxation (9) is admissible and*

$$\mathfrak{R}_T(\mathcal{F}|x_1,\dots,x_t) \le \mathbf{Rel}_T(\mathcal{F}|x_1,\dots,x_t) .$$

Furthermore, it leads to a parameter-free version of the Exponential Weights algorithm, defined on round $t+1$ by the mixed strategy

$$q_{t+1}(f) \propto \exp\left(-\lambda_t^* \sum_{s=1}^{t} \ell(f,x_s) \right)$$

with λ_t^* the optimal value in (9). The algorithm's regret is bounded by

$$\mathbf{Rel}_T(\mathcal{F}) \le 2\sqrt{2T\log|\mathcal{F}|} .$$

[1] It is somewhat cumbersome to write out the indices on $\mathbf{x}_{s-t}(\epsilon_{t+1:s-1})$ in (7), so we will instead use $\mathbf{x}_s(\epsilon)$ for $s = 1,\dots,T-t$, whenever this does not cause confusion.

The Chernoff-Cramèr inequality tells us that (9) is the tightest possible relaxation. The proof of Proposition 3 reveals that the only inequality is the softmax which is also present in the proof of the maximal inequality for a finite collection of random variables. In this way, exponential weights is an algorithmic realization of a maximal inequality for a finite collection of random variables. The connection between probabilistic (or concentration) inequalities and algorithms runs much deeper.

We point out that the exponential-weights algorithm arising from the relaxation (9) is a *parameter-free* algorithm. The learning rate λ^* can be optimized (via one-dimensional line search) at each iteration with almost no cost. This can lead to improved performance as compared to the classical methods that set a particular schedule for the learning rate.

Mirror Descent In the setting of online linear optimization, the loss is $\ell(f,x) = \langle f, x\rangle$. Suppose $\mathcal{F}$ is a unit ball in some Banach space and $\mathcal{X}$ is the dual. Let $\|\cdot\|$ be some $(2,C)$-smooth norm on $\mathcal{X}$ (in the Euclidean case, $C = 2$). Using the notation $\tilde{x}_{t-1} = \sum_{s=1}^{t-1} x_s$, a straightforward upper bound on sequential Rademacher complexity is the following relaxation:

$$\mathbf{Rel}_T\left(\mathcal{F}|x_1,\dots,x_t\right) = \sqrt{\|\tilde{x}_{t-1}\|^2 + \left\langle \nabla \tfrac{1}{2}\|\tilde{x}_{t-1}\|^2, x_t\right\rangle + C(T-t+1)} \tag{10}$$

Proposition 4. *The relaxation* (10) *is admissible and*

$$\mathfrak{R}_T(\mathcal{F}|x_1,\dots,x_t) \le \mathbf{Rel}_T\left(\mathcal{F}|x_1,\dots,x_t\right) \ .$$

Furthermore, it leads to the Mirror Descent algorithm with regret at most $\mathbf{Rel}_T\left(\mathcal{F}\right) \le \sqrt{2CT}$.

An important feature of the algorithms we just proposed is the absence of any parameters, as the step size is tuned automatically. We had chosen Exponential Weights and Mirror Descent for illustration because these methods are well-known. Our aim at this point was to show that the associated relaxations arise naturally (typically with a few steps of algebra) from the sequential Rademacher complexity. More examples are included later in the paper. It should now be clear that upper bounds, such as the Dudley Entropy integral, can be turned into a relaxation, provided that admissibility is proved. Our ideas have semblance of those in Statistics, where an information-theoretic complexity can be used for defining penalization methods.

4 Localized Complexities and the Localized-Meta Algorithm

The localized analysis plays an important role in Statistical Learning Theory. The basic idea is that better rates can be proved for empirical risk minimization when one considers the empirical process in the vicinity of the target hypothesis [13, 4]. Through this, localization gives *extra information* by shrinking the size of the set which needs to be analyzed. What does it mean to localize in online learning? As we obtain more data, we can rule out parts of $\mathcal{F}$ as those that are unlikely to become the leaders. This observation indeed gives rise to faster rates. Let us develop a general framework of localization and then illustrate it on examples. We emphasize that the localization ideas will be developed at an abstract level where no assumptions are placed on the loss function or the sets $\mathcal{F}$ and $\mathcal{X}$.

Given any $x_1,\dots,x_t \in \mathcal{X}$, for any $k \ge 1$ define

$$\mathcal{F}^k(x_1,\dots,x_t) = \left\{ f \in \mathcal{F} : \exists\ x_{t+1},\dots,x_{t+k} \in \mathcal{X} \ \text{ s.t. } \ \sum_{i=1}^{t+k} \ell(f,x_i) = \inf_{f\in\mathcal{F}} \sum_{i=1}^{t+k} \ell(f,x_i) \right\} \ .$$

That is, given the instances $x_1,\dots,x_t$, the set $\mathcal{F}^k(x_1,\dots,x_t)$ is the set of elements that could be the minimizers of cumulative loss on $t+k$ instances, the first t of which are $x_1,\dots,x_t$ and the remaining k arbitrary. We shall refer to minimizers of cumulative loss as *empirical risk minimizers* (or, ERM).

Henceforth, we shall use the notation $\tilde{k}_j \triangleq \sum_{i=1}^{j} k_i$. We now consider subdividing T into blocks of time $k_1, \ldots, k_m \in [T]$ such that $\tilde{k}_m = T$. With this notation, $\tilde{k}_i$ is the last time in the ith block. We then have regret upper bounded as

$$\begin{aligned}\sum_{t=1}^{T} \ell(f_t, x_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f, x_t) &\le \sum_{t=1}^{T} \ell(f_t, x_t) - \sum_{i=1}^{m} \inf_{f\in\mathcal{F}^{k_i}\left(x_1,\ldots,x_{\tilde{k}_{i-1}}\right)} \sum_{t=\tilde{k}_{i-1}+1}^{\tilde{k}_i} \ell(f, x_t) \\ &= \sum_{i=1}^{m} \left(\sum_{t=\tilde{k}_{i-1}+1}^{\tilde{k}_i} \ell(f, x_t) - \inf_{f\in\mathcal{F}^{k_i}\left(x_1,\ldots,x_{\tilde{k}_{i-1}}\right)} \sum_{t=\tilde{k}_{i-1}+1}^{\tilde{k}_i} \ell(f, x_t) \right) \\ &= \sum_{i=1}^{m} \mathbf{Reg}_{k_i}\left(x_{\tilde{k}_{i-1}+1}, \ldots, x_{\tilde{k}_i}, f_{\tilde{k}_{i-1}+1}, \ldots, f_{\tilde{k}_i}, \mathcal{F}^{k_i}(x_1, \ldots, x_{\tilde{k}_{i-1}})\right)\end{aligned} \tag{11}$$

The short inductive proof of inequality (11) is given in Appendix, Lemma 26.

Hence, one can decompose the online learning game into blocks of m successive games. The crucial point to notice is that at the i^{th} block, we do not compete with the best hypothesis in all of $\mathcal{F}$ but rather only $\mathcal{F}^{k_i}(x_1, \ldots, x_{\tilde{k}_{i-1}})$. Hence, if we only consider strategies that only pick from the set $\mathcal{F}^{k_i}(x_1, \ldots, x_{\tilde{k}_{i-1}})$ when playing in the corresponding block i, we only weaken the player. As a consequence, we have that

$$\mathcal{V}_T(\mathcal{F}) \le \sum_{i=1}^{m} \mathcal{V}_T\left(\mathcal{F}^{k_i}(x_1, \ldots, x_{\tilde{k}_{i-1}})\right) . \tag{12}$$

It is this localization based on history that could lead to possibly faster rates. While the "blocking" idea often appears in the literature (for instance, in the form of a doubling trick, as described below), the process is usually "restarted" from scratch by considering all of $\mathcal{F}$. Notice further that one need not choose all $k_1, \ldots, k_m$ in advance. The player can choose k_i based on history $x_1, \ldots, x_{\tilde{k}_{i-1}}$ and then use, for instance, the Meta-Algorithm introduced in the previous section to play the game within block k_i using the localized class $\mathcal{F}^{k_i}(x_1, \ldots, x_{\tilde{k}_{i-1}})$. Such adaptive procedures will be considered in Section 5, but presently we assume that the block sizes $k_1, \ldots, k_m$ are fixed.

While the successive localizations using subsets $\mathcal{F}^{k_i}(x_1, \ldots, x_{\tilde{k}_{i-1}})$ can provide an algorithm with possibly better performance, specifying and analyzing the localized subset $\mathcal{F}^{k_i}(x_1, \ldots, x_{\tilde{k}_{i-1}})$ exactly might not be possible. In such a case, one can instead use

$$\mathcal{F}_r(x_1, \ldots, x_{\tilde{k}_{i-1}}) = \left\{ f \in \mathcal{F} : P\left(f \mid x_1, \ldots, x_{\tilde{k}_{i-1}}\right) \le r \right\}$$

where P is some "property" of f given data. This definition echoes the definition of the set of r-minimizers of empirical or expected risk in Statistical Learning. Further, for a given k define

$$r(k; x_1, \ldots, x_t) = \inf\{r : \mathcal{F}^k(x_1, \ldots, x_t) \subset \mathcal{F}_r(x_1, \ldots, x_t)\}$$

the smallest "radius" such that $\mathcal{F}_r$ includes the set of potential minimizers over the next k time steps. Of course, if the property P does not enforce localization, the bounds are not going to exhibit any improvement, so P needs to be chosen carefully for a particular problem of interest. We have the following algorithm:

Algorithm 2 Localized Meta-Algorithm

Parameters : Relaxation **Rel**
Initialize $t = 0$ and blocks $k_1, \ldots, k_m$ s.t. $\sum_{i=1}^{m} k_i = T$
for $i = 1$ to m **do**
  Play k_i rounds using $\mathbf{MetAlgo}\left(\mathcal{F}_{r(k_i; x_1, \ldots, x_t)}\right)$ and set $t = t + k_i$
end for

Lemma 5. *The regret of the Localized Meta-Algorithm is bounded as*

$$\mathbf{Reg}_T(x_1,\dots,x_T) \le \sum_{i=1}^{m} \mathbf{Rel}_{k_i}\left(\mathcal{F}_{r\left(k_i;x_1,\dots,x_{\tilde{k}_{i-1}}\right)}\right)$$

Note that the above lemma points to local sequential complexities for online learning problems that can lead to possibly fast rates. In particular, if sequential Rademacher complexity is used as the relaxation in the Localized Meta-Algorithm, we get a bound in terms of *local sequential Rademacher complexities*.

Local Sequential Complexities

The following corollary is a direct consequence of Lemma 5.

Corollary 6 (Local Sequential Rademacher Complexity)**.** *For any property P and any $k_1,\dots,k_m \in \mathbb{N}$ such that $\sum_{i=1}^m k_i = T$, we have that :*

$$\mathcal{V}_T(\mathcal{F}) \le \sup_{x_1,\dots,x_T} \sum_{i=1}^{m} \mathfrak{R}_{k_i}\left(\mathcal{F}_{r\left(k_i;x_1,\dots,x_{\tilde{k}_{i-1}}\right)}\right)$$

Clearly, the sequential Rademacher complexities in the above bound can be replaced with other sequential complexity measures of the localized classes that are upper bounds on the sequential Rademacher complexities. For instance, one can replace each Rademacher complexity $\mathfrak{R}_{k_i}$ by covering number based bounds of the local classes, such as the analogues of the Dudley Entropy Integral bounds developed in the sequential setting in [16]. Once can also use, for instance, fat-shattering dimension based complexity measures for these local classes.

Example : Doubling trick

The doubling trick can be seen as a particular blocking strategy with $k_i = 2^{i-1}$ so that

$$\mathbf{Reg}_T(x_1,\dots,x_T) \le \sum_{i=1}^{\lceil \log_2 T\rceil+1} \mathbf{Rel}_{2^{i-1}}(\mathcal{F})$$

Now if **Rel** is such that for any t, $\mathbf{Rel}_t(\mathcal{F}) \le t^p$ for some p then the regret is upper bounded by $\frac{T^p-2^{-p}}{1-2^{-p}}$. The main advantage of the doubling trick is of course that we do not need to know T in advance.

Example : Strongly Convex Loss

To illustrate the idea of localization, consider online convex optimization with 1-Lipschitz λ-strongly convex functions $x_t : \mathcal{F} \mapsto \mathbb{R}$ (that is, $\ell(f,x) = x(f)$). Define

$$\mathbf{Rel}_T(\mathcal{F}|x_1,\dots,x_t) = -\inf_{f\in\mathcal{F}} \sum_{i=1}^{t} x_i(f) + (T-t) \inf_{f\in\mathcal{F}} \sup_{f'\in\mathcal{F}} \|f-f'\|$$

An easy Lemma 27 in the Appendix shows that this relaxation is admissible. Notice that this relaxation grows linearly with block size and is by itself quite bad. However, with blocking and localization, the relaxation gives an optimal bound for strongly convex objectives. To see this note that for $k = 1$, any minimizer of $\sum_{i=1}^{t+1} x_i(f)$ has to be close to the minimizer $\hat{f}_t$ of $\sum_{i=1}^{t} x_i(f)$, due to strong convexity of the functions. In other words, the property

$$P(f|x_1,\dots,x_t) = \|f-\hat{f}_t\|$$

with $r_t = 1/(\lambda t)$ entails

$$\mathcal{F}^1(x_1,\dots,x_t) \subseteq \left\{f \in \mathcal{F} : \|f - \hat{f}_t\| \le 1/(\lambda t)\right\} = \mathcal{F}_{r_t}(x_1,\dots,x_t).$$

The relaxation for the block of size $k = 1$ is

$$\mathbf{Rel}_1\left(\mathcal{F}_{r_t}(x_1,\dots,x_t)\right) \le \inf_{f\in\mathcal{F}_{r_t}(x_1,\dots,x_t)} \sup_{f'\in\mathcal{F}_{r_t}(x_1,\dots,x_t)} \|f - f'\|,$$

the radius of the smallest ball containing the localized set $\mathcal{F}_{r_t}(x_1,\dots,x_t)$, and we immediately get

$$\mathbf{Reg}_T(x_1,\dots,x_T) \le \sum_{t=1}^{T} 1/(\lambda t) \le (1+\log(T))/\lambda \ .$$

We remark that this proof is different in spirit from the usual proofs of fast rates for strongly convex functions, and it demonstrates the power of localization.

5 Adaptive Procedures

There is a strong interest in developing methods that enjoy worst-case regret guarantees but also take advantage of the suboptimality of the sequence being played by Nature. An algorithm that is able to do so without knowing in advance that the sequence will have a certain property will be called *adaptive*. Imagine, for instance, running an experts algorithm, and one of the experts has gained such a lead that she is clearly the winner (that is, the empirical risk minimizer) at the end of the game. In this case, since we are to be compared with the leader at the end, we need not focus on anyone else, and regret for the remainder of the game is zero.

There has been previous work on exploiting particular ways in which sequences can be suboptimal. Examples include the Adaptive Gradient Descent of [5], Adaptive Hedge of [22], and the variance-based bounds of [8, 11] among others. We now give a generic method which incorporates the idea of localization in order to adaptively (and constantly) check whether the sequence being played is of optimal or suboptimal nature. Notice that, as before, we present the algorithm at the abstract level of the online game with some decision sets $\mathcal{F}$, $\mathcal{X}$, and some loss function ℓ.

The adaptive procedure below uses a subroutine $\mathbf{Block}(\{x_1,\dots,x_t\},\tau)$ which, given the history $\{x_1,\dots,x_t\}$, returns a subdivision of the next τ rounds into sub-blocks. The choice of the blocking strategy has to be made for the particular problem at hand, but, as we show in examples, one can often use very simple strategies.

Let us describe the adaptive procedure. First, for simplicity of exposition, we start with the doubling-size blocks. Here is what happens within each of these blocks. During each round the learner decides whether to stay in the same sub-block or to start a new one, as given by the blocking procedure **Block**. If started, the new sub-block uses the localized subset given the history of adversary's moves up until last round. Choosing to start a new sub-block corresponds to the realization of the learner that the sequence being presented so far is in fact suboptimal. The learner then incorporates this suboptimality into the localized procedure.

Lemma 7. *Given some admissible relaxation* **Rel**, *the regret of the adaptive localized meta-algorithm (Algorithm 3) is bounded as*

$$\mathbf{Reg}_T \le \sum_{i=1}^{\mathtt{nbl}} \mathbf{Rel}_{k_i^*}\left(\mathcal{F}_{r\left(k_i^*;x_1,\dots,x_{\tilde{k}_{i-1}^*}\right)}\right)$$

where `nbl` *is the number of blocks actually played and* k_i^**'s are adaptive block lengths defined within the algorithm. Further, irrespective of the blocking strategy* **Block** *used, if the relaxation* **Rel** *is such that for any* t, $\mathbf{Rel}_t(\mathcal{F}) \le t^p$ *for some* $p \in (0,1]$, *then the worst case regret is always bounded as*

$$\mathbf{Reg}_T \le (T^p - 2^{-p})/(1-2^{-p}) \ .$$

Algorithm 3 Adaptive Localized Meta-Algorithm

Parameters : Relaxation **Rel** and block size calculator **Block**.
Initialize $t = 1$ and $\texttt{nbl} = 1$, and suppose $T = 2^c - 1$ for some $c \geq 2$.
for $i = 1$ to c **do**
  $G = \mathbf{Rel}_{2^i}\left(\mathcal{F}_r(2^i; x_1, \ldots, x_{t-1})\right)$ `% guaranteed value of relaxation`
  $m = 1, \texttt{curr} = 1$ and $K_1 = 2^i$
  while $\texttt{curr} \leq 2^i$ and $t \leq T$ **do**
    $(\kappa_1, \ldots, \kappa_{m'}) = \mathbf{Block}\left(\{x_1, \ldots, x_t\}, 2^i - \texttt{curr}\right)$ `% blocking for remainder of` 2^i
    if $G > \sup_{x_{t+1}, \ldots, x_{2^{i+1}-1}} \sum_{j=1}^{m'} \mathbf{Rel}_{\kappa_j}\left(\mathcal{F}_{r(\kappa_j; x_1, \ldots, x_{t+\tilde{\kappa}_{j-1}})}\right)$ **then**
      $k^*_{\texttt{nbl}} = \kappa_1$, $K = (\kappa_2, \ldots, \kappa_{m'})$, $m = m' - 1$ `% if better value, accept new blocking`
    else
      $k^*_{\texttt{nbl}} = K_1$, $K = (K_2, \ldots, K_m)$, $m = m - 1$ `% else continue with current blocking`
    end if
    Play $k^*_{\texttt{nbl}}$ rounds using $\mathbf{MetAlgo}(\mathcal{F}_{r(k^*_{\texttt{nbl}}; x_1, \ldots, x_t)})$
    $t = t + k^*_{\texttt{nbl}}$, $\texttt{curr} = \texttt{curr} + k^*_{\texttt{nbl}}$, $\texttt{nbl} = \texttt{nbl} + 1$
    Let

$$G = \sup_{x_{t+1}, \ldots, x_{2^{i+1}-1}} \sum_{j=1}^{m} \mathbf{Rel}_{K_j}\left(\mathcal{F}_{r(K_j; x_1, \ldots, x_{t+\sum_{i=1}^{j-1} K_i})}\right)$$

  end while
end for

We now demonstrate that the adaptive algorithm in fact takes advantage of sub-optimality in several situations that have been previously studied in the literature. On the conceptual level, adaptive localization allows us to view several fast rate results under the same umbrella.

Example: Adaptive Gradient Descent Consider the online convex optimization scenario. Following the setup of [5], suppose the learner encounters a sequence of convex functions x_t with the strong convexity parameter σ_t, potentially zero, with respect to a $(2, C)$-smooth norm $\|\cdot\|$. The goal is to adapt to the actual sequence of functions presented by the adversary. Let us invoke the Adaptive Localized Meta-Algorithm with a rather simple blocking strategy

$$\mathbf{Block}\left(\{x_1, \ldots, x_t\}, k\right) = \begin{cases} (k) & \text{if } \sqrt{k} > \tilde{\sigma}_t \\ (1, 1, \ldots, 1) & \text{otherwise} \end{cases}$$

where $\tilde{\sigma}_t = \sum_{s=1}^{t} \sigma_s$. This blocking strategy either says "use all of the next k rounds as one block", or "make each of the next k time step into separate blocks". Let $\hat{f}_t$ be the empirical minimizer at the start of the block (that is after t rounds), and let $y_t = \nabla x_t(f_t)$. Then we can use the localization

$$\mathcal{F}_{r(k; x_1, \ldots, x_t)} = \left\{f \in \mathcal{F} : \|f - \hat{f}_t\| \leq 2 \min\left\{1, k/\tilde{\sigma}_t\right\}\right\}$$

and relaxation

$$\mathbf{Rel}_k\left(\mathcal{F}_{r(k; x_1, \ldots, x_t)} \middle| y_1, \ldots, y_i\right) = -\left\langle \hat{f}_t, \tilde{y}_i \right\rangle + 2 \min\left\{1, k/\tilde{\sigma}_t\right\}\left(\|\tilde{y}_{i-1}\|^2 + \left\langle \nabla \tfrac{1}{2}\|\tilde{y}_{i-1}\|^2, y_i \right\rangle + C(k - i + 1)\right)^{1/2}$$

where $\tilde{y}_{i-1} = \sum_{j=1}^{i-1} y_j$. For the above relaxation we can show that the corresponding update at round $t + i$ is given by

$$f_{t+i} = \hat{f}_t - \max\left\{1, \frac{k}{\tilde{\sigma}_t}\right\} \frac{-\nabla \frac{1}{2}\|\bar{x}_{i-1}\|^2}{\sqrt{\|\bar{x}_{i-1}\|^2 + C(k - i + 1)}}$$

where k is the length of the current block. The next lemma shows that the proposed adaptive gradient descent recovers the results of [5]. The method is a mixture of Follow the Leader -style algorithm and a Gradient Descent -style algorithm.

Lemma 8. *The relaxation specified above is admissible. Suppose the adversary plays 1-Lipchitz convex functions $x_1, \dots, x_T$ such that for any $t \in [T]$, $\sum_{i=1}^{t} x_i$ is $\tilde{\sigma}_t$-strongly convex, and further suppose that for some $B \le 1$, we have that $\tilde{\sigma}_t = Bt^{\alpha}$. Then, for the blocking strategy specified above,*

1. *If $\alpha \le 1/2$ then $\mathbf{Reg}_T \le O\left(\sqrt{T}\right)$*
2. *If $1 > \alpha > 1/2$ then $\mathbf{Reg}_T \le O(\frac{T^{1-\alpha}}{B})$*
3. *If $\alpha = 1$ then $\mathbf{Reg}_T \le O\left(\frac{\log T}{B}\right)$*

Example: Adaptive Experts We now turn to the setting of Adaptive Hedge or Exponential Weights algorithm similar to the one studied in [22]. Consider the following situation: for all time steps after some τ, there is an element (or, expert) f that is the best by a margin k over the next-best choice in $\mathcal{F}$ in terms of the (unnormalized) cumulative loss, and it remains to be the winner until the end. Let us use the localization

$$\mathcal{F}_{r(k;x_1,\dots,x_t)} = \left\{ f \in \mathcal{F} \ : \ \sum_{i=1}^{t} \ell(f, x_i) - \min_{f \in \mathcal{F}} \sum_{i=1}^{t} \ell(f, x_i) \le k \right\} \ ,$$

the set of functions closer than the margin to the ERM. Let

$$\hat{\mathcal{F}}_t = \left\{ f \in \mathcal{F} \ : \ \sum_{i=1}^{t} \ell(f, x_i) = \min_{f \in \mathcal{F}} \sum_{i=1}^{t} \ell(f, x_i) \right\}$$

be the set of empirical minimizers at time t. We use the blocking strategy

$$\mathbf{Block}(\{x_1, \dots, x_t\}, k) = (j, k-j) \quad \text{where} \quad j = \left\lfloor \min_{f \notin \hat{\mathcal{F}}_t} \sum_{i=1}^{t} \ell(f, x_i) - \min_{f \in \hat{\mathcal{F}}_t} \sum_{i=1}^{t} \ell(f, x_i) \right\rfloor \tag{13}$$

which says that the size of the next block is given by the gap between empirical minimizer(s) and non-minimizers. The idea behind the proof and the blocking strategy is simple. If it happens at the start a new block that there is a large gap between the current leader and the next expert, then for the number of rounds approximately equal to this gap we can play a new block and not suffer any extra regret.

Consider the relaxation (9) used for the Exponential Weights algorithm.

Lemma 9. *Suppose that there exists a single best expert*

$$\hat{f}_T = \arg\min_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f, x_t),$$

and that for some $k \ge 1$ there exists $\tau \in [T]$ such that for all $t > \tau$ and all $f \ne \hat{f}_T$ the partial cumulative loss

$$\sum_{i=1}^{t} \ell(f, x_i) - \sum_{i=1}^{t} \ell(\hat{f}_T, x_i) \ge k \ .$$

Then the regret of Algorithm 3 with the Exponential Weights relaxation, the blocking strategy (13) and the localization mentioned above is bounded as

$$\mathbf{Reg}_T \le 4 \min\left\{ \tau, \sqrt{\tau \log(|\mathcal{F}|)} \right\}$$

While we demonstrated a very simple example, the algorithm is adaptive more generally. Lemma 9 considers the assumption that a single expert becomes a clear winner after τ rounds, with margin of k. Even when there is no clear winner throughout the game, we can still achieve low regret. For instance, this happens if only a few elements of $\mathcal{F}$ have low cumulative loss throughout the game and the rest of $\mathcal{F}$ suffers heavy loss. Then the algorithm adapts to the suboptimality and gives regret bound with the dominating term depending logarithmically only on the cardinality of the "good" choices in the set $\mathcal{F}$. Similar ideas appear in [10], and will be investigated in more generality in the full version of the paper.

Example: Adapting to the Data Norm Recall that the set $\mathcal{F}^k(x_1,\ldots,x_t)$ is the subset of functions in $\mathcal{F}$ that are possible empirical risk minimizers when we consider $x_1,\ldots,x_{t+k}$ for some $x_{t+1},\ldots,x_{t+k}$ that can occur in the future. Now, given history $x_1,\ldots,x_t$ and a possible future sequence $x_{t+1},\ldots,x_{t+k}$, if $\hat{f}_{t+k}$ is an ERM for $x_1,\ldots,x_{t+k}$ and $\hat{f}_t$ is an ERM for $x_1,\ldots,x_t$ then

$$\begin{aligned}\sum_{i=1}^{t}\ell(\hat{f}_{t+k},x_i)-\sum_{i=1}^{t}\ell(\hat{f}_t,x_i) &= \sum_{i=1}^{t+k}\ell(\hat{f}_{t+k},x_i)-\sum_{i=1}^{t+k}\ell(\hat{f}_t,x_i)+\sum_{i=t+1}^{t+k}\ell(\hat{f}_t,x_i)-\sum_{i=t+1}^{t+k}\ell(\hat{f}_{t+k},x_i)\\ &\le 0+\sup_{x_{t+1},\ldots,x_{t+k}}\left\{\sum_{i=t+1}^{t+k}\ell(\hat{f}_t,x_i)-\sum_{i=t+1}^{t+k}\ell(\hat{f}_{t+k},x_i)\right\}.\end{aligned}$$

Hence, we see that it suffices to consider localizations

$$\mathcal{F}_{r(k;x_1,\ldots,x_t)}=\left\{f\in\mathcal{F}\ :\ \sum_{i=1}^{t}\ell(f,x_i)-\sum_{i=1}^{t}\ell(\hat{f}_t,x_i)\le\sup_{x_{t+1},\ldots,x_{t+k}}\left\{\sum_{i=t+1}^{t+k}\ell(\hat{f}_t,x_i)-\sum_{i=t+1}^{t+k}\ell(f,x_i)\right\}\right\}.$$

If we consider online convex Lipschitz learning problems where $\mathcal{F}=\{f:\|f\|\le 1\}$ and loss is convex in f and is such that $\|\nabla\ell(f,x)\|_*\le 1$ in the dual norm $\|\cdot\|_*$, using the above argument we can use localization

$$\mathcal{F}_{r(k;x_1,\ldots,x_t)}=\left\{f\in\mathcal{F}\ :\ \sum_{i=1}^{t}\ell(f,x_i)-\sum_{i=1}^{t}\ell(\hat{f}_t,x_i)\le k\left\|f-\hat{f}_t\right\|\right\}. \tag{14}$$

Further, using Taylor approximation we can pass to the localization

$$\mathcal{F}_{r(k;x_1,\ldots,x_t)}=\left\{f\in\mathcal{F}\ :\ \frac{1}{2}\left\|f-\hat{f}_t\right\|^2_{x_1,\ldots,x_T}\le k\left\|f-\hat{f}_t\right\|\right\} \tag{15}$$

where $\|f\|^2_{x_1,\ldots,x_T}=f^\top H_t f$, and H_t is the Hessian of the function $g(f)=\sum_{i=1}^t\ell(f,x_i)$. Notice that the earlier example where we adapt to strong convexity of the loss is a special case of the above localization where we lower bound the data-dependent norm (Hessian-based norm) by the ℓ_2 norm times the smallest eigenvalue. If for instance we are faced with η-exp-concave losses, such as the squared loss, the data-dependent norm can be again lower bounded by

$$\|f\|^2_{x_1,\ldots,x_T}\ge\eta f^\top\left(\sum_{i=1}^{t}\nabla_i\right)\left(\sum_{i=1}^{t}\nabla_i\right)^\top f$$

and so we can use localization based on outer products of sum of gradients. We then do not "pay" for those directions in which the adversary has not played, thus adapting to the *effective dimension* of the sequence of plays.

In general, for online convex optimization problems one can use localizations given in Equations (14) or (15). The localization in Equation (14) is applicable even in the linear setting, and if it so happens that the adversary mainly plays in a one dimensional sub-space, then the algorithm automatically adapts to the adversary and yields faster rates for regret. As already mentioned, the example of adaptive gradient descent is a special case of localization in Equation (15). Of course, one needs to provide also an appropriate blocking strategy. A possible general blocking strategy could be :

$$\textbf{Block}(\{x_1,\ldots,x_t\},k)=(j,k-j),\quad\text{where}\quad j=\operatorname*{argmin}_{j\in\{0,\ldots,k\}}\left\{\textbf{Rel}_j\left(\mathcal{F}_{r(x_1,\ldots,x_t)}\right)+\sup_{x_{t+1},\ldots,x_{t+j}}\textbf{Rel}_{k-j}\left(\mathcal{F}_{r(x_1,\ldots,x_{t+k})}\right)\right\}.$$

In the remainder of the paper, we develop new algorithms to show the universality of our approach. One could try to argue that the introduction of the notion of a relaxation has not alleviated the burden of algorithm development, as we simply pushed the work into magically coming up with a relaxation. We would like to stress that this is not so. A key observation is that a relaxation does not appear out of thin air, but rather as an upper bound on the sequential Rademacher complexity. Thus, a general recipe is to start with a problem at hand and develop a sequence of upper bounds until one obtains a computationally feasible one, or until other desired properties are satisfied. Exactly for this purpose, the proofs in the appendix *derive* the relaxations rather than just present them as something given. Since one would follow the same upper bounding steps to prove an upper bound on the value of the game, *the derivation of the relaxation and the proof of the regret bound go hand-in-hand*. For this reason, we sometimes omit the explicit mention of a regret bound for the sake of conciseness: the algorithms enjoy the same regret bound as that obtained by the corresponding non-constructive proof of the upper bound.

6 Classification

We start by considering the problem of supervised learning, where $\mathcal{X}$ is the space of instances and $\mathcal{Y}$ the space of responses (labels). There are two closely related protocols for the online interaction between the learner and Nature, so let us outline them. The "proper" version of supervised learning follows the protocol presented in Section 2: at time t, the learner selects $f_t \in \mathcal{F}$, Nature simultaneously selects $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$, and the learner suffers the loss $\ell(f(x_t), y_t)$. The "improper" version is as follows: at time t, Nature chooses $x_t \in \mathcal{X}$ and presents it to the learner as "side information", the learner then picks $\hat{y}_t \in \mathcal{Y}$ and Nature simultaneously chooses $y_t \in \mathcal{Y}$. In the improper version, the loss of the learner is $\ell(\hat{y}_t, y_t)$, and it is easy to see that we may equivalently state this protocol as the learner choosing any function $f_t \in \mathcal{Y}^{\mathcal{X}}$ (not necessarily in $\mathcal{F}$), and Nature simultaneously choosing (x_t, y_t). We mostly focus on the "improper" version of supervised learning, as the distinction does not make any difference in any of the bounds.

For the improper version of supervised learning, we may write the value in (1) as

$$\mathcal{V}_T(\mathcal{F}) = \sup_{x_1 \in \mathcal{X}} \inf_{q_1 \in \Delta(\mathcal{Y})} \sup_{y_1 \in \mathcal{X}} \mathbb{E}_{\hat{y}_1 \sim q_1} \dots \sup_{x_T \in \mathcal{X}} \inf_{q_T \in \Delta(\mathcal{Y})} \sup_{y_T \in \mathcal{X}} \mathbb{E}_{\hat{y}_T \sim q_T} \left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) \right]$$

and a relaxation $\mathbf{Rel}\,()$ is admissible if for any $(x_1, y_1) \dots, (x_T, y_T) \in \mathcal{X} \times \mathcal{Y}$,

$$\sup_{x \in \mathcal{X}} \inf_{q \in \Delta(\mathcal{Y})} \sup_{y \in \mathcal{Y}} \left\{ \mathop{\mathbb{E}}_{\hat{y} \sim q} \ell(\hat{y}, y) + \mathbf{Rel}_T\left(\mathcal{F} \middle| \{(x_i, y_i)\}_{i=1}^{t}, (x, y)\right) \right\} \le \mathbf{Rel}_T\left(\mathcal{F} \middle| \{(x_i, y_i)\}_{i=1}^{t}\right) \tag{16}$$

and

$$\mathbf{Rel}_T\left(\mathcal{F} \middle| \{(x_i, y_i)\}_{i=1}^{T}\right) \ge - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t).$$

Let us now focus on binary label prediction, that is $\mathcal{Y} = \{\pm 1\}$. In this case, the supremum over y in (16) becomes a maximum over two values. Let us now take the absolute loss $\ell(\hat{y}, y) = |\hat{y} - y| = 1 - \hat{y}y$. We can see that the optimal randomized strategy, given the side information x, is given by (16) as

$$\operatorname*{argmin}_{q \in \Delta(\mathcal{Y})} \max \left\{ 1 - q + \mathbf{Rel}_T\left(\mathcal{F} \middle| \{(x_i, y_i)\}_{i=1}^{t}, (x, 1)\right), 1 + q + \mathbf{Rel}_T\left(\mathcal{F} \middle| \{(x_i, y_i)\}_{i=1}^{t}, (x, -1)\right) \right\}$$

which is achieved by setting the two expressions equal to each other:

$$q = \frac{1}{2}\left\{\mathbf{Rel}_T\left(\mathcal{F}\middle|\{(x_i,y_i)\}_{i=1}^t,(x,1)\right) - \mathbf{Rel}_T\left(\mathcal{F}\middle|\{(x_i,y_i)\}_{i=1}^t,(x,-1)\right)\right\} \tag{17}$$

This result will be specialized in the latter sections for particular relaxations $\mathbf{Rel}\,()$ and extended beyond absolute loss. We remark that the extension to k-class prediction is immediate and involves taking a maximum over k terms in (16).

6.1 Algorithms Based on the Littlestone's Dimension

Consider the problem of binary prediction, as described above. Further, assume that $\mathcal{F}$ has a finite Littlestone's dimension $\mathrm{Ldim}(\mathcal{F})$ [14, 6]. Suppose the loss function is $\ell(\hat{y},y) = |\hat{y}-y|$, and consider the "mixed" conditional Rademacher complexity

$$\sup_{\mathbf{x}} \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}}\left\{2\sum_{i=1}^{T-t}\epsilon_i f(\mathbf{x}_i(\epsilon)) - \sum_{i=1}^{t}|f(x_i)-y_i|\right\} \tag{18}$$

as a possible relaxation. Observe that the above complexity is defined with the loss function removed (in a contraction-style argument [16]) in the terms involving the "future", in contrast with the definition (7). The latter is defined with loss functions on both the "future" and the "past" terms. In general, if we can pass from the sequential Rademacher complexity over the loss class $\ell(\mathcal{F})$ to the sequential Rademacher complexity of the base class $\mathcal{F}$, we may attempt to do so step-by-step by using the "mixed" type of sequential Rademacher complexity as in (18). This idea shall be used several times later in this paper.

The admissibility condition (16) with the conditional sequential Rademacher (18) as a relaxation would require us to upper bound

$$\sup_{x_t}\inf_{q_t\in[-1,1]}\max_{y_t\in\{\pm1\}}\left\{\mathop{\mathbb{E}}_{\hat{y}_t\sim q_t}|\hat{y}_t-y_t| + \sup_{\mathbf{x}}\mathbb{E}_\epsilon\sup_{f\in\mathcal{F}}\left\{2\sum_{i=1}^{T-t}\epsilon_i f(\mathbf{x}_i(\epsilon)) - \sum_{i=1}^{t}|f(x_i)-y_i|\right\}\right\} \tag{19}$$

We observe that the supremum over $\mathbf{x}$ is preventing us from obtaining a concise algorithm. We need to further "relax" this supremum, and the idea is to pass to a finite cover of $\mathcal{F}$ on the given tree $\mathbf{x}$ and then proceed as in the Exponential Weights example for a finite collection of experts. This leads to an upper bound on (18) and gives rise to algorithms similar in spirit to those developed in [6], but with more attractive computational properties and defined more concisely.

Define the function $g(d,t) = \sum_{i=0}^{d}\binom{t}{i}$, which is shown in [16] to be the maximum size of an exact (zero) cover for a function class with the Littlestone's dimension $\mathrm{Ldim} = d$. Given $\{(x_1,y_t),\dots,(x_t,y_t)\}$ and $\sigma = (\sigma_1,\dots,\sigma_t)\in\{\pm1\}^t$, let

$$\mathcal{F}_t(\sigma) = \{f\in\mathcal{F} : f(x_i) = \sigma_i \quad \forall i\le t\},$$

the subset of functions that agree with the signs given by σ on the "past" data and let

$$\mathcal{F}|_{x_1,\dots,x_t} \triangleq \mathcal{F}|_{x^t} \triangleq \{(f(x_1),\dots,f(x_t)) : f\in\mathcal{F}\}$$

be the projection of $\mathcal{F}$ onto $x_1,\dots,x_t$. Denote $L_t(f) = \sum_{i=1}^t|f(x_i)-y_i|$ and $L_t(\sigma) = \sum_{i=1}^t|\sigma_i-y_i|$ for $\sigma\in\{\pm1\}^t$. The following proposition gives a relaxation and two algorithms, both of which achieve the $O(\sqrt{\mathrm{Ldim}(\mathcal{F})T\log T})$ regret bound proved in [6], yet both different from the algorithm in that paper.

Proposition 10. *The relaxation*

$$\mathbf{Rel}_T\left(\mathcal{F}\middle|(x^t,y^t)\right) = \frac{1}{\lambda}\log\left(\sum_{\sigma\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma)),T-t)\exp\{-\lambda L_t(\sigma)\}\right) + 2\lambda(T-t)\ .$$

is admissible and leads to an admissible algorithm

$$q_t(+1) = \frac{\sum_{(\sigma,+1)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,+1)), T-t)\exp\{-\lambda L_{t-1}(\sigma)\}}{\sum_{(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)), T-t)\exp\{-\lambda L_{t-1}(\sigma)\}}, \tag{20}$$

with $q_t(-1) = 1 - q_t(+1)$. *An alternative method for the same relaxation and the same regret guarantee is to predict the label* y_t *according to a distribution with mean*

$$q_t = \frac{1}{2\lambda}\log\frac{\sum_{(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)), T-t)\exp\{-\lambda L_{t-1}(\sigma)\}\exp\{-\lambda(1-\sigma_t)\}}{\sum_{(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)), T-t)\exp\{-\lambda L_{t-1}(\sigma)\}\exp\{-\lambda(1+\sigma_t)\}} \tag{21}$$

There is a very close correspondence between the proof of Proposition 10 and the proof of the combinatorial lemma of [16], the analogue of the Vapnik-Chervonenkis-Sauer-Shelah result.

The two algorithms presented above show two alternatives: one through employing the properties of exponential weights, and the other is through the solution in (17). The merits of the two approaches remain to be explored. In particular, it appears that the method based on (17) can lead to some non-trivial new algorithms, distinct from the more common exponential weighting technique.

7 Randomized Algorithms and Follow the Perturbed Leader

We now develop a class of admissible randomized methods that arise through sampling. Consider the objective

$$\inf_{q\in\Delta(\mathcal{F})}\sup_{x\in\mathcal{X}}\left\{\mathop{\mathbb{E}}_{f\sim q}[\ell(f,x)] + \mathbf{Rel}_T(\mathcal{F}|x_1,\ldots,x_{t-1},x)\right\}$$

given by a relaxation $\mathbf{Rel}(\,)$. If $\mathbf{Rel}(\,)$ is the sequential (or classical) Rademacher complexity, it involves an expectation over sequences of coin flips, and this computation (coupled with optimization for each sequence) can be prohibitively expensive. More generally, $\mathbf{Rel}(\,)$ might involve an expectation over possible ways in which the future might be realized. In such cases, we may consider a rather simple "random playout" strategy: draw the random sequence and solve only one optimization problem for that random sequence. The ideas of random playout have been discussed previously in the literature for estimating the utility of a move in a game (see also [3]). In this section we show that, in fact, the random playout strategy has a solid basis: for the examples we consider, it satisfies admissibility. Furthermore, we show that Follow the Perturbed Leader is an example of such a randomized strategy.

Let us informally describe the general idea, as the key steps might be hard to trace in the proofs. Suppose our objective is of the form

$$S(q) = \sup_x\left(\mathbb{E}_{f\sim q}\Psi(f,x) + \mathbb{E}_{w\sim p}\Phi(w,x)\right)$$

for some functions Ψ and Φ, and q a mixed strategy. We have in mind the situation where the first term is the instantaneous loss at the present round, and the second term is the expected cost for the future. Consider a randomized strategy $\tilde{q}$ which is defined by first randomly drawing $w\sim p$ and then computing

$$f(w) \triangleq \operatorname*{argmin}_f \sup_x(\Psi(f,x) + \Phi(w,x))$$

for the random draw w. We then verify that

$$\begin{aligned} S(\tilde{q}) &= \sup_x\left(\mathbb{E}_{f\sim\tilde{q}}\Psi(f,x) + \mathbb{E}_{w\sim p}\Phi(w,x)\right) = \sup_x\left(\mathbb{E}_{w\sim p}\Psi(f(w),x) + \mathbb{E}_{w\sim p}\Phi(w,x)\right) \\ &\le \mathbb{E}_{w\sim p}\sup_x\left(\Psi(f(w),x) + \Phi(w,x)\right) = \mathbb{E}_{w\sim p}\inf_f\sup_x\left(\Psi(f,x) + \Phi(w,x)\right)\ . \end{aligned}$$

What makes the proof of admissibility possible is that the infimum in the last expression is inside the expectation over w rather than outside. We can then appeal to the minimax theorem to prove admissibility.

In our examples, Ψ is the loss at round t and Φ is the relaxation term, such as the sequential Rademacher complexity. In Section 7.4 we show that, if we can compute the "worst" tree $\mathbf{x}$, we can randomly draw a path and use it for our randomized strategy. Note that the worst-case trees are closely related to random walks of maximal variation, and our method thus points to an intriguing connection between regret minimization and random walks (see also [3, 15] for related ideas).

Interestingly, in many learning problems it turns out that the sequential Rademacher complexity and the classical Rademacher complexity are within a constant factor of each other. In such cases, the function Φ does not involve the supremum over a tree, and the randomized method only needs to draw a sequence of coin flips and compute a solution to an optimization problem slightly more complicated than ERM.

In particular, the sequential and classical Rademacher complexities can be related for linear classes in finite-dimensional spaces. Online linear optimization is then a natural application of the randomized method we propose. Indeed, we show that Follow the Perturbed Leader (FPL) algorithm [12] arises in this way. We note that FPL has been previously considered as a rather unorthodox algorithm providing some kind of regularization via randomization. Our analysis shows that it arises through a natural relaxation based on the sequential (and thus the classical) Rademacher complexity, coupled with the random playout idea. As a new algorithmic contribution, we provide a version of the FPL algorithm for the case of the decision sets being ℓ_2 balls, with a regret bound that is *independent of the dimension*. We also provide an FPL-style method for the combination of ℓ_1 and ℓ_∞ balls. To the best of our knowledge, these results are novel.

In the later sections, we provide a novel randomized method for the Trace Norm Completion problem, and a novel randomized method for the setting of static experts and transductive learning. In general, the techniques we develop might in future provide computationally feasible randomized algorithms where deterministic ones are too computationally demanding.

7.1 When Sequential and Classical Rademacher Complexities are Related

The assumption below implies that the sequential Rademacher complexity and the classical Rademacher complexity are within constant factor C of each other. We will later verify that this assumption holds in the examples we consider.

Assumption 1. *There exists a distribution $D \in \Delta(\mathcal{X})$ and constant $C \geq 2$ such that for any $t \in [T]$ and given any $x_1, \ldots, x_{t-1}, x_{t+1}, \ldots, x_T \in \mathcal{X}$ and any $\epsilon_{t+1}, \ldots, \epsilon_T \in \{\pm 1\}$,*

$$\sup_{p \in \Delta(\mathcal{X})} \mathbb{E}_{x_t \sim p} \sup_{f \in \mathcal{F}} \left[C \sum_{i=t+1}^{T} \epsilon_i \ell(f, x_i) - L_{t-1}(f) + \mathbb{E}_{x \sim p} [\ell(f, x)] - \ell(f, x_t) \right]$$
$$\leq \mathbb{E}_{\epsilon_t, x_t \sim D} \sup_{f \in \mathcal{F}} \left[C \sum_{i=t}^{T} \epsilon_i \ell(f, x_i) - L_{t-1}(f) \right] \tag{22}$$

where ϵ_t is an independent Rademacher random variable and $L_{t-1}(f) = \sum_{i=1}^{t-1} \ell(f, x_i)$.

Under the above assumption one can use the following relaxation

$$\mathbf{Rel}_T(\mathcal{F} | x_1, \ldots, x_t) = \mathbb{E}_{x_{t+1}, \ldots x_T \sim D} \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left[C \sum_{i=t+1}^{T} \epsilon_i \ell(f, x_i) - \sum_{i=1}^{t} \ell(f, x_i) \right] \tag{23}$$

which is a partially symmetrized version of the classical Rademacher averages.

The proof of admissibility for the randomized methods based on this relaxation is quite curious – the forecaster can be seen as mimicking the sequential Rademacher complexity by sampling from the "equivalently bad" classical Rademacher complexity under the specific distribution D given by the above assumption.

Lemma 11. *Under Assumption 1, the relaxation in Eq. (23) is admissible and a randomized strategy that ensures admissibility is given by: at time t, draw $x_{t+1}, \ldots, x_T \sim D$ and Rademacher random variables $\epsilon = (\epsilon_{t+1}, \ldots, \epsilon_T)$ and then :*

1. *In the case the loss ℓ is convex in its first argument and the set $\mathcal{F}$ is convex and compact, define*

$$f_t = \operatorname*{argmin}_{g \in \mathcal{F}} \sup_{x \in \mathcal{X}} \left\{ \ell(g, x) + \sup_{f \in \mathcal{F}} \left\{ C \sum_{i=t+1}^{T} \epsilon_i \ell(f, x_i) - \sum_{i=1}^{t-1} \ell(f, x_i) - \ell(f, x) \right\} \right\} \tag{24}$$

2. *In the case of non-convex loss, sample f_t from the distribution*

$$\hat{q}_t = \operatorname*{argmin}_{\hat{q} \in \Delta(\mathcal{F})} \sup_{x \in \mathcal{X}} \left\{ \mathop{\mathbb{E}}_{f \sim \hat{q}} \left[\ell(f, x) \right] + \sup_{f \in \mathcal{F}} \left\{ C \sum_{i=t+1}^{T} \epsilon_i \ell(f, x_i) - \sum_{i=1}^{t-1} \ell(f, x_i) - \ell(f, x) \right\} \right\} \tag{25}$$

The expected regret for the method is bounded by the classical Rademacher complexity:

$$\mathbb{E}\left[\mathbf{Reg}_T\right] \le C \; \mathbb{E}_{x_{1:T} \sim D} \mathop{\mathbb{E}}_{\epsilon} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t \ell(f, x_t) \right],$$

Of particular interest are the settings of static experts and transductive learning, which we consider in Section 8. In the transductive case, the x_t's are pre-specified before the game, and in the static expert case – effectively absent. In these cases, as we show below, there is no explicit distribution D and we only need to sample the random signs ϵ's. We easily see that in these cases, the expected regret bound is simply two times the transductive Rademacher complexity.

7.2 Linear Loss

The idea of sampling from a fixed distribution is particularly appealing in the case of linear loss, $\ell(f, x) = \langle f, x \rangle$. Suppose $\mathcal{X}$ is a unit ball in some norm $\|\cdot\|$ in a vector space B, and $\mathcal{F}$ is a unit ball in the dual norm $\|\cdot\|_*$. Assumption 1 then becomes

Assumption 2. *There exists a distribution $D \in \Delta(\mathcal{X})$ and constant $C \ge 2$ such that for any $t \in [T]$ and given any $x_1, \ldots, x_{t-1}, x_{t+1}, \ldots, x_T \in \mathcal{X}$ and any $\epsilon_{t+1}, \ldots, \epsilon_T \in \{\pm 1\}$,*

$$\sup_{p \in \Delta(\mathcal{X})} \mathop{\mathbb{E}}_{x_t \sim p} \left\| C \sum_{i=t+1}^{T} \epsilon_i x_i - \sum_{i=1}^{t-1} x_i + \mathop{\mathbb{E}}_{x \sim p}\left[x\right] - x_t \right\| \le \mathop{\mathbb{E}}_{\epsilon_t, x_t \sim D} \left\| C \sum_{i=t}^{T} \epsilon_i x_i - \sum_{i=1}^{t-1} x_i \right\| \tag{26}$$

For (26) to hold it is enough to ensure that

$$\sup_{p \in \Delta(\mathcal{X})} \mathop{\mathbb{E}}_{x_t \sim p} \left\| w + \mathop{\mathbb{E}}_{x \sim p}\left[x\right] - x_t \right\| \le \mathop{\mathbb{E}}_{\epsilon_t, x_t \sim D} \left\| w + C \epsilon_t x_t \right\| \tag{27}$$

for any $w \in B$.

At round t, the generic algorithm specified by Lemma 25 draws fresh Rademacher random variables ϵ and $x_{t+1}, \ldots, x_T \sim D$ and picks

$$f_t = \operatorname*{argmin}_{f \in \mathcal{F}} \sup_{x \in \mathcal{X}} \left\{ \langle f, x \rangle + \left\| C \sum_{i=t+1}^{T} \epsilon_i x_i - \sum_{i=1}^{t-1} x_i - x \right\| \right\} \tag{28}$$

We now look at specific examples of ℓ_2/ℓ_2 and ℓ_1/ℓ_∞ cases and provide closed form solution of the randomized algorithms.

Example : ℓ_1/ℓ_∞ Follow the Perturbed Leader:

Here, we consider the setting similar to that in [12]. Let $\mathcal{F} \subset \mathbb{R}^N$ be the ℓ_1 unit ball and $\mathcal{X}$ the (dual) ℓ_∞ unit ball in $\mathbb{R}^N$. In [12], $\mathcal{F}$ is the probability simplex and $\mathcal{X} = [0,1]^N$ but these are subsumed by the ℓ_1/ℓ_∞ case. We claim that:

Lemma 12. *Assumption 2 is satisfied with a distribution D that is uniform on the vertices of the cube $\{\pm 1\}^N$ and $C = 6$.*

In fact, one can pick any symmetric distribution D on the real line and use D^N for the perturbation. Assumption 2 is then satisfied, as we show in the following lemma.

Lemma 13. *If D is any symmetric distribution over the real line, then Assumption 2 is satisfied by using the product distribution D^N. The constant C required is any $C \geq 6/\mathbb{E}_{x\sim D}|x|$.*

The above lemma is especially attractive when used with standard normal distribution because in that case as sum of normal random variables is again normal. Hence, instead of drawing $x_{t+1}, \ldots, x_T \sim N(0,1)$ on round t, one can simply draw just one vector $X_t \sim N(0, \sqrt{T-t})$ and use it for perturbation. In this case constant C is bounded by 8.

While we have provided simple distributions to use for perturbation, the form of update in Equation (28) is not in a convenient form. The following lemma shows a simple Follow the Perturbed Leader type algorithm with the associated regret bound.

Lemma 14. *Suppose $\mathcal{F}$ is the ℓ_1^N unit ball and $\mathcal{X}$ is the dual ℓ_∞^N unit ball, and let D be any symmetric distribution. Consider the randomized algorithm that at each round t freshly draws Rademacher random variables $\epsilon_{t+1}, \ldots, \epsilon_T$ and freshly draws $x_{t+1}, \ldots, x_T \sim D^N$ (each co-ordinate drawn independently from D) and picks*

$$f_t = \operatorname*{argmin}_{f\in\mathcal{F}} \left\langle f, \sum_{i=1}^{t-1} x_i - C \sum_{i=t+1}^{T} \epsilon_i x_i \right\rangle$$

where $C = 6/\mathbb{E}_{x\sim D}[|x|]$. The randomized algorithm enjoys a bound on the expected regret given by

$$\mathbb{E}[\mathbf{Reg}_T] \leq C \mathop{\mathbb{E}}_{x_{1:T}\sim D^N} \mathbb{E}_\epsilon \left\| \sum_{t=1}^{T} \epsilon_t x_t \right\|_\infty + 4 \sum_{t=1}^{T} \mathbf{P}_{y_{t+1:T}\sim D} \left(C \left| \sum_{i=t+1}^{T} y_i \right| \leq 4 \right)$$

Notice that for D being the $\{\pm 1\}$ coin flips or standard normal distribution, the probability

$$\mathbf{P}_{y_{t+1},\ldots,y_T\sim D} \left(C \left| \sum_{i=t+1}^{T} y_i \right| \leq 4 \right)$$

is exponentially small in $T-t$ and so $\sum_{t=1}^{T} \mathbf{P}_{y_{t+1},\ldots,y_T\sim D} \left(C \left| \sum_{i=t+1}^{T} y_i \right| \leq 4 \right)$ is bounded by a constant. For these cases, we have

$$\mathbb{E}[\mathbf{Reg}_T] \leq O\left(\mathop{\mathbb{E}}_{x_{1:T}\sim D^N} \mathbb{E}_\epsilon \left\| \sum_{t=1}^{T} \epsilon_t x_t \right\|_\infty \right) = O\left(\sqrt{T \log N} \right)$$

This yields the logarithmic dependence on the dimension, matching that of the Exponential Weights algorithm.

Example : ℓ_2/ℓ_2 Follow the Perturbed Leader:

We now consider the case when $\mathcal{F}$ and $\mathcal{X}$ are both the unit ℓ_2 ball. We can use as perturbation the uniform distribution on the surface of unit sphere, as the following lemma shows. This result was already hinted at in [2], as the random draw from the unit sphere is likely to produce an orthogonal direction, yielding a strategy close to optimal. However, we do not require dimensionality to be high for the result to hold.

Lemma 15. *Let $\mathcal{X}$ and $\mathcal{F}$ be unit balls in Euclidean norm. Then Assumption 2 is satisfied with a uniform distribution D on the surface of the unit sphere with constant $C = 4\sqrt{2}$.*

Again as in the previous example the form of update in Equation (28) is not in a convenient form and this is addressed in the following lemma.

Lemma 16. *Let $\mathcal{X}$ and $\mathcal{F}$ be unit balls in Euclidean norm, and D be the uniform distribution on the surface of the unit sphere. Consider the randomized algorithm that at each round (say round t) freshly draws $x_{t+1}, \ldots, x_T \sim D$ and picks*

$$f_t = \frac{-\sum_{i=1}^{t-1} x_i + C \sum_{i=t+1}^{T} x_i}{\sqrt{\left\| -\sum_{i=1}^{t-1} x_i + C \sum_{i=t+1}^{T} \epsilon_i x_i \right\|_2^2 + 1}}$$

where $C = 4\sqrt{2}$. The randomized algorithm enjoys a bound on the expected regret given by

$$\mathbb{E}\left[\mathbf{Reg}_T\right] \le C \; \mathbb{E}_{x_1,\ldots,x_T \sim D} \left\| \sum_{t=1}^{T} x_t \right\|_2 \le 4\sqrt{2T}$$

Importantly, the bound does not depend on the dimensionality of the space. To the best of our knowledge, this is the first such result for Follow the Perturbed Leader style algorithms.

Remark 1. *The FPL methods developed in [12, 7] assume that the adversary is oblivious. With this simplification, the algorithms can reuse the same random perturbation drawn at the beginning of the game. It is then argued in [7] that the methods also work for non-oblivious opponents since the FPL strategy is fully determined by the outcomes played by the adversary [7, Remark 4.2]. In contrast, our proofs directly deal with the adaptive adversary.*

7.3 Supervised Learning

For completeness, let us state a version of Assumption 1 for the case of supervised learning. That is, the side information x_t is presented to the learner, who then picks $\hat{y}_t$ and observes the outcome y_t.

Assumption 3. *There exists a distribution $D \in \Delta(\mathcal{X} \times \mathcal{Y})$ and constant $C \ge 2$ such that for any $t \in [T]$ and given any $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1}), (x_{t+1}, y_{t+1}), \ldots, (x_T, y_T) \in \mathcal{X} \times \mathcal{Y}$ and any $\epsilon_{t+1}, \ldots, \epsilon_T \in \{\pm 1\}$,*

$$\sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta(\mathcal{Y})} \mathbb{E}_{y_t \sim p_t} \sup_{f \in \mathcal{F}} \left[C \sum_{i=t+1}^{T} \epsilon_i \ell(f(x_i), y_i) - L_{t-1}(f) + \mathbb{E}_{y \sim p_t}\left[\ell(f(x_t), y)\right] - \ell(f(x_t), y_t) \right]$$
$$\le \mathbb{E}_{\epsilon_t, (x_t, y_t) \sim D} \sup_{f \in \mathcal{F}} \left[C \sum_{i=t}^{T} \epsilon_i \ell(f(x_i), y_i) - L_{t-1}(f) \right],$$

where ϵ_t is an independent Rademacher random variable and $L_{t-1}(f) = \sum_{i=1}^{t-1} \ell(f(x_i), y_i)$.

Under Assumption 3, we can use the following relaxation:

$$\mathbf{Rel}_T\left(\mathcal{F} | (x_1, y_1), \ldots, (x_t, y_t)\right) = \mathbb{E}_{\substack{\epsilon_{t+1:T} \\ (x_{t+1}, y_{t+1}), \ldots (x_T, y_T) \sim D}} \sup_{f \in \mathcal{F}} \left[C \sum_{i=t+1}^{T} \epsilon_i \ell(f(x_i), y_i) - \sum_{i=1}^{t} \ell(f(x_i), y_i) \right] \tag{29}$$

Lemma 17. *Under Assumption 3, the relaxation in Eq. (29) is admissible and a randomized strategy that ensures admissibility is given by: at time t, draw $(x_{t+1}, y_{t+1}), \ldots, (x_T, y_T) \sim D$ and Rademacher random variables $\epsilon_{t+1}, \ldots, \epsilon_T$ and then :*

1. *In the case the loss ℓ is convex in its first argument, define*

$$\hat{y}_t = \operatorname*{argmin}_{\hat{y}\in[-B,B]} \sup_{y_t\in\mathcal{Y}} \left\{ \ell(\hat{y}, y_t) + \sup_{f\in\mathcal{F}} \left[C \sum_{i=t+1}^{T} \epsilon_i \ell(f(x_i), y_i) - \sum_{i=1}^{t} \ell(f(x_i), y_i) \right] \right\} \tag{30}$$

and

2. *In the case of non-convex loss, pick $\hat{y}_t$ from the distribution*

$$\hat{q}_t = \operatorname*{argmin}_{\hat{q}\in\Delta([-B,B])} \sup_{y_t\in\mathcal{Y}} \left\{ \mathop{\mathbb{E}}_{\hat{y}\sim\hat{q}} [\ell(\hat{y}, y_t)] + \sup_{f\in\mathcal{F}} \left[C \sum_{i=t+1}^{T} \epsilon_i \ell(f(x_i), y_i) - \sum_{i=1}^{t} \ell(f(x_i), y_i) \right] \right\} \tag{31}$$

The expected regret bound of the method (in both cases) is

$$\mathbb{E}[\mathbf{Reg}_T] \le C \mathop{\mathbb{E}}_{(x_1,y_1),\ldots,(x_T,y_T)\sim D} \mathop{\mathbb{E}}_{\epsilon} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} \epsilon_t \ell(f(x_t), y_t) \right]$$

7.4 Random Walks with Trees

We can also define randomized algorithms without the assumption that the classical and the sequential Rademacher complexities are close. Instead, we assume that we have a black-box access to a procedure that on round t returns the "worst-case" tree $\mathbf{x}^t$ of depth $T-t$.

Lemma 18. *Given any $x_1, \ldots, x_{t-1}$ let*

$$\mathbf{x}^t \triangleq \operatorname*{argmax}_{\mathbf{x}} \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \left[2 \sum_{i=t+1}^{T} \epsilon_i \ell(f, \mathbf{x}_i(\epsilon)) - \sum_{i=1}^{t} \ell(f, x_i) \right] . \tag{32}$$

Consider the randomized strategy where at round t we first draw $\epsilon_{t+1}, \ldots, \epsilon_T$ uniformly at random and then further draw our move f_t according to the distribution

$$q_t(\epsilon) = \operatorname*{argmin}_{q\in\Delta(\mathcal{F})} \sup_{x_t} \left\{ \mathop{\mathbb{E}}_{f_t\sim q} [\ell(f_t, x_t)] + \sup_{f\in\mathcal{F}} \left[2 \sum_{i=t+1}^{T} \epsilon_i \ell(f, \mathbf{x}_i^t(\epsilon)) - \sum_{i=1}^{t} \ell(f, x_i) \right] \right\} \tag{33}$$

The expected regret of this randomized strategy is bounded by sequential Rademacher complexity:

$$\mathbb{E}[\mathbf{Reg}_T] \le \mathfrak{R}_T(\mathcal{F}) .$$

Thus, if for any given history $x_1, \ldots, x_{t-1}$ we can compute $\mathbf{x}^t$ in (32), or even just draw directly a random path $\mathbf{x}_1^t(\epsilon), \ldots, \mathbf{x}_{T-t}^t(\epsilon)$ on each round, then we obtain a randomized strategy that in expectation can guarantee a regret bound equal to sequential Rademacher complexity. Also notice that whenever the optimal strategy in (33) is deterministic (e.g. in the online convex optimization scenario), one does not need the double randomization. Instead, in such situations one can directly draw $\epsilon_1, \ldots, \epsilon_{T-t}$ and use

$$f_t(\epsilon) = \operatorname*{argmin}_{f_t\in\mathcal{F}} \sup_{x_t} \left\{ \ell(f_t, x_t) + \sup_{f\in\mathcal{F}} \left\{ 2 \sum_{i=t+1}^{T} \epsilon_i \ell(f, \mathbf{x}_i^t(\epsilon)) - \sum_{i=1}^{t} \ell(f, x_i) \right\} \right\}$$

8 Static Experts with Convex Losses and Transductive Online Learning

We show how to recover a variant of the R^2 forecaster of [9], for static experts and transductive online learning. At each round, the learner makes a prediction $q_t \in [-1, 1]$, observes the outcome $y_t \in [-1, 1]$, and

suffers convex L-Lipschitz loss $\ell(q_t, y_t)$. Regret is defined as the difference between learner's cumulative loss and $\inf_{f\in F}\sum_{t=1}^{T}\ell(f[t], y_t)$, where $F \subset [-1,1]^T$ can be seen as a set of static experts. The transductive setting is equivalent to this: the sequence of x_t's is known before the game starts, and hence the *effective* function class is once again a subset of $[-1,1]^T$.

It turns out that in the static experts case, sequential Rademacher complexity boils down to the classical Rademacher complexity (see [18]), and thus the relaxation in (17) can be taken to be the classical, rather than sequential, Rademacher averages. This is also the reason that an efficient implementation by sampling is possible. Furthermore, for the absolute loss, the factor of 2 that appears in the sequential Rademacher complexity is not needed. For general convex loss, one possible relaxation is just a conditional version of the classical Rademacher averages:

$$\mathbf{Rel}_T\left(F|y_1,\ldots,y_t\right) = \mathbb{E}_{\epsilon_{t+1:T}}\sup_{f\in F}\left[2L\sum_{s=t+1}^{T}\epsilon_s f[s] - L_t(f)\right] \tag{34}$$

where $L_t(f) = \sum_{s=1}^{t}\ell(f[s], y_s)$. This relaxation can be shown to be admissible.

First, consider the case of absolute loss $\ell(q_t, y_t) = |q_t - y_t|$ and binary-valued outcomes $y_t \in \{\pm 1\}$. In this case, the solution in (17) yields the algorithm

$$q_t = \frac{1}{2}\mathbb{E}_{\epsilon_{t+1:T}}\left[\sup_{f\in F}\left(\sum_{s=t+1}^{T}\epsilon_s f[s] - L_{t-1}(f) + f[t]\right) - \sup_{f\in F}\left(\sum_{s=t+1}^{T}\epsilon_s f[s] - L_{t-1}(f) - f[t]\right)\right]$$

which corresponds to the well-known minimax optimal forecaster for static experts with absolute loss [7]. Plugging in this value of q_t into Eq. (16) proves admissibility, and thus the regret guarantee of this method is equal to the classical Rademacher complexity.

We now derive two variants of the R^2 forecaster for the more general case of L-Lipschitz loss and $y_t \in [-1,1]$.

First Alternative : If (34) is used as a relaxation, the calculation of prediction $\hat{y}_t$ involves a supremum over $f \in F$ with (potentially nonlinear) loss functions of instances seen so far. In some cases this optimization might be hard and it might be preferable if the supremum only involves terms *linear* in f. This is the idea behind he first method we present. To this end we start by noting that by convexity

$$\sum_{t=1}^{T}\ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t), y_t) \le \sum_{t=1}^{T}\partial\ell(\hat{y}_t, y_t)\cdot\hat{y}_t - \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\partial\ell(\hat{y}_t, y_t)\cdot f[t]$$

Now given the above, one can consider an alternative online learning problem which, if we solve, also solves the original problem. That is, consider the online learning problem with the new loss

$$\ell'(\hat{y}, r) = r\cdot\hat{y}$$

In this alternative game, we first pick prediction $\hat{y}_t$ (deterministically), next the adversary picks r_t (corresponding to $r_t = \partial\ell(\hat{y}_t, y_t)$ for choice of y_t picked by adversary). Now note that ℓ' is indeed convex in its first argument and is L Lipschitz because $|\partial\ell(\hat{y}_t, y_t)| \le L$. This is a one dimensional convex learning game where we pick $\hat{y}_t$ and regret is given by

$$\mathbf{Reg}_T = \sum_{t=1}^{T}\partial\ell(\hat{y}_t, y_t)\cdot\hat{y}_t - \inf_{f\in F}\sum_{t=1}^{T}\partial\ell(\hat{y}_t, y_t)\cdot f[t]$$

One can consider the relaxation

$$\mathbf{Rel}_T\left(F|\partial\ell(\hat{y}_1, y_1),\ldots,\partial\ell(\hat{y}_t, y_t)\right) = \mathbb{E}_{\epsilon_{t+1:T}}\sup_{f\in F}\left[2L\sum_{i=t+1}^{T}\epsilon_i f[t] - \sum_{i=1}^{t}\partial\ell(\hat{y}_i, y_i)\cdot f[i]\right] \tag{35}$$

as a linearized form of (34). At round t, the prediction of the algorithm is then

$$\hat{y}_t = \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in F}\left\{\sum_{i=t+1}^{T}\epsilon_i f[i] - \frac{1}{2L}\sum_{i=1}^{t-1}\partial\ell(\hat{y}_i, y_i)f[i] + \frac{1}{2}f[t]\right\} - \sup_{f\in F}\left\{\sum_{i=t+1}^{T}\epsilon_i f[i] - \frac{1}{2L}\sum_{i=1}^{t-1}\partial\ell(\hat{y}_i, y_i)f[i] - \frac{1}{2}f[t]\right\}\right] \tag{36}$$

Lemma 19. *The relaxation in Equation (35) is admissible with respect to the prediction strategy specified in Equation (36). Further the regret of the strategy is bounded as*

$$\mathbf{Reg}_T \le 2L \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in F}\sum_{t=1}^{T}\epsilon_t f[t]\right]$$

The presented algorithm is similar in principle to R^2, with the main difference that R^2 computes the infima over a sum of absolute losses, while here we have a more manageable linearized objective. Note that while we need to evaluate the expectation over ϵ's on each round, we can estimate $\hat{y}_t$ by sampling ϵ's and using McDiarmid's inequality to argue that, with enough draws, our estimate is close to $\hat{y}_t$ with high probability. What is interesting, we can develop a randomized method that only draws one sequence of ϵ's per step, as shown next.

Second Alternative : Consider the non-linearized relaxation

$$\mathbf{Rel}_T\left(F|y_1,\ldots,y_t\right) = \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in F}\ 2L\sum_{i=t+1}^{T}\epsilon_i f[i] - \sum_{i=1}^{t}\ell(f[i], y_i)\right] \tag{37}$$

already given in (34). We now present a randomized method based on the ideas of Section 7: at round t we first draw $\epsilon_{t+1},\ldots,\epsilon_T$ and predict

$$\hat{y}_t(\epsilon) = \left(\inf_{f\in F}\left\{-\sum_{i=t+1}^{T}\epsilon_i f[i] + \frac{1}{2L}\sum_{i=1}^{t-1}\ell(f[i], y_i) + \frac{1}{2}f[t]\right\} - \inf_{f\in F}\left\{-\sum_{i=t+1}^{T}\epsilon_i f[i] + \frac{1}{2L}\sum_{i=1}^{t-1}\ell(f[i], y_i) - \frac{1}{2}f[t]\right\}\right) \tag{38}$$

We show that this predictor in expectation enjoys regret bound of the transductive Rademacher complexity. More specifically we have the following lemma.

Lemma 20. *The relaxation specified in Equation (37) is admissible w.r.t. the randomized prediction strategy specified in Equation (38). Further the expected regret of the randomized strategy is bounded as*

$$\mathbb{E}\left[\mathbf{Reg}_T\right] \le 2L \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in F}\sum_{t=1}^{T}\epsilon_t f[t]\right]$$

In the next section, we employ both alternatives to develop novel algorithms for matrix completion.

9 Matrix Completion

Consider the problem of predicting unknown entries in a matrix (as in collaborative filtering). We focus here on an online formulation, where at each round t the adversary picks an entry in an $m \times n$ matrix and a value y_t for that entry (we shall assume without loss of generality that $n \ge m$). The learner then chooses a predicted value $\hat{y}_t$, and suffers loss $\ell(y_t, \hat{y}_t)$, which we shall assume to be ρ-Lipschitz. We define our regret with respect to the class $\mathcal{F}$ which we will take to be the set of all matrices whose trace-norm is at most B

(namely, we can use any such matrix to predict just by returning its relevant entry at each round). Usually, one sets B to be on the order of $\sqrt{mn}$.

We consider here a transductive version, where the sequence of entry locations is known in advance, and only the entry values are unknown. We show how to develop an algorithm whose regret is bounded by the (transductive) Rademacher complexity of $\mathcal{F}$. We note that in Theorem 6 of [19], this complexity was shown to be at most order $B\sqrt{n}$ independent of T. Moreover, in [9], it was shown that for algorithms with such guarantees, and whose play each round does not depend on the order of future entries, under mild conditions on the loss function one can get the same regret even in the "fully" online case where the set of entry locations is unknown in advance. Algorithmically, all we need to do is pretend we are in a transductive game where the sequence of entries is all $m \times n$ entries, in some arbitrary order. In this section we use the two alternatives provided for transductive learning problem in the previous subsection and provide two alternatives for the matrix completion problem.

We note that both variants proposed here improve on the one provided by the R^2 forecaster in [9], since that algorithm competes against the smaller class $\mathcal{F}'$ of matrices with bounded trace-norm *and* bounded individual entries. In contrast, our algorithm provides similar regret guarantees against the larger class of matrices only whose trace-norm is bounded. Moreover, the variants are also computationally more efficient.

First Alternative : The algorithm we now present is obtained by using the first method for online tranductive learning proposed in the previous section. The relaxation in Equation (35) for the specific problem at hand is given by,

$$\mathbf{Rel}_{\mathcal{F}}(y_1, \ldots, y_t) = B \; \mathop{\mathbb{E}}_{\epsilon} \left[\left\| 2\rho \sum_{i=t+1}^{T} \epsilon_i x_i - \sum_{i=1}^{t} \partial\ell(\hat{y}_i, y_i) x_i \right\|_{\sigma} \right] \tag{39}$$

In the above $\|\cdot\|_{\sigma}$ stands for the spectral norm and each x_i is a matrix with a 1 at some specific position and 0 elsewhere. That is x_i at round i can be seen as the entry of the matrix which we are asked to fill in at round i. The prediction at round t returned by the algorithm is given by Equation (36) which for this problem is given by

$$\hat{y}_t = B \; \mathop{\mathbb{E}}_{\epsilon} \left[\left(\left\| \sum_{i=t+1}^{T} \epsilon_i x_i - \frac{1}{2\rho} \sum_{i=1}^{t-1} \partial\ell(\hat{y}_i, y_i) x_i + \frac{1}{2} x_t \right\|_{\sigma} - \left\| \sum_{i=t+1}^{T} \epsilon_i x_i - \frac{1}{2\rho} \sum_{i=1}^{t-1} \partial\ell(\hat{y}_i, y_i) x_i - \frac{1}{2} x_t \right\|_{\sigma} \right) \right]$$

Notice that the algorithm only involves calculation of spectral norms on each round which can be done efficiently. Again as mentioned in previous subsection, one can evaluate the expectation over random signs by sampling ϵ's on each round.

Second Alternative : The second algorithm is obtained from the second alternative for online transductive learning with convex losses in the previous section. The relaxation given in Equation (37) for the case of matrix completion problem with trace norm constraint is given by:

$$\mathbf{Rel}_T(\mathcal{F}|y_1, \ldots, y_t) = \mathop{\mathbb{E}}_{\epsilon} \left[\sup_{f : \|f\|_{\Sigma} \leq B} 2\rho \sum_{i=t+1}^{T} \epsilon_i \langle f, x_i \rangle - \sum_{i=1}^{t} \ell(\langle f, x_i \rangle, y_i) \right]$$

where $\|\cdot\|_{\Sigma}$ stands for the race norm of the $m \times n$ matrix f and each x_i is a matrix with a 1 at some specific position and 0 elsewhere. That is x_i at round i can be seen as the entry of the matrix which we are asked to fill in at round i. We use $\langle f, x \rangle$ to represent the generalized inner product of the two matrices. Since we only take inner products with respect to the matrices x_i, each $\langle f, x_i \rangle$ is simply the value of matrix f at the position specified by x_i's. The prediction at a matrix entry corresponding to position x_t is given by first drawing random $\{\pm 1\}$ valued ϵ's and then applying Equation (38) to the problem at hand, yielding

$$\hat{y}_t(\epsilon) = \inf_{\|f\|_{\Sigma} \leq B} \left\{ - \sum_{i=t+1}^{T} \epsilon_i \langle f, x_i \rangle + \frac{1}{2\rho} \sum_{i=1}^{t-1} \ell(\langle f, x_i \rangle, y_i) + \frac{1}{2} \langle f, x_t \rangle \right\} - \inf_{\|f\|_{\Sigma} \leq B} \left\{ - \sum_{i=t+1}^{T} \epsilon_i \langle f, x_i \rangle + \frac{1}{2\rho} \sum_{i=1}^{t-1} \ell(\langle f, x_i \rangle, y_i) - \frac{1}{2} \langle f, x_t \rangle \right\}$$

Notice that the above involves solving two trace norm constrained convex optimization problems per round. As a simple corollary of Lemma 20 we get the following bound on expected regret of the algorithm.

Corollary 21. *For the randomized prediction strategy specified above, the expected regret is bounded as*

$$\mathbb{E}\left[\mathbf{Reg}_T\right] \le 2B\ \rho\ \mathop{\mathbb{E}}_{\epsilon}\left[\left\|\sum_{t=1}^{T}\epsilon_t x_t\right\|_\sigma\right] \le O\left(B\ \rho\ \left(\sqrt{m}+\sqrt{n}\right)\right)$$

The last inequality in the above corollary is using Theorem 6 in [19].

Corollary 22. *For the predictions $\hat{y}_t$ specified above, the regret is bounded as*

$$\mathbf{Reg}_T \le O\left(B\ \rho\ \left(\sqrt{m}+\sqrt{n}\right)\right)$$

10 More Examples

10.1 Constrained Adversaries

We now show that algorithms can be also developed for situations when the adversary is constrained in the choices per step. Such constrained problems have been treated in a general non-algorithmic way in [18], and we picked the case of variation-constrained adversary for illustration. It is shown in [18] that the value of the game where the adversary is constrained to keep the next move x_t within σ_t from the average of the past moves $\frac{1}{t-1}\sum_{s=1}^{t-1} x_s$ is upper bounded as

$$\mathcal{V}_T \le 2 \sup_{(\mathbf{x},\mathbf{x}')\epsilon\mathcal{T}} \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\epsilon\mathcal{F}}\sum_{t=1}^{T}\epsilon_t\left(\langle f,\mathbf{x}_t(\epsilon)\rangle - \frac{1}{t-1}\sum_{\tau=1}^{t-1}\langle f,\chi_\tau(\epsilon_\tau)\rangle\right)\right] \tag{40}$$

where the supremum is over $\mathbf{x},\mathbf{x}'$ trees satisfying the above mentioned constraint per step, and the selector $\chi_t(\epsilon_t)$ is defined as $\mathbf{x}_t(\epsilon)$ if $\epsilon_t = -1$ and $\mathbf{x}'_t(\epsilon)$ otherwise. In our algorithmic framework, this leads to the following problem that needs to be solved at each step:

$$\inf_{f_t}\sup_{x_t}\left\{\langle f_t,x_t\rangle + 2\sup_{(\mathbf{x},\mathbf{x}')\epsilon\mathcal{T}}\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\epsilon\mathcal{F}}\left\langle f,\sum_{s=t+1}^{T}\epsilon_s\left(\mathbf{x}_s(\epsilon)-\frac{1}{s-t}\sum_{\tau=t+1}^{s-1}\chi_\tau(\epsilon_\tau)\right)-\sum_{r=1}^{t}x_r\right\rangle\right]\right\}$$

where the supremum is taken over x_t such that the constraint $C(x_1,\ldots,x_t)$ is satisfied and $\mathcal{T}$ is the set of trees that satisfy the constraints as continuation of the prefix $x_1,\ldots,x_t$. While this expression gives rise to an algorithm, we are aiming for a more computationally feasible method. In fact, passing to an upper bound on the sequential Rademacher complexity yields the following result.

Lemma 23. *The following relaxation is admissible and upper bounds the constrained sequential complexity*

$$\mathbf{Rel}_T\left(\mathcal{F}|x_1,\ldots,x_t\right) = \frac{2\sqrt{2}R}{\sqrt{\lambda}}\sqrt{\left\|\sum_{r=1}^{t}x_r\right\|^2 + C\sum_{s=t+1}^{T}\sigma_s^2}$$

Furthermore, the an admissible algorithm for this relaxation is Mirror Descent with a step size given at time $t \ge 2$ by

$$\frac{\left(1+\frac{1}{t-1}\right)^2}{2\sqrt{\|\tilde{x}_{t-1}\|^2 + C\sum_{s=t}^{T}\sigma_s^2}}$$

10.2 Universal Mirror Descent

In [20] it is shown that for the problem of general online convex optimization, the Mirror Descent algorithm is universal and near optimal (up to poly-log factors). Specifically, it is shown that there always exists an appropriate function Ψ such that the Mirror Descent algorithm using this function, along with an appropriate step size, gives the near optimal rate. Moreover, it is shown in [20] that one can use function Ψ whose convex conjugate is given by

$$\Psi^*(x) = \sup_{\mathbf{x}} \mathop{\mathbb{E}}_{\epsilon} \left[\left\| x + \sum_{i=1}^{T-t} \epsilon_i \mathbf{x}_i(\epsilon) \right\|^p - C \sum_{i=1}^{T-t} \mathop{\mathbb{E}}_{\epsilon} \left[\|\mathbf{x}_i(\epsilon)\|^p \right] \right], \tag{41}$$

as the "universal regularizer" for the Mirror Descent algorithm. We now show that this function arises rather naturally from the sequential Rademacher relaxation and, moreover, the Mirror Descent algorithm itself arises from this relaxation.

Let us denote the convex cost functions chosen by the adversary as ℓ_t, and let x_t be the subgradients $x_t = \nabla \ell_t(f_t)$ of the convex functions.

Lemma 24. *The relaxation*

$$\mathbf{Rel}_T(\mathcal{F}|x_1, \ldots, x_t) = \left(\Psi^* \left(\sum_{i=1}^{t-1} x_i \right) + \left\langle \nabla \Psi^* \left(\sum_{i=1}^{t-1} x_i \right), x_t \right\rangle + C(T-t+1) \right)^{1/p}$$

is an upper bound on the conditional sequential Rademacher complexity. Further, whenever for some $p' > p$ we have that $\mathcal{V}_T(\mathcal{F}) \le (CT)^{1/p'}$, then the relaxation is admissible and leads to a form of Mirror Descent algorithm with regret bounded as

$$\mathbf{Reg}_T \le (CT)^{1/p}$$

It is remarkable that the universal regularizer and the Mirror Descent algorithm arise naturally, in a few steps of algebra, as upper bounds on the sequential Rademacher complexity.

A PROOFS

Proof of Proposition 1. By definition,

$$\sum_{t=1}^{T} \mathbb{E}_{f_t \sim q_t} \ell(f_t, x_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f, x_t) \le \sum_{t=1}^{T} \mathbb{E}_{f_t \sim q_t} \ell(f_t, x_t) + \mathbf{Rel}_T\left(\mathcal{F} | x_1, \dots, x_T\right) \ .$$

Peeling off the T-th expected loss, we have

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}_{f_t \sim q_t} \ell(f_t, x_t) + \mathbf{Rel}_T\left(\mathcal{F} | x_1, \dots, x_T\right) &\le \sum_{t=1}^{T-1} \mathbb{E}_{f_t \sim q_t} \ell(f_t, x_t) + \{\mathbb{E}_{f_t \sim q_t} \ell(f_t, x_t) + \mathbf{Rel}_T\left(\mathcal{F} | x_1, \dots, x_T\right)\} \\
&\le \sum_{t=1}^{T-1} \mathbb{E}_{f_t \sim q_t} \ell(f_t, x_t) + \mathbf{Rel}_T\left(\mathcal{F} | x_1, \dots, x_{T-1}\right)
\end{aligned}$$

where we used the fact that q_T is an admissible algorithm for this relaxation, and thus the last inequality holds for any choice x_T of the opponent. Repeating the process, we obtain

$$\sum_{t=1}^{T} \mathbb{E}_{f_t \sim q_t} \ell(f_t, x_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f, x_t) \le \mathbf{Rel}_T\left(\mathcal{F}\right) \ .$$

We remark that the left-hand side of this inequality is random, while the right-hand side is not. Since the inequality holds for any realization of the process, it also holds in expectation. The inequality

$$\mathcal{V}_T(\mathcal{F}) \le \mathbf{Rel}_T\left(\mathcal{F}\right)$$

holds by unwinding the value recursively and using admissibility of the relaxation. The high-probability bound is an immediate consequences of (6) and the Hoeffding-Azuma inequality for bounded martingales. The last statement is immediate. □

Proof of Proposition 2. Denote $L_t(f) = \sum_{s=1}^{t} \ell(f, x_s)$. The first step of the proof is an application of the minimax theorem (we assume the necessary conditions hold):

$$\begin{aligned}
&\inf_{q_t \in \Delta(\mathcal{F})} \sup_{x_t \in \mathcal{X}} \left\{ \mathop{\mathbb{E}}_{f_t \sim q_t} [\ell(f_t, x_t)] + \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_{t+1:T}} \sup_{f \in \mathcal{F}} \left[2 \sum_{s=t+1}^{T} \epsilon_s \ell(f, \mathbf{x}_{s-t}(\epsilon_{t+1:s-1})) - L_t(f) \right] \right\} \\
&= \sup_{p_t \in \Delta(\mathcal{X})} \inf_{f_t \in \mathcal{F}} \left\{ \mathop{\mathbb{E}}_{x_t \sim p_t} [\ell(f_t, x_t)] + \mathop{\mathbb{E}}_{x_t \sim p_t} \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_{t+1:T}} \sup_{f \in \mathcal{F}} \left[2 \sum_{s=t+1}^{T} \epsilon_s \ell(f, \mathbf{x}_{s-t}(\epsilon_{t+1:s-1})) - L_t(f) \right] \right\}
\end{aligned}$$

For any $p_t \in \Delta(\mathcal{X})$, the infimum over f_t of the above expression is equal to

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_t \sim p_t} \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_{t+1:T}} \sup_{f \in \mathcal{F}} \left[2 \sum_{s=t+1}^{T} \epsilon_s \ell(f, \mathbf{x}_{s-t}(\epsilon_{t+1:s-1})) - L_{t-1}(f) + \inf_{f_t \in \mathcal{F}} \mathop{\mathbb{E}}_{x_t \sim p_t} [\ell(f_t, x_t)] - \ell(f, x_t) \right] \\
&\le \mathop{\mathbb{E}}_{x_t \sim p_t} \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_{t+1:T}} \sup_{f \in \mathcal{F}} \left[2 \sum_{s=t+1}^{T} \epsilon_s \ell(f, \mathbf{x}_{s-t}(\epsilon_{t+1:s-1})) - L_{t-1}(f) + \mathop{\mathbb{E}}_{x_t \sim p_t} [\ell(f, x_t)] - \ell(f, x_t) \right] \\
&\le \mathop{\mathbb{E}}_{x_t, x_t' \sim p_t} \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_{t+1:T}} \sup_{f \in \mathcal{F}} \left[2 \sum_{s=t+1}^{T} \epsilon_s \ell(f, \mathbf{x}_{s-t}(\epsilon_{t+1:s-1})) - L_{t-1}(f) + \ell(f, x_t') - \ell(f, x_t) \right]
\end{aligned}$$

We now argue that the independent x_t and x_t' have the same distribution p_t, and thus we can introduce a random sign ϵ_t. The above expression then equals to

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_t, x_t' \sim p_t} \mathbb{E}_{\epsilon_t} \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_{t+1:T}} \sup_{f \in \mathcal{F}} \left[2 \sum_{s=t+1}^{T} \epsilon_s \ell(f, \mathbf{x}_{s-t}(\epsilon_{t+1:s-1})) - L_{t-1}(f) + \epsilon_t(\ell(f, x_t') - \ell(f, x_t)) \right] \\
&\le \sup_{x_t, x_t' \in \mathcal{X}} \mathbb{E}_{\epsilon_t} \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_{t+1:T}} \sup_{f \in \mathcal{F}} \left[2 \sum_{s=t+1}^{T} \epsilon_s \ell(f, \mathbf{x}_{s-t}(\epsilon_{t+1:s-1})) - L_{t-1}(f) + \epsilon_t(\ell(f, x_t') - \ell(f, x_t)) \right]
\end{aligned}$$

where we upper bounded the expectation by the supremum. Splitting the resulting expression into two parts, we arrive at the upper bound of

$$2 \sup_{x_t \in \mathcal{X}} \mathbb{E} \sup_{\epsilon_t} \mathbb{E}_{\epsilon_{t+1:T}} \sup_{\mathbf{x}} \sup_{f \in \mathcal{F}} \left[\sum_{s=t+1}^{T} \epsilon_s \ell(f, \mathbf{x}_{s-t}(\epsilon_{t+1:s-1})) - \frac{1}{2} L_{t-1}(f) + \epsilon_t \ell(f, x_t) \right] = \mathfrak{R}_T(\mathcal{F}|x_1, \dots, x_{t-1}) \ .$$

The last equality is easy to verify, as we are effectively adding a root x_t to the two subtrees, for $\epsilon_t = +1$ and $\epsilon_t = -1$, respectively.

One can see that the proof of admissibility corresponds to one step minimax swap and symmetrization in the proof of [16]. In contrast, in the latter paper, all T minimax swaps are performed at once, followed by T symmetrization steps. □

Proof of Proposition 3. Let us first prove that the relaxation is admissible with the Exponential Weights algorithm as an admissible algorithm. Let $L_t(f) = \sum_{i=1}^{t} \ell(f, x_i)$. Let λ^* be the optimal value in the definition of $\mathbf{Rel}_T(\mathcal{F}|x_1, \dots, x_{t-1})$. Then

$$\begin{aligned} &\inf_{q_t \in \Delta(\mathcal{F})} \sup_{x_t \in \mathcal{X}} \left\{ \mathop{\mathbb{E}}_{f \sim q_t} [\ell(f, x_t)] + \mathbf{Rel}_T(\mathcal{F}|x_1, \dots, x_t) \right\} \\ &\leq \inf_{q_t \in \Delta(\mathcal{F})} \sup_{x_t \in \mathcal{X}} \left\{ \mathop{\mathbb{E}}_{f \sim q_t} [\ell(f, x_t)] + \frac{1}{\lambda^*} \log \left(\sum_{f \in \mathcal{F}} \exp\left(-\lambda^* L_t(f)\right) \right) + 2\lambda^* (T - t) \right\} \end{aligned}$$

Let us upper bound the infimum by a particular choice of q which is the exponential weights distribution

$$q_t(f) = \exp(-\lambda^* L_{t-1}(f))/Z_{t-1}$$

where $Z_{t-1} = \sum_{f \in \mathcal{F}} \exp\left(-\lambda^* L_{t-1}(f)\right)$. By [7, Lemma A.1],

$$\begin{aligned} \frac{1}{\lambda^*} \log \left(\sum_{f \in \mathcal{F}} \exp\left(-\lambda^* L_t(f)\right) \right) &= \frac{1}{\lambda^*} \log \left(\mathbb{E}_{f \sim q_t} \exp\left(-\lambda^* \ell(f, x_t)\right) \right) + \frac{1}{\lambda^*} \log Z_{t-1} \\ &\leq -\mathbb{E}_{f \sim q_t} \ell(f, x_t) + \frac{\lambda^*}{2} + \frac{1}{\lambda^*} \log Z_{t-1} \end{aligned}$$

Hence,

$$\begin{aligned} \inf_{q_t \in \Delta(\mathcal{F})} \sup_{x_t \in \mathcal{X}} \left\{ \mathop{\mathbb{E}}_{f \sim q_t} [\ell(f, x_t)] + \mathbf{Rel}_T(\mathcal{F}|x_1, \dots, x_t) \right\} &\leq \frac{1}{\lambda^*} \log \left(\sum_{f \in \mathcal{F}} \exp\left(-\lambda^* L_{t-1}(f)\right) \right) + 2\lambda^* (T - t + 1) \\ &= \mathbf{Rel}_T(\mathcal{F}|x_1, \dots, x_{t-1}) \end{aligned}$$

by the optimality of λ^*. The bound can be improved by a factor of 2 for some loss functions, since it will disappear from the definition of sequential Rademacher complexity.

We conclude that the Exponential Weights algorithm is an admissible strategy for the relaxation (9). The final regret bound follows immediately from the bound on sequential Rademacher complexity (which, in this case, is simply the supremum of a martingale difference process indexed by N elements – see e.g. [16]).

Arriving at the relaxation We now show that the Exponential Weights relaxation arises naturally as an upper bound on sequential Rademacher complexity of a finite class. For any $\lambda > 0$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in\mathcal{F}}\left\{2\sum_{i=1}^{T-t}\epsilon_i\ell(f,\mathbf{x}_i(\epsilon)) - L_t(f)\right\}\right] &\le \frac{1}{\lambda}\log\left(\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in\mathcal{F}}\exp\left(2\lambda\sum_{i=1}^{T-t}\epsilon_i\ell(f,\mathbf{x}_i(\epsilon)) - \lambda L_t(f)\right)\right]\right)\\
&\le \frac{1}{\lambda}\log\left(\mathop{\mathbb{E}}_{\epsilon}\left[\sum_{f\in\mathcal{F}}\exp\left(2\lambda\sum_{i=1}^{T-t}\epsilon_i\ell(f,\mathbf{x}_i(\epsilon)) - \lambda L_t(f)\right)\right]\right)\\
&= \frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\lambda L_t(f)\right)\mathop{\mathbb{E}}_{\epsilon}\left[\prod_{i=1}^{T-t}\exp\left(2\lambda\epsilon_i\ell(f,\mathbf{x}_i(\epsilon))\right)\right]\right)
\end{aligned}$$

Since, conditioned on $\epsilon_1,\dots,\epsilon_{i-1}$, the random variable $\epsilon_i\ell(f,\mathbf{x}_i(\epsilon))$ is subgaussian, we can upper bound the expected value of the product, peeling one random variable at a time from the end (see [16] for the proof). We arrive at the upper bound

$$\begin{aligned}
&\frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\lambda L_t(f)\right)\times\exp\left(2\lambda^2\max_{\epsilon_1,\dots\epsilon_{T-t}\in\{\pm1\}}\sum_{i=1}^{T-t}\ell(f,\mathbf{x}_i(\epsilon))^2\right)\right)\\
&\le \frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\lambda L_t(f) + 2\lambda^2\max_{\epsilon_1,\dots\epsilon_{T-t}\in\{\pm1\}}\sum_{i=1}^{T-t}\ell(f,\mathbf{x}_i(\epsilon))^2\right)\right)\\
&\le \frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\lambda L_t(f)\right)\right) + 2\lambda\sup_{\mathbf{x}}\sup_{f\in\mathcal{F}}\max_{\epsilon_1,\dots\epsilon_{T-t}\in\{\pm1\}}\sum_{i=1}^{T-t}\ell(f,\mathbf{x}_i(\epsilon))^2
\end{aligned}$$

The last term, representing the "worst future", is upper bounded by $2\lambda(T-t)$, assuming that the losses are bounded by 1. This removes the $\mathbf{x}$ tree and leads to the relaxation (9) and a computationally tractable algorithm. □

Proof of Proposition 4. The argument can be seen as a generalization of the Euclidean proof in [2] to general smooth norms. Let $\tilde{x}_{t-1} = \sum_{i=1}^{t-1} x_i$. The optimal algorithm for the relaxation (10) is

$$f_t^* = \mathop{\mathrm{argmin}}_{f\in\mathcal{F}}\left\{\sup_{x_t\in\mathcal{X}}\left\{\langle f, x_t\rangle + \sqrt{\|\tilde{x}_{t-1}\|^2 + \left\langle\nabla\tfrac{1}{2}\|\tilde{x}_{t-1}\|^2, x_t\right\rangle + C(T-t+1)}\right\}\right\} \tag{42}$$

Instead, let

$$f_t = -\frac{\nabla\frac{1}{2}\|\tilde{x}_{t-1}\|^2}{2\sqrt{\|\tilde{x}_{t-1}\|^2 + C(T-t+1)}}\,. \tag{43}$$

Plugging this choice into the admissibility condition (5), we get

$$\sup_{x_t\in\mathcal{X}}\left\{-\frac{\left\langle\nabla\frac{1}{2}\|\tilde{x}_{t-1}\|^2, x_t\right\rangle}{2\sqrt{A}} + \sqrt{A + \left\langle\nabla\tfrac{1}{2}\|\tilde{x}_{t-1}\|^2, x_t\right\rangle}\right\}$$

where $A = \|\tilde{x}_{t-1}\|^2 + C(T-t+1)$. It can be easily verified that an expression of the form $-x/(2y) + \sqrt{y+x}$ is maximized at $x = 0$ for a positive y. With these values,

$$\begin{aligned}
\inf_{f_t\in\mathcal{F}}\left\{\sup_{x_t\in\mathcal{X}}\left\{\langle f_t, x_t\rangle + \left(\|\tilde{x}_{t-1}\|^2 + \left\langle\nabla\tfrac{1}{2}\|\tilde{x}_{t-1}\|^2, x_t\right\rangle + C(T-t+1)\right)^{1/2}\right\}\right\} &= \left(\|\tilde{x}_{t-1}\|^2 + C(T-t+1)\right)^{1/2}\\
&\le \left(\|\bar{x}_{t-2}\|^2 + \left\langle\nabla\tfrac{1}{2}\|\bar{x}_{t-2}\|^2, x_{t-1}\right\rangle + C(T-t+2)\right)^{1/2} = \mathbf{Rel}_T\left(\mathcal{F}|x_1,\dots,x_{t-1}\right)
\end{aligned}$$

Hence, the choice (43) is an admissible algorithm for the relaxation (10).

Evidently, the above proof of admissibility is very simple, but it might seem that we pulled the algorithm (43) out of a hat. We now show that in fact one can derive this algorithm as a solution $f_t^\star$ in (42). The proof below is not required for admissibility, and we only include it for completeness. The proof uses the fact that for *any* norm $\|\cdot\|$,

$$\left\langle \nabla \tfrac{1}{2}\|x\|^2, x \right\rangle = \|x\|^2 \ . \tag{44}$$

To prove this, observe that by convexity $\|0\| \geq \|x\| + \langle \nabla\|x\|, 0-x\rangle$ and $\|2x\| \geq \|x\| + \langle \nabla\|x\|, 2x-x\rangle$ implying $\langle \nabla\|x\|, x\rangle = \|x\|$. On the other hand, by the chain rule, $\nabla \frac{1}{2}\|x\|^2 = \|x\| \cdot \nabla\|x\|$, thus implying (44).

Let

$$K \triangleq \mathrm{Kernel}(\nabla\|\tilde{x}_{t-1}\|^2) = \left\{h : \left\langle \nabla \left\|\tilde{x}_{t-1}\right\|^2, h \right\rangle = 0\right\} \ , \ K' \triangleq \mathrm{Kernel}(\tilde{x}_{t-1}) = \{h : \langle h, \tilde{x}_{t-1}\rangle = 0\} \, .$$

We first claim that x_t can always be written as $x_t = \beta\tilde{x}_{t-1} + \gamma y$ for some $y \in K$ and for some scalars β, γ. Indeed, suppose that $x_t = \beta\tilde{x}_{t-1} + \gamma y + z$ for some $y \in K$ and $z \notin K$. Then we can rewrite x_t as

$$x_t = (\beta + \delta)\, \tilde{x}_{t-1} + (\gamma y - \delta\tilde{x}_{t-1} + z)$$

where $\delta = \frac{\left\langle \nabla \frac{1}{2}\|\tilde{x}_{t-1}\|^2, z \right\rangle}{\|\tilde{x}_{t-1}\|^2}$. It is enough to check that $(\gamma y - \delta\tilde{x}_{t-1} + z) \in K$. Indeed, using (44),

$$\left\langle \nabla \left\|\tilde{x}_{t-1}\right\|^2, -\delta\tilde{x}_{t-1} + z \right\rangle = -\delta \left\|\tilde{x}_{t-1}\right\|^2 + \left\langle \nabla \left\|\tilde{x}_{t-1}\right\|^2, z \right\rangle = 0 \ .$$

An analogous proof shows that we may always decompose any f_t as $f_t = -\alpha\nabla\frac{1}{2}\|\tilde{x}_{t-1}\|^2 + g$ for some $g \in K'$ and a scalar α. Hence,

$$\begin{aligned} &\langle f_t, x_t\rangle + \left(\|\tilde{x}_{t-1}\|^2 + \left\langle \nabla\tfrac{1}{2}\|\tilde{x}_{t-1}\|^2, x_t \right\rangle + C(T-t+1)\right)^{1/2} \\ &= -\alpha\beta\|\tilde{x}_{t-1}\|^2 + \gamma\langle g, y\rangle + \left(\|\tilde{x}_{t-1}\|^2 + \beta\|\tilde{x}_{t-1}\|^2 + C(T-t+1)\right)^{1/2} \end{aligned} \tag{45}$$

Given any $f_t = -\alpha\nabla\frac{1}{2}\left\|\tilde{x}_{t-1}\right\|^2 + g$, x_t can be picked with $y \in K$ that satisfies $\langle g, y\rangle \geq 0$. One can always do this because if for some y', $\langle g, y'\rangle < 0$ by picking $y = -y'$ we can ensure that $\langle g, y\rangle \geq 0$. Hence the minimizer f_t must be once such that $f_t = -\alpha\nabla\frac{1}{2}\left\|\tilde{x}_{t-1}\right\|^2$ and thus $\langle g, y\rangle = 0$. Now, it must be that $\alpha \geq 0$ so that x_t either increases the first term or second term but not both. Hence we conclude that $f_t = -\alpha\nabla\frac{1}{2}\left\|\tilde{x}_{t-1}\right\|^2$ for some $\alpha \geq 0$. It remains to determine the optimal α. Given such an f_t, the sup over x_t can be written as supremum over β of a concave function, which gives rise to the derivative condition

$$-\alpha\left\|\tilde{x}_{t-1}\right\|^2 + \frac{\left\|\tilde{x}_{t-1}\right\|^2}{2\sqrt{\left\|\tilde{x}_{t-1}\right\|^2 + \beta\left\|\tilde{x}_{t-1}\right\|^2 + C(T-t+1)}} = 0$$

At this point it is clear that the value of

$$\alpha = \frac{1}{2\sqrt{\left\|\tilde{x}_{t-1}\right\|^2 + C(T-t+1)}} \tag{46}$$

forces $\beta = 0$. Let us in fact show that this value is optimal. We have

$$\frac{1}{4\alpha^2} = \left\|\tilde{x}_{t-1}\right\|^2 + \beta\left\|\tilde{x}_{t-1}\right\|^2 + C(T-t+1)$$

Plugging this value of β back, we optimize

$$\frac{1}{4\alpha} + \alpha\left\|\tilde{x}_{t-1}\right\|^2 + \alpha C(T-t+1)$$

over α and obtain the value given in (46). With this value, we have the familiar update (43). Plugging back the value of α, we find that $\beta = 0$. We conclude that f_t defined in (43) in fact coincides with the optimal solution (42).

Arriving at the Relaxation The derivation of the relaxation is immediate:

$$\mathfrak{R}_T(\mathcal{F}|x_1,\dots,x_t) = \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_{t+1:T}} \left\| \sum_{s=t+1}^{T} \epsilon_s \mathbf{x}_{s-t}(\epsilon_{t+1:s-1}) - \sum_{s=1}^{t} x_s \right\| \tag{47}$$

$$\leq \sup_{\mathbf{x}} \sqrt{\mathbb{E}_{\epsilon_{t+1:T}} \left\| \sum_{s=t+1}^{T} \epsilon_s \mathbf{x}_{s-t}(\epsilon_{t+1:s-1}) - \sum_{s=1}^{t} x_s \right\|^2} \tag{48}$$

$$\leq \sup_{\mathbf{x}} \sqrt{\left\| \sum_{s=1}^{t} x_s \right\|^2 + C\mathbb{E}_{\epsilon_{t+1:T}} \sum_{s=t+1}^{T} \|\epsilon_s \mathbf{x}_{s-t}(\epsilon_{t+1:s-1})\|^2} \tag{49}$$

where the last step is due to the smoothness of the norm and the fact that the first-order terms disappear under the expectation. The sum of norms is now upper bounded by $T-t$, thus removing the dependence on the "future", and we arrive at

$$\sqrt{\left\| \sum_{s=1}^{t} x_s \right\|^2 + C(T-t)} \leq \sqrt{\left\| \sum_{s=1}^{t-1} x_s \right\|^2 + \left\langle \nabla \tfrac{1}{2} \left\| \sum_{s=1}^{t-1} x_s \right\|^2, x_t \right\rangle + C(T-t+1)}$$

as a relaxation on the sequential Rademacher complexity. □

Proof of Lemma 8. We shall first establish the admissibility of the relaxation specified. To show admissibility, let us first check the initial condition:

$$\begin{aligned}
\mathbf{Rel}_k\left(\mathcal{F}_{r(k;x_1,\dots,x_t)}|y_1,\dots,y_k\right) &= -\left\langle \hat{f}_t, \tilde{y}_k \right\rangle + 2\min\left\{1, \frac{k}{\tilde{\sigma}_t}\right\} \sqrt{\left\| \sum_{j=1}^{k-1} y_j \right\|^2 + \left\langle \nabla \tfrac{1}{2} \left\| \sum_{j=1}^{k-1} y_j \right\|^2, y_k \right\rangle + C} \\
&\geq -\left\langle \hat{f}_t, \tilde{y}_k \right\rangle + 2\min\left\{1, \frac{k}{\tilde{\sigma}_t}\right\} \sqrt{\|\tilde{y}_k\|^2} \\
&\geq -\left\langle \hat{f}_t, \tilde{y}_k \right\rangle + \sup_{f:\left\|f-\hat{f}_t\right\| \leq 2\min\{1,\frac{k}{\tilde{\sigma}_t}\}} \left\langle f - \hat{f}_t, -\tilde{y}_k \right\rangle \\
&\geq -\inf_{f:\left\|f-\hat{f}_t\right\| \leq 2\min\{1,\frac{k}{\tilde{\sigma}_t}\}} \sum_{j=1}^{k} \langle f, y_j \rangle
\end{aligned}$$

Now, for the recurrence, we have

$$\begin{aligned}
&\langle f_i, y_i \rangle + \sup_{\mathbf{y}} \mathbb{E}_{\epsilon} \left[\sup_{f:\|f-\hat{f}_t\| \leq 2\min\left\{1,\frac{k}{\tilde{\sigma}_t}\right\}} \left\langle f, \sum_{j=1}^{k-i} \epsilon_j \mathbf{y}_j(\epsilon) - \sum_{j=1}^{i} y_j \right\rangle \right] \\
&= \langle f_i, y_i \rangle - \left\langle \hat{f}_t, \sum_{j=1}^{i} y_j \right\rangle + \sup_{\mathbf{y}} \mathbb{E}_{\epsilon} \left[\sup_{f:\|f-\hat{f}_t\| \leq 2\min\left\{1,\frac{k}{\tilde{\sigma}_t}\right\}} \left\langle f - \hat{f}_t, \sum_{j=1}^{k-i} \epsilon_j \mathbf{y}_j(\epsilon) - \sum_{j=1}^{i} y_j \right\rangle \right] \\
&\leq \langle f_i, y_i \rangle - \left\langle \hat{f}_t, \sum_{j=1}^{i} y_j \right\rangle + 2\min\left\{1, \frac{k}{\tilde{\sigma}_t}\right\} \sup_{\mathbf{y}} \mathbb{E}_{\epsilon} \left[\left\| \sum_{j=1}^{k-i} \epsilon_j \mathbf{y}_j(\epsilon) - \sum_{j=1}^{i} y_j \right\| \right]
\end{aligned}$$

$$\leq \langle f_i, y_i \rangle - \left\langle \hat{f}_t, \sum_{j=1}^{i} y_j \right\rangle + 2\min\left\{1, \frac{k}{\tilde{\sigma}_t}\right\} \sqrt{\|\tilde{y}_i\|^2 + C(k-i)}$$

$$\leq \langle f_i, y_i \rangle - \langle \hat{f}_t, \tilde{y}_i \rangle + 2\min\left\{1, \frac{k}{\tilde{\sigma}_t}\right\} \sqrt{\|\tilde{y}_i\|^2 + \left\langle \nabla \tfrac{1}{2} \|\tilde{y}_{i-1}\|^2, \mathbf{y}_i \right\rangle + C(k-i+1)}$$

$$= \langle f_i - \hat{f}_t, y_i \rangle - \langle \hat{f}_t, \tilde{y}_{i-1} \rangle + 2\min\left\{1, \frac{k}{\tilde{\sigma}_t}\right\} \sqrt{\|\tilde{y}_{i-1}\|^2 + \left\langle \nabla \tfrac{1}{2} \|\tilde{y}_{i-1}\|^2, \mathbf{x}_i \right\rangle + C(k-i+1)}$$

and we start block at $\hat{f}_t$. For the first block, this value is 0 but later on it is the empirical risk minimizer. We therefore get a mixture of Follow the Leader (FTL) and Gradient Descent (GD) algorithms. If block size is 1, we get FTL only, and when the block size is T we get GD only. In general, however, the resulting method is an interesting mixture of the two. Using the arguments of Proposition 10, the update in the block is given by

$$f_{t+i} = \hat{f}_t - \max\left\{1, \frac{k}{\tilde{\sigma}_t}\right\} \frac{-\nabla \tfrac{1}{2} \|\tilde{y}_{i-1}^2\|}{\sqrt{\|\tilde{y}_{i-1}\|^2 + C(k-i+1)}}$$

Now that we have shown the admissibility of the relaxation and the form of update obtained by the relaxation we turn to the bounds on the regret specified in the lemma. We shall provide these bounds using Lemma 7. We will split the analysis to two cases, one when $\alpha > 1/2$ and other when $\alpha \leq 1/2$.

Case $\alpha > \frac{1}{2}$:
To start note that since we initialize the block lengths with the doubling trick, that is initialize block lengths as $1, 2, 4, \ldots$ hence, after t rounds the maximum length of current block say k can be at most $2t$ and so $\sqrt{k} \leq \sqrt{2t}$. Now let us first consider the case when $\alpha > \frac{1}{2}$. In this case, since $\tilde{\sigma}_t = Bt^\alpha$, we can conclude that the condition $\tilde{\sigma}_t \geq \sqrt{k}$ is satisfied as long as $t^{\alpha - \frac{1}{2}} \geq \frac{\sqrt{2}}{B}$. Since we are considering the case when $\alpha > \frac{1}{2}$ we can conclude that for all rounds larger than $\sqrt{2}/B$, the blocking strategy always picks block size of 1. Hence applying Lemma 7 we conclude that in the case when $1 > \alpha > 1/2$ (or when $\alpha = 1/2$ and $B \geq \sqrt{2}$),

$$\mathbf{Reg}_T \leq \sum_{t=1}^{T} \frac{1}{\tilde{\sigma}_t} = \sum_{t=1}^{T} \frac{1}{Bt^\alpha} = O(T^{1-\alpha}/B)$$

Also note that for the case when $\alpha = 1$, the summation is bounded by $O(\log T)$ and so

$$\mathbf{Reg}_T \leq \sum_{t=1}^{T} \frac{1}{\tilde{\sigma}_t} = \sum_{t=1}^{T} \frac{1}{Bt^\alpha} = O(\log T/B)$$

Case $\alpha \leq \frac{1}{2}$:
Now we consider the case when $\alpha < 1/2$. Say we are at start of some block $t = 2^m$. The initial block length then is $2t$ by the doubling trick initialization. Now within this block, the adaptive algorithm continues with this current block until the point when the square-root of the remaining number of rounds in the block say k becomes smaller than $\tilde{\sigma}_{t+(2t-k)}$. That is until

$$\sqrt{k} \leq B(3t-k)^\alpha \tag{50}$$

The regret on this block can be bounded using Lemma 7 (notice that here we use the lemma for the algorithm within a sub-block initialized by the doubling trick rather than on the entire T rounds). The regret on this

block is bounded as :

$$\begin{aligned}
\mathbf{Rel}_{2t-k}\left(\mathcal{F}_{r(x_1,\ldots,x_t)}\right) + \sum_{i=2t-k+1}^{2t} \mathbf{Rel}_1\left(\mathcal{F}_{r(x_1,\ldots,x_i)}\right) &\le \sqrt{2t-k} + \sum_{j=2t-k+1}^{2t} \frac{1}{Bj^\alpha} \\
&\le \sqrt{2t} + \sum_{j=2t-k+1}^{2t} \frac{1}{Bj^\alpha} \\
&\le \sqrt{2t} + \frac{1}{B}\left((2t+1)^{1-\alpha} - (2t-k+1)^{1-\alpha}\right) \\
&\le \sqrt{2t} + \frac{k^{1-\alpha}}{B} \\
&\le \sqrt{2t} + \frac{B^{2(1-\alpha)}(3t)^{2\alpha(1-\alpha)}}{B} \qquad \text{(using Eq. (50))} \\
&\le \sqrt{2t} + B^{2(1-\alpha)-1}\sqrt{3t} \\
&\le \sqrt{12\ t}
\end{aligned}$$

Hence overall regret is bounded as

$$\mathbf{Reg}_T \le \sum_{i=1}^{\lceil \log_2 T\rceil+1} \sqrt{12\ \times 2^{i-1}} \le \sqrt{12} \sum_{i=1}^{\lceil \log_2 T\rceil+1} 2^{(i-1)/2} \le O(\sqrt{T})$$

This concludes the proof. □

Proof of Lemma 9. Notice that by the doubling trick for the first at most 2τ rounds we simply play the experts algorithm, thus suffering a maximum regret that is the minimum of τ and $4\sqrt{\tau \log |\mathcal{F}|}$. After these initial number of rounds, consider any round t at which we start a new block with the blocking strategy described above. The first sub-block given by the blocking strategy is of length at most k, thanks to our assumption about the gap between the leader and the second-best action. Clearly the minimizer of the cumulative loss up to t rounds already played, $\operatorname{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{t} \ell(f, x_i)$, is going to be the leader at least for the next k rounds. Hence for this block we suffer no regret. Now when we use the same blocking strategy repeatedly, due to the same reasoning, we end up playing the same leader for the rest of the game only in chunks of size k, and thus suffer no regret for the rest of the game. □

Proof of Proposition 10. We would like to show that, with the distribution $q_t^\star$ defined in (20),

$$\max_{y_t\in\{\pm1\}} \left\{ \mathop{\mathbb{E}}_{\hat{y}_t\sim q_t^\star} |\hat{y}_t - y_t| + \mathbf{Rel}_T\left(\mathcal{F}\middle|(x^t, y^t)\right)\right\} \le \mathbf{Rel}_T\left(\mathcal{F}\middle|(x^{t-1}, y^{t-1})\right)$$

for any $x_t \in \mathcal{X}$. Let $\sigma \in \{\pm1\}^{t-1}$ and $\sigma_t \in \{\pm1\}$. We have

$$\begin{aligned}
&\mathbf{Rel}_T\left(\mathcal{F}\middle|(x^t, y^t)\right) - 2\lambda(T-t) \\
&= \frac{1}{\lambda}\log\left(\sum_{(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)), T-t)\exp\{-\lambda L_{t-1}(\sigma)\}\exp\{-\lambda|\sigma_t - y_t|\}\right) \\
&\le \frac{1}{\lambda}\log\left(\sum_{\sigma_t\in\{\pm1\}} \exp\{-\lambda|\sigma_t - y_t|\} \sum_{\sigma:(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)), T-t)\exp\{-\lambda L_{t-1}(\sigma)\}\right)
\end{aligned}$$

Just as in the proof of Proposition 3, we may think of the two choices σ_t as the two experts whose weighting $q_t^\star$ is given by the sum involving the Littlestone's dimension of subsets of $\mathcal{F}$. Introducing the normalization

term, we arrive at the upper bound

$$\frac{1}{\lambda}\log\left(\mathbb{E}_{\sigma_t\sim q_t^*}\exp\left\{-\lambda|\sigma_t-y_t|\right\}\right)+\frac{1}{\lambda}\log\left(\sum_{\sigma_t\in\{\pm1\}}\sum_{\sigma:(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)),T-t)\exp\left\{-\lambda L_{t-1}(\sigma)\right\}\right)$$
$$\le -\mathbb{E}_{\sigma_t\sim q_t^*}|\sigma_t-y_t|+2\lambda+\frac{1}{\lambda}\log\left(\sum_{\sigma_t\in\{\pm1\}}\sum_{\sigma:(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)),T-t)\exp\left\{-\lambda L_{t-1}(\sigma)\right\}\right)$$

The last step is due to Lemma A.1 in [7]. It remains to show that the log normalization term is upper bounded by the relaxation at the previous step:

$$\begin{aligned}
&\frac{1}{\lambda}\log\left(\sum_{\sigma_t\in\{\pm1\}}\sum_{\sigma:(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)),T-t)\exp\left\{-\lambda L_{t-1}(\sigma)\right\}\right)\\
&\le\frac{1}{\lambda}\log\left(\sum_{\sigma\in\mathcal{F}|_{x^{t-1}}}\exp\left\{-\lambda L_{t-1}(\sigma)\right\}\sum_{\sigma_t\in\{\pm1\}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)),T-t)\right)\\
&\le\frac{1}{\lambda}\log\left(\sum_{\sigma\in\mathcal{F}|_{x^{t-1}}}\exp\left\{-\lambda L_{t-1}(\sigma)\right\}g(\mathrm{Ldim}(\mathcal{F}_{t-1}(\sigma)),T-t+1)\right)\\
&=\mathbf{Rel}_T\left(\mathcal{F}\middle|(x^{t-1},y^{t-1})\right)
\end{aligned}$$

To justify the last inequality, note that $\mathcal{F}_{t-1}(\sigma)=\mathcal{F}_t(\sigma,+1)\cup\mathcal{F}_t(\sigma,-1)$ and at most one of $\mathcal{F}_t(\sigma,+1)$ or $\mathcal{F}_t(\sigma,-1)$ can have Littlestone's dimension $\mathrm{Ldim}(\mathcal{F}_{t-1}(\sigma))$. We now appeal to the recursion

$$g(d,T-t)+g(d-1,T-t)\le g(d,T-t+1)$$

where $g(d,T-t)$ is the size of the zero cover for a class with Littlestone's dimension d on the worst-case tree of depth $T-t$ (see [16]). This completes the proof of admissibility.

Alternative Method Let us now derive the algorithm given in (21) and prove its admissibility. Once again, consider the optimization problem

$$\max_{y_t\in\{\pm1\}}\left\{\mathop{\mathbb{E}}_{\hat{y}_t\sim q_t^*}|\hat{y}_t-y_t|+\mathbf{Rel}_T\left(\mathcal{F}\middle|(x^t,y^t)\right)\right\}$$

with the relaxation

$$\mathbf{Rel}_T\left(\mathcal{F}\middle|(x^t,y^t)\right)=\frac{1}{\lambda}\log\left(\sum_{\sigma\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma)),T-t)\exp\left\{-\lambda L_t(\sigma)\right\}\right)+\frac{\lambda}{2}(T-t)$$

The maximum can be written explicitly, as in Section 6:

$$\begin{aligned}
\max\Bigg\{&1-q_t^*+\frac{1}{\lambda}\log\left(\sum_{(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)),T-t)\exp\left\{-\lambda L_{t-1}(\sigma)\right\}\exp\left\{-\lambda(1-\sigma_t)\right\}\right),\\
&1+q_t^*+\frac{1}{\lambda}\log\left(\sum_{(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)),T-t)\exp\left\{-\lambda L_{t-1}(\sigma)\right\}\exp\left\{-\lambda(1+\sigma_t)\right\}\right)\Bigg\}
\end{aligned}$$

where we have dropped the $\frac{\lambda}{2}(T-t)$ term from both sides. Equating the two values, we obtain

$$2q_t^*=\frac{1}{\lambda}\log\frac{\sum_{(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)),T-t)\exp\left\{-\lambda L_{t-1}(\sigma)\right\}\exp\left\{-\lambda(1-\sigma_t)\right\}}{\sum_{(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)),T-t)\exp\left\{-\lambda L_{t-1}(\sigma)\right\}\exp\left\{-\lambda(1+\sigma_t)\right\}}$$

The resulting value becomes

$$
\begin{aligned}
&1+\frac{\lambda}{2}(T-t)+\frac{1}{2\lambda}\log\left\{\sum_{(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)),T-t)\exp\left\{-\lambda L_{t-1}(\sigma)\right\}\exp\left\{-\lambda(1-\sigma_t)\right\}\right\}\\
&\qquad+\frac{1}{2\lambda}\log\left\{\sum_{(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)),T-t)\exp\left\{-\lambda L_{t-1}(\sigma)\right\}\exp\left\{-\lambda(1+\sigma_t)\right\}\right\}\\
&=1+\frac{\lambda}{2}(T-t)+\frac{1}{\lambda}\mathbb{E}_\epsilon\log\left\{\sum_{(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)),T-t)\exp\left\{-\lambda L_{t-1}(\sigma)\right\}\exp\left\{-\lambda(1-\epsilon\sigma_t)\right\}\right\}\\
&\le 1+\frac{\lambda}{2}(T-t)+\frac{1}{\lambda}\log\left\{\sum_{(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)),T-t)\exp\left\{-\lambda L_{t-1}(\sigma)\right\}\mathbb{E}_\epsilon\exp\left\{-\lambda(1-\epsilon\sigma_t)\right\}\right\}
\end{aligned}
$$

for a Rademacher random variable $\epsilon\in\{\pm1\}$. Now,

$$\mathbb{E}_\epsilon\exp\left\{-\lambda(1-\epsilon\sigma_t)\right\}=e^{-\lambda}\mathbb{E}_\epsilon e^{\lambda\epsilon\sigma_t}\le e^{-\lambda}e^{\lambda^2/2}$$

Substituting this into the above expression, we obtain an upper bound of

$$\frac{\lambda}{2}(T-t+1)+\frac{1}{\lambda}\log\left\{\sum_{(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}} g(\mathrm{Ldim}(\mathcal{F}_t(\sigma,\sigma_t)),T-t)\exp\left\{-\lambda L_{t-1}(\sigma)\right\}\right\}$$

which completes the proof of admissibility using the same combinatorial argument as in the earlier part of the proof.

Arriving at the Relaxation Finally, we show that the relaxation we use arises naturally as an upper bound on the sequential Rademacher complexity. Fix a tree $\mathbf{x}$. Let $\sigma\in\{\pm1\}^{t-1}$ be a sequence of signs. Observe that given history $x^t=(x_1,\ldots,x_t)$, the signs $\epsilon\in\{\pm1\}^{T-t}$, and a tree $\mathbf{x}$, the function class $\mathcal{F}$ takes on only a finite number of possible values (σ,σ_t,ω) on $(x^t,\mathbf{x}(\epsilon))$. Here, $\mathbf{x}(\epsilon)$ denotes the sequences of values along the path ϵ. We have,

$$
\begin{aligned}
\sup_{\mathbf{x}}\mathbb{E}_\epsilon\sup_{f\in\mathcal{F}}\left\{2\sum_{i=1}^{T-t}\epsilon_i f(\mathbf{x}_i(\epsilon))-\sum_{i=1}^{t}|f(x_i)-y_i|\right\}&=\sup_{\mathbf{x}}\mathbb{E}_\epsilon\max_{\sigma_t\in\{\pm1\}}\max_{(\sigma,\omega):(\sigma,\sigma_t,\omega)\in\mathcal{F}|_{(x^t,\mathbf{x}(\epsilon))}}\left\{2\sum_{i=1}^{T-t}\epsilon_i\omega_i-\sum_{i=1}^{t}|\sigma_i-y_i|\right\}\\
&\le\sup_{\mathbf{x}}\mathbb{E}_\epsilon\max_{\sigma_t\in\{\pm1\}}\max_{\sigma:(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}}\max_{\mathbf{v}\in V(\mathcal{F}(\sigma,\sigma_t),\mathbf{x})}\left\{2\sum_{i=1}^{T-t}\epsilon_i\mathbf{v}_i(\epsilon)-\sum_{i=1}^{t}|\sigma_i-y_i|\right\}
\end{aligned}
$$

where $\mathcal{F}|_{(x^t,\mathbf{x}(\epsilon))}$ is the projection of $\mathcal{F}$ onto $(x^t,\mathbf{x}(\epsilon))$, $\mathcal{F}(\sigma,\sigma_t)=\{f\in\mathcal{F}:f(x^t)=(\sigma,\sigma_t)\}$, and $V(\mathcal{F}(\sigma,\sigma_t),\mathbf{x})$ is the zero-cover of the set $\mathcal{F}(\sigma,\sigma_t)$ on the tree $\mathbf{x}$. We then have the following relaxation:

$$\frac{1}{\lambda}\log\left(\sup_{\mathbf{x}}\mathbb{E}_\epsilon\sum_{\sigma_t\in\{\pm1\}}\sum_{\sigma:(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}}\sum_{\mathbf{v}\in V(\mathcal{F}(\sigma,\sigma_t),\mathbf{x})}\exp\left\{2\lambda\sum_{i=1}^{T-t}\epsilon_i\mathbf{v}_i(\epsilon)-\lambda L_t(\sigma,\sigma_t)\right\}\right)$$

where $L_t(\sigma,\sigma_t)=\sum_{i=1}^{t}|\sigma_i-y_i|$. The latter quantity can be factorized:

$$
\begin{aligned}
&\frac{1}{\lambda}\log\left(\sup_{\mathbf{x}}\sum_{\sigma_t\in\{\pm1\}}\sum_{\sigma:(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}}\exp\left\{-\lambda L_t(\sigma,\sigma_t)\right\}\mathbb{E}_\epsilon\sum_{\mathbf{v}\in V(\mathcal{F}(\sigma,\sigma_t),\mathbf{x})}\exp\left\{2\lambda\sum_{i=1}^{T-t}\epsilon_i\mathbf{v}_i(\epsilon)\right\}\right)\\
&\le\frac{1}{\lambda}\log\left(\sup_{\mathbf{x}}\sum_{\sigma_t\in\{\pm1\}}\sum_{\sigma:(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}}\exp\left\{-\lambda L_t(\sigma,\sigma_t)\right\}\mathrm{card}(V(\mathcal{F}(\sigma,\sigma_t),\mathbf{x}))\exp\left\{2\lambda^2(T-t)\right\}\right)\\
&\le\frac{1}{\lambda}\log\left(\sum_{\sigma_t\in\{\pm1\}}\exp\left\{-\lambda|\sigma_t-y_t|\right\}\sum_{\sigma:(\sigma,\sigma_t)\in\mathcal{F}|_{x^t}}g(\mathrm{Ldim}(\mathcal{F}(\sigma,\sigma_t)),T-t)\exp\left\{-\lambda L_{t-1}(\sigma)\right\}\right)+2\lambda(T-t)\ .
\end{aligned}
$$

This concludes the derivation of the relaxation.

□

Proof of Lemma 11. We first exhibit the proof for the convex loss case. To show admissibility using the particular randomized strategy q_t given in the lemma, we need to show that

$$\sup_{x_t}\left\{\mathop{\mathbb{E}}_{f\sim q_t}\left[\ell(f,x_t)\right]+\mathbf{Rel}_T\left(\mathcal{F}|x_1,\dots,x_t\right)\right\}\le\mathbf{Rel}_T\left(\mathcal{F}|x_1,\dots,x_{t-1}\right)$$

The strategy q_t proposed by the lemma is such that we first draw $x_{t+1},\dots,x_T\sim D$ and $\epsilon_{t+1},\dots\epsilon_T$ Rademacher random variables, and then based on this sample pick $f_t=f_t(x_{t+1:T},\epsilon_{t+1:T})$ as in (24). Hence,

$$\begin{aligned}
&\sup_{x_t}\left\{\mathop{\mathbb{E}}_{f\sim q_t}\left[\ell(f,x_t)\right]+\mathbf{Rel}_T\left(\mathcal{F}|x_1,\dots,x_t\right)\right\}\\
&\quad=\sup_{x_t}\left\{\mathop{\mathbb{E}}_{\substack{\epsilon_{t+1:T}\\x_{t+1:T}}}\ell(f_t,x)+\mathop{\mathbb{E}}_{\substack{\epsilon_{t+1:T}\\x_{t+1:T}}}\sup_{f\in\mathcal{F}}\left[C\sum_{i=t+1}^{T}\epsilon_i\ell(f,x_i)-L_t(f)\right]\right\}\\
&\quad\le\mathop{\mathbb{E}}_{\substack{\epsilon_{t+1:T}\\x_{t+1:T}}}\sup_{x_t}\left\{\ell(f_t,x)+\sup_{f\in\mathcal{F}}\left[C\sum_{i=t+1}^{T}\epsilon_i\ell(f,x_i)-L_t(f)\right]\right\}
\end{aligned}$$

where $L_t(f)=\sum_{i=1}^{t}\ell(f,x_i)$. Observe that our strategy "matched the randomness" arising from the relaxation! Now, with f_t defined as

$$f_t=\mathop{\mathrm{argmin}}_{g\in\mathcal{F}}\sup_{x_t\in\mathcal{X}}\left\{\ell(g,x_t)+\sup_{f\in\mathcal{F}}\left[C\sum_{i=t+1}^{T}\epsilon_i\ell(f,x_i)-L_t(f)\right]\right\}$$

for any given $x_{t+1:T},\epsilon_{t+1:T}$, we have

$$\sup_{x_t}\left\{\ell(f_t,x_t)+\sup_{f\in\mathcal{F}}\left[C\sum_{i=t+1}^{T}\epsilon_i\ell(f,x_i)-L_t(f)\right]\right\}=\inf_{g\in\mathcal{F}}\sup_{x_t}\left\{\ell(g,x_t)+\sup_{f\in\mathcal{F}}\left[C\sum_{i=t+1}^{T}\epsilon_i\ell(f,x_i)-L_t(f)\right]\right\}$$

We can conclude that for this choice of q_t,

$$\begin{aligned}
&\sup_{x_t}\left\{\mathop{\mathbb{E}}_{f\sim q_t}\left[\ell(f,x_t)\right]+\mathbf{Rel}_T\left(\mathcal{F}|x_1,\dots,x_t\right)\right\}\le\mathop{\mathbb{E}}_{\substack{\epsilon_{t+1:T}\\x_{t+1:T}}}\inf_{g\in\mathcal{F}}\sup_{x_t}\left\{\ell(g,x_t)+\sup_{f\in\mathcal{F}}\left[C\sum_{i=t+1}^{T}\epsilon_i\ell(f,x_i)-L_t(f)\right]\right\}\\
&\quad=\mathop{\mathbb{E}}_{\substack{\epsilon_{t+1:T}\\x_{t+1:T}}}\inf_{g\in\mathcal{F}}\sup_{p_t\in\Delta(\mathcal{X})}\mathop{\mathbb{E}}_{x_t\sim p_t}\left[\ell(g,x_t)+\sup_{f\in\mathcal{F}}\left[C\sum_{i=t+1}^{T}\epsilon_i\ell(f,x_i)-L_t(f)\right]\right]\\
&\quad=\mathop{\mathbb{E}}_{\substack{\epsilon_{t+1:T}\\x_{t+1:T}}}\sup_{p\in\Delta(\mathcal{X})}\inf_{g\in\mathcal{F}}\left\{\mathop{\mathbb{E}}_{x_t\sim p}\left[\ell(g,x_t)\right]+\mathop{\mathbb{E}}_{x_t\sim p}\left[\sup_{f\in\mathcal{F}}C\sum_{i=t+1}^{T}\epsilon_i\ell(f,x_i)-L_t(f)\right]\right\}
\end{aligned}$$

In the last step we appealed to the minimax theorem which holds as loss is convex in g and $\mathcal{F}$ is a compact convex set and the term in the expectation is linear in p_t, as it is an expectation. The last expression can

be written as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{\epsilon_{t+1:T}\\ x_{t+1:T}}} \sup_{p\in\Delta(\mathcal{X})} \mathbb{E}_{x_t\sim p} \sup_{f\in\mathcal{F}}\left[C\sum_{i=t+1}^{T}\epsilon_i\ell(f,x_i) - L_{t-1}(f) + \inf_{g\in\mathcal{F}} \mathop{\mathbb{E}}_{x_t\sim p}[\ell(g,x_t)] - \ell(f,x_t)\right]\\
&\le \mathop{\mathbb{E}}_{\substack{\epsilon_{t+1:T}\\ x_{t+1:T}}} \sup_{p\in\Delta(\mathcal{X})} \mathbb{E}_{x_t\sim p} \sup_{f\in\mathcal{F}}\left[C\sum_{i=t+1}^{T}\epsilon_i\ell(f,x_i) - L_{t-1}(f) + \mathop{\mathbb{E}}_{x_t\sim p}[\ell(f,x_t)] - \ell(f,x_t)\right]\\
&\le \mathop{\mathbb{E}}_{\substack{\epsilon_{t+1:T}\\ x_{t+1:T}}} \mathbb{E}_{x_t\sim D}\mathbb{E}_{\epsilon_t} \sup_{f\in\mathcal{F}}\left[C\sum_{i=t+1}^{T}\epsilon_i\ell(f,x_i) - L_{t-1}(f) + C\epsilon_t\ell(f,x_t)\right]\\
&= \mathbf{Rel}_T\left(\mathcal{F}|x_1,\dots,x_{t-1}\right)
\end{aligned}$$

Last inequality is by Assumption 1, using which we can replace a draw from supremum over distributions by a draw from the "equivalently bad" fixed distribution D by suffering an extra factor of C multiplied to that random instance.

The key step where we needed convexity was to use minimax theorem to swap infimum and supremum inside the expectation. In general the minimax theorem need not hold. In the non-convex scenario this is the reason we add the extra randomization through $\hat{q}_t$. The non-convex case has a similar proof except that we have expectation w.r.t. $\hat{q}_t$ extra on each round which essentially convexifies our loss and thus allows us to appeal to the minimax theorem. □

Proof of Lemma 12. Let $w \in \mathbb{R}^N$ be arbitrary. Throughout this proof, let $\epsilon \in \{\pm 1\}$ be a single Rademacher random variable, rather than a vector. To prove (27), observe that

$$\begin{aligned}
\sup_{p\in\Delta(\mathcal{X})} \mathop{\mathbb{E}}_{x_t\sim p}\left\|w + \mathop{\mathbb{E}}_{x\sim p}[x] - x_t\right\|_\infty &\le \sup_{p\in\Delta(\mathcal{X})} \mathop{\mathbb{E}}_{x,x'\sim p}\|w + x' - x\|_\infty\\
&= \sup_{p\in\Delta(\mathcal{X})} \mathop{\mathbb{E}}_{x,x'\sim p}\mathbb{E}_\epsilon\|w + \epsilon(x'-x)\|_\infty\\
&\le \sup_{x,x'\in\mathcal{X}} \mathbb{E}_\epsilon\|w + \epsilon(x'-x)\|_\infty\\
&\le \sup_{x'\in\mathcal{X}} \mathbb{E}_\epsilon\|w/2 + \epsilon x'\|_\infty + \sup_{x\in\mathcal{X}}\mathbb{E}_\epsilon\|w/2 - \epsilon x\|_\infty\\
&= \sup_{x\in\mathcal{X}}\mathbb{E}_\epsilon \max_{i\in[N]}|w_i + 2\epsilon x_i|
\end{aligned}$$

The supremum over $x \in \mathcal{X}$ is achieved at the vertices of $\mathcal{X}$ since the expected maximum is a convex function. It remains to prove the identity

$$\max_{x\in\{\pm1\}^N} \mathbb{E}_\epsilon \max_{i\in[N]}|w_i + 2\epsilon x_i| \le \mathop{\mathbb{E}}_{x\sim D}\mathop{\mathbb{E}}_{\epsilon} \max_{i\in[N]}|w_i + 6\epsilon x_i| \tag{51}$$

Let $i^* = \operatorname{argmax}_i |w_i|$ and $j^* = \operatorname{argmax}_{i\neq i^*} |w_i|$ be the coordinates with largest and second-largest magnitude. If $|w_{i^*}| - |w_{j^*}| \ge 4$, the statement follows since, for any $x \in \{\pm1\}^N$ and $\epsilon \in \{\pm1\}$,

$$\max_{i\neq i^*}|w_i + 2\epsilon x_i| \le \max_{i\neq i^*}|w_i| + 2 \le |w_{i^*}| - 2 \le |w_{i^*} + 2\epsilon x_{i^*}|,$$

and thus

$$\max_{x\in\{\pm1\}^N} \mathbb{E}_\epsilon \max_{i\in[N]}|w_i + 2\epsilon x_i| = \max_{x\in\{\pm1\}^N} \mathbb{E}_\epsilon |w_{i^*} + 2\epsilon x_{i^*}| = |w_{i^*}| = \mathbb{E}_{x,\epsilon}|w_{i^*} + 6\epsilon x_{i^*}| \le \mathbb{E}_{x,\epsilon}\max_i |w_i + 6\epsilon x_i|.$$

It remains to consider the case when $|w_{i^*}| - |w_{j^*}| < 4$. We have that

$$\begin{aligned}
\mathbb{E}_{x,\epsilon}\max_{i\in[N]}|w_i + 6\epsilon x_i| \ge \mathbb{E}_{x,\epsilon}\max_{i\in\{i^*,j^*\}}|w_i + 6\epsilon x_i| &\ge \frac{1}{2}(|w_{i^*}| + 6) + \frac{1}{4}(|w_{i^*}| - 6) + \frac{1}{4}(|w_{j^*}| + 6) \ge |w_{i^*}| + 2 &(52)\\
&\ge \max_{x\in\{\pm1\}^N}\mathbb{E}_\epsilon\max_{i\in[N]}|w_i + 2\epsilon x_i|\,, &(53)
\end{aligned}$$

where 1/2 is the probability that $\epsilon x_{i^*} = sign(w_{i^*})$, the second event of probability 1/4 is the event that $\epsilon x_{i^*} \neq sign(w_{i^*})$ and $\epsilon x_{j^*} \neq sign(w_{j^*})$, while the third event of probability 1/4 is that $\epsilon x_{i^*} \neq sign(w_{i^*})$ and $\epsilon x_{j^*} = sign(w_{j^*})$. □

Proof of Lemma 13. Let $w \in \mathbb{R}^N$ be arbitrary. Just as in the proof of Lemma 12, we need to show

$$\max_{x\in\{\pm 1\}^N} \mathbb{E}_\epsilon \max_{i\in[N]} |w_i + 2\epsilon x_i| \leq \mathop{\mathbb{E}}_{x\sim D} \mathop{\mathbb{E}}_{\epsilon} \max_{i\in[N]} |w_i + C\epsilon x_i| \tag{54}$$

Let $i^* = \operatorname{argmax}_i |w_i|$ and $j^* = \operatorname{argmax}_{i\neq i^*} |w_i|$ be the coordinates with largest and second-largest magnitude. If $|w_{i^*}| - |w_{j^*}| \geq 4$, the statement follows exactly as in Lemma 12. It remains to consider the case when $|w_{i^*}| - |w_{j^*}| < 4$. In this case first note that,

$$\max_{x\in\{\pm 1\}^N} \mathbb{E}_\epsilon \max_{i\in[N]} |w_i + 2\epsilon x_i| \leq |w_{i^*}| + 2$$

On the other hand, since the distribution we consider is symmetric, with probability 1/2 its sign is negative and with remaining probability positive. Define $\sigma_{i^*} = \operatorname{sign}(x_{i^*})$, $\sigma_{j^*} = \operatorname{sign}(x_{j^*})$, $\tau_{i^*} = \operatorname{sign}(w_{i^*})$, and $\tau_{j^*} = \operatorname{sign}(w_{j^*})$. Since each coordinate is drawn i.i.d., using conditional expectations we have,

$$\begin{aligned}
&\mathbb{E}_{x,\epsilon} \max_i |w_i + C\epsilon x_i| = \mathbb{E}_x \max_i |w_i + Cx_i| \\
&\geq \frac{\mathbb{E}_x\left[|w_{i^*} + Cx_{i^*}| \mid \sigma_{i^*} = \tau_{i^*}\right]}{2} + \frac{\mathbb{E}_x\left[|w_{j^*} + Cx_{j^*}| \mid \sigma_{i^*} \neq \tau_{i^*}, \sigma_{j^*} = \tau_{j^*}\right]}{4} + \frac{\mathbb{E}\left[|w_{i^*} + Cx_{i^*}| \mid \sigma_{i^*} \neq \tau_{i^*}, \sigma_{j^*} \neq \tau_{j^*}\right]}{4} \\
&\geq \frac{\mathbb{E}_x\left[|w_{i^*}| + C|x_{i^*}| \mid \sigma_{i^*} = \tau_{i^*}\right]}{2} + \frac{\mathbb{E}_x\left[|w_{j^*}| + C|x_{j^*}| \mid \sigma_{i^*} \neq \tau_{i^*}, \sigma_{j^*} = \tau_{j^*}\right]}{4} + \frac{\mathbb{E}\left[|w_{i^*}| - C|x_{i^*}| \mid \sigma_{i^*} \neq \tau_{i^*}, \sigma_{j^*} \neq \tau_{j^*}\right]}{4} \\
&= \frac{\mathbb{E}\left[|w_{i^*}| + C|x_{i^*}| \mid \sigma_{i^*} = \tau_{i^*}\right]}{2} + \frac{\mathbb{E}\left[|w_{j^*}| + C|x_{j^*}| \mid \sigma_{j^*} = \tau_{j^*}\right]}{4} + \frac{\mathbb{E}\left[|w_{i^*}| - C|x_{i^*}| \mid \sigma_{i^*} \neq \tau_{i^*}\right]}{4} \\
&= \frac{|w_{i^*}| + C\mathbb{E}\left[|x_{i^*}| \mid \sigma_{i^*} = \tau_{i^*}\right]}{2} + \frac{|w_{j^*}| + C\mathbb{E}\left[|x_{j^*}| \mid \sigma_{j^*} = \tau_{j^*}\right]}{4} + \frac{|w_{i^*}| - C\mathbb{E}\left[|x_{i^*}| \mid \sigma_{i^*} \neq \tau_{i^*}\right]}{4} \\
&= \frac{2|w_{i^*}| + |w_{j^*}| + 3C\mathbb{E}\left[|x_{i^*}| \mid \sigma_{i^*} = \tau_{i^*}\right]}{4} + \frac{|w_{i^*}| - C\mathbb{E}\left[|x_{i^*}| \mid \sigma_{i^*} \neq \tau_{i^*}\right]}{4} \\
&= \frac{3|w_{i^*}| + |w_{j^*}| + 2C\mathbb{E}\left[|x_{i^*}| \mid \sigma_{i^*} = \tau_{i^*}\right]}{4}
\end{aligned}$$

Now since we are in the case when $|w_{i^*}| - |w_{j^*}| < 4$ we see that

$$\mathbb{E}_{x,\epsilon} \max_i |w_i + C\epsilon x_i| \geq \frac{3|w_{i^*}| + |w_{j^*}| + 2C\mathbb{E}\left[|x_{i^*}| \mid \sigma_{i^*} = \tau_{i^*}\right]}{4} \geq \frac{4|w_{i^*}| + 2C\mathbb{E}\left[|x_{i^*}| \mid \sigma_{i^*} = \tau_{i^*}\right] - 4}{4}$$

On the other hand, as we already argued,

$$\max_{x\in\{\pm 1\}^N} \mathbb{E}_\epsilon \max_{i\in[N]} |w_i + 2\epsilon x_i| \leq |w_{i^*}| + 2$$

Hence, as long as

$$\frac{C\ \mathbb{E}\left[|x_{i^*}| \mid \sigma_{i^*} = \tau_{i^*}\right] - 2}{2} \geq 2$$

or, in other words, as long as

$$C \geq 6/\mathbb{E}\left[|x_i| \mid \operatorname{sign}(x_i) = \operatorname{sign}(w_i)\right] = 6/\mathop{\mathbb{E}}_x\left[|x|\right] \ ,$$

we have that

$$\max_{x\in\{\pm 1\}^N} \mathbb{E}_\epsilon \max_{i\in[N]} |w_i + 2\epsilon x_i| \leq \mathbb{E}_{x,\epsilon} \max_i |w_i + C\epsilon x_i| \ .$$

This concludes the proof. □

Lemma 25. *Consider the case when $\mathcal{X}$ is the ℓ_∞^N ball and $\mathcal{F}$ is the ℓ_1^N unit ball. Let $f^* = \operatorname{argmin}_{f\in\mathcal{F}} \langle f, R\rangle$, then for any random vector R,*

$$\mathop{\mathbb{E}}_{R}\left[\sup_{x\in\mathcal{X}}\left\{\langle f^*, x\rangle + \|R+x\|_\infty\right\}\right] \le \mathop{\mathbb{E}}_{R}\left[\inf_{f\in\mathcal{F}}\sup_{x}\left\{\langle f, x\rangle + \|R+x\|_\infty\right\}\right] + 4\ \mathbf{P}\left(\|R\|_\infty \le 4\right)$$

Proof. Let $f^* = \operatorname{argmin}_{f\in\mathcal{F}} \langle f, R\rangle$. We start by noting that for any $f' \in \mathcal{F}$,

$$\begin{aligned}
\sup_{x\in\mathcal{X}}\left\{\langle f', x\rangle + \|R+x\|_\infty\right\} &= \sup_{x\in\mathcal{X}}\left\{\langle f', x\rangle + \sup_{f\in\mathcal{F}}\langle f, R+x\rangle\right\} \\
&= \sup_{f\in\mathcal{F}}\sup_{x\in\mathcal{X}}\left\{\langle f', x\rangle + \langle f, R+x\rangle\right\} \\
&= \sup_{f\in\mathcal{F}}\left\{\sup_{x\in\mathcal{X}}\langle f'+f, x\rangle + \langle f, R\rangle\right\} \\
&= \sup_{f\in\mathcal{F}}\left\{\|f'+f\|_1 + \langle f, R\rangle\right\}
\end{aligned}$$

Hence note that

$$\begin{aligned}
\inf_{f'\in\mathcal{F}}\sup_{x\in\mathcal{X}}\left\{\langle f', x\rangle + \|R+x\|_\infty\right\} &= \inf_{f'\in\mathcal{F}}\sup_{f\in\mathcal{F}}\left\{\|f'+f\|_1 + \langle f, R\rangle\right\} && (55)\\
&\ge \inf_{f'\in\mathcal{F}}\left\{\|f'-f^*\|_1 - \langle f^*, R\rangle\right\} \ge \inf_{f'\in\mathcal{F}}\left\{\|f'-f^*\|_1 + \|R\|_\infty\right\} = \|R\|_\infty && (56)
\end{aligned}$$

On the other hand note that, f^* is the vertex of the ℓ_1 ball (any one which given by $\operatorname{argmin}_{i\in[d]} |R[i]|$ with sign opposite as sign of $R[i]$ on that vertex). Since the ℓ_1 ball is the convex hull of the $2d$ vertices, any vector $f \in \mathcal{F}$ can be written as $f = \alpha h - \beta f^*$ some $h \in \mathcal{F}$ such that $\|h\|_1 = 1$ and $\langle h, R\rangle = 0$ (which means that h is 0 on the maximal co-ordinate of R specified by f^*) and for some $\beta \in [-1,1]$, $\alpha \in [0,1]$ s.t. $\|\alpha h - \beta f^*\|_1 \le 1$. Further note that the constraint on α, β imposed by requiring that $\|\alpha h - \beta f^*\|_1 \le 1$ can be written as $\alpha + |\beta| \le 1$. Hence,

$$\begin{aligned}
\sup_{x\in\mathcal{X}}\left\{\langle f^*, x\rangle + \|R+x\|_\infty\right\} &= \sup_{f\in\mathcal{F}}\left\{\|f^*+f\|_1 + \langle f, R\rangle\right\} \\
&= \sup_{\alpha\in[0,1]}\ \sup_{h\perp f^*, \|h\|_1=1}\ \sup_{\beta\in[-1,1], \|\alpha h-\beta f^*\|_1\le 1}\left\{\|(1-\beta)f^* + \alpha h\|_1 + \beta\langle f^*, R\rangle + \alpha\langle h, R\rangle\right\} \\
&= \sup_{\alpha\in[0,1]}\ \sup_{h\perp f^*, \|h\|_1=1}\ \sup_{\beta\in[-1,1], \|\alpha h-\beta f^*\|_1\le 1}\left\{|1-\beta|\,\|f^*\|_1 + \alpha\|h\|_1 + \beta\|R\|_\infty\right\} \\
&= \sup_{\alpha\in[0,1]}\ \sup_{\beta\in[-1,1]:|\beta|+\alpha\le 1}\left\{|1-\beta| + \alpha + \beta\|R\|_\infty\right\} \\
&\le \sup_{\beta\in[-1,1]}\left\{|1-\beta| + 1 - |\beta| + \beta\|R\|_\infty\right\} \\
&\le \sup_{\beta\in[-1,1]}\left\{2|1-\beta| + \beta\|R\|_\infty\right\} \\
&= \sup_{\beta\in\{-1,1\}}\left\{2|1-\beta| + \beta\|R\|_\infty\right\} \\
&= \max\left\{\|R\|_\infty, 4 - \|R\|_\infty\right\} \\
&\le \|R\|_\infty + 4\ \mathbf{1}\left\{\|R\|_\infty \le 4\right\}
\end{aligned}$$

Hence combining with equation 55 we can conclude that

$$\begin{aligned}\mathop{\mathbb{E}}_{R}\left[\sup_{x}\left\{\langle f^*, x\rangle + \|R+x\|_\infty\right\}\right] &\le \mathop{\mathbb{E}}_{R}\left[\inf_{f\in\mathcal{F}}\sup_{x}\left\{\langle f, x\rangle + \|R+x\|_\infty\right\}\right] + 4\mathop{\mathbb{E}}_{R}\left[\mathbf{1}\left\{\|R\|_\infty \le 4\right\}\right] \\ &= \mathop{\mathbb{E}}_{R}\left[\inf_{f\in\mathcal{F}}\sup_{x}\left\{\langle f, x\rangle + \|R+x\|_\infty\right\}\right] + 4\,\mathbf{P}\left(\|R\|_\infty \le 4\right)\end{aligned}$$

□

Proof of Lemma 14. On any round t, the algorithm draws $\epsilon_{t+1}, \dots, \epsilon_T$ and $x_{t+1}, \dots, x_T \sim D^N$ and plays

$$f_t = \operatorname*{argmin}_{f\in\mathcal{F}} \left\langle f, \sum_{i=1}^{t-1} x_i - C\sum_{i=t+1}^{T} x_i \right\rangle$$

We shall show that this randomized algorithm is (almost) admissible w.r.t. the relaxation (with some small additional term at each step). We define the relaxation as

$$\mathbf{Rel}_T\left(\mathcal{F}|x_1,\dots,x_t\right) = \mathop{\mathbb{E}}_{x_{t+1},\dots x_T\sim D}\left[\left\|\sum_{i=1}^{t} x_i - C\sum_{i=t+1}^{T} x_i\right\|_\infty\right]$$

Proceeding just as in the proof of Lemma 11 note that, for our randomized strategy,

$$\begin{aligned}&\sup_{x}\left\{\mathop{\mathbb{E}}_{f\sim q_t}\left[\langle f, x\rangle\right] + \mathbf{Rel}_T\left(\mathcal{F}|x_1,\dots,x_t\right)\right\} \\ &\quad= \sup_{x}\left\{\mathop{\mathbb{E}}_{x_{t+1:T}\sim D^N}\left[\langle f_t, x\rangle\right] + \mathop{\mathbb{E}}_{x_{t+1:T}\sim D^N}\left[\left\|\sum_{i=1}^{t-1} x_i + x - C\sum_{i=t+1}^{T} x_i\right\|_\infty\right]\right\} \\ &\quad\le \mathop{\mathbb{E}}_{x_{t+1:T}\sim D^N}\left[\sup_{x}\left\{\langle f_t, x\rangle + \left\|\sum_{i=1}^{t-1} x_i + x - C\sum_{i=t+1}^{T} x_i\right\|_\infty\right\}\right] \qquad (57)\end{aligned}$$

In view of Lemma 25 (with $R = \sum_{i=1}^{t-1} x_i - C\sum_{i=t+1}^{T}\epsilon_i x_i$) we conclude that

$$\begin{aligned}&\mathop{\mathbb{E}}_{x_{t+1},\dots,x_T}\left[\sup_{x\in\mathcal{X}}\left\{\langle f_t, x\rangle + \left\|\sum_{i=1}^{t-1} x_i - C\sum_{i=t+1}^{T} x_i + x\right\|_\infty\right\}\right] \\ &\quad\le \mathop{\mathbb{E}}_{x_{t+1},\dots,x_T}\left[\inf_{f\in\mathcal{F}}\sup_{x}\left\{\langle f, x\rangle + \left\|\sum_{i=1}^{t-1} x_i - C\sum_{i=t+1}^{T} x_i + x\right\|_\infty\right\}\right] + 4\,\mathbf{P}\left(\left\|\sum_{i=1}^{t-1} x_i - C\sum_{i=t+1}^{T} x_i\right\|_\infty \le 4\right) \\ &\quad= \mathop{\mathbb{E}}_{x_{t+1},\dots,x_T}\left[\sup_{x}\left\{\langle f_t^*, x\rangle + \left\|\sum_{i=1}^{t-1} x_i - C\sum_{i=t+1}^{T} x_i + x\right\|_\infty\right\}\right] + 4\,\mathbf{P}\left(\left\|\sum_{i=1}^{t-1} x_i - C\sum_{i=t+1}^{T} x_i\right\|_\infty \le 4\right)\end{aligned}$$

where

$$f_t^* = \operatorname*{argmin}_{f\in\mathcal{F}}\sup_{x}\left\{\langle f, x\rangle + \left\|\sum_{i=1}^{t-1} x_i - C\sum_{i=t+1}^{T} x_i + x\right\|_\infty\right\}$$

Combining with Equation (57) we conclude that

$$\begin{aligned}&\sup_{x}\left\{\mathop{\mathbb{E}}_{f\sim q_t}\left[\langle f, x\rangle\right] + \mathbf{Rel}_T\left(\mathcal{F}|x_1,\dots,x_t\right)\right\} \\ &\quad\le \mathop{\mathbb{E}}_{x_{t+1},\dots,x_T}\left[\sup_{x}\left\{\langle f_t^*, x\rangle + \left\|\sum_{i=1}^{t-1} x_i - C\sum_{i=t+1}^{T} x_i + x\right\|_\infty\right\}\right] + 4\,\mathbf{P}\left(\left\|\sum_{i=1}^{t-1} x_i - C\sum_{i=t+1}^{T} x_i\right\|_\infty \le 4\right)\end{aligned}$$

Now, since

$$4\ \mathbf{P}\left(\left\|\sum_{i=1}^{t-1} x_i - C\sum_{i=t+1}^{T} x_i\right\|_\infty \le 4\right) \le 4\ \mathbf{P}\left(C\left\|\sum_{i=t+1}^{T} x_i\right\|_\infty \le 4\right) \le 4\ \mathbf{P}_{y_{t+1},\dots,y_T\sim D}\left(C\left|\sum_{i=t+1}^{T} y_i\right| \le 4\right)$$

we have

$$\sup_x \left\{ \mathop{\mathbb{E}}_{f\sim q_t}\left[\langle f, x\rangle\right] + \mathbf{Rel}_T\left(\mathcal{F}|x_1,\dots,x_t\right)\right\} \tag{58}$$

$$\le \mathop{\mathbb{E}}_{x_{t+1},\dots,x_T}\left[\sup_x\left\{\langle f_t^*, x\rangle + \left\|\sum_{i=1}^{t-1} x_i - C\sum_{i=t+1}^{T} x_i + x\right\|_\infty\right\}\right] + 4\ \mathbf{P}_{y_{t+1},\dots,y_T\sim D}\left(C\left|\sum_{i=t+1}^{T} y_i\right| \le 4\right) \tag{59}$$

In view of Lemma 13, Assumption 2 is satisfied by D^N with constant C. Further in the proof of Lemma 11 we already showed that whenever Assumption 2 is satisfied, the randomized strategy specified by f_t^* is admissible. More specifically we showed that

$$\mathop{\mathbb{E}}_{x_{t+1},\dots,x_T}\left[\sup_x\left\{\langle f_t^*, x\rangle + \left\|\sum_{i=1}^{t-1} x_i - C\sum_{i=t+1}^{T} x_i + x\right\|_\infty\right\}\right] \le \mathbf{Rel}_T\left(F|x_1,\dots,x_{t-1}\right)$$

and so using this in Equation (58) we conclude that for the randomized strategy in the statement of the lemma,

$$\sup_x \left\{ \mathop{\mathbb{E}}_{f\sim q_t}\left[\langle f, x\rangle\right] + \mathbf{Rel}_T\left(\mathcal{F}|x_1,\dots,x_t\right)\right\}$$
$$\le \mathbf{Rel}_T\left(F|x_1,\dots,x_{t-1}\right) + 4\ \mathbf{P}_{y_{t+1},\dots,y_T\sim D}\left(C\left|\sum_{i=t+1}^{T} y_i\right| \le 4\right)$$

Or in other words the randomized strategy proposed is admissible with an additional additive factor of $4\ \mathbf{P}_{y_{t+1},\dots,y_T\sim D}\left(C\left|\sum_{i=t+1}^{T} y_i\right| \le 4\right)$ at each time step t. Hence by Proposition 1 we have that for the randomized algorithm specified in the lemma,

$$\mathbb{E}\left[\mathbf{Reg}_T\right] \le \mathbf{Rel}_T\left(F\right) + 4\sum_{t=1}^{T}\mathbf{P}_{y_{t+1},\dots,y_T\sim D}\left(C\left|\sum_{i=t+1}^{T} y_i\right| \le 4\right)$$
$$= C\mathop{\mathbb{E}}_{x_1,\dots,x_T\sim D^N}\left[\left\|\sum_{t=1}^{T} x_t\right\|_\infty\right] + 4\sum_{t=1}^{T}\mathbf{P}_{y_{t+1},\dots,y_T\sim D}\left(C\left|\sum_{i=t+1}^{T} y_i\right| \le 4\right)$$

This concludes the proof. □

Proof of Lemma 15. Instead of using $C = 4\sqrt{2}$ and drawing uniformly from surface of unit sphere we can equivalently think of the constant as being 1 and drawing uniformly from surface of sphere of radius $4\sqrt{2}$. Let $\|\cdot\|$ stand for the Euclidean norm. To prove (27), first observe that

$$\sup_{p\in\Delta(\mathcal{X})}\mathop{\mathbb{E}}_{x_t\sim p}\left\|w + \mathop{\mathbb{E}}_{x\sim p}\left[x\right] - x_t\right\| \le \sup_{x\in\mathcal{X}}\mathop{\mathbb{E}}_{\epsilon}\left\|w + 2\epsilon x\right\| \tag{60}$$

for any $w \in B$. Further, using Jensen's inequality

$$\sup_{x\in\mathcal{X}}\mathop{\mathbb{E}}_{\epsilon}\left\|w + 2\epsilon x\right\| \le \sup_{x\in\mathcal{X}}\sqrt{\mathop{\mathbb{E}}_{\epsilon}\left\|w + 2\epsilon x\right\|^2} \le \sup_{x\in\mathcal{X}}\sqrt{\left\|w\right\|^2 + \mathop{\mathbb{E}}_{\epsilon}\left\|2\epsilon x\right\|^2} = \sqrt{\left\|w\right\|^2 + 4}$$

To prove the lemma, it is then enough to show that for $r = 4\sqrt{2}$

$$\mathbb{E}_{x\sim D}\left\|w + rx\right\| \ge \sqrt{\left\|w\right\|^2 + 4} \tag{61}$$

for any w, where we omitted ϵ since D is symmetric. This fact can be proved with the following geometric argument.

We define quadruplets $(w + z_1, w + z_2, w - z_1, w - z_2)$ of points on the sphere of radius r. Each quadruplets will have the property that

$$\frac{\|w + z_1\| + \|w + z_2\| + \|w - z_1\| + \|w - z_2\|}{4} \geq \sqrt{\|w\|^2 + 4} \tag{62}$$

for any w. We then argue that the uniform distribution can be decomposed into these quadruplets such that each point on the sphere occurs in only one quadruplet (except for a measure zero set when z_1 is aligned with $-w$), thus concluding that (61) holds true.

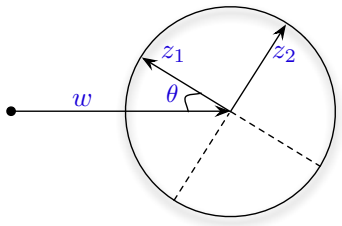

Figure 1: The two-dimensional construction for the proof of Lemma 15.

Pick any direction $w^\perp$ perpendicular to w. A quadruplet is defined by perpendicular vectors z_1 and z_2 which have length r and which lie in the plane spanned by $w, w^\perp$. Let θ be the angle between $-w$ and z_1. Since we are now dealing with a two dimensional plane spanned by w and $w^\perp$, we may as well assume that w is aligned with the positive x-axis, as in Figure 1. We write w for $\|w\|$. The coordinates of the quadruplet are

$$(w - r\cos(\theta), r\sin(\theta)), \quad (w + r\cos(\theta), -r\sin(\theta)), \quad (w + r\sin(\theta), r\cos(\theta)), \quad (w - r\sin(\theta), -r\cos(\theta))$$

For brevity, let $s = \sin(\theta), c = \cos(\theta)$. The desired inequality (62) then reads

$$\sqrt{w^2 - 8wc + r^2} + \sqrt{w^2 + 8wc + r^2} + \sqrt{w^2 + 8ws + r^2} + \sqrt{w^2 - 8ws + r^2} \geq 4\sqrt{w^2 + 4}$$

To prove that this inequality holds, we square both sides, keeping in mind that the terms are non-negative. The sum of four squares on the left hand side gives $4w^2 + 4r^2$. For the six cross terms, we can pass to a lower bound by replacing r^2 in each square root by r^2c^2 or r^2s^2, whichever completes the square. Then observe that

$$|w + rs| \cdot |w - rs| + |w + rc| \cdot |w - rc| = 2w^2 - r^2$$

while the other four cross terms

$$(|w + rs| \cdot |w - rc| + |w + rs| \cdot |w + rc|) + (|w - rs| \cdot |w + rc| + |w - rs| \cdot |w - rc|) \geq |w + rs| \cdot 2w + |w - rs| \cdot 2w \geq 4w^2$$

Doubling the cross terms gives a contribution of $2(6w^2 - r^2)$, while the sum of squares yielded $4w^2 + 4r^2$. The desired inequality is satisfied as long as $16w^2 + 2r^2 \geq 16(w^2 + 4)$, or $r \geq 4\sqrt{2}$.

□

Proof of Lemma 16. By Lemma 15, Assumption 2 is satisfied by distribution D with constant $C = 4\sqrt{2}$. Hence by Lemma 13 we can conclude that for the randomized algorithm which at round t freshly draws $x_{t+1}, \ldots, x_T \sim D$ and picks

$$f_t^* = \operatorname*{argmin}_{f \in \mathcal{F}} \sup_{x \in \mathcal{X}} \left\{ \langle f, x \rangle + \left\| -\sum_{i=1}^{t-1} x_i + 4\sqrt{2} \sum_{i=t+1}^{T} x_i - x \right\|_2 \right\}$$

(we dropped the ϵ's as the distribution is symmetric to start with) the expected regret is bounded as

$$\mathbb{E}\left[\mathbf{Reg}_T\right] \le 4\sqrt{2} \mathop{\mathbb{E}}_{x_1,\ldots,x_T\sim D}\left[\left\|\sum_{t=1}^{T} x_t\right\|_2\right] \le 4\sqrt{2T}$$

We claim that the strategy specified in the lemma that chooses

$$f_t = \frac{-\sum_{i=1}^{t-1} x_i + 4\sqrt{2}\sum_{i=t+1}^{T} x_i}{\sqrt{\left\|-\sum_{i=1}^{t-1} x_i + 4\sqrt{2}\sum_{i=t+1}^{T}\epsilon_i x_i\right\|_2^2 + 1}}$$

is the same as choosing f_t^*. To see this let us start by defining

$$\bar{x}_t = -\sum_{i=1}^{t-1} x_i + 4\sqrt{2}\sum_{i=t+1}^{T} x_i$$

Now note that

$$\begin{aligned}
f_t^* = \mathop{\mathrm{argmin}}_{f\in\mathcal{F}} \sup_{x\in\mathcal{X}}\left\{\langle f,x\rangle + \left\|-\sum_{i=1}^{t-1} x_i + 4\sqrt{2}\sum_{i=t+1}^{T} x_i - x\right\|_2\right\} &= \mathop{\mathrm{argmin}}_{f\in\mathcal{F}} \sup_{x\in\mathcal{X}}\left\{\langle f,x\rangle + \|\bar{x}_t - x\|_2\right\} \\
&= \mathop{\mathrm{argmin}}_{f\in\mathcal{F}} \sup_{x\in\mathcal{X}}\left\{\langle f,x\rangle + \sqrt{\|\bar{x}_t - x\|_2^2}\right\} \\
&= \mathop{\mathrm{argmin}}_{f\in\mathcal{F}} \sup_{x:\|x\|_2\le 1}\left\{\langle f,x\rangle + \sqrt{\|\bar{x}_t\|^2 - 2\langle \bar{x}_t, x\rangle + \|x\|_2^2}\right\} \\
&= \mathop{\mathrm{argmin}}_{f\in\mathcal{F}} \sup_{x:\|x\|_2=1}\left\{\langle f,x\rangle + \sqrt{\|\bar{x}_t\|^2 - 2\langle \bar{x}_t, x\rangle + 1}\right\}
\end{aligned}$$

However this argmin calculation is identical to the one in the proof of Proposition 4 (with $C = 1$ and $T - t = 0$) and the solution is given by

$$f_t^* = f_t = \frac{-\sum_{i=1}^{t-1} x_i + 4\sqrt{2}\sum_{i=t+1}^{T} x_i}{\sqrt{\left\|-\sum_{i=1}^{t-1} x_i + 4\sqrt{2}\sum_{i=t+1}^{T}\epsilon_i x_i\right\|_2^2 + 1}}$$

Thus we conclude the proof. □

Proof of Lemma 17. We first prove the statement for the convex case. To show admissibility using the particular randomized strategy given in the lemma, we need to show that for the randomized strategy specified by q_t,

$$\sup_{y_t}\left\{\mathop{\mathbb{E}}_{\hat{y}_t\sim q_t}\left[\ell(\hat{y}_t, y_t)\right] + \mathbf{Rel}_T\left(\mathcal{F}|(x_1,y_1),\ldots,(x_t,y_t)\right)\right\} \le \mathbf{Rel}_T\left(\mathcal{F}|(x_1,y_1),\ldots,(x_{t-1},y_{t-1})\right)$$

for any x_t. The strategy q_t proposed by the lemma is such that we first draw $(x_{t+1},y_{t+1}),\ldots,(x_T,y_T) \sim D$ and $\epsilon_{t+1},\ldots\epsilon_T$ Rademacher random variables, and then based on this sample pick $\hat{y}_t = \hat{y}_t(x_{t+1:T}, y_{t+1:T}, \epsilon_{t+1:T})$ as in (30). Hence,

$$\begin{aligned}
&\sup_{y_t}\left\{\mathop{\mathbb{E}}_{\hat{y}_t\sim q_t}\left[\ell(\hat{y}_t, y_t)\right] + \mathbf{Rel}_T\left(\mathcal{F}|(x_1,y_1),\ldots,(x_t,y_t)\right)\right\} \\
&\quad= \sup_{y_t}\left\{\mathop{\mathbb{E}}_{\substack{\epsilon_{t+1:T}\\(x_{t+1:T},y_{t+1:T})}} \ell(\hat{y}_t, y_t) + \mathop{\mathbb{E}}_{\substack{\epsilon_{t+1:T}\\(x_{t+1:T},y_{t+1:T})}} \sup_{f\in\mathcal{F}}\left[C\sum_{i=t+1}^{T}\epsilon_i\ell(f(x_i),y_i) - L_t(f)\right]\right\} \\
&\quad\le \mathop{\mathbb{E}}_{\substack{\epsilon_{t+1:T}\\(x_{t+1:T},y_{t+1:T})}} \sup_{y_t}\left\{\ell(\hat{y}_t, y_t) + \sup_{f\in\mathcal{F}}\left[C\sum_{i=t+1}^{T}\epsilon_i\ell(f(x_i),y_i) - L_t(f)\right]\right\} .
\end{aligned}$$

Now, with $\hat{y}_t$ in (30),

$$\begin{aligned}
&\sup_{y_t}\left\{\ell(\hat{y}_t,y_t)+\sup_{f\in\mathcal{F}}\left[C\sum_{i=t+1}^{T}\epsilon_i\ell(f(x_i),y_i)-L_t(f)\right]\right\}\\
&\quad=\inf_{\hat{y}\in[-B,B]}\sup_{y_t}\left\{\ell(\hat{y},y_t)+\sup_{f\in\mathcal{F}}\left[C\sum_{i=t+1}^{T}\epsilon_i\ell(f(x_i),y_i)-L_t(f)\right]\right\}\\
&\quad=\inf_{\hat{y}\in[-B,B]}\sup_{p_t}\mathbb{E}_{y_t\sim p_t}\left[\ell(\hat{y},y_t)+\sup_{f\in\mathcal{F}}\left[C\sum_{i=t+1}^{T}\epsilon_i\ell(f(x_i),y_i)-L_t(f)\right]\right]
\end{aligned}$$

Now we assume that the loss $\ell(\hat{y},y)$ is convex in the first argument (and bounded). Note that the term

$$\mathbb{E}_{y_t\sim p_t}\left[\ell(\hat{y},y_t)+\sup_{f\in\mathcal{F}}\left\{C\sum_{i=t+1}^{T}\epsilon_i\ell(f(x_i),y_i)-L_t(f)\right\}\right]$$

is linear in p_t and, due to convexity of loss, is convex in $\hat{y}_t$. Hence by the minimax theorem, for this choice of q_t, we conclude that

$$\begin{aligned}
&\sup_{y_t}\left\{\mathbb{E}_{\hat{y}_t\sim q_t}\left[\ell(\hat{y}_t,y_t)\right]+\mathbf{Rel}_T\left(\mathcal{F}|(x_1,y_1),\ldots,(x_t,y_t)\right)\right\}\\
&\quad\le\mathbb{E}_{\substack{\epsilon_{t+1:T}\\(x_{t+1},y_{t+1}),\ldots,(x_T,y_T)}}\inf_{\hat{y}_t\in[-B,B]}\sup_{p_t}\mathbb{E}_{y_t\sim p_t}\left[\ell(\hat{y}_t,y_t)+\sup_{f\in\mathcal{F}}\left\{C\sum_{i=t+1}^{T}\epsilon_i\ell(f(x_i),y_i)-L_t(f)\right\}\right]\\
&\quad=\mathbb{E}_{\substack{\epsilon_{t+1:T}\\(x_{t+1},y_{t+1}),\ldots,(x_T,y_T)}}\sup_{p_t}\inf_{\hat{y}_t\in[-B,B]}\mathbb{E}_{y_t\sim p_t}\left[\ell(\hat{y}_t,y_t)+\sup_{f\in\mathcal{F}}\left\{C\sum_{i=t+1}^{T}\epsilon_i\ell(f(x_i),y_i)-L_t(f)\right\}\right]
\end{aligned}$$

The last step above is due to the minimax theorem as the loss is convex in $\hat{y}_t$, the set $[-B,B]$ is compact, and the term is linear in p_t. The above expression is equal to

$$\begin{aligned}
&=\mathbb{E}\sup_{p_t}\mathbb{E}_{y_t\sim p_t}\sup_{f\in\mathcal{F}}\left\{C\sum_{i=t+1}^{T}\epsilon_i\ell(f(x_i),y_i)-\sum_{i=1}^{t-1}\ell(f(x_i),y_i)+\inf_{\hat{y}_t\in[-B,B]}\mathbb{E}_{y_t\sim p_t}\left[\ell(\hat{y}_t,y_t)\right]-\ell(f(x_t),y_t)\right\}\\
&\le\mathbb{E}\sup_{p_t}\mathbb{E}_{y_t\sim p_t}\sup_{f\in\mathcal{F}}\left\{C\sum_{i=t+1}^{T}\epsilon_i\ell(f(x_i),y_i)-L_{t-1}(f)+\inf_{g\in\mathcal{F}}\mathbb{E}_{y_t\sim p_t}\left[\ell(g(x_t),y_t)\right]-\ell(f(x_t),y_t)\right\}\\
&\le\mathbb{E}\sup_{p_t}\mathbb{E}_{y_t\sim p_t}\sup_{f\in\mathcal{F}}\left\{C\sum_{i=t+1}^{T}\epsilon_i\ell(f(x_i),y_i)-L_{t-1}(f)+\mathbb{E}_{y_t\sim p_t}\left[\ell(f(x_t),y_t)\right]-\ell(f(x_t),y_t)\right\}\\
&\le\mathbb{E}_{\substack{\epsilon_{t+1:T}\\(x_{t+1},y_{t+1}),\ldots,(x_T,y_T)}}\mathbb{E}_{(x_t,y_t)\sim D}\mathbb{E}_{\epsilon_t}\sup_{f\in\mathcal{F}}\left\{C\sum_{i=t+1}^{T}\epsilon_i\ell(f(x_i),y_i)-L_{t-1}(f)+C\epsilon_t\ell(f(x_t),y_t)\right\}\\
&=\mathbf{Rel}_T\left(\mathcal{F}|(x_1,y_1),\ldots,(x_{t-1},y_{t-1})\right)
\end{aligned}$$

The second part of the Lemma is proved analogously.

□

Proof of Lemma 18. Now let q_t be the randomized strategy where we draw $\epsilon_{t+1},\ldots,\epsilon_T$ uniformly at random and pick

$$q_t(\epsilon)=\operatorname*{argmin}_{q\in\Delta(\mathcal{F})}\sup_{x_t}\left\{\mathbb{E}_{f_t\in q}\left[\ell(f_t,x_t)\right]+\sup_{f\in\mathcal{F}}\left[2\sum_{i=t+1}^{T}\epsilon_i\ell(f,\mathbf{x}_i^t(\epsilon))-\sum_{i=1}^{t}\ell(f,x_i)\right]\right\}\tag{63}$$

With the definition of $\mathbf{x}^t$ in (32), and with the notation $L_t(f) = \sum_{i=1}^t \ell(f, x_i)$

$$\begin{aligned}
&\sup_{x_t}\left\{\mathop{\mathbb{E}}_{f_t\sim q_t}[\ell(f_t,x_t)] + \sup_{\mathbf{x}}\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}}\left[2\sum_{i=t+1}^T \epsilon_i\ell(f,\mathbf{x}_i(\epsilon)) - L_t(f)\right]\right\}\\
&\quad= \sup_{x_t}\left\{\mathop{\mathbb{E}}_{\epsilon}\left[\mathop{\mathbb{E}}_{f_t\sim q_t(\epsilon)}[\ell(f_t,x_t)]\right] + \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}}\left[2\sum_{i=t+1}^T \epsilon_i\ell(f,\mathbf{x}_i^t(\epsilon)) - L_t(f)\right]\right\}\\
&\quad\le \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{x_t}\left\{\mathop{\mathbb{E}}_{f_t\sim q_t(\epsilon)}[\ell(f_t,x_t)] + \sup_{f\in\mathcal{F}}\left[2\sum_{i=t+1}^T \epsilon_i\ell(f,\mathbf{x}_i^t(\epsilon)) - L_t(f)\right]\right\}\right]\\
&\quad= \mathop{\mathbb{E}}_{\epsilon}\left[\inf_{q_t\in\Delta(\mathcal{F})}\sup_{x_t}\left\{\mathop{\mathbb{E}}_{f_t\sim q_t}[\ell(f_t,x_t)] + \sup_{f\in\mathcal{F}}\left[2\sum_{i=t+1}^T \epsilon_i\ell(f,\mathbf{x}_i^t(\epsilon)) - L_t(f)\right]\right\}\right]
\end{aligned}$$

where the last step is due to the way we pick our predictor $f_t(\epsilon)$ given random draw of ϵ's in Equation (63). We now apply the minimax theorem, yielding the following upper bound on the term above:

$$\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{p_t\in\Delta(\mathcal{X})}\inf_{f_t\in\mathcal{F}}\left\{\mathop{\mathbb{E}}_{x_t\sim p_t}[\ell(f_t,x_t)] + \mathbb{E}_{x_t\sim p_t}\sup_{f\in\mathcal{F}}\left[2\sum_{i=t+1}^T \epsilon_i\ell(f,\mathbf{x}_i^t(\epsilon)) - L_t(f)\right]\right\}\right]$$

This expression can be re-written as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{p_t\in\Delta(\mathcal{X})}\left\{\mathbb{E}_{x_t\sim p_t}\sup_{f\in\mathcal{F}}\left[2\sum_{i=t+1}^T \epsilon_i\ell(f,\mathbf{x}_i^t(\epsilon)) - L_{t-1}(f) + \mathop{\mathbb{E}}_{x_t\sim p_t}[\ell(f,x_t)] - \ell(f,x_t)\right]\right\}\right]\\
&\le \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{p_t\in\Delta(\mathcal{X})}\left\{\mathbb{E}_{x_t,x_t'\sim p_t}\mathbb{E}_{\epsilon_t}\sup_{f\in\mathcal{F}}\left[2\sum_{i=t+1}^T \epsilon_i\ell(f,\mathbf{x}_i^t(\epsilon)) - L_{t-1}(f) + \epsilon_t\left(\ell(f,x_t) - \ell(f,x_t)\right)\right]\right\}\right]
\end{aligned}$$

By passing to the supremum over x_t, x_t', we get an upper bound

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{x_t,x_t'\in\mathcal{X}}\left\{\mathbb{E}_{\epsilon_t}\sup_{f\in\mathcal{F}}\left[2\sum_{i=t+1}^T \epsilon_i\ell(f,\mathbf{x}_i^t(\epsilon)) - L_{t-1}(f) + \epsilon_t\left(\ell(f,x_t) - \ell(f,x_t)\right)\right]\right\}\right]\\
&\le \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{x_t\in\mathcal{X}}\left\{\mathbb{E}_{\epsilon_t}\sup_{f\in\mathcal{F}}\left[2\sum_{i=t+1}^T \epsilon_i\ell(f,\mathbf{x}_i^t(\epsilon)) - L_{t-1}(f) + 2\epsilon_t\ell(f,x_t)\right]\right\}\right]\\
&\le \sup_{\tilde{\mathbf{x}}}\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{x_t\in\mathcal{X}}\left\{\mathbb{E}_{\epsilon_t}\sup_{f\in\mathcal{F}}\left[2\sum_{i=t+1}^T \epsilon_i\ell(f,\tilde{\mathbf{x}}_i(\epsilon)) - L_{t-1}(f) + 2\epsilon_t\ell(f,x_t)\right]\right\}\right]\\
&\le \sup_{\mathbf{x}}\mathbb{E}_\epsilon\sup_{f\in\mathcal{F}}\left[2\sum_{i=t}^T \epsilon_i\ell(f,\mathbf{x}_i(\epsilon)) - L_{t-1}(f)\right]
\end{aligned}$$

□

Proof of Lemma 19. We shall start by showing that the relaxation is admissible for the game where we pick prediction $\hat{y}_t$ and the adversary then directly picks the gradient $\partial\ell(\hat{y}_t, y_t)$. To this end note that

$$\begin{aligned}
&\inf_{\hat{y}_t}\sup_{\partial\ell(\hat{y}_t,y_t)}\left\{\partial\ell(\hat{y}_t,y_t)\cdot\hat{y}_t + \mathbf{Rel}_T\left(\mathcal{F}|\partial\ell(\hat{y}_1,y_1),\ldots,\partial\ell(\hat{y}_t,y_t)\right)\right\}\\
&\qquad= \inf_{\hat{y}_t}\sup_{\partial\ell(\hat{y}_t,y_t)}\left\{\partial\ell(\hat{y}_t,y_t)\cdot\hat{y}_t + \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in\mathcal{F}} 2L\sum_{i=t+1}^T \epsilon_i f[t] - \sum_{i=1}^t \partial\ell(\hat{y}_i,y_i)\cdot f[i]\right]\right\}\\
&\qquad\le \inf_{\hat{y}_t}\sup_{r_t\in[-L,L]}\left\{r_t\cdot\hat{y}_t + \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in\mathcal{F}} 2L\sum_{i=t+1}^T \epsilon_i f[t] - L_{t-1}(f) - r_t\cdot f[t]\right]\right\}
\end{aligned}$$

Let us use the notation $L_{t-1}(f) = \sum_{i=1}^{t-1} \partial\ell(\hat{y}_i, y_i) \cdot f[i]$ for the present proof. The supremum over $r_t \in [-L, L]$ is achieved at the endpoints since the expression is convex in r_t. Therefore, the last expression is equal to

$$\begin{aligned}
&\inf_{\hat{y}_t} \sup_{r_t \in \{-L,L\}} \left\{ r_t \cdot \hat{y}_t + \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left[2L \sum_{i=t+1}^{T} \epsilon_i f[t] - L_{t-1}(f) - r_t \cdot f[t] \right] \right\} \\
&= \inf_{\hat{y}_t} \sup_{p_t \in \Delta(\{-L,L\})} \mathop{\mathbb{E}}_{r_t \sim p_t} \left[r_t \cdot \hat{y}_t + \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left[2L \sum_{i=t+1}^{T} \epsilon_i f[t] - L_{t-1}(f) - r_t \cdot f[t] \right] \right] \\
&= \sup_{p_t \in \Delta(\{-L,L\})} \inf_{\hat{y}_t} \mathop{\mathbb{E}}_{r_t \sim p_t} \left[r_t \cdot \hat{y}_t + \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left[2L \sum_{i=t+1}^{T} \epsilon_i f[t] - L_{t-1}(f) - r_t \cdot f[t] \right] \right]
\end{aligned}$$

where the last step is due to the minimax theorem. The last quantity is equal to

$$\begin{aligned}
&\sup_{p_t \in \Delta(\{-L,L\})} \mathop{\mathbb{E}}_{\epsilon} \left[\mathop{\mathbb{E}}_{r_t \sim p_t} \left[\inf_{\hat{y}_t} \mathop{\mathbb{E}}_{r_t \sim p_t} [r_t] \cdot \hat{y}_t + \sup_{f \in \mathcal{F}} \left(2L \sum_{i=t+1}^{T} \epsilon_i f[t] - L_{t-1}(f) - r_t \cdot f[t] \right) \right] \right] \\
&\leq \sup_{p_t \in \Delta(\{-L,L\})} \mathop{\mathbb{E}}_{\epsilon} \left[\mathop{\mathbb{E}}_{r_t \sim p_t} \left[\sup_{f \in \mathcal{F}} \left(2L \sum_{i=t+1}^{T} \epsilon_i f[t] - L_{t-1}(f) + (\mathop{\mathbb{E}}_{r_t \sim p_t} [r_t] - r_t) \cdot f[t] \right) \right] \right] \\
&\leq \sup_{p_t \in \Delta(\{-L,L\})} \mathop{\mathbb{E}}_{r_t, r_t' \sim p_t} \left[\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left[2L \sum_{i=t+1}^{T} \epsilon_i f[t] - L_{t-1}(f) + (r_t' - r_t) \cdot f[t] \right] \right] \\
&= \sup_{p_t \in \Delta(\{-L,L\})} \mathop{\mathbb{E}}_{r_t, r_t' \sim p_t} \left[\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left[2L \sum_{i=t+1}^{T} \epsilon_i f[t] - L_{t-1}(f) + \epsilon_t (r_t' - r_t) \cdot f[t] \right] \right]
\end{aligned}$$

By passing to the worst-case choice of r_t, r_t' (which is achieved at the endpoints because of convexity), we obtain a further upper bound

$$\begin{aligned}
&\sup_{r_t, r_t' \in \{L,-L\}} \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left[2L \sum_{i=t+1}^{T} \epsilon_i f[t] - L_{t-1}(f) + \epsilon_t (r_t' - r_t) \cdot f[t] \right] \\
&\leq \sup_{r_t \in \{L,-L\}} \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left[2L \sum_{i=t+1}^{T} \epsilon_i f[t] - L_{t-1}(f) + 2\epsilon_t r_t \cdot f[t] \right] \\
&= \sup_{r_t \in \{L,-L\}} \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left[2L \sum_{i=t}^{T} \epsilon_i f[t] - L_{t-1}(f) \right] \\
&= \mathbf{Rel}_T \left(\mathcal{F} | \partial\ell(\hat{y}_1, y_1), \ldots, \partial\ell(\hat{y}_{t-1}, y_{t-1}) \right)
\end{aligned}$$

Thus we see that the relaxation is admissible. Now the corresponding prediction is given by

$$\begin{aligned}
\hat{y}_t &= \operatorname*{argmin}_{\hat{y}} \sup_{r_t \in [-L,L]} \left\{ r_t \hat{y} + \mathop{\mathbb{E}}_{\epsilon} \left[\sup_{f \in \mathcal{F}} \left\{ 2L \sum_{i=t+1}^{T} \epsilon_i f[i] - \sum_{i=1}^{t-1} \partial\ell(\hat{y}_i, y_i) f[i] - r_t f[t] \right\} \right] \right\} \\
&= \operatorname*{argmin}_{\hat{y}} \sup_{r_t \in [-L,L]} \left\{ r_t \hat{y} + \mathop{\mathbb{E}}_{\epsilon} \left[\sup_{f \in \mathcal{F}} \left\{ 2L \sum_{i=t+1}^{T} \epsilon_i f[i] - \sum_{i=1}^{t-1} \partial\ell(\hat{y}_i, y_i) f[i] - r_t f[t] \right\} \right] \right\} \\
&= \operatorname*{argmin}_{\hat{y}} \sup_{r_t \in \{-L,L\}} \left\{ r_t \hat{y} + \mathop{\mathbb{E}}_{\epsilon} \left[\sup_{f \in \mathcal{F}} \left\{ 2L \sum_{i=t+1}^{T} \epsilon_i f[i] - \sum_{i=1}^{t-1} \partial\ell(\hat{y}_i, y_i) f[i] - r_t f[t] \right\} \right] \right\}
\end{aligned}$$

The last step holds because of convexity of the term inside the supremum over r_t is convex in r_t and so the suprema is attained at the endpoints of the interval. The $\hat{y}_t$ above is attained when both terms of the supremum are equalized, that is for $\hat{y}_t$ is the prediction that satisfies :

$$\hat{y}_t = \mathop{\mathbb{E}}_{\epsilon} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{i=t+1}^{T} \epsilon_i f[i] - \frac{1}{2L} \sum_{i=1}^{t-1} \partial\ell(\hat{y}_i, y_i) f[i] + \frac{1}{2} f[t] \right\} - \sup_{f \in \mathcal{F}} \left\{ \sum_{i=t+1}^{T} \epsilon_i f[i] - \frac{1}{2L} \sum_{i=1}^{t-1} \partial\ell(\hat{y}_i, y_i) f[i] - \frac{1}{2} f[t] \right\} \right]$$

Finally since the relaxation is admissible we can conclude that the regret of the algorithm is bounded as

$$\mathbf{Reg}_T \le \mathbf{Rel}_T(\mathcal{F}) = 2\ L\ \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} \epsilon_t f[t]\right].$$

This concludes the proof. □

Proof of Lemma 20. The proof is similar to that of Lemma 19, with a few more twists. We want to establish admissibility of the relaxation given in (37) w.r.t. the randomized strategy q_t we provided. To this end note that

$$\begin{aligned}
&\sup_{y_t}\left\{\mathop{\mathbb{E}}_{\hat{y}_t\sim q_t}[\ell(\hat{y}_t, y_t)] + \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i] - L_t(f)\right\}\right]\right\} \\
&= \sup_{y_t}\left\{\mathop{\mathbb{E}}_{\epsilon}[\ell(\hat{y}_t(\epsilon), y_t)] + \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i] - L_t(f)\right\}\right]\right\} \\
&\le \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{y_t}\left\{\ell(\hat{y}_t(\epsilon), y_t) + \sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i] - L_t(f)\right\}\right\}\right]
\end{aligned}$$

by Jensen's inequality, with the usual notation $L_t(f) = \sum_{i=1}^{t}\ell(f[i], y_i)$. Further, by convexity of the loss, we may pass to the upper bound

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{y_t}\left\{\partial\ell(\hat{y}_t(\epsilon), y_t)\hat{y}_t(\epsilon) + \sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i] - L_{t-1}(f) - \partial\ell(\hat{y}_t(\epsilon), y_t)f[t]\right\}\right\}\right] \\
&\le \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{y_t}\left\{\mathop{\mathbb{E}}_{r_t}[r_t\cdot\hat{y}_t(\epsilon)] + \sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i] - L_{t-1}(f) - \mathop{\mathbb{E}}_{r_t}[r_t\cdot f[t]]\right\}\right\}\right]
\end{aligned}$$

where r_t is a $\{\pm L\}$-valued random variable with the mean $\partial\ell(\hat{y}_t(\epsilon), y_t)$. With the help of Jensen's inequality, and passing to the worst-case r_t (observe that this is legal for any given ϵ), we have an upper bound

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{y_t}\left\{\mathop{\mathbb{E}}_{r_t\sim\partial\ell(\hat{y}_t(\epsilon),y_t)}[r_t\cdot\hat{y}_t(\epsilon)] + \mathop{\mathbb{E}}_{r_t\sim\partial\ell(\hat{y}_t(\epsilon),y_t)}\left[\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i] - L_{t-1}(f) - r_t\cdot f[t]\right\}\right]\right\}\right] \\
&\le \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{r_t\in\{\pm L\}}\left\{r_t\cdot\hat{y}_t(\epsilon) + \sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i] - L_{t-1}(f) - r_t\cdot f[t]\right\}\right\}\right] \qquad (64)
\end{aligned}$$

Now the strategy we defined is

$$\hat{y}_t(\epsilon) = \mathop{\mathrm{argmin}}_{\hat{y}_t}\ \sup_{r_t\in\{\pm L\}}\left\{r_t\cdot\hat{y}_t(\epsilon) + \sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i] - \sum_{i=1}^{t-1}\ell(f[i], y_i) - r_t\cdot f[t]\right\}\right\}$$

which can be re-written as

$$\hat{y}_t(\epsilon) = \left(\sup_{f\in\mathcal{F}}\left\{\sum_{i=t+1}^{T}\epsilon_i f[i] - \frac{1}{2L}L_{t-1}(f) + \frac{1}{2}f[t]\right\} - \sup_{f\in\mathcal{F}}\left\{\sum_{i=t+1}^{T}\epsilon_i f[i] - \frac{1}{2L}L_{t-1}(f) - \frac{1}{2}f[t]\right\}\right)$$

By this choice of $\hat{y}_t(\epsilon)$, plugging back in Equation (64) we see that

$$\begin{aligned}
&\sup_{y_t}\left\{\mathop{\mathbb{E}}_{\hat{y}_t\sim q_t}\left[\ell(\hat{y}_t,y_t)\right]+\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i]-L_t(f)\right\}\right]\right\}\\
&\quad\le\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{r_t\in\{\pm L\}}\left\{r_t\cdot\hat{y}_t(\epsilon)+\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i]-L_{t-1}(f)-r_t\cdot f[t]\right\}\right\}\right]\\
&\quad=\mathop{\mathbb{E}}_{\epsilon}\left[\inf_{\hat{y}_t}\sup_{r_t\in\{\pm L\}}\left\{r_t\cdot\hat{y}_t+\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i]-L_{t-1}(f)-r_t\cdot f[t]\right\}\right\}\right]\\
&\quad=\mathop{\mathbb{E}}_{\epsilon}\left[\inf_{\hat{y}_t}\sup_{p_t\in\Delta(\{\pm L\})}\mathop{\mathbb{E}}_{r_t\sim p_t}\left\{r_t\cdot\hat{y}_t+\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i]-L_{t-1}(f)-r_t\cdot f[t]\right\}\right\}\right]
\end{aligned}$$

The expression inside the supremum is linear in p_t, as it is an expectation. Also note that the term is convex in $\hat{y}_t$, and the domain $\hat{y}_t\in[-\sup_{f\in\mathcal{F}}|f[t]|,\sup_{f\in\mathcal{F}}|f[t]|]$ is a bounded interval (hence, compact). We conclude that we can use the minimax theorem, yielding

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{p_t\in\Delta(\{\pm L\})}\inf_{\hat{y}_t}\mathop{\mathbb{E}}_{r_t\sim p_t}\left[r_t\cdot\hat{y}_t+\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i]-L_{t-1}(f)-r_t\cdot f[t]\right\}\right]\right]\\
&=\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{p_t\in\Delta(\{\pm L\})}\left\{\inf_{\hat{y}_t}\mathop{\mathbb{E}}_{r_t\sim p_t}\left[r_t\cdot\hat{y}_t\right]+\mathop{\mathbb{E}}_{r_t\sim p_t}\left[\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i]-L_{t-1}(f)-r_t\cdot f[t]\right\}\right]\right\}\right]\\
&=\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{p_t\in\Delta(\{\pm L\})}\left\{\mathop{\mathbb{E}}_{r_t\sim p_t}\left[\sup_{f\in\mathcal{F}}\left\{\inf_{\hat{y}_t}\mathop{\mathbb{E}}_{r_t\sim p_t}\left[r_t\cdot\hat{y}_t\right]+2L\sum_{i=t+1}^{T}\epsilon_i f[i]-L_{t-1}(f)-r_t\cdot f[t]\right\}\right]\right\}\right]\\
&\le\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{p_t\in\Delta(\{\pm L\})}\left\{\mathop{\mathbb{E}}_{r_t\sim p_t}\left[\sup_{f\in\mathcal{F}}\left\{\mathop{\mathbb{E}}_{r_t\sim p_t}\left[r_t\cdot f[t]\right]+2L\sum_{i=t+1}^{T}\epsilon_i f[i]-L_{t-1}(f)-r_t\cdot f[t]\right\}\right]\right\}\right]
\end{aligned}$$

In the last step, we replaced the infimum over $\hat{y}_t$ with $f[t]$, only increasing the quantity. Introducing an i.i.d. copy r_t' of r_t,

$$\begin{aligned}
&=\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{p_t\in\Delta(\{\pm L\})}\left\{\mathop{\mathbb{E}}_{r_t\sim p_t}\left[\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i]-L_{t-1}(f)+\left(\mathop{\mathbb{E}}_{r_t\sim p_t}\left[r_t\right]-r_t\right)\cdot f[t]\right\}\right]\right\}\right]\\
&\le\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{p_t\in\Delta(\{\pm L\})}\left\{\mathop{\mathbb{E}}_{r_t,r_t'\sim p_t}\left[\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i]-L_{t-1}(f)+(r_t'-r_t)\cdot f[t]\right\}\right]\right\}\right]
\end{aligned}$$

Introducing the random sign ϵ_t and passing to the supremum over r_t,r_t', yields the upper bound

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{p_t\in\Delta(\{\pm L\})}\left\{\mathbb{E}_{r_t,r_t'\sim p_t}\mathop{\mathbb{E}}_{\epsilon_t}\left[\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i]-L_{t-1}(f)+(r_t'-r_t)\cdot f[t]\right\}\right]\right\}\right]\\
&\le\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{r_t,r_t'\in\{\pm L\}}\left\{\mathop{\mathbb{E}}_{\epsilon_t}\left[\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i]-L_{t-1}(f)+\epsilon_t\left(r_t'-r_t\right)\cdot f[t]\right\}\right]\right\}\right]\\
&\le\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{r_t,r_t'\in\{\pm L\}}\left\{\mathop{\mathbb{E}}_{\epsilon_t}\left[\sup_{f\in\mathcal{F}}\left\{L\sum_{i=t+1}^{T}\epsilon_i f[i]-\frac{1}{2}L_{t-1}(f)+\epsilon_t r_t'\cdot f[t]\right\}\right]\right\}\right]\\
&\qquad+\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{r_t,r_t'\in\{\pm L\}}\left\{\mathop{\mathbb{E}}_{\epsilon_t}\left[\sup_{f\in\mathcal{F}}\left\{L\sum_{i=t+1}^{T}\epsilon_i f[i]-\frac{1}{2}L_{t-1}(f)-\epsilon_t r_t\cdot f[t]\right\}\right]\right\}\right]
\end{aligned}$$

In the above we split the term in the supremum as the sum of two terms one involving r_t and other r_t' (other

terms are equally split by dividing by 2), yielding

$$\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{r_t\in\{\pm L\}}\left\{\mathop{\mathbb{E}}_{\epsilon_t}\left[\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i]-L_{t-1}(f)+2\ \epsilon_t\ r_t\cdot f[t]\right\}\right]\right\}\right]$$

The above step used the fact that the first term only involved r'_t and second only r_t and further ϵ_t and $-\epsilon_t$ have the same distribution. Now finally noting that irrespective of whether r_t in the above supremum is L or $-L$, since it is multiplied by ϵ_t we obtain an upper bound

$$\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t}^{T}\epsilon_i f[i]-L_{t-1}(f)\right\}\right]$$

We conclude that the relaxation

$$\mathbf{Rel}_T\left(\mathcal{F}|y_1,\dots,y_t\right)=\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in\mathcal{F}}\left\{2L\sum_{i=t+1}^{T}\epsilon_i f[i]-L_t(f)\right\}\right]$$

is admissible and further the randomized strategy where on each round we first draw ϵ's and then set

$$\begin{aligned}\hat{y}_t(\epsilon)&=\left(\sup_{f\in\mathcal{F}}\left\{\sum_{i=t+1}^{T}\epsilon_i f[i]-\frac{1}{2L}L_{t-1}(f)+\frac{1}{2}f[t]\right\}-\sup_{f\in\mathcal{F}}\left\{\sum_{i=t+1}^{T}\epsilon_i f[i]-\frac{1}{2L}L_{t-1}(f)-\frac{1}{2}f[t]\right\}\right)\\&=\left(\inf_{f\in\mathcal{F}}\left\{-\sum_{i=t+1}^{T}\epsilon_i f[i]+\frac{1}{2L}L_{t-1}(f)+\frac{1}{2}f[t]\right\}-\inf_{f\in\mathcal{F}}\left\{-\sum_{i=t+1}^{T}\epsilon_i f[i]+\frac{1}{2L}L_{t-1}(f)-\frac{1}{2}f[t]\right\}\right)\end{aligned}$$

is an admissible strategy. Hence, the expected regret under the strategy is bounded as

$$\mathbb{E}\left[\mathbf{Reg}_T\right]\le\mathbf{Rel}_T\left(\mathcal{F}\right)=2L\ \mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{T}\epsilon_i f[i]\right]$$

which concludes the proof. □

Proof of Lemma 23. The proof is almost identical to the proof of admissibility for the Mirror Descent relaxation, so let us only point out the differences. Let $\tilde{x}_{t-1}=\sum_{i=1}^{t}x_i$ and $\mu_{t-1}=\frac{1}{t-1}\tilde{x}_{t-1}$. Using the fact that x_t is σ_t-close to μ_{t-1}, we expand

$$\left(\left\|\tilde{x}_t\right\|^2+C\sum_{s=t+1}^{T}\sigma_s^2\right)^{1/2}\le\left(\left\|\tilde{x}_{t-1}\left(\frac{t}{t-1}\right)\right\|^2+\left\langle\nabla\tfrac{1}{2}\left\|\left(\frac{t}{t-1}\right)\tilde{x}_{t-1}\right\|^2,x_t-\mu_{t-1}\right\rangle+C\sum_{s=t+1}^{T}\sigma_s^2\right)^{1/2}$$

As before, pick $x_t=\beta\tilde{x}_{t-1}+\gamma y$ for some $y\in\mathrm{Kernel}(\nabla\|\tilde{x}_{t-1}\|^2)$. The above expression under the square root then becomes

$$\|\tilde{x}_{t-1}\|^2+\underbrace{\left(\frac{1}{(t-1)^2}+\frac{2}{t-1}+\left(\frac{t}{t-1}\right)^2\left(\beta-\frac{1}{t-1}\right)\right)}_{\beta'}\|\tilde{x}_{t-1}\|^2+C\sum_{s=t+1}^{T}\sigma_s^2,$$

and the only difference from the expression in (45) is that we have a β' instead of β under the square root. Taking the derivatives, we see that

$$\alpha=\frac{\left(1+\frac{1}{t-1}\right)^2}{2\sqrt{\|\tilde{x}_{t-1}\|^2+C\sum_{s=t}^{T}\sigma_s^2}}$$

forces $\beta'=0$ and we conclude admissibility.

Arriving at the Relaxation We upper bound the sequential Rademacher complexity as

$$\frac{2}{\alpha}\sup_{(\mathbf{x},\mathbf{x}')\in\mathcal{T}}\mathbb{E}_{\epsilon}\left[\sup_{f\in\mathcal{F}}\left\langle f,\alpha\sum_{s=t+1}^{T}\epsilon_s\left(\mathbf{x}_s(\epsilon)-\frac{1}{s-t}\sum_{\tau=t+1}^{s-1}\chi_\tau(\epsilon_\tau)\right)-\sum_{r=1}^{t}x_r\right\rangle\right]$$

$$\leq\frac{2R^2}{\alpha}+\frac{\alpha}{\lambda}\sup_{(\mathbf{x},\mathbf{x}')}\mathbb{E}_{\epsilon}\left\|\sum_{s=t+1}^{T}\epsilon_s\left(\mathbf{x}_s(\epsilon)-\frac{1}{s-t}\sum_{\tau=t+1}^{s-1}\chi_\tau(\epsilon_\tau)\right)-\sum_{r=1}^{t}x_r\right\|^2 \tag{65}$$

$$\leq\frac{2\sqrt{2}R}{\sqrt{\lambda}}\sqrt{\sup_{(\mathbf{x},\mathbf{x}')}\mathbb{E}_{\epsilon}\left[\left\|\sum_{s=t+1}^{T}\epsilon_s\left(\mathbf{x}_s(\epsilon)-\frac{1}{s-t}\sum_{\tau=t+1}^{s-1}\chi_\tau(\epsilon_\tau)\right)-\sum_{r=1}^{t}x_r\right\|^2\right]} \tag{66}$$

$$\leq\frac{2\sqrt{2}R}{\sqrt{\lambda}}\sqrt{\left\|\sum_{r=1}^{t}x_r\right\|^2+\sup_{(\mathbf{x},\mathbf{x}')}C\sum_{s=t+1}^{T}\left\|\mathbf{x}_s(\epsilon)-\frac{1}{s-t}\sum_{\tau=t+1}^{s-1}\chi_\tau(\epsilon_\tau)\right\|^2} \tag{67}$$

Since $(\mathbf{x},\mathbf{x}')\in\mathcal{T}$ are pairs of tree such that for any $\epsilon\in\{\pm1\}^T$ and any $t\in[T]$.

$$C(x_1,\dots,x_t,\chi_1(\epsilon_1),\dots,\chi_{t-1}(\epsilon_{t-1}),\mathbf{x}_t(\epsilon))=1$$

we can conclude that for any $\epsilon\in\{\pm1\}^T$ and any $t\in[T]$,

$$\left\|\mathbf{x}_t(\epsilon)-\frac{1}{t-1}\sum_{\tau=1}^{t-1}\chi_\tau(\epsilon_\tau)\right\|\leq\sigma_t$$

□

Proof of Lemma 24. Then Sequential Rademacher complexity can be upper bounded as

$$\begin{aligned}
\sup_{\mathbf{x}}\mathbb{E}_{\epsilon}\left[\left\|\sum_{i=1}^{t}x_t+\sum_{i=1}^{T-t}\epsilon_i\mathbf{x}_i(\epsilon)\right\|\right]&\leq\sup_{\mathbf{x}}\left(\mathbb{E}_{\epsilon}\left[\left\|\sum_{i=1}^{t}x_t+\sum_{i=1}^{T-t}\epsilon_i\mathbf{x}_i(\epsilon)\right\|^p\right]\right)^{1/p}\\
&\leq\sup_{\mathbf{x}}\left(\mathbb{E}_{\epsilon}\left[\left\|\sum_{i=1}^{t}x_t+\sum_{i=1}^{T-t}\epsilon_i\mathbf{x}_i(\epsilon)\right\|^p-C\sum_{i=1}^{T-t}\mathbb{E}_{\epsilon}\left[\|\mathbf{x}_i(\epsilon)\|^p\right]\right]+C(T-t)\right)^{1/p}\\
&=\left(\Psi^*\left(\sum_{i=1}^{t}x_i\right)+C(T-t)\right)^{1/p}\\
&\leq\left(\Psi^*\left(\sum_{i=1}^{t-1}x_i\right)+\left\langle\nabla\Psi^*\left(\sum_{i=1}^{t-1}x_i\right),x_t\right\rangle+C(T-t+1)\right)^{1/p}
\end{aligned}$$

and admissibility is verified in a similar way to the 2-smooth case in the Section 3. Here we instead use p-smoothness which follows from result in [20]. The form of update specified by the relaxation in this case follows exactly the proof of Proposition 4, yielding

$$f_t=-\frac{\nabla\Psi^*(\sum_{j=1}^{t-1}x_i)}{p\left(\Psi^*(\sum_{j=1}^{t-1}x_i)+C(T-t+1)\right)^{1/p}}$$

□

Lemma 26. *The regret upper bound*

$$\sum_{t=1}^{T}\ell(f_t,x_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f,x_t)\leq\sum_{t=1}^{T}\ell(f_t,x_t)-\sum_{i=1}^{m}\inf_{f\in\mathcal{F}^{k_i}\left(x_1,\dots,x_{\tilde{k}_{i-1}}\right)}\sum_{t=\tilde{k}_{i-1}+1}^{\tilde{k}_i}\ell(f,x_t)\ . \tag{68}$$

is valid.

Proof of Lemma 26. To prove this inequality, it is enough to show that it holds for subdividing T into two blocks k_1 and k_2. Rearranging, we would like to show that

$$\inf_{f\in\mathcal{F}} \sum_{t=1}^{k_1} \ell(f, x_t) \; + \inf_{f\in\mathcal{F}^{k_2}(x_1,\ldots,x_{k_1})} \sum_{t=k_1+1}^{T} \ell(f, x_t) \le \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f, x_t)$$

for $k_1 + k_2 = T$. Observe, that the comparator term becomes only smaller if we pass to two instead of one infima, but we must check that no function f that minimizes the loss over both blocks (that is, the right hand side) is removed from being a potential minimizer over the second block. This is exactly the definition of $\mathcal{F}^{k_2}(x_1, \ldots, x_{k_1})$, and so the inequality is verified. We can now recurse and break up the first block in a similar manner, thus proving the statement of the lemma. □

Lemma 27. *The relaxation*

$$\mathbf{Rel}_T\left(\mathcal{F}|x_1,\ldots,x_t\right) = -\inf_{f\in\mathcal{F}} \sum_{i=1}^{t} x_i(f) + (T-t) \inf_{f\in\mathcal{F}} \sup_{f'\in\mathcal{F}} \|f - f'\|$$

is admissible.

Proof of Lemma 27. First,

$$\mathbf{Rel}_T\left(\mathcal{F}|x_1,\ldots,x_T\right) = -\inf_{f\in\mathcal{F}} \sum_{t=1}^{T} x_t(f).$$

As for admissibility,

$$\begin{aligned}
&\inf_{f_t\in\mathcal{F}} \sup_{x} \left\{x(f_t) + \mathbf{Rel}_T\left(\mathcal{F}|x_1,\ldots,x_{t-1},x\right)\right\} \\
&= \inf_{f_t\in\mathcal{F}} \sup_{x} \left\{x(f_t) - \inf_{f\in\mathcal{F}} \left\{\sum_{i=1}^{t-1} x_i(f) + x(f)\right\}\right\} + (T-t) \inf_{f\in\mathcal{F}} \sup_{f'\in\mathcal{F}} \|f - f'\| \\
&\le \inf_{f_t\in\mathcal{F}} \sup_{x} \left\{x(f_t) - \inf_{f\in\mathcal{F}} \sum_{i=1}^{t-1} x_i(f) - \inf_{f\in\mathcal{F}} x(f)\right\} + (T-t) \inf_{f\in\mathcal{F}} \sup_{f'\in\mathcal{F}} \|f - f'\| \\
&\le \inf_{f_t\in\mathcal{F}} \sup_{x} \left\{\sup_{f\in\mathcal{F}} \langle \nabla x, f_t - f\rangle - \inf_{f\in\mathcal{F}} \sum_{i=1}^{t-1} x_i(f)\right\} + (T-t) \inf_{f\in\mathcal{F}} \sup_{f'\in\mathcal{F}} \|f - f'\| \\
&\le \inf_{f_t\in\mathcal{F}} \left\{\sup_{f\in\mathcal{F}} \|f_t - f\| - \inf_{f\in\mathcal{F}} \sum_{i=1}^{t-1} x_i(f)\right\} + (T-t) \inf_{f\in\mathcal{F}} \sup_{f'\in\mathcal{F}} \|f - f'\| \\
&= \mathbf{Rel}_T\left(\mathcal{F}|x_1,\ldots,x_{t-1}\right)
\end{aligned}$$

□

Acknowledgements

We gratefully acknowledge the support of NSF under grants CAREER DMS-0954737 and CCF-1116928.

References

[1] J. Abernethy, A. Agarwal, P. L. Bartlett, and A. Rakhlin. A stochastic view of optimal regret through minimax duality. In *COLT '09*, 2009.

[2] J. Abernethy, P. L. Bartlett, A. Rakhlin, and A. Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of The Twenty First Annual Conference on Learning Theory*, 2008.

[3] J. Abernethy, M.K. Warmuth, and J. Yellin. Optimal strategies from random walks. In *COLT*, pages 437–445, 2008.

[4] P.L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[5] P.L. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. *Advances in Neural Information Processing Systems*, 20:65–72, 2007.

[6] S. Ben-David, D. Pál, and S. Shalev-Shwartz. Agnostic online learning. In *COLT*, 2009.

[7] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[8] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2):321–352, 2007.

[9] N. Cesa-Bianchi and O. Shamir. Efficient online learning via randomized rounding. In *NIPS*, 2011.

[10] K. Chaudhuri, Y. Freund, and D. Hsu. A parameter-free hedging algorithm. *Arxiv preprint arXiv:0903.2851*, 2009.

[11] E. Hazan and S. Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80(2):165–188, 2010.

[12] A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *J. Comput. Syst. Sci.*, 71(3):291–307, 2005.

[13] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

[14] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 04 1988.

[15] H. Narayanan and A. Rakhlin. Random walk approach to regret minimization. In *NIPS*, 2010.

[16] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In *NIPS*, 2010. Available at http://arxiv.org/abs/1006.1138.

[17] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Beyond regret. In *COLT*, 2011. Available at http://arxiv.org/abs/1011.3168.

[18] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. In *NIPS*, 2011. Available at http://arxiv.org/abs/1104.5070.

[19] O. Shamir and S. Shalev-Shwartz. Collaborative filtering with the trace norm: Learning, bounding, and transducing. In *COLT*, 2011.

[20] N. Srebro, K. Sridharan, and A. Tewari. On the universality of online mirror descent. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2645–2653. 2011.

[21] K. Sridharan and A. Tewari. Convex games in banach spaces. In *Proceedings of the 23nd Annual Conference on Learning Theory*, 2010.

[22] T. van Erven, P. Grünwald, W. M. Koolen, and S. de Rooij. Adaptive Hedge. In *Advances in Neural Information Processing Systems 24*, 2011.