2013

Localization and Adaptation in Online Learning

Alexander Rakhlin
University of Pennsylvania

Ohad Shamir

Karthik Sridharan
University of Pennsylvania

Recommended Citation

Rakhlin, A., Shamir, O., & Sridharan, K. (2013). Localization and Adaptation in Online Learning. *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics, 31* 516-526. Retrieved from http://repository.upenn.edu/statistics_papers/130

Localization and Adaptation in Online Learning

Abstract

We introduce a formalism of *localization* for online learning problems, which, similarly to statistical learning theory, can be used to obtain fast rates. In particular, we introduce local sequential Rademacher complexities and other local measures. Based on the idea of relaxations for deriving algorithms, we provide a template method that takes advantage of localization. Furthermore, we build a general adaptive method that can take advantage of the suboptimality of the observed sequence. We illustrate the utility of the introduced concepts on several problems. Among them is a novel upper bound on regret in terms of classical Rademacher complexity when the data are i.i.d.

Disciplines

Statistics and Probability

Localization and Adaptation in Online Learning

Alexander Rakhlin
University of Pennsylvania

Ohad Shamir
Microsoft Research

Karthik Sridharan
University of Pennsylvania

Abstract

We introduce a formalism of *localization* for online learning problems, which, similarly to statistical learning theory, can be used to obtain fast rates. In particular, we introduce local sequential Rademacher complexities and other local measures. Based on the idea of relaxations for deriving algorithms, we provide a template method that takes advantage of localization. Furthermore, we build a general adaptive method that can take advantage of the suboptimality of the observed sequence. We illustrate the utility of the introduced concepts on several problems. Among them is a novel upper bound on regret in terms of classical Rademacher complexity when the data are i.i.d.

1 Introduction

The online learning framework has been a popular alternative to the well-studied setting of statistical learning theory. In the latter, the i.i.d. assumption on data makes it possible to leverage the rich set of tools developed within statistics and probability theory. In contrast, the online learning framework [4] deals with adversarial sequences of data, or sequences with some non-i.i.d. structure [12].

One unsatisfying aspect of the developments in the online learning literature 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 a case-by-case basis, such as for online optimization of strongly convex functions. In this paper we show that a localized analysis can be performed at an abstract level, and it goes hand-in-hand with the idea of relaxations, introduced in [10]. 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 explain the idea of relaxations, introduced in [10], as well as the meta algorithm based on these relaxations, and present a few examples. Section 3 is devoted to a new formalism of localized complexities, and we present a basic localized meta algorithm. In Section 4, we combine the idea of localization and relaxations, thus showing how to obtain localized complexities. We show, in particular, that for strongly convex objectives, the regret is easily bounded through localization. Next, in Section 5, we present an adaptive method that constantly checks whether the sequence being played by the adversary is in fact minimax optimal and adapts accordingly. We show how this algorithm recovers known adaptive fast rate results. Furthermore, we demonstrate how local data-dependent norms arise naturally from our framework.

Notation: A set $\{x_1, \ldots, x_t\}$ 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 $\{1, \ldots, T\}$ 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 $(\epsilon_1, \ldots, \epsilon_T)$ of i.i.d.

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31. Copyright 2013 by the authors.

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 [11]). We often write $\mathbf{x}_t(\epsilon)$ instead of $\mathbf{x}_t(\epsilon_{1:t-1})$.

2 Relaxations and Meta-Algorithms

Let $\mathcal{F}$ be the set of learner's moves and $\mathcal{X}$ the set of possible outcomes (moves) chosen by Nature. The online learning problem follows the following protocol: 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*

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

where $\ell : \mathcal{F}\times\mathcal{X} \to \mathbb{R}$ is a known loss function which we assume is bounded by 1. Adopting the game-theoretic language, the online learning framework can be seen as a multi-stage two-player game with a payoff at the end of T rounds.

A *relaxation* **Rel** is a sequence of real-valued functions $\mathbf{Rel}_T(\mathcal{F}|x_1,\dots,x_t)$ 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 $\forall x_1,\dots,x_T \in \mathcal{X}$,

$$\begin{aligned}&\mathbf{Rel}_T(\mathcal{F}|x_1,\dots,x_t) \qquad (1)\\ &\geq \inf_{q\in\Delta(\mathcal{F})} \sup_{x\in\mathcal{X}} \left\{ \mathop{\mathbb{E}}_{f\sim q}[\ell(f,x)] + \mathbf{Rel}_T(\mathcal{F}|x_1,\dots,x_t,x) \right\}\end{aligned}$$

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

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

A strategy q that minimizes the expression in (1) defines an optimal algorithm for the relaxation **Rel**. This algorithm is given below under the name "Meta-Algorithm". However, minimization need not be exact: any q that satisfies the admissibility condition (1) is a valid method, and we will say that such an algorithm is *admissible with respect to the relaxation* **Rel**.

Algorithm 1 Meta-Algorithm **MetAlgo**

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

Proposition 1 ([10]). *Let* **Rel** *be an admissible relaxation. For any admissible algorithm with respect to* **Rel**, *including the* Meta-Algorithm, *irrespective of the strategy of the adversary,*

$$\sum_{t=1}^{T} \mathbb{E}_{f_t\sim q_t} \ell(f_t, x_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f, x_t) \leq \mathbf{Rel}_T(\mathcal{F}) \quad , \qquad (2)$$

and therefore, $\mathbb{E}[\mathbf{Reg}_T] \leq \mathbf{Rel}_T(\mathcal{F})$. *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* $1-\delta$,

$$\mathbf{Reg}_T \leq \mathbf{Rel}_T(\mathcal{F}) + (b-a)\sqrt{T/2\cdot\log(2/\delta)} \ .$$

Further, if for all $t \in [T]$, *the admissible strategies* q_t *are deterministic,* $\mathbf{Reg}_T \leq \mathbf{Rel}_T(\mathcal{F})$.

It was shown in [10] that the idea of relaxations unifies the vast majority of known online learning methods, including such unorthodox algorithms as Follow the Perturbed Leader. Moreover, a principled way of arriving at relaxations was shown: they naturally arise as upper bounds on the conditional value of the game. One of the tightest such upper bounds is achieved through symmetrization. The *conditional Sequential Rademacher complexity*

$$\begin{aligned}&\mathfrak{R}_T(\mathcal{F}|x_1,\dots,x_t) = \qquad (3)\\ &\sup_{\mathbf{x}} \mathop{\mathbb{E}}_{\epsilon_{t+1:T}} \sup_{f\in\mathcal{F}} \left[2\sum_{s=t+1}^{T} \epsilon_s \ell(f, \mathbf{x}_{s-t}(\epsilon_{t+1:s-1})) - \sum_{s=1}^{t} \ell(f, x_s) \right]\end{aligned}$$

can be shown to be an admissible relaxation [10]. 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 $\mathfrak{R}_T(\mathcal{F})$, which is

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

For computational purposes, further upper bounds (relaxations) on the conditional Rademacher complexity are sought in order to remove the supremum over the trees $\mathbf{x}$. Various techniques can be employed, including random playout, or moment-type inequalities as shown in the next example.

Suppose $\mathcal{F}$ is a finite class and $|\ell(f,x)| \leq 1$. The following relaxation is an upper bound on conditional sequential Rademacher complexity and it yields a parameter-free version of Exponential Weights:

$$\begin{aligned}&\mathbf{Rel}_T(\mathcal{F}|x_1,\dots,x_t) \qquad (4)\\ &= \inf_{\lambda>0} \left\{ \frac{1}{\lambda} \log\left(\sum_{f\in\mathcal{F}} \exp\left(-\lambda \sum_{i=1}^{t} \ell(f, x_i) \right) \right) + 2\lambda(T-t) \right\}\end{aligned}$$

This relaxation will be used later in the paper in the context of localized complexities.

3 Localization

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 [9, 2]. 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? The answer is, in fact, quite natural: As we obtain more data, we can rule out parts of $\mathcal{F}$ as those that are unlikely to be good solutions for the remainder of the learning game or for the next block of rounds. 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 ℓ or the sets $\mathcal{F}$ and $\mathcal{X}$.

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

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

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 $\tilde{k}_j \triangleq \sum_{i=1}^j k_i$. Consider subdividing T into blocks of time $k_1,\dots,k_m \in [T]$ 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

$$\begin{aligned} &\sum_{t=1}^T \ell(f_t,x_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^T \ell(f,x_t) \qquad (5) \\ &\le \sum_{t=1}^T \ell(f_t,x_t) - \sum_{i=1}^m \inf_{f\in\mathcal{F}^{k_i}\left(x_{1:\tilde{k}_{i-1}}\right)} \sum_{t=\tilde{k}_{i-1}+1}^{\tilde{k}_i} \ell(f,x_t) \\ &= \sum_{i=1}^m \left(\sum_{t=\tilde{k}_{i-1}+1}^{\tilde{k}_i} \ell(f,x_t) - \inf_{f\in\mathcal{F}^{k_i}\left(x_{1:\tilde{k}_{i-1}}\right)} \sum_{t=\tilde{k}_{i-1}+1}^{\tilde{k}_i} \ell(f,x_t) \right) \\ &= \sum_{i=1}^m \mathbf{Reg}_{k_i}\left(f_{\tilde{k}_{i-1}+1:\tilde{k}_i}, x_{\tilde{k}_{i-1}+1:\tilde{k}_i}, \mathcal{F}^{k_i}(x_{1:\tilde{k}_{i-1}}) \right) \end{aligned}$$

The short inductive proof of inequality (5) is given in Appendix, Lemma 8.

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 in $\mathcal{F}^{k_i}(x_1,\dots,x_{\tilde{k}_{i-1}})$. Further, if we only consider learner's strategies that pick from the set $\mathcal{F}^{k_i}(x_1,\dots,x_{\tilde{k}_{i-1}})$ when playing in the corresponding block i, we only weaken the learner, leading to the upper bound on regret.

We may take a minimax point of view [1, 11, 10]. The *value* $\mathcal{V}_T(\mathcal{F})$ of the game is defined as the best regret the learner can achieve if she and Nature play optimally. As a consequence of the above decomposition (5), we have that

$$\mathcal{V}_T(\mathcal{F}) \le \sum_{i=1}^m \mathcal{V}_{k_i}\left(\mathcal{F}^{k_i}(x_1,\dots,x_{\tilde{k}_{i-1}})\right) \qquad (6)$$

for any sequence $x_1,\dots,x_T$.

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 $k_1,\dots,k_m$ in advance. The player can choose k_i based on history $x_1,\dots,x_{\tilde{k}_{i-1}}$ and then use some learning algorithm to play the game within block k_i using the localized class $\mathcal{F}^{k_i}(x_1,\dots,x_{\tilde{k}_{i-1}})$. Such adaptive procedures will be considered in Section 5, but presently we assume that the block sizes $k_1,\dots,k_m$ are fixed.

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

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

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

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

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. Putting together all the ideas discussed so far, we have the following algorithm:

In the following lemma, we assume that the algorithm **MetAlgo** enjoys a regret bound of $\mathbf{Rel}_k(\mathcal{F}')$ for any number of rounds k and any subset $\mathcal{F}' \subseteq \mathcal{F}$.

Lemma 2. *For any choice of $k_1,\dots,k_m$ with $\sum_{i=1}^m k_i = T$, the regret of the Localized Meta-Algorithm is bounded as*

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

Algorithm 2 Localized Meta-Algorithm

```
Input: MetAlgo algorithm
Init. $t = 0$ and blocks $k_1, \dots, k_m$ s.t. $\sum_{i=1}^m k_i = T$
for $i = 1$ to $m$ do
    Play $k_i$ rounds using MetAlgo$\left(\mathcal{F}_{r(k_i; x_1, \dots, x_t)}\right)$
    and set $t = t + k_i$
end for
```

Of course, the above lemma still requires us to get a handle on regret over the localized subsets $\mathcal{F}_{r(k_i; x_1, \dots, x_t)}$. This is shown in the next section.

4 Local Sequential Complexities

We now combine the idea of relaxations and localization. As a start, we notice that 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*. The following corollary is a direct consequence of Lemma 2.

Corollary 3 (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, 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 [11]. One 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 $\mathbf{Rel}$ 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

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

An easy Lemma 9 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

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

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

$$\mathbf{Rel}_1(\mathcal{F}_{r_t}(x_{1:t})) \le \inf_{f \in \mathcal{F}_{r_t}(x_{1:t})} \sup_{f' \in \mathcal{F}_{r_t}(x_{1:t})} \| f - f' \|,$$

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 for strongly convex functions (e.g. [7]), and demonstrates the power of localization.

Example : IID Adversary

In this example we consider the case when the adversary outputs a sequence $x_1, \dots, x_T$ drawn iid from some fixed distribution $\mathbf{D}$ unknown to the learner. Recall the definition of the worst case classical (iid) Rademacher complexity:

$$\mathcal{R}_n(\mathcal{F}) = \sup_{\mathbf{D}} \mathop{\mathbb{E}}_{x_1, \dots, x_n \sim \mathbf{D}, \epsilon} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \ell(f, x_i) \right| \right],$$

where the supremum is over all distributions over $\mathcal{X}$. We now show that the idea of localization allows us to consider smaller subsets of $\mathcal{F}$ as the game progresses. This leads to a final regret bound given by the classical i.i.d. Rademacher complexity.

Lemma 4. *Without loss of generality assume $T = 2^m$. Fix a blocking strategy with $k_i = 2^i$, $i \in [m]$. Consider the empirical restriction property*

$$P(f \mid x_1, \dots, x_t) = \frac{1}{t}\sum_{i=1}^{t} \ell(f, x_i) - \inf_{f \in \mathcal{F}} \frac{1}{t}\sum_{i=1}^{t} \ell(f, x_i)$$

and a radius $r = 12\mathcal{R}_t(\mathcal{F}) + \sqrt{\frac{16 \log \frac{m}{\delta}}{t}}$. Under this localization, as long as loss is bounded by 1, with probability at least $1 - \delta$, the regret of the algorithm which chooses any element from the localized subset $\mathcal{F}_r(x_{1:\tilde{k}_{i-1}})$ is at most

$$44T\mathcal{R}_T(\mathcal{F}) + 21\sqrt{T \log \frac{\log(T)}{\delta}}.$$

To the best of our knowledge, the bound on regret in terms of i.i.d. Rademacher complexity in the case that Nature plays an i.i.d. sequence is novel, and it naturally arises from the idea of localization. We remark that regret for worst-case sequences (in the supervised scenario with absolute loss) has been shown in [11] to be characterized precisely by sequential (rather than iid) Rademacher complexity, which can be different from the iid complexity. Given the above lemma, we see that it is precisely the idea of localization that bridges the gap between i.i.d. sequences and worst-case sequences, as we are able to discard parts of $\mathcal{F}$, thanks to concentration of measure.

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 [3], Adaptive Hedge of [13], and the variance-based bounds of [5, 8] 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 ℓ.

The adaptive procedure below uses a subroutine **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 blocking 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.

Algorithm 3 Adaptive Localized Meta-Algorithm

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

Lemma 5. *Given some admissible relaxation* **Rel**, *the regret of the adaptive localized meta-algorithm (Algo-*

rithm 3) is bounded as

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

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*

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

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 [3], 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 simple blocking strategy

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

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

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

and relaxation

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

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

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

where k is the length of the current block. The next lemma shows that the proposed adaptive gradient descent recovers the results of [3]. The method is a mixture of Follow the Leader -style algorithm and a Gradient Descent -style algorithm.

Lemma 6. *The relaxation specified above is admissible. Suppose the adversary plays 1-Lipchitz convex functions $x_1,\ldots,x_T$ 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 [13]. 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

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

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

$$\mathbf{Block}(\{x_1,\ldots,x_t\},k) = (j, k-j) \qquad (7)$$

where

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

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 (4) used for the Exponential Weights algorithm.

Lemma 7. *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 $k \geq 1$ there exists $\tau \in [T]$ such that for all $t > \tau$ and all $f \neq \hat{f}_T$ the partial cumulative loss

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

Then for any loss function mapping to the interval $[0,1]$, *the regret of Algorithm 3 with the Exponential Weights relaxation, the blocking strategy (7) and the localization mentioned above is bounded as*

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

While we demonstrated a very simple example, the algorithm is adaptive more generally. Lemma 7 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 [6], 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

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

Hence, we see that it suffices to consider localizations

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

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

$$\mathcal{F}_{r(k; x_{1:t})} = \left\{ f \in \mathcal{F} : \sum_{i=1}^{t} \ell(f, x_i) - \ell(\hat{f}_t, x_i) \leq k \left\|f - \hat{f}_t\right\| \right\} \quad (8)$$

Further, using Taylor approximation we can pass to the localization

$$\mathcal{F}_{r(k; x_1, \ldots, x_t)} = \left\{ f \in \mathcal{F} : \tfrac{1}{2} \left\|f - \hat{f}_t\right\|^2_{x_1, \ldots, x_t} \leq k \left\|f - \hat{f}_t\right\| \right\} \quad (9)$$

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 lower bounded by

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

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.

Now notice that the set $\mathcal{F}_{r(k; x_1, \ldots, x_t)}$ consists of $f \in \mathcal{F}$ for which $\frac{1}{2}\left\|f - \hat{f}_t\right\|^2_{x_1, \ldots, x_T} \leq k \left\|f - \hat{f}_t\right\|$. However $f \in \mathcal{F}$ simply implies that that $\frac{1}{2}\left\|f - \hat{f}_t\right\|^2 \leq \left\|f - \hat{f}_t\right\|$ and so we can conclude that :

$$\begin{aligned}
&\mathcal{F}_{r(k; x_{1:t})} \\
&\subseteq \left\{ f \ : \ \frac{1}{2}\left(\left\|f - \hat{f}_t\right\|^2_{x_{1:t}} + \left\|f - \hat{f}_t\right\|^2 \right) \leq (k+1) \left\|f - \hat{f}_t\right\| \right\}
\end{aligned}$$

Therefore, one can use the above localized sets. For the Euclidean case the above becomes :

$$\mathcal{F}_{r(k; x_{1:t})} \subseteq \left\{ f \ : \ \frac{1}{2} \left\|f - \hat{f}_t\right\|^2_{H_t + I} \leq (k+1) \left\|f - \hat{f}_t\right\| \right\}$$

and one can use the above set to localize to the effective dimensionality of data so far. The associated blocking strategy for the adaptation we propose is

$$\begin{aligned}
&\mathbf{Block}\left(\{x_1, \ldots, x_t\}, k\right) \\
&= \begin{cases} (k) & \text{if } \sqrt{k} > \inf_{f \in \mathcal{F}} \frac{\left\|f - \hat{f}_t\right\|^2_{H_t + I}}{\left\|f - \hat{f}_t\right\|^2} \\ (1, 1, \ldots, 1) & \text{otherwise} \end{cases}
\end{aligned}$$

Notice that this blocking strategy automatically enjoys the same bound as in Lemma 6 for the setting in the

Lemma because $\frac{1}{2}\left\|\cdot\right\|_{H_t}^2 \geq \frac{\tilde{\sigma}_t}{2}\left\|\cdot\right\|_2^2$. However in general the bound could be much better as we do not just restrict based on minimal eigenvalue but rather the entire eigen-spectrum plays a role in the bound. For instance in the case the adversary plays exp-concave functions (Eg. square loss), this adaptive algorithm should enjoy much better bounds that depend on the eigen-spectrum of the data.

For the problem of general online convex optimization problems one can use localizations given in Equations (8) or (9). The localization in Equation (8) 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 (9). Of course, one needs to provide also an appropriate blocking strategy. A possible general blocking strategy could be

$$\mathbf{Block}(\{x_1,\ldots,x_t\},k) = (j, k-j)$$

where

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

6 Summary

In this paper we introduced a framework for studying localization and adaptation in the context of online learning. With the help of the generic relaxation mechanism from [10], we showed that the ideas of localization and adaptation can lead to new adaptive online learning algorithms that not only recover known fast rate results but also yield new and improved analyses.

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

[3] P.L. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. *NIPS*, 20:65–72, 2007.

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

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

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

[7] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Mach. Learn.*, 69(2-3):169–192, 2007.

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

[9] 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.

[10] A. Rakhlin, O. Shamir, and K. Sridharan. Relax and randomize: From value to algorithms. In *NIPS*, 2012.

[11] 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.

[12] 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.

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

A PROOFS

Proof of Lemma 4. For any $t \in [T]$ by symmetrization lemma we have that with probability $1-\delta$ over the sample,

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{t} \sum_{i=1}^{t} \ell(f, x_t) - \mathbb{E}\left[\ell(f,x)\right] \right| \le 2\mathcal{R}_t(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{t}}$$

Using this we conclude that with probability at least $1-\delta$,

$$\begin{aligned} \inf_{f\in\mathcal{F}} \frac{\sum_{i=1}^{t+k} \ell(f,x_i)}{t+k} &\le \inf_{f\in\mathcal{F}} \mathbb{E}\left[\ell(f,x)\right] + 2\mathcal{R}_{t+k}(\mathcal{F}) + \sqrt{\frac{\log\frac{1}{\delta}}{t+k}} \\ &\le \inf_{f\in\mathcal{F}} \frac{1}{t}\sum_{i=1}^{t} \ell(f,x_i) + 4\mathcal{R}_t(\mathcal{F}) + \sqrt{\frac{4\log\frac{1}{\delta}}{t}} \end{aligned}$$

and that with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} \left(\sum_{i=1}^{t} \ell(f,x_i) - \frac{t}{t+k} \right) \le 4t\mathcal{R}_t(\mathcal{F}) + \sqrt{4t\log\frac{1}{\delta}}$$

Hence, for any k and any t with probability $1-\delta$,

$$\begin{aligned} &\mathcal{F}^k(x_{1:t}) \subseteq \\ &\left\{ f : \sum_{i=1}^{t} \ell(f,x_i) - \inf_{f\in\mathcal{F}} \sum_{i=1}^{t} \ell(f,x_i) \le 8t\mathcal{R}_t(\mathcal{F}) + 4\sqrt{t\log\frac{1}{\delta}} \right\} \end{aligned}$$

This establishes the choice of the localized subsets: for each $j \in [m]$,

$$\begin{aligned} \mathcal{F}_r(x_{1:\tilde{k}_{j-1}}) := \Bigg\{ f \in\mathcal{F} : \sum_{i=1}^{\tilde{k}_{j-1}} \ell(f,x_i) - \inf_{f\in\mathcal{F}} \sum_{i=1}^{\tilde{k}_{j-1}} \ell(f,x_i) \\ \le 8\tilde{k}_{j-1}\mathcal{R}_{\tilde{k}_{j-1}}(\mathcal{F}) + 4\sqrt{\tilde{k}_{j-1}\log\frac{m}{\delta}} \Bigg\} \end{aligned}$$

Now in fact we argue that for the iid case once we localize as above it does not matter which elements of the localized set the meta-algorithm uses. That is the algorithm can simply pick any fixed $f_{\tilde{k}_j+1} \in \mathcal{F}^k(x_{1:\tilde{k}_j})$ for every round in the block. That is $f_t = f_{\tilde{k}_j+1}$ some arbitrary element of $\mathcal{F}^k(x_{1:\tilde{k}_j})$ for every $t \in [\tilde{k}_j+1, \dots, \tilde{k}_{j+1}]$. Hence we have that with probability at least $1-\delta$,

$$\begin{aligned} \sum_{t=1}^{T} \ell(f_t,x_t) &= \sum_{j=0}^{m-1} \sum_{i=\tilde{k}_j+1}^{\tilde{k}_{j+1}} \ell(f_{\tilde{k}_j+1}, x_i) \\ &\le \sum_{j=0}^{m-1} k_{j+1}\mathbb{E}\left[\ell(f_{\tilde{k}_j+1},x)\right] + \sum_{j=1}^{m} \sqrt{4k_j\log\frac{m}{\delta}} \end{aligned}$$

Now, using the fact that $f_{\tilde{k}_j+1}$ are almost ERM, the above expression is upper bounded by

$$\begin{aligned} &\left(\sum_{j=0}^{m-1} k_{j+1}\right) \inf_{f\in\mathcal{F}} \mathbb{E}\left[\ell(f,x)\right] + 12\sum_{j=0}^{m-1} k_{j+1}\mathcal{R}_{\tilde{k}_j}(\mathcal{F}) \\ &\quad + \sum_{j=1}^{m}\sqrt{4k_j\log\frac{m}{\delta}} + \sum_{j=1}^{m} k_j \sqrt{\frac{16\log\frac{m}{\delta}}{\tilde{k}_j}} \\ &\le T \inf_{f\in\mathcal{F}} \mathbb{E}\left[\ell(f,x)\right] + 12\sum_{j=0}^{m-1} k_{j+1}\mathcal{R}_{\tilde{k}_j}(\mathcal{F}) + 6\sum_{j=1}^{m}\sqrt{k_j\log\frac{m}{\delta}} \\ &\le \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f,x_t) + 12\sum_{j=0}^{m-1} k_{j+1}\mathcal{R}_{\tilde{k}_j}(\mathcal{F}) + 2T\mathcal{R}_T(\mathcal{F}) \\ &\quad + 6\sum_{j=1}^{m}\sqrt{k_j\log\frac{m}{\delta}} \end{aligned}$$

where we used $k_0 = 0$ and the convention, $\mathcal{R}_0(\mathcal{F}) = 1$. Now note that $\mathcal{R}_T(\mathcal{F})$ is always of order $1/\sqrt{T}$ or larger (by Kintchine's inequality). Hence using $k_j = 2^j$ we conclude that w.p. at least $1-\delta$,

$$\begin{aligned} &\sum_{t=1}^{T} \ell(f_t,x_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f,x_t) \\ &\qquad \le 44T\mathcal{R}_T(\mathcal{F}) + 21\sqrt{T\log\frac{\log(T)}{\delta}} \end{aligned}$$

□

Proof of Lemma 6. We start each block at $\hat{f}_t$. For the first block, $\hat{f}_t = 0$ and for later blocks, $\hat{f}_t$ is the empirical risk minimizer w.r.t. instances $x_1, \dots, x_t$. 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. We now start the proof by establishing the admissibility of the relaxation specified. To show admissibility, let us first check the initial condition:

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

Next we check the recursive inequality. To this end note that :

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

Hence note that :

$$\begin{aligned}
&\inf_{f_i : \|f_i\|} \sup_{y_i} \left\{ \langle f_i, y_i \rangle + \mathbf{Rel}_k\left(\mathcal{F}_{r(k;x_1,\ldots,x_t)} \middle| y_1,\ldots,y_i\right) \right\} \\
&= -\left\langle \hat{f}_t, \tilde{y}_{i-1} \right\rangle + \inf_{f_i : \|f_i\|} \sup_{y_i} \Bigg\{ \left\langle f_i - \hat{f}_t, y_i \right\rangle + 2\min\left\{1, \frac{k}{\tilde{\sigma}_t}\right\} \times \\
&\quad \times \sqrt{\left\|\tilde{y}_{i-1}\right\|^2 + \left\langle \nabla \tfrac{1}{2}\left\|\tilde{y}_{i-1}\right\|^2, y_i \right\rangle + C(k-i+1)} \Bigg\}
\end{aligned}$$

Writing $g_i = f_i - \hat{f}_t$, admissibility step and update form in terms of g_i is now identical to the admissibility and updates from [10], Proposition 4. Hence we concluded that the relaxation satisfies the recursive inequality and that the update in the block is given by

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

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 5. We will split the analysis to two cases, one when $\alpha > 1/2$ and other when $\alpha \le 1/2$.

Case $\alpha > \frac{1}{2}$:
The case when $\alpha > \frac{1}{2}$ is rather simple. This is because, at the beginning of each doubling block, the blocking strategy within that block is decided by checking if $\sqrt{2t} \le B(t)^\alpha$ for the $\alpha > 1/2$. However notice that since $\alpha > 1/2$ this inequality is never true (except for the initial constant number of rounds). Hence basically when $\alpha > 1/2$ we simply end up running gradient descent algorithm with doubling trick and initial vector within each doubling block given by the ERM so far. However since doubling trick guarantees a regret bound of $O(\sqrt{T})$ we can conclude the result for $\alpha > 1/2$.

Case $\alpha \le \frac{1}{2}$:
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

$$\sqrt{k} \le B(3t-k)^\alpha \tag{10}$$

The regret on this block can be bounded using Lemma 5 (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 :

$$\begin{aligned}
&\mathbf{Rel}_{2t-k}\left(\mathcal{F}_{r(x_1,\ldots,x_t)}\right) + \sum_{i=2t-k+1}^{2t} \mathbf{Rel}_1\left(\mathcal{F}_{r(x_1,\ldots,x_i)}\right) \\
&\le \sqrt{2t-k} + \sum_{j=2t-k+1}^{2t} \frac{1}{Bj^\alpha} \\
&\le \sqrt{2t} + \sum_{j=2t-k+1}^{2t} \frac{1}{Bj^\alpha} \\
&\le \sqrt{2t} + \frac{1}{B}\left((2t+1)^{1-\alpha} - (2t-k+1)^{1-\alpha}\right) \\
&\le \sqrt{2t} + \frac{k^{1-\alpha}}{B} \\
&\le \sqrt{2t} + \frac{B^{2(1-\alpha)}(3t)^{2\alpha(1-\alpha)}}{B} \quad \text{(using Eq. (10))} \\
&\le \sqrt{2t} + B^{2(1-\alpha)-1}\sqrt{3t} \\
&\le \sqrt{12\ t}
\end{aligned}$$

Hence overall regret is bounded as

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

This concludes the proof. □

Proof of Lemma 7. 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, $\underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^t \ell(f, x_i)$, 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. □

Lemma 8. *The regret upper bound*

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

is valid.

Proof of Lemma 8. 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

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

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 9. *The relaxation*

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

is admissible.

Proof of Lemma 9. 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,

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

The last quantity is upper bounded by

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

which, in turn, is upper bounded by

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

□