ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Online Learning Algorithms

Nicolò Cesa-Bianchi[1] and Francesco Orabona[2]

[1]Department of Computer Science and Data Science Research Center, Università degli Studi di Milano, Milano 20133, Italy; email: nicolo.cesa-bianchi@unimi.it

[2]Department of Electrical and Computer Engineering, Boston University, Boston, Massachusetts 02215, USA; email: francesco@orabona.com

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

Annu. Rev. Stat. Appl. 2021. 8:13.1–13.26

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-040620-035329

Copyright © 2021 by Annual Reviews.
All rights reserved

Keywords

regret minimization, convex optimization, regression, classification

Abstract

Online learning is a framework for the design and analysis of algorithms that build predictive models by processing data one at the time. Besides being computationally efficient, online algorithms enjoy theoretical performance guarantees that do not rely on statistical assumptions on the data source. In this review, we describe some of the most important algorithmic ideas behind online learning and explain the main mathematical tools for their analysis. Our reference framework is online convex optimization, a sequential version of convex optimization within which most online algorithms are formulated. More specifically, we provide an in-depth description of online mirror descent and follow the regularized leader, two of the most fundamental algorithms in online learning. As the tuning of parameters is a typically difficult task in sequential data analysis, in the last part of the review we focus on coin-betting, an information-theoretic approach to the design of parameter-free online algorithms with good theoretical guarantees.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

1. INTRODUCTION

The growing success of technologies based on machine learning is driven by the availability of massive data sets in digital format. Processing these large amounts of data poses computational challenges that are not always properly addressed by traditional statistical learning methods. For this reason, online or sequential learning, a framework specifically designed to cope with big data scenarios, has become a key tool in machine learning applications. Online algorithms go through the data points sequentially, using each new data point to adjust their predictive model or estimator. Typically, this adjustment is local, as it only involves the current model and the new data point. As each data point is often processed in constant time, this results in an overall running time scaling linearly with the number of data points. Besides the computational advantage, there are more reasons for which sequential learning may be preferred over other approaches. In many application domains—such as online advertising, digital markets, sensor networks, and mobile user applications—new data are generated at high rates. In these cases, the sequential adaptation process of online learning has the potential to capture subtle nonstationary features of the unknown data source.

A fundamental issue in machine learning is what mathematical assumptions on data sources are reasonable to make. Online learning advocates an approach in which the source is viewed as an arbitrary and unknown deterministic process. This is a radical departure from classical statistical approaches to sequential decision-making—such as Bayesian decision theory (Berger 2013) or Markov decision processes (Puterman 2014)—and finds its roots in the pioneering works on repeated games by Robbins (1951), Hannan (1957), and Blackwell (1956), where the data source consists of the opponent's plays in a two-person game. The theme of predicting individual, deterministic sequences also surfaced in other disciplines, including information theory (Cover 1967, Feder et al. 1992), and computer science (Borodin & El-Yaniv 2005). More recently, some of the online learning techniques, such as exponential weighted aggregation, have also appeared in the statistical literature (see, e.g., Dalalyan & Tsybakov 2008, Dalalyan & Salmon 2012, Rigollet & Tsybakov 2012).

Stripping the data source of any statistical assumption allows us to define a crisp, minimalistic framework for investigating the notion of algorithmic learning, where only the empirical properties of the observed data sequence matter. A substitute for the notion of statistical risk must then be introduced to define a notion of minimax optimality over the sequence of losses in a mathematically rigorous way. In view of that, one should note that in online learning we never measure the predictive power of a single model; rather, we consider the ensemble of models sequentially generated by the online learner while processing a data sequence. The notion of risk that we use to evaluate this ensemble is appropriately called sequential risk and measures the extent to which each model generated by the algorithm is able to predict the next element in the sequence. Sequential risk is thus associated with the behavior of an algorithm on an individual data sequence. Based on sequential risk, we then derive the notion of regret, which can be viewed as the online counterpart of the statistical excess risk measured with respect to a class of predictive models. The control of regret is the main goal in the analysis of online learning algorithms.

In the rest of this review, we introduce and describe some of the most fundamental online learning algorithms. Our goal is to explain the behavior of these algorithms through the analysis of their regret. Therefore, rather than going through as many as possible of the existing approaches to online learning, we prefer to focus on the conceptual foundations and the main proof techniques. We believe this is a more effective way to keeping alive the interest of someone who wants to know more about this exciting field of research.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

1.1. Online Convex Optimization

The standard framework for the study of parametric online learning with convex losses[1] is known as online convex optimization (OCO) (see, e.g., Shalev-Shwartz 2012, Hazan 2016, McMahan 2017, Orabona 2019). While it is also possible to design and study nonparametric online learning algorithms (e.g., Hazan & Megiddo 2007, De Rosa et al. 2015, Kuzborskij & Cesa-Bianchi 2017), we focus here on the more common parametric setting.

As we said earlier, the data-generating mechanism of online learning is an unknown and deterministic process. In OCO, the data process is replaced by a deterministic sequence of unknown and convex loss functions ℓ_t, evaluating the performance of the models incrementally generated by the algorithm. For example, if we want to cast linear regression in the OCO framework, then the loss functions ℓ_t take the form $\ell_t(\mathbf{w}) = (\mathbf{w}^\top \mathbf{x}_t - y_t)^2$, where $\mathbf{w} \in \mathbb{R}^d$ is a linear prediction model and $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots \in \mathbb{R}^d \times \mathbb{R}$ is the underlying deterministic data sequence.

Throughout this article, the model space $\mathbb{X}$ is a convex, closed, and nonempty subset of $\mathbb{R}^d$, while a loss function is any nonnegative and convex function $\ell_t : \mathbb{X} \to \mathbb{R}$. For any sequence ℓ_1, $\ell_2, \ldots$ of loss functions, an online learner A is a sequence $A_1, A_2, \ldots$ of mappings with range $\mathbb{X}$. The learner's model at time t is $\mathbf{w}_t = A_{t-1}(\ell_1, \ldots, \ell_{t-1})$, where $\mathbf{w}_1 = A_1 \in \mathbb{X}$ is the default initial model, and the learner's loss at time t is $\ell_t(\mathbf{w}_t)$. While the notation $\mathbf{w}_t = A_{t-1}(\ell_1, \ldots, \ell_{t-1})$ highlights the fact that $\mathbf{w}_{t+1}$ can depend on all the past losses, we are especially interested in cases where the update $\mathbf{w}_t \to \mathbf{w}_{t+1}$ can be done efficiently based only on $\mathbf{w}_t$ and ℓ_t. Also, for certain applications it makes sense to consider special cases of this framework, where additional assumptions besides convexity are made on the loss functions.

1.2. Regret and Sequential Risk

The performance of the learner is measured according to the regret

$$R_T(\mathbf{u}) = \sum_{t=1}^{T} \big(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u})\big) \qquad \text{for } \mathbf{u} \in \mathbb{X}.$$

Online learning is concerned with the design of algorithms for which $R_T(\mathbf{u})$ grows sublinearly in T for all $\mathbf{u} \in \mathbb{X}$ and irrespective to the loss sequence (the so-called no-regret property). The quantity $\frac{1}{T}\sum_{t=1}^{T} \ell_t(\mathbf{w}_t)$ is sometimes called sequential risk (as opposed to the classical statistical risk); thus, sublinear regret implies that the excess sequential risk,

$$\frac{1}{T}\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \frac{1}{T}\sum_{t=1}^{T} \ell_t(\mathbf{u}),$$

converges to zero for any $\mathbf{u} \in \mathbb{X}$.

Algorithms that enjoy the no-regret property can also be used to solve convex optimization problems $\min_{\mathbf{w}\in\mathbb{X}} f(\mathbf{w})$, which are viewed as instances of OCO with $\ell_t = f$ for all t. Indeed, Jensen's inequality shows that

$$f(\overline{\mathbf{w}}) - \min_{\mathbf{w}\in\mathbb{X}} f(\mathbf{w}) \le \frac{1}{T}\left(\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathbb{X}} \sum_{t=1}^{T} \ell_t(\mathbf{w})\right) \le \frac{R_T(\mathbf{w}^\star)}{T},$$

where $\overline{\mathbf{w}} = \frac{1}{T}(\mathbf{w}_1 + \cdots + \mathbf{w}_T)$ is the average of the iterates $\mathbf{w}_t$ and $\mathbf{w}^\star$ is the minimizer of f in $\mathbb{X}$. Similarly, we may also consider stochastic optimization problems $\min_{\mathbf{w}\in\mathbb{X}} \mathbb{E}[F(\mathbf{w}, \omega)]$, where

[1] Partial extensions of online learning to nonconvex losses have been recently considered by Agarwal et al. (2019).

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

the random variable $F(\mathbf{w}, \cdot)$ is the stochastic objective and $F(\cdot, \omega)$ is convex for all $\omega \in \Omega$. Given access to independent and identically distributed (i.i.d.) draws $\omega_1, \omega_2, \ldots$, we can solve stochastic optimization problems using a no-regret algorithm run with $\ell_t(\mathbf{w}) = F(\mathbf{w}, \omega_t)$. Let $\mathbf{w}^\star \in \operatorname{argmin}_{\mathbf{w}\in\mathbb{X}} \mathbb{E}[F(\mathbf{w}, \omega)]$ and observe that, using Jensen's inequality once more,

$$\mathbb{E}\big[F(\overline{\mathbf{w}}, \omega) - F(\mathbf{w}^\star, \omega)\big] \le \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\big(F(\mathbf{w}_t, \omega_t) - F(\mathbf{w}^\star, \omega_t)\big)\right]$$

$$= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\big(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}^\star)\big)\right] \le \mathbb{E}\left[\frac{R_T(\mathbf{w}^\star)}{T}\right].$$

In machine learning, stochastic optimization is typically used to solve empirical risk minimization problems:

$$\min_{\mathbf{w}\in\mathbb{X}} \frac{1}{m}\sum_{i=1}^{m} F\big(\mathbf{w}, \mathbf{z}_i\big), \qquad 1.$$

where $\mathbf{z}_1, \ldots, \mathbf{z}_m$ is a data set and $F(\mathbf{w}, \mathbf{z}_i)$ measures the loss of $\mathbf{w}$ on the data point $\mathbf{z}_i$. If $F(\cdot, \mathbf{z})$ is convex for all $\mathbf{z}$, then we may set $\ell_t = F(\cdot, \mathbf{Z}_t)$, where $\mathbf{Z}_1, \mathbf{Z}_2, \ldots$ are i.i.d. uniform draws from the data set.

1.3. Lower Bounds

Consider the easy case when $\mathbb{X}$ is a bounded[2] set with diameter D and all losses ℓ_t are Lipschitz on $\mathbb{X}$, how well can we control the regret R_T in this scenario? It turns out that the worst case for OCO occurs when all loss functions are linear. Interestingly, the proof uses a stochastic rather than deterministic loss process. More specifically, let $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{X}$ such that $\|\mathbf{v}_1 - \mathbf{v}_2\|_2 = D$ and define $\mathbf{z}_0 = \frac{\mathbf{v}_1 - \mathbf{v}_2}{\|\mathbf{v}_1 - \mathbf{v}_2\|_2}$. Stochastic losses $L_1, L_2, \ldots$ are defined by $L_t(\mathbf{w}) = \varepsilon_t L\, \mathbf{z}_0^\top \mathbf{w}$, where $\varepsilon_1, \varepsilon_2, \ldots$ are independent Rademacher random variables, i.e., $\mathbb{P}(\varepsilon_t = 1) = \mathbb{P}(\varepsilon_t = -1) = \frac{1}{2}$, and $L > 0$ is the Lipschitz constant for all the losses.

Now, fix any algorithm for OCO. Clearly, its regret satisfies

$$\max_{\varepsilon_1,\ldots,\varepsilon_T} \max_{\mathbf{u}\in\{\mathbf{v}_1,\mathbf{v}_2\}} R_T(\mathbf{u}) \ge \mathbb{E}\left[\max_{\mathbf{u}\in\{\mathbf{v}_1,\mathbf{v}_2\}} R_T(\mathbf{u})\right],$$

where the expectation is with respect to the random draw of $\varepsilon_1, \ldots, \varepsilon_T$. Moreover, since $\mathbb{E}[L_t(\mathbf{w})] = 0$ for all $\mathbf{w}$, we have

$$\mathbb{E}\left[\max_{\mathbf{u}\in\{\mathbf{v}_1,\mathbf{v}_2\}} R_T(\mathbf{u})\right] = \mathbb{E}\left[\max_{\mathbf{u}\in\{\mathbf{v}_1,\mathbf{v}_2\}} \sum_{t=1}^{T} L_t(\mathbf{u})\right]. \qquad 2.$$

Now, using the elementary identity $\max\{a, b\} = \frac{1}{2}(a + b + |a - b|)$ and Khintchine inequality (see, e.g., Cesa-Bianchi & Lugosi 2006, lemma 8.2), we obtain that the right-hand side of Equation 2 is equal to

$$\frac{L}{2}\mathbb{E}\left[\left|\sum_{t=1}^{T} \varepsilon_t \mathbf{z}_0^\top(\mathbf{v}_1 - \mathbf{v}_2)\right|\right] = \frac{LD}{2}\mathbb{E}\left[\left|\sum_{t=1}^{T}\varepsilon_t\right|\right] \ge \frac{LD}{4}\sqrt{2T}, \qquad 3.$$

[2]This can be a plausible assumption in many practical cases—for instance, in online linear regression when upper bounds on $\max_t \|\mathbf{x_t}\|_2$ and $\max_t |y_t|$ are known in advance.

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

where the equality is proven using $\mathbf{z}_0^\top(\mathbf{v}_1 - \mathbf{v}_2) = D$ due to our choice of $\mathbf{z}_0$. This shows that we cannot expect the regret to grow slower that $LD\sqrt{T}$—where D is the Euclidean diameter of $\mathbb{X}$ and L is the Lipschitz constant of the loss—unless the two main parameters of our setting, that is, the model space $\mathbb{X}$ and the loss process $\ell_1, \ell_2, \ldots$, enjoy some additional properties.

Lower bounds arguments based on the Khintchine inequality are rather common in online learning; Luo et al. (2016, theorem 1) provide an example close to the one presented here. The effectiveness of stochastic loss sequences to prove tight lower bounds in OCO settings in not accidental. To gain a better understanding of the connections between stochastic and online learning, readers are directed to the work of Rakhlin & Sridharan (A. Rakhlin & K. Sridharan, unpublished book draft), who study online learning as a minimax problem.

The game-theoretic roots of OCO are described by Cesa-Bianchi & Lugosi (2006). Since then, the interface between sequential optimization, game theory, and statistics has been intensively explored in many works, including the surveys by Shalev-Shwartz (2007), Hazan (2016), McMahan (2017), and Orabona (2019).

2. ONLINE MIRROR DESCENT

We now introduce the most popular algorithm for OCO, online mirror descent (OMD). OMD is the online version of the mirror descent algorithm of Nemirovsky & Yudin (1983) for convex optimization. Mirror descent is based on a generalization of projected gradient descent in which distances in the model space $\mathbb{X}$ are not necessarily measured using the Euclidean norm. This allows one to take advantage of specific geometrical properties that $\mathbb{X}$ may have. To see how this is done, we start from the iterates $\mathbf{w}_{t+1} = \Pi_{\mathbb{X}}(\mathbf{w}_t - \eta_t \nabla F(\mathbf{w}_t))$ of projected gradient descent on a convex and differentiable objective $F : \mathbb{R}^d \to \mathbb{R}$, where $\Pi_{\mathbb{X}}$ denotes the Euclidean projection onto $\mathbb{X}$ and $\eta_t > 0$ is the step size at time t. The expression defining the iterates can be rewritten in an equivalent optimization form,

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{X}} \ \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2 + \mathbf{w}^\top \nabla F(\mathbf{w}_t).$$

Mirror descent replaces the Euclidean norm in the above equation with a generalized distance or divergence D,

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{X}} \ \frac{1}{2\eta_t}\mathrm{D}(\mathbf{w}, \mathbf{w}_t) + \mathbf{w}^\top \nabla F(\mathbf{w}_t). \qquad 4.$$

Following Beck & Teboulle (2003), the divergences used by mirror descent are parameterized by mirror map functions $\psi : \mathbb{X} \to \mathbb{R}$ that are strictly convex and continuously differentiable on the interior of $\mathbb{X}$. Given such a ψ, the Bregman divergence $\mathrm{D}_\psi : \mathbb{X} \times \operatorname{int}\mathbb{X} \to \mathbb{R}$ is defined by

$$\mathrm{D}_\psi(\mathbf{u}, \mathbf{w}) = \psi(\mathbf{u}) - \psi(\mathbf{w}) - \nabla\psi(\mathbf{w})^\top(\mathbf{u} - \mathbf{w}).$$

Note that D_ψ is not necessarily symmetric and, since ψ is strictly convex, is always nonnegative and equals zero only when $\mathbf{w} = \mathbf{u}$. When ψ is also twice differentiable, then Taylor's theorem shows that

$$\mathrm{D}_\psi(\mathbf{u}, \mathbf{w}) = \frac{1}{2}(\mathbf{u} - \mathbf{w})^\top \nabla^2\psi(\mathbf{z})(\mathbf{u} - \mathbf{w})$$

for some $\mathbf{z}$ on the line segment joining $\mathbf{u}$ and $\mathbf{w}$. In other words, for ψ that are twice differentiable, the divergence locally behaves like a squared Mahalanobis distance. Just like in the Euclidean case, we can also write the mirror descent update (Equation 4) using a Bregman projection (see, e.g., Bubeck 2015, section 4).

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

The online version of mirror descent is now straightforward to obtain. In order to avoid considering iterates $\mathbf{w}_t$ on the boundary of $\mathbb{X}$, where $\mathrm{D}_\psi(\cdot, \mathbf{w}_t)$ is not defined, we restrict the argmin in Equation 4 to a convex and nonempty subset $\mathbb{V} \subseteq \operatorname{int} \mathbb{X}$ and, consequently, measure the regret $R_T(\mathbf{u})$ only for $\mathbf{u} \in \mathbb{V}$. Let $\mathbf{w}_1 \in \mathbb{V}$ and fix a sequence $\eta_1 \geq \eta_2 \geq \cdots > 0$ of step sizes. Now, for any sequence $\ell_1, \ell_2, \ldots$ of differentiable loss functions, the iterates of OMD are defined by

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{V}} \ \frac{1}{2\eta_t} \mathrm{D}_\psi(\mathbf{w}, \mathbf{w}_t) + \mathbf{w}^\top \mathbf{g}_t, \qquad 5.$$

where $\mathbf{g}_t = \nabla \ell_t(\mathbf{w}_t)$. The differentiability assumption for the losses ℓ_t can be relaxed to subdifferentiability, in which case $\mathbf{g}_t$ is any element of the subdifferential of ℓ_t at $\mathbf{w}_t$. This is useful because some popular loss functions, like the hinge loss $\ell_t(\mathbf{w}_t) = \max\{0, 1 - y_t \mathbf{w}^\top \mathbf{x}_t\}$ for binary classification ($y_t \in \{-1, 1\}$), are not everywhere differentiable. In the rest of this review, we use the same notation $\mathbf{g}_t$ to denote $\nabla \ell_t(\mathbf{w}_t)$ or any subgradient of ℓ_t at $\mathbf{w}_t$, according to whether ℓ_t is differentiable or only subdifferentiable.

One might wonder why we are using a linear approximation $\mathbf{w}^\top \mathbf{g}_t$ instead of the loss $\ell_t(\mathbf{w})$ in the update of OMD. Indeed, the variant where $\ell_t(\mathbf{w})$ replaces $\mathbf{w}^\top \mathbf{g}_t$ is called the proximal point method in the convex optimization literature and implicit update in the online learning literature (see Kivinen & Warmuth 1997, Kulis & Bartlett 2010, and also McMahan 2017, section 6). Connections between implicit updates and optimistic updates in saddle point optimization problems were investigated by Mokhtari et al. (2019) (see also Section 3). We now look at two important choices for the mirror map:

1. Euclidean: If $\psi = \frac{1}{2}\|\cdot\|_2^2$, then $\mathrm{D}_\psi(\mathbf{u}, \mathbf{w}) = \frac{1}{2}\|\mathbf{u} - \mathbf{w}\|_2^2$ and the OMD update becomes the online version of gradient descent (OGD) with Euclidean projection, $\mathbf{w}_{t+1} = \Pi_{\mathbb{V}}(\mathbf{w}_t - \eta_t \mathbf{g}_t)$.
2. Entropic: If $\mathbb{X}$ is the simplex of probability distributions over $\{1, \ldots, d\}$ and ψ is the negative entropy $\psi(\mathbf{w}) = \sum_i w_i \ln w_i$, then $\mathrm{D}_\psi(\mathbf{u}, \mathbf{w}) = \sum_i u_i \ln \frac{u_i}{w_i}$ is the Kullback–Leibler divergence (or cross entropy) and the OMD update becomes the exponentiated gradient (EG) algorithm of Kivinen & Warmuth (1997),

$$w_{t+1,i} = \frac{w_{t,i} e^{-\eta_t g_{t,i}}}{\sum_{j=1}^d w_{t,j} e^{-\eta_t g_{t,j}}} \qquad i = 1, \ldots, d,$$

   where $g_{t,i}$ are the components of the gradient $\mathbf{g}_t$. As in this case

$$\lim_{\mathbf{w} \to \operatorname{bd} \mathbb{X}} \left\| \nabla \psi(\mathbf{w}) \right\|_2 = \infty, \qquad 6.$$

   where $\operatorname{bd} \mathbb{X}$ is the boundary of $\mathbb{X}$, we can measure the regret $R_T(\mathbf{u})$ against $\mathbf{u} \in \mathbb{X}$ instead of restricting to $\mathbf{u} \in \mathbb{V} \subseteq \operatorname{int} \mathbb{X}$. In fact, under the latter condition, the update rule of OMD will never return a point on the boundary of $\mathbb{X}$. For instance, when losses are linear, $\sum_t \ell_t$ is always minimized at a $\mathbf{u}$ located on a corner of the simplex, and we can show that EG (OMD with entropic mirror map) has vanishing regret with respect to any $\mathbf{u}$ in the simplex $\mathbb{X}$, including the corners.

We now look at a different interpretation of the update rule of OMD. First, let a differentiable function $f : \mathbb{X} \to \mathbb{R}$ be μ-strongly convex on $\mathbb{V} \subseteq \operatorname{int} \mathbb{X}$ with respect to a norm $\|\cdot\|$ if, for all $\mathbf{u}, \mathbf{v} \in \mathbb{V}$, we have that $f(\mathbf{u}) \geq f(\mathbf{v}) + \nabla f(\mathbf{v})^\top (\mathbf{u} - \mathbf{v}) + \frac{\mu}{2}\|\mathbf{u} - \mathbf{v}\|^2$. Mirror maps ψ that are μ-strongly convex with respect to a norm $\|\cdot\|$ define Bregman divergences that grow faster than the square of the same norm, i.e., $\mathrm{D}_\psi(\mathbf{u}, \mathbf{w}) \geq \frac{\mu}{2}\|\mathbf{u} - \mathbf{w}\|^2$.

For strongly convex mirror maps, an equivalent way of writing Equation 5 is $\mathbf{w}_{t+1} = \nabla \psi_{\mathbb{V}}^\star (\nabla \psi_{\mathbb{V}}(\mathbf{w}_t) - \eta_t \mathbf{g}_t)$, where the function $\psi_{\mathbb{V}}$ is the restriction of ψ to $\mathbb{V}$, i.e., $\psi_{\mathbb{V}}(\mathbf{w}) = \psi(\mathbf{w})$

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

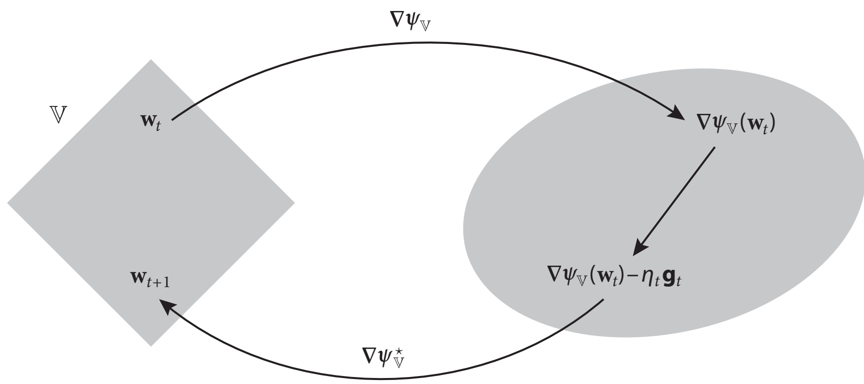

Figure 1

A step of online mirror descent (OMD) run with $\psi_{\mathbb{V}}$ (the restriction to the model space $\mathbb{V} \subseteq \mathbb{X}$ of a strongly convex mirror map ψ). The function $\psi^{\star}_{\mathbb{V}}$ is the Fenchel conjugate of $\psi_{\mathbb{V}}$. The vector $\mathbf{g}_t$ denotes the loss gradient $\nabla\ell_t(\mathbf{w}_t)$ (or any subgradient of ℓ_t at $\mathbf{w}_t$), and $\eta_t > 0$ is a variable step size.

if $\mathbf{w} \in \mathbb{V}$ and $\psi_{\mathbb{V}}(\mathbf{w}) = \infty$ otherwise. The function $\psi^{\star}_{\mathbb{V}} : \mathbb{R}^d \to \mathbb{R}$ is the Fenchel conjugate of $\psi_{\mathbb{V}}$, defined by

$$\psi^{\star}_{\mathbb{V}}(\boldsymbol{\theta}) = \max_{\mathbf{w}\in\mathbb{R}^d} \mathbf{w}^{\top}\boldsymbol{\theta} - \psi_{\mathbb{V}}(\mathbf{w}). \qquad 7.$$

Strong convexity ensures that $\psi^{\star}_{\mathbb{V}}$ is differentiable and that $\nabla\psi^{\star}_{\mathbb{V}}$ is the functional inverse of $\nabla\psi_{\mathbb{V}}$. The function $\psi_{\mathbb{V}}$ maps the iterates $\mathbf{w}_t$ to the dual space of gradients where a gradient step is performed. The inverse function $\psi^{\star}_{\mathbb{V}}$ maps back to the primal space of iterates. **Figure 1** illustrates the OMD update. Going back to the previous examples, we see that $\psi = \frac{1}{2}\left\|\cdot\right\|_2^2 = \psi^{\star}$. When ψ is the negative entropy, we instead have that $\psi^{\star}(\boldsymbol{\theta}) = \ln(\sum_{i=1}^{d} e^{\theta_i})$.

We now move on to analyze the regret of OMD and show us how the change of geometry caused by the choice of the strongly convex mirror map ψ affects the algorithm's performance. The first step in bounding the regret $R_T(\mathbf{u})$ consists in linearizing the loss using convexity, $\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u}) \le \mathbf{g}_t^{\top}(\mathbf{w}_t - \mathbf{u})$; recall that $\mathbf{g}_t = \nabla\ell_t(\mathbf{w}_t)$ for differentiable losses. We call the resulting upper bound on $R_T(\mathbf{u})$ linearized regret. Next, using the properties of the optimization form for the OMD iterates (Equation 4) and the μ-strong convexity of the mirror map with respect to $\|\cdot\|$, it takes a few steps to prove that

$$\eta_t\mathbf{g}_t^{\top}(\mathbf{w}_t - \mathbf{u}) \le \mathrm{D}_{\psi}(\mathbf{u}, \mathbf{w}_t) - \mathrm{D}_{\psi}(\mathbf{u}, \mathbf{w}_{t+1}) + \frac{\eta_t^2}{2\mu}\left\|\mathbf{g}_t\right\|_{\star}^2,$$

where $\|\cdot\|_{\star}$ is the dual norm of $\|\cdot\|$. Now, dividing both sides by $\eta_t > 0$ and summing over $t = 1, \ldots, T$ gives the following chain of inequalities:

$$\begin{aligned} R_T(\mathbf{u}) &= \sum_{t=1}^{T}\left(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u})\right) \\ &\le \sum_{t=1}^{T}\mathbf{g}_t^{\top}(\mathbf{w}_t - \mathbf{u}) && \text{(linearized regret)} \\ &\le \sum_{t=1}^{T}\left(\frac{\mathrm{D}_{\psi}(\mathbf{u}, \mathbf{w}_t)}{\eta_t} - \frac{\mathrm{D}_{\psi}(\mathbf{u}, \mathbf{w}_{t+1})}{\eta_t}\right) + \frac{1}{2\mu}\sum_{t=1}^{T}\eta_t\left\|\mathbf{g}_t\right\|_{\star}^2 \end{aligned}$$

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

$$= \frac{\mathrm{D}_\psi(\mathbf{u}, \mathbf{w}_1)}{\eta_1} - \frac{\mathrm{D}_\psi(\mathbf{u}, \mathbf{w}_{T+1})}{\eta_{T+1}} + \sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \mathrm{D}_\psi(\mathbf{u}, \mathbf{w}_{t+1}) + \frac{1}{2\mu} \sum_{t=1}^{T} \eta_t \left\| \mathbf{g}_t \right\|_\star^2$$

$$\leq \frac{D^2}{\eta_1} + \left(\frac{1}{\eta_T} - \frac{1}{\eta_1} \right) D^2 + \frac{1}{2\mu} \sum_{t=1}^{T} \eta_t \left\| \mathbf{g}_t \right\|_\star^2 \qquad \left[\text{where } D^2 = \max_{\mathbf{u},\mathbf{w} \in \mathbb{V}} \mathrm{D}_\psi(\mathbf{u}, \mathbf{w})\right]$$

$$= \frac{D^2}{\eta_T} + \frac{1}{2\mu} \sum_{t=1}^{T} \eta_t \left\| \mathbf{g}_t \right\|_\star^2 . \qquad 8.$$

We can now set $\eta_t = D\sqrt{\frac{\mu}{\sum_{s=1}^{t} \|\mathbf{g}_s\|_\star^2}}$ and obtain

$$R_T(\mathbf{u}) \leq 2D \sqrt{\frac{1}{\mu} \sum_{t=1}^{T} \left\| \mathbf{g}_t \right\|_\star^2}. \qquad 9.$$

Equipped with this result, we can see how the choice of the mirror map affects OMD performance. Consider first OGD, where $\mathbb{V}$ is the closed Euclidean ball of diameter U and $\psi = \frac{1}{2} \|\cdot\|_2^2$. If $\|\mathbf{g_t}\|_\infty \leq G$ for some $G > 0$, then $\left\|\mathbf{g}_t\right\|_\star^2 = \left\|\mathbf{g}_t\right\|_2^2 \leq G^2 d$ and OGD has a regret bound of the form $R_T(\mathbf{u}) \leq 2UG\sqrt{dT}$. Under these assumptions, losses have Lipschitz constant $L = \max_t \left\|\mathbf{g}_t\right\|_2 \leq G\sqrt{d}$, and so this regret bound matches the lower bound, Equation 3 in Section 1.3, up to constant factors. Therefore, in the Euclidean domain, OGD has essentially optimal dependence on time, diameter of the model space, and Lipschitz constant of the losses.

Next, consider EG, where $\mathbb{V}$ is the probability simplex and $\psi(\mathbf{w}) = \sum_i w_i \ln w_i$, which is 1-strongly convex with respect to $\|\cdot\|_1$. Here, we run into a problem because the diameter of the simplex is unbounded when measured using the cross entropy (i.e., the Bregman divergence D_ψ corresponding to the entropic mirror map). This prevents us from obtaining a constant upper bound D^2 on $\max_t \mathrm{D}_\psi(\mathbf{u}, \mathbf{w}_t)$. We can fix this by choosing a constant step size η, which allows us to transform the regret guarantee (Equation 8) into the following alternative bound:

$$R_T(\mathbf{u}) \leq \frac{\mathrm{D}_\psi(\mathbf{u}, \mathbf{w}_1)}{\eta} + \frac{\eta}{2\mu} \sum_{t=1}^{T} \left\| \mathbf{g}_t \right\|_\star^2 . \qquad 10.$$

Under the same assumption $\|\mathbf{g_t}\|_\infty \leq G$ as before, we get $\left\|\mathbf{g}_t\right\|_\star^2 = \left\|\mathbf{g}_t\right\|_\infty^2 \leq G^2$. Hence, choosing $\mathbf{w}_1 = (1/d, \dots, 1/d)$ as first iterate so that $\mathrm{D}_\psi(\mathbf{u}, \mathbf{w}_1) \leq \ln d$ and setting $\eta = \sqrt{(2 \ln d)/(G^2 T)}$ gives $R_T(\mathbf{u}) \leq G\sqrt{(T \ln d)/2}$. Using the squared Euclidean norm in the same setting would instead give $R_T(\mathbf{u}) \leq 2G\sqrt{dT}$. Note that for linear losses, the quantity $G\sqrt{(T \ln d)/2}$ is asymptotically minimax, including constants (see Cesa-Bianchi & Lugosi 2006, corollary 8.3). This shows the importance of matching the mirror map to the geometry of the model space. From a practical point of view, the logarithmic dependence on d in the regret guarantees that EG is robust to a large number of irrelevant features.

When the convex losses ℓ_t are induced by pairs $(\mathbf{x}_t, y_t)$, as in $\ell_t(\mathbf{w}) = (\mathbf{w}^\top \mathbf{x}_t - y_t)^2$, then loss gradients $\nabla \ell_t(\mathbf{w})$ are proportional to data points $\mathbf{x}_t$. In this case, bounds on $R_T(\mathbf{u})$ for OGD and EG depend on different products of dual norms: $(\max_t \|\mathbf{x_t}\|_2)\|\mathbf{u}\|_2$ for OGD and $(\max_t \|\mathbf{x_t}\|_\infty)\|\mathbf{u}\|_1$ for EG. Since $\|\cdot\|_\infty \leq \|\cdot\|_2 \leq \|\cdot\|_1$, neither bound dominates. However, for sparse $\mathbf{u}$, $\|\mathbf{u}\|_1$ approaches $\|\mathbf{u}\|_\infty$ and EG performs better than OGD.

The analysis of OGD can be easily adapted to derive a regret bound for the perceptron algorithm of Rosenblatt (1958) for binary classification, without the need to assume linear separability of the data sequence (see Freund & Schapire 1999). In the special case of separable data sequences, the regret bound reduces to the result originally proven in the well-known perceptron convergence theorem (Block 1962, Novikoff 1963).

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

If the loss functions are known to be μ-strongly convex and L-Lipschitz in $\mathbb{V}$, then the step size of OGD can be set more aggressively to $\eta_t = \frac{1}{\mu t}$ to exploit the curvature. The resulting bound on the regret is

$$R_T(\mathbf{u}) \leq \frac{L^2}{2\mu}\ln(T+1) \qquad \forall \mathbf{u} \in \mathbb{V}.$$

Note that there is no dependence on the diameter of the model space in this bound, only on the gradients of the losses (through the Lipschitz constant L). Many machine learning algorithms, including support vector machines (Cortes & Vapnik 1995), can be trained using stochastic optimization over a training set of data points $\mathbf{z}_1, \ldots, \mathbf{z}_m$ to minimize a strongly convex functional F (see Equation 1). A prime example of a stochastic gradient descent algorithm is OGD run on the sequence of strongly convex losses $\ell_t = F(\cdot, \mathbf{Z}_t)$, where $\mathbf{Z}_t$ is drawn at random from the data set. The regret analysis of OGD with strongly convex losses can be used to obtain rates of convergence to the minimum of F in Equation 1. This is done, for example, to analyze the Pegasos algorithm of Shalev-Shwartz et al. (2011).

The EG algorithm is an instance of the multiplicative update method, a technique that found applications in computer science (Littlestone & Warmuth 1994), game theory (Fudenberg & Levine 1995), information theory, statistics, and other disciplines (see, e.g., Cesa-Bianchi & Lugosi 2006, Arora et al. 2012b). In the special case of linear losses with uniformly bounded coefficients, EG is known as the hedge algorithm (Freund & Schapire 1997), and the corresponding setting is known as prediction with expert advice (Cesa-Bianchi et al. 1997).

2.1. The AdaGrad Algorithm

A variant of OMD that has become of widespread use as a stochastic gradient descent algorithm for training neural networks is the AdaGrad (adaptive gradient) algorithm, independently introduced by McMahan & Streeter (2010) and Duchi et al. (2011). For simplicity, we look at the so-called diagonal version of AdaGrad, which uses a coordinate-dependent step size.

Let $\mathbb{V}$ be the hyperrectangle $[a_1, b_1] \times \cdots \times [a_d, b_d]$ and $D_i = b_i - a_i$ for $i = 1, \ldots, d$. AdaGrad runs OMD with mirror map $\psi = \frac{1}{2}\|\cdot\|^2$ and projection onto $\mathbb{V}$. The iterates, including the projection step, can be written as $w_{t+1,i} = \max\{\min\{w_{t,i} - \eta_{t,i} g_{t,i}\}, a_i\}$ for $i = 1, \ldots, d$ and an arbitrary initial point $\mathbf{w}_1 \in \mathbb{V}$. The components of the step size are chosen as

$$\eta_{t,i} = \frac{D_i}{\sqrt{2\sum_{s=1}^t g_{s,i}^2}}.$$

The regret analysis is straightforward: After linearizing the losses so that $\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u}) \leq \mathbf{g}_t^\top (\mathbf{w}_t - \mathbf{u})$, one can simply perform the standard OMD analysis independently for each coordinate. By applying the bound (Equation 9) on each coordinate and then summing over coordinates, we obtain

$$R_T(\mathbf{u}) \leq \sum_{i=1}^d D_i \sqrt{2\sum_{t=1}^T g_{t,i}^2}. \qquad 11.$$

We can compare this bound to Equation 9 with $\mu = 1$ and $\|\cdot\|_\star = \|\cdot\|_2$. For simplicity, we take $\mathbb{V}$ to be the hypercube with $D_i = 1$, so that D in Equation 9 is equal to $\sqrt{d}$. Using Cauchy-Schwartz,

$$\sqrt{\sum_{t=1}^T \|\mathbf{g}_t\|_2^2} \leq \sum_{i=1}^d \sqrt{\sum_{t=1}^T g_{t,i}^2} \leq \sqrt{d}\sqrt{\sum_{t=1}^T \|\mathbf{g}_t\|_2^2},$$

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

where Equation 9 appears in the right-hand side and Equation 11 appears in the middle. Hence, ignoring the fact that the sequence of realized subgradients $\mathbf{g}_t$ is different for the two algorithms, with this choice of $\mathbb{V}$, AdaGrad can gain up to a factor of $\sqrt{d}$ in the regret bound with respect to plain OMD. Compared with the EG algorithm, which uses an entropic mirror map on the probability simplex, here the advantage is brought by a coordinate-dependent step size, which exploits the decomposability of the regret across the d coordinates granted by the geometry of $\mathbb{V}$.

Note that the specific choice of the step size makes the algorithm independent with respect to rescalings of the coordinates (Orabona & Pál 2018). This property is especially useful in neural network training, where the range of gradient components may vary a lot across the different layers.

3. FOLLOW THE REGULARIZED LEADER

A very natural online learning strategy for the OCO setting is follow the leader (FTL), which corresponds to predicting with the model minimizing the sum of the losses observed so far,

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{V}} \sum_{s=1}^{t} \ell_t(\mathbf{w}).$$

Here $\mathbf{w}_1 \in \mathbb{V}$ is any convex, closed, and nonemtpy subset of $\mathbb{R}^d$. When losses are strongly convex and Lipschitz in $\mathbb{V}$, FTL achieves $R_T(\mathbf{u}) = \mathcal{O}(\ln T)$ for all $\mathbf{u} \in \mathbb{V}$ (McMahan 2017, section 3.7). However, the curvature of each loss function is necessary to obtain a nontrivial performance. When losses are linear, FTL provably incurs a regret that grows linearly in T (see, e.g., Shalev-Shwartz 2012, example 2.2). This is caused by the intrinsic instability of the algorithm: In a nonstochastic setting, one can design the loss sequence so that the trajectory $\mathbf{w}_1, \mathbf{w}_2, \ldots$ of FTL models oscillates wildly in $\mathbb{V}$, a behavior that the adversary can exploit to increase the regret. Similarly to OMD, which achieves stability by forcing $\mathbf{w}_{t+1}$ to be not too far away from $\mathbf{w}_t$ (see Equation 5), FTL can be stabilized by adding a strictly convex regularization function. The resulting algorithm, appropriately called follow the regularized leader (FTRL), is a close relative of OMD. Indeed, the regularization functions used by FTRL are formally equivalent to OMD's mirror maps ψ. As FTRL is not formulated as a gradient descent method, we replace OMD's step sizes η_t with time-dependent regularizers ψ_t. For any sequence $\psi_1, \psi_2, \ldots$ of strictly convex regularizers, FTRL iterates are defined by

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{V}} \ \psi_{t+1}(\mathbf{w}) + \sum_{s=1}^{t} \ell_s(\mathbf{w}), \qquad 12.$$

where $\ell_1, \ell_2, \ldots$ is an arbitrary sequence of convex losses. From the viewpoint of regret minimization, we know that $\ell_s(\mathbf{w})$ can be replaced by $\mathbf{g}_s^\top \mathbf{w}$ (recall the linearization step in Section 2), where $\mathbf{g}_s$ is the gradient (or subgradient) of $\ell_s(\mathbf{w}_s)$. If we do that in Equation 12, we obtain

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{V}} \ \psi_{t+1}(\mathbf{w}) + \sum_{s=1}^{t} \mathbf{g}_s^\top \mathbf{w}. \qquad 13.$$

This is the version of FTRL we study in the rest of this section. Let $\psi_{\mathbb{V},t}$ be the restriction of ψ_t to $\mathbb{V}$ (see Section 2). If the regularizers $\psi_{\mathbb{V},t}$ are all strongly convex, then $\mathbf{w}_{t+1}$ in Equation 13 has the closed form

$$\mathbf{w}_{t+1} = \nabla \psi^\star_{\mathbb{V},t+1}\left(-\sum_{s=1}^{t} \mathbf{g}_s\right),$$

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

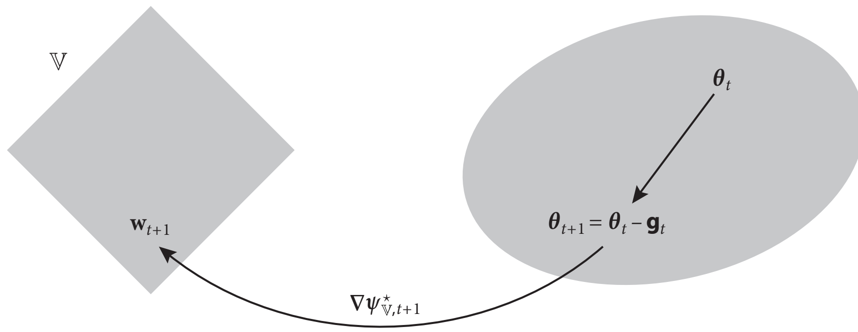

Figure 2

A step of follow the regularized leader (FTRL) run with linearized losses and time-dependent regularizers $\psi_{\mathbb{V},t}$ (these are the restrictions to the model space $\mathbb{V} \subseteq \mathbb{X}$ of strongly convex regularizers ψ_t). The functions $\psi^\star_{\mathbb{V},t}$ are the Fenchel conjugate of $\psi_{\mathbb{V},t}$. The vectors $\mathbf{g}_t$ denote the loss gradients $\nabla\ell_t(\mathbf{w}_t)$ (or any subgradient of ℓ_t at $\mathbf{w}_t$), while the state variables $\boldsymbol{\theta}_t$ are simply the sum of all past loss gradients, $\boldsymbol{\theta}_t = -(\mathbf{g}_1 + \cdots + \mathbf{g}_{t-1})$.

where $\psi^\star_{\mathbb{V},t+1}$ is the Fenchel conjugate of $\psi_{\mathbb{V},t+1}$ (differentiability of $\psi^\star_{\mathbb{V},t+1}$ is guaranteed by the strong convexity of $\psi_{\mathbb{V},t+1}$). By letting $\boldsymbol{\theta}_{t+1} = -(\mathbf{g}_1 + \cdots + \mathbf{g}_t)$, the FTRL iterates can be written as $\mathbf{w}_{t+1} = \nabla\psi^\star_{\mathbb{V},t+1}(\boldsymbol{\theta}_{t+1})$ (see **Figure 2**). By comparison, the OMD iterates for strongly convex mirror maps are written as $\mathbf{w}_{t+1} = \nabla\psi^\star_{\mathbb{V}}(\boldsymbol{\theta}'_{t+1})$, where $\boldsymbol{\theta}'_{t+1} = \nabla\psi_{\mathbb{V}}(\mathbf{w}_t) - \eta_t\mathbf{g}_t$.

Note that FTRL keeps a state variable $\boldsymbol{\theta}_t$ in the dual space of gradients. This is mapped to the primal space of iterates every time a prediction is needed. Instead, OMD keeps its state $\mathbf{w}_t$ in the primal space of iterates. This is then mapped to the dual space of gradients every time an update must be computed. Another difference is that gradients are all equally weighted in FTRL, whereas in OMD, each gradient $\mathbf{g}_t$ is weighted by a potentially different step size η_t.

FTRL was originally introduced by Abernethy et al. (2008), although the key ideas are contained in Shalev-Shwartz (2007). The version with linearized losses and $\psi_t = \psi = \frac{\eta}{2}\|\cdot\|_2^2$ was introduced by Zinkevich (2004) as a so-called lazy update variant of OMD.

In order to understand whether the two algorithms can produce the same sequence of iterates, we focus on the case when $\eta_t = \eta$ in OMD and $\psi_t = \psi/\eta$ in FTRL, for all t. Then we define the iterates of both OMD and FTRL through a common optimization problem $\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w}\in\mathbb{V}} F_{t+1}(\mathbf{w})$, where, for OMD and FTRL, respectively,

$$F_{t+1}(\mathbf{w}) = \mathrm{D}_\psi(\mathbf{w}, \mathbf{w}_t) + \eta\mathbf{g}_t^\top\mathbf{w}$$

$$F_{t+1}(\mathbf{w}) = \psi(\mathbf{w}) + \sum_{s=1}^{t}\mathbf{g}_s^\top\mathbf{w}.$$

When $\mathbb{V}$ is such that $\mathbf{w}_{t+1}$ satisfies $\nabla F_{t+1}(\mathbf{w}_{t+1}) = \mathbf{0}$ for both instances of F_{t+1}, a quick computation shows us that for both choices of F_{t+1}, the gradient vanishes when $\nabla\psi(\mathbf{w}_{t+1}) = -\eta\boldsymbol{\theta}_{t+1}$. Due to the strong convexity of ψ, this is equivalent to $\mathbf{w}_{t+1} = \nabla\psi^\star(-\eta\boldsymbol{\theta}_{t+1})$. Under these conditions, the two algorithms produce the same sequence of iterates. A concrete setting in which the two algorithms become identical is $\mathbb{V} \equiv \mathbb{R}^d$ and $\psi = \frac{1}{2}\|\cdot\|_2^2$.

FTRL enjoys a regret bound similar to the one we stated for OMD. Let $\psi : \mathbb{V} \to \mathbb{R}$ be a μ-strongly convex function with respect to a norm $\|\cdot\|$. Since FTRL is invariant to positive constants added to the regularizers, without loss of generality we may assume that $\min_{\mathbf{u}\in\mathbb{V}}\psi(\mathbf{u}) = 0$. For scaling factors $\beta_1 \geq \beta_2 \geq \cdots > 0$, let $\psi_t = \psi/\beta_t$. Then, for any sequence $\ell_1, \ell_2, \ldots$ of convex

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

losses, the regret of FTRL satisfies

$$R_T(\mathbf{u}) \le \frac{\psi(\mathbf{u})}{\beta_T} + \frac{1}{2\mu}\sum_{t=1}^{T}\beta_t \left\|\mathbf{g}_t\right\|_\star^2 \qquad \forall \mathbf{u} \in \mathbb{V}. \qquad 14.$$

This is very similar to the OMD bound (Equation 8). However, unlike OMD, where the step size η_t is used in the update $\mathbf{w}_t \to \mathbf{w}_{t+1}$, in the above formulation FTRL uses the scaling factor β_t in the update $\mathbf{w}_{t-1} \to \mathbf{w}_t$. In particular, in FTRL, β_t cannot depend on $\mathbf{g}_t$.

Given the similarities between OMD and FTRL, are there settings in which the latter should be preferred over the former? Consider, for instance, the analysis of the EG algorithm (i.e., OMD run with entropic mirror map and $\mathbb{V}$ set to the probability simplex). In this case, we cannot use OMD with a variable step size because of the unboundedness of the Bregman divergence. When FTRL is applied to the same setting, the regret bound specializes to

$$R_T(\mathbf{u}) \le \frac{\ln d}{\beta_T} + \frac{G^2}{2}\sum_{t=1}^{T}\beta_t,$$

where, we recall, G upper bounds $\|\mathbf{g_t}\|_\infty$ for all t. Taking $\beta_t = \sqrt{(\ln d)/(G^2 t)}$ then gives $R_T(\mathbf{u}) \le 2G\sqrt{T\ln d}$. Up to constants, this is the same as the bound we obtain for OMD when the step size is tuned with prior knowledge of the number of rounds T.

3.1. Online Newton Step

As mentioned at the beginning of this section, strongly convex losses are an easy case for online learning. A simple algorithm like FTL achieves logarithmic regret on any sequence of such losses whenever the strong convexity coefficients in the loss sequence are bounded away from zero. A natural question is then whether logarithmic regret is possible for convex loss functions that are not strongly convex but also not linear.

For any symmetric and positive semidefinite matrix M, introduce the seminorm $\|\cdot\|_M$ such that $\|\mathbf{w}\|_M^2 = \mathbf{w}^\top M\mathbf{w}$. As it turns out, the right curvature property sufficient to guarantee logarithmic regret is the following:

$$\ell_t(\mathbf{u}) \ge \ell_t(\mathbf{w}) + \mathbf{g}^\top(\mathbf{u}-\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{u}-\mathbf{w}\|_{\mathbf{g}\mathbf{g}^\top}^2 \qquad \mathbf{u},\mathbf{w} \in \mathbb{V} \qquad 15.$$

for some $\lambda > 0$, where $\mathbf{g} = \nabla\ell_t(\mathbf{w})$ (or any subgradient if ℓ_t is only subdifferentiable). In words, we require ℓ_t to be strongly convex only in the direction of its gradient (or in the direction of some of its subgradients). For example, the square loss $\ell_t(\mathbf{w}) = \frac{1}{2}(\mathbf{w}^\top\mathbf{x}_t - y_t)^2$ satisfies the property in Equation 15 for $\lambda \le \frac{1}{8C^2}$ whenever $|\mathbf{w}^\top\mathbf{x}_t|, |y_t| \le C$ (Hazan et al. 2007, lemma 3). Also, the logistic loss $\ell_t(\mathbf{w}) = \ln(1+\exp(-\mathbf{w}^\top\mathbf{x}_t))$ satisfies Equation 15 when $\mathbf{w}$ belongs to an Euclidean ball of fixed radius. More generally, any loss ℓ_t such that $e^{-\alpha\ell_t}$ is concave in $\mathbb{V}$ for some $\alpha > 0$ satisfies Equation 15 with $\lambda \le \frac{1}{2}\min\{\frac{1}{GD},\alpha\}$, where D is the Euclidean diameter of $\mathbb{V}$ and G is a bound on $\max_{\mathbf{w}\in\mathbb{V}}\left\|\nabla\ell_t(\mathbf{w})\right\|$. Such losses are said to be α-exp-concave in $\mathbb{V}$ (Kivinen & Warmuth 1999). If ℓ_t is twice differentiable, then α-exp-concavity in $\mathbb{V}$ is equivalent to $\nabla^2\ell_t(\mathbf{w}) - \alpha\nabla\ell_t(\mathbf{w})\nabla\ell_t(\mathbf{w})^\top$ being positive semidefinite for all $\mathbf{w} \in \mathbb{V}$.

Hazan et al. (2007, theorem 6) prove an $\mathcal{O}(d\ln T)$ regret bound on FTL for all sequences of loss functions ℓ_t satisfying Equation 15 for some $\lambda > 0$. They do so by introducing the quadratic approximation of each loss function $\widehat{\ell_t}(\mathbf{w}) = \ell_t(\mathbf{w}_t) + \mathbf{g}_t^\top(\mathbf{w}-\mathbf{w}_t) + \frac{\lambda}{2}\|\mathbf{w}-\mathbf{w}_t\|_{\mathbf{g}_t\mathbf{g}_t^\top}^2$, where $\mathbf{g}_t = \nabla\ell_t(\mathbf{w}_t)$ and $\mathbf{w}_1, \mathbf{w}_2, \ldots \in \mathbb{V}$ are defined by

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{V}} \sum_{s=1}^{t}\widehat{\ell_s}(\mathbf{w}) \qquad 16.$$

Review in Advance first posted on
November 9, 2020. (Changes may
still occur before final publication.)

($\mathbf{w}_1$ can be defined arbitrarily). Observe that Equation 16 is just FTL run on the losses $\widehat{\ell}_t$. Moreover, $\widehat{\ell}_t(\mathbf{w}_t) = \ell_t(\mathbf{w}_t)$, and because of Equation 15, $\widehat{\ell}_t(\mathbf{u}) \le \ell_t(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{V}$. This implies that the regret of FTL with respect to the original losses ℓ_t satisfies

$$R_T(\mathbf{u}) \le \sum_{t=1}^{T} \widehat{\ell}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \widehat{\ell}_t(\mathbf{u}) \qquad \forall \mathbf{u} \in \mathbb{V}.$$

The rest of the proof uses the special properties of the functions $\widehat{\ell}_t$ to derive the desired $\mathcal{O}(d \ln T)$ bound on the regret. Hazan et al. (2007, lemma 4) also prove that Equation 16 is equivalent to

$$\begin{aligned} \mathbf{w}'_{t+1} &= S_t^+ \sum_{s=1}^{t} \left(\mathbf{g}_s^\top \mathbf{w}_s - \frac{1}{\lambda} \right) \mathbf{g}_s, \\ \mathbf{w}_{t+1} &= \operatorname*{argmin}_{\mathbf{w} \in \mathbb{V}} \left\| \mathbf{w} - \mathbf{w}'_{t+1} \right\|_{S_t}, \end{aligned} \qquad 17.$$

where S_t^+ denotes the Moore-Penrose pseudoinverse of $S_t = \mathbf{g}_1 \mathbf{g}_1^\top + \cdots + \mathbf{g}_t \mathbf{g}_t^\top$.

A similar $\mathcal{O}(d \ln T)$ bound can be proven through a more general proof, this time using the FTRL framework with linearized losses. Define the regularizers $\psi_1, \psi_2, \ldots$ given by

$$\psi_1(\mathbf{w}) = \frac{\|\mathbf{w}\|_2^2}{2} \qquad \text{and} \qquad \psi_{t+1}(\mathbf{w}) = \psi_t(\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}_t - \mathbf{w}\|^2_{\mathbf{g}_t \mathbf{g}_t^\top}, \qquad 18.$$

where $\mathbf{g}_t = \nabla \ell_t(\mathbf{w}_t)$ and the $\mathbf{w}_t$ are the FTRL iterates defined in Equation 13,

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{V}} \ \psi_{t+1}(\mathbf{w}) + \sum_{s=1}^{t} \mathbf{g}_s^\top \mathbf{w}. \qquad 19.$$

An equivalent and more explicit form for the iterates defined in Equation 19, which brings out the similarity with Equation 17, is

$$\begin{aligned} \mathbf{w}'_{t+1} &= A_t^{-1} \sum_{s=1}^{t} (\mathbf{g}_s^\top \mathbf{w}_s - 1) \mathbf{g}_s, \\ \mathbf{w}_{t+1} &= \operatorname*{argmin}_{\mathbf{w} \in \mathbb{V}} \left\| \mathbf{w} - \mathbf{w}'_{t+1} \right\|_{A_t}, \end{aligned} \qquad 20.$$

where $A_0 = I$ and $A_t = A_{t-1} + \lambda \mathbf{g}_t \mathbf{g}_t^\top$.

Each regularizer ψ_t of the form in Equation 18 is 1-strongly convex with respect to the norm $\|\cdot\|_{A_{t-1}}$. Using this property and the special recursive form of these regularizers, one can prove for this algorithm a regret bound of the following form: for any sequence $\ell_1, \ell_2, \ldots$ of losses satisfying Equation 15,

$$R_T(\mathbf{u}) \le \frac{\|\mathbf{u}\|_2^2}{2} + \frac{1}{2} \sum_{t=1}^{T} \left\| \mathbf{g}_t \right\|^2_{A_t^{-1}} \le \frac{\|\mathbf{u}\|_2^2}{2} + \frac{d}{2\lambda} \ln \left(1 + \frac{\lambda G T}{d} \right), \qquad 21.$$

where $\max_t \|\mathbf{g_t}\|_2 \le G$ and we used a standard majorization for bounding the sum of terms $\left\| \mathbf{g}_t \right\|^2_{A_t^{-1}}$ (see, e.g., Cesa-Bianchi & Lugosi 2006, lemma 11.11 and theorem 11.7).

One may also wonder whether the same logarithmic regret bound for losses satisfying Equation 15 could be achieved using OMD instead of FTL/FTRL as we did in this section. Noting that, for M symmetric and positive definite, the divergence associated with $\psi = \frac{1}{2} \|\cdot\|_M^2$ is $\mathrm{D}(\mathbf{u}, \mathbf{w}) = \frac{1}{2} \|\mathbf{u} - \mathbf{w}\|_M^2$, we may introduce the following instance of OMD:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{V}} \ \frac{1}{2} \|\mathbf{w} - \mathbf{w}_t\|_{A_t}^2 + \mathbf{w}^\top \mathbf{g}_t,$$

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

where $A_0 = I, A_t = A_{t-1} + \lambda \mathbf{g}_t \mathbf{g}_t^\top$, and $\mathbf{g}_t = \nabla \ell(\mathbf{w}_t)$. The closed-form expression is

$$\begin{aligned} \mathbf{w}'_{t+1} &= A_t^{-1} \mathbf{g}_t, \\ \mathbf{w}_{t+1} &= \operatorname*{argmin}_{\mathbf{w} \in \mathbb{V}} \left\| \mathbf{w} - \mathbf{w}'_{t+1} \right\|_{A_t}. \end{aligned} \qquad 22.$$

Note that this algorithm is rather different from both Equation 17 and Equation 20. Its regret analysis is relatively simple, although it is not derived as a special case of the general OMD analysis (which uses constant mirror maps). Fix any sequence $\ell_1, \ell_2, \ldots$ of loss functions satisfying Equation 15. Since $\mathbf{w}_{t+1}$ is the projection of $\mathbf{w}'_{t+1}$ onto $\mathbb{V}$, the update in Equation 22 ensures that

$$\|\mathbf{w}_{t+1} - \mathbf{u}\|_{A_t}^2 \le \left\|\mathbf{w}'_{t+1} - \mathbf{u}\right\|_{A_t}^2 = \|\mathbf{w}_t - \mathbf{u}\|_{A_t}^2 + \mathbf{g}_t^\top A_t^{-1} \mathbf{g}_t - 2\mathbf{g}_t^\top (\mathbf{w}_t - \mathbf{u}) \qquad 23.$$

for all $\mathbf{u} \in \mathbb{V}$. By the curvature property (Equation 15), we then have that

$$\begin{aligned} 2R_T(\mathbf{u}) &\le 2 \sum_{t=1}^{T} \left(\mathbf{g}_t^\top (\mathbf{w}_t - \mathbf{u}) - \lambda \|\mathbf{w}_t - \mathbf{u}\|_{\mathbf{g}_t \mathbf{g}_t^\top}^2 \right) \\ &\le \sum_{t=1}^{T} \left(\|\mathbf{g}_t\|_{A_t^{-1}} + \|\mathbf{w}_t - \mathbf{u}\|_{A_t}^2 - \|\mathbf{w}_{t+1} - \mathbf{u}\|_{A_t}^2 - \lambda \|\mathbf{w}_t - \mathbf{u}\|_{\mathbf{g}_t \mathbf{g}_t^\top}^2 \right) \quad \text{(using Equation 23)} \\ &= \sum_{t=1}^{T} \|\mathbf{g}_t\|_{A_t^{-1}}^2 + \|\mathbf{u}\|_{A_0}^2 + \sum_{t=1}^{T} (\mathbf{w}_t - \mathbf{u})^\top (A_t - A_{t-1} - \lambda \mathbf{g}_t \mathbf{g}_t^\top)(\mathbf{w}_t - \mathbf{u}) \\ &= \sum_{t=1}^{T} \|\mathbf{g}_t\|_{A_t^{-1}}^2 + \|\mathbf{u}\|_2^2, \end{aligned}$$

using the definition of A_t. Now note that the above is exactly equivalent to the bound in Equation 21, which we proved for FTRL.

3.2. Online Linear Regression

An important special case of online learning is (unconstrained) online linear regression, where $\mathbb{V} = \mathbb{R}^d$ and the losses $\ell_t(\mathbf{w}) = \frac{1}{2}\left(\mathbf{w}^\top \mathbf{x}_t - y_t\right)^2$ are induced by an arbitrary and deterministic sequence $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots$ of data points $\mathbf{x}_t \in \mathbb{R}^d$ and values $y_t \in \mathbb{R}$. The online version of the classical ridge regression algorithm by Hoerl & Kennard (2000) is an instance of FTRL without linearized losses (see Equation 12):

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^d} \ \frac{1}{2} \|\mathbf{w}\|_2^2 + \frac{1}{2} \sum_{s=1}^{t} \left(\mathbf{w}^\top \mathbf{x}_s - y_s\right)^2. \qquad 24.$$

Letting $A_t = I + \sum_{s=1}^{t} \mathbf{x}_s \mathbf{x}_s^\top$, where I is the $d \times d$ identity matrix, the ridge regression iterates (Equation 24) can be written in closed form as $\mathbf{w}_{t+1} = A_t^{-1}(y_1 \mathbf{x}_1 + \cdots + y_t \mathbf{x}_t)$. Online ridge regression enjoys the following regret bound (Cesa-Bianchi & Lugosi 2006, theorem 11.7):

$$R_T(\mathbf{u}) \le \frac{\|\mathbf{u}\|^2}{2} + d \ln\left(1 + \frac{T}{d}\left(\max_{t=1,\ldots,T} \|\mathbf{x}_t\|_2^2\right)\right)\left(\max_{t=1,\ldots,T} \ell_t(\mathbf{w}_t)\right) \qquad \forall \mathbf{u} \in \mathbb{R}^d.$$

This result shows a regret potentially logarithmic in time, except for the extraneous quantity $\max_t \ell_t(\mathbf{w}_t)$ in place of the correct scaling factor $\max_t y_t^2$. If we knew an upper bound Y on $\max_t |y_t|$, then we could obtain the desired scaling by clipping predictions $\mathbf{w}_t^\top \mathbf{x}_t$ in the interval $[-Y, Y]$ (Vovk 2001, theorem 4). Luckily, there is a better fix, which does not require any preliminary knowledge about $\max_t |y_t|$: Simply add to the objective function in Equation 24 an extra loss term associated

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

with the data point $\mathbf{x}_{t+1}$ and the value $y_{t+1} = 0$. As we see next, this has the effect of shrinking (toward zero) the linear predictions $\mathbf{w}_t^\top \mathbf{x}_t$, thus adding stability to the algorithm. Cesa-Bianchi & Lugosi (2006) call the resulting algorithm VAW, after Vovk (2001) and Azoury & Warmuth (2001), who independently introduced it. VAW iterates are defined by

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{1}{2}\sum_{s=1}^{t}\left(\mathbf{w}^\top\mathbf{x}_s - y_s\right)^2 + \frac{1}{2}\left(\mathbf{w}^\top\mathbf{x}_{t+1}\right)^2. \qquad 25.$$

The closed form expression is simply $\mathbf{w}_{t+1} = A_{t+1}^{-1}(y_1\mathbf{x}_1 + \cdots + y_t\mathbf{x}_t)$. In order to appreciate how predictions are shrunk by the addition of the term $\frac{1}{2}(\mathbf{w}^\top\mathbf{x}_{t+1})^2$, let $\widehat{y}_{t+1}^{\mathrm{RR}} = \mathbf{w}_{t+1}^\top\mathbf{x}_{t+1}$ be the ridge regression prediction computed via Equation 24. Then, the VAW prediction $\widehat{y}_{t+1}^{\mathrm{VAW}} = \mathbf{w}_{t+1}^\top\mathbf{x}_{t+1}$ computed via Equation 25 satisfies

$$\widehat{y}_{t+1}^{\mathrm{VAW}} = \frac{\widehat{y}_{t+1}^{\mathrm{RR}}}{1 + \mathbf{x}_{t+1}^\top A_t^{-1}\mathbf{x}_{t+1}}.$$

Clearly, $|\widehat{y}_{t+1}^{\mathrm{VAW}}| < |\widehat{y}^{\mathrm{RR}}|$ whenever $\|\mathbf{x}_{\mathbf{t+1}}\|_2 > 0$. It is interesting to compare VAW to a Gaussian process (GP) for regression (Rasmussen & Williams 2005). It is known that GP predicts with $\widehat{y}_{t+1}^{\mathrm{RR}}$; however, it also returns an estimate of the variance of the prediction that depends on $\mathbf{x}_{t+1}^\top A_t^{-1}\mathbf{x}_{t+1}$. Hence, VAW shrinks more the points that are assigned a high variance estimate by a GP. The regret analysis of WAV shows that

$$R_T(\mathbf{u}) \le \frac{\|\mathbf{u}\|^2}{2} + \left(\max_{t=1,\dots,T} y_t^2\right)\frac{d}{2}\ln\left(1 + \frac{T}{d}\left(\max_{t=1,\dots,T}\|\mathbf{x}_t\|_2^2\right)\right) \qquad \forall \mathbf{u}\in\mathbb{R}^d. \qquad 26.$$

Note that this bound simultaneously holds for any sequence $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots \in \mathbb{R}^d \times \mathbb{R}$, for any time horizon T, and for any linear comparator $\mathbf{u} \in \mathbb{R}^d$. Moreover, as proven in Vovk (2001, section 3.3), the bound in Equation 26 is asymptotically optimal, including the leading constant.

3.3. Optimistic Updates

The iterates of VAW are computed by adding to the expression in the argmin an extra loss term $\frac{1}{2}(\mathbf{w}^\top\mathbf{x}_{t+1})^2$ that predicts the next label y_{t+1} to be zero. This idea can be generalized to what is known in the literature as an optimistic update (Chiang et al. 2012, Rakhlin & Sridharan 2013). Given any sequence $\ell_1, \ell_2, \ldots$ of convex losses, FTRL with linearized losses and optimistic updates uses the iterates

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{V}} \psi_{t+1}(\mathbf{w}) + \sum_{s=1}^{t}\mathbf{g}_s^\top\mathbf{w} + \widehat{\mathbf{g}}_{t+1}^\top\mathbf{w},$$

where $\mathbf{g}_s = \nabla\ell_s(\mathbf{w}_s)$, $\widehat{\mathbf{g}}_{t+1}$ is a guess for $\nabla\ell_{t+1}(\mathbf{w}_{t+1})$, and $\psi_1, \psi_2, \ldots$ are arbitrary μ-strongly convex regularizers with respect to a norm $\|\cdot\|$. One can show that, for all $\mathbf{u} \in \mathbb{V}$,

$$R_T(\mathbf{u}) \le \psi_{T+1}(\mathbf{u}) - \psi_1(\mathbf{w}_1) + \frac{1}{2\mu}\sum_{t=1}^{T}\Big(\left\|\mathbf{g}_t - \widehat{\mathbf{g}}_t\right\|_\star + \psi_t(\mathbf{w}_{t+1}) - \psi_{t+1}(\mathbf{w}_{t+1})\Big). \qquad 27.$$

We now describe two concrete and simple examples. First, consider $\widehat{\mathbf{g}}_{t+1} = \frac{1}{t}\sum_{s=1}^{t}\mathbf{g}_s$ with $\widehat{\mathbf{g}}_1 = \mathbf{0}$ and—for the sake of simplicity—assume the convex losses are 1-Lipschitz in $\mathbb{V}$, so that $\|\mathbf{g}_\mathbf{t}\|_2 \le 1$ for all $\mathbf{g}_t$ generated in Equation 27. Take $\psi_t = \frac{\mu}{2}\|\cdot\|_2^2$ for all t. Then, for all $\mathbf{u} \in \mathbb{V}$,

$$R_T(\mathbf{u}) \le \frac{\mu}{2}\|\mathbf{u}\|_2^2 + \frac{1}{2\mu}\sum_{t=1}^{T}\left\|\mathbf{g}_t - \widehat{\mathbf{g}}_t\right\|_2^2.$$

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

Now note that $\ell_t = \frac{1}{2}\left\|\mathbf{g}_t - \cdot\right\|_2^2$ is a 1-strongly convex loss, and $\widehat{\mathbf{g}}_1, \widehat{\mathbf{g}}_2, \ldots$ correspond to the predictions of FTL on the sequence $\ell_1, \ell_2, \ldots$ of such losses. Therefore, using the FTL analysis for strongly convex losses mentioned at the beginning of this section, we have that

$$\sum_{t=1}^{T}\left\|\mathbf{g}_t - \widehat{\mathbf{g}}_t\right\|_2^2 - \min_{\mathbf{g}\,:\,\|\mathbf{g}\|_2\le 1}\sum_{t=1}^{T}\left\|\mathbf{g}_t - \mathbf{g}\right\|_2^2 = \mathcal{O}(\ln T).$$

It is easy to see that the minimizer in the above expression is $\overline{\mathbf{g}} = \frac{1}{T}\sum_t \mathbf{g}_t$. So, we conclude

$$R_T(\mathbf{u}) \le \frac{\mu}{2}\|\mathbf{u}\|_2^2 + \frac{1}{2\mu}\sum_{t=1}^{T}\left\|\mathbf{g}_t - \overline{\mathbf{g}}\right\|_2^2 + \mathcal{O}(\ln T).$$

Hence, the regret is bounded in terms of the cumulative empirical variance of the loss gradients. Using more sophisticated time-dependent regularizers, the regret can be bounded by an expression sublinear in the cumulative variance. We mention a similar result in the next example, where we choose $\widehat{\mathbf{g}}_{t+1} = \mathbf{g}_t$. In other words, we guess the next loss gradient to be similar to the current one. This is only beneficial under the additional assumption that losses are smooth, which is equivalent to assuming that gradients are Lipschitz. Then, choosing regularizers $\psi_t = \beta_t\psi$, where ψ is a base 1-strongly convex regularizer with respect to a norm $\|\cdot\|$ and $\beta_t > 0$ is a scaling factor, one can prove that the regret is bounded by an expression of the order of (see Orabona 2019)

$$\sqrt{1 + \sum_{t=2}^{T}\left\|\nabla\ell_t(\mathbf{w}_{t-1}) - \nabla\ell_{t-1}(\mathbf{w}_{t-1})\right\|_\star^2}.$$

Bounds of this form were first proven by Chiang et al. (2012). Optimistic updates have been also applied to show fast rates for regret minimization problems in game theory by Syrgkanis et al. (2015) and Foster et al. (2016).

4. UNCONSTRAINED ONLINE CONVEX OPTIMIZATION

As discussed in previous sections, both OMD and FTRL enjoy a regret of order $\mathcal{O}(D\sqrt{T})$ for convex and Lipschitz losses, where D bounds the diameter of $\mathbb{V}$ according to the divergence D_ψ (for OMD) or to the range of the base regularizer Ψ (for FTRL) (see Equations 8 and 14). If $\mathbb{V}$ is unbounded, say $\mathbb{V} = \mathbb{R}^d$, and losses are convex Lipschitz, we can still run OGD with $\mathbf{w}_1 = \mathbf{0}$ and fixed step size $\eta = \alpha/\sqrt{T}$ for $\alpha > 0$. Using Equation 10, we get the following upper bound on the regret:

$$R_T(\mathbf{u}) \le \frac{1}{2}\left(\frac{\|\mathbf{u}\|_2^2}{\alpha} + \alpha\right)\sqrt{T} \qquad \forall \mathbf{u} \in \mathbb{R}^d. \tag{28}$$

By tuning α optimally with respect to $\|\mathbf{u}\|_2^2$, we could get the bound $R_T(\mathbf{u}) \le \|\mathbf{u}\|_2\sqrt{T}$. This is equivalent to what we would get by running OGD with projection in the Euclidean ball of radius $U = \|\mathbf{u}\|_2$. However, this bound cannot be simultaneously achieved for all $\mathbf{u}$ because the algorithms must be run with only one choice of their parameters. The problem we study in this section is whether we can get a bound better than Equation 28 in the unconstrained setting. In order to answer this question, we explore the connection between OCO and the problem of sequentially betting on the elements of a deterministic sequence of values.

The fact that certain instances of online learning could be phrased as a gambling problem has been known since the works of Kelly (1956), Cover (1974), and Feder (1991). It was only recently,

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

however, that Orabona & Pál (2016) realized that certain betting strategies could be used to derive OCO algorithms with no parameters to tune, which could be also used for solving unconstrained OCO in an optimal way.

The betting game we are interested in is parameterized by an unknown deterministic sequence $x_1, x_2, \ldots \in [-1, 1]$ of real numbers. The bettor starts out with an initial wealth of $C_0 = \varepsilon > 0$. In each round $t = 1, 2, \ldots$ of the game:

1. The bettor bets $\alpha_t \in [-1, 1]$, whose absolute value is the fraction of their current wealth C_{t-1} they are betting, and whose sign indicates which sign we bet x_t will have.
2. The next value $x_t \in [-1, 1]$ is revealed.
3. The bettor's return is $x_t \times \alpha_t C_{t-1} \in \mathbb{R}$.

Note that the bettor's wealth changes in each step t according to $C_t = (1 + \alpha_t x_t)C_{t-1}$.

In order to show how to apply a betting strategy to any OCO problems, we first consider the 1-dimensional case with losses $\ell_1, \ell_2, \ldots$ defined on $\mathbb{R}$ and having uniformly bounded derivatives $|\ell_t'| \le 1$ for all t. First of all, recall that, because of the convexity of losses, for any $u \in \mathbb{R}$ the regret $R_T(u)$ of any online algorithm predicting with $w_1, w_2, \ldots \in \mathbb{R}$ is upper bounded by $\sum_{t=1}^T (w_t - u)g_t$, where g_t is the derivative of ℓ_t at w_t. In order to use a betting strategy to solve the online problem, we set $w_t = \alpha_t C_{t-1}$. We then have $C_t = C_{t-1} + w_t x_t$, which implies $C_T = \varepsilon + \sum_{t=1}^T w_t x_t$. If we now set $x_t = -g_t \in [-1, 1]$ (by our assumption on the losses), we get

$$C_T = \varepsilon \prod_{t=1}^T (1 + \alpha_t x_t) = \varepsilon + \sum_{t=1}^T w_t x_t = \varepsilon - \sum_{t=1}^T w_t g_t \quad \forall x_1, \ldots, x_T. \tag{29}$$

Next, we prove that a lower bound on the wealth C_T can be used to upper bound $R_T(u)$ for all $u \in \mathbb{R}$. In particular, assume that we have a betting strategy $\alpha_1, \alpha_2, \ldots \in \mathbb{R}$ such that

$$C_T \ge \phi\left(\sum_{t=1}^T x_t\right) = \phi\left(-\sum_{t=1}^T g_t\right) \tag{30}$$

for some convex real function ϕ. Then, for any $u \in \mathbb{R}$ we have

$$\begin{aligned}
R_T(u) &\le \sum_{t=1}^T (w_t - u)g_t \\
&= \sum_{t=1}^T -ug_t - \left(\varepsilon - \sum_{t=1}^T w_t g_t\right) + \varepsilon \\
&= \sum_{t=1}^T -ug_t - C_T + \varepsilon && \text{(using Equation 29)} \\
&\le \sum_{t=1}^T -ug_t - \phi\left(-\sum_{t=1}^T g_t\right) + \varepsilon && \text{(by our assumption on } C_T\text{)} \\
&\le \sup_{\theta \in \mathbb{R}} \theta u - \phi(\theta) + \varepsilon \\
&= \phi^\star(u) + \varepsilon,
\end{aligned}$$

using the definition of Fenchel conjugate, Equation 7. Now that we have established a formal connection between online prediction (in the one-dimensional case) and betting, we must pick a good betting strategy, one that has a small regret in the betting game. We measure this regret

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

against the class of strategies that always bet the best constant $\alpha \in [-1, 1]$. After T steps, the final wealth of each such strategy is $C_T(\alpha) = \varepsilon \prod_{t=1}^{T}(1 + \alpha x_t)$. Since wealth changes multiplicatively, it is natural to define the regret of a betting strategy $\alpha_1, \alpha_2, \dots$ using the logarithm of the wealth,

$$R_T(\alpha) = \ln\left(\varepsilon \prod_{t=1}^{T}(1 + \alpha x_t)\right) - \ln\left(\varepsilon \prod_{t=1}^{T}(1 + \alpha_t x_t)\right) = \sum_{t=1}^{T} \ln(1 + \alpha x_t) - \sum_{t=1}^{T} \ln(1 + \alpha_t x_t).$$

In the special case of Boolean values $x_t \in \{-1, 1\}$, the optimal regret $R_T(\alpha)$ was very precisely determined (including constants) by Shtarkov (1987), who established that

$$R_T(\alpha) = \frac{1}{2}\ln T + \frac{1}{2}\ln\frac{\pi}{2} + o(1) \qquad \forall \alpha \in [-1, 1].$$

Unfortunately, Shtarkov's strategy is not efficiently computable. A more efficient (and nearly optimal) strategy was introduced a bit earlier by Krichevsky & Trofimov (1981). The Krichevsky–Trofimov (KT) strategy has a regret of

$$R_T(\alpha) = \frac{1}{2}\ln T + \Theta(1) \qquad \forall \alpha \in [-1, 1],$$

which is optimal only up to constants. When $x_t \in [-1, 1]$, the KT strategy takes the simple form $\alpha_1 = 0$ and $\alpha_t = (x_1 + \cdots + x_{t-1})/t$. Orabona & Pál (2016) proved that the wealth of the KT strategy satisfies

$$C_T \geq \frac{\varepsilon}{c\sqrt{T}} \exp\left(\frac{1}{2T}\left(\sum_{t=1}^{T} x_t\right)^2\right)$$

for some universal constant $c > 0$.

Going back to the one-dimensional OCO case and recalling Equation 29 and $x_t = -g_t \in [-1, 1]$, we see that the KT strategy generates predictions of the form

$$w_t = \alpha_t C_{t-1} = -\left(\frac{1}{t}\sum_{s=1}^{t-1} g_s\right)\left(\varepsilon - \sum_{s=1}^{t-1} g_s w_s\right). \tag{31}$$

Using Equation 30 and our bound $R_T(u) \leq \phi^\star(u) + \varepsilon$ on the regret, we can compute $\phi^\star$ for our case and derive the bound

$$R_T(u) \leq |u|\sqrt{T \ln(u^2 c^2 T + 1)} + 1 \qquad \forall u \in \mathbb{R}, \tag{32}$$

where, for simplicity, we set $\varepsilon = 1$. Recall that this bound holds, simultaneously for all $u \in \mathbb{R}$, under the condition that $|\nabla \ell_t| \leq 1$ for all t. Under the same conditions, OGD guarantees a bound on the regret $R_T(u)$ of worse order $u^2\sqrt{T}$ (see Equation 28).

This result has been extended to $\mathbb{R}^d$ (and more generally to Banach spaces) by Cutkosky & Orabona (2018) using a simple trick: In order to bound the regret $R_T(\mathbf{u})$ against any $\mathbf{u} \in \mathbb{R}^d$ we use the one-dimensional KT strategy to learn the length $\|\mathbf{u}\|$ of $\mathbf{u}$ and OMD with model space $\mathbb{V} \equiv \left\{\mathbf{u} \in \mathbb{R}^d \,:\, \|\mathbf{u}\| \leq 1\right\}$ to learn the direction $\mathbf{u}/\|\mathbf{u}\|$ of $\mathbf{u}$, where $\|\cdot\|$ is any desired norm. This is done as follows: Let w_t be the prediction (Equation 31) of KT and let $\mathbf{v}_t$ be the prediction of OMD. Then, the combined algorithm predicts $w_t\mathbf{v}_t$ and, upon receiving the gradient $\mathbf{g}_t = \nabla \ell_t(w_t\mathbf{v}_t)$, feeds the derivative $\mathbf{g}_t^\top \mathbf{v}_t$ to KT and the gradient $\mathbf{g}_t$ to OMD.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

To analyze the regret of this combined KT-OMD strategy, note that $\|\mathbf{v_t}\| \le 1$ for all t because OMD projects onto $\mathbb{V}$. Therefore, $|\mathbf{g}_t^\top \mathbf{v}_t| \le \|\mathbf{g}\|_\star$. So, by linearizing the regret, we can write

$$\begin{aligned}
R_T(\mathbf{u}) &= \sum_{t=1}^T \ell_t(w_t\mathbf{v}_t) - \sum_{t=1}^T \ell_t(\mathbf{u}) \\
&\le \sum_{t=1}^T \mathbf{g}_t^\top (w_t\mathbf{v}_t - \mathbf{u}) \\
&= \sum_{t=1}^T \left(w_t\mathbf{g}_t^\top \mathbf{v}_t - \|\mathbf{u}\|\,\mathbf{g}_t^\top \mathbf{v}_t\right) + \sum_{t=1}^T \left(\|\mathbf{u}\|\,\mathbf{g}_t^\top \mathbf{v}_t - \mathbf{g}_t^\top \mathbf{u}\right) \\
&= \sum_{t=1}^T \left(w_t\mathbf{g}_t^\top \mathbf{v}_t - \|\mathbf{u}\|\,\mathbf{g}_t^\top \mathbf{v}_t\right) + \|\mathbf{u}\| \sum_{t=1}^T \left(\mathbf{g}_t^\top \mathbf{v}_t - \mathbf{g}_t^\top \frac{\mathbf{u}}{\|\mathbf{u}\|}\right).
\end{aligned}$$

Now note that the first sum is the linearized regret of KT against $\|\mathbf{u}\|$, where $\mathbf{g}_t^\top \mathbf{v}_t$ are the loss derivatives. The second sum (ignoring the $\|\mathbf{u}\|$ factor) is instead the linearized regret of OMD against $\mathbf{u}/\|\mathbf{u}\|$ with loss gradients $\mathbf{g}_t$.

Since KT requires the loss derivatives to belong to $[-1, 1]$, we can apply the above bound when $|\mathbf{g}_t^\top \mathbf{v}_t| \in [0, 1]$, which holds when ℓ_t is 1-Lipschitz for all t. Assuming that, and in comparison with Equation 28, we consider OGD run on the unit Euclidean ball with step size $\eta_t = 1/\sqrt{t}$ and KT run with $\varepsilon = 1$. Then, using Equation 9 for OGD and Equation 32 for KT, we obtain that the regret of KT-OGD satisfies

$$R_T(\mathbf{u}) = \mathcal{O}\left(\left(\sqrt{\ln\left(\|\mathbf{u}\|_2^2\, T + 1\right)} + 1\right)\|\mathbf{u}\|_2 \sqrt{T} + 1\right) \qquad \forall \mathbf{u} \in \mathbb{R}^d. \tag{33}$$

Note that, except for the logarithmic factor $\sqrt{\ln\left(\|\mathbf{u}\|_2^2\, T + 1\right)} + 1$, this bound matches the bound $\|\mathbf{u}\|_2 \sqrt{T}$ for OGD tuned with the unknown knowledge of $\|\mathbf{u}\|$. Indeed, this logarithmic term is the price we have to pay for the adaptivity of the algorithm to $\|\mathbf{u}\|$, as Streeter & McMahan (2012) showed it to be unavoidable in the unconstrained setting.

Note that parameter-free algorithms like KT-OMD are useful also when $\mathbf{u}$ lives in a bounded subset $\mathbb{V} \subset \mathbb{R}^d$, but the Bregman divergence D_ψ is unbounded in $\mathbb{V}$. The prime example for this scenario is the entropic mirror map ψ on the simplex $\mathbb{V}$, where D_ψ is the cross-entropy. The regret bound (Equation 10) for the corresponding instance of OMD (which we called EG) has the form

$$R_T(\mathbf{u}) = \mathcal{O}\left(\frac{\mathrm{D}_\psi(\mathbf{u}, \mathbf{w}_1)}{\eta} + \eta T\right) \qquad \forall \mathbf{u} \in \mathbb{V}$$

for any fixed choice of the step size η. In the setting of prediction with expert advice (a special case of OCO), an implicitly defined parameter-free version of EG for linear losses called NormalHedge was introduced by Chaudhuri et al. (2009). The bound was later improved by Chernov & Vovk (2010) with another implicitly defined algorithm. Using a different reduction to betting strategies, Orabona & Pál (2016) proved a bound of order $\sqrt{\left(\mathrm{D}_\psi(\mathbf{u}, \mathbf{w}_1) + 1\right) T}$ for all $\mathbf{u}$ in the simplex using a closed form update. Note that, as before, this bound is equal to the bound that one would get by tuning EG using the (unknown) knowledge of $\mathrm{D}_\psi(\mathbf{u}, \mathbf{w}_1)$. Further improvement were obtained by Koolen & Van Erven (2015), who introduced an algorithm called Squint.

One of the nice features of Equation 33 is that $R_T(\mathbf{0})$ does not depend on T. So we can then run multiple instances of the combined KT-OMD strategy using a different mirror map for each OMD instance. This allows us to bound the regret as if we ran the algorithm using the best

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

mirror map in hindsight (where best is relative to the linearized regret). Suppose, for example, we run two instances of KT-OMD and predict using the sum $\mathbf{w}_t = \mathbf{w}_t^{(1)} + \mathbf{w}_t^{(2)}$ of their predictions. Then, linearizing the regret, we obtain

$$R_T(\mathbf{u}) \leq \sum_{t=1}^{T} \mathbf{g}_t^\top (\mathbf{w}_t - \mathbf{u}) = \min_{\substack{\mathbf{v},\mathbf{z}\in\mathbb{R}^d \\ \mathbf{v}+\mathbf{z}=\mathbf{u}}} \left(\sum_{t=1}^{T} \mathbf{g}_t^\top \left(\mathbf{w}_t^{(1)} - \mathbf{v}\right) + \sum_{t=1}^{T} \mathbf{g}_t^\top \left(\mathbf{w}_t^{(2)} - \mathbf{z}\right) \right)$$

$$= \min \left\{ \sum_{t=1}^{T} \mathbf{g}_t^\top \left(\mathbf{w}_t^{(1)} - \mathbf{u}\right), \sum_{t=1}^{T} \mathbf{g}_t^\top \left(\mathbf{w}_t^{(2)} - \mathbf{u}\right) \right\}.$$

For instance, if we run OMD using the p-norm[3] mirror map $\psi = \frac{1}{2}\|\cdot\|_p^2$ for $1 < p \leq 2$, which is $(p-1)$-strongly convex with respect to the same norm, then $R_T(\mathbf{u})$ is bounded in terms of $\|\mathbf{u}\|_p$ and $\|\mathbf{g_t}\|_q$, where q is the conjugate coefficient of p, satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Interestingly, one can show (Gentile & Littlestone 1999) that choosing $p = 2\ln d/(2\ln d - 1)$ (for $d \geq 3$) gives a regret bound of the form $\left(\|\mathbf{u}\|_1^2/\alpha + \alpha\right) G\sqrt{T\ln d}$ for all $\mathbf{u} \in \mathbb{R}^d$ and for $\max_t \|\mathbf{g_t}\|_\infty \leq G$. As this is very similar to the bound obtained by EG when $\mathbb{V}$ is the simplex, we see that by choosing $p \in (1, 2]$ we can interpolate between OGD and EG. By running multiple instances of KT-OMD, where the OMD instances use different values of p, one can derive a regret bound almost as good as the p-norm OMD run with the best value of p.

5. OTHER NOTIONS OF REGRET

Bounding the regret $R_T(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{V}$ may not be crucial in some practical applications. For example, if the loss sequence $\ell_1, \ell_2, \ldots$ is such that no $\mathbf{u} \in \mathbb{V}$ achieves a small cumulative loss $\ell_1(\mathbf{u}) + \ell_2(\mathbf{u}) + \cdots$, then regret bounds may not be at all helpful in telling good algorithms from bad ones. This lack of a single good minimizer in $\mathbb{V}$ of the cumulative loss is likely to occur when the loss sequence is generated by a highly nonstationary data sequence, possibly affected by seasonalities and other disturbances. In this case, regret should be replaced by more robust measures, allowing better comparators than fixed elements of $\mathbb{V}$. In what follows, D_2 is the Euclidean diameter of $\mathbb{V}$ and L is the Lipschitz constant of the convex loss functions in the sequence $\ell_1, \ldots, \ell_T$.

5.1. Dynamic Regret

A notion of regret which captures nonstationary comparators is that of dynamic regret (Herbster & Warmuth 1998b),

$$R_T^{\text{dyn}}(\mathbf{u}_1, \ldots, \mathbf{u}_T) = \sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \sum_{t=1}^{T} \ell_t(\mathbf{u}_t) \qquad \text{where } \mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathbb{V}.$$

Note that $R_T^{\text{dyn}}(\mathbf{u}, \ldots, \mathbf{u}) = R_T(\mathbf{u})$, so dynamic regret includes standard regret as a special case. Zhang et al. (2018a) show a general lower bound on dynamic regret of the form $\Omega\left(L\sqrt{(D_2 + \Pi_{2,T})D_2 T}\right)$, where

$$\Pi_{p,T} = \sum_{t=1}^{T-1} \|\mathbf{u}_{t+1} - \mathbf{u}_t\|_p$$

is the p-norm path-length function, measuring the nonstationarity of the comparator sequence $\mathbf{u}_1, \ldots, \mathbf{u}_T$. When $\mathbf{u}_1 = \cdots = \mathbf{u}_T$ then $\Pi_{p,T} = 0$. In this case, the lower bound on the dynamic regret reduces (for $p = 2$) to the lower bound $\Omega(L_2 D_2 \sqrt{T})$ on the standard regret proven in Section 1.3.

[3] Recall the definition of p-norm of a vector $\mathbf{u} \in \mathbb{R}^d$, $\|\mathbf{u}\|_p = (|u_1|^p + \cdots + |u_d|^p)^{1/p}$.

Review in Advance first posted on
November 9, 2020. (Changes may
still occur before final publication.)

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

Herbster & Warmuth (1998b) (see also Cesa-Bianchi & Lugosi 2006, theorem 11.4) prove upper bounds on the dynamic regret of OMD run with the p-norm mirror map (for $1 < p \leq 2$) of the form

$$R_T^{\text{dyn}}(\mathbf{u}_1, \ldots, \mathbf{u}_T) \leq \frac{\|\mathbf{u}_T\|_p^2 + \Pi_{p,T} D_p}{\eta} + \eta L_q^2 T \qquad \forall \mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathbb{V},$$

where D_p is the diameter of $\mathbb{V}$ measured using the p-norm, and L_q is the Lipschitz constant (with respect to the dual q-norm) of the loss functions in the sequence $\ell_1, \ldots, \ell_T$ —Zinkevich (2003) shows similar results in the special case of $p = 2$. Choosing the step size as $\eta = 1/(L_q\sqrt{T})$ gives a suboptimal upper bound of the order of $L_q(D_p^2 + \Pi_{p,T} D_p)\sqrt{T}$. In the Euclidean case ($p = 2$), Zhang et al. (2018a) use the Hedge algorithm (a special case of EG) to aggregate $\mathcal{O}(\ln T)$ instances of OGD, each one run with a different choice of η to guess the desired value of $\Pi_{2,T}$ (up to a constant factor). They prove the dynamic regret bound $\mathcal{O}\left(L_2\sqrt{(D_2 + \Pi_{2,T})D_2 T}\right)$, matching the lower bound up to constants.

5.2. Adaptive Regret

A different view on the theme of nonstationary comparators is offered by the notion of adaptive regret (Hazan & Seshadhri 2007). Adaptive regret evaluates the performance of the online algorithm against that of the best fixed comparator in any interval of time. Formally,

$$R_{\tau,T}^{\text{ada}} = \max_{s=1,\ldots,T-\tau+1} \left(\sum_{t=s}^{s+\tau-1} \ell_t(\mathbf{w}_t) - \min_{\mathbf{u}\in\mathbb{V}} \sum_{t=s}^{s+\tau-1} \ell_t(\mathbf{u}) \right), \qquad \text{where } \tau \in \{1, \ldots, T\}.$$

In their paper, Hazan & Seshadhri (2007) use a harder notion of adaptive regret, namely $\max_\tau R_{\tau,T}^{\text{ada}}$. They show an online algorithm whose regret grows in T like $\sqrt{T(\ln T)^3}$. In a follow-up paper, Daniely et al. (2015) devise an online algorithm with the adaptive regret bound

$$R_{\tau,T}^{\text{ada}}(\mathbf{u}) = \mathcal{O}\left((DL + \ln T)\sqrt{\tau}\right) \qquad \mathbf{u} \in \mathbb{V}, \qquad 34.$$

where D is the Euclidean diameter of $\mathbb{V}$ and L is the Lipschitz constant (with respect to the Euclidean norm) of the loss functions in the sequence $\ell_1, \ldots, \ell_T$. This result is improved by Jun et al. (2017), who show the better bound $R_{\tau,T}^{\text{ada}}(\mathbf{u}) = \mathcal{O}((DL + \sqrt{\ln T})\sqrt{\tau})$ using the betting framework described in Section 4.

Most of the online algorithms for minimizing adaptive regret work by combining several instances of an online algorithm for the standard notion of regret. Each instance is run in a specific interval of time, where the set of intervals is carefully designed so that the overall number of instances to be run is $\mathcal{O}(\ln T)$. These instances are then combined using an algorithm based on the framework of prediction with expert advice (Cesa-Bianchi et al. 1997), where each instance is viewed as an expert. As instances typically run for less than T time steps, Jun et al. (2017) combine the betting framework with the sleeping experts model (Freund et al. 1997), which allows only for a subset of the experts to be active at any point in time.

Although the algorithm of Jun et al. (2017) is designed to minimize adaptive regret, Zhang et al. (2018b) show that the same algorithm can be also used to prove the following dynamic regret result:

$$R_T^{\text{dyn}}(\mathbf{u}_1, \ldots, \mathbf{u}_T) = \mathcal{O}\left(DL \max\left\{\sqrt{T \ln T},\, T^{2/3} V_T^{1/3} (\ln T)^{1/3}\right\}\right), \qquad 35.$$

where

$$V_T = \sum_{t=2}^{T} \sup_{\mathbf{u}\in\mathbb{V}} \left\| \ell_t(\mathbf{u}) - \ell_{t-1}(\mathbf{u}) \right\|_2$$

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

measures the variation of the loss sequence $\ell_1, \ldots, \ell_T$. As shown by Besbes et al. (2015), the dependence on V_T in Equation 35 is not improvable.

A result relating dynamic regret to adaptive regret is proven by Zhang et al. (2018b), who show that

$$R_T^{\mathrm{dyn}}(\mathbf{u}_1, \ldots, \mathbf{u}_T) \le \min_{\mathcal{P}_T} \sum_{\mathcal{I} \in \mathcal{P}_T} \left(R^{\mathrm{ada}}_{|\mathcal{I}|,T} + 2|\mathcal{I}| V_T(\mathcal{I}) \right),$$

where the min is taken over all partitions $\mathcal{P}_T$ of $\{1, \ldots, T\}$ in intervals $\mathcal{I} = \{t_r, \ldots, t_s\}$ of consecutive time steps, with $1 \le r \le s \le T$. The quantity

$$V_T(\mathcal{I}) = \sum_{t \in \mathcal{I}} \sup_{\mathbf{u} \in \mathbb{V}} \left\| \ell_t(\mathbf{u}) - \ell_{t-1}(\mathbf{u}) \right\|_2$$

is the variation of the loss sequence within the time interval. This is later extended by Zhang et al. (2020), who prove

$$R_T^{\mathrm{dyn}}(\mathbf{u}_1, \ldots, \mathbf{u}_T) \le \min_{\mathcal{P}_T} \sum_{\mathcal{I} \in \mathcal{P}_T} \left(R^{\mathrm{ada}}_{|\mathcal{I}|,T} + L|\mathcal{I}| \Pi_T(\mathcal{I}) \right) \qquad \forall \mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathbb{V},$$

where

$$\Pi_T(\mathcal{I}) = \sum_{t \in \mathcal{I}} \|\mathbf{u}_{t+1} - \mathbf{u}_t\|_2$$

is the path length over the interval $\mathcal{I}$. However, when combined with known bounds on the dynamic regret, this bound does not give the optimal bound $\mathcal{O}\left(L_2\sqrt{(D_2 + \Pi_{2,T})D_2 T}\right)$ for dynamic regret. Zhang et al. (2020) also derive algorithms simultaneously minimizing adaptive and dynamic regret.

Some of these notions of regret were originally introduced in the setting of prediction with expert advice (i.e., OCO with linear losses, where $\mathbb{V}$ is equal to the probability simplex and regret is measured against the corners of the simplex, where linear functions are minimized). In that framework, dynamic regret is known as tracking or shifting regret (Herbster & Warmuth 1998a). Other notions of regret, instead, are mostly studied in the experts framework. For example, policy regret (Arora et al. 2012a) applies to settings where the loss function ℓ_t at each time t depends not only on the current model $\mathbf{w}_t$ but also on the past models $\mathbf{w}_{t-s}$, where s spans a window in the past (whose size H could potentially depend on t). These loss functions can be used to model natural scenarios, such as the switching cost scenario where $H = 1$ and $\ell_t(\mathbf{w}_t, \mathbf{w}_{t+1}) = c$ whenever $\mathbf{w}_t \ne \mathbf{w}_{t-1}$ (Kalai & Vempala 2005). Swap regret (Blum & Mansour 2007) instead measures regret against a set of modification rules. Each modification rule F is an operator on the set $\{1, \ldots, d\}$ of coordinates. The instantaneous regret at time t against F of an algorithm choosing $\mathbf{w}$ in the simplex is $\ell_t(\mathbf{w}) - \ell_t(\mathbf{w}^{(F)})$, where $w_i^{(F)} = \sum_{j:F(j)=i} w_j$. Note that when $F(j) = i$ for all $j = 1, \ldots, d$, swap regret against F reduces to standard regret against the ith corner $\mathbf{e}_i$ of the simplex, that is, $\ell_t(\mathbf{w}^{(F)}) = \ell_t(\mathbf{e}_i)$ for all $\mathbf{w}$. Swap regret is especially important when using online learning algorithms to approximate equilibria in games (Cesa-Bianchi & Lugosi 2006).

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

ACKNOWLEDGMENTS

N.C.-B. is partially supported by the MIUR PRIN grant Algorithms, Games, and Digital Markets (ALGADIMAR). F.O. is partially supported by the National Science Foundation under grant no. 1925930, Collaborative Research: TRIPODS Institute for Optimization and Learning.

LITERATURE CITED

Abernethy J, Hazan E, Rakhlin A. 2008. Competing in the dark: an efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory*, pp. 263–73. Madison, WI: Omnipress

Agarwal N, Gonen A, Hazan E. 2019. Learning in non-convex games with an optimization oracle. In *Proceedings of the 32nd Annual Conference on Learning Theory*. Brookline, MA: Microtome. **http://proceedings.mlr.press/v99/agarwal19a/agarwal19a.pdf**

Arora R, Dekel O, Tewari A. 2012a. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the 29th International Conference on Machine Learning*. Madison, WI: Omnipress. **http://icml.cc/2012/papers/749.pdf**

Arora S, Hazan E, Kale S. 2012b. The multiplicative weights update method: a meta-algorithm and applications. *Theory Comput*. 8:121–64

Azoury KS, Warmuth MK. 2001. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.* 43:211–46

Beck A, Teboulle M. 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.* 31:167–75

Berger JO. 2013. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer

Besbes O, Gur Y, Zeevi A. 2015. Nonstationary stochastic optimization. *Oper. Res.* 63:1227–44

Blackwell D. 1956. An analog of the minimax theorem for vector payoffs. *Pac. J. Math.* 6:1–8

Block HD. 1962. The perceptron: a model for brain functioning. *Rev. Mod. Phys.* 34:123

Blum A, Mansour Y. 2007. From external to internal regret. *J. Mach. Learn. Res.* 8:1307–24

Borodin A, El-Yaniv R. 2005. *Online Computation and Competitive Analysis*. Cambridge, UK: Cambridge Univ. Press

Bubeck S. 2015. Convex optimization: algorithms and complexity. *Found. Trends Mach. Learn.* 8:231–357

Cesa-Bianchi N, Freund Y, Haussler D, Helmbold DP, Schapire RE, Warmuth MK. 1997. How to use expert advice. *J. ACM* 44:427–85

Cesa-Bianchi N, Lugosi G. 2006. *Prediction, Learning, and Games*. Cambridge, UK: Cambridge Univ. Press

Chaudhuri K, Freund Y, Hsu DJ. 2009. A parameter-free hedging algorithm. In *Advances in Neural Information Processing Systems 22*, ed. Y Bengio, D Schuurmans, JD Lafferty, CKI Williams, A Culotta, pp. 297–305. Red Hook, NY: Curran

Chernov A, Vovk V. 2010. Prediction with advice of unknown number of experts. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, ed. P Grunwald, P Spirtes, pp. 117–25. Arlington, VA: AUAI

Chiang CK, Yang T, Lee CJ, Mahdavi M, Lu CJ, et al. 2012. Online optimization with gradual variations. *JMLR Worksh. Conf. Proc.* 23:6.1–20

Cortes C, Vapnik V. 1995. Support-vector networks. *Mach. Learn.* 20:273–97

Cover T. 1967. Behaviour of sequential predictors of binary sequences. In *Transactions of the Fourth Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes*, pp. 263–72. Prague: Czechoslov. Acad. Sci.

Cover TM. 1974. *Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin*. Tech. Rep. 12, Dep. Stat., Stanford Univ., Stanford, CA

Cutkosky A, Orabona F. 2018. Black-box reductions for parameter-free online learning in Banach spaces. *Proc. Mach. Learn. Res.* 75:1–37

Dalalyan A, Tsybakov AB. 2008. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.* 72:39–61

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

Dalalyan AS, Salmon J. 2012. Sharp oracle inequalities for aggregation of affine estimators. *Ann. Stat.* 40:2327–55

Daniely A, Gonen A, Shalev-Shwartz S. 2015. Strongly adaptive online learning. *Proc. Mach. Learn. Res.* 37:1405–11

De Rosa R, Orabona F, Cesa-Bianchi N. 2015. The ABACOC algorithm: a novel approach for nonparametric classification of data streams. In *2015 IEEE International Conference on Data Mining*, pp. 733–38. New York: IEEE

Duchi J, Hazan E, Singer Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* 12:2121–59

Feder M. 1991. Gambling using a finite state machine. *IEEE Trans. Inf. Theory* 37:1459–65

Feder M, Merhav N, Gutman M. 1992. Universal prediction of individual sequences. *IEEE Trans. Inf. Theory* 38:1258–70

Foster DJ, Li Z, Lykouris T, Sridharan K, Tardos E. 2016. Learning in games: robustness of fast convergence. In *Advances in Neural Information Processing Systems 29*, ed. DD Lee, U von Luxburg, R Garnett, M Sugiyama, I Guyon, pp. 4734–42. Red Hook, NY: Curran

Freund Y, Schapire RE. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55:119–39

Freund Y, Schapire RE. 1999. Large margin classification using the perceptron algorithm. *Mach. Learn.* 37:277–96

Freund Y, Schapire RE, Singer Y, Warmuth MK. 1997. Using and combining predictors that specialize. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, pp. 334–43. New York: ACM

Fudenberg D, Levine DK. 1995. Consistency and cautious fictitious play. *J. Econ. Dyn. Control* 19:1065–89

Gentile C, Littlestone N. 1999. The robustness of the p-norm algorithms. In *Proceedings of the 12th Annual Conference on Computational Learning Theory*, pp. 1–11. New York: ACM

Hannan J. 1957. Approximation to Bayes risk in repeated play. *Contrib. Theory Games* 3:97–139

Hazan E. 2016. Introduction to online convex optimization. *Found. Trends Optim.* 2:157–325

Hazan E, Agarwal A, Kale S. 2007. Logarithmic regret algorithms for online convex optimization. *Mach. Learn.* 69:169–92

Hazan E, Megiddo N. 2007. Online learning with prior knowledge. In *Learning Theory: 20th Annual Conference on Learning Theory, COLT 2007*, ed. NH Bshouty, C Gentile, pp. 499–513. New York: Springer

Hazan E, Seshadhri C. 2007. Adaptive algorithms for online decision problems. *Electron. Colloq. Comput. Complexity* 14:88

Herbster M, Warmuth MK. 1998a. Tracking the best expert. *Mach. Learn.* 32:151–78

Herbster M, Warmuth MK. 1998b. Tracking the best regressor. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pp. 24–31. New York: ACM

Hoerl AE, Kennard RW. 2000. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 42:80–86

Jun KS, Orabona F, Wright S, Willett R. 2017. Improved strongly adaptive online learning using coin betting. *Proc. Mach. Learn. Res.* 54:943–51

Kalai A, Vempala S. 2005. Efficient algorithms for online decision problems. *J. Comput. Syst. Sci.* 71:291–307

Kelly J. 1956. A new interpretation of information rate. *Bell Syst. Technol. J.* 35:917–26

Kivinen J, Warmuth MK. 1997. Exponentiated gradient versus gradient descent for linear predictors. *Inform. Comput.* 132:1–63

Kivinen J, Warmuth MK. 1999. Averaging expert predictions. In *EuroCOLT 1999: Computational Learning Theory*, ed. P Fischer, HU Simon, pp. 153–67. New York: Springer

Koolen WM, Van Erven T. 2015. Second-order quantile methods for experts and combinatorial games. *Proc. Mach. Learn. Res.* 40:1764–66

Krichevsky R, Trofimov V. 1981. The performance of universal encoding. *IEEE Trans. Inform. Theory* 27:199–207

Kulis B, Bartlett PL. 2010. Implicit online learning. In *Proceedings of the 27th International Conference on Machine Learning*, ed. J Fürnkranz, T Joachims, pp. 575–82. Madison, WI: Omnipress

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

Kuzborskij I, Cesa-Bianchi N. 2017. Nonparametric online regression while learning the metric. In *Advances in Neural Information Processing Systems 30*, ed. I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, et al., pp. 667–76. Red Hook, NY: Curran

Littlestone N, Warmuth MK. 1994. The weighted majority algorithm. *Inform. Comput.* 108:212–61

Luo H, Agarwal A, Cesa-Bianchi N, Langford J. 2016. Efficient second order online learning by sketching. In *Advances in Neural Information Processing Systems 29*, ed. DD Lee, M Sugiyama, UV Luxburg, I Guyon, R Garnett, pp. 902–10. Red Hook, NY: Curran

McMahan HB. 2017. A survey of algorithms and analysis for adaptive online learning. *J. Mach. Learn. Res.* 18:1–50

McMahan HB, Streeter MJ. 2010. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Annual Conference on Learning Theory*, ed. AT Kalai, M Mohri, pp. 244–56. Madison, WI: Omnipress

Mokhtari A, Ozdaglar A, Pattathil S. 2019. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: proximal point approach. arXiv:1901.08511 [math.OC]

Nemirovsky AS, Yudin DB. 1983. *Problem Complexity and Method Efficiency in Optimization*. New York: Wiley

Novikoff AB. 1963. *On convergence proofs for perceptrons*. Tech. rep., SRI, Menlo Park, CA

Orabona F. 2019. A modern introduction to online learning. arXiv:1912.13213 [cs.LG]

Orabona F, Pál D. 2016. Coin betting and parameter-free online learning. In *Advances in Neural Information Processing Systems 29*, ed. DD Lee, M Sugiyama, UV Luxburg, I Guyon, R Garnett, pp. 577–85. Red Hook, NY: Curran

Orabona F, Pál D. 2018. Scale-free online learning. *Theor. Comput. Sci.* 716:50–69

Puterman ML. 2014. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York: Wiley

Rakhlin A, Sridharan K. 2013. Online learning with predictable sequences. *JMLR Worsh. Conf. Proc.* 30:993–1019

Rasmussen CE, Williams CKI. 2005. *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press

Rigollet P, Tsybakov AB. 2012. Sparse estimation by exponential weighting. *Stat. Sci.* 27:558–75

Robbins H. 1951. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*, ed. J Neyman, pp. 131–49. Berkeley: Univ. Calif. Press

Rosenblatt F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65:386

Shalev-Shwartz S. 2007. *Online learning: theory, algorithms, and applications*. Ph.D. Thesis, Hebrew Univ. Jerusalem

Shalev-Shwartz S. 2012. Online learning and online convex optimization. *Found. Trends Mach. Learn.* 4:107–94

Shalev-Shwartz S, Singer Y, Srebro N, Cotter A. 2011. Pegasos: primal estimated sub-gradient solver for SVM. *Math. Program.* 127:3–30

Shtarkov YM. 1987. Universal sequential coding of single messages. *Probl. Pereda. Inform.* 23:3–17

Streeter M, McMahan B. 2012. No-regret algorithms for unconstrained online convex optimization. In *Advances in Neural Information Processing Systems 25*, ed. F Periera, CJC Burges, L Bottou, KQ Weinberger, pp. 2402–10. Red Hook, NY: Curran

Syrgkanis V, Agarwal A, Luo H, Schapire RE. 2015. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems 28*, ed. C Cortes, ND Lawrence, DD Lee, M Sugiyama, R Garnett, pp. 2989–97. Red Hook, NY: Curran

Vovk V. 2001. Competitive on-line statistics. *Int. Stat. Rev.* 69:213–48

Zhang L, Lu S, Yang T. 2020. Minimizing dynamic regret and adaptive regret simultaneously. arXiv:2002.02085 [cs.LG]

Zhang L, Lu S, Zhou ZH. 2018a. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems 31*, ed. S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett, pp. 1323–33. Red Hook, NY: Curran

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)

Zhang L, Yang T, Zhou ZH, et al. 2018b. Dynamic regret of strongly adaptive methods. *Proc. Mach. Learn. Res.* 80:5882–91

Zinkevich M. 2003. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, ed. T Fawcett, N Mishra, pp. 928–36. Menlo Park, CA: AAAI

Zinkevich M. 2004. *Theoretical guarantees for algorithms in multi-agent settings*. PhD Thesis, Sch. Comput. Sci., Carnegie Mellon University, Pittsburgh, PA

Annu. Rev. Stat. Appl. 2021.8. Downloaded from www.annualreviews.org
Access provided by University of Liverpool on 11/13/20. For personal use only.

Review in Advance first posted on November 9, 2020. (Changes may still occur before final publication.)