On the Universality of Online Mirror Descent

Nathan Srebro
TTIC
nati@ttic.edu

Karthik Sridharan
TTIC
karthik@ttic.edu

Ambuj Tewari
University of Texas at Austin
ambuj@cs.utexas.edu

Abstract

We show that for a general class of convex online learning problems, Mirror Descent can always achieve a (nearly) optimal regret guarantee.

1 Introduction

Mirror Descent is a first-order optimization procedure which generalizes the classic Gradient Descent procedure to non-Euclidean geometries by relying on a "distance generating function" specific to the geometry (the squared ℓ_2-norm in the case of standard Gradient Descent) [14, 4]. Mirror Descent is also applicable, and has been analyzed, in a stochastic optimization setting [9] and in an online setting, where it can ensure bounded online regret [20]. In fact, many classical online learning algorithms can be viewed as instantiations or variants of Online Mirror Descent, generally either with the Euclidean geometry (e.g. the Perceptron algorithm [5] and Online Gradient Descent [27]), or in the simplex (ℓ_1 geometry), using an entropic distance generating function (Winnow [13] and Multiplicative Weights / Online Exponentiated Gradient algorithm [11]). More recently, the Online Mirror Descent framework has been applied, with appropriate distance generating functions derived for a variety of new learning problems like multi-task learning and other matrix learning problems [10], online PCA [26] etc.

In this paper, we show that Online Mirror Descent is, in a sense, *universal*. That is, for any convex online learning problem, of a general form (specified in Section 2), if the problem is online learnable, then it is online learnable, with a nearly optimal regret rate, using Online Mirror Descent, with an appropriate distance generating function. Since Mirror descent is a first order method and often has simple and computationally efficient update rules, this makes the result especially attractive. Viewing online learning as a sequentially repeated game, this means that Online Mirror Descent is a near optimal strategy, guaranteeing an outcome very close to the value of the game.

In order to show such universality, we first generalize and refine the standard Mirror Descent analysis to situations where the constraint set is not the dual of the data domain, obtaining a general upper bound on the regret of Online Mirror Descent in terms of the existence of an appropriate uniformly convex distance generating function (Section 3). We then extend the notion of a *martingale type* of a Banach space to be sensitive to both the constraint set and the data domain, and building on results of [24], we relate the value of the online learning repeated game to this generalized notion of martingale type (Section 4). Finally, again building on and generalizing the work of [16], we show how having appropriate martingale type guarantees the existence of a good uniformly convex function (Section 5), that in turn establishes the desired nearly-optimal guarantee on Online Mirror Descent (Section 6). We mainly build on the analysis of [24], who related the value of the online game to the notion of martingale type of a Banach space and uniform convexity when the constraint set and data domain are dual to each other. The main technical advance here is a non-trivial generalization of their analysis (as well as the Mirror Descent analysis) to the more general situation where the constraint set and data domain are chosen independently of each other. In Section 7 several examples are provided that demostrate the use of our analysis.

Mirror Descent was initially introduced as a first order deterministic optimization procedure, with an ℓ_p constraint and a matching ℓ_q Lipschitz assumption ($1 \leq p \leq 2, 1/q + 1/p = 1$), was shown to be optimal in terms of the *number of exact gradient evaluations* [15]. Shalev-Shwartz and Singer later observed that the online version of Mirror Descent, again with an ℓ_p bound and matching ℓ_q Lipschitz assumption ($1 \leq p \leq 2, 1/q + 1/p = 1$), is also optimal in terms

of the worst-case (adversarial) online regret. In fact, in such scenarios stochastic Mirror Descent is also optimal in terms of the number of samples used. We emphasize that although in most, if not all, settings known to us these three notions of optimality coincide, here we focus only on the worst-case online regret.

Sridharan and Tewri [24] generalized the optimality of online Mirror Descent (w.r.t. regret) to scenarios where learner is constrained to a unit ball of an arbitrary Banach space (not necessarily and ℓ_p space) and the objective functions have sub-gradients that lie in the *dual ball* of the space—for reasons that will become clear shortly, we refer to this as the *data domain*. However, often we encounter problems where the constraint set and data domain are not dual balls, but rather are arbitrary convex subsets. In this paper, we explore this more general, "non-dual", variant, and show that also in such scenarios online Mirror Descent is (nearly) optimal in terms of the (asymptotic) worst-case online regret.

2 Online Convex Learning Problem

An online convex learning problem can be viewed as a multi-round repeated game where on round t, the learner first picks a vector (predictor) $\mathbf{w}_t$ from some fixed set $\mathcal{W}$, which is a closed convex subset of a vector space $\mathcal{B}$. Next, the adversary picks a convex cost function $f_t : \mathcal{W} \mapsto \mathbb{R}$ from a class of convex functions $\mathcal{F}$. At the end of the round, the learner pays instantaneous cost $f_t(\mathbf{w}_t)$. We refer to the strategy used by the learner to pick the f_t's as an *online learning algorithm*. More formally, an online learning algorithm $\mathcal{A}$ for the problem is specified by the mapping $\mathcal{A} : \bigcup_{n\in\mathbb{N}} \mathcal{F}^{n-1} \mapsto \mathcal{W}$. The regret of the algorithm $\mathcal{A}$ for a given sequence of cost functions $f_1, \ldots, f_n$ is given by

$$\mathbf{R}_n(\mathcal{A}, f_1, \ldots, f_n) = \frac{1}{n}\sum_{t=1}^{n} f_t(\mathcal{A}(f_{1:t-1})) - \inf_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{t=1}^{n} f_t(\mathbf{w}) \,.$$

The goal of the learner (or the online learning algorithm), is to minimize the regret for any n.

In this paper, we consider cost function classes $\mathcal{F}$ specified by a convex subset $\mathcal{X} \subset \mathcal{B}^\star$ of the dual space $\mathcal{B}^\star$. We consider various types of classes, where for all of them, subgradients[1] of the functions in $\mathcal{F}$ lie inside $\mathcal{X}$ (we use the notation $\langle \mathbf{x}, \mathbf{w}\rangle$ to mean applying linear functional $\mathbf{x} \in \mathcal{B}^\star$ on $\mathbf{w} \in \mathcal{B}$) :

$$\mathcal{F}_{\mathrm{Lip}}(\mathcal{X}) = \{f : f \text{ is convex } \forall \mathbf{w} \in \mathcal{W}, \nabla f(\mathbf{w}) \in \mathcal{X}\}\,, \qquad \mathcal{F}_{\mathrm{lin}}(\mathcal{X}) = \{\mathbf{w} \mapsto \langle \mathbf{x}, \mathbf{w}\rangle : \mathbf{x} \in \mathcal{X}\}\,,$$
$$\mathcal{F}_{\mathrm{sup}}(\mathcal{X}) = \{\mathbf{w} \mapsto |\langle \mathbf{x}, \mathbf{w}\rangle - y| : \mathbf{x} \in \mathcal{X}, y \in [-b, b]\}$$

The value of the game is then the best possible worst-case regret guarantee an algorithm can enjoy. Formally :

$$\mathcal{V}_n(\mathcal{F}, \mathcal{X}, \mathcal{W}) = \inf_{\mathcal{A}} \sup_{f_{1:n}\in\mathcal{F}(\mathcal{X})} \mathbf{R}_n(\mathcal{A}, f_{1:n}) \tag{1}$$

It is well known that the value of a game for all the above sets $\mathcal{F}$ is the same. More generally:

Proposition 1. *If for a convex function class $\mathcal{F}$, we have that $\forall f \in \mathcal{F}, \mathbf{w} \in \mathcal{W}, \nabla f(\mathbf{w}) \in \mathcal{X}$ then,*

$$\mathcal{V}_n(\mathcal{F}, \mathcal{X}, \mathcal{W}) \le \mathcal{V}_n(\mathcal{F}_{\mathrm{lin}}, \mathcal{X}, \mathcal{W})$$

Furthermore, $$\mathcal{V}_n(\mathcal{F}_{\mathrm{Lip}}, \mathcal{X}, \mathcal{W}) = \mathcal{V}_n(\mathcal{F}_{\mathrm{sup}}, \mathcal{X}, \mathcal{W}) = \mathcal{V}_n(\mathcal{F}_{\mathrm{lin}}, \mathcal{X}, \mathcal{W})$$

That is, the value for any class $\mathcal{F}$ for which subgradients are in $\mathcal{X}$, is upper bounded by the value of the class of linear functionals in $\mathcal{W}$, see e.g. [1]. In particular, this includes the class $\mathcal{F}_{\mathrm{Lip}}$ which is the class of *all* functions with subgradients in $\mathcal{X}$, and since $\mathcal{F}_{\mathrm{lin}}(\mathcal{X}) \subset \mathcal{F}_{\mathrm{Lip}}(\mathcal{X})$ we get the first equality. The second equality is shown in [18].

The class $\mathcal{F}_{\mathrm{sup}}(\mathcal{X})$ corresponds to linear prediction with an absolute-difference loss, and thus its value is the best possible guarantee for online supervised learning with this loss. We can define more generally a class $\mathcal{F}_\ell = \{\ell(\langle \mathbf{x}, \mathbf{w}\rangle, y) : \mathbf{x} \in \mathcal{X}, y \in [-b, b]\}$ for any 1-Lipschitz loss ℓ, and this class would also be of the desired type, with its value upper bounded by $\mathcal{V}_n(\mathcal{F}_{\mathrm{lin}}, \mathcal{X}, \mathcal{W})$. In fact, this setting includes supervised learning fairly generally, including problems such as multitask learning and matrix completion, where in all cases $\mathcal{X}$ specifies the data domain[2]. The equality in the above proposition can also be extended to other commonly occurring convex loss function classes like the hinge loss class with some extra constant factors.

[1] Throughout we commit to a slight abuse of notation, with $\nabla f(\mathbf{w})$ indicating some sub-gradient of f at $\mathbf{w}$ and $\nabla f(\mathbf{w}) \in \mathcal{X}$ meaning that at least one of the sub-gradients is in $\mathcal{X}$.

[2] Any convex supervised learning problem can be viewed as linear classification with some convex constraint $\mathcal{W}$ on predictors.

Owing to Proposition 1, we can focus our attention on the class $\mathcal{F}_{\text{lin}}$ (as other two behave similarly), and use shorthand

$$\mathcal{V}_n(\mathcal{W}, \mathcal{X}) := \mathcal{V}_n(\mathcal{F}_{\text{lin}}, \mathcal{X}, \mathcal{W}) \tag{2}$$

Henceforth the term value without any qualification refers to value of the linear game. Further, for any $p \in [1, 2]$ let,

$$V_p := \inf \left\{ V \;\middle|\; \forall n \in \mathbb{N}, \mathcal{V}_n(\mathcal{W}, \mathcal{X}) \leq V n^{-\left(1-\frac{1}{p}\right)} \right\} \tag{3}$$

Most prior work on online learning and optimization considers the case when $\mathcal{W}$ is the unit ball of some Banach space, and $\mathcal{X}$ is the unit ball of the dual space, i.e. $\mathcal{W}$ and $\mathcal{X}$ are related to each other through duality. In this work, however, we analyze the general problem where $\mathcal{X} \in \mathcal{B}^\star$ is not necessarily the dual ball of $\mathcal{W}$. It will be convenient for us to relate the notions of a convex set and a corresponding norm. The Minkowski functional of a subset $\mathcal{K}$ of a vector space $\mathcal{V}$ is defined as $\|\mathbf{v}\|_{\mathcal{K}} := \inf \{\alpha > 0 : \mathbf{v} \in \alpha\mathcal{K}\}$. If $\mathcal{K}$ is convex and centrally symmetric (i.e. $\mathcal{K} = -\mathcal{K}$), then $\|\cdot\|_{\mathcal{K}}$ is a semi-norm. **Throughout this paper, we will require that $\mathcal{W}$ and $\mathcal{X}$ are convex and centrally symmetric.** Further, if the set $\mathcal{K}$ is bounded then $\|\cdot\|_{\mathcal{K}}$ is a norm. Although not strictly required for our results, for simplicity we will assume $\mathcal{W}$ and $\mathcal{X}$ are are such that $\|\cdot\|_{\mathcal{W}}$ and $\|\cdot\|_{\mathcal{X}}$ (the Minkowski functionals of the sets $\mathcal{W}$ and $\mathcal{X}$) are norms. Even though we do this for simplicity, we remark that all the results go through for semi-norms. We use $\mathcal{X}^\star$ and $\mathcal{W}^\star$ to represent the dual of balls $\mathcal{X}$ and $\mathcal{W}$ respectively, i.e. the unit balls of the dual norms $\|\cdot\|_{\mathcal{X}}^*$ and $\|\cdot\|_{\mathcal{W}}^*$.

3 Mirror Descent and Uniform Convexity

A key tool in the analysis mirror descent is the notion of strong convexity, or more generally uniform convexity:

Definition 1. *$\Psi : \mathcal{B} \to \mathbb{R}$ is q-uniformly convex w.r.t. $\|\cdot\|$ if for any $\mathbf{w}, \mathbf{w}' \in \mathcal{B}$:*

$$\forall_{\alpha \in [0,1]} \;\; \Psi\left(\alpha\mathbf{w} + (1-\alpha)\mathbf{w}'\right) \leq \alpha\Psi(\mathbf{w}) + (1-\alpha)\Psi(\mathbf{w}') - \tfrac{\alpha(1-\alpha)}{q} \|\mathbf{w} - \mathbf{w}'\|^q$$

We emphasize that in the definition above, the norm $\|.\|$ and the subset $\mathcal{W}$ need not be related, and we only require uniform convexity inside $\mathcal{W}$. This allows us to relate a norm with a non-matching "ball". To this end define,

$$D_p := \inf \left\{ \left(\sup_{\mathbf{w} \in \mathcal{W}} \Psi(\mathbf{w}) \right)^{\frac{p-1}{p}} \;\middle|\; \Psi : \mathcal{W} \mapsto \mathbb{R}^+ \text{ is } \tfrac{p}{p-1}\text{-uniformly convex w.r.t. } \|\cdot\|_{\mathcal{X}^*}, \Psi(0) = 0 \right\}$$

Given a function Ψ, the Mirror Descent algorithm, $\mathcal{A}_{\text{MD}}$ is given by

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} \Delta_\Psi\left(\mathbf{w} | \mathbf{w}_t\right) + \eta \langle \nabla f_t(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t \rangle \tag{4}$$

$$\text{or equivalently} \qquad \mathbf{w}'_{t+1} = \nabla\Psi^*\left(\nabla\Psi(\mathbf{w}_t) - \eta \nabla f_t(\mathbf{w}_t)\right), \;\; \mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} \Delta_\Psi\left(\mathbf{w} \middle| \mathbf{w}'_{t+1}\right) \tag{5}$$

where $\Delta_\Psi\left(\mathbf{w}|\mathbf{w}'\right) := \Psi(\mathbf{w}) - \Psi(\mathbf{w}') - \langle \nabla\Psi(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \rangle$ is the Bregman divergence and Ψ^* is the convex conjugate of Ψ. As an example notice that when $\Psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$ then we get the gradient descent algorithm and when $\mathcal{W}$ is the d dimensional simplex and $\Psi(\mathbf{w}) = \sum_{i=1}^d \mathbf{w}_i \log(1/\mathbf{w}_i)$ then we get the multiplicative weights update algorithm.

Lemma 2. *Let $\Psi : \mathcal{B} \mapsto \mathbb{R}$ be non-negative and q-uniformly convex w.r.t. norm $\|\cdot\|_{\mathcal{X}^*}$. For the Mirror Descent algorithm with this Ψ, using $\mathbf{w}_1 = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} \Psi(\mathbf{w})$ and $\eta = \left(\frac{\sup_{\mathbf{w} \in \mathcal{W}} \Psi(\mathbf{w})}{nB}\right)^{1/p}$ we can guarantee that for any $f_1, \ldots, f_n$ s.t. $\frac{1}{n}\sum_{t=1}^n \|\nabla f_t\|_{\mathcal{X}}^p \leq 1$ (where $p = \frac{q}{q-1}$),*

$$\mathbf{R}(\mathcal{A}_{\text{MD}}, f_1, \ldots, f_n) \leq 2\left(\frac{\sup_{\mathbf{w} \in \mathcal{W}} \Psi(\mathbf{w})}{n}\right)^{\frac{1}{q}}.$$

Note that in our case we have $\nabla f \in \mathcal{X}$, i.e. $\|\nabla f\|_{\mathcal{X}} \leq 1$, and so certainly $\frac{1}{n}\sum_{t=1}^n \|\nabla f_t\|_{\mathcal{X}}^p \leq 1$. Similarly to the value of the game, for any $p \in [1, 2]$, we define:

$$\text{MD}_p := \inf \left\{ D : \exists \Psi, \eta \text{ s.t. } \forall n \in \mathbb{N}, \sup_{f_{1:n} \in \mathcal{F}(\mathcal{X})} \mathbf{R}_n(\mathcal{A}_{\text{MD}}, f_{1:n}) \leq D n^{-(1-\frac{1}{p})} \right\} \tag{6}$$

where the Mirror Descent algorithm in the above definition is run with the corresponding Ψ and η. The constant MD_p is a characterization of the best guarantee the Mirror Descent algorithm can provide. Lemma 2 therefore implies:

Corollary 3. $V_p \leq \mathrm{MD}_p \leq 2D_p$.

Proof. The first inequality is by the definition of V_p and MD_p. Second inequality follows from previous lemma. □

The Mirror Descent bound suggests that as long as we can find an appropriate function Ψ that is uniformly convex w.r.t. $\|\cdot\|_{\mathcal{X}}^*$ we can get a diminishing regret guarantee. This suggests constructing the following function:

$$\tilde{\Psi}_q := \operatorname*{argmin}_{\substack{\psi:\psi \text{ is } q\text{-uniformly convex} \\ \text{w.r.t. } \|\cdot\|_{\mathcal{X}^*} \text{ on } \mathcal{W} \text{ and } \psi \geq 0}} \sup_{\mathbf{w}\in\mathcal{W}} \Psi(\mathbf{w}) \,. \tag{7}$$

If no q-uniformly convex function exists then $\tilde{\Psi}_q = \infty$ is assumed by default. The above function is in a sense the best choice for the Mirror Descent bound in (2). The question then is: when can we find such appropriate functions and what is the best rate we can guarantee using Mirror Descent?

4 Martingale Type and Value

In [24], it was shown that the concept of the *Martingale type* (also sometimes called the *Haar type*) of a Banach space and optimal rates for online convex optimization problem, where $\mathcal{X}$ and $\mathcal{W}$ are duals of each other, are closely related. In this section we extend the classic notion of Martingale type of a Banach space (see for instance [16]) to one that accounts for the pair $(\mathcal{W}^\star, \mathcal{X})$. Before we proceed with the definitions we would like to introduce a few necessary notations. First, throughout we shall use $\epsilon \in \{\pm 1\}^{\mathbb{N}}$ to represent infinite sequence of signs drawn uniformly at random (i.e. each ϵ_i has equal probability of being $+1$ or -1). Also throughout $(\mathbf{x}_n)_{n\in\mathbb{N}}$ represents a sequence of mappings where each $\mathbf{x}_n : \{\pm 1\}^{n-1} \mapsto \mathcal{B}^\star$. We shall commit to the abuse of notation and use $\mathbf{x}_n(\epsilon)$ to represent $\mathbf{x}_n(\epsilon) = \mathbf{x}_n(\epsilon_1, \ldots, \epsilon_{n-1})$ (i.e. although we used entire ϵ as argument, $\mathbf{x}_n$ only depends on first $n-1$ signs). We are now ready to give the extended definition of Martingale type (or M-type) of a pair $(\mathcal{W}^\star, \mathcal{X})$.

Definition 2. *A pair $(\mathcal{W}^\star, \mathcal{X})$ of subsets of a vector space $\mathcal{B}^\star$ is said to be of M-type p if there exists a constant $C \geq 1$ such that for all sequence of mappings $(\mathbf{x}_n)_{n\geq 1}$ where each $\mathbf{x}_n : \{\pm 1\}^{n-1} \mapsto \mathcal{B}^\star$ and any $\mathbf{x}_0 \in \mathcal{B}^\star$:*

$$\sup_n \mathbb{E}\left[\left\|\mathbf{x}_0 + \sum_{i=1}^n \epsilon_i \mathbf{x}_i(\epsilon)\right\|_{\mathcal{W}^\star}^p\right] \leq C^p \left(\|\mathbf{x}_0\|_{\mathcal{X}}^p + \sum_{n\geq 1} \mathbb{E}\left[\|\mathbf{x}_n(\epsilon)\|_{\mathcal{X}}^p\right]\right) \tag{8}$$

The concept is called Martingale type because $(\epsilon_n \mathbf{x}_n(\epsilon))_{n\in\mathbb{N}}$ is a martingale difference sequence and it can be shown that rate of convergence of martingales in Banach spaces is governed by the rate of convergence of martingales of the form $Z_n = \mathbf{x}_0 + \sum_{i=1}^n \epsilon_i \mathbf{x}_i(\epsilon)$ (which are incidentally called Walsh-Paley martingales). We point the reader to [16, 17] for more details. Further, for any $p \in [1, 2]$ we also define,

$$C_p := \inf\left\{ C \;\middle|\; \forall \mathbf{x}_0 \in \mathcal{B}^\star, \forall (\mathbf{x}_n)_{n\in\mathbb{N}},\ \sup_n \mathbb{E}\left[\left\|\mathbf{x}_0 + \sum_{i=1}^n \epsilon_i \mathbf{x}_i(\epsilon)\right\|_{\mathcal{W}^\star}^p\right] \leq C^p \left(\|\mathbf{x}_0\|_{\mathcal{X}}^p + \sum_{n\geq 1} \mathbb{E}\,\|\mathbf{x}_n(\epsilon)\|_{\mathcal{X}}^p\right)\right\}$$

C_p is useful in determining if the pair $(\mathcal{W}^\star, \mathcal{X})$ has Martingale type p.

The results of [24, 18] showing that a Martingale type implies low regret, actually apply also for "non-matching" $\mathcal{W}$ and $\mathcal{X}$ and, in our notation, imply that $V_p \leq 2C_p$. Specifically we have the following theorem from [24, 18] :

Theorem 4. *[24, 18] For any $\mathcal{W} \in \mathcal{B}$ and any $\mathcal{X} \in \mathcal{B}^\star$ and any $n \geq 1$,*

$$\sup_{\mathbf{x}} \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbf{x}_i(\epsilon)\right\|_{\mathcal{W}^\star}\right] \leq \mathcal{V}_n(\mathcal{W}, \mathcal{X}) \leq 2 \sup_{\mathbf{x}} \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbf{x}_i(\epsilon)\right\|_{\mathcal{W}^\star}\right]$$

where the supremum above is over sequence of mappings $(\mathbf{x}_n)_{n\geq 1}$ where each $\mathbf{x}_n : \{\pm 1\}^{n-1} \mapsto \mathcal{X}$.

Our main interest here will is in establishing that low regret implies Martingale type. To do so, we start with the above theorem to relate value of the online convex optimization game to rate of convergence of martingales in the Banach space. We then extend the result of Pisier in [16] to the "non-matching" setting combining it with the above theorem to finally get :

Lemma 5. *If for some $r \in (1,2]$ there exists a constant $D > 0$ such that for any n,*

$$\mathcal{V}_n(\mathcal{W}, \mathcal{X}) \leq Dn^{-(1-\frac{1}{r})}$$

then for all $p < r$, we can conclude that any $\mathbf{x}_0 \in \mathcal{B}^\star$ and any $\mathcal{B}^\star$ sequence of mappings $(\mathbf{x}_n)_{n\geq 1}$ where each $\mathbf{x}_n : \{\pm 1\}^{n-1} \mapsto \mathcal{B}^\star$ will satisfy :

$$\sup_n \mathbb{E}\left[\left\|\mathbf{x}_0 + \sum_{i=1}^n \epsilon_i \mathbf{x}_i(\epsilon)\right\|_{\mathcal{W}^\star}^p\right] \leq \left(\frac{1104\, D}{(r-p)^2}\right)^p \left(\|\mathbf{x}_0\|_{\mathcal{X}}^p + \sum_{i\geq 1} \mathbb{E}\left[\|\mathbf{x}_i(\epsilon)\|_{\mathcal{X}}^p\right]\right)$$

That is, the pair $(\mathcal{W}, \mathcal{X})$ is of martingale type p.

The following corollary is an easy consequence of the above lemma.

Corollary 6. *For any $p \in [1,2]$ and any $p' < p$: $C_{p'} \leq \frac{1104\, V_p}{(p-p')^2}$*

5 Uniform Convexity and Martingale Type

The classical notion of Martingale type plays a central role in the study of geometry of Banach spaces. In [16], it was shown that a Banach space has Martingale type p (the classical notion) if and only if uniformly convex functions with certain properties exist on that space (w.r.t. the norm of that Banach space). In this section, we extend this result and show how the Martingale type of a pair $(\mathcal{W}^\star, \mathcal{X})$ are related to existence of certain uniformly convex functions. Specifically, the following theorem shows that the notion of Martingale type of pair $(\mathcal{W}^\star, \mathcal{X})$ is equivalent to the existence of a non-negative function that is uniformly convex w.r.t. the norm $\|\cdot\|_{\mathcal{X}^\star}$.

Lemma 7. *If, for some $p \in (1,2]$, there exists a constant $C > 0$, such that for all sequences of mappings $(\mathbf{x}_n)_{n\geq 1}$ where each $\mathbf{x}_n : \{\pm 1\}^{n-1} \mapsto \mathcal{B}^\star$ and any $\mathbf{x}_0 \in \mathcal{B}^\star$:*

$$\sup_n \mathbb{E}\left[\left\|\mathbf{x}_0 + \sum_{i=1}^n \epsilon_i \mathbf{x}_i(\epsilon)\right\|_{\mathcal{W}^\star}^p\right] \leq C^p \left(\|\mathbf{x}_0\|_{\mathcal{X}}^p + \sum_{n\geq 1} \mathbb{E}\left[\|\mathbf{x}_n(\epsilon)\|_{\mathcal{X}}^p\right]\right)$$

(i.e. $(\mathcal{W}^\star, \mathcal{X})$ has Martingale type p), then there exists a convex function $\Psi : \mathcal{B} \mapsto \mathbb{R}^+$ with $\Psi(0) = 0$, that is q-uniformly convex w.r.t. norm $\|\cdot\|_{\mathcal{X}^\star}$ s.t. $\forall \mathbf{w} \in \mathcal{B}$, $\frac{1}{q}\|\mathbf{w}\|_{\mathcal{X}^\star}^q \leq \Psi(\mathbf{w}) \leq \frac{C^q}{q}\|\mathbf{w}\|_{\mathcal{W}}^q$.

The following corollary follows directly from the above lemma.

Corollary 8. *For any $p \in [1,2]$, $D_p \leq C_p$.*

The proof of Lemma 7 goes further and gives a specific uniformly convex function Ψ satisfying the desired requirement (i.e. establishing $D_p \leq C_p$) under the assumptions of the previous lemma:

$$\Psi_q^*(\mathbf{x}) := \sup\left\{\frac{1}{C^p}\sup_n \mathbb{E}\left[\left\|\mathbf{x} + \sum_{i=1}^n \epsilon_i \mathbf{x}_i(\epsilon)\right\|_{\mathcal{W}^\star}^p\right] - \sum_{i\geq 1}\mathbb{E}\left[\|\mathbf{x}_i(\epsilon)\|_{\mathcal{X}}^p\right]\right\} \;,\; \Psi_q := (\Psi_q^*)^* \,. \tag{9}$$

where the supremum above is over sequences $(\mathbf{x}_n)_{n\in\mathbb{N}}$ and $p = \frac{q}{q-1}$.

6 Optimality of Mirror Descent

In the Section 3, we saw that if we can find an appropriate uniformly convex function to use in the mirror descent algorithm, we can guarantee diminishing regret. However the pending question there was when can we find such a function and what is the rate we can gaurantee. In Section 4 we introduced the extended notion of Martingale type of a pair $(\mathcal{W}^\star, \mathcal{X})$ and how it related to the value of the game. Then, in Section 5, we saw how the concept of M-type related to existence of certain uniformly convex functions. We can now combine these results to show that the mirror descent algorithm is a universal online learning algorithm for convex learning problems. Specifically we show that whenever a problem is online learnable, the mirror descent algorithm can guarantee near optimal rates:

Theorem 9. *If for some constant $V > 0$ and some $q \in [2, \infty)$, $\mathcal{V}_n(\mathcal{W}, \mathcal{X}) \leq V n^{-\frac{1}{q}}$ for all n, then for any $n > e^{q-1}$, there exists regularizer function Ψ and step-size η, such that the regret of the mirror descent algorithm using Ψ against any $f_1, \ldots, f_n$ chosen by the adversary is bounded as:*

$$\mathbf{R}_n(\mathcal{A}_{\mathrm{MD}}, f_{1:n}) \leq 6002\, V\, \log^2(n)\, n^{-\frac{1}{q}} \tag{10}$$

Proof. Combining Mirror descent guarantee in Lemma 2, Lemma 7 and the lower bound in Lemma 5 with $p = \frac{q}{q-1} - \frac{1}{\log(n)}$ we get the above statement. □

The above Theorem tells us that, with appropriate Ψ and learning rate η, mirror descent will obtain regret at most a factor of $6002 \log(n)$ from the best possible worst-case upper bound. We would like to point out that the constant V in the value of the game appears linearly and there is no other problem or space related hidden constants in the bound.

The following figure summarizes the relationship between the various constants. The arrow mark from $C_{p'}$ to C_p indicates that for any n, all the quantities are within $\log^2 n$ factor of each other.

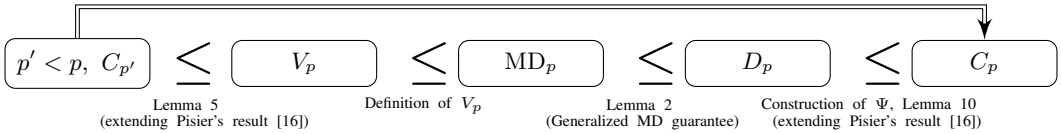

Figure 1: Relationship between the various constants

We now provide some general guidelines that will help us in picking out appropriate function Ψ for mirror descent. First we note that though the function Ψ_q in the construction (9) need not be such that $(q\Psi_q(\mathbf{w}))^{1/q}$ is a norm, with a simple modification as noted in [17] we can make it a norm. This basically tells us that the pair $(\mathcal{W}, \mathcal{X})$ is online learnable, if and only if we can sandwich a q-uniformly convex norm in-between $\mathcal{X}^\star$ and a scaled version of $\mathcal{W}$ (for some $q < \infty$). Also note that by definition of uniform convexity, if any function Ψ is q-uniformly convex w.r.t. some norm $\|\cdot\|$ and we have that $\|\cdot\| \geq c\,\|\cdot\|_{\mathcal{X}}$, then $\frac{\Psi(\cdot)}{c^q}$ is q-uniformly convex w.r.t. norm $\|\cdot\|_{\mathcal{X}}$. These two observations together suggest that given pair $(\mathcal{W}, \mathcal{X})$ what we need to do is find a norm $\|\cdot\|$ in between $\|\cdot\|_{\mathcal{X}}^\star$ and $C\,\|\cdot\|_{\mathcal{W}}$ ($C < \infty$, smaller the C better the bound) such that $\|\cdot\|^q$ is q-uniformly convex w.r.t $\|\cdot\|$.

7 Examples

We demonstrate our results on several online learning problems, specified by $\mathcal{W}$ and $\mathcal{X}$.

ℓ_p non-dual pairs It is usual in the literature to consider the case when $\mathcal{W}$ is the unit ball of the ℓ_p norm in some finite dimension d while $\mathcal{X}$ is taken to be the unit ball of the dual norm ℓ_q where p, q are Hölder conjugate exponents. Using the machinery developed in this paper, it becomes effortless to consider the non-dual case when $\mathcal{W}$ is the unit ball B_{p_1} of some ℓ_{p_1} norm while $\mathcal{X}$ is the unit ball B_{p_2} for *arbitrary* p_1, p_2 in $[1, \infty]$. We shall use q_1 and q_2 to represent Holder conjugates of p_1 and p_2. Before we proceed we first note that for any $r \in (1, 2]$, $\psi_r(\mathbf{w}) := \frac{1}{2(r-1)}\|\mathbf{w}\|_r^2$ is 2-uniformly w.r.t. norm $\|\cdot\|_r$ (see for instance [25]). On the other hand by Clarkson's inequality, we have that for $r \in (2, \infty)$, $\psi_r(\mathbf{w}) := \frac{2^r}{r}\|\mathbf{w}\|_r^r$ is r-uniformly convex w.r.t. $\|\cdot\|_r$. Putting it together we see that for any $r \in (1, \infty)$, the function ψ_r defined above, is Q-uniformly convex w.r.t $\|\cdot\|_r$ for $Q = \max\{r, 2\}$. The basic technique idea is to be to select ψ_r based on the guidelines in the end of the previous section. Finally we show that using $\tilde{\psi}_r := d^{Q \max\{\frac{1}{q_2} - \frac{1}{r}, 0\}}\psi_r$ in Mirror descent Lemma 2 yields the bound that for any $f_1, \ldots, f_n \in \mathcal{F}$:

$$\mathbf{R}_n(\mathcal{A}_{\mathrm{MD}}, f_{1:n}) \leq \frac{2\max\{2, \frac{1}{\sqrt{2(r-1)}}\} d^{\max\{\frac{1}{q_2} - \frac{1}{r}, 0\} + \max\{\frac{1}{r} - \frac{1}{p_1}, 0\}}}{n^{1/\max\{r,2\}}}$$

The following table summarizes the scenarios where a value of $r = 2$, i.e. a rate of $D_2/\sqrt{n}$, is possible, and lists the corresponding values of D_2 (up to numeric constant of at most 16):

p_1 Range	$q_2 = \frac{p_2}{p_2-1}$ Range	D_2
$1 \le p_1 \le 2$	$q_2 > 2$	1
$1 \le p_1 \le 2$	$p_1 \le q_2 \le 2$	$\sqrt{p_2 - 1}$
$1 \le p_1 \le 2$	$1 \le q_2 < p_1$	$d^{1/q_2 - 1/p_1}\sqrt{p_2 - 1}$
$p_1 > 2$	$q_2 > 2$	$d^{(1/2 - 1/p_1)}$
$p_1 > 2$	$1 \le q_2 \le 2$	$d^{(1/q_2 - 1/p_1)}$
$1 \le p_1 \le 2$	$q_2 = \infty$	$\sqrt{\log(d)}$

Note that the first two rows are dimension free, and so apply also in infinite-dimensional settings, whereas in the other scenarios, D_2 is finite only when the dimension is finite. An interesting phenomena occurs when d is ∞, $p_1 > 2$ and $q_2 \ge p_1$. In this case $D_2 = \infty$ and so one cant expect a rate of $O(\frac{1}{\sqrt{n}})$. However we have $D_{p_2} < 16$ and so can still get a rate of $n^{-\frac{1}{q_2}}$.

Ball et al [3] tightly calculate the constants of strong convexity of squared ℓ_p norms, establishing the tightness of D_2 when $p_1 = p_2$. By extending their constructions it is also possible to show tightness (up to a factor of 16) for all other values in the table. Also, Agarwal et al [2] recently showed lower bounds on the sample complexity of stochastic optimization when $p_1 = \infty$ and p_2 is arbitrary—their lower bounds match the last two rows in the table.

Non-dual Schatten norm pairs in finite dimensions Exactly the same analysis as above can be carried out for Schatten p-norms, i.e. when $\mathcal{W} = B_{S(p_1)}$, $\mathcal{X} = B_{S(p_2)}$ are the unit balls of Schatten p-norm (the p-norm of the singular values) for matrix of dimensions $d_1 \times d_2$. We get the same results as in the table above (as upper bounds on D_2), with $d = \min\{d_1, d_2\}$. These results again follow using similar arguments as ℓ_p case and tight constants for strong convexity parameters of the Schatten norm from [3].

Non-dual group norm pairs in finite dimensions In applications such as multitask learning, groups norms such as $\|\mathbf{w}\|_{q,1}$ are often used on matrices $\mathbf{w} \in \mathbb{R}^{k \times d}$ where $(q, 1)$ norm means taking the ℓ_1-norm of the ℓ_q-norms of the columns of $\mathbf{w}$. Popular choices include $q = 2, \infty$. Here, it may be quite unnatural to use the dual norm (p, ∞) to define the space $\mathcal{X}$ where the data lives. For instance, we might want to consider $\mathcal{W} = B_{(q,1)}$ and $\mathcal{X} = B_{(\infty,\infty)} = B_\infty$. In such a case we can calculate that $D_2(\mathcal{W}, \mathcal{X}) = \Theta(k^{1-\frac{1}{q}}\sqrt{\log(d)})$ using $\Psi(\mathbf{w}) = \frac{1}{q+r-2}\|\mathbf{w}\|_{q,r}^2$ where $r = \frac{\log\ d}{\log\ d-1}$.

Max Norm Max-norm has been proposed as a convex matrix regularizer for application such as matrix completion [21]. In the online version of the matrix completion problem at each time step one element of the matrix is revealed, corresponding to $\mathcal{X}$ being the set of all matrices with a single element being 1 and the rest 0. Since we need $\mathcal{X}$ to be convex we can take the absolute convex hull of this set and use $\mathcal{X}$ to be the unit element-wise ℓ_1 ball. Its dual is $\|W\|_{\mathcal{X}^\star} = \max_{i,j}|W_{i,j}|$. On the other hand given a matrix W, its max-norm is given by $\|W\|_{\max} = \min_{U,V:W=UV^\top}(\max_i \|U_i\|_2)\left(\max_j \|V_j\|_2\right)$. The set $\mathcal{W}$ is the unit ball under the max norm. As noted in [22] the max-norm ball is equivalent, up to a factor two, to the convex hull of all rank one sign matrices. Let us now make a more general observation. Consider any set $\mathcal{W} = \mathrm{abscvx}(\{\mathbf{w}_1, \ldots, \mathbf{w}_K\})$, the absolute convex hull of K points $\mathbf{w}_1, \ldots, \mathbf{w}_K \in \mathcal{B}$. In this case, the Minkowski norm for this $\mathcal{W}$ is given by $\|\mathbf{w}\|_{\mathcal{W}} := \inf_{\alpha_1,\ldots,\alpha_K:\mathbf{w}=\sum_{i=1}^K \alpha_i\mathbf{w}_i}\sum_{i=1}^K|\alpha_i|$. In this case, for any $q \in (1, 2]$, if we define the norm $\|\mathbf{w}\|_{\mathcal{W},q} = \inf_{\alpha_1,\ldots,\alpha_K:\mathbf{w}=\sum_{i=1}^K \alpha_i\mathbf{w}_i}\left(\sum_{i=1}^K|\alpha_i|^q\right)^{1/q}$, then the function $\Psi(\mathbf{w}) = \frac{1}{2(q-1)}\|\mathbf{w}\|_{\mathcal{W},q}^2$ is 2-uniformly convex w.r.t. $\|\cdot\|_{\mathcal{W},q}$ (similar to $\ell_1 - \ell_q$ case). Further if we use $q = \frac{\log K}{\log K - 1}$, then $\sup_{\mathbf{w}\in\mathcal{W}}\sqrt{\Psi(\mathbf{w})} = O(\sqrt{\log K})$ and so $D_2 = \sqrt{\log K}$. For the max norm case the norm is equivalent to the norm got by the taking the absolute convex hull of the set of all rank one sign matrices. Cardinality of this set is of course 2^{N+M}. Hence using the above proposition and noting that $\mathcal{X}^\star$ is the unit ball of $|\cdot|_\infty$ we see that Ψ is obviously 2-uniformly convex w.r.t. $\|\cdot\|_{\mathcal{X}^\star}$ and so we get a regret bound $O\left(\sqrt{\frac{M+N}{n}}\right)$. This matches the stochastic (PAC) learning guarantee [22], and is the first guarantee we are aware of for the max norm matrix completion problem in the online setting.

8 Conclusion and Discussion

In this paper we showed that for a general class of convex online learning problems, there always exists a distance generating function Ψ such that Mirror Descent using this function achieves a near-optimal regret guarantee. This

shows that a fairly simple first-order method, in which each iteration requires a gradient computation and a prox-map computation, is sufficient for online learning in a very general sense. Of course, the main challenge is deriving distance generating functions appropriate for specific problems—although we give two mathematical expressions for such functions, in equations (7) and (9), neither is particularly tractable in general. In the end of Section 6 we do give some general guidelines for choosing the right distance generating function. However obtaining a more explicit and simple procedure at least for reasonable Banach spaces is a very interesting question.

Furthermore, for the Mirror Descent procedure to be efficient, the prox-map of the distance generating function must be efficiently computable, which means that even though a Mirror Descent procedure is always theoretically possible, we might in practice choose to use a non-optimal distance generating function, or even a non-MD procedure. Furthermore, we might also find other properties of $\mathbf{w}$ desirable, such as sparsity, which would bias us toward alternative methods [12, 7]. Nevertheless, in most instances that we are aware of, Mirror Descent, or slight variations of it, is truly an optimal procedure and this is formalized and rigorously establish here.

In terms of the generality of the problems we handle, we required that the constraint set $\mathcal{W}$ be convex, but this seems unavoidable if we wish to obtain efficient algorithms (at least in general). Furthermore, we know that in terms of worst-case behavior, both in the stochastic and in the online setting, for convex cost functions, the value is unchained when the convex hull of a non-convex constraint set [18]. The requirement that the data domain $\mathcal{X}$ be convex is perhaps more restrictive, since even with non-convex data domain, the objective is still convex. Such non-convex $\mathcal{X}$ are certainly relevant in many applications, e.g. when the data is sparse, or when $\mathbf{x} \in \mathcal{X}$ is an indicator, as in matrix completion problems and total variation regularization. In the total variation regularization problem, $\mathcal{W}$ is the set of all functions on the interval $[0, 1]$ with total variation bounded by 1 which is in fact a Banach space. However set $\mathcal{X}$ we consider here is not the entire dual ball and in fact is neither convex nor symmetric. It only consists of evaluations of the functions in $\mathcal{W}$ at points on interval $[0, 1]$ and one can consider a supervised learning problem where the goal is to use the set of all functions with bounded variations to predict targets which take on values in $[-1, 1]$. Although the total-variation problem is not learnable, the matrix completion problem certainly is of much interest. In the matrix completion case, taking the convex hull of $\mathcal{X}$ does not seem to change the value, but we are unaware of neither a guarantee that the value of the game is unchanged when a non-convex $\mathcal{X}$ is replaced by its convex hull, nor of an example where the value does change—it would certainly be useful to understand this issue. We view the requirement that $\mathcal{W}$ and $\mathcal{X}$ be symmetric around the origin as less restrictive and mostly a matter of convenience.

We also focused on a specific form of the cost class $\mathcal{F}$, which beyond the almost unavoidable assumption of convexity, is taken to be constrained through the cost sub-gradients. This is general enough for considering supervised learning with an arbitrary convex loss in a worst-case setting, as the sub-gradients in this case exactly correspond to the data points, and so restricting $\mathcal{F}$ through its sub gradients corresponds to restricting the data domain. Following Proposition 1, any optimality result for $\mathcal{F}_{\mathrm{Lip}}$ also applies to $\mathcal{F}_{\sup}$, and this statement can also be easily extended to any other reasonable loss function, including the hinge-loss, smooth loss functions such as the logistic loss, and even strongly-convex loss functions such as the squared loss (in this context, note that a strongly convex scalar function for supervised learning does *not* translate to a strongly convex optimization problem in the worst case). Going beyond a worst-case formulation of supervised learning, one might consider online repeated games with other constraints on $\mathcal{F}$, such as strong convexity, or even constraints on $\{f_t\}$ as a sequence, such as requiring low average error or conditions on the covariance of the data—these are beyond the scope of the current paper.

Even for the statistical learning setting, online methods along with online to batch conversion are often preferred due to their efficiency especially in high dimensional problems. In fact for ℓ_p spaces in the dual case, using lower bounds on the sample complexity for statistical learning of these problems, one can show that for large dimensional problems, mirror descent is an optimal procedure even for the statistical learning problem. We would like to consider the question of whether Mirror Descent is optimal for stochastic convex optimization (convex statistical learning) setting [9, 19, 23] in general. Establishing such universality would have significant implications, as it would indicate that any learnable (convex) problem, is learnable using a one-pass first-order online method (i.e. Stochastic Approximation approach).

References

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

[2] Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization.

[3] Keith Ball, Eric A. Carlen, and Elliott H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Invent. Math.*, 115:463–482, 1994.

[4] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[5] H. D. Block. The perceptron: A model for brain functioning. *Reviews of Modern Physics*, 34:123–135, 1962. Reprinted in "Neurocomputing" by Anderson and Rosenfeld.

[6] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Sparse and low-rank matrix decompositions. In *IFAC Symposium on System Identification*, 2009.

[7] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[8] Ali Jalali, Pradeep Ravikumar, Sujay Sanghavi, and Chao Ruan. A Dirty Model for Multi-task Learning. In *NIPS*, December 2010.

[9] A. Juditsky, G. Lan, A. Nemirovski, and A. Shapiro. Stochastic approximation approach to stochastic programming. *SIAM J. Optim*, 19(4):1574–1609, 2009.

[10] Sham M. Kakade, Shai Shalev-shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization, 2010.

[11] J. Kivinen and M. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–64, January 1997.

[12] J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. In *Advances in Neural Information Processing Systems 21*, pages 905–912, 2009.

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

[14] A. Nemirovski and D. Yudin. On cesaro's convergence of the gradient descent method for finding saddle points of convex-concave functions. *Doklady Akademii Nauk SSSR*, 239(4), 1978.

[15] A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. Nauka Publishers, Moscow, 1978.

[16] G. Pisier. Martingales with values in uniformly convex spaces. *Israel Journal of Mathematics*, 20(3–4):326–350, 1975.

[17] G. Pisier. Martingales in banach spaces (in connection with type and cotype). *Winter School/IHP Graduate Course*, 2011.

[18] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Random averages, combinatorial parameters, and learnability. *NIPS*, 2010.

[19] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *COLT*, 2009.

[20] S. Shalev-Shwartz and Y. Singer. Convex repeated games and fenchel duality. *Advances in Neural Information Processing Systems*, 19:1265, 2007.

[21] Nathan Srebro, Jason D. M. Rennie, and Tommi S. Jaakola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, pages 1329–1336. MIT Press, 2005.

[22] Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *Proceedings of the 18th Annual Conference on Learning Theory*, pages 545–560. Springer-Verlag, 2005.

[23] Nathan Srebro and Ambuj Tewari. Stochastic optimization for machine learning. In *ICML 2010, tutorial*, 2010.

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

[25] S.Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, Hebrew University of Jerusalem, 2007.

[26] Manfred K. Warmuth and Dima Kuzmin. Randomized online pca algorithms with regret bounds that are logarithmic in the dimension, 2007.

[27] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, 2003.

[28] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.