On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization

Sham M. Kakade and **Shai Shalev-Shwartz** and **Ambuj Tewari**
Toyota Technological Institute—Chicago, USA
{sham,shai,tewari}@tti-c.org

Abstract

We show that a function is strongly convex with respect to some norm if and only if its conjugate function is strongly smooth with respect to the dual norm. This result has already been found to be a key component in deriving and analyzing several learning algorithms. Utilizing this duality, we isolate a single inequality which seamlessly implies both generalization bounds and online regret bounds; and we show how to construct strongly convex functions over matrices based on strongly convex functions over vectors. The newly constructed functions (over matrices) inherit the strong convexity properties of the underlying vector functions. We demonstrate the potential of this framework by analyzing several learning algorithms including group Lasso, kernel learning, and online control with adversarial quadratic costs.

1 Introduction

As we tackle more challenging learning problems, there is an increasing need for algorithms which efficiently impose more sophisticated forms of prior knowledge. Examples include: the group Lasso problem (for "shared" feature selection across problems), kernel learning, multi-class prediction, and multi-task learning. A central question here is to understand the generalization ability of such algorithms in terms of the attendant complexity restrictions imposed by the algorithm (such analyses often illuminate the nature in which our prior knowledge is being imposed).

There is growing body of work suggesting that the notion of *strong* convexity is a fundamental tool in designing and analyzing (the regret or generalization ability of) a wide range of learning algorithms (which we discuss in the next Subsection). The underlying intuition for this is as follows: Most of our efficient algorithms (both in the batch and online settings) impose some complexity control via the use of some *strictly* convex penalty function (either explicitly via a regularizer or implicitly in the design of an online update rule). Central to understanding these algorithms is the manner in which these penalty functions are strictly convex, i.e. the behavior of the "gap" by which these convex functions lie above their tangent planes (which is strictly positive for strictly convex functions). Here, the notion of strong convexity provides one means to characterize this gap in terms of some general norm (rather than just Euclidean).

This work examines the notion of strong convexity from a duality perspective. We show a function is strongly convex with respect to some norm if and only if its (Fenchel) conjugate function is strongly smooth with respect to its dual norm. Roughly speaking, this notion of smoothness (defined precisely later) provides a second order upper bound of the conjugate function, which has already been found to be a key component in deriving and analyzing several learning algorithms. Using this relationship, we are able to characterize a number of matrix based penalty functions, of recent interest, as being strongly convex functions, which allows us to immediately derive online algorithms and generalization bounds when using such functions.

We now briefly discuss related work and our contributions.

1.1 Related work

The notion of strong convexity takes its roots in optimization based on ideas in Nemirovski and Yudin [1978] (where it was defined with respect to the Euclidean norm) — the generalization to arbitrary norms was by Nesterov [2005]. Relatively recently, its use in machine learning has been two fold: in deriving regret bounds for online algorithm and generalization bounds in batch settings.

The duality of strong convexity and strong smoothness was first used by Shalev-Shwartz and Singer [2006], Shalev-Shwartz [2007] in the context of deriving low regret online algorithms. Here, once we choose a particular strongly convex penalty function, we immediately have a family of algorithms along with a regret bound for these algorithms that is in terms of a certain strong convexity parameter. A variety of algorithms (and regret bounds) can be seen as special cases.

A similar technique, in which the Hessian is directly bounded, is described by Grove et al. [2001], Shalev-Shwartz and Singer [2007]. Another related approach involved bounding a Bregman divergence [Kivinen and Warmuth, 1997, 2001, Gentile, 2002] (see Cesa-Bianchi and Lugosi [2006] for a detailed survey). Another interesting application of the very same duality is for deriving and analyzing boosting algorithms [Shalev-Shwartz and Singer, 2008].

More recently, Kakade et al. [2008] showed how to use the very same duality for bounding the Rademacher complexity of classes of linear predictors. That the Rademacher complexity is closely related to Fenchel duality was shown in Meir and Zhang [2003], and the work in Kakade et al. [2008] made the further connection to strong convexity. Again, under this characterization, a number of generalization and margin bounds (for methods which use linear prediction) are immediate corollaries, as one only needs to specify the strong convexity parameter from which these bounds easily follow (see Kakade et al. [2008] for details).

The concept of strong smoothness (essentially a second order upper bound on a function) has also been in play in a different literature, for the analysis of the concentration of martingales in *smooth* Banach spaces [Pinelis, 1994, Pisier, 1975]. This body of work seeks to understand the concentration properties of a random variable $||X_t||$, where X_t is a (vector valued) martingale and $||\cdot||$ is a smooth norm, say an L_p-norm.

Recently, Juditsky and Nemirovski [2008] proved that a *norm* is strongly convex if and only if its conjugate is strongly smooth. This duality was useful in deriving concentration properties of a random variable $||M||$, where now M is a random matrix. The norms considered here were the (Schatten) L_p-matrix norms (where $||M||_p$ is the L_p norm of the singular values of M) and certain "block" composed norms (such as the $||\cdot||_{2,q}$ norm).

1.2 Our Contributions

The first contribution of this paper is to further distill this theory of strong convexity. While Shalev-Shwartz [2007] have shown that strong convexity (of general functions) implies strong smoothness, here we show that the other direction also holds and thus the two notions are equivalent. This result generalizes the recent results of Juditsky and Nemirovski [2008] to functions rather than norms. This generalization has a number of consequences which this work explores. For example, in Corollary 7, we isolate an important inequality that follows from the strong-convexity/strong-smoothness duality and show that this inequality alone seamlessly yields regret bounds and Rademacher bounds.

The second contribution of this paper is in deriving new families of strongly convex (smooth) functions. To do so, we rely and further generalize the recent results of Juditsky and Nemirovski [2008]. In particular, we obtain a strongly convex function over matrices based on strongly convex vector functions, which leads to a number of corollaries relevant to problems of recent interest.

Furthermore, this characterization allows us to place a wider class of online algorithms (along with regret bounds) as special cases of the general primal-dual framework developed in Shalev-Shwartz [2007]. Examples which are now immediate corollaries include: online PCA [Warmuth and Kuzmin, 2006], the perceptron algorithm derived with a Schatten norm complexity function [Cavallanti et al., 2008], and the multi-task algorithm of Agarwal et al. [2008]. These corollaries follow once we characterize a certain strong convexity parameter (along with the derivative of the conjugate function) — here, a family of online algorithms all enjoy the same regret bound, with no further analysis required.

Finally, we use the generality of our results for obtaining new (and sharper) generalization bounds for various applications, including the group Lasso and kernel learning. In the former, we are able to show how the $||\cdot||_{2,1}$ "group" norm enjoys certain (shared) feature selection properties (with only logarithmic dependence on the number of features). In the latter, we show how kernel learning (learning a kernel as a convex combination of base kernels) has only a mild dependence on the number of base kernels used (only logarithmic).

2 The duality of strong convexity and strong smoothness

2.1 Preliminaries

Here, we briefly recall some key definitions from convex analysis that are useful throughout the paper (for details, see any of the several excellent references on the subject, e.g. Borwein and Lewis [2006], Rockafellar [1970]).

We consider convex functions $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$, where $\mathcal{X}$ is a Euclidean vector space equipped with an inner product $\langle \cdot, \cdot \rangle$. We denote $\mathbb{R}^* = \mathbb{R} \cup \{\infty\}$.

Definition 1 *Given a convex function $f : \mathcal{X} \to \mathbb{R}^*$, its subdifferential at $x \in \mathcal{X}$, denoted by $\partial f(x)$, is defined as,*

$$\partial f(x) := \{y \in \mathcal{X} \ : \ \forall z, \ f(x+z) \geq f(x) + \langle y, z \rangle\}$$

Definition 2 *Given a convex function $f : \mathcal{X} \to \mathbb{R}^*$, the Fenchel conjugate $f^\star : \mathcal{X} \to \mathbb{R}^*$ is defined as*

$$f^\star(y) := \sup_{x \in \mathcal{X}} \langle x, y \rangle - f(x)$$

We also deal with a variety of norms in this paper. Recall the definition of the dual norm.

Definition 3 *Given a norm $\|\cdot\|$ on $\mathcal{X}$, its dual $\|\cdot\|_\star$ is the norm (also on $\mathcal{X}$) defined as,*

$$\|y\|_\star := \sup\{\langle x, y \rangle \ : \ \|x\| \leq 1\}$$

An important property of the dual norm is that the Fenchel conjugate of the function $\frac{1}{2}\|x\|^2$ is $\frac{1}{2}\|y\|_\star^2$.

The definition of Fenchel conjugate implies

$$\forall x, y, \ f(x) + f^\star(y) \geq \langle x, y \rangle \ ,$$

which is known as the Fenchel-Young inequality. An equivalent and useful definition of the subdifferential can be given in terms of the Fenchel conjugate,

$$\partial f(x) = \{y \in \mathcal{X} \ : \ f(x) + f^*(y) = \langle x, y \rangle\}$$

2.2 Main result

Recall that the domain of a function $f : \mathcal{X} \to \mathbb{R}^*$ is the set of x such that $f(x) < \infty$ (allowing f to take infinite values is the effective way to restrict its domain to a proper subset of $\mathcal{X}$). We first define strong convexity.

Definition 4 *A function $f : \mathcal{X} \to \mathbb{R}^*$ is β-strongly convex w.r.t. a norm $\|\cdot\|$ if for all x, y in the relative interior of the domain of f and $\alpha \in (0,1)$ we have*

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) - \tfrac{1}{2}\beta\alpha(1-\alpha)\|x-y\|^2$$

We now define strong smoothness. Note that a strongly smooth function f is always finite.

Definition 5 *A function $f : \mathcal{X} \to \mathbb{R}$ is β-strongly smooth w.r.t. a norm $\|\cdot\|$ if f is everywhere differentiable and if for all x, y we have*

$$f(x+y) \leq f(x) + \langle \nabla f(x), y \rangle + \frac{1}{2}\beta\|y\|^2$$

The following central theorem shows that strong convexity and strong smoothness are dual properties. Recall that the biconjugate $f^{\star\star}$ equals f if and only if f is closed and convex.

Theorem 6 *(Strong/Smooth Duality) Assume that f is a closed and convex function. Then f is β-strongly convex w.r.t. a norm $\|\cdot\|$ if and only if $f^\star$ is $\frac{1}{\beta}$-strongly smooth w.r.t. the dual norm $\|\cdot\|_\star$.*

Subtly, note that while the domain of a strongly convex function f may be a proper subset of $\mathcal{X}$ (important for a number of settings), its conjugate $f^\star$ always has a domain which is $\mathcal{X}$ (since $f^\star$ is strongly smooth then it is finite and everywhere differentiable).

The proof is provided in the appendix.

2.3 Machine learning implications of the strong-convexity / strong-smoothness duality

The following direct corollary of Thm. 6 is central in proving both regret and generalization bounds.

Corollary 7 *If f is β strongly convex w.r.t. $\|\cdot\|$ and $f^\star(\mathbf{0}) = 0$, then, for any sequence $v_1, \dots, v_n$ and for any u we have*

$$\begin{aligned}\sum_{i=1}^{n} \langle v_i, u \rangle - f(u) &\leq f^\star(v_{1:n}) \\ &\leq \sum_{i=1}^{n} \langle \nabla f^\star(v_{1:i-1}), v_i \rangle + \frac{1}{2\beta}\sum_{i=1}^{n}\|v_i\|_\star^2\end{aligned}$$

where $v_{1:i}$ denotes the sum $\sum_{j=1}^{i} v_j$.

Proof: The first inequality is Fenchel-Young and the second is from the definition of smoothness by induction. ■

From this we can easily obtain regret bounds and Rademacher bounds.

2.3.1 Regret Bound

Algorithm 1 provides one common algorithm (Follow the Regularized Leader) which achieves the following regret bound. It is one of a family of algorithms which enjoys the same regret bound (see Shalev-Shwartz [2007]).

Theorem 8 *(Regret) Suppose Algorithm 1 is used with a function f that is β-strongly convex w.r.t. a norm $\|\cdot\|$ on S and has $f^\star(\mathbf{0}) = 0$. Suppose the loss functions l_t are convex and V-Lipschitz w.r.t. the dual norm $\|\cdot\|_\star$. Then, the algorithm run with any positive η enjoys the regret bound,*

$$\sum_{t=1}^{T} l_t(w_t) - \min_{u\in S}\sum_{t=1}^{T} l_t(u) \leq \frac{\max_{u\in S} f(u)}{\eta} + \frac{\eta V^2 T}{2\beta}$$

Algorithm 1 Follow the Regularized Leader

```
$w_1 \leftarrow \nabla f^\star(\mathbf{0})$
for $t = 1$ to $T$ do
    Play $w_t \in S$
    Receive $l_t$ and pick $v_t \in \partial l_t(w_t)$
    $w_{t+1} \leftarrow \nabla f^\star\left(-\eta \sum_{s=1}^{t} v_t\right)$
end for
```

Proof: Apply Corollary 7 to the sequence $-\eta v_1, \dots, -\eta v_T$ to get, for all u,

$$-\eta\sum_{t=1}^{T}\langle v_t, u\rangle - f(u) \leq -\eta\sum_{t=1}^{T}\langle v_t, w_t\rangle + \frac{1}{2\beta}\sum_{t=1}^{T}\|\eta v_t\|_\star^2 .$$

Using the that l_t is V-Lipschitz, we get $\|v_t\|_\star \leq V$. Plugging this into the inequality above and rearranging gives,

$$\sum_{t=1}^{T}\langle v_t, w_t - u\rangle \leq \frac{f(u)}{\eta} + \frac{\eta V^2 T}{2\beta} .$$

By convexity of l_t, $l_t(w_t) - l_t(u) \leq \langle v_t, w_t - u\rangle$. Therefore,

$$\sum_{t=1}^{T} l_t(w_t) - \sum_{t=1}^{T} l_t(u) \leq \frac{f(u)}{\eta} + \frac{\eta V^2 T}{2\beta} .$$

Take the max over $u \in S$ on both sides to finish the proof. ■

2.3.2 Rademacher Bound

Let $\mathcal{X}$ be an input space. Let $\mathcal{T} = (X_1, \dots, X_n)$ be a dataset consisting of i.i.d. samples from some fixed distribution on $\mathcal{X}$. For a class of real valued functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$, define its Rademacher complexity on $\mathcal{T}$ to be

$$\mathcal{R}_{\mathcal{T}}(\mathcal{F}) := \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right] .$$

Here, the expectation is over ϵ_i's, which are i.i.d. Rademacher random variables, i.e. $\mathbb{P}(\epsilon_i = -1) = \mathbb{P}(\epsilon_1 = +1) = \frac{1}{2}$. Since $\mathcal{T}$ is random, this is a random variable. We can also take expectation over the choice of $\mathcal{T}$ and define,

$$\mathcal{R}_n(\mathcal{F}) := \mathbb{E}\left[\mathcal{R}_{\mathcal{T}}(\mathcal{F})\right]$$

which gives us a number that is a function of the input space, the function class and the sample size n. It is well known that bounds on Rademacher complexity of a class immediately yield generalization bounds for classifiers picked from that class. Recently, Kakade et al. [2008] proved Rademacher complexity bounds for classes consisting of linear predictors using strong convexity arguments. We now give a quick proof of their main result using Corollary 7. This proof is essentially the same as their original proof but highlights the importance of Corollary 7.

Theorem 9 *(Generalization) Let f be a β-strongly convex function w.r.t. a norm $\|\cdot\|$ on S such that $f^\star(\mathbf{0}) = 0$. Let*

$\mathcal{X} = \{x : \|x\|_\star \le X\}$ and $\mathcal{W} = \{w : f(w) \le f_{\max}\}$. Consider the class of linear functions,

$$\mathcal{F} = \{x \mapsto \langle w, x\rangle : w \in \mathcal{W}\} .$$

Then, for any dataset $\mathcal{T} \in \mathcal{X}^n$, we have

$$\mathcal{R}_{\mathcal{T}}(\mathcal{F}) \le X\sqrt{\frac{2f_{\max}}{\beta n}} .$$

Therefore, the same bound holds for $\mathcal{R}_n(\mathcal{F})$.

Proof: Let $\lambda > 0$. Apply Corollary 7 with $u = w$ and $v_i = \lambda\epsilon_i X_i$ to get,

$$\begin{aligned}\sup_{w\in\mathcal{W}} \sum_{i=1}^n \langle w, \lambda\epsilon_i X_i\rangle &\le \frac{\lambda^2}{2\beta}\sum_{i=1}^n \|\epsilon_i X_i\|_\star^2 + \sup_{w\in\mathcal{W}} f(w) \\ &\quad + \sum_{i=1}^n \langle \nabla f^\star(v_{1:i-1}), \epsilon_i X_i\rangle \\ &\le \frac{\lambda^2 X^2 n}{2\beta} + f_{\max} \\ &\quad + \sum_{i=1}^n \langle \nabla f^\star(v_{1:i-1}), \epsilon_i X_i\rangle .\end{aligned}$$

Now take expectation on both sides. The left hand side is $n\lambda\mathcal{R}_{\mathcal{T}}(\mathcal{F})$ and the last term above becomes zero. Dividing throughout by $n\lambda$, we get,

$$\mathcal{R}_{\mathcal{T}}(\mathcal{F}) \le \frac{\lambda X^2}{2\beta} + \frac{f_{\max}}{n\lambda} .$$

Optimizing over λ gives us the result. ■

3 Examples of strongly convex matrix functions

We now provide examples of strongly convex functions over matrices, which have a number of algorithmic implications. We begin by understanding the strong convexity properties of functions which only depend on the singular values of the matrix — this class includes the Schatten norms. We then turn to understanding norms of matrices which constructed in a certain "group" manner, where a norm is first applied to each column of the matrix (to obtain a vector) and then a norm is applied to this resultant vector. This class for norms include the $||\cdot||_{2,1}$ norm of recent interest (e.g. for the group Lasso).

We start by presenting tools useful for analyzing matrices, borrowing heavily from Lewis [1995] and Juditsky and Nemirovski [2008]. In fact, Juditsky and Nemirovski [2008] already proved the strong smoothness for Schatten p-norms. We provide additional results for the entropy based matrix functions. Also, our results on the group norms are more general that those in Juditsky and Nemirovski [2008].

3.1 Convex analysis of matrix functions

We consider the vector space $\mathcal{X} = \mathbb{R}^{m\times n}$ of real matrices of size $m \times n$ and the vector space $\mathcal{X} = \mathbb{S}^n$ of symmetric matrices of size $n \times n$, both equipped with the inner product,

$$\langle X, Y\rangle := \mathrm{Tr}(X^\top Y) .$$

Recall that any matrix $X \in \mathbb{R}^{m\times n}$ can be decomposed as,

$$X = U\mathrm{Diag}(\sigma(X))V$$

where $\sigma(X)$ denotes the vector $(\sigma_1, \sigma_2, \ldots \sigma_l)$ ($l = \min\{m, n\}$), where $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_l \ge 0$ are the singular values of X arranged in non-increasing order, and $U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n}$ are orthogonal matrices. Also, any matrix $X \in \mathbb{S}^n$ can be decomposed as,

$$X = U\mathrm{Diag}(\lambda(X))U^\top$$

where $\lambda(X) = (\lambda_1, \lambda_2, \ldots \lambda_n)$, where $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_n$ are the eigenvalues of X arranged in non-increasing order, and U is an orthogonal matrix. Two important results relate matrix inner products to inner products between singular (and eigen-) values

Theorem 10 (von Neumann) *Any two matrices $X, Y \in \mathbb{R}^{m\times n}$ satisfy the inequality*

$$\langle X, Y\rangle \le \langle \sigma(X), \sigma(Y)\rangle .$$

Equality holds above, if and only if, there exist orthogonal U, V such that

$$X = U\mathrm{Diag}(\sigma(X))V \qquad Y = U\mathrm{Diag}(\sigma(Y))V .$$

Theorem 11 (Fan) *Any two matrices $X, Y \in \mathbb{S}^n$ satisfy the inequality*

$$\langle X, Y\rangle \le \langle \lambda(X), \lambda(Y)\rangle .$$

Equality holds above, if and only if, there exists orthogonal U such that

$$X = U\mathrm{Diag}(\lambda(X))U^\top \qquad Y = U\mathrm{Diag}(\lambda(Y))U^\top .$$

We say that a function $g : \mathbb{R}^n \to \mathbb{R}^*$ is symmetric if $g(x)$ is invariant under arbitrary permutations of the components of x. We say g is absolutely symmetric if $g(x)$ is invariant under arbitrary permutations and sign changes of the components of x.

Given a function $f : \mathbb{R}^l \to \mathbb{R}^*$, we can define a function $f \circ \sigma : \mathbb{R}^{m\times n} \to \mathbb{R}^*$ as,

$$(f \circ \sigma)(X) := f(\sigma(X)) .$$

Similarly, given a function $g : \mathbb{R}^n \to \mathbb{R}^*$, we can define a function $g \circ \lambda : \mathbb{S}^n \to \mathbb{R}^*$ as,

$$(g \circ \lambda)(X) := g(\lambda(X)) .$$

This allows us to define functions over matrices starting from functions over vectors. Note that when we use $f \circ \sigma$ we are assuming that $\mathcal{X} = \mathbb{R}^{m\times n}$ and for $g \circ \lambda$ we have $\mathcal{X} = \mathbb{S}^n$. The following result allows us to immediately compute the conjugate of $f \circ \sigma$ and $g \circ \lambda$ in terms of the conjugates of f and g respectively.

Theorem 12 *(Lewis [1995]) Let $f : \mathbb{R}^l \to \mathbb{R}^*$ be an absolutely symmetric function. Then,*

$$(f \circ \sigma)^\star = f^\star \circ \sigma .$$

Let $g : \mathbb{R}^n \to \mathbb{R}^*$ be a symmetric function. Then,

$$(g \circ \lambda)^\star = g^\star \circ \lambda .$$

Proof: Lewis [1995] proves the case for singular values. For the eigenvalue case, the proof is entirely analogous to that in Lewis [1995], except that Fan's inequality is used instead of von Neumann's inequality. ■

Using this general result, we are able to define certain matrix norms.

Corollary 13 *(Matrix norms) Let $f : \mathbb{R}^l \to \mathbb{R}^*$ be absolutely symmetric. Then if $f = \|\cdot\|$ is a norm on $\mathbb{R}^l$ then $f \circ \sigma = \|\sigma(\cdot)\|$ is a norm on $\mathbb{R}^{m\times n}$. Further, the dual of this norm is $\|\sigma(\cdot)\|_\star$.*

Let $g : \mathbb{R}^n \to \mathbb{R}^*$ be symmetric. Then if $g = \|\cdot\|$ is a norm on $\mathbb{R}^n$ then $g \circ \lambda = \|\lambda(\cdot)\|$ is a norm on $\mathbb{S}^n$. Further, the dual of this norm is $\|\lambda(\cdot)\|_\star$.

Another nice result allows us to compute subdifferentials of $f \circ \sigma$ and $g \circ \lambda$ (note that elements in the subdifferential of $f \circ \sigma$ and $g \circ \lambda$ are matrices) from the subdifferentials of f and g respectively.

Theorem 14 *(Lewis [1995]) Let $f : \mathbb{R}^l \to \mathbb{R}^*$ be absolutely symmetric and $X \in \mathbb{R}^{m\times n}$. Then,*

$$\partial(f \circ \sigma)(X) = \{U\mathrm{Diag}(\mu)V^\top \ : \ \mu \in \partial f(\sigma(X)) \\ U, V \ \textit{orthogonal}, \ X = U\mathrm{Diag}(\sigma(X))V^\top\}$$

Let $g : \mathbb{R}^n \to \mathbb{R}^*$ be symmetric and $X \in \mathbb{S}^n$. Then,

$$\partial(g \circ \lambda)(X) = \{U\mathrm{Diag}(\mu)U^\top \ : \ \mu \in \partial g(\lambda(X)) \\ U \ \textit{orthogonal}, \ X = U\mathrm{Diag}(\lambda(X))U^\top\}$$

Proof: Again, Lewis [1995] proves the case for singular values. For the eigenvalue case, again, the proof is identical to that in Lewis [1995], except that Fan's inequality is used instead of von Neumann's inequality. ■

Our final tool is a technical result from Juditsky and Nemirovski [2008].

Lemma 15 *(Juditsky and Nemirovski [2008]) Let Δ be an open interval. Suppose $\phi : \Delta \to \mathbb{R}^*$ is a twice differentiable convex function such that ϕ'' is monotonically non-decreasing. Let $\mathbb{S}_n(\Delta)$ be the set of all symmetric $n \times n$ matrices with eigenvalues in Δ. Define the function $F : \mathbb{S}^n(\Delta) \to \mathbb{R}^*$*

$$F(X) = \sum_{i=1}^n \phi(\lambda_i(X))$$

and let

$$f(t) = F(X + tH)$$

for some $X \in \mathbb{S}^n(\Delta), H \in \mathbb{S}^n$. Then, we have,

$$f''(0) \le 2\sum_{i=1}^n \phi''(\lambda_i(X))\lambda_i(H)^2 \ .$$

Proof: This follows directly from Proposition 3.1 in Juditsky and Nemirovski [2008]. ■

3.2 Strongly convex matrix functions

We first provide results on functions which only depend on the singular values of a matrix and then provide results on group norms.

Unitarily invariant matrix functions. Our first result is on the p-Schatten norm $\|X\|_{S(p)} := \|\sigma(X)\|_p$ (which follows from results in Juditsky and Nemirovski [2008]) and on an entropy-based matrix function.

Theorem 16 *(Schatten and entropic matrix functions)*

- *Define $F(X) = \sum_i \lambda_i(X)\log(\lambda_i(X))$ on its domain:*
  $$\{X \in \mathbb{S}^n \ : \ X \succeq 0, \ \mathrm{Tr}(X) = 1\},$$
  i.e. the set of symmetric positive semidefinite matrices with trace 1, and $F(X) = \infty$ elsewhere (on $\mathbb{S}^n$). We have that $F(X)$ is $1/2$-strongly convex w.r.t. the trace norm $\|\lambda(X)\|_1$.
- *For $p \in [1,2]$, the function $F(X) = \frac{1}{2}\|\sigma(X)\|_p^2$ is $\min\{\frac{1}{2}, p-1\}$-strongly convex w.r.t. the p-Schatten norm $\|X\|_{S(p)} := \|\sigma(X)\|_p$.*

Proof: For the first part, we prove that the function $(g \circ \lambda)(X)$ is 2-smooth on $\mathbb{S}_n$ w.r.t. $\|\lambda(X)\|_\infty$ where

$$g(\mathbf{x}) = \log\left(\sum_{i=1}^n \exp(x_i)\right) \ .$$

Since g is symmetric, by Thm. 12, $(g \circ \lambda)^\star$ is $g^\star \circ \lambda$, where $g^\star$ can be shown to be the function

$$g^\star(\mathbf{x}) = \sum_{i=1}^n x_i \log x_i$$

with domain $\{\mathbf{x} \ge \mathbf{0} \ : \ \sum_i x_i = 1\}$ and $g^\star(\mathbf{x}) = \infty$ elsewhere. Note that by Thm. 6, 2-smoothness of $(g \circ \lambda)$ implies $1/2$-strong convexity of $(g \circ \lambda)^\star$.

Let us now prove 2-smoothness of $g \circ \lambda$. Fix arbitrary $X, H \in \mathbb{S}^n$, and define

$$f(t) = \sum_{i=1}^n \exp(\lambda_i(X + tH))$$

and let $h(t) = \log(f(t))$. Note that $h(t) = (g \circ \lambda)(X + tH)$. To prove 2-smoothness of $g \circ \lambda$, it suffices to prove

$$h''(0) \le 2\|\lambda(H)\|_\infty^2 \ .$$

By the chain rule,

$$h''(t) = -\frac{(f'(t))^2}{f(t)^2} + \frac{f''(t)}{f(t)} \ .$$

The first term in non-positive and therefore $h''(0) \le f''(0)/f(0)$. By Lemma 15 (with $\phi(x) = \exp(x)$),

$$\begin{aligned} f''(0) &\le 2\sum_{i=1}^n \exp(\lambda_i(X))\lambda_i(H)^2 \\ &\le 2\|\lambda(H)\|_\infty^2 \sum_{i=1}^n \exp(\lambda_i(X)) \\ &= 2\|\lambda(H)\|_\infty^2 f(0) \ , \end{aligned}$$

whence $h''(0) \le f''(0)/f(0) \le 2\|\lambda(H)\|_\infty^2$.

For the second part, let q be the dual exponent of p, i.e. $1/q + 1/p = 1$. Note that $\|\sigma(X)\|_p$ and $\|\sigma(X)\|_q$ are dual norms by Corollary 13. Now we use the result [Juditsky and Nemirovski, 2008, Example 3.3] which says that $\frac{1}{2}\|\sigma(X)\|_q^2$ is $\max\{2, q-1\}$-smooth w.r.t. $\|\sigma(X)\|_q$. Hence, by Thm. 6, $\frac{1}{2}\|\sigma(X)\|_p^2$ is $\min\{\frac{1}{2}, p-1\}$-strongly convex w.r.t. $\|\sigma(X)\|_p$. ■

Group Norms. Let $X = (X_1 X_2 \ldots X_n)$ be a $m \times n$ real matrix with columns $X_i \in \mathbb{R}^m$. Given norms Ψ and Φ on $\mathbb{R}^m$ and $\mathbb{R}^n$, we define the norm $\|X\|_{\Psi,\Phi}$ as

$$\|X\|_{\Psi,\Phi} := \Phi(\Psi(X_1), \ldots, \Psi(X_n)) \ .$$

That is, we apply Ψ to each column of X to get a vector in $\mathbb{R}^n$ to which we apply the norm Φ to get the value of $\|X\|_{\Psi,\Phi}$. It is easy to check that this is indeed a norm.

An important special case is when $\Phi = \|\cdot\|_r$ and $\Psi = \|\cdot\|_p$ for $r, p \ge 1$. In this case, we denote the norm $\|\cdot\|_{\Psi,\Phi}$ by $\|\cdot\|_{p,r}$.

The dual of $\|\cdot\|_{\Psi,\Phi}$ is also easily calculated from the duals $\Psi_\star$ and $\Phi_\star$ of Ψ and Φ under a mild condition on Φ.

Lemma 17 *Let Φ be an absolutely symmetric norm on $\mathbb{R}^n$. Then*

$$(\|\cdot\|_{\Psi,\Phi})_\star = \|\cdot\|_{\Psi_\star,\Phi_\star}$$

Proof: See Sec. A.2 in the Appendix. ■

We now state our main theorem for group norms. A special case of this theorem (when $\Phi = \|\cdot\|_s, s \ge 2$) appeared in Juditsky and Nemirovski [2008]. We not only generalize their result but also provide a simpler proof.

Theorem 18 *(Group Norms) Let Ψ, Φ be absolutely symmetric norms on $\mathbb{R}^m, \mathbb{R}^n$. Let $\Phi^2 \circ \surd : \mathbb{R}^n \to \mathbb{R}^*$ denote the following function,*

$$(\Phi^2 \circ \surd)(x) := \Phi^2(\sqrt{x_1}, \ldots, \sqrt{x_n}) \ .$$

Suppose, $(\Phi^2 \circ \surd)$ is a norm on $\mathbb{R}^n$. Further, let the functions Ψ^2 and Φ^2 be σ_1- and σ_2-smooth w.r.t. Ψ and Φ respectively. Then, $\|\cdot\|_{\Psi,\Phi}^2$ is $(\sigma_1 + \sigma_2)$-smooth w.r.t. $\|\cdot\|_{\Psi,\Phi}$.

Proof: See Sec. A.2 in the Appendix. ■

Lemma 17 implies that $(\|\cdot\|_{p,r})_\star$ is $\|\cdot\|_{q,s}$ where $1/p + 1/q = 1$ and $1/r + 1/s = 1$. Moreover, when $s \ge 2$, $\|\cdot\|_s^2 \circ \surd$ is simply $\|\cdot\|_{\frac{s}{2}}$ which is a norm. Thm. 18 now gives us the following corollary.

Corollary 19 *Let $q, s \ge 2$. The function $\frac{1}{2}\|\cdot\|_{q,s}^2$ is $(q+s-2)$-smooth w.r.t. $\|\cdot\|_{q,s}$ on $\mathbb{R}^{m\times n}$.*

4 Applications

The potential of this framework is that once we characterize the β-strong convexity of our penalty function F (e.g over matrices or other abstract convex sets), then we often immediately obtain both a family of online algorithms (along with their regret bounds) and generalization bounds. In fact, for matrix based penalty functions, a number of dedicated previous algorithms/regret bounds are now special cases of the results herein (using the family of algorithms described in Shalev-Shwartz [2007]), including online PCA [Warmuth and Kuzmin, 2006], the perceptron algorithm derived with a Schatten norm [Cavallanti et al., 2008], and the multi-task algorithm (using the $||\cdot||_{2,1}$ group norm) of Agarwal et al. [2008]. Note that in order to derive an online algorithm with penalty F (e.g. as in Algorithm 1), we must specify $\nabla F^\star$, which is often straightforward to compute using the calculus of certain matrix functions discussed in Subsection 3.1.

We now demonstrate how to obtain a few generalization and regret bounds for problems of recent interest.

4.1 Group Lasso

Consider the setting of k-multivariate regression or classification problems, where the dataset consists of i.i.d. pairs $(\mathbf{x}_i, \mathbf{y}_i)$ where $\mathbf{x}_i \in \mathbb{R}^d$ is an example vector and $\mathbf{y}_i \in \mathbb{R}^k$ are the responses for k different problems. To predict the k responses, we learn a matrix $W \in \mathbb{R}^{k\times d}$ such that $W\mathbf{x}$ is a good predictor of $\mathbf{y}$. The rows $W_{i,\cdot}$ $(1 \le i \le k)$ are the linear predictors for the individual problems. If the same features are going to be relevant across the k problems, then natural *block regularization* schemes have been already proposed in the literature [Yuan and Lin, 2006]. With the squared loss these schemes try to solve an optimization problem of the form,

$$\min_W \frac{1}{n}\sum_{i=1}^n \|\mathbf{y}_i - W\mathbf{x}_i\|_2^2 + \lambda\|W\|_{p,r} \qquad (1)$$

for some $p, r \in [1, 2]$ and $\lambda > 0$. For some choices of (p, r) there is no coupling across problems. For example, the choices $(2, 2)$ and $(1, 1)$ exactly correspond to solving k independent L_2- and L_1-regularized problems respectively. However, for other choices, we get more interesting coupling of the k problems. For example, the group Lasso choice sets $p = 2$ and $r = 1$. That is, we take the L_2-norm of the k columns of W and add them up.

Let us focus on the constrained form of the group Lasso problem Eq. (1),

$$\min_W \frac{1}{n}\sum_{i=1}^n \|\mathbf{y}_i - W\mathbf{x}_i\|_2^2 \ \text{ s.t.} \qquad \|W\|_{2,1} \le \bar{W}_{2,1}$$

for some $\bar{W}_{2,1} > 0$.

In order to obtain generalization bounds for the solution of this problem, we need to control the Rademacher complexity of the function class,

$$\mathcal{F} = \{(\mathbf{x}, \mathbf{y}) \mapsto \|\mathbf{y} - W\mathbf{x}\|_2^2 \ : \ \|W\|_{2,1} \le \bar{W}_{2,1}\} \ . \qquad (2)$$

Theorem 20 *(Group Lasso) Let the distribution of $\mathbf{x}, \mathbf{y}$ be such that $\|\mathbf{x}\|_\infty \le X_\infty$ and $\|\mathbf{y}\|_2 \le Y_2$ a.s. Then, for the class defined above in Eq. (2), we have*

$$\mathcal{R}_n(\mathcal{F}) \le \frac{\left(Y_2 + e\bar{W}_{2,1}X_\infty\sqrt{\log d}\right)^2}{\sqrt{n}}$$

Note that this bound shows feature selection properties of the group Lasso, in the following sense: if there are q relevant (shared) features across all problems (whose weights are bounded), then the above bound scales as $O(\frac{q^2 \log d}{\sqrt{n}})$. The above bound directly leads to a generalization bound.

Proof: We have

$$\begin{aligned}\|\mathbf{y} - W\mathbf{x}\|_2^2 &= \mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top W\mathbf{x} + \mathbf{x}^\top W^\top W\mathbf{x} \\ &= \mathbf{y}^\top\mathbf{y} - 2\mathrm{Tr}\left(\mathbf{y}^\top W\mathbf{x}\right) + \mathrm{Tr}\left(\mathbf{x}^\top W^\top W\mathbf{x}\right) \\ &= \mathbf{y}^\top\mathbf{y} - 2\mathrm{Tr}\left(\mathbf{x}\mathbf{y}^\top W\right) + \mathrm{Tr}\left(W^\top W\mathbf{x}\mathbf{x}^\top\right) \\ &= \mathbf{y}^\top\mathbf{y} - 2\left\langle \mathbf{y}\mathbf{x}^\top, W\right\rangle + \left\langle W^\top W, \mathbf{x}\mathbf{x}^\top\right\rangle\end{aligned}$$

where the inner products appearing in the last line are matrix inner products. Now consider the classes:

$$\mathcal{F}_1 = \{(\mathbf{x},\mathbf{y}) \mapsto 2\left\langle W, \mathbf{y}\mathbf{x}^\top\right\rangle \;:\; \|W\|_{2,1} \le \bar{W}_{2,1}\}\;,$$

$$\mathcal{F}_2 = \{(\mathbf{x},\mathbf{y}) \mapsto \left\langle W^\top W, \mathbf{x}\mathbf{x}^\top\right\rangle \;:\; \|W\|_{2,1} \le \bar{W}_{2,1}\}\;.$$

It is straightforward to show:

$$\mathcal{R}_n(\mathcal{F}) \le \mathcal{R}_n(\mathcal{F}_1) + \mathcal{R}_n(\mathcal{F}_2)\;. \tag{3}$$

For $\mathcal{F}_1$, we use Thm. 9 with $\|\cdot\| = \|\cdot\|_{2,r}$ for $r \in (1,2]$ and $f(W) = \frac{1}{2}\|W\|_{2,r}^2$. Let $1/r + 1/s = 1$, so that $s \in [2,\infty)$. By Corollary 19, $\frac{1}{2}\|\cdot\|_{2,s}^2$ is s-smooth. Hence, by Thm. 6, its conjugate $\frac{1}{2}\|\cdot\|_{2,r}^2$ is $1/s$-strongly convex. Moreover $\|\mathbf{y}\mathbf{x}^\top\|_{2,s} \le d^{1/s}Y_2X_\infty$. Now, Thm. 9 gives us,

$$\mathcal{R}_n(\mathcal{F}_1) \le 2d^{1/s}Y_2X_\infty\bar{W}_{2,1}\sqrt{\frac{s}{n}}\;.$$

Setting $s = \log d$ gives,

$$\mathcal{R}_n(\mathcal{F}_1) \le 2eY_2X_\infty\bar{W}_{2,1}\sqrt{\frac{\log d}{n}}\;. \tag{4}$$

For $\mathcal{F}_2$, note that

$$\begin{aligned}\|W^\top W\|_{1,1} &= \sum_{i,j} |\left\langle W_{\cdot,i}, W_{\cdot,j}\right\rangle| \\ &\le \sum_{i,j} \|W_{\cdot,i}\|_2 \cdot \|W_{\cdot,j}\|_2 \\ &= \sum_i \|W_{\cdot,i}\|_2 \cdot \sum_j \|W_{\cdot,j}\|_2 \\ &= \|W\|_{2,1}^2\;.\end{aligned}$$

Also, $\|\mathbf{x}\mathbf{x}^\top\|_{\infty,\infty} \le X_\infty^2$. Now, using the L_∞/L_1 result from Sec. 3.1 in Kakade et al. [2008], we get

$$\mathcal{R}_n(\mathcal{F}_2) \le X_\infty^2\bar{W}_{2,1}^2\sqrt{\frac{2\log d}{n}}\;. \tag{5}$$

The result follows from Eq. (3), with Eq. (4) and Eq. (5). ■

4.2 Kernel Learning

We briefly review the kernel learning setting first explored in Lanckriet et al. [2004]. Let $\mathcal{X}$ be an input space and let $\mathcal{T} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathcal{X}^n$ be the training dataset. Kernel algorithms work with the space of linear functions,

$$\left\{\mathbf{x} \mapsto \sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \mathbf{x}) \;:\; \alpha_i \in \mathbb{R}\right\}\;.$$

In kernel learning, we consider a kernel *family* $\mathcal{K}$ and consider the class,

$$\left\{\mathbf{x} \mapsto \sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \mathbf{x}) \;:\; K \in \mathcal{K},\; \alpha_i \in \mathbb{R}\right\}\;.$$

In particular, we can choose a finite set $\{K_1, \ldots, K_k\}$ of base kernels and consider the convex combinations,

$$\mathcal{K}_c^+ = \left\{\sum_{j=1}^k \mu_j K_j \;:\; \mu_j \ge 0,\; \sum_{j=1}^k \mu_j = 1\right\}\;.$$

This is the unconstrained function class. In applications, one constrains the function class in some way. The class considered in Lanckriet et al. [2004] is

$$\mathcal{F}_{\mathcal{K}_c^+} = \left\{\mathbf{x} \mapsto \sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \cdot) \;:\; K = \sum_{j=1}^k \mu_j K_j,\; \mu_j \ge 0, \right.$$
$$\left.\sum_{j=1}^k \mu_j = 1,\; \boldsymbol{\alpha}^\top K(\mathcal{T})\boldsymbol{\alpha} \le 1/\gamma^2\right\} \tag{6}$$

where $\gamma > 0$ is a margin parameter and $K(\mathcal{T})_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$ is the $n \times n$ Gram matrix of the kernel K on the dataset $\mathcal{T}$.

Theorem 21 *(Kernel learning) Consider the class $\mathcal{F}_{\mathcal{K}_c^+}$ defined in Eq. (6). Let $K_j(\mathbf{x},\mathbf{x}) \le B$ for $1 \le j \le k$ and $\mathbf{x} \in \mathcal{X}$. Then,*

$$\mathcal{R}_\mathcal{T}(\mathcal{F}_{\mathcal{K}_c^+}) \le e\sqrt{\frac{B\log k}{\gamma^2 n}}\;.$$

Before we present the proof, first note that the dependence on the number of features, k, is rather mild (only logarithmic) — implying that we can learn a kernel as a (convex) combination of a rather large number of base kernels.

Also, let us discuss how the above improves upon the prior bounds provided by Lanckriet et al. [2004] and Srebro and Ben-David [2006] (neither of which had logarithmic k dependence). The former proves a bound of $O\left(\sqrt{\frac{Bk}{\gamma^2 n}}\right)$ which is quite inferior to our bound. We cannot compare our bound directly to the bound in Srebro and Ben-David [2006] as they do not work with Rademacher complexities. However, if one compares the resulting generalization error bounds, then their bound is

$$O\left(\sqrt{\frac{k\log\frac{n^3B}{\gamma^2 k} + \frac{B}{\gamma^2}\log\frac{\gamma n}{\sqrt{B}}\log\frac{nB}{\gamma^2}}{n}}\right)$$

and ours is

$$O\left(\sqrt{\frac{B\log k}{\gamma^2 n}}\right)\;.$$

If $k \ge n$, their bound is vacuous (while ours is still meaningful). If $k \le n$, our bound is better.

Proof: Let $\mathcal{H}_j$ be the RKHS of K_j,

$$\mathcal{H}_j = \left\{\sum_{i=1}^l \alpha_i K_j(\tilde{\mathbf{x}}_i, \cdot) \;:\; l > 0,\; \tilde{\mathbf{x}}_i \in \mathcal{X},\; \boldsymbol{\alpha} \in \mathbb{R}^l\right\}\;.$$

equipped with the inner product

$$\left\langle \sum_{i=1}^l \alpha_i K_j(\tilde{\mathbf{x}}_i, \cdot), \sum_{j=1}^m \alpha'_i K_j(\tilde{\mathbf{x}}'_j, \cdot)\right\rangle_{\mathcal{H}_j} = \sum_{i,j} \alpha_i\alpha'_j K_j(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}'_j)$$

Consider the space $\mathcal{H} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_k$ equipped with the inner product,

$$\langle \vec{u}, \vec{v} \rangle := \sum_{i=1}^{k} \langle u_i, v_i \rangle_{\mathcal{H}_i} .$$

Let r, s be dual exponents with $r \in (1, 2]$, $s \in [2, \infty)$. For $\vec{w} \in \mathcal{H}$, let $\| \cdot \|_{2,r}$ be the norm defined by

$$\|\vec{w}\|_{2,r} = \left(\sum_{i=1}^{k} \|w_i\|_{\mathcal{H}_i}^{r} \right)^{\frac{1}{r}} .$$

We now claim that

$$\mathcal{F}_{\mathcal{K}_c^+} \subseteq \mathcal{F}_r \tag{7}$$

where

$$\mathcal{F}_r := \{ \mathbf{x} \mapsto \left\langle \vec{w}, \vec{\phi}(\mathbf{x}) \right\rangle \, : \, \vec{w} \in \mathcal{H}, \ \|\vec{w}\|_{2,r} \le 1/\gamma \} ,$$

and

$$\vec{\phi}(\mathbf{x}) = (K_1(\mathbf{x}, \cdot), \ldots, K_k(\mathbf{x}, \cdot)) \in \mathcal{H} .$$

To see this, pick arbitrary an arbitrary f in $\mathcal{F}_{\mathcal{K}_c^+}$. Thus, for some α_i's and μ_j's, we have,

$$\begin{aligned} f(\mathbf{x}) &= \sum_{i=1}^{n} \alpha_i \left(\sum_{j=1}^{K} \mu_j K_j(\mathbf{x}_i, \mathbf{x}) \right) \\ &= \sum_{j=1}^{k} \sum_{i=1}^{n} \mu_j \alpha_i K_j(\mathbf{x}_i, \mathbf{x}) \\ &= \left\langle \vec{w}, \vec{\phi}(\mathbf{x}) \right\rangle \end{aligned}$$

where $\vec{w} \in \mathcal{H}$ is such that

$$w_j = \sum_{i=1}^{n} \mu_j \alpha_i K_j(\mathbf{x}_i, \cdot) \in \mathcal{H}_j .$$

Moreover,

$$\begin{aligned} \|\vec{w}\|_{2,r}^2 &\le \|\vec{w}\|_{2,1}^2 \\ &= \left(\sum_{j=1}^{k} \left\| \sum_{i=1}^{n} \mu_j \alpha_i K_j(\mathbf{x}_i, \cdot) \right\|_{\mathcal{H}_j} \right)^2 \\ &= \left(\sum_{j=1}^{k} \mu_j \left\| \sum_{i=1}^{n} \alpha_i K_j(\mathbf{x}_i, \cdot) \right\|_{\mathcal{H}_j} \right)^2 \\ &\le \sum_{j=1}^{k} \mu_j \left\| \sum_{i=1}^{n} \alpha_i K_j(\mathbf{x}_i, \cdot) \right\|_{\mathcal{H}_j}^2 \\ &= \sum_{j=1}^{k} \mu_j \boldsymbol{\alpha}^\top K_j(\mathcal{T}) \boldsymbol{\alpha} \\ &= \boldsymbol{\alpha}^\top \left(\sum_{j=1}^{k} \mu_j K_j(\mathcal{T}) \right) \boldsymbol{\alpha} \\ &\le 1/\gamma^2 , \end{aligned}$$

where the second inequality is by Cauchy-Schwarz. Hence, Eq. (7) holds.

Since $\|K_j(\mathbf{x}, \cdot)\|_{\mathcal{H}_j} \le \sqrt{B}$, we also have $\|\vec{\phi}(\mathbf{x})\|_{2,s} \le k^{1/s}\sqrt{B}$ for any $\mathbf{x} \in \mathcal{X}$. By Corollary 19, $\frac{1}{2}\| \cdot \|_{2,s}$ is s-smooth. Hence, by Thm. 6, its conjugate $\frac{1}{2}\| \cdot \|_{2,r}^2$ is $1/s$-strongly convex. Now, we use Thm. 9 with $X = k^{1/s}\sqrt{B}$, $f_{\max} = \frac{1}{2}\gamma^2$ and $\beta = 1/s$, to get

$$\mathcal{R}_n(\mathcal{F}_{\mathcal{K}_c^+}) \le \mathcal{R}_n(\mathcal{F}_r) \le k^{1/s} \sqrt{\frac{Bs}{\gamma^2 n}} .$$

Setting $s = \log k$ finishes the proof. ■

4.3 Online Control with Quadratic Costs

Consider a finite horizon control problem where at each time step t, the learner (or controller in this context) has to choose a "control" direction $\mathbf{u}_t \in \mathbb{R}^n$, $\|\mathbf{u}_t\|_2 = 1$. However, instead of a fixed quadratic cost function, as is assumed in Linear-Quadratic control, the cost function is chosen adversarially. More specifically, at each time step t, the adversary chooses a positive semidefinite matrix $C_t \in \mathbb{S}^n$ and the learner incurs the cost $\mathbf{u}_t^\top C_t \mathbf{u}_t$. As is usual, we define the learner's regret after T time steps to be

$$\sum_{t=1}^{T} \mathbf{u}_t^\top C_t \mathbf{u}_t - \inf_{\mathbf{u} \, : \, \|\mathbf{u}\|_2 = 1} \sum_{t=1}^{T} \mathbf{u}^\top C_t \mathbf{u} .$$

Note that this is similar to the online variance minimization problem described in Warmuth and Kuzmin [2006]. Suppose Algorithm 2 is run with

$$F(X) = \sum_{i=1}^{n} \lambda_i(X) \log(n \lambda_i(X)) . \tag{8}$$

Since $F = f \circ \lambda$ for $f(x) = \sum_i x_i \log(n x_i)$, we have, by Thm. 12, $F^\star = f^\star \circ \lambda$ where $f^\star(y) = \log\left((\sum_i \exp(y_i))/n \right)$. Also, by Thm. 14,if $X = U \mathrm{Diag}(\lambda(X)) U^\top$ then

$$\begin{aligned} \nabla F^\star(X) &= U \mathrm{Diag}(\nabla f^\star(\lambda(X))) U^\top \\ &= U \mathrm{Diag}\left(\frac{\exp(\lambda_1(X))}{Z}, \ldots, \frac{\exp(\lambda_n(X))}{Z} \right) U^\top \end{aligned}$$

where $Z = \sum_{i=1}^{n} \exp(\lambda_i(X))$.

Theorem 22 *(Control) Suppose Algorithm 2 is run with the choice of F given in Eq. (8) and the sequence C_t is such that $C_t \in \mathbb{S}^n$, $\|\lambda(C_t)\|_\infty \le K$. Then, we have, for any $\eta > 0$,*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{u}_t^\top C_t \mathbf{u}_t - \inf_{\mathbf{u} \, : \, \|\mathbf{u}\|_2 = 1} \sum_{t=1}^{T} \mathbf{u}^\top C_t \mathbf{u} \right] \le \frac{\log n}{\eta} + \eta K^2 T .$$

Proof: Note that

$$\begin{aligned} \mathbb{E}\left[\mathbf{u}_t^\top C_t \mathbf{u}_t \right] &= \mathbb{E}\left[\sum_{i=1}^{n} \lambda_{t,i} \mathbf{v}_{t,i}^\top C_t \mathbf{v}_{t,i} \right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n} \lambda_{t,i} \left\langle C_t, \mathbf{v}_{t,i} \mathbf{v}_{t,i}^\top \right\rangle \right] \\ &= \mathbb{E}\left[\left\langle C_t, \sum_{i=1}^{n} \lambda_{t,i} \mathbf{v}_{t,i} \mathbf{v}_{t,i}^\top \right\rangle \right] \\ &= \mathbb{E}\left[\langle C_t, W_t \rangle \right] \end{aligned}$$

Algorithm 2 Online Control with Quadratic Costs

```
$S_1 \leftarrow \mathbf{0}$
$W_1 \leftarrow \mathrm{Diag}(1/n, 1/n, \ldots, 1/n)$

for $t = 1$ to $T$ do
    Compute the eigen-decomposition
        $W_t = \sum_{i=1}^{n} \lambda_{t,i} \mathbf{v}_{t,i} \mathbf{v}_{t,i}^\top$
    Choose $i_t$ such that $\mathbb{P}(i_t = j) = \lambda_{t,j}$
    Apply "control" $\mathbf{u}_t = \mathbf{v}_{t,i_t}$
    Receive cost matrix $C_t \in \mathbb{S}^n$
    Incur cost $\mathbf{u}_t^\top C_t \mathbf{u}_t$
    $S_{t+1} \leftarrow S_t + C_t$
    $W_{t+1} \leftarrow \nabla F^\star(-\eta S_{t+1})$
end for
```

Note that $F^\star(\mathbf{0}) = 0$. Now, Thm. 8 along with Thm. 16 gives us,

$$\sum_{t=1}^{T} \langle C_t, W_t \rangle - \min_{W \in S} \sum_{t=1}^{T} \langle C_t, W \rangle \leq \frac{\log n}{\eta} + \frac{\eta K^2 T}{2 \cdot \frac{1}{2}} ,$$

where S is the set of symmetric positive semidefinite matrices with trace 1. Note that if $\|\mathbf{u}\|_2 = 1$, then $\mathbf{u}\mathbf{u}^\top \in S$ and $\left\langle C_t, \mathbf{u}\mathbf{u}^\top \right\rangle = \mathbf{u}^\top C_t \mathbf{u}$. Therefore,

$$\sum_{t=1}^{T} \langle C_t, W_t \rangle - \min_{\|\mathbf{u}\|_2 = 1} \sum_{t=1}^{T} \mathbf{u}^\top C_t \mathbf{u} \leq \frac{\log n}{\eta} + \eta K^2 T .$$

Taking expectation now proves the result. ■

References

Alekh Agarwal, Alexander Rakhlin, and Peter Bartlett. Matrix regularization techniques for online multitask learning. Technical report, EECS Department, University of California, Berkeley, 2008.

J. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization*. Springer, 2006.

G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Linear algorithms for online multitask classification. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, pages 251–262, 2008.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

C. Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53(3), 2002.

A. J. Grove, N. Littlestone, and D. Schuurmans. General convergence results for linear discriminant updates. *Machine Learning*, 43(3):173–210, 2001.

A. Juditsky and A. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *submitted to Annals of Probability*, 2008.

S.M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 22*, 2008.

J. Kivinen and M. Warmuth. Relative loss bounds for multidimensional regression problems. *Journal of Machine Learning*, 45(3):301–329, July 2001.

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

G.R.G. Lanckriet, N. Cristianini, P.L. Bartlett, L. El Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(2):173–183, 1995.

R. Meir and T. Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

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

Y. Nesterov. Primal-dual subgradient methods for convex problems. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 2005.

I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *Ann. Probab*, 22(4):1679–1706, 1994.

G. Pisier. Martingales with values in uniformly convex spaces. *Israel J. Math.*, 20: 326–350, 1975.

R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University, 2007.

S. Shalev-Shwartz and Y. Singer. Convex repeated games and Fenchel duality. In *Advances in Neural Information Processing Systems 20*, 2006.

S. Shalev-Shwartz and Y. Singer. A primal-dual perspective of online learning algorithms. *Machine Learning Journal*, 2007.

S. Shalev-Shwartz and Y. Singer. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2008.

N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, pages 169–183, 2006.

M. Warmuth and D. Kuzmin. Online variance minimization. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2006.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.

A Technical Lemmas and Proofs

A.1 The duality of strong convexity and strong smoothness

Proof:[of Thm. 6] First, [Shalev-Shwartz, 2007, Lemma 15] yields the claim $1 \Rightarrow 2$. It is left to prove that f is strongly convex assuming that $f^\star$ is strongly smooth. For simplicity assume that $\beta = 1$. Denote $g(y) = f^\star(x + y) - (f^\star(x) + \langle \nabla f^\star(x), y \rangle)$. By the smoothness assumption, $g(y) \leq \frac{1}{2}\|y\|_\star^2$. This implies that $g^\star(a) \geq \frac{1}{2}\|a\|^2$ because of [Shalev-Shwartz and Singer, 2008, Lemma 19] and that the conjugate of half squared norm is half squared of the dual norm. Using the definition of g we have

$$\begin{aligned} g^\star(a) &= \sup_y \langle y, a \rangle - g(y) \\ &= \sup_y \langle y, a \rangle - (f^\star(x + y) - (f^\star(x) + \langle \nabla f^\star(x), y \rangle)) \\ &= \sup_y \langle y, a + \nabla f^\star(x) \rangle - f^\star(x + y) + f^\star(x) \\ &= \sup_z \langle z - x, a + \nabla f^\star(x) \rangle - f^\star(z) + f^\star(x) \\ &= f(a + \nabla f^\star(x)) + f^\star(x) - \langle x, a + \nabla f^\star(x) \rangle \end{aligned}$$

where we have used that $f^{\star\star} = f$, in the last step. Denote $u = \nabla f^\star(x)$. From the equality in Fenchel-Young (e.g. [Shalev-Shwartz and Singer, 2008, Lemma 17]) we obtain that $\langle x, u \rangle = f^\star(x) + f(u)$ and thus

$$g^\star(a) = f(a + u) - f(u) - \langle x, a \rangle \ .$$

Combining with $g^\star(a) \geq \frac{1}{2}\|a\|^2$, we have

$$f(a + u) - f(u) - \langle x, a \rangle \ \geq \ \frac{1}{2}\|a\|^2 \ , \tag{9}$$

which holds for all a, x, with $u = \nabla f^\star(x)$.

Now let us prove that for any point u' in the relative interior of the domain of f that if $x \in \partial f(u')$ then $u' = \nabla f^\star(x)$. Let $u := \nabla f^\star(x)$ and we must show that $u' = u$. By Fenchel-Young, we have that $\langle x, u' \rangle = f^\star(x) + f(u')$, and, again by Fenchel-Young (and $f^{\star\star} = f$), we have $\langle x, u \rangle =$

$f^\star(x)+f(u)$. We can now apply Equation Eq. (9), to obtain:

$$\begin{aligned} 0 &= \langle x,u\rangle - f(u) - (\langle x,u'\rangle - f(u')) \\ &= f(u') - f(u) - \langle x,u'-u\rangle \ \geq\ \frac{1}{2}\|u'-u\|^2 , \end{aligned}$$

which implies that $u' = \nabla f^\star(x)$.

Next, let u_1, u_2 be two points in the relative interior of the domain of f, let $\alpha \in (0,1)$, and let $u = \alpha u_1 + (1-\alpha)u_2$. Let $x \in \partial f(u)$ (which is non-empty [1]). We have that $u = \nabla f^\star(x)$, by the previous argument. Now we are able to apply Equation Eq. (9) twice, once with $a = u_1 - u$ and once with $a = u_2 - u$ (and both with x) to obtain

$$\begin{aligned} f(u_1) - f(u) - \langle x, u_1 - u\rangle &\geq \frac{1}{2}\|u_1 - u\|^2 \\ f(u_2) - f(u) - \langle x, u_2 - u\rangle &\geq \frac{1}{2}\|u_2 - u\|^2 \end{aligned}$$

Finally, summing up the above two equations with coefficients α and $1-\alpha$ we obtain that f is strongly convex. ■

A.2 Group Norms

Proof:[of Lemma 17] Recall that, by definition,

$$\begin{aligned} (\|Y\|_{\Psi,\Phi})_\star &= \sup\{\langle X,Y\rangle \ : \ \|X\|_{\Psi,\Phi} \leq 1\} \\ &= \sup\{\sum_{i=1}^n \langle X_i, Y_i\rangle \ : \ \|X\|_{\Psi,\Phi} \leq 1\} \qquad (10) \end{aligned}$$

We first prove $(\|Y\|_{\Psi,\Phi})_\star \leq \|Y\|_{\Psi_\star,\Phi_\star}$. Let $\vec{\Psi}(X)$ be a shorthand for $(\Psi(X_1),\ldots,\Psi(X_n))$. Now, we have,

$$\begin{aligned} \sum_{i=1}^n \langle X_i, Y_i\rangle &\leq \sum_{i=1}^n \Psi(X_i)\Psi_\star(Y_i) \\ &= \left\langle \vec{\Psi}(X), \vec{\Psi_\star}(Y)\right\rangle \\ &\leq \Phi(\vec{\Psi}(X)) \cdot \Phi_\star(\vec{\Psi_\star}(Y)) \\ &= \|X\|_{\Psi,\Phi} \cdot \|Y\|_{\Psi_\star,\Phi_\star} . \end{aligned}$$

Thus, the sup in Eq. (10) is no more than $\|Y\|_{\Psi_\star,\Phi_\star}$.

To prove $(\|Y\|_{\Psi,\Phi})_\star \geq \|Y\|_{\Psi_\star,\Phi_\star}$, let $\theta \in \mathbb{R}^n$ be such that $\Phi(\theta) = 1$ and

$$\left\langle \theta, \vec{\Psi_\star}(Y)\right\rangle = \Phi_\star(\vec{\Psi_\star}(Y)) = \|Y\|_{\Psi_\star,\Phi_\star} .$$

Further, for each i, let $Z_i \in \mathbb{R}^m$ be such that $\Psi(Z_i) = 1$ and $\langle Z_i, Y_i\rangle = \Psi_\star(Y_i)$. Now, let X be the matrix with columns $X_i = \theta_i Z_i$. Then, we have

$$\begin{aligned} \sum_{i=1}^n \langle X_i, Y_i\rangle &= \sum_{i=1}^n \theta_i \langle Z_i, Y_i\rangle \\ &= \sum_{i=1}^n \theta_i \Psi_\star(Y_i) \\ &= \left\langle \theta, \vec{\Psi_\star}(Y)\right\rangle \\ &= \|Y\|_{\Psi_\star,\Phi_\star} . \end{aligned}$$

Furthermore, since $\Psi(X_i) = |\theta_i|\Psi(Z_i) = |\theta_i|$ and Φ is absolutely symmetric, we have $\|X\|_{\Psi,\Phi} = \Phi(|\theta|) = \Phi(\theta) = 1$. Thus, the sup in Eq. (10) is at least $\|Y\|_{\Psi_\star,\Phi_\star}$. ■

Proof:[of Thm. 18] Note that an equivalent definition of σ-smoothness of a function f w.r.t. a norm $\|\cdot\|$ is that, for all x,y and $\alpha \in [0,1]$, we have

$$\begin{aligned} f(\alpha x + (1-\alpha)y) \geq \alpha f(x) + (1-\alpha)f(y) \\ - \frac{1}{2}\sigma\alpha(1-\alpha)\|x-y\|^2 . \end{aligned}$$

Let $X, Y \in \mathbb{R}^{m\times n}$ be arbitrary matrices with columns X_i and Y_i respectively. We need to prove

$$\begin{aligned} \|(1-\alpha)X + \alpha Y\|_{\Psi,\Phi}^2 \geq \alpha\|X\|_{\Psi,\Phi}^2 + (1-\alpha)\|Y\|_{\Psi,\Phi}^2 \\ - \frac{1}{2}(\sigma_1+\sigma_2)\alpha(1-\alpha)\|X-Y\|_{\Psi,\Phi}^2 . \quad (11) \end{aligned}$$

We have,

$$\begin{aligned} &\|(1-\alpha)X + \alpha Y\|_{\Psi,\Phi}^2 \\ &= \Phi^2(\ldots, \Psi(\alpha X_i + (1-\alpha)Y_i), \ldots) \\ &= (\Phi^2 \circ \surd)(\ldots, \Psi^2(\alpha X_i + (1-\alpha)Y_i), \ldots) \\ &\geq (\Phi^2 \circ \surd)(\ldots, \alpha\Psi^2(X_i) + (1-\alpha)\Psi^2(Y_i) \\ &\quad - \frac{1}{2}\sigma_1\alpha(1-\alpha)\Psi^2(X_i - Y_i), \ldots) \\ &\geq (\Phi^2 \circ \surd)(\ldots, \alpha\Psi^2(X_i) + (1-\alpha)\Psi^2(Y_i), \ldots) \\ &\quad - \frac{1}{2}\sigma_1\alpha(1-\alpha)(\Phi^2 \circ \surd)(\ldots, \Psi^2(X_i - Y_i), \ldots) \\ &= \Phi^2(\ldots, \sqrt{\alpha\Psi^2(X_i) + (1-\alpha)\Psi^2(Y_i)}, \ldots) \\ &\quad - \frac{1}{2}\sigma_1\alpha(1-\alpha)\|X-Y\|_{\Psi,\Phi}^2 . \qquad (12) \end{aligned}$$

Now, we use that, for any $x, y \geq 0$ and $\alpha \in [0,1]$, we have

$$\sqrt{\alpha x^2 + (1-\alpha)y^2} \geq \alpha x + (1-\alpha)y .$$

Thus, we have

$$\begin{aligned} &\Phi^2(\ldots, \sqrt{\alpha\Psi^2(X_i) + (1-\alpha)\Psi^2(Y_i)}, \ldots) \\ &\geq \Phi^2(\ldots, \alpha\Psi(X_i) + (1-\alpha)\Psi(Y_i), \ldots) \\ &\geq \alpha\Phi^2(\ldots, \Psi(X_i), \ldots) + (1-\alpha)\Phi^2(\ldots, \Psi(Y_i), \ldots) \\ &\quad - \frac{1}{2}\sigma_2\alpha(1-\alpha)\Phi^2(\ldots, \Psi(X_i) - \Psi(Y_i), \ldots) \\ &\geq \alpha\|X\|_{\Psi,\Phi}^2 + (1-\alpha)\|Y\|_{\Psi,\Phi}^2 \\ &\quad - \frac{1}{2}\sigma_2\alpha(1-\alpha)\Phi^2(\ldots, \Psi(X_i - Y_i), \ldots) \\ &= \alpha\|X\|_{\Psi,\Phi}^2 + (1-\alpha)\|Y\|_{\Psi,\Phi}^2 \\ &\quad - \frac{1}{2}\sigma_2\alpha(1-\alpha)\|X-Y\|_{\Psi,\Phi}^2 \end{aligned}$$

Plugging this into Eq. (12) proves Eq. (11). ■

[1] The set $\partial f(u)$ is not empty for all u in the relative interior of the domain of f. See the relative max formula in [Borwein and Lewis, 2006, page 42] or [Rockafellar, 1970, page 253]. If u is not in the interior of f, then $\partial f(u)$ is empty. But, a function is defined to be essentially strictly convex if it is strictly convex on any subset of $\{u : \partial f(u) \neq \emptyset\}$. The last set is called the domain of ∂f and it contains the relative interior of the domain of f, so we're ok here.