On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization

Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari
Toyota Technological Institute — Chicago

Abstract

We provide sharp bounds for Rademacher and Gaussian complexities of (constrained) linear classes. These bounds make short work of providing a number of corollaries including: risk bounds for linear prediction (including settings where the weight vectors are constrained by either L_2 or L_1 constraints), margin bounds (including both L_2 and L_1 margins, along with more general notions based on relative entropy), a proof of the PAC-Bayes theorem, and L_2 covering numbers (with L_p norm constraints and relative entropy constraints). In addition to providing a unified analysis, the results herein provide some of the sharpest risk and margin bounds (improving upon a number of previous results). Interestingly, our results show that the uniform convergence rates of empirical risk minimization algorithms tightly match the regret bounds of online learning algorithms for linear prediction (up to a constant factor of 2).

1 Introduction

Linear prediction is the cornerstone of an extensive number of machine learning algorithms, including SVM's, logistic and linear regression, the lasso, boosting, etc. A paramount question is to understand the generalization ability of these algorithms in terms of the attendant complexity restrictions imposed by the algorithm. For example, for the sparse methods (e.g. regularizing based on L_1 norm of the weight vector) we seek generalization bounds in terms of the sparsity level. For margin based methods (e.g. SVMs or boosting), we seek generalization bounds in terms of either the L_2 or L_1 margins. The focus of this paper is to provide a more unified analysis for methods which use linear prediction.

Given a training set $\{(x_i, y_i)\}_{i=1}^n$, the paradigm is to compute a weight vector $\hat{\mathbf{w}}$ which minimizes the F-regularized ℓ-risk. More specifically,

$$\hat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} \ell(\langle \mathbf{w}, x_i \rangle, y_i) + \lambda F(\mathbf{w}) \tag{1}$$

where ℓ is the loss function and F is the regularizer. In the closely related dual version of this problem, we have:

$$\hat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}:F(\mathbf{w})\le c} \frac{1}{n}\sum_{i=1}^{n} \ell(\langle \mathbf{w}, x_i \rangle, y_i) \tag{2}$$

where, instead of regularizing, a hard restriction over the parameter space is imposed (by the constant c). This works provides generalization bounds for an extensive family of regularization functions F. We do this via providing sharp Rademacher complexity bounds on the class of linear predictions. Using standard results in the literature, we can obtain both generalization bounds (a la [Bartlett and Mendelson, 2002]) and margin bounds (a la Koltchinskii and Panchenko [2002]).

A staggering number of results have focused on this problem in varied special cases. Perhaps the most extensively studied are margin bounds for the 0-1 loss. For L_2-margins (relevant for SVM's, perceptron based algorithms, etc.), the sharpest bounds are those provided by Bartlett and Mendelson [2002] (using Rademacher complexities) and Langford and Shawe-Taylor [2003], McAllester [2003] (using the PAC-Bayes theorem). For L_1-margins (relevant for Boosting, winnow, etc), bounds are provided by Schapire et al. [1998] (using a self-contained analysis) and Langford et al. [2001] (using PAC-Bayes, with a different analysis). Another active line of work is on sparse methods — particularly methods which impose sparsity via L_1 regularization (in lieu of the non-convex L_0 norm). For L_1 regularization, Ng [2004] provides generalization bounds for this case, which follow from the covering number bounds of Zhang [2002]. However, these bounds are only stated as polynomial in the relevant quantities (dependencies are not provided).

Previous to this work, the most unified framework for providing generalization bounds for linear prediction stem from the covering number bounds in Zhang [2002]. Using these covering number bounds, Zhang [2002] derives margin bounds in a variety of cases. However, providing sharp generalization bounds for problems with L_1 regularization (or L_1 constraints in the dual) requires more delicate arguments. As mentioned, Ng [2004] provides bounds for this case, but the techniques used by Ng [2004] would result in rather loose dependencies (the dependence on m would be $m^{\frac{1}{4}}$). We discuss this later in Section 4.

In this work, we provide sharp bounds for Rademacher and Gaussian complexities of linear classes, with respect to a strongly convex complexity function F (the regularizer in Equation 1). These bounds make short work of providing a number of corollaries:

- **Risk Bounds:** We provide sharp generalization bounds for the optimization problem in Equation 2, when the loss function is Lipschitz continuous. As a special case, obtain generalization bounds when we impose L_1 constraints (or regularization) on the weights. These bounds are much sharper than those provided by Ng [2004], and, even using a sharper analysis using Dudley's entropy integral with the covering number bounds of Zhang [2002] would lead to extra $\log n$ factors.
- **Margin Bounds:** Our bounds directly lead to bounds for both L_2 and L_1 margins, and, more generally, for L_p margins. For the L_2 case, Bartlett and Mendelson [2002] provides essentially the same bound. For the L_1 margin case, our bound improves upon all previous bounds by removing all $\log n$ factors from Schapire et al. [1998], Langford et al. [2001], Zhang [2002]. There are also generalizations which can be stated in terms of the relative entropy as in Langford et al. [2001], Zhang [2002] (and these bounds are also sharper than previous bounds, in the general case).
- **PAC-Bayes theorem:** As a simple corollary, we are able to derive a (slightly sharper) version of the original PAC-Bayes theorem.
- **Covering number bounds:** By Sudakov's minoration, we also obtain upper bounds on the L_2 covering numbers in a variety of cases, which are sharper for than those implied by Zhang [2002].
- **Fast Rates:** Appealing to the recent results in Anonymous [2008], our complexity bounds can also be used in understanding the performance of the regularized minimizer, as in Equation 1, for certain cases.

The proof techniques for providing these bounds are rooted in convex duality — reminiscent of the proofs for deriving regret bounds for online learning algorithms. Interestingly, the risk bounds we provide closely match the regret bounds for online learning algorithms (up to a constant factor of 2), thus showing that the uniform converge rates of empirical risk minimization algorithms tightly match the regret bounds of online learning algorithms (for linear prediction). The Discussion provides this more detailed comparison.

2 Preliminaries

Our input space, $\mathcal{X}$, is a subset of a vector space, and our output space is $\mathcal{Y}$. Our samples $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ are distributed according to some unknown distribution P. The inner product between vectors $\mathbf{x}$ and $\mathbf{w}$ is denoted by $\langle \mathbf{w}, \mathbf{x} \rangle$, where $\mathbf{w} \in S$ (here, S is a subset of the dual space to our input vector space). A norm of a vector x is denoted by $||x||$, and the dual norm is defined as $\|\mathbf{w}\|_\star = \sup\{\langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{x}\| \le 1\}$. We further assume that for all $\mathbf{x} \in \mathcal{X}$, $\|\mathbf{x}\| \le X$.

Let $\ell : \mathbf{R} \times \mathcal{Y} \to \mathbb{R}^+$ be our loss function of interest. Throughout we shall consider linear predictors of form $\langle \mathbf{w}, \mathbf{x} \rangle$. The expected of loss of $\mathbf{w}$ is denoted by $\mathcal{L}(\mathbf{w}) = \mathbb{E}[\ell(\langle \mathbf{w}, \mathbf{x} \rangle, y)]$. As usual, we are provided with an sequence of i.i.d. samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, and our goal is to minimize our expected loss. We denote the empirical loss as $\hat{\mathcal{L}}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \ell(\langle \mathbf{w}, \mathbf{x}_i \rangle, y_i)$.

The restriction we make on our complexity function F is that it is a strongly convex function. In particular, we assume it is strongly convex with respect to our dual norm: a function $F : S \to \mathbb{R}$ is said to be σ-strongly convex w.r.t. to $\|\cdot\|_*$ iff $\forall \mathbf{u}, \mathbf{v} \in S$, $\forall \alpha \in [0, 1]$, we have

$$F(\alpha \mathbf{u} + (1-\alpha)\mathbf{v}) \le \alpha F(\mathbf{u}) + (1-\alpha)F(\mathbf{v}) - \frac{\sigma}{2}\alpha(1-\alpha)\|\mathbf{u} - \mathbf{v}\|_*^2 .$$

See Shalev-Shwartz and Singer [2006] for more discussion on this generalized definition of strong convexity.

Recall the definition of the Rademacher and Gaussian complexity of a function class $\mathcal{F}$,

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)\epsilon_i\right] \qquad \mathcal{G}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)\epsilon_i\right]$$

where, in the former, ϵ_i independently takes values in $\{-1, +1\}$ with equal probability, and, in the latter, ϵ_i are independent, standard normal random variables. In both expectations, $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are i.i.d.

There are number of methods in the literature to use Rademacher complexities to obtain either generalization bounds or margin bounds. Bartlett and Mendelson [2002] provide a generalization bound for Lipschitz loss functions. For binary prediction, the results in Koltchinskii and Panchenko [2002] provide means to obtain margin bounds through Rademacher complexities, which we use later. Typically, we desire an upper bound on the Rademacher Complexity that decreases with n.

3 Complexities of Linear Function Classes

Given a subset $\mathcal{W} \subseteq S$, define the associated class of linear functions $\mathcal{F}_\mathcal{W}$ as $\mathcal{F}_\mathcal{W} := \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \mathbf{w} \in \mathcal{W}\}$. Our main theorem bounds the complexity of $\mathcal{F}_\mathcal{W}$ for certain sets $\mathcal{W}$.

Theorem 1. *(Complexity Bounds) Let S be a closed convex set and let $F : S \to \mathbb{R}$ be σ-strongly convex w.r.t. $\|\cdot\|_*$. Further, let $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\| \le X\}$. Define $\mathcal{W} = \{\mathbf{w} \in S : F(\mathbf{w}) \le W_*^2\}$. Then, we have*

$$\mathcal{R}_n(\mathcal{F}_\mathcal{W}) \le XW_*\sqrt{\frac{2}{\sigma n}} \qquad , \qquad \mathcal{G}_n(\mathcal{F}_\mathcal{W}) \le XW_*\sqrt{\frac{2}{\sigma n}} .$$

Interestingly, these complexity bounds precisely match the regret bounds for online learning algorithms (for linear prediction), a point which we return to in the Discussion. We first provide a few examples, before proving this result.

3.1 Examples

(1) L_p/L_q **norms.** Let $S = \mathbb{R}^d$. Take $\|\cdot\|, \|\cdot\|_*$ to be the L_p, L_q norms for $p \in [2, \infty)$, $1/p + 1/q = 1$, where $\|\mathbf{x}\|_p := \left(\sum_{j=1}^d |\mathbf{x}_i|^p\right)^{1/p}$. Choose $F(\mathbf{w}) = \|\cdot\|_q^2$ and note that it is $2(q-1)$-strongly convex on $\mathbb{R}^d$ w.r.t. itself. Set $\mathcal{X}$, $\mathcal{W}$ as in Theorem 1. Then, we have

$$\mathcal{R}_n(\mathcal{F}_\mathcal{W}) \le XW_*\sqrt{\frac{p-1}{n}} . \tag{3}$$

(2) L_∞/L_1 **norms.** Let $S = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_1 = W_1 , \ \mathbf{w}_j \ge 0\}$ be the W_1-scaled probability simplex. Take $\|\cdot\|$, $\|\cdot\|_*$ to be the L_∞, L_1 norms, $\|\mathbf{x}\|_\infty = \max_{1\le j\le d} |\mathbf{x}_j|$. Fix a probability distribution $\mu > 0$ and let $F(\mathbf{w}) = \mathrm{entro}_\mu(\mathbf{w}) := \sum_j (\mathbf{w}_j/W_1)\log(\mathbf{w}_j/(W_1\mu_j))$. For any μ, $\mathrm{entro}_\mu(\mathbf{w})$ is $1/W_1^2$-strongly convex on S w.r.t. $\|\cdot\|_1$. Set $\mathcal{X}$ as in Theorem 1 and let $\mathcal{W}(E) = \{\mathbf{w} \in S : \mathrm{entro}_\mu(\mathbf{w}) \le E\}$. Then, we have

$$\mathcal{R}_n(\mathcal{F}_{\mathcal{W}(E)}) \le XW_1\sqrt{\frac{2E}{n}} . \tag{4}$$

Let $\mathcal{W} := \mathcal{W}(\log d)$ with uniform μ and note that it is the entire scaled probability simplex. Then

$$\mathcal{R}_n(\mathcal{F}_{\mathcal{W}}) \leq XW_1\sqrt{\frac{2\log d}{n}} \ . \tag{5}$$

The restriction $\mathbf{w}_j \geq 0$ can be removed in the definition of S by the standard trick of doubling the dimension of $\mathbf{x}$ to include negated copies of each coordinate. So, if we have $S = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_1 \leq W_1\}$ and we set $\mathcal{X}$ as above and $\mathcal{W} = S$, then we get $\mathcal{R}_n(\mathcal{F}_{\mathcal{W}}) \leq XW_1\sqrt{2\log(2d)/n}$.

(3) Smooth norms. A norm is $(2, D)$-smooth on S if for any $\mathbf{x}, \mathbf{y} \in S$,

$$\frac{d^2}{dt^2}\|\mathbf{x} + t\mathbf{y}\|^2 \leq 2D^2\|\mathbf{y}\|^2 \ .$$

Let $\|\cdot\|$ be a $(2, D)$-smooth norm and $\|\cdot\|_*$ be its dual. Lemma 11 in the appendix proves that $\|\cdot\|_*$ is $2/D^2$-strongly convex w.r.t. itself. Set $\mathcal{X}$, $\mathcal{W}$ as in Theorem 1. Then, we have

$$\mathcal{R}_n(\mathcal{F}_{\mathcal{W}}) \leq \frac{XW_*D}{\sqrt{n}} \ . \tag{6}$$

(4) Bregman divergences. For a strongly convex F, define the *Bregman divergence* $\Delta_F(\mathbf{w}\|\mathbf{v}) := F(\mathbf{w}) - F(\mathbf{v}) - \langle \nabla F(\mathbf{v}), \mathbf{w} - \mathbf{v}\rangle$. It is interesting to note that Theorem 1 is still valid if we choose $W_* = \{\mathbf{w} \in S : \Delta_F(\mathbf{w}\|\mathbf{v}) \leq W_*^2\}$ for some fixed $\mathbf{v} \in S$. This is because the Bregman divergence $\Delta_F(\cdot\|\mathbf{v})$ inherits the strong convexity of F.

Except for (5), none of the above bounds depend explicitly on the dimension of the underlying space and hence can be easily extended to infinite dimensional spaces under appropriate assumptions.

3.2 The Proof

First, some background on convex duality is in order. The Fenchel conjugate of $F : S \to \mathbb{R}$ is defined as:

$$F^*(\boldsymbol{\theta}) := \sup_{\mathbf{w}\in S} \langle \mathbf{w}, \boldsymbol{\theta}\rangle - F(\mathbf{w}) \ .$$

A simple consequence of this definition is Fenchel-Young inequality,

$$\forall \boldsymbol{\theta}, \mathbf{w} \in S, \ \langle \mathbf{w}, \boldsymbol{\theta}\rangle \leq F(\mathbf{w}) + F^*(\boldsymbol{\theta}) \ .$$

If F is σ-strongly convex, then F^* is differentiable and

$$\forall \boldsymbol{\theta}, \boldsymbol{\eta}, \ F^*(\boldsymbol{\theta} + \boldsymbol{\eta}) \leq F^*(\boldsymbol{\theta}) + \langle \nabla F^*(\boldsymbol{\theta}), \boldsymbol{\eta}\rangle + \frac{1}{2\sigma}\|\boldsymbol{\eta}\|_*^2 \ . \tag{7}$$

See the Appendix in Shalev-Shwartz [2007] for proof. Using this inequality we can control the expectation of F^* applied to a sum of independent random variables.

Lemma 2. *Let S be a closed convex set and let $F : S \to \mathbb{R}$ be σ-strongly convex w.r.t. $\|\cdot\|_*$. Let Z_i be mean zero independent random vectors such that $\mathbb{E}[\|Z_i\|^2] \leq V^2$. Define $S_i := \sum_{j\leq i} Z_i$. Then $F^*(S_i) - iV^2/2\sigma$ is a supermartingale. Therefore, $\mathbb{E}[F^*(S_n)] \leq nV^2/2\sigma$.*

Proof. Inequality (7) gives,

$$F^*(S_{i-1} + Z_i) \leq F^*(S_i) + \langle \nabla F^*(S_{i-1}), Z_i\rangle + \frac{1}{2\sigma}\|Z_i\|_*^2 \ .$$

Taking conditional expectation w.r.t. $Z_1, \ldots, Z_{i-1}$ and noting that $\mathbb{E}_{i-1}[Z_i] = 0$ and $\mathbb{E}_{i-1}[\|Z_i\|_*^2] \leq V^2$, we get

$$\mathbb{E}_{i-1}[F^*(S_i)] \leq F^*(S_{i-1}) + 0 + \frac{V^2}{2\sigma}$$

where $\mathbb{E}_{i-1}[\cdot]$ abbreviates $\mathbb{E}[\cdot \mid Z_1, \ldots, Z_{i-1}]$. □

Now we are ready to complete the proof:

Proof. Fix $\mathbf{x}_1, \ldots, \mathbf{x}_n$ such that $\|\mathbf{x}_i\| \le X$. Let $\boldsymbol{\theta} = \frac{1}{n}\sum_i \epsilon_i \mathbf{x}_i$ where ϵ_i's are i.i.d. Rademacher or Gaussian random variables (our proof only requires that $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i^2] = 1$). Choose arbitrary $\lambda > 0$. By Fenchel's inequality, we have $\langle \mathbf{w}, \lambda\boldsymbol{\theta}\rangle \le F(\mathbf{w}) + F^*(\lambda\boldsymbol{\theta})$ which implies

$$\langle \mathbf{w}, \boldsymbol{\theta}\rangle \le \frac{F(\mathbf{w})}{\lambda} + \frac{F^*(\lambda\boldsymbol{\theta})}{\lambda} .$$

Since, $F(\mathbf{w}) \le W_*^2$ for all $\mathbf{w} \in \mathcal{W}$, we have

$$\sup_{\mathbf{w}\in\mathcal{W}} \langle \mathbf{w}, \boldsymbol{\theta}\rangle \le \frac{W_*^2}{\lambda} + \frac{F^*(\lambda\boldsymbol{\theta})}{\lambda} .$$

Taking expectation (w.r.t. ϵ_i's), we get

$$\mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \langle \mathbf{w}, \boldsymbol{\theta}\rangle\right] \le \frac{W_*^2}{\lambda} + \frac{1}{\lambda}\mathbb{E}\left[F^*(\lambda\boldsymbol{\theta})\right] .$$

Now set $Z_i = \frac{\lambda\epsilon_i\mathbf{x}_i}{n}$ (so that $S_n = \lambda\boldsymbol{\theta}$) and note that the conditions of Lemma 2 are satisfied with $V^2 = \lambda^2 B^2/n^2$ and hence $\mathbb{E}[F^*(\lambda\boldsymbol{\theta})] \le \frac{\lambda^2 X^2}{2\sigma n}$. Plugging this above, we have

$$\mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \langle \mathbf{w}, \boldsymbol{\theta}\rangle\right] \le \frac{W_*^2}{\lambda} + \frac{\lambda X^2}{2\sigma n} .$$

Setting $\lambda = \sqrt{\frac{2\sigma n W_*^2}{X^2}}$ gives

$$\mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \langle \mathbf{w}, \boldsymbol{\theta}\rangle\right] \le X W_* \sqrt{\frac{2}{\sigma n}} .$$

which completes the proof. □

4 Corollaries

4.1 Risk Bounds

We now provided generalization error bounds for any Lipschitz loss function ℓ, with Lipschitz constant L_ℓ. The main reason we are interested in bounding Rademacher complexities is so that we can obtain generalization error bounds. To this end we first we first state the following theorem.

Theorem 3. (Bartlett and Mendelson [2002]) *For a Lipschitz loss ℓ bounded by c, for any $\delta > 0$ with probability at least $1 - \delta$ simultaneously for all $f \in \mathcal{F}$ we have that,*

$$\mathcal{L}(f) \le \hat{\mathcal{L}}(f) + 2L_\ell \mathcal{R}_n(\mathcal{F}) + c\sqrt{\frac{\log(1/\delta)}{2n}}$$

Now based on the above theorem and the bounds on Rademacher complexity proved in previous section, we obtain the following corollaries.

Corollary 4. *Each of the following statements holds with probability at least $1-\delta$ over the sample:*

- *Let $\mathcal{W}$ be as in the L_p/L_q* **norms** *example. For all $\mathbf{w} \in \mathcal{W}$,*

$$\mathcal{L}(\mathbf{w}) \le \hat{\mathcal{L}}(\mathbf{w}) + 2L_\ell X W_* \sqrt{\frac{p-1}{n}} + L_\ell X W_* \sqrt{\frac{\log(1/\delta)}{2n}}$$

- *Let $\mathcal{W}$ be as in the L_∞/L_1* **norms** *example. For all $\mathbf{w} \in \mathcal{W}$,*

$$\mathcal{L}(\hat{\mathbf{w}}) \le \hat{\mathcal{L}}(\mathbf{w}) + 2L_\ell X W_1 \sqrt{\frac{2\log(d)}{n}} + L_\ell X W_1 \sqrt{\frac{\log(1/\delta)}{2n}}$$

Ng [2004] provides bounds for methods which use L_1 regularization. These bounds are only stated as polynomial bounds, and, the methods used (covering number techniques from Pollard [1984] and covering number bounds from Zhang [2002]) would provide rather loose bounds (the n dependence would be $n^{\frac{1}{4}}$). In fact, even a more careful analysis via Dudley's entropy integral using the covering numbers from Zhang [2002] would result in a worse bound (with additional $\log n$ factors). The above argument is sharp and rather direct.

4.2 Margin Bounds

In this section we restrict ourselves to binary classification where $\mathcal{Y} = \{+1, -1\}$. Our prediction is given by $\text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)$. The zero-one loss function is given by $\ell(\langle \mathbf{w}, \mathbf{x} \rangle, y) = \mathbf{1}[y \langle \mathbf{w}, \mathbf{x} \rangle \leq 0]$. Denote the fraction of the data having γ-margin mistakes by $K_\gamma(f) := \frac{|\{i : y_i f(\mathbf{x}_i) < \gamma\}|}{n}$. We now demonstrate how to get improved margin bounds using the upper bounds for the Rademacher Complexity derived in Section 3. To this end we first state the following theorem which is a variant of Theorem 2 in Koltchinskii and Panchenko [2002].

Theorem 5. (Koltchinskii and Panchenko [2002]) *Consider an arbitrary function class $\mathcal{F}$ such that $\forall f \in \mathcal{F}$ we have $\sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x})| \leq C$. Then, with probability at least $1 - \delta$ over the sample, for all margins $\gamma > 0$ and all $f \in \mathcal{F}$ we have,*

$$\mathcal{L}(f) \leq K_\gamma(f) + 4\frac{\mathcal{R}_n(\mathcal{F})}{\gamma} + \sqrt{\frac{\log(\log_2 \frac{4C}{\gamma})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \, .$$

Proof. Proof is provided in the appendix. □

Now based on the above, we get the following corollary which will directly imply the margin bounds we are aiming for. The bound for the $p = 2$ case has been used to explain the performance of SVMs. Our bound essentially matches the best known bound [Bartlett and Mendelson, 2002] which was an improvement over previous bounds [Bartlett and Shawe-Taylor, 1999] proved using fat-shattering dimension estimates. For the L_∞/L_1 case, our bound improves the best known bound [Schapire et al., 1998] by removing a factor of $\sqrt{\log n}$.

Corollary 6. *(L_p Margins) Each of the following statements holds with probability at least $1 - \delta$ over the sample:*

- *Let $\mathcal{W}$ be as in the L_p/L_q* **norms** *example. For all $\gamma > 0$, $\mathbf{w} \in \mathcal{W}$,*

$$\mathcal{L}(\mathbf{w}) \leq K_\gamma(\mathbf{w}) + 4\frac{XW_*}{\gamma}\sqrt{\frac{p-1}{n}} + \sqrt{\frac{\log(\log_2 \frac{4XW_*}{\gamma})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

- *Let $\mathcal{W}$ be as in the L_∞/L_1* **norms** *example. For all $\gamma > 0$, $\mathbf{w} \in \mathcal{W}$,*

$$\mathcal{L}(\mathbf{w}) \leq K_\gamma(\mathbf{w}) + 4\frac{XW_1}{\gamma}\sqrt{\frac{2\log(d)}{n}} + \sqrt{\frac{\log(\log_2 \frac{4XW_1}{\gamma})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

The following result improves the best known results of the same kind, [Langford et al., 2001, Theorem 5] and [Zhang, 2002, Theorem 7], by removing a factor of $\sqrt{\log n}$. These results themselves were an improvement over previous results obtained using fat-shattering dimension estimates.

Corollary 7. *(Entropy Based Margins) Let $\mathcal{X}$ be such that for all $\mathbf{x} \in \mathcal{X}$, $\|\mathbf{x}\|_\infty \leq X$. Consider the class $\mathcal{W} = \{w \in \mathbb{R}^d \ : \ \|\mathbf{w}\|_1 \leq W_1\}$. Fix an arbitrary prior μ. We have that with probability at least $1 - \delta$ over the sample, for all margins $\gamma > 0$ and all weight vector $\mathbf{w} \in \mathcal{W}$,*

$$\mathcal{L}(\mathbf{w}) \leq K_\gamma(\mathbf{w}) + 8.5\,\frac{XW_1}{\gamma}\sqrt{\frac{\text{entro}_\mu(\mathbf{w}) + 2.5}{n}} + \sqrt{\frac{\log(\log_2 \frac{4XW_1}{\gamma})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

where $\text{entro}_\mu(\mathbf{w}) := \sum_i \frac{|\mathbf{w}_i|}{\|\mathbf{w}\|_1}\log(\frac{|\mathbf{w}_i|}{\mu_i \|\mathbf{w}\|_1})$

Proof. Proof is provided in the appendix. □

4.3 PAC-Bayes Theorem

We now show that (a form of) the PAC Bayesian theorem [McAllester, 1999] is a consequence of Theorem 1. In the PAC Bayesian theorem, we have a set of hypothesis (possibly infinite) $\mathcal{C}$. We choose some prior distribution over this hypothesis set say μ, and after observing the training data, we choose any arbitrary posterior ν and the loss we are interested in is $\ell_\nu(\mathbf{x}, y) = \mathbb{E}_{c\sim\nu}\ell(c, \mathbf{x}, y)$ that is basically the expectation of the loss when hypothesis $c \in \mathcal{C}$ are drawn i.i.d. using distribution ν. Note that in this section we are considering a more general form of the loss.

The key observation as that we can view $\ell_\nu(\mathbf{x})$ as the inner product $\langle d\nu(\cdot), \ell(\cdot, \mathbf{x}, y)\rangle$ between the measure $d\nu(\cdot)$ and the loss $\ell(\cdot, x)$. This leads to the following straightforward corollary.

Corollary 8. *(PAC-Bayes) For a fixed prior μ over the hypothesis set $\mathcal{C}$, and any loss bounded by 1, with probability at least $1-\delta$ over the sample, simultaneously for all choice of posteriors ν over $\mathcal{C}$ we have that,*

$$\mathcal{L}_\nu \leq \hat{\mathcal{L}}_\nu + 4.5\sqrt{\frac{\max\{\mathrm{KL}(\nu\|\mu), 2\}}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{8}$$

Proof. Proof is provided in the appendix. □

Interestingly, this result is an improvement over the original statement, in which the last term was $\sqrt{\log(n/\delta)/n}$. Our bound removes this extra $\log(n)$ factor, so, in the regime where we fix ν and examine large n, this bound is sharper. We note that our goal was not to prove the PAC-Bayes theorem, and we have made little attempt to optimize the constants.

4.4 Covering Number Bounds

It is worth noting that using Sudakov's minoration results we can obtain upper bound on the L_2 (and hence also L_1) covering numbers using the Gaussian complexities. The following is a direct corollary of the Sudakov minoration theorem for Gaussian complexities (Theorem 3.18, Page 80 of Ledoux and Talagrand [1991]).

Corollary 9. *Let $\mathcal{F}_\mathcal{W}$ be the function class from Theorem 1. There exists a universal constant $K > 0$ such that its L_2 covering number is bounded as follows:*

$$\forall \epsilon > 0 \quad \log(\mathcal{N}_2(\mathcal{F}_\mathcal{W}, \epsilon, n)) \leq \frac{2K^2X^2W_*^2}{\sigma\epsilon^2}$$

This bound is sharper than those that could be derived from the $\mathcal{N}_\infty$ covering number bounds of Zhang [2002].

4.5 Fast Rates

Let F be 1-strongly convex w.r.t. the L_q norm $\|\cdot\|_q$. Define

$$\mathcal{L}_\lambda(\mathbf{w}) := \mathbb{E}\left[\ell(\langle \mathbf{w}, \mathbf{x}\rangle, y)\right] + \lambda F(\mathbf{w}) \,,$$
$$\hat{\mathcal{L}}_\lambda(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n}\ell(\langle \mathbf{w}, \mathbf{x}_i\rangle, y_i) + \lambda F(\mathbf{w}) \,,$$

and $\mathbf{w}^\star := \operatorname{argmin} \mathcal{L}_\lambda(\mathbf{w})$ and $\hat{\mathbf{w}} := \operatorname{argmin} \hat{\mathcal{L}}_\lambda(\mathbf{w})$. Note that these are unconstrained minimizers over all of $\mathbb{R}^d$. Then equations (3) and (4) combined with the results recently proved in Anonymous [2008], we obtain the following guarantee with probability at least $1-\delta$,

$$\mathcal{L}_\lambda(\hat{\mathbf{w}}) - \mathcal{L}_\lambda(\mathbf{w}^\star) \leq O\left(\frac{(p-1) + \log(1/\delta)}{\lambda n}\right)$$

Note that the above statements help analyze the regularized minimizer (Equation 1) directly.

5 Discussion: Relations to Online, Regret Minimizing, Algorithms

In this section, we make a further assumption that loss $\ell(\langle \mathbf{w}, \mathbf{x}\rangle, y)$ is convex in $\mathbf{w}$. We now show that in the online setting that the regret bounds for linear prediction closely match our risk bounds. The algorithm we consider performs the update,

$$\mathbf{w}_{t+1} = \nabla F^{-1}(\nabla F(\mathbf{w}_t) - \eta \nabla_\mathbf{w}\ell(\langle \mathbf{w}_t, \mathbf{x}_t\rangle, y_t)) \tag{9}$$

This algorithm captures both gradient updates, multiplicative updates, and updates based on the L_p norms, through appropriate choices of F. See Shalev-Shwartz [2007] for discussion.

For the algorithm given by the above update, the following theorem is a bound on the cumulative regret. It is a corrolary of of Theorem 1 in Shalev-Shwartz and Singer [2006] (and also of Corollary 1 in Shalev-Shwartz [2007]), applied to our linear case.

Corollary 10. *(Shalev-Shwartz and Singer [2006]) Let S be a closed convex set and let $F : S \to \mathbb{R}$ be σ-strongly convex w.r.t. $\|\cdot\|_*$. Further, let $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\| \le X\}$ and $\mathcal{W} = \{\mathbf{w} \in S : F(\mathbf{w}) \le W_*^2\}$. Then for the update given by Equation 9 if we start with $\mathbf{w}_1 = \operatorname{argmin} F(\mathbf{w})$, we have that for all sequences $\{(\mathbf{x}_t, y_t)\}_{t=1}^n$,*

$$\sum_{t=1}^{n} \ell(\langle \mathbf{w}_t, \mathbf{x}_t \rangle, y_t) - \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} \sum_{t=1}^{n} \ell(\langle \mathbf{w}, \mathbf{x}_t \rangle, y_t) \le L_\ell X W_* \sqrt{\frac{2n}{\sigma}}$$

For completeness, we provide a direct proof in the Appendix. Interestingly, the regret above is precisely our complexity bounds (when $L_\ell = 1$). Also, our risk bounds are a factor of 2 worse, essentially due to the symmetrization step used in proving Theorem 3.

References

A. Anonymous. Fast rates for regularized objectives, 2008. submitted to NIPS 2008.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

P. L. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, pages 43–54. MIT Press, 1999.

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

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

J. Langford and J. Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems 15*, pages 423–430, 2003.

J. Langford, M. Seeger, and Nimrod Megiddo. An improved predictive accuracy bound for averaging classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 290–297, 2001.

M. Ledoux and M. Talagrand. *Probability in Banach spaces: Isoperimetry and processes*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3)*. Springer-Verlag, 1991.

David A. McAllester. Simplified PAC-Bayesian margin bounds. In *Proceedings of the Sixteenth Annual Conference on Computational Learning Theory*, pages 203–215, 2003.

David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 164–170, 1999.

A.Y. Ng. Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.

David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, October 1998.

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. Technical report, The Hebrew University, 2006. Available at http://www.cs.huji.ac.il/∼shais.

M. Warmuth and A. K. Jagota. Continuous versus discrete-time non-linear gradient descent: Relative loss bounds and convergence. In *Fifth International Symposium on Artificial Intelligence and Mathematics*, 1997.

T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

Appendix

Proof of Theorem 5. Let $\ell_\gamma(t)$ be defined as

$$\ell_\gamma(t) := \begin{cases} 1 & t \le 0\,, \\ 1 - \frac{t}{\gamma} & 0 < t < \gamma\,, \\ 0 & t \ge \gamma\,. \end{cases}$$

For $i = 0, 1, \ldots$, set $\gamma_i = C/2^i$ and $\delta_i = \delta/(i+1)^2$. Let $\mathcal{L}_\gamma$ and $\hat{\mathcal{L}}_\gamma$ denote the expected and empirical loss when the loss function is ℓ_γ. Let $\mathcal{L}$ and $\hat{\mathcal{L}}$ denote the corresponding quantities for the 0-1 loss. Applying Theorem 3 to the loss function ℓ_{γ_i} which has Lipschitz constant $1/\gamma_i$, we get, with probability at least $1 - \delta_i$, for all $f \in \mathcal{F}$,

$$\mathcal{L}_{\gamma_i}(f) \le \hat{\mathcal{L}}_{\gamma_i}(f) + \frac{2}{\gamma_i}\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta_i)}{2n}}\,.$$

Note that for any $\gamma > 0$ and any f, $\mathcal{L}(f) \le \mathcal{L}_\gamma(f)$ and $\hat{\mathcal{L}}_\gamma \le K_\gamma(f)$. So with probability at least $1 - \delta_i$, for all $f \in \mathcal{F}$,

$$\mathcal{L}(f) \le K_{\gamma_i}(f) + \frac{2}{\gamma_i}\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta_i)}{2n}}\,.$$

Taking union bound over $i \ge 0$, we get that with probability at least $1 - \pi^2\delta/6 \ge 1 - 2\delta$, for all $i \ge 0$ and all $f \in \mathcal{F}$,

$$\mathcal{L}(f) \le K_{\gamma_i}(f) + \frac{2}{\gamma_i}\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta_i)}{2n}}\,.$$

Suppose the above event indeed happens. Note that, since $|f(x)| \le C$, $\hat{\mathcal{L}}_C(f) = 1$ and so that bound claimed in the Theorem holds trivially for $\gamma \ge C$. For $\gamma < C$ find the i such that $\gamma_i \le \gamma < \gamma_{i-1}$. Note that this means $i \le \log_2(C/\gamma) + 1$. Since $K_{\gamma_i}(f) \le K_\gamma(f)$, $1/\gamma_i \le 2/\gamma$ and $\log(1/\delta_i) \le \log(1/\delta) + 2\log(\log_2(4C/\gamma))$, we have

$$\mathcal{L}(f) \le K_\gamma(f) + \frac{4}{\gamma}\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta) + 2\log(\log_2(4C/\gamma))}{2n}}\,.$$

□

Proof of Corollary 7. Consider the linear function class $\mathcal{F}_\mathcal{W}$ where $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^d \,:\, \mathrm{entro}_\mu(\mathbf{w}) \le E,\ \|\mathbf{w}\|_1 \le W_1\}$. Note that our Theorem 1 cannot bound the Rademacher complexity of this class directly as $\mathrm{entro}_\mu(\mathbf{w})$ is not strongly convex on an L_1 ball. But there is simple trick to deal with this. Simply write $\langle \mathbf{w}, \mathbf{u}\rangle$ as $\langle \tilde{\mathbf{w}}, \overline{\mathbf{u}}\rangle$ where $\overline{\mathbf{u}} = (\mathbf{u}, -\mathbf{u})$, $\tilde{\mathbf{w}} = (\mathbf{w}_+, \mathbf{w}_-)$ and $\mathbf{w}_+$ and $\mathbf{w}_-$ are the positive and negative parts of $\mathbf{w}$. Note that $\tilde{\mathbf{w}} \ge 0$ and $\|\tilde{\mathbf{w}}\|_1 = \|\mathbf{w}\|_1$. Let $\mu' = (\mu/2, \mu/2)$. Since $\|\overline{\mathbf{u}}\|_\infty = \|\mathbf{u}\|_\infty$ and $\mathrm{entro}_{\mu'}(\tilde{\mathbf{w}}) = \mathrm{entro}_\mu(\mathbf{w}) + \log 2$, we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{\substack{\|\mathbf{w}\|_1 \le W_1 \\ \mathrm{entro}_\mu(\mathbf{w}) \le E}} \frac{1}{n}\left\langle \mathbf{w}, \sum_{i=1}^n \epsilon_i \mathbf{x}_i \right\rangle\right] &\le \mathbb{E}\left[\sup_{\substack{\|\tilde{\mathbf{w}}\|_1 \le W_1,\ \tilde{\mathbf{w}} \ge 0 \\ \mathrm{entro}_{\mu'}(\tilde{\mathbf{w}}) \le E + \log 2}} \frac{1}{n}\left\langle \tilde{\mathbf{w}}, \overline{\sum_{i=1}^n \epsilon_i \mathbf{x}_i} \right\rangle\right] \\
&\le \mathbb{E}\left[\sup_{\substack{\|\tilde{\mathbf{w}}\|_1 = W_1,\ \tilde{\mathbf{w}} \ge 0 \\ \mathrm{entro}_{\mu'}(\tilde{\mathbf{w}}) \le E + \log 2}} \frac{1}{n}\left\langle \tilde{\mathbf{w}}, \sum_{i=1}^n \epsilon_i \overline{\mathbf{x}_i} \right\rangle\right] \\
&\le X W_1 \sqrt{\frac{2(E + \log 2)}{n}}\,,
\end{aligned}$$

where we have applied (4) in the last step.

Setting ℓ to be the 0-1 loss and using the Theorem 5 with probability at least $1 - \delta$, for all $\mathbf{w} \in \mathcal{W}$,

$$\mathcal{L}(\mathbf{w}) \le K_\gamma(\mathbf{w}) + \frac{4}{\gamma} X W_1 \sqrt{\frac{2(E + \log 2)}{n}} + \sqrt{\frac{\log(\log_2(X W_1/\gamma))}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Now using the union bound technique used in the proof of Corollary 8 below we prove that with probability $1-\delta$ the above statement holds simultaneously for all $j \geq 0$ with $E = a2^j$ and $\delta_j = \delta/2^{j+1}$. Then, for all $\mathbf{w}$ with $\|\mathbf{w}\|_1 \leq W_1$, we have

$$\begin{aligned}\mathcal{L}(\mathbf{w}) \leq & K_\gamma(\mathbf{w}) + \frac{8}{\gamma} X W_1 \sqrt{\frac{\max\{\mathrm{entro}_\mu(\mathbf{w}), 2\} + \frac{\log 2}{2})}{n}} \\ & + \sqrt{\frac{\log(\max\{\mathrm{entro}_\mu(\mathbf{w}), 2\})}{2n}} + \sqrt{\frac{\log(\log_2(XW_1/\gamma))}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}\end{aligned}$$

Now since $8\sqrt{\max\{x,2\} + \frac{\log 2}{2}} + \sqrt{\log(\max\{x,2\})/2} \leq 8.5\sqrt{\max\{x,2\} + \frac{\log 2}{2}}$, $\max\{x,2\} \leq x+2$ and $\log 2 \leq 1$, we get the result. □

Proof of Corollary 8. Note that by Theorem 3 we have, that for some fixed prior μ (over concept class $\mathcal{C}$), with probability at least $1-\delta$ over sample, for all distributions ν over concept class $\mathcal{C}$ such that $\mathrm{KL}(\nu\|\mu) \leq E$ we have that

$$\mathcal{L}_\nu \leq \hat{\mathcal{L}}_\nu + 2\mathbb{E}\left[\sup_{\nu\,:\,\mathrm{KL}(\nu\|\mu)\leq E} \frac{1}{n}\sum_{i=1}^n \ell_\nu(\mathbf{x}_i, y_i)\epsilon_i\right] + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Now note that $\ell_\nu(\mathbf{x}_i, y_i)$ can be written as $\ell_\nu(\mathbf{x}_i, y_i) = \langle d\nu(\cdot), \ell(\cdot, \mathbf{x}_i, y_i)\rangle$. Further note that $\mathrm{KL}(\nu\|\mu)$ is strongly convex with respect to the L_1 norm and we assumed loss is bounded by 1 and so we have,

$$\mathcal{L}_\nu \leq \hat{\mathcal{L}}_\nu + 2\sqrt{\frac{2E}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Now to get the required statement we take a union bound over possible values of E. To do this consider some $a > 0$, let $\Gamma_0 = \{\nu\ :\ \mathrm{KL}(\nu\|\mu) \leq a\}$ and for $j \geq 1$, let $\Gamma_j = \{\nu\ :\ a2^{j-1} < \mathrm{KL}(\nu\|\mu) \leq a2^j\}$. Also for any $\delta > 0$ let $\delta_j = 2^{-(j+1)}\delta$. Clearly $\sum_{j=0}^\infty \delta_j = \delta$ and hence union bound gives us that with probability at least $1-\delta$ over sample, for any distribution ν,

$$\mathcal{L}_\nu \leq \hat{\mathcal{L}}_\nu + 2\sqrt{\frac{2(a\,2^j)}{n}} + \sqrt{\frac{\log(1/\delta_j)}{2n}}$$

where j is the smallest index such that $\mathrm{KL}(\nu\|\mu) \leq a2^j$. However note that by our choice of partitions we have that for any $j \geq 1$, $a2^j \leq 2\,\mathrm{KL}(\nu\|\mu)$. Also note that

$$\delta_j = 2^{-(j+1)}\delta \geq \frac{a\,\delta}{4\,\mathrm{KL}(\nu\|\mu)}$$

For $j = 0$ clearly $\delta_j = \delta/2$ and $a2^j = a$ and so we get that with probability at least $1-\delta$ over sample, for any distribution ν,

$$\mathcal{L}_\nu \leq \hat{\mathcal{L}}_\nu + 4\sqrt{\frac{\max\{\mathrm{KL}(\nu\|\mu), a/2\}}{n}} + \sqrt{\frac{\log(\max\{\frac{4\,\mathrm{KL}(\nu\|\mu)}{a}, 2\})}{2n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Now take, say, $a = 4$ and note that $4\sqrt{\max\{x,2\}} + \sqrt{\log(\max\{x,2\})/2} \leq 4.5\sqrt{\max\{x,2\}}$ and so we get that

$$\mathcal{L}_\nu \leq \hat{\mathcal{L}}_\nu + 4.5\,\sqrt{\frac{\max\{\mathrm{KL}(\nu\|\mu), 2\}}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

□

Proof of Theorem 10. Warmuth and Jagota [1997] showed that for the algorithm given by Equation 9 the following bound holds,

$$\sum_{j=1}^n \ell(\langle \mathbf{w}_j, \mathbf{x}_j\rangle, y_j) \leq \sum_{j=1}^n \ell(\langle \mathbf{u}, \mathbf{x}_j\rangle, y_j) + \frac{1}{\eta}\Delta_F(\mathbf{u}\|\mathbf{w}_1) + \frac{1}{\eta}\sum_{j=1}^n \Delta_F(\mathbf{w}_{t+1}\|\mathbf{w}_t) \quad (10)$$

Now since $\mathbf{w}_1 = \operatorname{argmin} F(\mathbf{w})$ we see that $\nabla F(\mathbf{w}_1) = 0$. Hence we have

$$\begin{aligned}\Delta_F(\mathbf{u}\|\mathbf{w}_1) &= F(\mathbf{u}) - F(\mathbf{w}_1) + \langle \nabla F(\mathbf{w}_1), \mathbf{u} - \mathbf{w}_1 \rangle \\ &= F(\mathbf{u}) - F(\mathbf{w}_1) \\ &\leq F(\mathbf{u}) \leq W_*^2 \end{aligned} \tag{11}$$

To analyze the second term in the bound we first note that (Proposition 11.1 of Cesa-Bianchi and Lugosi [2006]).

$$\Delta_F(\mathbf{w}_j\|\mathbf{w}_{j+1}) = \Delta_{F^*}(\nabla F(\mathbf{w}_j)\|\nabla F(\mathbf{w}_{j+1}))$$

Now by Equation 7 we see that

$$\Delta_{F^*}(\nabla F(\mathbf{w}_j)\|\nabla F(\mathbf{w}_{j+1})) \leq \frac{1}{2\sigma}\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{j+1})\|^2$$

However by the update rule (9) we have that we have that

$$\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{j+1}) = \eta \nabla_{\mathbf{w}} \ell(\langle \mathbf{w}_j, \mathbf{x}_j \rangle, y_j)$$

Hence,

$$\Delta_{F^*}(\nabla F(\mathbf{w}_j)\|\nabla F(\mathbf{w}_{j+1})) \leq \frac{\eta^2}{2\sigma}\|\nabla_{\mathbf{w}} \ell(\langle \mathbf{w}_j, \mathbf{x}_j \rangle, y_j)\|^2$$

Now clearly

$$\nabla_{\mathbf{w}} \ell(\langle \mathbf{w}_j, \mathbf{x}_j \rangle, y_j) = \frac{\partial \ell(\langle \mathbf{w}_j, \mathbf{x}_j \rangle, y_j)}{\partial \langle \mathbf{w}_j, \mathbf{x}_j \rangle}\mathbf{x}_j$$

Hence we have that

$$\|\nabla_{\mathbf{w}} \ell(\langle \mathbf{w}_j, \mathbf{x}_j \rangle, y_j)\|^2 \leq L_\ell^2 X^2$$

Therefore we get that

$$\|\nabla_{\mathbf{w}} \ell(\langle \mathbf{w}_j, \mathbf{x}_j \rangle, y_j)\|^2 \leq \frac{\eta^2 L_\ell^2 X^2}{2\sigma}$$

Now using the above and (11) in the bound given in (10) we get that

$$\sum_{j=1}^{n} \ell(\langle \mathbf{w}_j, \mathbf{x}_j \rangle, y_j) \leq \sum_{j=1}^{n} \ell(\langle \mathbf{u}, \mathbf{x}_j \rangle, y_j) + \frac{1}{\eta} W_*^2 + \eta \frac{n L_\ell^2 X^2}{2\sigma}$$

Now using $\eta = \frac{W_* \sqrt{2\sigma}}{\sqrt{n} L_\ell X}$ we get the required statement. □

Lemma 11. *Let $\|\cdot\|, \|\cdot\|_*$ be a pair of dual norms defined on $\mathbb{R}^d$ that are twice differentiable. If $\|\cdot\|$ is $(2, D)$ smooth then $\|\cdot\|_*$ is $\frac{2}{D^2}$-strongly convex w.r.t. itself.*

Proof. Let $F(\mathbf{x})$ denote $\|\mathbf{x}\|^2$. Then the convex conjugate $F^*(\mathbf{w})$ of F is $\frac{1}{4}\|\mathbf{w}\|_*^2$. Now we have,

$$\nabla F(\nabla F^*(\mathbf{w})) = \mathbf{w} \ .$$

Applying chain rule for vector differentiation, we get

$$\nabla^2 F(\nabla F^*(\mathbf{w})) \cdot \nabla^2 F^*(\mathbf{w}) = \mathbf{I} \ . \tag{12}$$

By definition of $(2, D)$-smoothness of $\|\cdot\|$,

$$\forall \mathbf{y}, \mathbf{x},\ \langle \mathbf{y}, \nabla^2 F(\mathbf{x}) \mathbf{y} \rangle \leq 2D^2 \|\mathbf{y}\|^2 \ .$$

Set $\mathbf{x} = \nabla F^*(\mathbf{w})$ and $\mathbf{A} = \nabla^2 F(\mathbf{x})/2D^2$, so that we have,

$$\forall \mathbf{y}, \|\mathbf{y}\|_{\mathbf{A}}^2 \leq \|\mathbf{y}\|^2 \ ,$$

where $\|\mathbf{y}\|_A := \sqrt{\langle \mathbf{y}, \mathbf{A}\mathbf{y} \rangle}$. The dual of $\|\cdot\|_{\mathbf{A}}$ is $\|\cdot\|_{\mathbf{A}^{-1}}$. Moreover, if a norm dominates another norm then its dual is in turn dominated by the dual of the other. So, we get

$$\forall \mathbf{v},\ \|\mathbf{v}\|_{\mathbf{A}^{-1}}^2 \geq \|\mathbf{v}\|_*^2 \ .$$

From (12), $\mathbf{A}^{-1} = 2D^2 \nabla^2 F^*(\mathbf{w})$. Plugging this in, we get

$$\forall \mathbf{v},\ \langle \mathbf{v}, 2D^2 \nabla^2 F^*(\mathbf{w}) \mathbf{v} \rangle \geq \|\mathbf{v}\|_*^2 \ .$$

Now, noting that that $F^*(\mathbf{w}) = \frac{1}{4}\|\mathbf{w}\|_*^2$, we get

$$\forall \mathbf{v},\ \langle \mathbf{v}, \nabla^2 \|\mathbf{w}\|_*^2\, \mathbf{v} \rangle \geq \frac{2}{D^2}\|\mathbf{v}\|_*^2 \ ,$$

which proves the lemma. □