A Short Survey of Generalization and Oracle Bounds Obtained for a Lipschitz Objective with Regularization

Thierry Bazier-Matte

February 12, 2017

1 Introduction, Notation and Asumptions

This document is an attempt at regrouping under a unified notation and assumptions a number of results from machine learning, statistical learning and high-dimensional statistics when considering a loss function of the form $(q, z) \mapsto \ell(y\, q^T x)$, with ℓ a γ-Lipschitz loss function. This particular form of loss readlily applies to single-asset portfolio optimization (where one maximizes $u(r\, q^T x)$ and to SVM objective (where the loss is given by $(1 - y\, q^T x)_+$). We first define a number of solutions whose properties will be studied in the following sections:

$$R(q) = \boldsymbol{E}_{X,Y}\, \ell(Y\, q^T X)\,; \tag{1}$$

$$\hat{R}(q) = n^{-1} \sum_{i=1}^{n} \ell(y_i\, q^T x_i)\,; \tag{2}$$

$$\hat{q}_2 = \arg\min_q \left\{ \hat{R}(q) + \lambda_2 \|q\|_2^2 \right\}\ ; \tag{3}$$

$$q_2^\star = \arg\min_q \left\{ R(q) + \lambda_2 \|q\|_2^2 \right\}\ ; \tag{4}$$

$$\hat{q}_{12} = \arg\min_q \left\{ \hat{R}(q) + \lambda_1 \|q\|_1 + \lambda_2 \|q\|_2^2 \right\}\ ; \tag{5}$$

$$q_{12}^\star = \arg\min_q \left\{ R(q) + \lambda_1 \|q\|_1 + \lambda_2 \|q\|_2^2 \right\}\ ; \tag{6}$$

In particular, we will be interested in establishing generalization bounds g_1 and oracle bounds g_2, that is,

$$R(\hat{q}) \leq \hat{R}(\hat{q}) + g_1(n, p, \delta)$$

and

$$R(\hat{q}) \leq R(q^\star) + g_2(n, p, \delta).$$

In particular, we will try to make the bounds explicitly depending on the sample size n, the dimensionality of the problem p (that is, the cardinality of X) and the confidence level $1-\delta$.

We now formalize a number of assumptions.

Assumption 1. *The loss function ℓ is γ-Lipschitz, i.e., $|\ell(z_1)-\ell(-z_2)| \leq \gamma|z_1-z_2|$.*

Assumption 2. *The random variable Y rests on a bounded support, i.e., $|Y| \leq \bar{y}$.*

Assumption 3. *Every feature vector is standardized, i.e., for any $j \in \{1,\dots,p\}$, $\boldsymbol{E}X_i = 0$, $\operatorname{Var} X_i = 1$. In particular, this implies that $\boldsymbol{E}X_i^2 = 1$, so that $\boldsymbol{E}\|X\|_2^2 = \sum_{i=1}^p \boldsymbol{E}X_i^2 = p$.*

2 Lemmas

In this section we prove a number of lemmas whose results will be recurrent in proving various theorems presented in the following sections.

Lemma 1. *The following bound holds with probability $1-\delta_X$:*

$$\|X\|_2^2 \leq \frac{p}{\delta_X}.$$

Let $\xi^2 = O(p)$ denote this bound.

Proof. The result simply follows from the fact that $\boldsymbol{E}\|X\|_2^2 = p$ and by applying Markov's inequality. □

Remark. Note that this result can be made more precise if one has a better understanding of the features. For example, if one considers only bounded features, or only subgaussian features then Hoeffding's or Bernstein's inequalities apply [**Todo: Is it really Bernstein?**], thus providing tighter guarantees. However, we want to stress that the $\|X\|_2^2 = O(p)$ is the important result to take away from Lemma 1.

Lemma 2. *Let $\hat{q}$ denote either $\hat{q}_2$ or $\hat{q}_{12}$. Then with probability $1-\delta_X$, the following bound holds:*

$$\|\hat{q}\|_2 \leq \frac{\gamma\bar{y}\xi}{2\lambda_2}.$$

Let $B_q = O(\sqrt{p})$ denote this bound.

Corollary 1. *The loss suffered from applying either $\hat{q}_2$ or $\hat{q}_{12}$ is bounded by*

$$\ell(y\,\hat{q}^T x) \leq \gamma\bar{y}B_q\xi.$$

Let $B_\ell = O(p)$ denote this bound.

[Todo: Add proof.]

Lemma 3. *This lemma concerns specifically the $\hat{q}_2$ case. Let $R_\lambda(q) = R(q) + \lambda\|q\|_2^2$ and let $q_\lambda^\star$ be the theoretical minimizer of R_λ. Then R_λ is 2λ-strongly convex, i.e.,*

$$\lambda\|q - q_\lambda^\star\|_2^2 \leq R_\lambda(q) - R_\lambda(q_\lambda^\star)$$

Furthermore, if there exists a function g such that

$$R_\lambda(\hat{q}) - R_\lambda(q_\lambda^\star) \leq g(n, p, \delta),$$

then we have the following oracle bound on R:

$$R(\hat{q}) \leq R(q^\star) + \lambda\|q^\star\|_2^2 + 2g + 2\lambda B_{q_2}\sqrt{g/\lambda}.$$

Proof. First note that from the second hypothesis and the triangle inequality we have

$$R(\hat{q}) - R(q_\lambda^\star) \leq g + \lambda\left(\|q_\lambda^\star\|_2^2 - \|\hat{q}\|_2^2\right) \leq g + \lambda\left(2\|\hat{q}\|_2\|q_\lambda^\star - \hat{q}\|_2 + \|q_\lambda^\star - \hat{q}\|_2^2\right).$$

Next, the second hypothesis combined with the first one yields that $\|\hat{q} - q_\lambda^\star\|_2 \leq \sqrt{g/\lambda}$. Also, Lemma 2 implies that $\|\hat{q}\|_2$ and $\|q_\lambda^\star\|_2$ are bounded by B_{q_2}, so therefore

$$R(\hat{q}) - R(q_\lambda^\star) \leq 2g + 2\lambda B_{q_2}\sqrt{g/\lambda}.$$

Next, by definition of $q_\lambda^\star$, we have

$$R(q_\lambda^\star) + \lambda\|q_\lambda^\star\|_2^2 \leq R(q^\star) + \lambda\|q^\star\|_2^2,$$

it follows that

$$R(q_\lambda^\star) - R(q^\star) \leq \lambda\|q^\star\|_2^2 - \lambda\|q_\lambda^\star\|_2^2 \leq \lambda\|q^\star\|_2^2,$$

which combined to the equality

$$R(\hat{q}) = R(q^\star) + R(\hat{q}) - R(q_\lambda^\star) + R(q_\lambda^\star) - R(q^\star)$$

yields the claimed result. □

3 Generalization Bounds

This section applies for any $\hat{q}$, although because $\|\hat{q}_1\|$ is possibly unbounded, the results presented are only finite when $\hat{q}$ is either $\hat{q}_2$ or $\hat{q}_{12}$.

3.1 Pseudo-Dimension Bound

The first generalization bound relies on the concept of pseudo-dimension Pdim defined for a family of functions. Even though we shall not define it precisely, we will make use of some theorems. For references on the concept, please consult [6, 1, 8].

First, some definitions are in order.

Definition. The family of linear decisions $\mathcal{Q}$ is the following set of functions:

$$\mathcal{Q} = \{q : (\mathcal{X}, \mathcal{Y}) \to \mathscr{R} \,|\, q(x, y) = y\, q^T x\}.$$

Remark. In particular, $\hat{q}^T \in \mathcal{Q}$.

Definition. The familiy of losses $\mathcal{L}$ associated to $\mathcal{Q}$ is the following set of functions:

$$\mathcal{L} = \{\ell_q : (\mathcal{X}, \mathcal{Y}) \to \mathscr{R} \,|\, \ell_q(x, y) = \ell(q(x, y))\}.$$

Proposition 1. $\mathrm{Pdim}(\mathcal{Q}) = p + 1$.

Proof. Theorem 10.4 from [6] indicates that the family of hyperplanes in $\mathscr{R}^m$, *i.e.*,

$$\mathcal{W} = \{x \mapsto w^T x\}$$

has $\mathrm{Pdim}(\mathcal{W}) = m + 1$. But $\mathcal{Q}$ can also be considered as the family of hyperplanes since it only differs by a scaling factor of y. This yields the claimed result. □

Proposition 2. $\mathrm{Pdim}(\mathcal{L}) = p + 1$.

Proof. We see that $\mathcal{L} = \ell \circ \mathcal{Q}$. But by Exercise 10.1 of [6], since ℓ is monotonic, the result follows. □

Theorem 1. *Let $\hat{q}$ be either $\hat{q}_2$ or $\hat{q}_{12}$. Then the following bound holds with probability $1 - \delta$:*

$$\begin{aligned} R(\hat{q}) &\leq \hat{R}(\hat{q}) + B_{q_2}\left(\sqrt{\frac{2p\log(en/p)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}\right) \\ &\leq \hat{R}(\hat{q}) + O\left(\frac{p\sqrt{p\log(n/p)}}{n}\right), \end{aligned}$$

with e the Euler constant.

Proof. See Theorem 10.6 from [6] in conjunction with the two previous propositions. □

3.2 Rademacher Bound

Here again we will use the concept of Rademacher complexity $\hat{\mathfrak{R}}_n$ without defining it.

Let $\mathcal{F}_{\tilde{\mathcal{Q}}}$ be a slightly different family of functions taking in their input only from the $\mathcal{X}$ space:

$$\mathcal{F}_{\tilde{\mathcal{Q}}} = \{x \mapsto q^T x : q \in \tilde{\mathcal{Q}}\},$$

and

$$\tilde{\mathcal{Q}} = \{q \in \mathscr{R}^p : \|q\|_2 \leq B_q\}.$$

We will use results from [5] and [2] in order to derive our next bound.

Lemma. *The Rademacher complexity of $\tilde{\mathcal{Q}}$ is bounded:*

$$\hat{\mathfrak{R}}_n(\mathcal{F}_{\tilde{\mathcal{Q}}}) \leq \xi B_q \sqrt{\frac{1}{n}}.$$

Proof. This results comes from Theorem 3 in [5]. Here is a verbatim transcription of the theorem.

Theorem. *Let $\mathcal{F}_{\mathcal{W}} = \{x \mapsto \langle w, x\rangle : w \in \mathcal{W}\}$. Let S be a closed convex set and let $F : S \to \mathscr{R}$ be a σ-strongly convex w.r.t. $\|\cdot\|_\star$ s.t. $\inf_{w\in S} F(w) = 0$. Further, let $\mathcal{X} = \{x : \|x\| \leq X\}$. Define $\mathcal{W} = \{w \in S : F(w) \leq W_\star^2\}$. Then, we have*

$$\hat{\mathfrak{R}}_n(\mathcal{F}_{\mathcal{W}}) \leq X W_\star \sqrt{\frac{2}{\sigma n}}.$$

We can directly apply the above theorem to our case by taking $F = \|\cdot\|^2$ and noticing it is 2-strongly convex. Therefore, we have $\mathcal{W} = \tilde{\mathcal{Q}}$ if we set $W_\star^2 = B_q^2$, which leads to the claimed result. □

Next, we inoke generalization theorem proved by Bartlett and Mendelson in 2002 [2].

Theorem 2. *With probability $1 - \delta$, the following bound holds:*

$$\begin{aligned} R(\hat{q}) &\leq \hat{R}(\hat{q}) + \frac{2\gamma\xi B_q}{\sqrt{n}} + B_\ell \sqrt{\frac{\log(1/\delta)}{2n}} \\ &\leq \hat{R}(\hat{q}) + O\left(\frac{p}{\sqrt{n}}\right). \end{aligned}$$

Proof. See Theorem 1 in [5] in addition to Corollary 1. □

3.3 Stability Bound

The next bound was introduced by [3].

Lemma. *The β-stability of the ERM algorithm leading to $\hat{q}$ is bounded:*

$$\beta \leq \frac{(\gamma\bar{y}\xi)^2}{\lambda_2 n} = O\left(\frac{p}{n}\right).$$

Proof. See Proposition 11.1 in [6] with $\sigma = \gamma\bar{y}$. □

Theorem 3. *With probability $1-\delta$, the following bound holds:*

$$\begin{aligned} R(\hat{q}) &\leq \hat{R}(\hat{q}) + \beta + (2n\beta + B_\ell)\sqrt{\frac{log(1/\delta)}{2n}} \\ &\leq \hat{R}(\hat{q}) + O\left(\frac{p}{\sqrt{n}}\right). \end{aligned}$$

Proof. Directly by applying Theorem 11.1 from [6] with the above lemma and Corollary 1. □

4 Oracle Bounds

4.1 Fast Rate Bound

The next bound is due to [7]. The "fast" rate means that $\hat{q} \to q_\lambda^\star$ at a $O(1/n)$ rate due to the strong convexity of the regularizer.

Theorem 4. *The following bound holds with probability $1-\delta$:*

$$\begin{aligned} R(\hat{q}) &\leq R(q^\star) + \lambda\|q^\star\|_2^2 + \frac{8\gamma^2\xi^2(32+\log(1/\delta))}{\lambda n} + 8\gamma\lambda\xi B_q\sqrt{\frac{32+\log(1/\delta)}{\lambda n}} \\ &\leq R(q^\star) + \lambda\|q^\star\|_2^2 + O\left(\frac{p}{\sqrt{n}}\right). \end{aligned}$$

Proof. As shown in Theorem 1 of [7], the following bound holds with probability $1-\delta$:

$$R_\lambda(\hat{q}) \leq R_\lambda(q_\lambda^\star) + \frac{4\gamma^2\xi^2(32+\log(1/\delta))}{\lambda n}.$$

Applying the result of Lemma 3 yields the result. □

4.2 Empirical Process Bound

The next bound has been developed using the machinery developed for the high dimensional statistical theory. Note that unlike the previous bounds where the p dependancy was a consequence of $\xi = O(\sqrt{p})$, the dependancy on p relies on the so-called contraction inequality. Another particularity of the bound is its looseness: whereas previous bounds holded with exponential confidence [**Todo: Rephrase.**], this bound is actually concerns the expectation of the suboptimality, and Markov's inequality only yields a weak bound.

Theorem 5. *The following bound holds with probability* $1 - \delta$*:*

$$\begin{aligned}
R(\hat{q}) &\leq R(q^\star) + \lambda \|q^\star\|_2^2 + \frac{512\gamma^2}{\lambda\delta^2}\frac{p}{n} + 32\lambda B_q \frac{\gamma}{\delta\sqrt{\lambda}}\sqrt{\frac{p}{n}} \\
&\leq R(q^\star) + \lambda \|q^\star\|_2^2 + O\left(\frac{p}{\sqrt{n}}\right).
\end{aligned}$$

Proof. Follows from the quadratic margin of $\mathscr{E}_2$ (with constant λ_2) around $q_2^\star$, Lemma 6.6 and Lemma 14.19 in [4] in conjunction with Markov's inequality, followed by Lemma 3. [**Todo: More details?**] □

4.3 Elastic Net Penalization

Although high dimensional statistics was developed with a "fixed number of *real* features, high number of unimportant features" philosophy (especially useful in certain fields like biology), we can use its results on the solution $\hat{q}_{12}$, but we have to suppose that all considered features are useful. This leads to a worsening of $O(\log p)$ of previous bounds. This is because, by considering $q_\lambda^\star$ as the objective, we know that none of its components is zero, almost surely. Although we could also consider a single linear penalty (thus leading to a lasso generalized linear model), our lack of knowledge of curvature at $q^\star$ would yield a bound with unknown constant.

Theorem 6. *Let*

$$\lambda_1 = 16\gamma\left(4\sqrt{\frac{2\log 2p}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}\right).$$

Then the following bound holds with probability $1 - \delta$*:*

$$\begin{aligned}
R(\hat{q}) &\leq R(q_\lambda^\star) + \lambda_2 \|q^\star\|_2^2 + \frac{32\lambda_1^2 p}{\lambda_2} + 4\lambda_2 B_q \frac{\lambda_1}{\sqrt{\lambda_2}}\sqrt{p} \\
&\leq R(q_\lambda^\star) + \lambda_2 \|q^\star\|_2^2 + O\left(\frac{p\log p}{\sqrt{n}}\right).
\end{aligned}$$

Proof. The proof first follows from

$$R(\hat{q}) - R(q_\lambda^\star) \leq \frac{16\lambda_1 p}{\lambda_2}$$

(Corrolary 6.3 in [4] using quadratic margin with constant $c = \lambda_2$) in conjunction to Example 14.2 also in [4]. $\square$

References

[1] M. Anthony and P. L. Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[2] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[3] O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[4] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[5] 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*, pages 793–800, 2009.

[6] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[7] K. Sridharan, S. Shalev-Shwartz, and N. Srebro. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems*, pages 1545–1552, 2009.

[8] M. Vidyasagar. *Learning and generalisation: with applications to neural networks*. Springer Science & Business Media, 2013.