Fast Convergence Rates for Excess Regularized Risk with Application to SVM

Karthik Sridharan, Nathan Srebro, Shai Shalev-Shwartz

1 Introduction

We consider the stochastic minimization problem

$$\operatorname*{argmin}_{\mathbf{w}\in\mathbf{W}} F(\mathbf{w}) \tag{1}$$

where

$$F(\mathbf{w}) = \mathbf{E}_\theta \left[f(\mathbf{w};\theta) \right] \tag{2}$$

is the expectation of a random function $f(\cdot;\theta)$, and the optimization is based on a sample $f(\cdot;\theta_1),\ldots,f(\cdot;\theta_n)$ of i.i.d. instantiations of $f(\cdot;\theta)$. Instead of thinking of a "random function", one can think of a fixed and known function $f(\mathbf{w};\theta)$, and a distribution over the parameter θ. A special case is the familiar prediction setting where $\theta = (\mathbf{x}, y)$ is an instance-label pair and, e.g., $f(\mathbf{w};\mathbf{x},y) = \ell(\langle \mathbf{w},\mathbf{x}\rangle, y)$ for some loss function ℓ. However, we would like to emphasize that we do not assume any metric or other structure on the parameter θ.

In this paper, we are given a random sample of functions (i.e. parameters θ) and can find the empirical minimizer

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \hat{F}(\mathbf{w}) \tag{3}$$

where $\hat{F}$ is the empirical average

$$\hat{F}(\mathbf{w}) = \hat{\mathbb{E}}\left[f(\mathbf{w};\theta) \right] = \tfrac{1}{n}\sum_{i=1}^{n} f(\mathbf{w};\theta_i). \tag{4}$$

Our guarantees are then on the expected value $F(\hat{\mathbf{w}})$ (generalization performance in the prediction setting) of the empirical minimizer. In fact, we do not need to assume that we perform the minimization (3) exactly: we provide uniform (over all $\mathbf{w} \in \mathbf{W}$) guarantees on the expected regret

$$F(\mathbf{w}) - F(\mathbf{w}^\star) \tag{5}$$

in terms of the empirical regret

$$\hat{F}(\mathbf{w}) - \hat{F}(\mathbf{w}^\star). \tag{6}$$

This is a stronger type of result than what can be obtained with online-to-batch conversions, as it applies to any possible solution $\mathbf{w}$, and not only some specific algorithmically defined solution. For example, it can be used to analyze the performance of approximate minimizers obtained through approximate optimization techniques.

2 Concentration of Strongly convex Linear Functions

We first introduce two conditions under which we show that the rate of convergence is of order $1/n$. The first condition assumes that the function at hand can be decomposed into a part only dependent on parameter $\mathbf{w}$ and into a part that is Lipschitz w.r.t. $\langle \mathbf{w}, P(\theta) \rangle$ where P is a fixed function of θ

Condition 1. *For all $\theta \in \Theta$ and $\mathbf{w} \in \mathbf{W}$:*

$$f(w, \theta) = h(w, \theta) + N(w)$$

where N is a function of $\mathbf{w}$ alone and $\forall \theta$, $\|P(\theta)\| \leq B_x$ Also for all $\theta \in \Theta$ and $\mathbf{w}_1, \mathbf{w}_2 \in \mathbf{W}$:

$$|h(\mathbf{w}_1, \theta) - h(\mathbf{w}_2, \theta)| \leq L \left|\langle \mathbf{w}_1 - \mathbf{w}_2, P(\theta) \rangle\right|$$
$$|N(\mathbf{w}_1) - N(\mathbf{w}_2)| \leq k \|\mathbf{w}_1 - \mathbf{w}_2\|$$

Note that the condition implies that $|f(\mathbf{w}_1, \theta) - f(w_2, \theta)| \leq (LB_x + k)\|w_1 - w_2\|$ The second condition we require is that the function in expectation is strongly convex.

Condition 2. *$F(\mathbf{w})$ is λ-strongly convex. That is, for all $\mathbf{w}_1, \mathbf{w}_2 \in \mathbf{W}$:*

$$F(\tfrac{\mathbf{w}_1 + \mathbf{w}_2}{2}) + \tfrac{\lambda}{8} \left\|\mathbf{w}_1 - \mathbf{w}_2\right\|^2 \leq \tfrac{1}{2}(F(\mathbf{w}_1) + F(\mathbf{w}_2))$$

Condition 2 ensures that:

$$F(\mathbf{w}) \geq F(\mathbf{w}^\star) + \tfrac{\lambda}{2} \left\|\mathbf{w} - \mathbf{w}^\star\right\|^2 \tag{7}$$

for all $\mathbf{w} \in \mathbf{W}$. Recall that $\mathbf{w}^\star = \arg\min_{\mathbf{w}} F(\mathbf{w})$. Our results will actually only use (7).

Condition 2 only requires that the expected function $F(\mathbf{w})$ is strongly convex. Of course, requiring that $f(\mathbf{w}; \theta)$ is λ-strongly convex for all θ (with respect to $\mathbf{w}$) is enough to ensure the condition.

Theorem 1. *If $f(\mathbf{w}, \theta)$ and $\mathbf{W}$ satisfy Conditions 1 and 2, then, for any $\delta > 0$ and any $a > 0$, with probability of at least $1 - \delta$ over the sample, we have that for all $\mathbf{w} \in \mathbf{W}$:*

$$F(\mathbf{w}) - F(\mathbf{w}^\star) \leq (1 + a)(\hat{F}(\mathbf{w}) - \hat{F}(\mathbf{w}^\star)) + O\left(\frac{(1 + \frac{1}{a})L^2(1 + \log(1/\delta))}{\lambda n}\right).$$

as well as:

$$\hat{F}(\mathbf{w}) - \hat{F}(\mathbf{w}^\star) \leq (1 + a)(F(\mathbf{w}) - F(\mathbf{w}^\star)) + O\left(\frac{(1 + \frac{1}{a})L^2(1 + \log(1/\delta))}{\lambda n}\right).$$

2.1 Proof of Theorem 1

To prove the theorem we use techniques of reweighing and peeling following **?** (**?**). First, we need the following definitions. Let $g : \mathbf{W} \times \Theta \to \mathbb{R}$ be the function

$$g(\mathbf{w}, \theta) = f(\mathbf{w}; \theta) - f(\mathbf{w}^\star; \theta) \ . \tag{8}$$

We also define the class of functions

$$\mathcal{G} = \{g_{\mathbf{w}}(\theta) = g(\mathbf{w}, \theta) : w \in \mathbf{W}\} \ . \tag{9}$$

We next define a weighted class of functions. To do so, let $\mathbb{Z}_+$ be the set of non-negative integers and let $\gamma \geq 1$ be some fixed scalar to be set later. For any $r > 0$ we define

$$\mathcal{G}_r \ = \ \left\{ \tfrac{g}{\gamma^k} : g \in \mathcal{G}, k = \min\{k' \in \mathbb{Z}_+ : \mathbb{E}\left[g\right] \leq r\gamma^{k'}\} \right\} \tag{10}$$

In other words, for any function $g \in \mathcal{G}$, we have a function $g_r \in \mathcal{G}_r$ such that g_r is just a scaled version of g and the scaling factor ensures that

$$\mathbb{E}\left[g_r\right] \leq r \ . \tag{11}$$

The following lemma upper bounds the range of the functions in $\mathcal{G}_r$.

Lemma 2. *Assume that the conditions stated in Theorem 1 hold and let $\mathcal{G}_r$ be as defined in* (10)*. Then, for all $g_r \in \mathcal{G}_r$ and $\theta \in \Theta$ we have $|g_r(\theta)| \leq L\sqrt{2r/\lambda}$.*

Proof. Using the definition of g_r we have that there exists $w \in \mathbf{W}$ such that

$$g_r(\theta) = \frac{f(\mathbf{w}, \theta) - f(\mathbf{w}^\star, \theta)}{\gamma^k} \ .$$

Since $f(\mathbf{w}, \theta)$ is L-Lipschitz w.r.t. w we get that

$$|f(\mathbf{w}, \theta) - f(\mathbf{w}^\star, \theta)| \leq L \left\|w - w^\star\right\| .$$

Combining the above two equations we obtain that,

$$|g_r(\theta)| \ \leq \ \frac{L \left\|w - w^\star\right\|}{\gamma^k} \ . \tag{12}$$

Next, we obtain a bound on $\|w - w^\star\|$. To do so, we rewrite equation (7):

$$\|\mathbf{w} - \mathbf{w}^\star\| \leq \sqrt{\tfrac{2}{\lambda}(F(\mathbf{w}) - F(\mathbf{w}^\star))} \ . \tag{13}$$

Using again the definition of F we know that there exists $g \in \mathcal{G}$ such that $g_r = g/\gamma^k$ and

$$F(\mathbf{w}) - F(\mathbf{w}^\star) = \mathbb{E}\left[g\right] = \gamma^k \, \mathbb{E}\left[g_r\right] . \tag{14}$$

Using (11) we know that $\mathbb{E}\left[g_r\right] \leq r$. Combining this with (14) and (13) we obtain that

$$\|\mathbf{w} - \mathbf{w}^\star\| \leq \sqrt{\tfrac{2}{\lambda}\gamma^k \, r} \ .$$

Combining the above with (12) yields

$$|g_r(\theta)| \;\leq\; \frac{L\sqrt{2r/\lambda}}{\gamma^{k/2}} \,.$$

Finally, we note that $\gamma > 1$ which concludes our proof. □

Denote $c(r) = L\sqrt{2r/\lambda}$. Using Theorem 3.2 of **?** (**?**) we have that with probability of at least $1-\delta$,

$$\sup_{g_r \in \mathcal{G}_r} |\mathbb{E}\left[g_r\right] - \hat{\mathbb{E}}\left[g_r\right]| \leq 2\mathbb{E}\left[R_n(\mathcal{G}_r)\right] + c(r)\sqrt{\frac{2\log(1/\delta)}{n}} \,. \tag{15}$$

The following lemma bounds the Rademacher complexity $R_n(\mathcal{G}_r)$ of the class $\mathcal{G}_r$.

Lemma 3. $R_n(\mathcal{G}_r) \;=\; O\left(\frac{L\,\log^2(n)\,\sqrt{r}}{\sqrt{\lambda\, n}}\right) \,.$

Proof. We prove the lemma using Lemma **??** and the peeling technique (**?**). First, it is convenient to center g and define the function $\tilde{g} : \tilde{\mathbf{W}} \times \Theta \to \mathbb{R}$ where $\tilde{\mathbf{W}} = \{\tilde{\mathbf{w}} : (\mathbf{w}^\star - \tilde{\mathbf{w}}) \in W\}$ and

$$\tilde{g}(\tilde{\mathbf{w}}, \theta) := f(\mathbf{w}^\star - \tilde{\mathbf{w}}, \theta) - f(\mathbf{w}^\star, \theta) = g(\mathbf{w}^\star - \tilde{\mathbf{w}}, \theta) \,.$$

We also use the notation $\tilde{G} \;=\; \{\tilde{g}_{\tilde{\mathbf{w}}}(\theta) \;=\; \tilde{g}(\tilde{\mathbf{w}}, \theta) \;:\; \tilde{\mathbf{w}} \;\in\; \tilde{\mathbf{W}}\}$ and $\tilde{G}_r \;=\; \left\{\frac{\tilde{g}_{\tilde{\mathbf{w}}}}{\gamma^{k(\tilde{\mathbf{w}})}} : \tilde{\mathbf{w}} \in \tilde{\mathbf{W}}\right\}$ where

$$k(\tilde{\mathbf{w}}) \;=\; \min\{k' \in \mathbb{Z}_+ \;:\; \mathbb{E}\left[g(\mathbf{w}^\star - \tilde{\mathbf{w}}, \cdot)\right] \leq r\gamma^{k'}\} \,.$$

Comparing the definitions of $\tilde{G}_r$ and $\mathcal{G}_r$ we note that these two classes of functions are identical and thus $R_n(\mathcal{G}_r) = R_n(\tilde{G}_r)$. Therefore, from now on we focus on the problem of upper bounding $R_n(\tilde{G}_r)$.

It is convenient to divide the class $\tilde{G}$ into rings according to the expectation. Formally, for any $0 < a < b$ we define

$$\tilde{G}(a,b) \;=\; \{\tilde{g}_{\tilde{\mathbf{w}}} : \mathbb{E}\left[\tilde{g}_{\tilde{\mathbf{w}}}\right] \in [a,b]\} \,.$$

We also let

$$\tilde{\mathbf{W}}(a,b) \;=\; \{\tilde{\mathbf{w}} \in \tilde{\mathbf{W}} : \mathbb{E}\left[\tilde{g}_{\tilde{\mathbf{w}}}\right] \in [a,b]\} \,.$$

Consider the set $\tilde{\mathbf{W}}(0,r')$ for some arbitrary $r' > 0$. If $\tilde{\mathbf{w}} \in \tilde{\mathbf{W}}(0,r')$ then we have

$$r' \geq \mathbb{E}\left[\tilde{g}_{\tilde{\mathbf{w}}}\right] = \mathbb{E}\left[g_{\mathbf{w}^\star - \tilde{\mathbf{w}}}\right] = F(\mathbf{w}^\star - \tilde{\mathbf{w}}) - F(\mathbf{w}^\star) \,.$$

Using (7) we know that

$$F(\mathbf{w}^\star - \tilde{\mathbf{w}}) - F(\mathbf{w}^\star) \geq \tfrac{\lambda}{2}\|\tilde{\mathbf{w}}\|^2 \,.$$

Therefore, overall we have shown that

$$\forall \tilde{\mathbf{w}} \in \tilde{\mathbf{W}}(0,r'), \;\; \|\tilde{\mathbf{w}}\| \leq \sqrt{\tfrac{2r'}{\lambda}} \,.$$

This establishes Condition **??**, with $B = \sqrt{2r'/\lambda}$, for $\tilde{\mathbf{W}}$. It is straightforward to verify that since the Lipschitz and convexity Condition 1 holds for f it also holds for $\tilde{g}$ (strictly speaking, we should be discussing its convex hull $\mathrm{conv}\,\tilde{\mathbf{W}}$, on which the norm is of course also bounded, so that convexity inside it can be properly defined). We can therefore apply Lemma **??** and get that there exists a constant C such that

$$R_n(\tilde{G}(0,r')) \;\leq\; \frac{C\,L\,\sqrt{r'}\,\log^2(n)}{\sqrt{\lambda\,n}}\;. \tag{16}$$

We now get back to the problem of bounding $R_n(\tilde{G}_r)$. To do so, we note that

$$\tilde{G}_r \;=\; \cup_{j=0}^{\infty}\,\{\tfrac{\tilde{g}_{\tilde{\mathbf{w}}}}{\gamma^j} : \tilde{\mathbf{w}} \in \tilde{\mathbf{W}}, k(\tilde{\mathbf{w}}) = j\}\;.$$

Therefore

$$\begin{aligned} R_n(\tilde{G}_r) &\leq \sum_{j=0}^{\infty} R_n\left(\{\tfrac{\tilde{g}_{\tilde{\mathbf{w}}}}{\gamma^j} : \tilde{\mathbf{w}} \in \tilde{\mathbf{W}}, k(\tilde{\mathbf{w}}) = j\}\right) \\ &\leq \sum_{j=0}^{\infty} \gamma^{-j} R_n(\tilde{G}(0, r\,\gamma^j)) \\ &\leq \frac{C\,L\,\log^2(n)\,\sqrt{r}}{\sqrt{\lambda\,n}} \sum_{j=0}^{\infty} \gamma^{-j/2} \end{aligned}$$

Finally, setting $\gamma = 4$ concludes our proof. □

Equipped with the above We are now ready to prove Theorem 1.

Proof of Theorem 1 *contd.* Given any $r > 0$, note that for any $g \in \mathcal{G}$ we have for the corresponding $g_r = g/\gamma^k \in \mathcal{G}_r$ such that,

$$\mathbb{E}\,[g_r] - \hat{\mathbb{E}}\,[g_r] \leq \sup_{g'_r \in \mathcal{G}_r} (\mathbb{E}\,[g'_r] - \hat{\mathbb{E}}\,[g'_r])\;.$$

and so

$$\mathbb{E}\,[g] - \hat{\mathbb{E}}\,[g] \leq \gamma^k \sup_{g'_r \in \mathcal{G}_r} (\mathbb{E}\,[g'_r] - \hat{\mathbb{E}}\,[g'_r])\;.$$

Now we consider two possible cases, the first when $k = 0$ and the second when $k > 0$. If $k = 0$ then

$$\mathbb{E}\,[g] - \hat{\mathbb{E}}\,[g] \leq \sup_{g'_r \in \mathcal{G}_r} (\mathbb{E}\,[g'_r] - \hat{\mathbb{E}}\,[g'_r])\;. \tag{17}$$

When $k > 0$ then, since $k = \min\{k' \in \mathbb{Z}_+ : \mathbb{E}\,[g] \leq r\gamma^{k'}\}$, we have that $\gamma^{k-1} r \leq \mathbb{E}\,[g]$. Hence,

$$\mathbb{E}\,[g] - \hat{\mathbb{E}}\,[g] \leq \frac{\gamma \mathbb{E}\,[g]}{r} \sup_{g'_r \in \mathcal{G}_r} (\mathbb{E}\,[g'_r] - \hat{\mathbb{E}}\,[g'_r])\;. \tag{18}$$

For a given $K > 1$, we will select an $r > 0$ such that

$$\sup_{g'_r \in \mathcal{G}_r} \left(\mathbb{E}\left[g'_r\right] - \hat{\mathbb{E}}\left[g'_r\right] \right) \leq \tfrac{r}{\gamma K} \,. \tag{19}$$

In that case when $k = 0$, from (17) we have that

$$\mathbb{E}\left[g\right] - \hat{\mathbb{E}}\left[g\right] \leq \tfrac{r}{\gamma K} \,.$$

and in the case when $k > 0$ we have using (18) that

$$\mathbb{E}\left[g\right] - \hat{\mathbb{E}}\left[g\right] \leq \tfrac{\mathbb{E}[g]}{K}$$

or equivalently

$$\mathbb{E}\left[g\right] \leq \tfrac{K}{K-1} \hat{\mathbb{E}}\left[g\right] .$$

Combining both the case when $k = 0$ and $k > 0$ we see that

$$\mathbb{E}\left[g\right] \leq \tfrac{K}{K-1} \hat{\mathbb{E}}\left[g\right] + \tfrac{r}{\gamma K} \,. \tag{20}$$

All we need to do now is to find an r that satisfiescan the condition given in (19). To do so, we use (15), the upper bound on $c(r)$ given in Lemma 2, and the upper bound on the Rademacher complexity given in Lemma 3. We get that there exists some constant C such that with probability of at least $1 - \delta$ we have

$$\sup_{g_r \in \mathcal{G}_r} \left| \mathbb{E}\left[g_r\right] - \hat{\mathbb{E}}\left[g_r\right] \right| \leq C\, L \sqrt{\frac{r}{\lambda\, n}} \left(\log^2(n) + \sqrt{\log(\tfrac{1}{\delta})} \right)$$

Hence, the condition given in (19) is satisfied if the following holds

$$\frac{r}{\gamma k} \geq C\, L \sqrt{\frac{r}{\lambda\, n}} \left(\log^2(n) + \sqrt{\log(\tfrac{1}{\delta})} \right)$$

or equivalently if

$$r = \frac{(C \gamma K L)^2}{\lambda\, n} \left(\log^2(n) + \sqrt{\log(\tfrac{1}{\delta})} \right)^2 .$$

Substituting the above value of r into (20) yields

$$\mathbb{E}\left[g\right] \leq \tfrac{K}{K-1} \hat{\mathbb{E}}\left[g\right] + \frac{C^2 L^2 \gamma K}{\lambda\, n} \left(\log^2(n) + \sqrt{\log(\tfrac{1}{\delta})} \right)^2 .$$

Finally, writing $a = 1/(K-1)$ concludes the proof of the first inequality. For the second inequality we start by noting that for all $g \in \mathcal{G}$,

$$\hat{\mathbb{E}}\left[g\right] - \mathbb{E}\left[g\right] \leq \gamma^k \sup_{g'_r \in \mathcal{G}_r} \left(\left| \hat{\mathbb{E}}\left[g'_r\right] - \mathbb{E}\left[g'_r\right] \right| \right) .$$

and then essentially repeat the same steps as above with appropriate changes, $\square$

3 A $1/n$ Rate for the SVM Objective

We consider the problem of L_2-regularized linear prediction of target labels $y \in \mathcal{Y}$, using vectors $\mathbf{x}$ in a subset $\mathcal{X}$ of a Hilbert space, with $\|\mathbf{x}\| \leq B_x$. Predictions are given by $\langle \mathbf{w}, \mathbf{x} \rangle$, and we shall consider a convex loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ that is L_ℓ-Lipschitz in its first argument. An example of such a loss function is the hinge loss used in SVMs:

$$\ell(z, y) = [1 - yz]_+ = \max(1 - yz, 0) \tag{21}$$

In this section we aim at analyzing the rate of convergence of the regularized risk, e.g. the SVM objective, rather than the risk directly. To this end, we define the regularized objective function,

$$f(\mathbf{w}; \mathbf{x}, y) = \frac{\lambda}{2} \|\mathbf{w}\|^2 + \ell(\langle \mathbf{w}, \mathbf{x} \rangle, y) \tag{22}$$

As before, we use $F(\mathbf{w})$ and $\hat{F}(\mathbf{w})$ to denote the expectation and empirical averages of the objective function (22). When $\ell(\cdot, \cdot)$ is the hinge-loss, the empirical average $\hat{F}(\mathbf{w})$ is precisely the SVM training objective.

Since we take $\ell(z, y)$ to be convex in z, then $\mathbf{w} \mapsto \ell(\langle \mathbf{w}, \mathbf{x} \rangle, y)$ is also convex in $\mathbf{w}$ for all $\mathbf{x}, y$. Adding the λ-strongly convex function $\mathbf{w} \mapsto \frac{\lambda}{2} \|\mathbf{w}\|^2$, we see that f is λ-strongly convex w.r.t. $\mathbf{w}$, establishing Condition 2.

We will also need to assume that the loss is non-negative and that $\ell(0, y)$ is bounded for all y. This later condition is trivially true when the label space $\mathcal{Y}$ is finite, and holds for any reasonable loss function and a bounded label space. To reduce excess notation, we will just assume $\ell(0, y) \leq 1$, as is the case for the hinge-loss. This implies that

$$\mathbb{E}\left[\ell(\langle \mathbf{w}^\star, \mathbf{x} \rangle, y)\right] + \tfrac{\lambda}{2} \|\mathbf{w}^\star\|^2 = F(\mathbf{w}^\star) \leq F(0) \leq 1 \tag{23}$$

and hence we conclude that $\|\mathbf{w}^\star\| \leq \sqrt{\frac{2}{\lambda}}$. We will therefore restrict our analysis only to weight vectors whose norm is bounded by $\sqrt{\frac{2}{\lambda}}$. That is, we take:

$$\mathbf{W} = \left\{ \mathbf{w} \middle| \|\mathbf{w}\| \leq \sqrt{\tfrac{2}{\lambda}} \right\} \tag{24}$$

We already saw that $f(\mathbf{w}; \mathbf{x}, y)$ is convex w.r.t. $\mathbf{w}$. To bound the Lipschitz constant of f, note that $\mathbf{w} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$ is $\|x\|$-Lipschitz, and so $\mathbf{w} \mapsto \ell(\langle \mathbf{w}, \mathbf{x} \rangle; y)$ is $L_\ell \cdot B_x$-Lipschitz for all $\|x\| \leq B_x$. To this we must add the Lipschitz constant of $\frac{\lambda}{2} \|\mathbf{w}\|^2$, which is $\frac{\lambda}{2}(2 \sup \|\mathbf{w}\|) \leq \frac{\lambda}{2}(2\sqrt{\frac{2}{\lambda}}) = \sqrt{2\lambda}$. We therefore establish Condition 1 with

$$L \leq (L_\ell B_x + \sqrt{2\lambda})$$

Having established Conditions 1 and 2, we can can apply Theorem 1. We further note that the additional factor of $\log^2(n)$ in Theorem 1 results from bounding the Rademacher complexity using the fat shattering dimension. However, for objective functions of the form (22), we can bound the Rademacher complexity directly using

standard results on the Rademacher complexity of linear predictors, and avoid the spurious log-factors: In equation (16) of Lemma 3, instead of using Lemma **??** we can use standard results on linear predictors to bound $R_n(\tilde{G}(0,r')) \leq 2\frac{L\sqrt{r'}}{\sqrt{\lambda n}}$. We can therefore conclude:

Theorem 4. *Let f be of the form* (22)*, with ℓ convex, L_ℓ-Lipschitz, non-negative, and $\ell(0,y) \leq 1$, and $\|\mathbf{x}\| \leq B_x$. Then for any $a, \delta > 0$, with probability of at least $1-\delta$ over the sample, we have that for all $\mathbf{w}$ with $\|\mathbf{w}\| \leq \sqrt{2/\lambda}$:*

$$F(\mathbf{w}) - F(\mathbf{w}^\star) \leq (1+a)(\hat{F}(\mathbf{w}) - \hat{F}(\mathbf{w}^\star)) + O\left(\frac{(1+\frac{1}{a})(L_\ell^2 B_x^2 + \lambda)(\log(1/\delta))}{\lambda n}\right).$$

as well as:

$$\hat{F}(\mathbf{w}) - \hat{F}(\mathbf{w}^\star) \leq (1+a)(F(\mathbf{w}) - F(\mathbf{w}^\star)) + O\left(\frac{(1+\frac{1}{a})(L_\ell^2 B_x^2 + \lambda)(\log(1/\delta))}{\lambda n}\right).$$

Using the above we see that for SVMs (i.e. ℓ is the hinge loss) $L_\ell = 1$ and so for $\lambda = O(B_x^2)$ fixed, we can conclude that the generalization regret (in terms of the SVM objective) of the empirical optimizer $\hat{\mathbf{w}}$ converges as $O(\frac{B_x^2 \log(1/\delta)}{\lambda n})$. Studying the convergence rate of the SVM objective allows us to better understand and appreciate analysis of computational optimization approaches for this objective, as well as obtain oracle inequalities on the generalization loss of $\hat{\mathbf{w}}$.

4 Oracle Inequalities for SVMs

In this Section we apply the concentration results of Section 3 to obtain an oracle inequality on the generalization error of an approximate minimizer of the SVM objective. The basic setup is the same as in Section 3. We are now mostly concerned with the generalization error (i.e. expected loss—as we are dealing here only with convex losses, we slightly abuse standard terminology, and refer to the "loss" also as "error") :

$$\mathcal{L}(\mathbf{w}) = \mathbb{E}\left[\ell(\langle \mathbf{w}, \mathbf{x}\rangle, y)\right] \tag{25}$$

We assume, as an oracle assumption, that there exists a good predictor $\mathbf{w}_o$ with low norm $\|\mathbf{w}_o\|$ and which attains low generalization error $\mathcal{L}(\mathbf{w}_o)$. We now want to analyze the generalization error $\mathcal{L}(\hat{\mathbf{w}})$ of the empirical minimizer of the SVM objective $\hat{F}(\mathbf{w})$, or perhaps an approximate minimizer of this objective.

For any $\lambda > 0$ and $\mathbf{w} \in W$, we first decompose:

$$\begin{aligned}\mathcal{L}(\mathbf{w}) = \mathcal{L}(\mathbf{w}_o) &\\ + (F(\mathbf{w}) - F(\mathbf{w}^\star)) &\\ + (F(\mathbf{w}^\star) - F(\mathbf{w}_o)) &\\ + \tfrac{\lambda}{2}\|\mathbf{w}_o\|^2 - \tfrac{\lambda}{2}\|\mathbf{w}\|^2 &\end{aligned} \tag{26}$$

Consider an optimization algorithm for $\hat{F}(\mathbf{w})$ that is guaranteed to find $\tilde{\mathbf{w}}$ such that $\hat{F}(\tilde{\mathbf{w}}) \leq \min \hat{F}(\mathbf{w}) + \epsilon_{\text{opt}}$. Now, using Theorem 4 (ensuring $\lambda = O(B_x^2)$), and setting $a = 1$, we have, with probability at least $1 - \delta$:

$$F(\tilde{\mathbf{w}}) - F(\mathbf{w}^\star) \leq 2\epsilon_{\text{opt}} + O(\frac{B_x^2 \log(1/\delta)}{\lambda n}) \,. \tag{27}$$

Optimizing to within $\epsilon_{\text{opt}} = O(\frac{B_x^2}{\lambda n})$ is then enough to ensure

$$F(\tilde{\mathbf{w}}) - F(\mathbf{w}^\star) = O(\frac{B_x^2 \log(1/\delta)}{\lambda n}) \,. \tag{28}$$

We can use (28) to bound the second term of (26). The optimality of $\mathbf{w}^\star$ guarantees us that $F(\mathbf{w}^\star) - F(\mathbf{w}_o) \leq 0$ and certainly $\frac{\lambda}{2} \left\|\tilde{\mathbf{w}}\right\|^2 \geq 0$. We therefore obtain:

$$\mathcal{L}(\tilde{\mathbf{w}}) \leq \mathcal{L}(\mathbf{w}_o) + \tfrac{\lambda}{2} \left\|\mathbf{w}_o\right\|^2 + O\left(\frac{B_x^2 \log(1/\delta)}{\lambda n}\right) \tag{29}$$

This might seem like a rate of $1/n$ on the generalization error, but we need to choose λ so as to balance the second and third terms. The optimal choice for λ is

$$\lambda_n = O\left(\frac{B_x \sqrt{\log(1/\delta)}}{\left\|\mathbf{w}_o\right\| \sqrt{n}}\right) \,. \tag{30}$$

which yields:

$$\mathcal{L}(\tilde{\mathbf{w}}) \leq \mathcal{L}(\mathbf{w}_o) + O\left(\sqrt{\frac{B_x^2 \left\|\mathbf{w}_o\right\|^2 \log(1/\delta)}{n}}\right) \tag{31}$$

Even though we rely only on a single fixed $\mathbf{w}_o$, the choice of λ should change as the sample size increases.

It is interesting to repeat the analysis of this Section using the more standard result:

$$F(\mathbf{w}) - F(\mathbf{w}^\star) \leq \hat{F}(\mathbf{w}) - \hat{F}(\mathbf{w}^\star) + O\left(\sqrt{\frac{B_x^2}{\lambda n}}\right) \tag{32}$$

where we ignore the dependence on δ, and use $\left\|\mathbf{w}\right\| \leq \sqrt{2/\lambda}$. The bound (32) is immediate from concentration results on the unregularized loss, as the regularization terms cancel out. However, using (32) instead of Theorem 4 in our analysis yields the oracle inequality:

$$\mathcal{L}(\tilde{\mathbf{w}}) \leq \mathcal{L}(\mathbf{w}_o) + O\left(\left(\frac{B_x^2 \left\|\mathbf{w}_o\right\|^2 \log^2(1/\delta)}{n}\right)^{1/3}\right) \tag{33}$$

The oracle analysis studied here is very simple—our oracle assumption involves only a single predictor $\mathbf{w}_o$, and we make no assumptions about the kernel or the noise. We note that a more sophisticated analysis has been carried out by Steinwart et al.

(2006). Steinwart et al. shows that rates faster than $1/\sqrt{n}$ are possible under certain conditions on noise and complexity of kernel class. In Steinwart et al.'s analyses the estimation rates (i.e. rates for expected regularized risk) are given in terms of the approximation error quantity $\frac{\lambda}{2}\|\mathbf{w}^\star\|^2 + \mathcal{L}(\mathbf{w}^\star) - \mathcal{L}^*$ where $\mathcal{L}^*$ is the Bayes risk. In our result we consider the estimation rate for regularized objective independent of the approximation error.

References

Steinwart, I., Hush, D., & Scovel, C. (2006). A new concentration result for regularized risk minimizers. *High-dimensional Probability IV, in IMS Lecture Notes*, *51*, 260–275.