A Generalization bounds for unbounded losses

When the class of functions is not bounded, a single function can take arbitrarily large values with arbitrarily small probabilities. This is the main issue for deriving uniform convergence bounds for unbounded losses. This problem can be avoided either by assuming the existence of an envelope, that is a single non-negative function with a finite expectation lying above the absolute value of the loss of every function in the hypothesis set [12, 21, 13, 22, 15], or by assuming that some moment of the function losses is bounded [26, 27]. Our example in the simple case of Gaussians where the function w is exponential shows that no envelope function would be suitable for the problem of importance weighting. Thus, in view of the critical role played by the second moment of the importance weight, we have chosen to favor the assumption that the second moment is bounded, as in that example. A similar analysis can be given for other moments.

Here, we give two-sided generalization bounds for unbounded losses with finite second moments. The one-sided version of our bounds coincides with that of [26, 27] modulo a constant factor of $\sqrt{2}$, but the proofs given by Vapnik in both books seem to be incorrect.[1] The core component of our proof is based on a different technique, which is simpler and easy to check.

In what follows, we use the notation $\widehat{\Pr}$ to denote the empirical distribution based on a finite sample of size m, and $\widehat{\mathrm{E}}$ to denote the expectation based on $\widehat{\Pr}$. The following theorem reduces the problem of bounding $\sup_{h\in H}(\mathrm{E}[L_h]-\widehat{\mathrm{E}}[L_h])/\sqrt{\mathrm{E}[L_h^2]}$ to that of a standard relative deviation bound for classification.

Theorem 5. *For any loss function L (not necessarily bounded) and hypothesis set H such that $0<\mathrm{E}[L_h^2]<+\infty$ for all $h\in H$, the following two inequalities hold:*

$$\Pr\left[\sup_{h\in H}\frac{\mathrm{E}[L_h]-\widehat{\mathrm{E}}[L_h]}{\sqrt{\mathrm{E}[L_h^2]}}>\epsilon\sqrt{2+\log\frac{1}{\epsilon}}\right]\leq\Pr\left[\sup_{h\in H,t\in\mathbb{R}}\frac{\Pr[L_h>t]-\widehat{\Pr}[L_h>t]}{\sqrt{\Pr[L_h>t]}}>\epsilon\right].$$

$$\Pr\left[\sup_{h\in H}\frac{\widehat{\mathrm{E}}[L_h]-\mathrm{E}[L_h]}{\sqrt{\widehat{\mathrm{E}}[L_h^2]}}>\epsilon\sqrt{2+\log\frac{1}{\epsilon}}\right]\leq\Pr\left[\sup_{h\in H,t\in\mathbb{R}}\frac{\widehat{\Pr}[L_h>t]-\Pr[L_h>t]}{\sqrt{\widehat{\Pr}[L_h>t]}}>\epsilon\right].$$

Proof. We prove the first statement. The second statement can be shown in a very similar way.

Fix $\epsilon>0$ and assume that for any $h\in H$ and $t\geq 0$, the following holds:

$$\frac{\Pr[L_h>t]-\widehat{\Pr}[L_h>t]}{\sqrt{\Pr[L_h>t]}}\leq\epsilon. \tag{12}$$

We show that this implies that for any $h\in H$, $\frac{\mathrm{E}[L_h]-\widehat{\mathrm{E}}[L_h]}{\sqrt{\mathrm{E}[L_h^2]}}\leq\epsilon\sqrt{2+\log\frac{1}{\epsilon}}$. By the properties of the Lebesgue integral, we can write

$$\mathrm{E}[L_h]=\int_0^{+\infty}\Pr[L_h>t]\,dt \quad\text{and}\quad \widehat{\mathrm{E}}[L_h]=\int_0^{+\infty}\widehat{\Pr}[L_h>t]\,dt,$$

and similarly,

$$\mathrm{E}[L_h^2]=\int_0^{+\infty}\Pr[L_h^2>t]\,dt=\int_0^{+\infty}2t\Pr[L_h>t]\,dt.$$

In what follows, we use the shorter notation $I=\mathrm{E}[L_h^2]$. Let $t_1=\sqrt{\frac{I}{2}\frac{1}{\epsilon}}$. To bound $\mathrm{E}[L_h]-\widehat{\mathrm{E}}[L_h]$, we simply bound $\Pr[L_h>t]-\widehat{\Pr}[L_h>t]$ by $\Pr[L_h>t]$ for large values of t, that is $t>t_1$, and

---

[1] In [26][p.204-206], statement (5.37) cannot be derived from assumption (5.35), contrarily to what is claimed by the author, and in general does not hold: the first integral in (5.37) is restricted to a sub-domain and is thus smaller than the integral of (5.35). Furthermore, the main statement claimed in Section (5.6.2) is not valid. In [27][p.200-202], the author invokes the *Lagrange method* to show the main inequality, but the proof steps are not mathematically justified. Even with our best efforts, we could not justify some of the steps and strongly believe the proof not to be correct. In particular, the way function z is concluded to be equal to one over the first interval is suspicious and not based on a mathematical proof.

use inequality (12) for smaller values of t:

$$\mathrm{E}[L_h] - \widehat{\mathrm{E}}[L_h] = \int_0^{+\infty} \Pr[L_h > t] - \widehat{\Pr}[L_h > t]\, dt \leq \int_0^{t_1} \epsilon \sqrt{\Pr[L_h > t]} dt + \int_{t_1}^{+\infty} \Pr[L_h > t] dt.$$

For relatively small values of t, $\Pr[L_h > t]$ is close to one. Thus, if we define t_0 by $t_0 = \sqrt{\frac{I}{2}}$, we can write

$$\mathrm{E}[L_h] - \widehat{\mathrm{E}}[L_h] \leq \int_0^{t_0} \epsilon\, dt + \int_{t_0}^{t_1} \epsilon \sqrt{\Pr[L_h > t]} dt + \int_{t_1}^{+\infty} \Pr[L_h > t] dt = \int_0^{+\infty} f(t) g(t)\, dt,$$

with

$$f(t) = \begin{cases} (2I)^{1/4}\, \epsilon & \text{if } 0 \leq t \leq t_0 \\ \sqrt{2t \Pr[L_h > t]}\, \epsilon & \text{if } t_0 \leq t \leq t_1 \\ \sqrt{2t \Pr[L_h > t]}\, \epsilon & \text{if } t_1 \leq t. \end{cases} \qquad g(t) = \begin{cases} \frac{1}{(2I)^{1/4}} & \text{if } 0 \leq t \leq t_0 \\ \frac{1}{\sqrt{2t}} & \text{if } t_0 \leq t \leq t_1 \\ \sqrt{\frac{\Pr[L_h > t]}{2t}} \frac{1}{\epsilon} & \text{if } t_1 \leq t. \end{cases}$$

Now, by the Cauchy-Schwarz inequality,

$$\mathrm{E}[L_h] - \widehat{\mathrm{E}}[L_h] \leq \sqrt{\int_0^{+\infty} f(t)^2\, dt} \sqrt{\int_0^{+\infty} g(t)^2\, dt}.$$

The first integral on the right-hand side can be bounded as follows:

$$\int_0^{+\infty} f(t)^2\, dt = \int_0^{t_0} \sqrt{2I} \epsilon^2\, dt + \int_{t_0}^{+\infty} 2t \Pr[L_h > t] \epsilon^2\, dt \leq \sqrt{2I}\, t_0\, \epsilon^2 + \epsilon^2 I = 2\epsilon^2 I,$$

and, since $t_1/t_0 = 1/\epsilon$, the second one can be computed and bounded following

$$\begin{aligned} \int_0^{+\infty} g(t)^2\, dt &= \int_0^{t_0} \frac{dt}{\sqrt{2I}} + \int_{t_0}^{t_1} \frac{dt}{2t} + \int_{t_1}^{+\infty} \frac{\Pr[L_h > t]}{2t\epsilon^2} dt \\ &= \frac{1}{2} + \frac{1}{2} \log \frac{1}{\epsilon} + \int_{t_1}^{+\infty} \frac{2t \Pr[L_h > t]}{4t^2 \epsilon^2} dt \\ &\leq \frac{1}{2} + \frac{1}{2} \log \frac{1}{\epsilon} + \int_{t_1}^{+\infty} \frac{2t \Pr[L_h > t]}{4t_1^2 \epsilon^2} dt \leq \frac{1}{2} + \frac{1}{2} \log \frac{1}{\epsilon} + \frac{I}{4t_1^2 \epsilon^2} = 1 + \frac{1}{2} \log \frac{1}{\epsilon}. \end{aligned}$$

Combining the bounds obtained for these integrals yields directly

$$\mathrm{E}[L_h] - \widehat{\mathrm{E}}[L_h] \leq \sqrt{2\epsilon^2 I \left(1 + \frac{1}{2} \log \frac{1}{\epsilon}\right)} = \epsilon \sqrt{\left(2 + \log \frac{1}{\epsilon}\right)} \sqrt{I},$$

which concludes the proof of the theorem. □

We will use the following relative deviation bound for classification of [1, 27]. Such relative deviation results give sharper generalization bounds for binary classification.

Theorem 6 ([1]). *Let L be the binary classification loss. Then, for any hypothesis set H of real-valued functions, the following inequality holds:*

$$\Pr\left[\sup_{h \in H} \frac{R(h) - \widehat{R}(h)}{\sqrt{R(h)}} > \epsilon\right] \leq 4\Pi_H(2m) \exp\left(-\frac{m\epsilon^2}{4}\right),$$

where $\Pi_H(m)$ is the value of the growth function (maximum number of classifications) for a sample of size m, using the hypothesis set H.

It is not hard to show using the same proof as that of [1, 27] that a similar guarantee holds for the left side: $\Pr[\sup_{h \in H} \frac{\widehat{R}(h) - R(h)}{\sqrt{\widehat{R}(h)}} > \epsilon] \leq 4\Pi_H(2m) \exp\left(-\frac{m\epsilon^2}{4}\right)$. Combining these results with Theorem 5 yields directly the following.

Theorem 7. *Let H be a hypothesis set of real-valued functions and L a loss function (not necessarily bounded) such that for all $h \in H$, $0 < \mathrm{E}[L_h^2(x)] < +\infty$. Then, the following holds:*

$$\Pr\left[\sup_{h\in H} \frac{\mathrm{E}[L_h(x)] - \widehat{\mathrm{E}}[L_h(x)]}{\sqrt{\mathrm{E}[L_h^2(x)]}} > \epsilon\sqrt{2 + \log\frac{1}{\epsilon}}\right] \leq 4\Pi_H(2m)\exp\left(-\frac{m\epsilon^2}{4}\right).$$

$$\Pr\left[\sup_{h\in H} \frac{\widehat{\mathrm{E}}[L_h(x)] - \mathrm{E}[L_h(x)]}{\sqrt{\widehat{\mathrm{E}}[L_h^2(x)]}} > \epsilon\sqrt{2 + \log\frac{1}{\epsilon}}\right] \leq 4\Pi_H(2m)\exp\left(-\frac{m\epsilon^2}{4}\right).$$

Theorem 8. *Let H be a hypothesis set of real-valued functions and L a loss function (not necessarily bounded) such that for all $h \in H$, $0 < \mathrm{E}[L_h^2(x)] < +\infty$. Assume that that $\mathrm{Pdim}(\{L_h(x)\colon h \in H\}) = d < \infty$. Then, the following holds:*

$$\Pr\left[\sup_{h\in H} \frac{\mathrm{E}[L_h(x)] - \widehat{\mathrm{E}}[L_h(x)]}{\sqrt{\mathrm{E}[L_h^2(x)]}} > \epsilon\sqrt{2 + \log\frac{1}{\epsilon}}\right] \leq 4\exp\left(d\log\frac{2em}{d} - \frac{m\epsilon^2}{4}\right).$$

Proof. The results follows immediately by Sauer's lemma and the fact that the VC dimension of the family $\{\mathrm{sgn}(L_h(x)-t)\colon h \in H, t \in \mathbb{R}\}$ is precisely the pseudo-dimension of $\{L_h(x)\colon h \in H\}$. □

The following corollary gives a simpler form of this bound.

Corollary 1. *Let H be a hypothesis set of real-valued functions and L a loss function (not necessarily bounded) such that for all $h \in H$, $0<\mathrm{E}[L_h^2(x)]<+\infty$. Assume that that $\mathrm{Pdim}(\{L_h(x)\colon h\in H\})=d<\infty$. Then, the following holds:*

$$\Pr\left[\sup_{h\in H} \frac{\mathrm{E}[L_h(x)] - \widehat{\mathrm{E}}[L_h(x)]}{\sqrt{\mathrm{E}[L_h^2(x)]}} > \epsilon\right] \leq 4\exp\left(d\log\frac{2em}{d} - \frac{m\epsilon^{8/3}}{4^{5/3}}\right).$$

Proof. It is not hard to show that $3/4 = \min_\beta\{\beta\colon \forall\epsilon \in [0,1], \epsilon\sqrt{1 + \frac{1}{2}\log\frac{1}{\epsilon}} \leq e^\beta\}$ by studying the function $\epsilon \mapsto \epsilon\sqrt{1 + \frac{1}{2}\log\frac{1}{\epsilon}} - e^\beta$. This, combined with Theorem 7, gives the result. □

The following two-sided bound results directly from Corollary 1 and a similar bound for the other side that can be derived in the same way from Theorem 7.

Corollary 2. *Let H be a hypothesis set of real-valued functions and L a loss function (not necessarily bounded) such that for all $h \in H$, $0<\mathrm{E}[L_h^2(x)]<+\infty$. Assume that that $\mathrm{Pdim}(\{L_h(x)\colon h\in H\}) = d < \infty$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$, the following holds:*

$$\left|\mathrm{E}[L_h(x)] - \widehat{\mathrm{E}}[L_h(x)]\right| \leq 2^{5/4}\max\left\{\sqrt{\mathrm{E}[L_h^2(x)]}, \sqrt{\widehat{\mathrm{E}}[L_h^2(x)]}\right\} \sqrt[\frac{3}{8}]{\frac{d\log\frac{2me}{d} + \log\frac{8}{\delta}}{m}}.$$

B General lower bound based on maximum variance

Variants of the following result are known in the folklore of the learning theory community. We give a full proof below.

Theorem 9. *Let G denote a family of functions taking values in $[0,1]$. For $g\in G$, let $\sigma^2(g)$ denote the variance of g and $\sigma(G) = \sup_{g\in G}\sigma(g)$. Assume that $\frac{1}{m} \leq \sigma^2(G) < +\infty$. Then, the following inequality holds:*

$$\Pr\left[\sup_{g\in G}\left[\frac{\left|\,\mathrm{E}[g] - \widehat{\mathrm{E}}[g]\right|}{\sigma(G)}\right] \geq \frac{1}{2\sqrt{m}}\right] \geq \frac{2}{41^2}, \tag{13}$$

where the probability is taken over samples of size m.

Proof. Fix $\epsilon \in (0,1)$. By definition of the supremum, there exists $g \in G$ with $\sigma(g) \geq (1-\epsilon)\sigma(G)$. Let $Z = \frac{\mathrm{E}[g]-\widehat{\mathrm{E}}[g]}{\sigma(g)}$. By definition, $\mathrm{E}[Z^2] = \frac{1}{m^2}\,\mathrm{E}[\sum_{i=1}^{m} \frac{(\mathrm{E}[g]-g(x_i))^2}{\sigma^2(g)}] = \frac{1}{m}$. Thus,

$$\begin{aligned}\frac{1}{m} = \mathrm{E}[Z^2] &= \mathrm{E}[Z^2 1_{Z\in[0,1/(2\sqrt{m})]}] + \mathrm{E}[Z^2 1_{Z\in[1/(2\sqrt{m}),u/(\sqrt{m})]}] + \mathrm{E}[Z^2 1_{Z\geq u/(\sqrt{m})}] \\ &\leq \frac{1}{4m} + \frac{u^2}{m}\Pr[|Z| \geq 1/(2\sqrt{m})] + \mathrm{E}[Z^2 1_{Z>u/(\sqrt{m})}],\end{aligned}$$

which gives

$$\Pr[|Z| \geq 1/(2\sqrt{m})] \geq \frac{3}{4u^2} - \frac{m}{u^2}\,\mathrm{E}[Z^2 1_{Z>u/(\sqrt{m})}]. \tag{14}$$

Now, by the property of the Lebesgue integral,

$$\begin{aligned} m\,\mathrm{E}[Z^2 1_{Z>\frac{u}{\sqrt{m}}}] &= \int_0^{+\infty} \Pr[mZ^2 1_{Z>\frac{u}{\sqrt{m}}} > t]dt = \int_0^{u^2} \Pr\left[Z > \frac{u}{\sqrt{m}}\right]dt + \int_{u^2}^{+\infty} \Pr\left[Z > \sqrt{\frac{t}{m}}\right]dt \\ &= u^2 \Pr\left[Z > \sqrt{\frac{u^2}{m}}\right] + \int_{u^2}^{+\infty} \Pr\left[Z > \sqrt{\frac{t}{m}}\right]dt. \end{aligned}$$

By Bernstein's inequality,

$$\Pr\left[Z > \sqrt{\frac{t}{m}}\right] = \Pr[\left|\,\mathrm{E}[g] - \widehat{\mathrm{E}}[g]\right| > \sqrt{\frac{t}{m}}\sigma(g)] \leq \exp\left[\frac{-m\,\frac{t}{m}\,\sigma^2(g)}{2\sigma^2(g) + 2/3\sqrt{\frac{t}{m}}\sigma(g)}\right].$$

Using the assumption $1/\sqrt{m} \leq \sigma(g)$ gives

$$\Pr\left[Z > \sqrt{\frac{t}{m}}\right] \leq \exp\left[\frac{-t\sigma^2(g)}{2\sigma^2(g) + 2/3\sqrt{t}\sigma^2(g)}\right] = \exp\left[\frac{-t}{2 + 2/3\sqrt{t}}\right] \leq \exp(-3/8\sqrt{t}).$$

Thus,

$$m\,\mathrm{E}[Z^2 1_{Z>\frac{u}{\sqrt{m}}}] \leq u^2 e^{-3/8u} + \int_{u^2}^{+\infty} e^{-3/8\sqrt{t}}dt = u^2 e^{-3/8u} + \int_{u}^{+\infty} 2te^{-3/8t}dt.$$

An integration in parts leads to

$$m\,\mathrm{E}[Z^2 1_{Z>\frac{u}{\sqrt{m}}}] \leq (u^2 + 16/3u + 128/9)\exp(-3/8u).$$

For $u \geq 41/2$, $(u^2 + 16/3u + 128/9)\exp(-3/8u) \leq 1/4$. Thus, by (14), for $u \geq 41/2$,

$$\Pr[|Z| \geq 1/(2\sqrt{m})] \geq \frac{3}{4u^2} - \frac{1}{4u^2} = \frac{1}{2u^2}.$$

Thus, for all $\epsilon \in (0,1)$,

$$\Pr\left[\sup_{g\in G}\left[\frac{\left|\,\mathrm{E}[g] - \widehat{\mathrm{E}}[g]\right|}{\sigma(G)}\right] \geq \frac{1-\epsilon}{2\sqrt{m}}\right] \geq \frac{2}{41^2},$$

which concludes the proof. □

Acknowledgments

Author YM was at Google Research NY during part of the research. YM's work was supported in part by a grant from the MoS grant No. 3-6797, by a grant from the ISF (grant No. 709/09) and grant No. 2008-321 from the BSF, and by the IST Programme PASCAL2 NoE, IST-2007-216886. The work of author MM was partly supported by a Google Research Award.