Advanced Course in Machine Learning Spring 2010

Generalization bounds

Handouts are jointly prepared by Shie Mannor and Shai Shalev-Shwartz

The problem of characterizing learnability is the most basic question of learning theory. A fundamental and long-standing answer, formally proven for supervised classification and regression, is that learnability is equivalent to uniform convergence, and that if a problem is learnable, it is learnable via empirical risk minimization. Furthermore, for the problem of binary classification, uniform convergence is equivalent to finite VC dimension.

In this lecture we will talk about other methods for obtaining generalization bounds and establishing learnability. We start with PAC-Bayes bounds which can be though of as an extension to Minimum Description Length (MDL) bounds and Occam's razor. Next, we discuss a compression bound which states that if a learning algorithm only uses a small fraction of the training set to form its hypothesis then it generalizes. Finally, we turn to online-to-batch conversions. In the next lecture we will discuss the "General Learning Setting" (introduced by Vapnik), which includes most statistical learning problems as special cases.

1 Setup

We now return to the familiar learning setup. We assume that the data is distributed over some input space $\mathcal{X}$ and that the labels are in some space $\mathcal{Y}$. Generically, we equip $\mathcal{X} \times \mathcal{Y}$ with a probability distribution $\mathcal{D}$. We let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be some loss function. The hypothesis class is $\mathcal{H}$, where $h \in \mathcal{H}$ is a function from $\mathcal{X}$ to $\mathcal{Y}$. We define

$$L(h) = \mathbb{E}_{\mathcal{X} \times \mathcal{Y} \sim \mathcal{D}} \ell(h(x), y).$$

2 PAC-Bayes

There are several paradigms for preventing overfitting. One popular approach is to restrict the search space to a hypothesis class with bounded VC dimension or a bounded Rademacher complexity. Another approach is the Minimum Description Length (MDL) and Occam bounds in which we allow a potentially very large hypothesis class but define a hierarchy over hypotheses and prefer to choose hypotheses that appear higher in the hierarchy. The PAC-Bayesian approach further generalizes this idea.

As in the MDL paradigm, we define a hierarchy over hypotheses in our class $\mathcal{H}$. Now, the hierarchy takes the form of a prior distribution over $\mathcal{H}$. That is, we assign a probability (or density if $\mathcal{H}$ is continuous) $P(h) \geq 0$ for each $h \in \mathcal{H}$ and refer to $P(h)$ as the prior score of h. Following the Bayesian reasoning approach, the output of the learning algorithm is not necessarily a single hypothesis. Instead, the learning process defines a posterior probability over $\mathcal{H}$, which we denote Q. One can think on Q as defining a randomized prediction rule as follows. Whenever we get a new instance $\mathbf{x}$, we randomly pick a hypothesis $h \in \mathcal{H}$ according to Q and predict $h(\mathbf{x})$. We analyze the expected performance of the probabilistic prediction rule, namely, we are interested in bounding

$$\mathop{\mathbb{E}}_{h \sim Q} L(h) \ .$$

The following theorem tells us that the difference between the generalization loss and the empirical loss of a posterior Q is bounded by an expression that depends on the Kullback-Leibler divergence between Q and the prior distribution P. The Kullback-Leibler is a natural measure of the distance between two distributions and it has various usage in statistics and information theory. The theorem suggests that if we would like to minimize the generalization loss of Q we should jointly minimize both the empirical loss of Q and the Kullback-Leibler distance between Q and the prior distribution.

Theorem 1 *Let $\mathcal{D}$ be an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{H}$ be a hypothesis class and let ℓ be a loss function such that for all h and $\mathbf{z}$ we have $\ell(h, z) \in [0, 1]$. Let P be a prior distribution over $\mathcal{H}$ and let $\delta \in (0, 1)$. Then, with probability of at least $1 - \delta$ over the choice of an i.i.d. training set $S = \{z_1, \ldots, z_m\}$ sampled according to $\mathcal{D}$, for all distributions Q over $\mathcal{H}$ (even such that depend on S), we have*

$$\mathop{\mathbb{E}}_{h\sim Q}[L(h)] \le \mathop{\mathbb{E}}_{h\sim Q}[L_S(h)] + \sqrt{\frac{D(Q||P) + \ln m/\delta}{2(m-1)}} ,$$

where

$$D(Q||P) = \mathop{\mathbb{E}}_{h\sim Q}[\ln(Q(h)/P(h))]$$

is the Kullback-Leibler divergence.

Proof For any function $f(S)$, using Markov's inequality:

$$\mathop{\mathbb{P}}_{S}[f(S) \ge \epsilon] = \mathop{\mathbb{P}}_{S}[e^{f(S)} \ge e^{\epsilon}] \le \frac{\mathbb{E}_S[e^{f(S)}]}{e^{\epsilon}} . \tag{1}$$

Let $\Delta(h) = L(h) - L_S(h)$. We will apply Eq. (1) with the function

$$f(S) = \sup_{Q}\ 2(m-1) \mathop{\mathbb{E}}_{h\sim Q}(\Delta(h))^2 - D(Q||P) .$$

We now turn to bound $\mathbb{E}_S[e^{f(S)}]$. The main trick is to upper bound $f(S)$ by using an expression that does not depend on Q but rather depend on the prior probability P. To do so, fix some S and note that from the definition of $D(Q||P)$ we get that for all Q,

$$\begin{aligned} 2(m-1) \mathop{\mathbb{E}}_{h\sim Q}(\Delta(h))^2 - D(Q||P) &= \mathop{\mathbb{E}}_{h\sim Q}[\ln(e^{2(m-1)\Delta(h)^2} P(h)/Q(h))] \\ &\le \ln \mathop{\mathbb{E}}_{h\sim Q}[e^{2(m-1)\Delta(h)^2} P(h)/Q(h)] \\ &= \ln \mathop{\mathbb{E}}_{h\sim P}[e^{2(m-1)\Delta(h)^2}] , \end{aligned} \tag{2}$$

where the inequality follows from Jensen's inequality and the concavity of the log function. Therefore,

$$\mathop{\mathbb{E}}_{S}[e^{f(S)}] \le \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{h\sim P}[e^{2(m-1)\Delta(h)^2}] \tag{3}$$

The advantage of the expression on the right-hand side stems from the fact that we can switch orders of expectation (because P is a prior that does not depend on S) and get that

$$\mathop{\mathbb{E}}_{S}[e^{f(S)}] \le \mathop{\mathbb{E}}_{h\sim P} \mathop{\mathbb{E}}_{S}[e^{2(m-1)\Delta(h)^2}] . \tag{4}$$

Next, we show that for all h we have $\mathbb{E}_S[e^{2(m-1)\Delta(h)^2}] \le m$. To do so, recall that Hoeffding's inequality tells us that

$$\mathop{\mathbb{P}}_{S}[\Delta(h) \ge \epsilon] \le e^{-2m\epsilon^2} .$$

In Exercise 1 we show that this implies that $\mathbb{E}_S[e^{2(m-1)\Delta(h)^2}] \le m$. Combining this with Eq. (4) and plugging into Eq. (2) we get

$$\mathop{\mathbb{P}}_{S}[f(S) \ge \epsilon] \ \le\ \frac{m}{e^{\epsilon}} . \tag{5}$$

Denote the right-hand side of the above δ, thus $\epsilon = \ln(m/\delta)$, and we therefore obtain that with probability of at least $1 - \delta$ we have that for all Q

$$2(m-1) \mathop{\mathbb{E}}_{h\sim Q}(\Delta(h))^2 - D(Q||P) \le \epsilon = \ln(m/\delta) .$$

Rearranging the above and using Jensen's inequality again (the function x^2 is convex) we conclude that

$$\left(\mathop{\mathbb{E}}_{h\sim Q}\Delta(h)\right)^2 \le \mathop{\mathbb{E}}_{h\sim Q}(\Delta(h))^2 \le \frac{\ln(m/\delta)+D(Q||P)}{2(m-1)} . \tag{6}$$

■

Exercise

1. Let X be a random variable that satisfies $\mathbb{P}[X \ge \epsilon] \le e^{-2m\epsilon^2}$. Prove that $\mathbb{E}[e^{2(m-1)X^2}] \le m$.

3 Compression Bounds

Consider an algorithm which receives a training set of m examples but whose output hypothesis can be determined by only observing a subset of k examples. Informally, in this case we can use the error on the rest $m-k$ examples as an estimator of the generalization error. This is formalized in the following theorem that is stated for classification loss but is valid for any bounded loss. We define $L_I(h)$ to be the empirical loss of hypothesis h on a subset I of the samples.

Theorem 2 *Let $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ be a sequence of examples and let $A(S)$ be a learning rule. Let $I = (i_1, \ldots, i_k)$ be a sequence of indices from $[m]$ such that $k < m$ and let J be the rest of the indices in $[m]$. Assume that there is a deterministic mapping from $(\mathbf{x}_{i_1}, y_{i_1}), \ldots, (x_{i_k}, y_{i_k})$ to a hypothesis h_I such that for all $j \in J$ we have $A(S)(\mathbf{x}_j) = h(\mathbf{x}_j)$. Then, with probability of at least $1-\delta$ we have*

$$L_{\mathcal{D}}(A(S)) \ \le\ L_{S_J}(A(S)) + \sqrt{\frac{(k+1)\log(m/\delta)}{m-k}} .$$

Proof

$$\begin{aligned}
&\mathbb{P}[L_{\mathcal{D}}(A(S)) - L_{S_J}(A(S)) \ge \epsilon] \\
&\quad \le \mathbb{P}[\exists I \text{ s.t. } L_{\mathcal{D}}(h_I) - L_{S_J}(h_I) \ge \epsilon] \\
&\quad \le \sum_{k=1}^{m-1} \sum_{I:|I|=k} \mathbb{P}[L_{\mathcal{D}}(h_I) - L_{S_J}(h_I) \ge \epsilon] \\
&\quad = \sum_{k=1}^{m-1} \sum_{I:|I|=k} \mathbb{P}_{S_I}\,\mathbb{P}_{S_J|S_I}[L_{\mathcal{D}}(h_I) - L_{S_J}(h_I) \ge \epsilon]
\end{aligned}$$

By Hoeffding we know that

$$\mathbb{P}_{S_J|S_I}[L_{\mathcal{D}}(h_I) - L_{S_J}(h_I) \ge \epsilon] \ \le\ e^{-|J|\epsilon^2}$$

Therefore, we obtain the bound

$$\sum_{I:|I|=k} \mathbb{P}_{S_I}\,\mathbb{P}_{S_J|S_I}[L_{\mathcal{D}}(h_I) - L_{S_J}(h_I) \ge \epsilon] \le m^k e^{-|J|\epsilon^2} .$$

The right-hand side of the above is bounded above by δ/m provided that

$$\epsilon \le \sqrt{\frac{(k+1)\log(m/\delta)}{|J|}} .$$

For such ϵ we obtain

$$\mathbb{P}[L_{\mathcal{D}}(A(S)) - L_{S_J}(A(S)) \geq \epsilon] \leq \sum_{k=1}^{m-1} \delta/m \leq \delta ,$$

which concludes our proof. ■

Exercises

1. Apply the compression bound given in Theorem 2 to derive a generalization bound for the Perceptron algorithm.
2. A compression scheme of size k for a concept class C picks from any set of examples consistent with some $h \in C$ a subset of at most k examples that "represents" a hypothesis consistent with the whole original training set. Theorem 2 tells us that the existence of a compression scheme guarantees generalization and thus learnability. Prove that the VC dimension of C should be bounded (and in fact, based on the bounds, it should equal k). The opposite question is an open problem proposed by Manfred Warmuth: Does any concept class of VC dimension d has a compression scheme of size d?

4 Online to batch conversions

In this section we show that an online algorithm that attains low regret can be converted into a batch learning algorithm that attains low risk. Such online-to-batch conversions are interesting both from the practical as from the theoretical perspective.

Recall that we assume that the sequence of examples are independently and identically distributed according to an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. To emphasize the above fact and to simplify our notation, we denote by Z_t the tth example in the sequence and use the shorthand

$$Z_1^j \;=\; (Z_1, \ldots, Z_j) \;=\; ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_j, y_j)) \;.$$

We denote the tth hypothesis that the online learning algorithm generates by h_t. Note that h_t is a function of Z_1^{t-1} and thus it is a random variable (w.r.t. $\mathcal{D}$ but also randomization by the online algorithm). We denote the average loss of the online algorithm by

$$M_T(Z_1^T) \;=\; \frac{1}{T} \sum_{t=1}^{T} \ell(h_t, Z_t) \;. \tag{7}$$

We often omit the dependence of M_T on Z_1^T and use the shorthand M_T for denoting $M_T(Z_1^T)$.

The rest of this section is organized as follows. In Section 4.1 we show that the expected value of M_T equals the expected value of $\frac{1}{T}\sum_{t=1}^{T} L_{\mathcal{D}}(h_t)$. Thus, the online loss is an un-biased estimator for the average risk of the ensemble $(h_1, \ldots, h_T)$. Next, in Section 4.2 we underscore that regret bounds (i.e., bounds on M_T) can yield bounds on the average risk of $(h_1, \ldots, h_T)$. Therefore, there exists at least one hypothesis in the ensemble $(h_1, \ldots, h_T)$ whose risk is low. Since our goal in batch learning is typically to output a *single* hypothesis (with low risk), we must find a way to choose a single good hypothesis from the ensemble. In Section 4.3, we discuss several simple procedures for choosing a single good hypothesis from the ensemble.

4.1 Online loss and ensemble's risk

Our first theorem shows that the expected value of M_T equals the expected value of the risk of the ensemble $(h_1, \ldots, h_T)$.

Theorem 3 *Let $Z_1, \ldots, Z_T$ be a sequence of independent random variables, each of which is distributed according to a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. Let $h_1, \ldots, h_T$ be the sequence of hypotheses generated by an online algorithm when running on the sequence $Z_1, \ldots, Z_T$, and let $L_{\mathcal{D}}(h) = \mathbb{E}_{Z \sim \mathcal{D}}[\ell(h, Z)]$. Then,*

$$\mathbb{E}_{Z_1,\ldots,Z_T}\left[\tfrac{1}{T}\sum_{t=1}^{T} L_{\mathcal{D}}(h_t)\right] \;=\; \mathop{\mathbb{E}}_{Z_1,\ldots,Z_T}\left[\tfrac{1}{T}\sum_{t=1}^{T} \ell(h_t, Z_t)\right] \;.$$

Proof Using the linearity of expectation and the fact that h_t only depends on Z_1^{t-1} we have,

$$\mathop{\mathbb{E}}_{Z_1^T}[\frac{1}{T}\sum_{t=1}^{T}\ell(h_t, Z_t)] \;=\; \frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{Z_1^T}[\ell(h_t, Z_t)] \;=\; \frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{Z_1^t}[\ell(h_t, Z_t)] \,. \tag{8}$$

Recall that the law of total expectation implies that for any two random variables R_1, R_2, and a function f, $\mathbb{E}_{R_1}[f(R_1)] = \mathbb{E}_{R_2}\mathbb{E}_{R_1}[f(R_1)|R_2]$. Setting $R_1 = Z_1^t$ and $R_2 = Z_1^{t-1}$ we get that

$$\mathbb{E}_{Z_1^t}[\ell(h_t, Z_t)] \;=\; \mathbb{E}_{Z_1^{t-1}}[\mathbb{E}_{Z_1^t}[\ell(h_t, Z_t)|Z_1^{t-1}]] \;=\; \mathbb{E}_{Z_1^{t-1}}[L_{\mathcal{D}}(h_t)] \;=\; \mathbb{E}_{Z_1^T}[L_{\mathcal{D}}(h_t)] \;.$$

Combining the above with Eq. (8) concludes our proof. ■

The above theorem tells us that in expectation, the online loss equals the average risk of the ensemble of hypotheses generated by the online algorithm. The next step is to use our regret bounds from previous lectures to derive bounds on the batch loss.

4.2 From Regret Bounds to Risk Bounds

In the previous section we analyzed the risks of the hypotheses generated by an online learning algorithm based on the average online loss, M_T. In previous lectures we analyzed the regret of online algorithms by bounding the online loss, M_T, in terms of the loss of any competing hypothesis in $\mathcal{H}$. In particular, the bound holds for the hypothesis in $\mathcal{H}$ whose risk is minimal. Formally, assume that the minimum risk is achievable and denote by $h^\star$ a hypothesis s.t. $L_{\mathcal{D}}(h^\star) = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$. In this section, we derive bounds on M_T in terms of $L_{\mathcal{D}}(h^\star)$.

In the simplest form of regret bounds, there exists a deterministic function $B : \mathbb{N} \to \mathbb{R}$ such that

$$\forall h \in \mathcal{H},\; \frac{1}{T}\sum_{t=1}^{T}\ell(h_t, Z_t) \;\le\; \frac{1}{T}\sum_{t=1}^{T}\ell(h, Z_t) + \frac{B(T)}{T} \;. \tag{9}$$

E.g., for online convex optimization we had $B(T) = O(\sqrt{T})$.

We start by deriving a bound on the average risk of the ensemble of hypotheses that the online algorithm generates.

Theorem 4 *Assume that the condition stated in Theorem 3 holds and that the online algorithm satisfies Eq. (9). Then,*

$$\mathbb{E}_{Z_1^T}\left[\tfrac{1}{T}\sum_{t=1}^{T} L_{\mathcal{D}}(h_t)\right] \;\le\; L_{\mathcal{D}}(h^\star) + \frac{B(T)}{T} \;.$$

Proof Taking expectation of the inequality given in Eq. (9) we obtain

$$E_{Z_1^T}\left[\frac{1}{T}\sum_{t=1}^{T}\ell(h_t, Z_t)\right] \;\le\; \mathbb{E}_{Z_1^T}\left[\frac{1}{T}\sum_{t=1}^{T}\ell(h^\star, Z_t)\right] + \frac{B(T)}{T} \;. \tag{10}$$

Since $h^\star$ does not depend on the choice of Z_1^T, we have,

$$\mathbb{E}_{Z_1^T}\left[\frac{1}{T}\sum_{t=1}^{T}\ell(h^\star, Z_t)\right] \;=\; \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{Z_1^T}[\ell(h^\star, Z_t)] \;=\; \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{Z_t}[\ell(h^\star, Z_t)] = \; L_{\mathcal{D}}(h^\star)\;. \qquad (11)$$

Combining the above with Eq. (10) and Theorem 3 we conclude our proof. ■

4.3 Choosing a hypothesis from the ensemble

In the previous section we showed that regret bounds yield bounds on the average risk of $(h_1, \ldots, h_T)$. The goal of this section is two-fold. First, we need to output a single hypothesis and thus we must choose a single good hypothesis from $(h_1, \ldots, h_T)$. Second, the analysis in the previous section focuses on the expected value of the risk whereas in practice we have a single training set. Therefore, we must analyze the concentration properties of our online-to-batch conversion schemes.

4.3.1 Last (with random stopping time)

The simplest conversion scheme runs the online algorithm on a sequence of r examples and returns the last hypothesis h_r. To analyze the risk of h_r we assume that r is chosen uniformly at random from $\{1, 2, \ldots, T\}$, where T is a predefined integer. The following lemma shows that the bounds on $\mathbb{E}[\frac{1}{T}\sum_t L_{\mathcal{D}}(h_t)]$ we derived in the previous section can be transformed into a bound on $L_{\mathcal{D}}(h_r)$.

Lemma 1 *Assume that the conditions stated in Theorem 3 hold. Let $h^\star$ be a hypothesis in $\mathcal{H}$ whose risk is minimal and let $\delta \in (0, 1)$. Assume that there exists a scalar α such that*

$$\mathbb{E}_{Z_1^T}\left[\tfrac{1}{T}\sum_{t=1}^{T} L_{\mathcal{D}}(h_t)\right] \;\leq\; L_{\mathcal{D}}(h^\star) + \alpha\;.$$

Let $r \in [T]$ and assume that r is uniformly chosen at random from $[T]$. Then, with a probability of at least $1 - \delta$ over the choices of Z_1^T and r we have

$$L_{\mathcal{D}}(h_r) \;\leq\; L_{\mathcal{D}}(h^\star) + \frac{\alpha}{\delta}\;\;.$$

Proof Let R be the random variable $(L_{\mathcal{D}}(h_r) - L_{\mathcal{D}}(h^\star))$. From the definition of $h^\star$ as the minimizer of $L_{\mathcal{D}}(h)$ we clearly have that R is a non-negative random variable. In addition, the assumption in the lemma implies that $\mathbb{E}[R] \leq \alpha$. Thus, from the Markov inequality

$$\mathbb{P}[R \geq a] \leq \frac{\mathbb{E}[R]}{a} \leq \frac{\alpha}{a}\;\;.$$

Setting $\frac{\alpha}{a} = \delta$ we conclude our proof. ■

Combining the above lemma with Theorem 4 we obtain the following:

Corollary 1 *Assume that the conditions stated in Theorem 4 hold. Let $h^\star$ be a hypothesis in $\mathcal{H}$ whose risk is minimal and let $\delta \in (0, 1)$. Then, with a probability of at least $1 - \delta$ over the choices of $(Z_1, \ldots, Z_T)$ and the index r we have that*

$$L_{\mathcal{D}}(h_r) \;\leq\; L_{\mathcal{D}}(h^\star) + \frac{B(T)}{\delta\, T}\;\;.$$

Corollary 1 implies that by running the online algorithm on the first r examples and outputting h_r we obtain a batch learning algorithm with a guaranteed risk bound. However, the concentration bound given in Corollary 1 depends linearly on $1/\delta$, where δ is the confidence parameter. Our next conversion scheme is preferable when we are aiming for very high confidence.

4.3.2 Validation

In the validation conversion scheme, we first pick a subset of hypotheses from $h_1, \ldots, h_T$ and then use a fresh validation set to decide which hypothesis to output. We start by using a simple amplification technique (a.k.a. boosting the confidence), to construct a few candidate hypotheses such that with confidence $1 - \delta$ the risk of at least one of the hypotheses is low.

Theorem 5 *Assume that the conditions stated in Theorem 3 hold. Let s be a positive integer. Assume that we reset the online algorithm after each block of T/s examples. Let $h'_1, \ldots, h'_s$ be a sequence of hypotheses where for each $i \in \{1, 2, \ldots, s\}$, the hypothesis h'_i is picked uniformly at random from $\{h_{i\,s+1}, \ldots, h_{i\,s+s}\}$. Then with a probability of at least $1 - e^{-s}$, there exists $i \in \{1, 2, \ldots, s\}$ such that*

$$L_{\mathcal{D}}(h'_i) \;\leq\; L_{\mathcal{D}}(h^\star) + \frac{e\; s\, B(T/s)}{T} \;\; .$$

Proof Using Corollary 1 with a confidence value of $1/e$, we know that for all $i \in [s]$, with a probability of at least $1 - 1/e$ we have that

$$L_{\mathcal{D}}(h'_i) - L_{\mathcal{D}}(h^\star) \leq e\,\alpha(T/s) \tag{12}$$

where $\alpha(k) = B(k)/k$. Therefore, the probability that for *all* blocks the above inequality does not hold is at most e^{-s}. Hence, with a probability of at least $1 - e^{-s}$, at least one of the hypotheses h'_i satisfies Eq. (12). ■

The above theorem tells us that there exists at least one hypothesis $h_g \in \{h'_1, \ldots, h'_s\}$ such that $L_{\mathcal{D}}(h_g)$ is small. To find such a good hypothesis we can use a validation set and choose the hypothesis whose loss on the validation set is minimal. Formally, let $Z'_1, \ldots, Z'_m$ be a sequence of random variables that represents a fresh validation set and let h_o be the hypothesis in $\{h'_1, \ldots, h'_s\}$ whose loss over $Z'_1, \ldots, Z'_m$ is minimal. Applying standard generalization bounds (e.g., Eq. (21) in Boucheron Bousquet and Lugosi, 2005) on the finite hypothesis class $\{h'_1, \ldots, h'_s\}$ we obtain that there exists a constant C such that

$$L_{\mathcal{D}}(h_o) - L_{\mathcal{D}}(h_g) \;\leq\; C\left(\sqrt{L_{\mathcal{D}}(h_g)\,\frac{\log(s)\,\log(m) + \ln\left(\frac{1}{\delta}\right)}{m}} + \frac{\log(s)\,\log(m) + \ln\left(\frac{1}{\delta}\right)}{m}\right) \;\; .$$

4.3.3 Averaging

If the set of hypotheses is convex and the loss function, ℓ, is convex with respect to h then the risk function is also convex with respect to h. Therefore, Jensen's inequality implies that

$$L_{\mathcal{D}}\left(\tfrac{1}{T}\sum_{t=1}^{T} h_t\right) \;\leq\; \tfrac{1}{T}\sum_{t=1}^{T} L_{\mathcal{D}}(h_t) \;\; .$$

We can use the above inequality in conjunction with Theorem 4 to derive bounds on the expected risk of the averaged hypothesis $\bar{h} = \frac{1}{T}\sum_{t=1}^{T} h_t$. In particular, the bounds we derived in Corollary 1 hold for $\bar{h}$ as well. If we want to have very high confidence, we can use the amplification technique described in the previous conversion scheme. Alternatively, we can use Azuma's inequality to obtain a bound that holds with high probability. This is left as an exercise.