Math 547: Mathematical Foundations of Statistical Learning Theory

Lectures 1–3 — Week 10 + Lecture 1 of Week 11

Instructor: S. Minsker *Scribe: S. Minsker*

1.1 Bounded Difference (McDiarmid's) Inequality

At this point, we have the tools to control the size of expected supremum $\mathbb{E}\|P_n - P\|_{\mathcal{F}}$. Result in this section allows one to show that the random variable $\|P_n - P\|_{\mathcal{C}}$ concentrates tightly around its expectation.

Let $\mathbb{Z}(x_1, \dots, x_n)$ be a function of n variables. Assume that it satisfies the following assumption:

Assumption (Bounded difference)**.** For all $1 \le j \le n$,

$$\sup_{x_1,\dots,x_j,x_j',\dots,x_n} |\mathbb{Z}(\mathrm{x}_1,\dots,\mathrm{x}_j,\dots,\mathrm{x}_n) - \mathbb{Z}(\mathrm{x}_1,\dots,\mathrm{x}_j',\dots,\mathrm{x}_n)| \le \mathrm{c}_j.$$

In essence, the bounded difference condition controls the fluctuations of $\mathbb{Z}$ with respect to each variable separately.

Theorem 1. Let $X_1, \dots, X_n$ be i.i.d. random variables. Assume that $\mathbb{Z}(X_1, \dots, X_n)$ satisfies the bounded difference assumption. Then

$$\Pr\left(|\mathbb{Z}(X_1,\dots,X_n) - \mathbb{E}\mathbb{Z}(X_1,\dots,X_n)| \ge t\right) \le 2e^{-\frac{t^2}{2\sum_{j=1}^n c_j^2}}.$$

Remark 1. A more careful argument (for instance, given in the original paper by McDiarmid) allows to obtain the tail bound $2e^{-\frac{2t^2}{\sum_{j=1}^n c_j^2}}$.

Example 1. Let $\mathbb{Z}(X_1, \dots, X_n) = \sum_{j=1}^n X_j$. Then

$$\mathbb{Z}(X_1,\dots,X_j,\dots,X_n) - \mathbb{Z}(X_1,\dots,X_j',\dots,X_n) = X_j - X_j'.$$

Hence, if $|X_j| \le M \implies |X_j - X_j'| \le 2M$, and the bounded difference inequality becomes Hoeffding's inequality.

Example 2. This is one of the main examples we are interested in. Let

$$\mathbb{Z} = \|P_n - P\|_{\mathcal{C}} = \sup_{C\in\mathcal{C}} \left| \frac{1}{n}\sum_{j=1}^n I\{X_j \in C\} - P(C) \right|.$$

Then $\mathbb{Z}$ satisfies the bounded difference properly with $c_j = \frac{1}{n}$. Indeed,

$$\sup_{C\in\mathcal{C}} \left| \frac{1}{n}\sum_{j=1}^n I\{X_j \in C\} - P(C) + \frac{1}{n} I\{X_j' \in C\} - \frac{1}{n} I\{X_j' \in C\} \right|$$
$$\le \sup_{C\in\mathcal{C}} \left| \frac{1}{n}\sum_{j\ne i}^n I\{X_j \in C\} + I\{X_j' \in C\} - P(C) \right| + \left| \frac{I\{X_j \in C\}}{n} - \frac{I\{X_j' \in C\}}{n} \right|.$$

Proof of the Theorem. The proof is based on the *margingale decomposition technique*. To this end, we will find random variables $V_1, \ldots, V_n$ such that

(1) $\mathbb{Z}(X_1, \ldots, X_n) - \mathbb{E}\mathbb{Z}(X_1, \ldots, X_n) = \sum_{J=1}^{n} V_j$;

(2) $\mathbb{E}[V_j | V_1, \ldots, V_{j-1}] = 0$;

(3) $V_j = V_j(X_1, \ldots, X_j)$ (in other words, V_j is a function of $X_1, \ldots, X_j$ only.

Recall the definition of a martingale sequence:

Definition 1 (Martingale)**.** A sequence $(Y_j, \mathcal{F}_j)_{j \geq 1}$, where X_j is a random variable and $\mathcal{F}_j$ is a σ-algebra, is a martingale if

(1) $\mathbb{E}\,|Y_j| < \infty$;

(2) Y_j is $\mathcal{F}_j$ – measurable;

(3) $\mathbb{E}(Y_j | \mathcal{F}_{j-1}) = Y_{j-1}$.

Note that the partial sums $\sum_{j=1}^{k} V_j$ form a martingale, assuming that V_j's satisfy the requirements stated above.

We will define $V_1, \ldots, V_n$ as follows: let $(Y_1, \ldots, Y_n)$ be an independent copy of $(X_1, \ldots, X_n)$, and set

$$V_j = \mathbb{E}_Y[\mathbb{Z}(X_1, \ldots, X_j, Y_{j+1}, \ldots, Y_n)] - \mathbb{E}_Y\left[\mathbb{Z}(X_1, \ldots, X_{j-1}, Y_j, \ldots, Y_n)\right].$$

From this definition, it is clear that $|V_j| \leq c_j$ by the bounded difference property of $\mathbb{Z}$. Moreover, note that V_j can be equivalently written as

$$V_j = \mathbb{E}\left[\mathbb{Z}(X_1, \ldots, X_n) | X_1, \ldots, X_j\right] - \mathbb{E}\left[\mathbb{Z}(X_1, \ldots, X_n) | X_1, \ldots, X_{j-1}\right].$$

It is easy to check that $\sum_{j=1}^{n} V_j = \mathbb{Z} - \mathbb{E}Z$ and that V_j depends only on $X_1, \ldots, X_j$. To show the second property, note that by the "tower property" of conditional expectations,

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\mathbb{Z} | X_1, \ldots, X_j\right] - \mathbb{E}\left[\mathbb{Z}(X_1, \ldots, X_{j-1}) \,|\, X_1, \ldots, X_{j-1}\right] \\ = \mathbb{E}\left[\mathbb{Z} | X_1, \ldots, X_{j-1}\right] - \mathbb{E}\left[\mathbb{Z} | X_1, \ldots, X_{j-1}\right] = 0.\end{aligned}$$

The rest of the proof proceeds a in way that is similar to the proofs of Hoeffding's and Bernstein's inequalities (that is, estimating the moment generating function). Let $\lambda > 0$ (to be chosen later), and note that

$$\begin{aligned}\Pr\left(\mathbb{Z} - \mathbb{E}\mathbb{Z} > t\right) &= \Pr\left(\lambda(\mathbb{Z} - \mathbb{E}\mathbb{Z}) \geq \lambda t\right) \\ &= \Pr\left(e^{\lambda(\mathbb{Z} - \mathbb{E}\mathbb{Z})} \geq e^{\lambda t}\right) \leq e^{-\lambda t}\mathbb{E}e^{\lambda(\mathbb{Z} - \mathbb{E}\mathbb{Z})} \\ &= e^{-\lambda t}\mathbb{E}e^{\lambda \sum_{j=1}^{n} V_j}.\end{aligned}$$

Next,

$$\begin{aligned}
\mathbb{E}e^{\lambda\sum_{j=1}^{n}V_j} &= \mathbb{E}\mathbb{E}\left[e^{\lambda\sum_{j=1}^{n}V_j}|X_1,\ldots,X_{n-1}\right] = \mathbb{E}\mathbb{E}\left[e^{\lambda\sum_{j=1}^{n-1}V_j}e^{\lambda V_n}|X_1,\ldots,X_{n-1}\right] \\
&= \mathbb{E}\left(e^{\lambda\sum_{j=1}^{n-1}V_j}\mathbb{E}\left[e^{\lambda V_n}|X_1,\ldots,X_{n-1}\right]\right) \leq e^{\frac{\lambda c_n^2}{2}}\mathbb{E}e^{-\lambda\sum_{j=1}^{n-1}V_j} \\
&= e^{\frac{\lambda c_n^2}{2}}\mathbb{E}\mathbb{E}\left[e^{\lambda\sum_{j=1}^{n-2}V_j}e^{\lambda V_{n-1}}|X_1,\ldots,X_{n-2}\right] \\
&= e^{\frac{\lambda c_n^2}{2}}\mathbb{E}\left(e^{\lambda\sum_{j=1}^{n-2}V_j}\mathbb{E}\left[e^{\lambda V_{n-1}}|X_1,\ldots,X_{n-2}\right]\right) \\
&\leq e^{\frac{\lambda^2 c_n^2}{2}}e^{\frac{\lambda^2 c_{n-1}^2}{2}}\mathbb{E}e^{\lambda\sum_{j=1}^{n-2}V_j} \leq \ldots \leq e^{\frac{\lambda^2}{2}\sum_{j=1}^{n}c_j^2}.
\end{aligned}$$

We used the property that $|V_j| \leq c_j$ almost surely, and the fact that bounded zero-mean random variables are sub-Gaussian, implying that $\mathbb{E}\left[e^{\lambda V_j}|X_1,\ldots,X_{j-1}\right] \leq e^{\lambda^2 c_j^2/2}$. Hence,

$$\Pr(\mathbb{Z}-\mathbb{E}\mathbb{Z}\geq t) \leq e^{-\lambda t+\frac{\lambda^2}{2}\sum_{j=1}^{n}c_j^2}$$

The minimum of $f(\lambda) = -\lambda t + \frac{\lambda^2}{2}\sum_{j=1}^{n}c_j^2$ is attained for $\lambda_* = \frac{t}{\sum_{j=1}^{n}c_j^2}$, and the final result follows.

The inequality for the lower tail $\Pr(\mathbb{Z}-\mathbb{E}\mathbb{Z}\leq -t)$ is obtained from the same argument applied to $-Z$ instead of Z. □

1.2 Generalization error bounds for Adaboost

We are almost ready to go back to the first weeks of the course and establish a bound on the generalization error of AdaBoost. To this end, we will need one more technical result called "Talagrand's contraction inequality" that is stated below. Its (very nice!) proof is given in the end of these notes.

We will say that $\varphi : \mathbb{R} \mapsto \mathbb{R}$ is a *contraction* whenever $|\varphi(x)-\varphi(y)| \leq |x-y|$ for all $x, y \in \mathbb{R}$.

Theorem 2 (M. Talagrand)**.** Assume that Φ be a convex function that is nondecreasing on $[0,\infty)$, and let φ_j be contractions for $j = 1,\ldots,n$ such that $\varphi_j(0) = 0$. Then for any $T \subseteq \mathbb{R}^n$

$$\mathbb{E}\Phi\left(\frac{1}{2}\sup_{t\in T}\left|\sum_{j=1}^{n}\varepsilon_j\varphi(t_j)\right|\right) \leq \mathbb{E}\Phi\left(\sup_{t\in T}\left|\sum_{j=1}^{n}\varepsilon_j t_j\right|\right).$$

Here, $t = (t_1,\ldots,t_n)$. Before we proceed to the proof, let's look at a useful example.

Example 3. Let $\mathcal{F} = \{f : S \mapsto \mathbb{R}\}$, and let φ be a contraction. Define $\varphi\circ\mathcal{F} := \{\varphi(f),\ f \in \mathcal{F}\}$. Finally, given $x_1,\ldots x_n \in S$, let $T := \{f(x_1),\ldots f(x_n), f \in \mathcal{F}\}$. Then

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\phi(f(x_j))\right| \leq 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j f(x_j)\right|.$$

For instance, of $\varphi(x) = x^2$ and $|f| \le U$ for $f \in \mathcal{F}$, then $\varphi(x) = \frac{\varphi(x)}{2U}$ is a contraction on $[-U, U]$. Hence,

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j f^2(x_j) \right| \le 4U \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j f(x_j) \right|$$

In particular, we can replace $x_1, \ldots, x_n$ by random variables $X_1, \ldots, X_n$, and get that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j \phi(f(X_j)) \right| \le 2\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j f(X_j) \right|$$

Now we proceed with our main goal – controlling the generalization error of AdaBoost. Let $\mathcal{F}$ be the "base class" – class of binary classifiers of VC dimension V, and let

$$\hat{g}_T = \frac{\sum_{j=1}^{T} \alpha_j f_j}{\sum_{j=1}^{T} \alpha_j}$$

be the output of Adaboost after T steps; note that $\hat{g}_T \in \mathrm{co}(\mathcal{F})$, the convex hull of $\mathcal{F}$.

Theorem 3. With probability$\ge 1 - e^{-t}$, for any $\theta > 0$

$$\Pr\left(Y \ne \mathrm{sign}\left(\hat{g_T}(X)\right)\right) \le \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}\left\{Y_j \hat{g_T}(X_j) \le \theta\right\} + K\left(\frac{1}{\theta}\sqrt{\frac{V}{n}} + \sqrt{\frac{t}{n}}\right)$$

where K is an absolute constant.

Proof. Let

$$\varphi_\theta(x) = \begin{cases} 1, & x \le 0 \\ 1 - \frac{x}{\theta}, & 0 \le x \le \theta, \\ 0, & x \ge \theta \end{cases}$$

and note that φ_θ is Lipschitz continuous with Lipschitz constant $\frac{1}{\theta}$. Moreover, it is easy to see that $\mathbb{I}\{x \le 0\} \le \varphi_\theta(x) \le \mathbb{I}\{x \le \theta\}$ for all $x \in \mathbb{R}$. Consequently,

$$\begin{aligned} \Pr\left(Y \hat{g}_T(X) \le 0\right) &= \mathbb{E}\mathbb{I}\left\{Y \hat{g}_T(X) \le 0\right\} \le \mathbb{E}\varphi_\theta\left(Y \hat{g}_T(X)\right) \\ &= \mathbb{E}\varphi_\theta\left(Y \hat{g}_T(X)\right) \pm \left(\frac{1}{n} \sum_{j=1}^{n} \varphi_\theta\left(Y_j \hat{g}_T(X_j)\right)\right) \\ &\le \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}\left\{Y_j \hat{g}_T(X_j) \le \theta\right\} + \underbrace{\sup_{g \in \mathrm{co}(\mathcal{F})} \left| \frac{1}{n} \sum_{j=1}^{n} \varphi_\theta\left(Y_j g(X_j)\right) - \mathbb{E}\varphi_\theta\left(Y g(X)\right) \right|}_{Z((X_1,Y_1),\ldots,(X_n,Y_n))}. \end{aligned}$$

Notice that Z satisfies the bounded difference property with $c_j = \frac{1}{n}$. Hence, the bounded difference inequality implies that

$$Z \le \mathbb{E}Z + \sqrt{\frac{t}{n}} \text{ with probability } \ge 1 - e^{-2t}.$$

It remains to estimate $\mathbb{E}Z$. To this end, we apply the symmetrization and contraction inequalities to get that

$$\mathbb{E}\sup_{g\in\mathrm{co}(\mathcal{F})}\left|\frac{1}{n}\sum_{j=1}^{n}(\varphi_n(Y_jg(X_j)))-\mathbb{E}\varphi_\theta\left(Y_jg(X_j)\right)\right|\le\frac{2}{\theta}\mathbb{E}\sup_{g\in\mathrm{co}(\mathcal{F})}\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\varphi_\theta\left(Y_jg(X_j)\right)\right|$$
$$\le\frac{4}{\theta}\mathbb{E}\sup_{g\in\mathrm{co}(\mathcal{F})}\left|\frac{1}{n}\sum_{1}^{n}\varepsilon_jY_jg(X_j)\right|.$$

Since the maximum of a linear function over a convex set is always attained at one of the extreme points,

$$\mathbb{E}\sup_{g\in\mathrm{co}(\mathcal{F})}\left|\frac{1}{n}\sum_{j=1}^{n}\tilde{\varepsilon}_jg(X_j)\right|=\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n}\tilde{\varepsilon}_jf(X_j)\right|,$$

where $\tilde{\varepsilon}_j=\varepsilon_jY_j,\ j=1,\dots,n$ is a sequence of Rademacher random variables since ε_j's and Y_j's are independent. It implies that

$$\mathbb{E}\sup_{g\in\mathrm{co}(\mathcal{F})}\left|\frac{1}{n}\sum_{j=1}^{n}\left(\varphi_n\left(Y_jg(X_j)\right)\right)-\mathbb{E}\varphi_\theta(Y_jg(X_j))\right|\le\frac{4}{\theta}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n}\tilde{\varepsilon}_jf(X_j)\right|$$
$$\le\frac{K}{\theta}\sqrt{\frac{V(\mathcal{F})}{n}},$$

where K is an absolute constant; the last inequality follows from Dudley's entropy integral bound and Haussler's theorem. □

Finally, we will address the size of the term $\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\left\{Y_j\hat{g_T}\left(X_j\right)\le\theta\right\}$. Assume that γ-weak learnability assumption holds, meaning that for some $\gamma>0$

$$\forall\omega_1,...,\omega_n\ge0,\ \sum_{j=1}^{n}\omega_j=1,\ \exists f\in\mathcal{F},\ \text{such that}\ \sum_{j=1}^{n}\omega_j\mathbb{I}\left\{Y_j\neq f(X_j)\right\}\le1/2-\gamma.$$

Theorem 4. Let $\hat{g}_T$ be the output of AdaBoost after T steps, and assume that γ-weak learnability assumption holds (for some, possibly unknown, $\gamma>0$). Then $\forall\theta\in(0,1)$,

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\left\{Y_j\hat{g}_T(X_j)\le\theta\right\}\le2^T\left(\sqrt{(1-2\gamma)^{1-\theta}(1+2\gamma)^{1+\theta}}\right)^T.$$

Remark 2. As $\theta\to0$, the right-hand side of the inequality converges to $\left(\sqrt{(\frac{1}{2}-\gamma)(\frac{1}{2}+\gamma)}\right)^T$ that goes to 0 exponentially fast with T. Hence, for θ small enough, it is bounded by $C(\gamma,\theta)^T$ for some constant $C(\gamma,\theta)<1$.

Exercise. Prove that $C(\gamma,\theta)<1$ if $\theta<\gamma$.

Sketch of the proof. Assume that $Y\hat{g_T}(X) \leq \theta$, then

$$\exp\left(-Y\sum_{j=1}^{T}\alpha_j f_j(X) + \theta\sum_{j=1}^{T}\alpha_j\right) \geq 1.$$

Next,

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\left\{Y_j\hat{g_T}(X_j) \leq \theta\right\} &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\left\{Y_j\hat{g_T}(X_j) - \theta \leq 0\right\} \\ &\leq \frac{1}{n}\sum_{j=1}^{n}\exp\left(\theta\sum_{i=1}^{T}\alpha_i - Y_j\sum_{i=1}^{T}\alpha_i f_i(X_j)\right) \\ &= \frac{1}{n}\exp\left(\theta\sum_{i=1}^{T}\alpha_i\right)\sum_{j=1}^{n}\exp\left(-Y_j\sum_{i=1}^{T}\alpha_i f_i(X_j)\right).\end{aligned}$$

The rest of the proof (estimation of the term $\sum_{j=1}^{n}\exp\left(-Y_j\sum_{i=1}^{T}\alpha_i f_i(X_j)\right)$) proceeds as before, see Theorem 1 in the notes for Week 4. □

1.3 Proof of Talagrand's Contraction Inequality

1. First, we will show that for an *arbitrary map* $A : T \subset \mathbb{R}^n \mapsto \mathbb{R}$,

$$\mathbb{E}\Phi\left(\sup_{t\in T}\left[A(t) + \sum_{j=1}^{n}\varepsilon_j\varphi(t_j)\right]\right) \leq \mathbb{E}\Phi\left(\sup_{t\in T}\left[A(t) + \sum_{j=1}^{n}\varepsilon_j t_j\right]\right)$$

2. It is sufficient to consider the case $n = 1$. Indeed, if

$$\mathbb{E}\Phi\left(\sup_{t\in T}\left[A(t) + \varepsilon\varphi(t_n)\right]\right) \leq \mathbb{E}\Phi\left(\sup_{t\in T}\left[A(t) + \varepsilon t_n\right]\right)$$

then

$$\begin{aligned}&\mathbb{E}_{\varepsilon_1,\ldots\varepsilon_{n-1}}\mathbb{E}_{\varepsilon_n}\Phi\left(\sup_{t\in T}\left[\underbrace{A(t) + \sum_{j=1}^{n-1}\varepsilon_j\varphi(t_j)}_{\widehat{A}(t)} + \varepsilon_n\varphi(t_n)\right]\right) \\ &\leq \mathbb{E}_{\varepsilon_1,\ldots\varepsilon_{n-1}}\mathbb{E}_{\varepsilon_n}\Phi\left(\sup_{t\in T}\left[A(t) + \sum_{j=1}^{n-1}\varepsilon_j\varphi_j(t) + \varepsilon_n t_n\right]\right) \\ &\leq \ldots \leq \mathbb{E}\Phi\left(\sup_{t\in T}\left[A(t) + \sum_{j=1}^{n}\varepsilon_j t_j\right]\right).\end{aligned}$$

Recall that we want to show that $\mathbb{E}\Phi\left(\sup_{t\in T}\left[A\left(t\right)+\varepsilon\varphi\left(t_n\right)\right]\right)\leq\mathbb{E}\Phi\left(\sup_{t\in T}\left[A\left(t\right)+\varepsilon t_n\right]\right)$. This is equivalent to

$$\frac{1}{2}\Phi\left(\sup_{t\in T}\left[A\left(t\right)-\varphi\left(t_n\right)\right]\right)+\frac{1}{2}\Phi\left(\sup_{t\in T}\left[A\left(t\right)+\varphi\left(t_n\right)\right]\right)$$
$$\leq\frac{1}{2}\Phi\left(\sup_{t\in T}\left[A\left(t\right)-t_n\right]\right)+\frac{1}{2}\Phi\left(\sup_{t\in T}\left[A\left(t\right)+t_n\right]\right)$$

Let $s_1:=A(t),\ s_2:=t_n$, and $\tilde{T}=\{(s_1,s_2),\ t\in T\}$. We then need to show that

$$\Phi\left(\sup_{s_1,s_2\in\tilde{T}}\left[s_1-\varphi\left(s_2\right)\right]\right)+\Phi\left(\sup_{s_1,s_2\in\tilde{T}}\left[s_1+\varphi\left(s_2\right)\right]\right)$$
$$\leq\Phi\left(\sup_{s_1,s_2\in\tilde{T}}\left(s_1-s_2\right)\right)+\Phi\left(\sup_{s_1,s_2\in\tilde{T}}\left(s_1+s_2\right)\right)$$

Let $\bar{t}_1,\ \bar{t}_2$ be such that $\sup_{s_1,s_2}\left(t_1+\varphi\left(t_2\right)\right)=\bar{t}_1+\varphi\left(\bar{t}_2\right)$, and $\bar{s}_1,\ \bar{s}_2$ – such that $\sup_{s_1,s_2}\left(s_1-\varphi\left(s_2\right)\right)=\bar{s}_1-\varphi\left(\bar{s}_2\right)$. We will consider several cases:

(a) Assume that $\bar{t}_2\geq 0,\ \bar{s}_2\geq 0,\quad \bar{t}_2\geq\bar{s}_2$. We will show that

$$\Phi\left(\bar{t}_1+\varphi\left(\bar{t}_2\right)\right)+\Phi\left(\bar{s}_1-\varphi\left(\bar{s}_2\right)\right)\leq\Phi\left(\bar{t}_1+\bar{t}_2\right)+\Phi\left(\bar{s}_1-\bar{s}_2\right),$$

which suffices as the right-hand side is clearly at most $\Phi\left(\sup_{s_1,s_2\in\tilde{T}}\left(s_1-s_2\right)\right)+\Phi\left(\sup_{s_1,s_2\in\tilde{T}}\left(s_1+s_2\right)\right)$. Let $a=\bar{t}_1+\varphi\left(\bar{t}_2\right),\ b=\bar{t}_1+\bar{t}_2,\ c=\bar{s}_1-\bar{s}_2$ and $d=\bar{s}_1-\varphi\left(\bar{s}_2\right)$, and note that $a\leq b, c\leq d$ since φ is a contraction. Next,

$$b-a=\bar{t}_2-\varphi\left(\bar{t}_2\right)\geq\bar{s}_2-\varphi\left(\bar{s}_2\right)=d-c,$$

hence $\varphi\left(\bar{t}_2\right)-\varphi\left(\bar{s}_2\right)\leq\bar{t}_2-\bar{s}_2$. Moreover,

$$a=\bar{t}_1+\varphi\left(\bar{t}_2\right)\geq\bar{s}_1+\varphi\left(\bar{s}_2\right)\geq\bar{s}_1-\bar{s}_2=c.$$

We used the definition of $\bar{t}_1$ and $\bar{t}_2$ above, together with the fact that ϕ is a contraction. As Φ is nondecreasing and convex, these inequalities imply that the increment of Φ over $[a,b]$ is larger than the increment over $[c,d]$, or

$$\Phi\left(b\right)-\Phi\left(a\right)\geq\Phi\left(d\right)-\Phi\left(c\right).$$

The case $\bar{t}_2\geq 0,\ \bar{s}_2\geq 0,\ \bar{t}_2\leq\bar{s}_2$ is similar after renaming the variables $(\bar{t}_1,\bar{t}_2)\leftrightarrow(\bar{s}_1,\bar{s}_2)$, and replacing φ by $-\varphi$.

(b) The case $\bar{t}_2\leq 0,\ \bar{s}_2\leq 0$ can be reduced to the previous one by using the transformation $(\bar{t}_1,\bar{t}_2)\mapsto(\bar{t}_1,-\bar{t}_2)$, $(\bar{s}_1,\bar{s}_2)\to(\bar{s}_1,-\bar{s}_2)$ and $\tilde{\varphi}(x)=\varphi(-x)$.

(c) Finally, in the case $\bar{t}_2\geq 0,\ \bar{s}_2\leq 0$, observe that

$$\varphi\left(t_2\right)\leq t_2,\ -\varphi\left(s_2\right)\leq -S_2,$$

hence monotonicity of Φ implies that

$$\Phi\left(\bar{t}_1+\varphi\left(\bar{t}_2\right)\right)+\Phi\left(\bar{s}_1-\varphi\left(\bar{s}_2\right)\right)\leq\Phi\left(\bar{t}_1+\bar{t}_2\right)+\Phi\left(\bar{s}_1-\bar{s}_2\right).$$

3. Let $a_+ := \max(a, 0)$. To finish the proof, note that

$$\begin{aligned}
\mathbb{E}\Phi\left(\frac{1}{2}\sup_{t\in T}\left|\sum_{j=1}^{n}\varepsilon_j\varphi\left(t_j\right)\right|\right) &= \mathbb{E}\Phi\left(\frac{1}{2}\left(\sup_{t\in T}\sum_{j=1}^{n}\varepsilon_j\varphi\left(t_j\right)\right)_+ + \frac{1}{2}\left(\sup_{t\in T}\sum_{j=1}^{n}\left(-\varepsilon_j\right)\varphi\left(t_j\right)\right)_+\right) \\
&\le \frac{1}{2}\mathbb{E}\Phi\left(\left(\sup_{t\in T}\sum_{j=1}^{n}\varepsilon_j\varphi\left(t_j\right)\right)_+\right) + \frac{1}{2}\mathbb{E}\Phi\left(\left(\sup_{t\in T}\sum_{j=1}^{n}\left(-\varepsilon_j\right)\varphi\left(t_j\right)\right)_+\right) \\
&= \mathbb{E}\Phi\left(\left(\sup_{t\in T}\sum_{j=1}^{n}\varepsilon_j\varphi\left(t_j\right)\right)_+\right)
\end{aligned}$$

As the function $\tilde{\Phi}(a) := \Phi\left(a_+\right)$ is still non-decreasing and convex,

$$\mathbb{E}\Phi\left(\left(\sup_{t\in T}\sum_{j=1}^{n}\varepsilon_j\varphi\left(t_j\right)\right)_+\right) \le \mathbb{E}\Phi\left(\sup_{t\in T}\left(\sum_{j=1}^{n}\varepsilon_j t_j\right)_+\right) \le \mathbb{E}\Phi\left(\sup_{t\in T}\left|\sum_{j=1}^{n}\varepsilon_j t_j\right|\right),$$

where the last inequality follows from the monotonicity of Φ.