A Bit of Sequence Prediction: Lec 2

Linear Betting Game

For $t = 1$ to n

- Learner has to place a bet of amount $|\hat{y}_t|$ on either team A or team B (sign of $\hat{y}_t$ tells us which team we bet on)
- Outcome of the round is revealed as $y_t \in \{\pm 1\}$
- Learner looses money $\ell_t = -y_t \cdot \hat{y}_t$

End For

Goal: given $\phi : \{\pm 1\}^n \mapsto \mathbb{R}$ can we guarantee:

$$\sum_{t=1}^{n} (-y_t \cdot \hat{y}_t) \leq \phi(y_1, \ldots, y_n)$$

Linear Betting Result

Lemma

For any $\phi : \{\pm 1\}^n \mapsto \mathbb{R}$, *there exists a strategy with guarantee that*

$$\sum_{t=1}^{n} -(\hat{y}_t \cdot y_t) \leq \phi(y_1, \ldots, y_n)$$

If and only if

$$\mathbb{E}\left[\phi(\epsilon_1, \ldots, \epsilon_n)\right] \geq 0$$

and the strategy achieving this is $\hat{y}_t = \frac{1}{2}\mathbb{E}\left[\phi(y_1, \ldots, y_{t-1}, -1, \epsilon_{t+1}, \ldots, \epsilon_n) - \phi(y_1, \ldots, y_{t-1}, +1, \epsilon_{t+1}, \ldots, \epsilon_n)\right]$.

Proof:
Again the reverse direction is easy just plug in random outcomes.
For the other direction is similar to Cover's result proof, only this time we don't need restriction on ϕ since we can place any magnitude bet.

Betting Example

- We are given that m teams are playing in pairs n matches against each other.
- Assume that the pairs that are going to play for the n rounds are announced in advance as $(i_1, j_1), \ldots, (i_n, j_n)$.
- Benchmark we want to consider is one where, after the fact, we assign scores $w[1], \ldots, w[m]$ to each of the m teams and when teams i, j play, the benchmark makes a bet of $w[i] - w[j]$
- Maximum bet value allowed by the benchmark is some value B

Goal: Strategy to minimize

$$\mathrm{Reg}_n := \sum_{t=1}^{n} -(\hat{y}_t \cdot y_t) - \min_{w \in \mathbb{R}^m : \max_{i,j} w[i] - w[j] \leq B} \sum_{t=1}^{n} (-y_t \cdot (w[i_t] - w[j_t]))$$

Betting Example

- Define
  $\phi(y_1,\ldots,y_n) = \min_{w \in \mathbb{R}^m : \max_{i,j} w[i]-w[j] \le B} \sum_{t=1}^{n} \left(-y_t \cdot (w[i_t] - w[j_t])\right) + C_n$
- Using the Lemma lets write out C_n

$$
\begin{aligned}
C_n &= -\mathbb{E}_{\epsilon_1,\ldots,\epsilon_n}\left[\min_{w \in \mathbb{R}^m : \max_{i,j} w[i]-w[j] \le B} \sum_{t=1}^{n} \left(-\epsilon_t \cdot (w[i_t] - w[j_t])\right)\right] \\
&= \mathbb{E}_{\epsilon_1,\ldots,\epsilon_n}\left[\max_{w \in \mathbb{R}^m : \max_{i,j} w[i]-w[j] \le B} \sum_{t=1}^{n} \left(\epsilon_t \cdot (w[i_t] - w[j_t])\right)\right] \\
&= \mathbb{E}_{\epsilon_1,\ldots,\epsilon_n}\left[\max_{w \in [0,B]^m} \sum_{t=1}^{n} \left(\epsilon_t \cdot (w[i_t] - w[j_t])\right)\right] \\
&= \mathbb{E}_{\epsilon_1,\ldots,\epsilon_n}\left[\max_{w \in [0,B]^m} \sum_{i=1}^{m} \sum_{t=1}^{n} \epsilon_t \cdot w[i] \left(\mathbf{1}\{i_t = i\} - \mathbf{1}\{j_t = i\}\right)\right] \\
&= \mathbb{E}_{\epsilon_1,\ldots,\epsilon_n}\left[\sum_{i=1}^{m} \max_{w[i] \in [0,B]} \sum_{t=1}^{n} \epsilon_t \cdot w[i] \left(\mathbf{1}\{i_t = i\} - \mathbf{1}\{j_t = i\}\right)\right] \\
&= B\, \mathbb{E}_{\epsilon_1,\ldots,\epsilon_n}\left[\sum_{i=1}^{m} \max\left\{\sum_{t=1}^{n_i} \epsilon_t, 0\right\}\right] \le \frac{B}{2} \sum_{i=1}^{m} \sqrt{n_i} \le \frac{B}{2}\sqrt{mn}
\end{aligned}
$$

What if we didn’t know the which pairs play in advance?

BETTING GAME WITH ARBITRARY CONTEXTS

For $t = 1$ to n

- Context $x_t \in \mathcal{X}$ is provided.
- Learner has to place a bet of amount $|\hat{y}_t|$ on either team A or team B (sign of $\hat{y}_t$ tells us which team we bet on)
- Outcome of the round is revealed as $y_t \in \{\pm 1\}$
- Learner looses money $\ell_t = -y_t \cdot \hat{y}_t$

End For

Goal: given $\phi : \mathcal{X}^n \times \{\pm 1\}^n \mapsto \mathbb{R}$ can we guarantee:

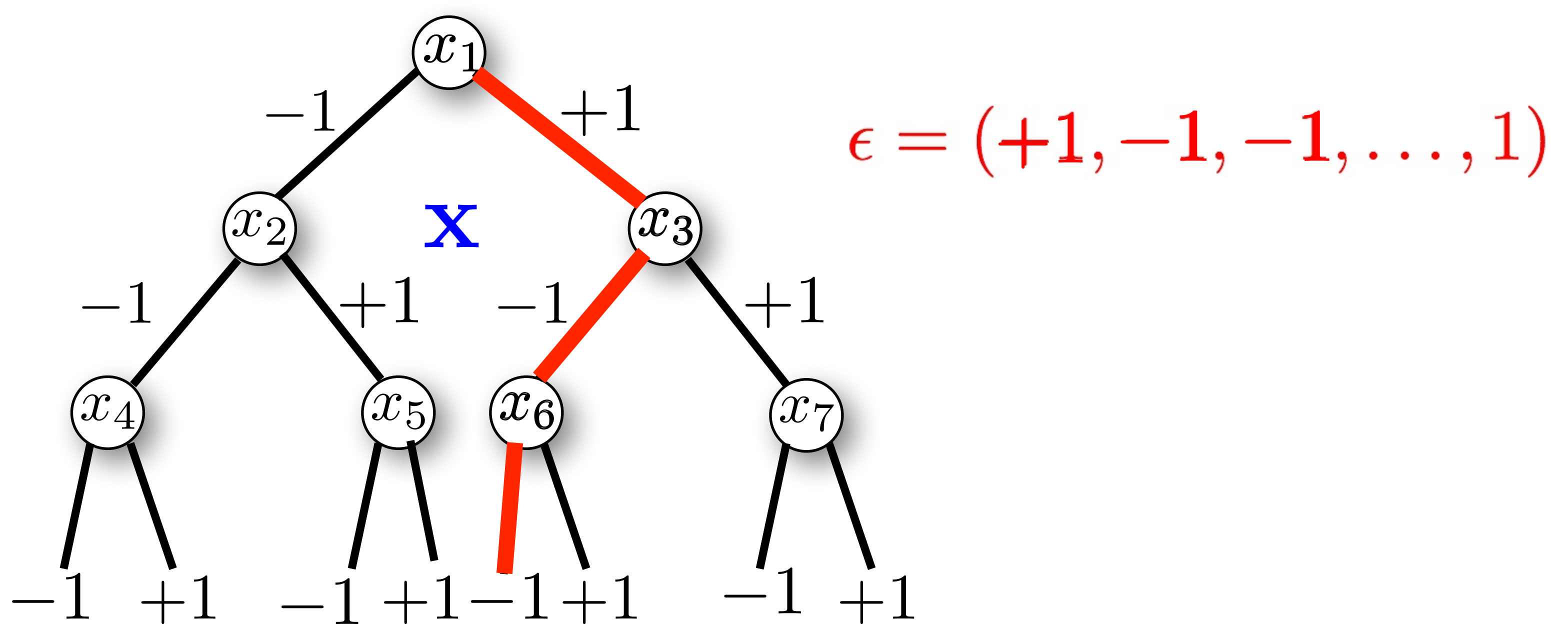
$$\sum_{t=1}^{n} (-y_t \cdot \hat{y}_t) \leq \phi(x_1, \ldots, x_n, y_1, \ldots, y_n)$$

Eg. $x_t = (i_t, j_t)$ teams playing can be chosen arbitrarily as we go.

Meet the Trees

Definition

An $\mathcal{X}$ valued binary tree is a sequence of maapping $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ where $\mathbf{x}_t : \{\pm 1\}^{t-1} \mapsto \mathcal{X}$. Here $\mathbf{x}_1 \in \mathcal{X}$ is a constant



$\mathbf{x}_1 = x_1$ $\quad\quad$ $\mathbf{x}_2(+1) = x_3$ $\quad\quad$ $\mathbf{x}_3(+1, -1) = x_6$

Arbitrary Covariates Result

Lemma

For any $\phi : \mathcal{X}^n \times \{\pm 1\}^n \mapsto \mathbb{R}$*, there exists a strategy with can guarantee that*

$$\sum_{t=1}^{n} -(\hat{y}_t \cdot y_t) \leq \phi(x_1, y_1, \ldots, x_n, y_n)$$

If and only if

$$\inf_{\mathbf{x}} \mathbb{E}\left[\phi(\mathbf{x}_1, \mathbf{x}_2(\epsilon_1), \ldots, \mathbf{x}_n(\epsilon_1, \ldots, \epsilon_{n-1}), \epsilon_1, \ldots, \epsilon_n)\right] \geq 0$$

Arbitrary Covariates Result

- If we consider the example

$$\phi(x_1, \ldots, x_n, y_1, \ldots, y_n) = \min_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} -y_t \cdot f(x_t) \right\} + C_n(\mathcal{F})$$

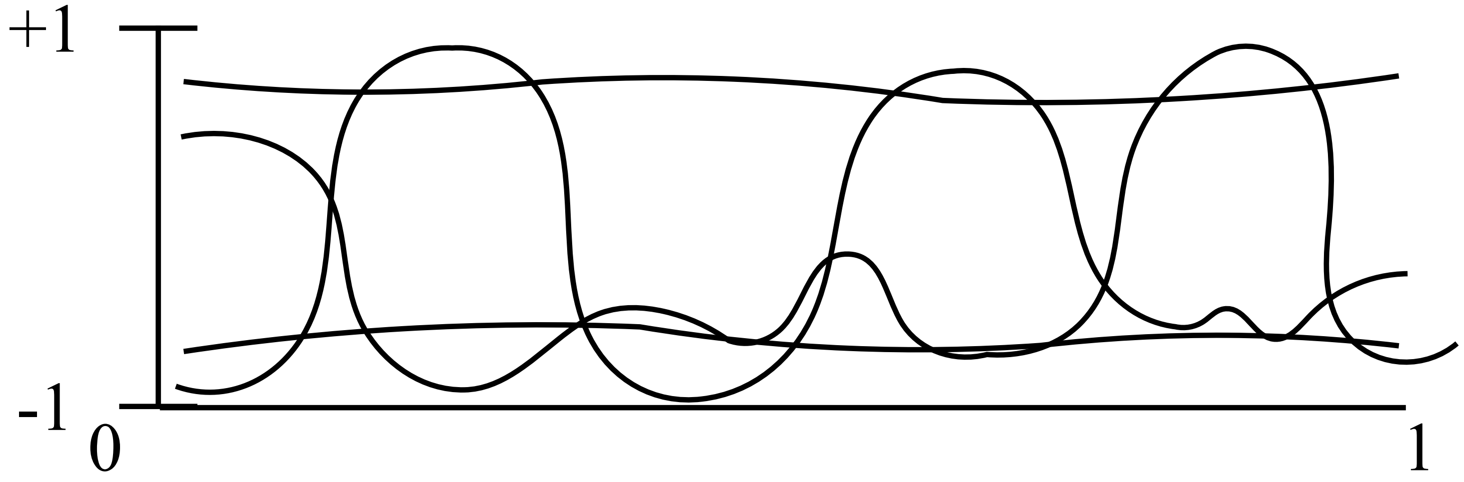
- In this case, using the lemma we get:

$$\begin{aligned} C_n(\mathcal{F}) &= -\mathbb{E}_\epsilon \left[\min_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} -\epsilon_t \cdot f(\mathbf{x}_t(\epsilon_1, \ldots, \epsilon_{t-1})) \right\} \right] \\ &= \mathbb{E}_\epsilon \left[\max_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t \cdot f(\mathbf{x}_t(\epsilon_1, \ldots, \epsilon_{t-1})) \right\} \right] := \mathrm{Rad}_n(\mathcal{F}) \end{aligned}$$

Rademacher Complexity

Rademacher Complexity

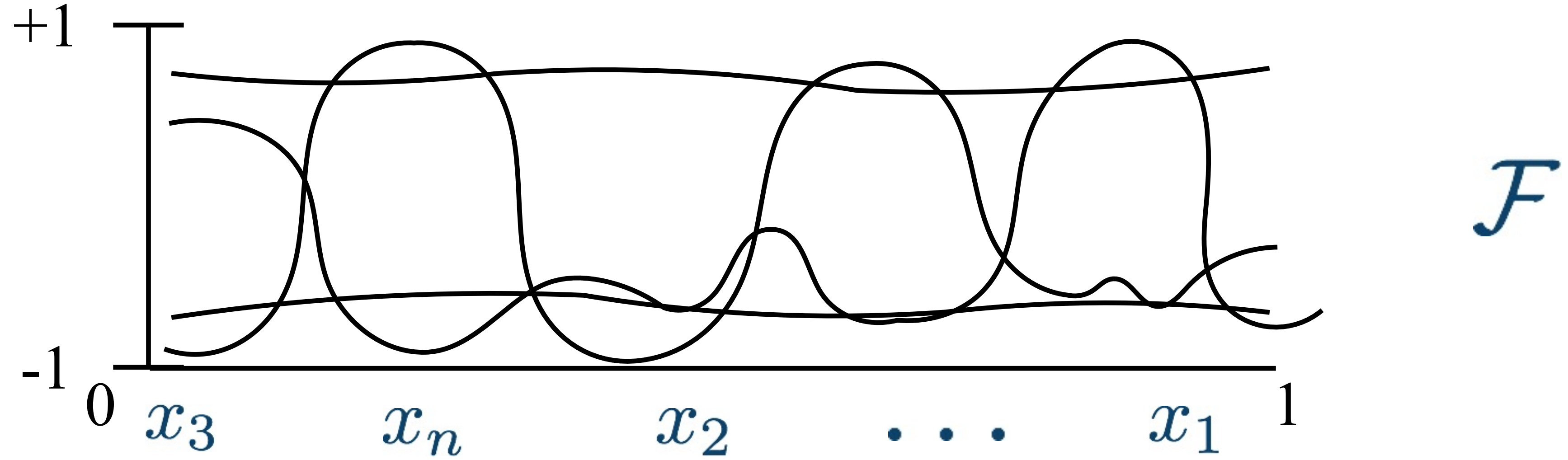
$$\text{Example : } \mathcal{X} = [0,1],\ \mathcal{Y} = [-1,1]$$



$\mathcal{F}$

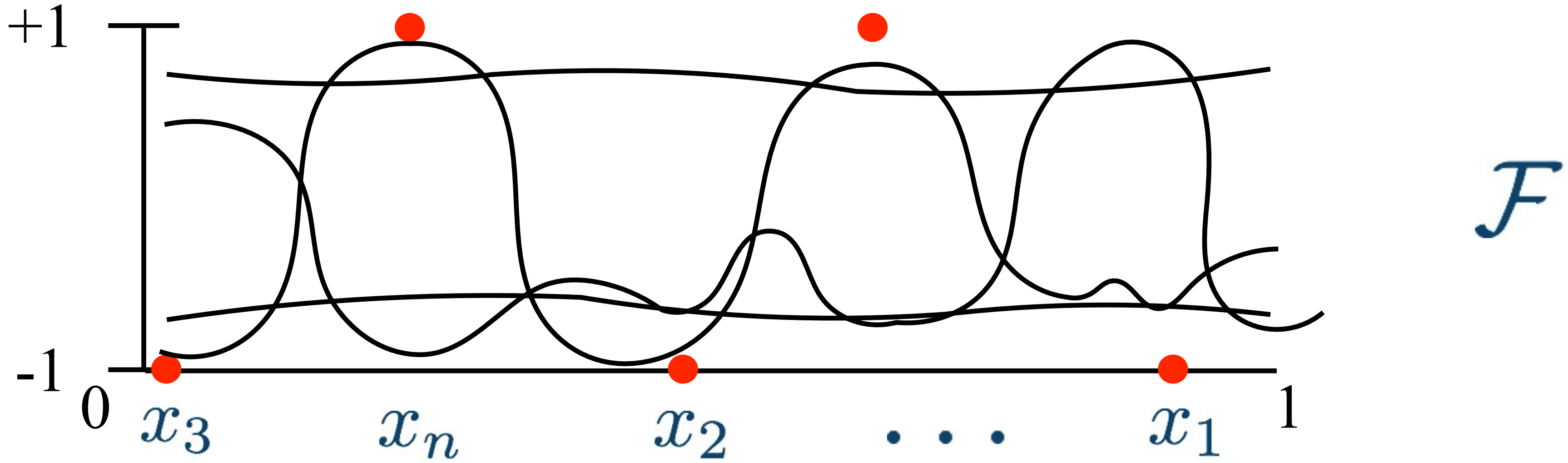
RADEMACHER COMPLEXITY

$$\text{Example : } \mathcal{X} = [0,1],\ \mathcal{Y} = [-1,1]$$



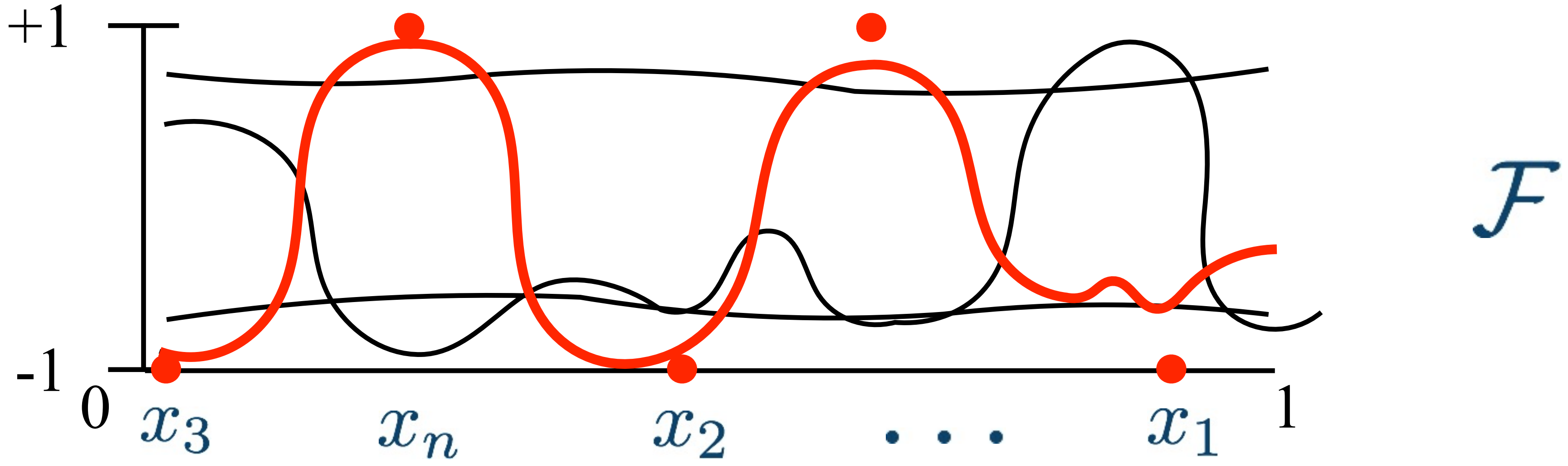
Rademacher Complexity

$$\text{Example : } \mathcal{X} = [0,1],\ \mathcal{Y} = [-1,1]$$

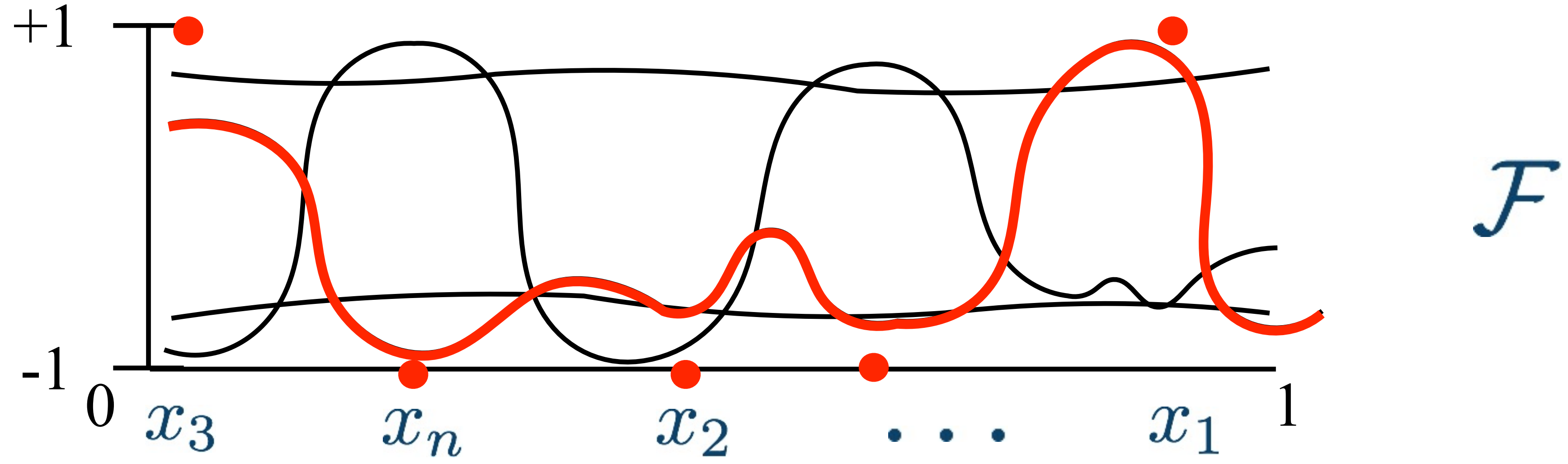


Rademacher Complexity

$$\text{Example}: \mathcal{X} = [0,1],\ \mathcal{Y} = [-1,1]$$

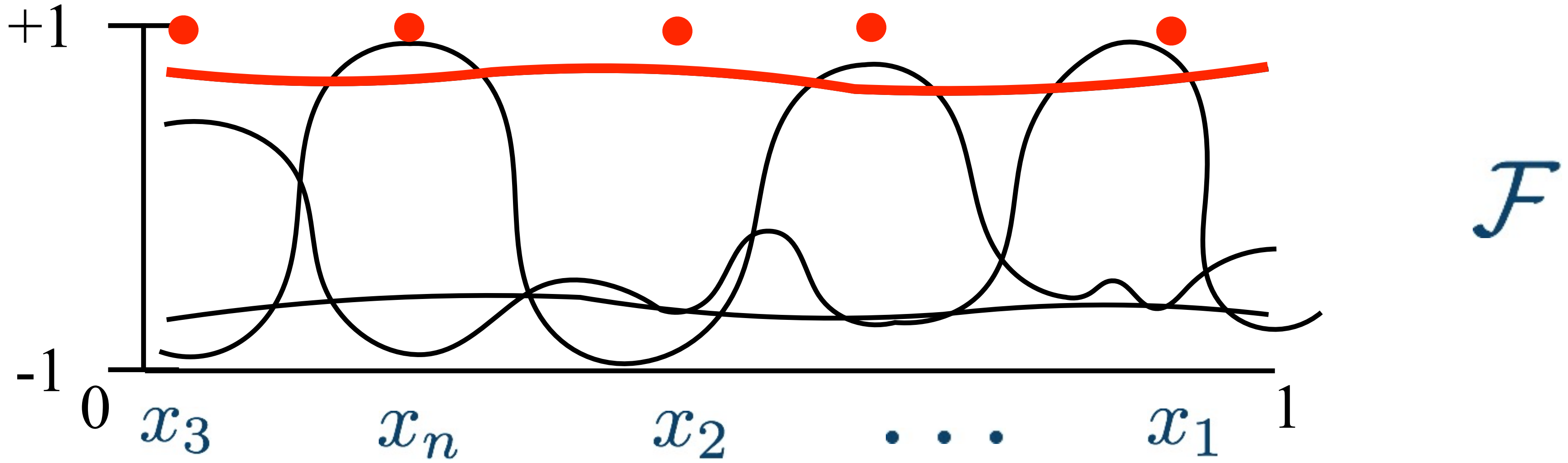


Example : $\mathcal{X} = [0, 1], \ \mathcal{Y} = [-1, 1]$



Rademacher Complexity

$$\text{Example}: \mathcal{X} = [0,1],\ \mathcal{Y} = [-1,1]$$



Rademacher Complexity

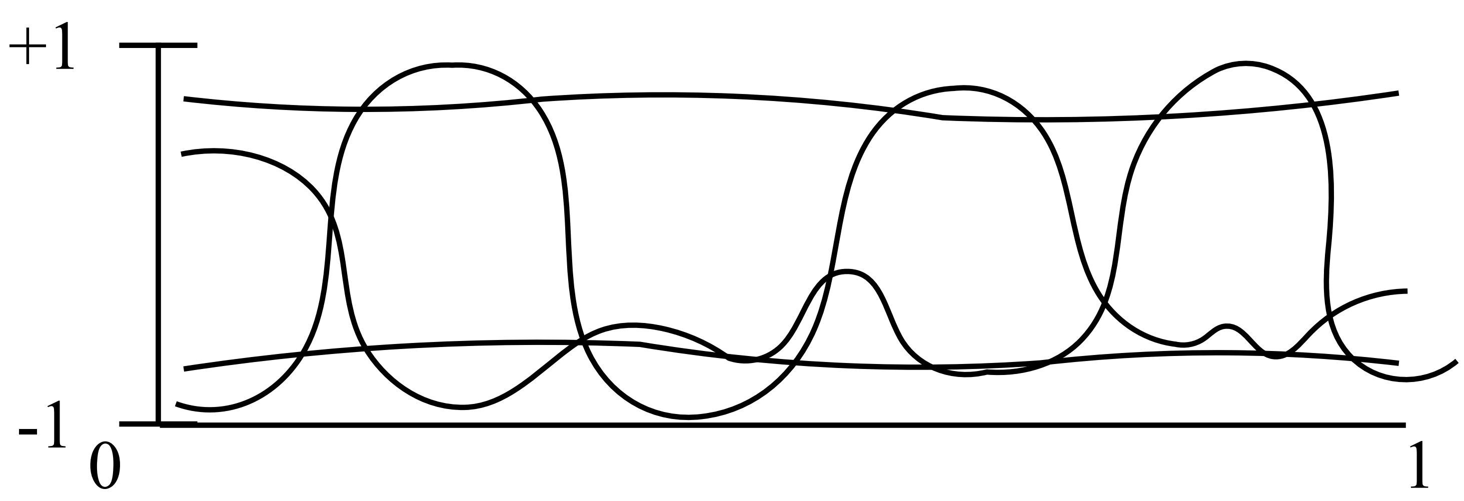
$$\mathcal{R}_n(\mathcal{F}) := \sup_{x_1,\dots,x_n} \mathbb{E}_{\epsilon}\left[\frac{2}{n}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n} \epsilon_t f(x_t)\right|\right]$$

Rademacher Complexity

$$\mathcal{R}_n(\mathcal{F}) := \sup_{x_1,\ldots,x_n} \mathbb{E}_\epsilon \left[\frac{2}{n} \sup_{f\in\mathcal{F}} \left| \sum_{t=1}^{n} \epsilon_t f(x_t) \right| \right]$$

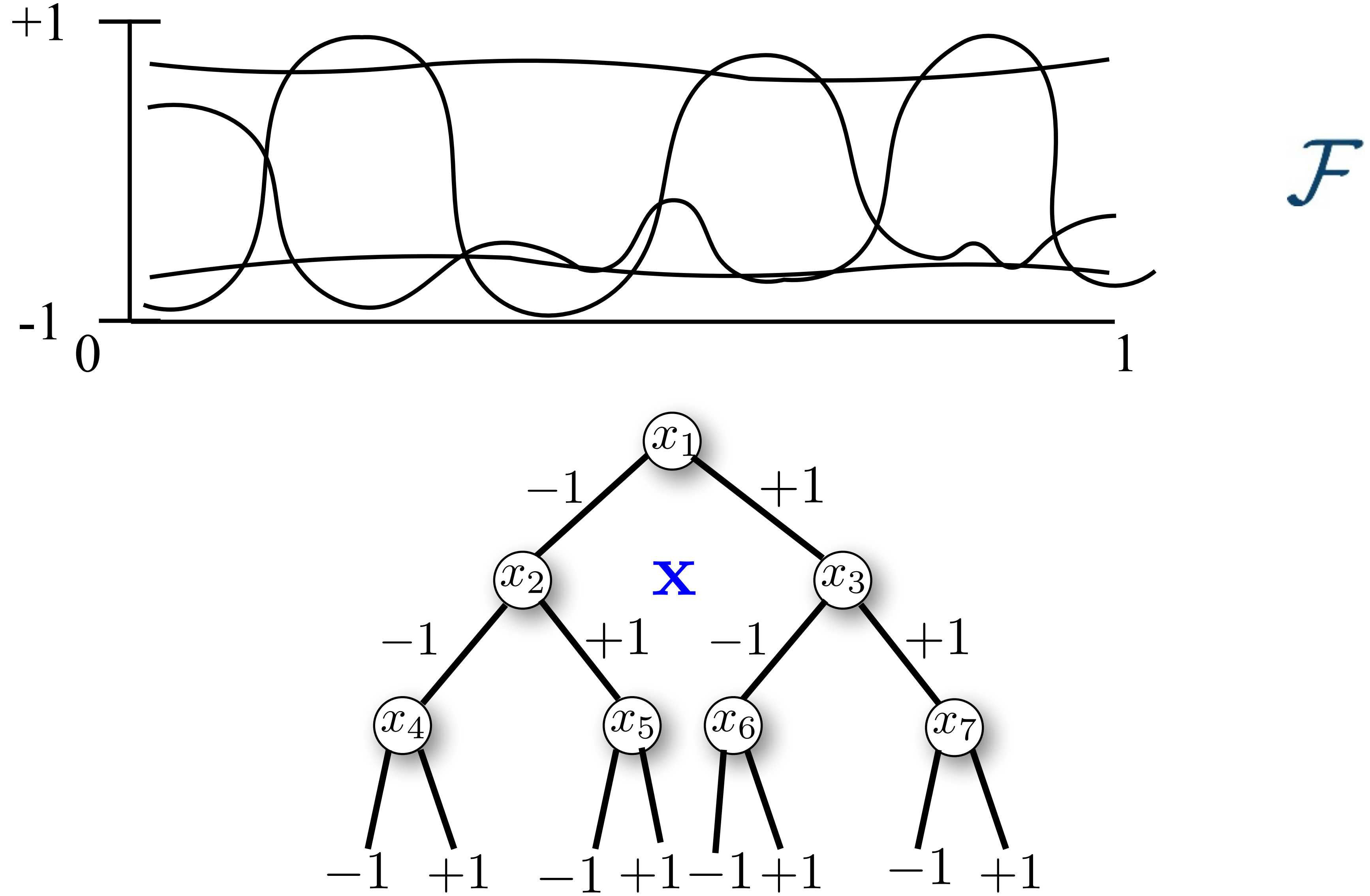
sequence random signs max correlation

Sequential Rademacher Complexity

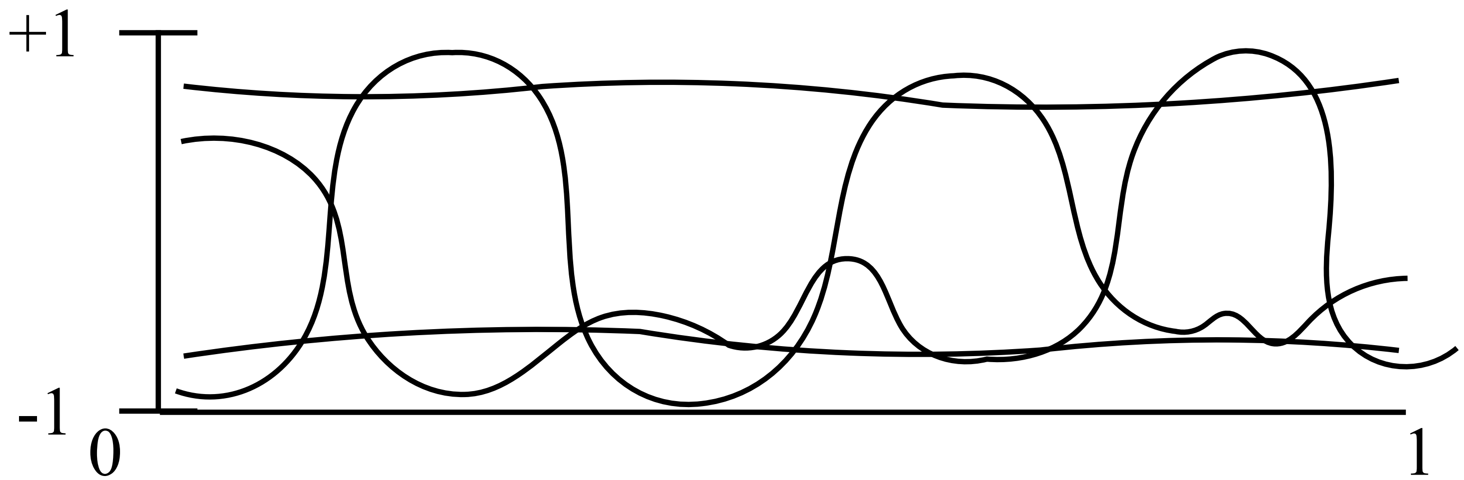


$\mathcal{F}$

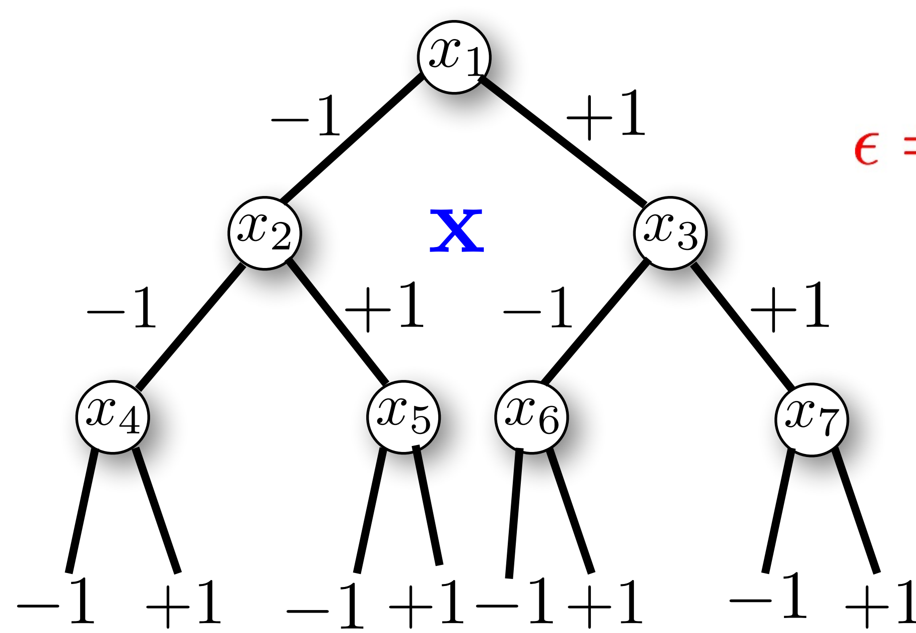
Sequential Rademacher Complexity



Sequential Rademacher Complexity

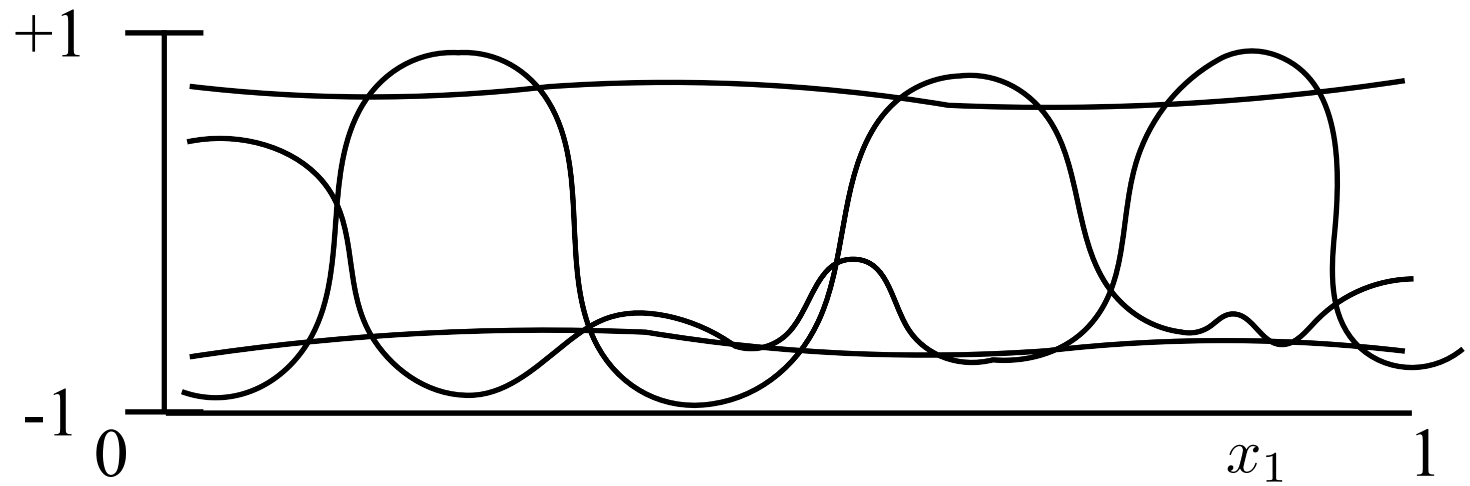


$\mathcal{F}$

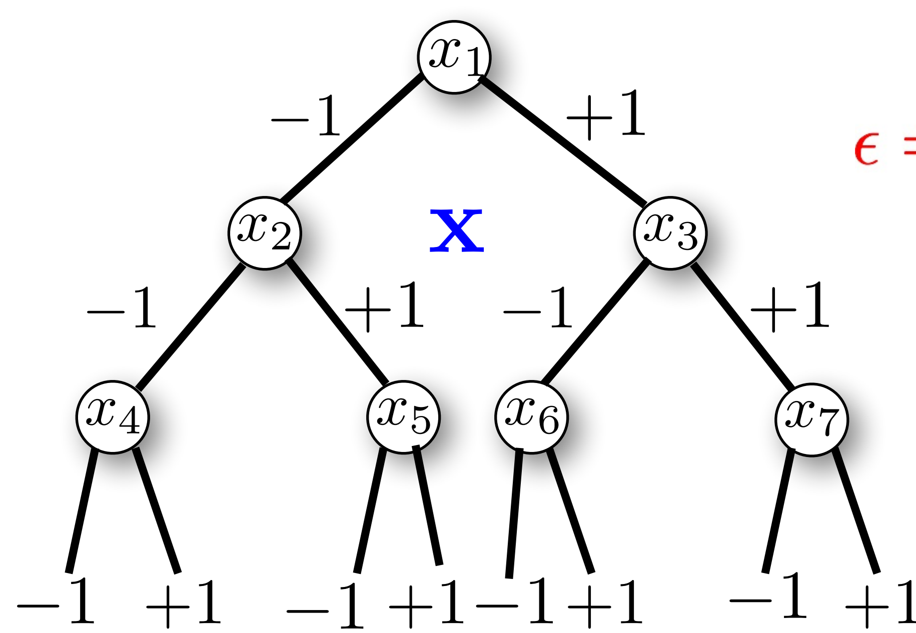


$\epsilon = (+1, -1, -1, \ldots, 1)$

SEQUENTIAL RADEMACHER COMPLEXITY

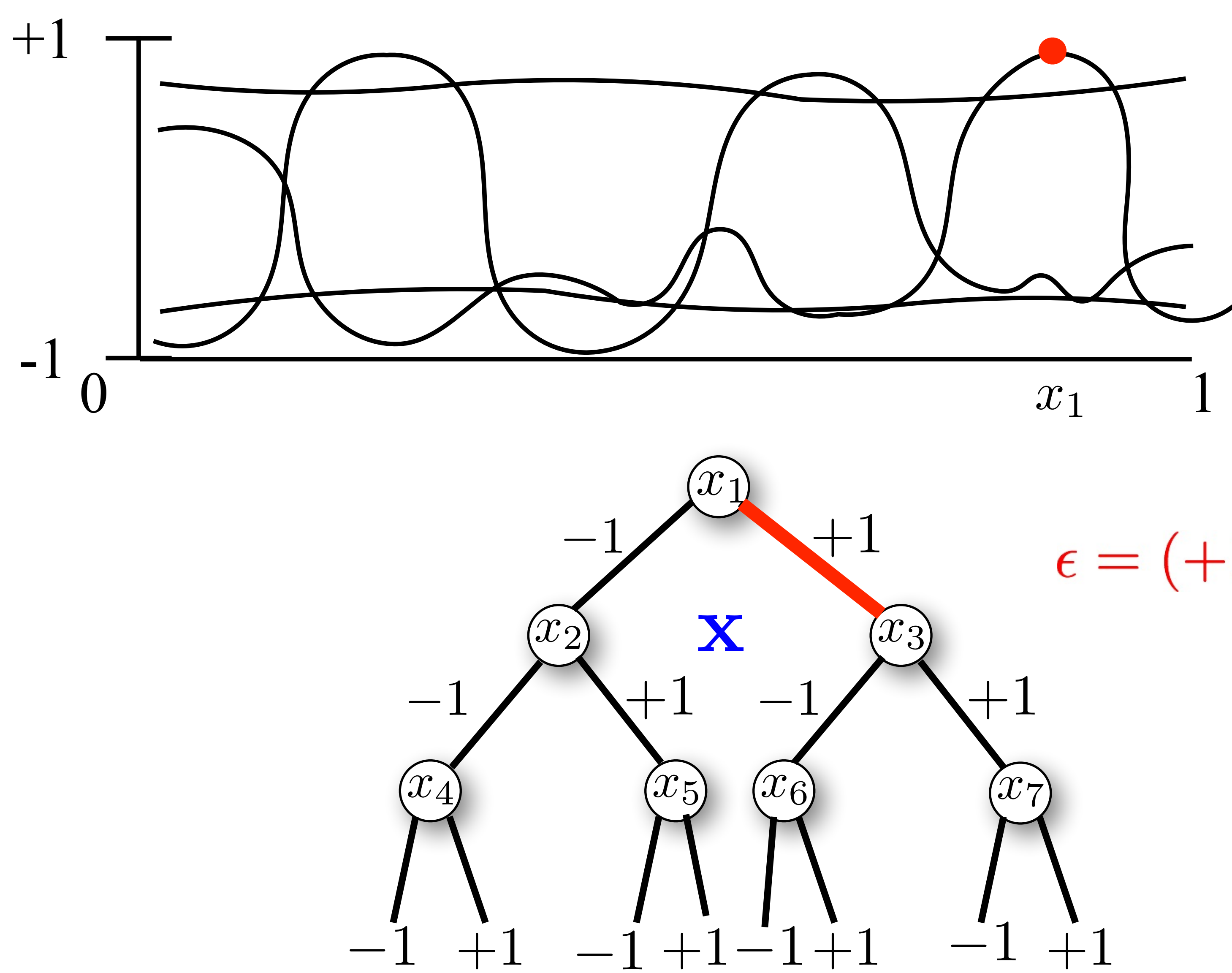


$\mathcal{F}$



$\epsilon = (+1, -1, -1, \ldots, 1)$

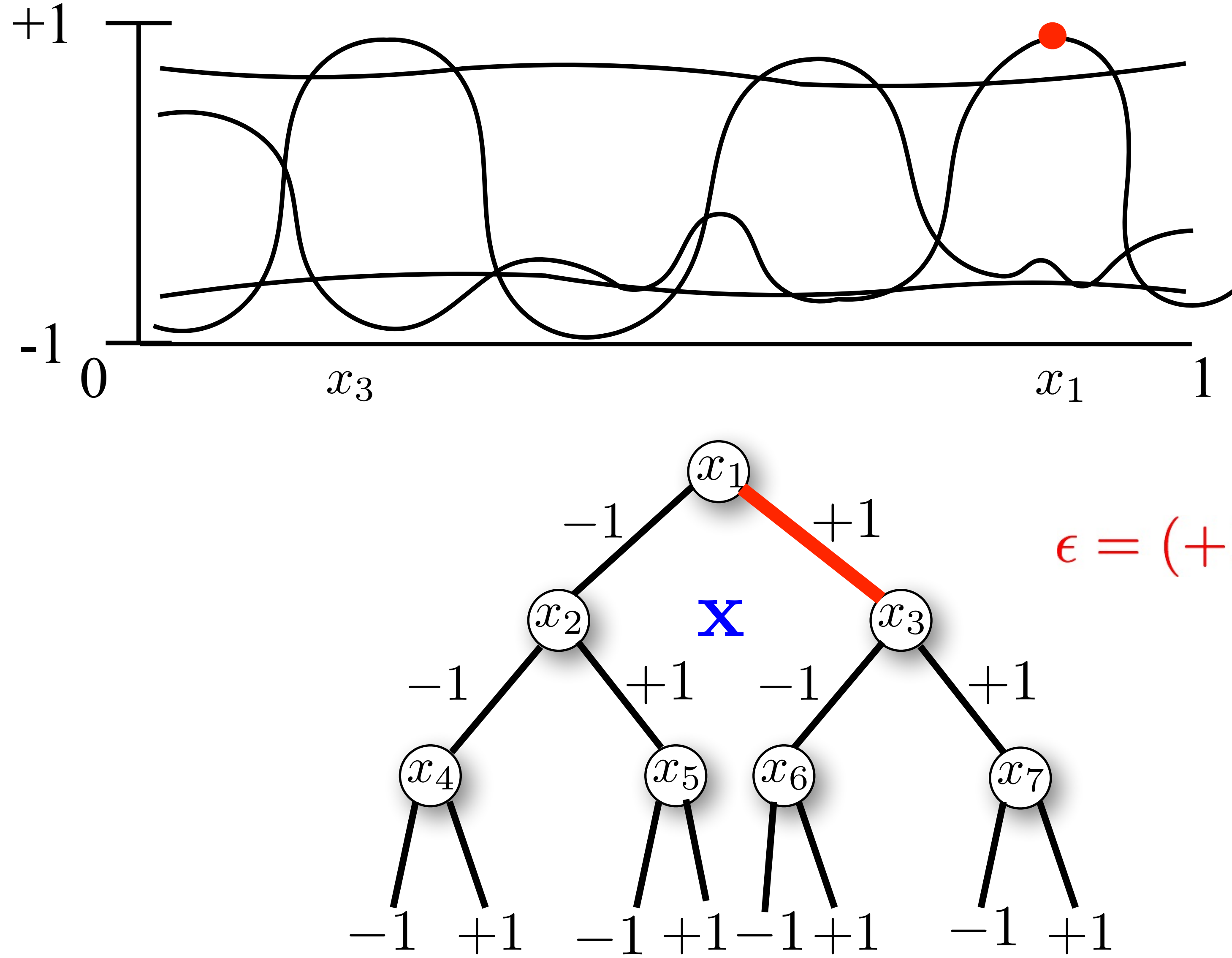
Sequential Rademacher Complexity



$\mathcal{F}$

$\epsilon = (+1, -1, -1, \ldots, 1)$

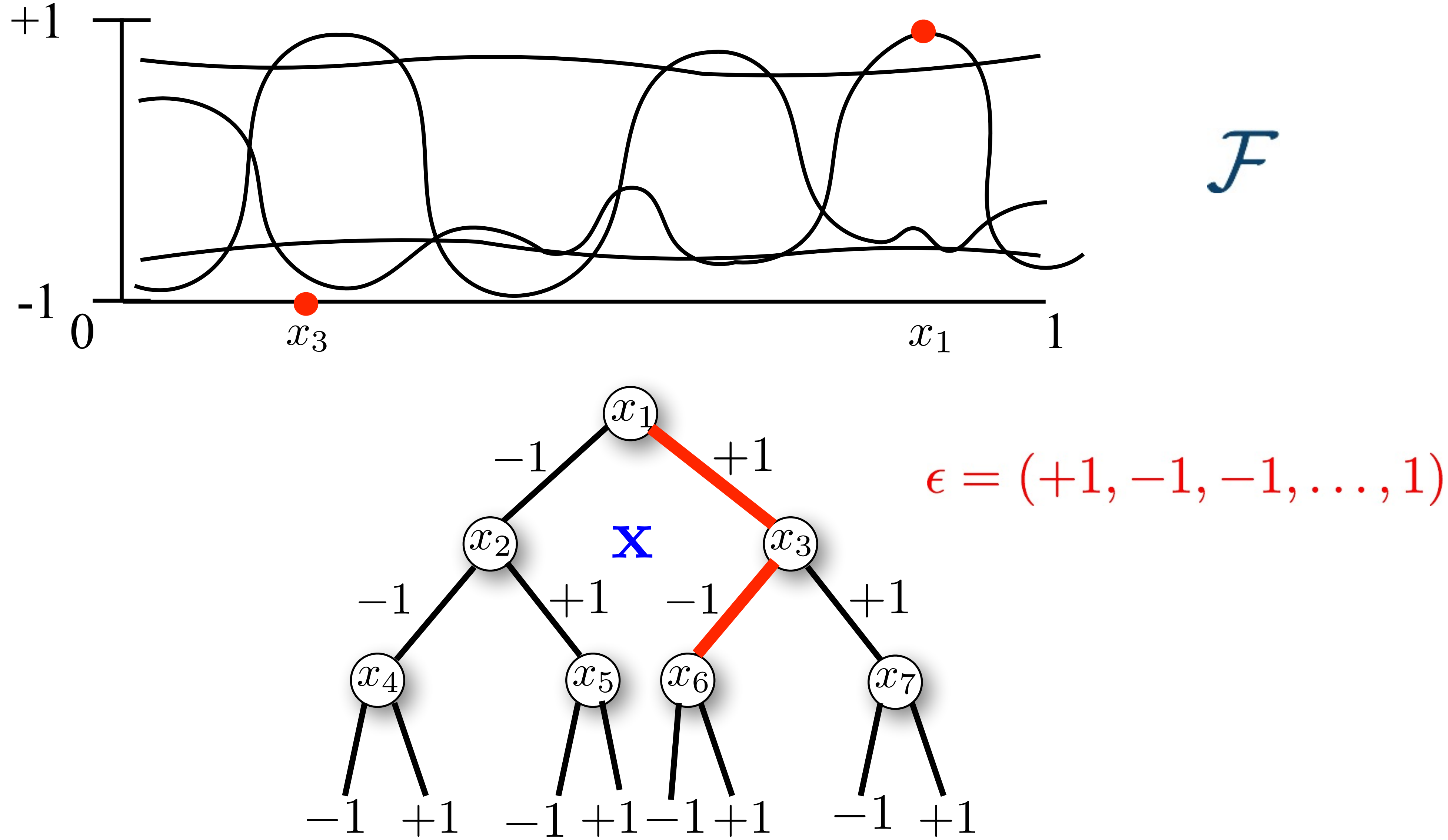
Sequential Rademacher Complexity



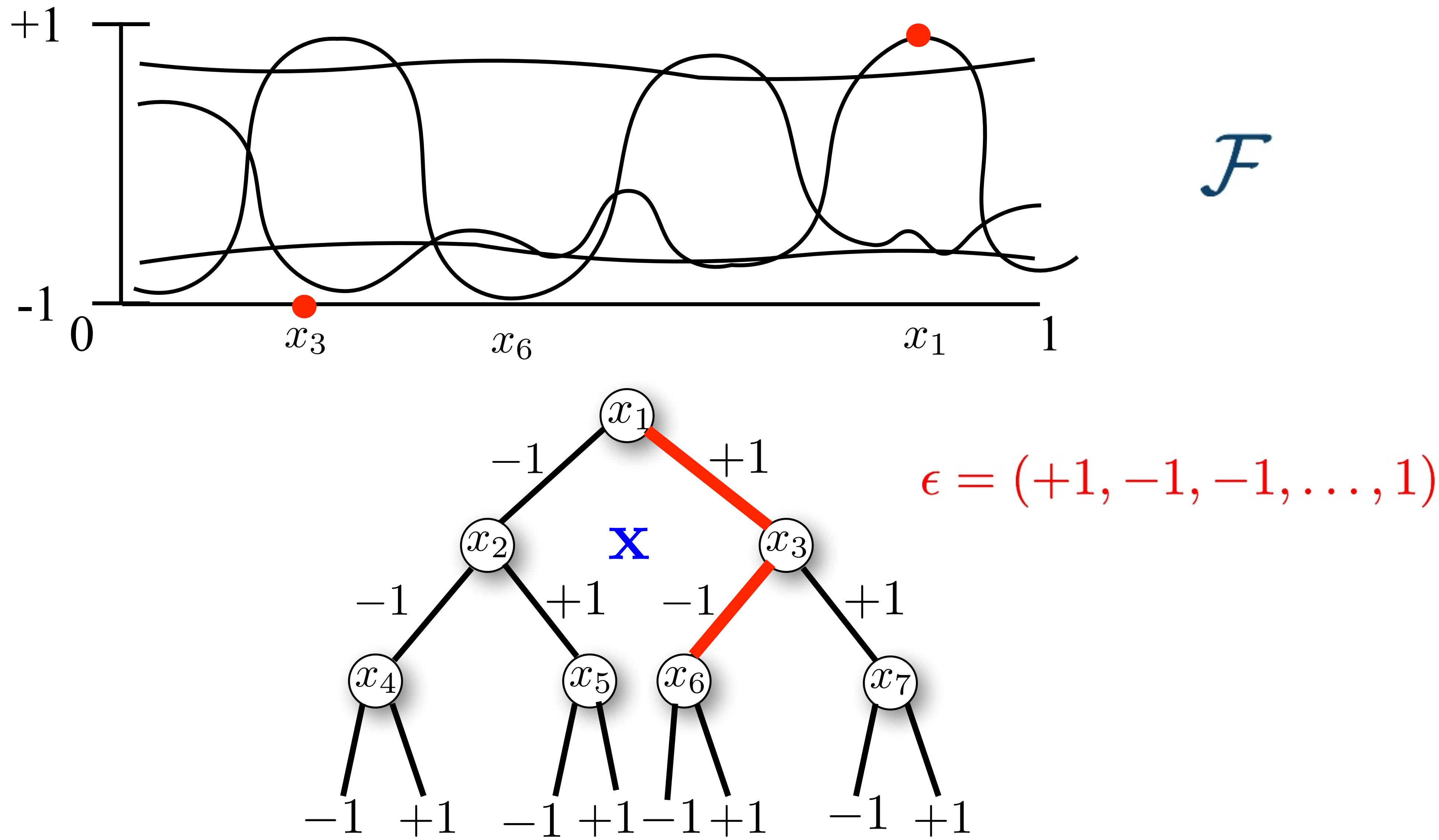
$\mathcal{F}$

$\epsilon = (+1, -1, -1, \ldots, 1)$

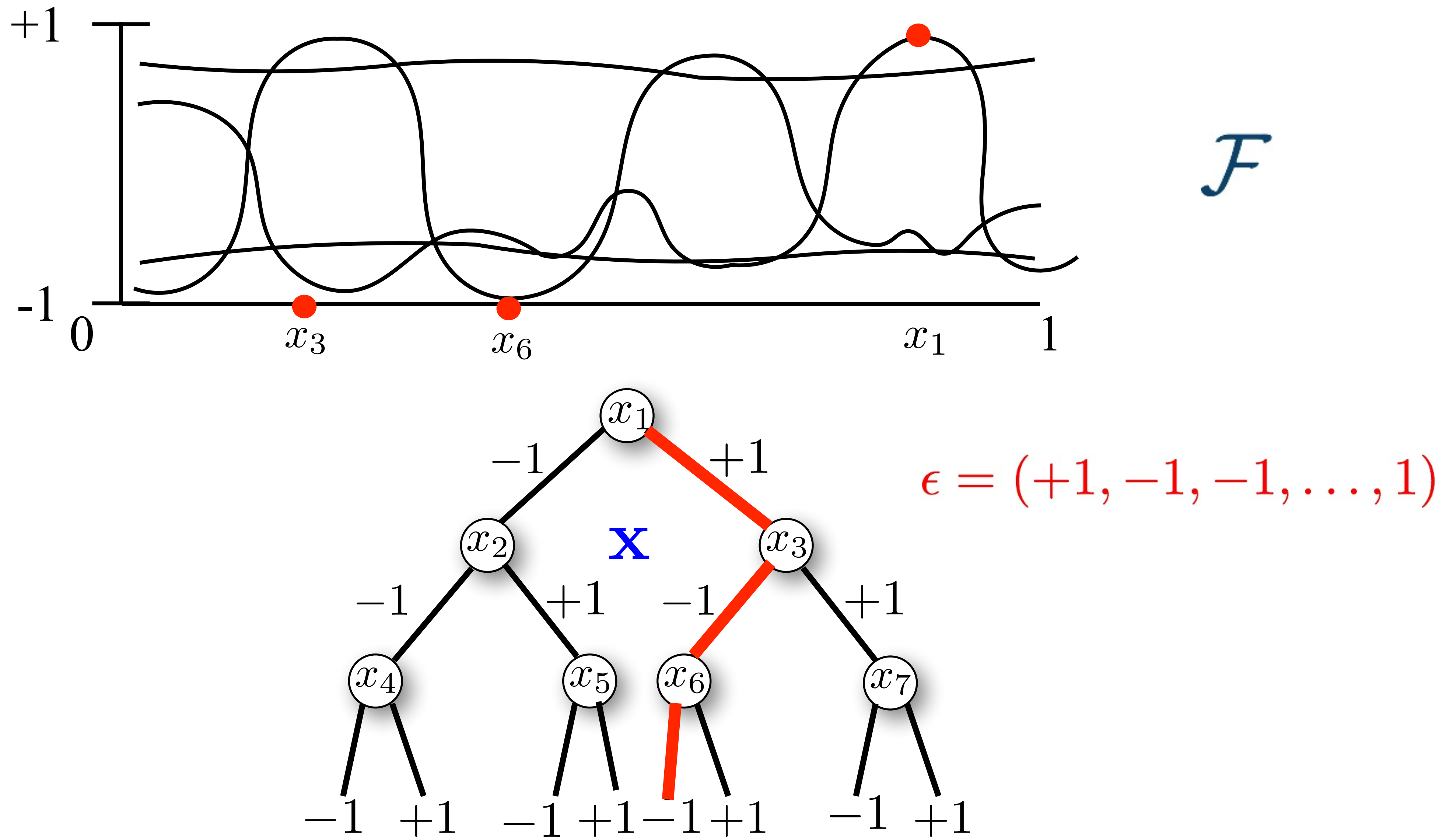
Sequential Rademacher Complexity



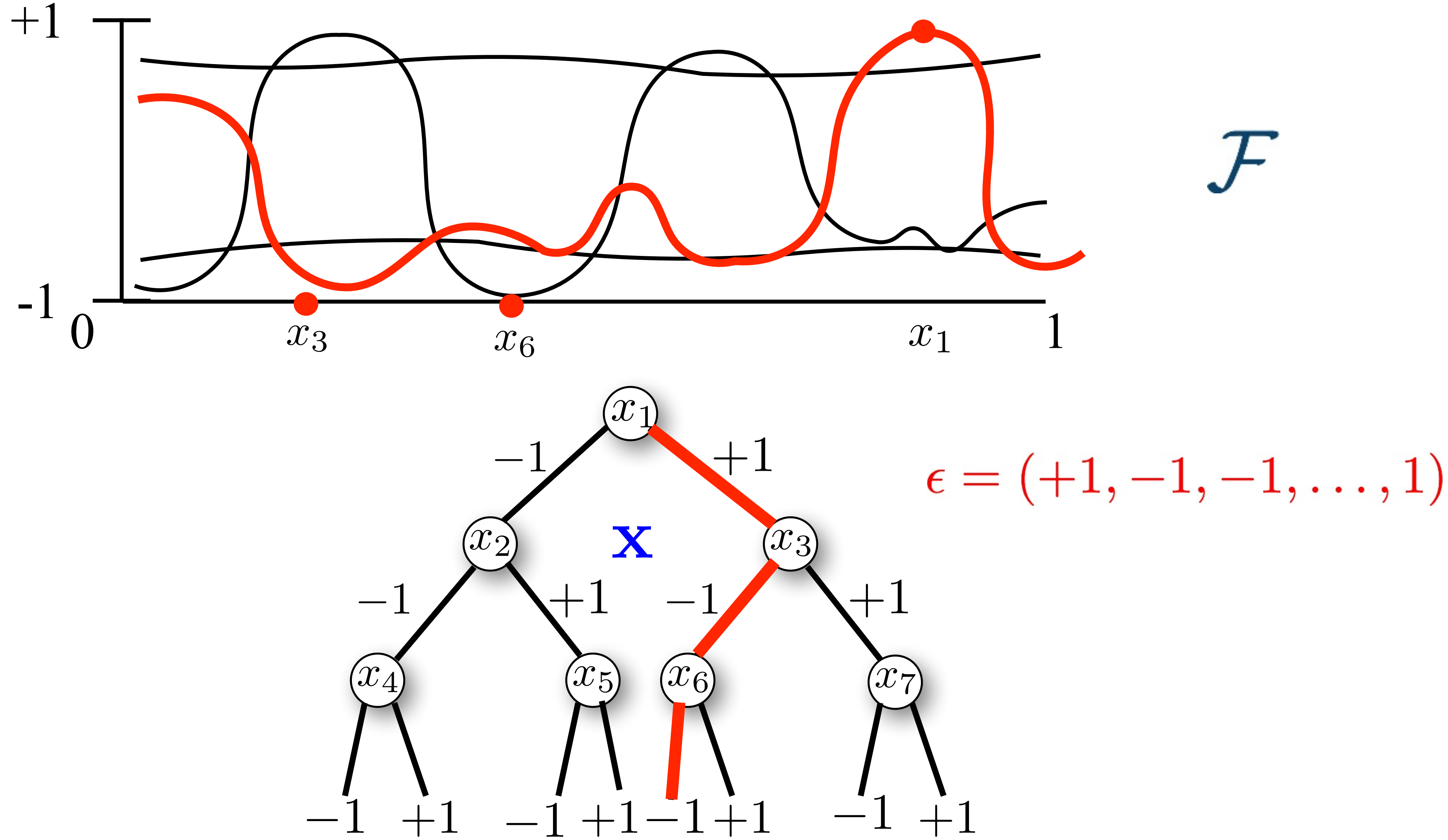
SEQUENTIAL RADEMACHER COMPLEXITY



Sequential Rademacher Complexity



Sequential Rademacher Complexity



Online Supervised Learning

Given: model class $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$, convex L Lipschitz loss $\ell : \mathbb{R} \times \mathcal{Z} \mapsto \mathbb{R}$

For $t = 1$ to n

- Context $x_t \in \mathcal{X}$ is provided.
- Learner picks prediction $\hat{y}_t \in \mathbb{R}$
- Outcome of the round $z_t \in \mathcal{Z}$ is revealed
- Learner suffers loss $\ell(\hat{y}_t, z_t)$

End For

Goal: minimize regret

$$\mathrm{Reg}_n(\mathcal{F}) = \sum_{t=1}^{n} \ell(\hat{y}_t, z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), z_t)$$

Online Supervised Learning: Reduction

$$
\begin{aligned}
\mathrm{Reg}_n(\mathcal{F}) &= \sum_{t=1}^{n} \ell(\hat{y}_t, z_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), z_t) \\
&= \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \left(\ell(\hat{y}_t, z_t) - \ell(f(x_t), z_t)\right) \\
&\le \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \partial\ell(\hat{y}_t, z_t) \cdot (\hat{y}_t - f(x_t)) \\
&= L \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \mathbb{E}_{b_t \sim \frac{1+\partial\ell(\hat{y}_t, z_t)/L}{2}} \left[b_t \cdot (\hat{y}_t - f(x_t))\right] \\
&\le L\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} b_t \cdot (\hat{y}_t - f(x_t))\right] \\
&= L\mathbb{E}\left[\sum_{t=1}^{n} b_t \cdot \hat{y}_t - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} b_t \cdot f(x_t)\right]
\end{aligned}
$$

Online Supervised Learning: Reduction

- So online supervised learning with any convex, L-Lipschitz loss we get guarantee that

$$\mathrm{Reg}_n(\mathcal{F}) \leq L\,\mathrm{Rad}_n(\mathcal{F})$$

- With a more complicated proof technique, one can show the same result but without requiring convexity of loss ℓ.