Chapter 3.5: Model Selection

Geerten Koers

Reading group *Weak Convergence and Empirical Processes*

2020-09-21

1. Model selection: General Result

2. Model selection: Statistical Learning

Model selection: General Result

- Let $\mathcal{K}$ be a countable set;

Model selection: General Result

- Let $\mathcal{K}$ be a countable set;
- For $k \in \mathcal{K}$, let Θ_k be a subset of a metric space (Θ, d);

Model selection: General Result

- Let $\mathcal{K}$ be a countable set;
- For $k \in \mathcal{K}$, let Θ_k be a subset of a metric space (Θ, d);
- For $\theta \in \Theta$, let $m_\theta : \mathcal{X} \to \mathbb{R}$ be measurable;

Model selection: General Result

- Let $\mathcal{K}$ be a countable set;
- For $k \in \mathcal{K}$, let Θ_k be a subset of a metric space (Θ, d);
- For $\theta \in \Theta$, let $m_\theta : \mathcal{X} \to \mathbb{R}$ be measurable;
- We want to estimate $\theta^* := \text{argmin}_{\theta \in \Theta} P m_\theta$.

Model selection: General Result

- Let $\mathcal{K}$ be a countable set;
- For $k \in \mathcal{K}$, let Θ_k be a subset of a metric space (Θ, d);
- For $\theta \in \Theta$, let $m_\theta : \mathcal{X} \to \mathbb{R}$ be measurable;
- We want to estimate $\theta^* := \text{argmin}_{\theta \in \Theta} P m_\theta$.
- For $k \in \mathcal{K}$, define $\hat{\theta}_k := \text{argmin}_{\theta \in \Theta_k} \mathbb{P}_n m_\theta$;

Model selection: General Result

- Let $\mathcal{K}$ be a countable set;
- For $k \in \mathcal{K}$, let Θ_k be a subset of a metric space (Θ, d);
- For $\theta \in \Theta$, let $m_\theta : \mathcal{X} \to \mathbb{R}$ be measurable;
- We want to estimate $\theta^* := \text{argmin}_{\theta \in \Theta} P m_\theta$.
- For $k \in \mathcal{K}$, define $\hat{\theta}_k := \text{argmin}_{\theta \in \Theta_k} \mathbb{P}_n m_\theta$;
- For $J : \mathcal{K} \to [0, \infty)$, define $\hat{k} := \text{argmin}_{k \in \mathcal{K}} (\mathbb{P}_n m_{\hat{\theta}_k} + J(k))$;

Model selection: General Result

- Let $\mathcal{K}$ be a countable set;
- For $k \in \mathcal{K}$, let Θ_k be a subset of a metric space (Θ, d);
- For $\theta \in \Theta$, let $m_\theta : \mathcal{X} \to \mathbb{R}$ be measurable;
- We want to estimate $\theta^* := \text{argmin}_{\theta \in \Theta} P m_\theta$.
- For $k \in \mathcal{K}$, define $\hat{\theta}_k := \text{argmin}_{\theta \in \Theta_k} \mathbb{P}_n m_\theta$;
- For $J : \mathcal{K} \to [0, \infty)$, define $\hat{k} := \text{argmin}_{k \in \mathcal{K}} (\mathbb{P}_n m_{\hat{\theta}_k} + J(k))$;
- Estimate by $\hat{\theta} := \hat{\theta}_{\hat{k}} = \text{argmin}_{\theta \in \cup_k \Theta_k} \left(\mathbb{P}_n m_\theta + \sum_k J(k) 1_{\theta \in \Theta_k} \right)$.

Model selection: General Result

- Theorem 3.4.1 shows that the rate of convergence is described by $\phi_{n,k}(\delta_{n,k}) \lesssim \sqrt{n}\delta_{n,k}^2$.

Model selection: General Result

- Theorem 3.4.1 shows that the rate of convergence is described by $\phi_{n,k}(\delta_{n,k}) \lesssim \sqrt{n}\delta_{n,k}^2$.
- Within Θ_k, the best estimator may be defined as $\theta_k^* := \text{argmin}_{\theta \in \Theta_k} Pm_\theta$.

Model selection: General Result

- Theorem 3.4.1 shows that the rate of convergence is described by $\phi_{n,k}(\delta_{n,k}) \lesssim \sqrt{n}\delta_{n,k}^2$.
- Within Θ_k, the best estimator may be defined as $\theta_k^* := \text{argmin}_{\theta \in \Theta_k} Pm_\theta$.
- The best estimator may also be defined as $\bar{\theta}_k = \text{argmin}_{\theta \in \Theta_k} d^2(\theta, \theta^*)$.

Model selection: General Result

- Theorem 3.4.1 shows that the rate of convergence is described by $\phi_{n,k}(\delta_{n,k}) \lesssim \sqrt{n}\delta_{n,k}^2$.
- Within Θ_k, the best estimator may be defined as $\theta_k^* := \text{argmin}_{\theta \in \Theta_k} Pm_\theta$.
- The best estimator may also be defined as $\bar{\theta}_k = \text{argmin}_{\theta \in \Theta_k} d^2(\theta, \theta^*)$.
- Only using Θ_k, we hope for a rate of order $O_P(\delta_{n,k}) + d(\bar{\theta}_k, \theta^*)$.

Model selection: General Result

- Theorem 3.4.1 shows that the rate of convergence is described by $\phi_{n,k}(\delta_{n,k}) \lesssim \sqrt{n}\delta_{n,k}^2$.
- Within Θ_k, the best estimator may be defined as $\theta_k^* := \text{argmin}_{\theta \in \Theta_k} Pm_\theta$.
- The best estimator may also be defined as $\bar{\theta}_k = \text{argmin}_{\theta \in \Theta_k} d^2(\theta, \theta^*)$.
- Only using Θ_k, we hope for a rate of order $O_P(\delta_{n,k}) + d(\bar{\theta}_k, \theta^*)$.
- Using multiple models Θ_k at once, we could hope for a rate of order $\inf_k(\delta_{n,k} + d(\bar{\theta}_k, \theta^*))$.

Model selection: General Result

- Theorem 3.4.1 shows that the rate of convergence is described by $\phi_{n,k}(\delta_{n,k}) \lesssim \sqrt{n}\delta_{n,k}^2$.
- Within Θ_k, the best estimator may be defined as $\theta_k^* := \text{argmin}_{\theta\in\Theta_k} Pm_\theta$.
- The best estimator may also be defined as $\bar{\theta}_k = \text{argmin}_{\theta\in\Theta_k} d^2(\theta, \theta^*)$.
- Only using Θ_k, we hope for a rate of order $O_P(\delta_{n,k}) + d(\bar{\theta}_k, \theta^*)$.
- Using multiple models Θ_k at once, we could hope for a rate of order $\inf_k(\delta_{n,k} + d(\bar{\theta}_k, \theta^*))$.
- This may be unrealistic.

Model selection: General Result

This is an M-estimator, use Theorem 3.4.1*!*

Model selection: General Result

This is an M-estimator, use Theorem 3.4.1*!*

We want to compare the rate of convergence to the rate of convergence of the best model in the list (even though it is unknown beforehand).

Model selection: General Result

This is an M-estimator, use Theorem 3.4.1!

We want to compare the rate of convergence to the rate of convergence of the best model in the list (even though it is unknown beforehand).

Definition

(x_k, B_k) are <u>co-monotone</u> if $x_k B_{k'} \leq x_k B_k \vee x_{k'} B_{k'}$ for all $k, k' \in \mathcal{K}$.

Model selection: General Result

Theorem (3.5.4 1/2)

Let $\{m_\theta : \theta \in \Theta\}$ *be a class of measurable functions* $m_\theta : \chi \to \mathbb{R}$ *indexed by a metric space* (Θ, d) *such that, for some constants* B_k,

$$\begin{aligned} P(m_\theta - m_{\theta'})^2 &\leq d^2(\theta, \theta'), && \theta, \theta' \in \Theta, \\ \|m_\theta\|_\infty &\leq B_k, && \theta \in \Theta_k. \end{aligned}$$

Model selection: General Result

Theorem (3.5.4 1/2)

Let $\{m_\theta : \theta \in \Theta\}$ be a class of measurable functions $m_\theta : \chi \to \mathbb{R}$ indexed by a metric space (Θ, d) such that, for some constants B_k,

$$\begin{aligned} P(m_\theta - m_{\theta'})^2 &\leq d^2(\theta, \theta'), & \theta, \theta' \in \Theta, \\ \|m_\theta\|_\infty &\leq B_k, & \theta \in \Theta_k. \end{aligned}$$

Furthermore, assume that for given positive constants C_k,

$$\begin{aligned} C_k P(m_\theta - m_{\theta^*}) &\geq d^2(\theta, \theta^*), \qquad \theta \in \Theta_k, \\ E^* \sup_{\theta \in \Theta_k : d(\theta, \bar{\theta}_k) < \delta} \mathbb{G}_n(m_{\bar{\theta}_k} - m_\theta) &\leq \phi_{n,k}(\delta), \end{aligned}$$

for function $\phi_{n,k} : (0, \infty) \to \mathbb{R}$ such that $\delta \mapsto \phi_{n,k}(\delta)/\delta$ is nonincreasing, and for every $\delta \geq \underline{\delta}_n$.

Model selection: General Result

Theorem (3.5.4 2/2)

Let $J : \mathcal{K} \to [0, \infty)$ *satisfy*

$$J(k) \gtrsim \delta_{n,k}^2 C_k D_k^2 + \frac{1}{n}(B_k + C_k)(x_k + \log(B_k + C_k))$$

for $\delta_{n,k}$ *satisfying* $\phi_{n,k}(\delta_{n,k}) \leq D_k \sqrt{n}\delta_{n,k}^2$ *and* $(x_k : k \in \mathcal{K})$ *numbers such that* $\sum_k e^{-x_k} \leq 1$, *and where the proportionality constant is universal.*

Model selection: General Result

Theorem (3.5.4 2/2)

Let $J : \mathcal{K} \to [0, \infty)$ satisfy

$$J(k) \gtrsim \delta_{n,k}^2 C_k D_k^2 + \frac{1}{n}(B_k + C_k)(x_k + \log(B_k + C_k))$$

for $\delta_{n,k}$ satisfying $\phi_{n,k}(\delta_{n,k}) \leq D_k \sqrt{n}\delta_{n,k}^2$ and $(x_k : k \in \mathcal{K})$ numbers such that $\sum_k e^{-x_k} \leq 1$, and where the proportionality constant is universal. Assume that the x_k are chosen so that the numbers (x_k, B_k) and (x_k, C_k) are co-monotone. Then

$$E^* P(m_{\hat{\theta}} - m_{\theta^*}) \lesssim \inf_{k \in \mathcal{K}} \left(P(m_{\theta_k^*} - m_{\theta^*}) + J(k) + \frac{1}{n} \right).$$

Model selection: General Result

1. The risk $E^*P(m_{\hat{\theta}} - m_{\theta^*})$ should be interpreted as a square distance, so the rate is at most $1/\sqrt{n}$.

Model selection: General Result

1. The risk $E^* P(m_{\hat{\theta}} - m_{\theta^*})$ should be interpreted as a square distance, so the rate is at most $1/\sqrt{n}$.
2. If B_k and C_k are bounded in k, then the penalty contributes $\delta_{n,k}^2$ and x_k/n to the upper bound. The first is the inverse rate of convergence for just the model Θ_k, the second is the penalty for using multiple models.

$$J(k) \gtrsim \delta_{n,k}^2 C_k D_k^2 + \frac{1}{n}(B_k + C_k)(x_k + \log(B_k + C_k))$$

Model selection: General Result

1. The risk $E^* P(m_{\hat{\theta}} - m_{\theta^*})$ should be interpreted as a square distance, so the rate is at most $1/\sqrt{n}$.
2. If B_k and C_k are bounded in k, then the penalty contributes $\delta_{n,k}^2$ and x_k/n to the upper bound. The first is the inverse rate of convergence for just the model Θ_k, the second is the penalty for using multiple models.
3. If $\mathcal{K} = \mathbb{N}$ and $x_k = k$, this can correspond to a model class with a single model for each dimension $k \in \mathbb{N}$.

Model selection: General Result

1. The risk $E^* P(m_{\hat{\theta}} - m_{\theta^*})$ should be interpreted as a square distance, so the rate is at most $1/\sqrt{n}$.

2. If B_k and C_k are bounded in k, then the penalty contributes $\delta_{n,k}^2$ and x_k/n to the upper bound. The first is the inverse rate of convergence for just the model Θ_k, the second is the penalty for using multiple models.

3. If $\mathcal{K} = \mathbb{N}$ and $x_k = k$, this can correspond to a model class with a single model for each dimension $k \in \mathbb{N}$.
4. If $|\mathcal{K}| \leq n^r$, then we may choose $x_k = r \log n$ for all k. Then the loss is of order $\log n/n$.

Model selection: General Result

1. The risk $E^*P(m_{\hat{\theta}} - m_{\theta^*})$ should be interpreted as a square distance, so the rate is at most $1/\sqrt{n}$.
2. If B_k and C_k are bounded in k, then the penalty contributes $\delta_{n,k}^2$ and x_k/n to the upper bound. The first is the inverse rate of convergence for just the model Θ_k, the second is the penalty for using multiple models.
3. If $\mathcal{K} = \mathbb{N}$ and $x_k = k$, this can correspond to a model class with a single model for each dimension $k \in \mathbb{N}$.
4. If $|\mathcal{K}| \leq n^r$, then we may choose $x_k = r \log n$ for all k. Then the loss is of order $\log n/n$.
5. $\delta_{n,k}$, C_k and D_k in the penalty refer to the true distribution P, so we need to construct penalty terms applicable to a broad class of true distributions.

Model selection: General Result

Lemma (3.5.9)

Let $\{m_\theta : \theta \in \Theta\}$ *be a class of measurable functions* $m_\theta : \mathcal{X} \to [-B, B] \subset \mathbb{R}$ *indexed by a metric space* (Θ, d) *such that, for some function* $\phi : (0, \infty) \to \mathbb{R}$ *such that* $\delta \mapsto \phi(\delta)/\delta$ *is nonincreasing, and every* $\delta \geq \underline{\delta}_n$*,*

$$P(m_\theta - m_{\theta_0})^2 \leq d^2(\theta, \theta_0),$$

$$E^* \sup_{\theta : d(\theta, \theta_0) < \delta} (\mathbb{G}_n(m_\theta - m_{\theta_0}))_+ \leq \phi(\delta).$$

Model selection: General Result

Lemma (3.5.9)

Let $\{m_\theta : \theta \in \Theta\}$ *be a class of measurable functions* $m_\theta : \mathcal{X} \to [-B, B] \subset \mathbb{R}$ *indexed by a metric space* (Θ, d) *such that, for some function* $\phi : (0, \infty) \to \mathbb{R}$ *such that* $\delta \mapsto \phi(\delta)/\delta$ *is nonincreasing, and every* $\delta \geq \underline{\delta}_n$,

$$P(m_\theta - m_{\theta_0})^2 \leq d^2(\theta, \theta_0),$$
$$E^* \sup_{\theta : d(\theta,\theta_0) < \delta} (\mathbb{G}_n(m_\theta - m_{\theta_0}))_+ \leq \phi(\delta).$$

Then for any $\delta_n \geq \underline{\delta}_n$ *with* $\phi(\delta_n) \leq D^2 \sqrt{n} \delta_n^2$, *every* $0 < \eta < D$ *and every* $x > 0$, *with probability at least* $1 - e^{-x}$,

$$\forall \theta \in \Theta : \frac{1}{135\sqrt{n}} \mathbb{G}_n(m_\theta - m_{\theta_0}) \leq \eta d^2(\theta, \theta_0) + \frac{\delta_n^2 D}{\eta} + \left(B + \frac{1}{\eta}\right) \frac{x}{n}.$$

Model selection: General Result

Proof Lemma 3.5.9 1/4.

Assume $m_{\theta_0} \equiv 0$ by increasing the constant B to $2B$.

Model selection: General Result

Proof Lemma 3.5.9 1/4.

Assume $m_{\theta_0} \equiv 0$ by increasing the constant B to $2B$. For $c \geq 1$ we know that $\phi(c\delta)/(c\delta) \leq \phi(\delta)/\delta$, so $\phi(c\delta) \leq c\phi(\delta)$.

Proof Lemma 3.5.9 1/4.

Assume $m_{\theta_0} \equiv 0$ by increasing the constant B to $2B$. For $c \geq 1$ we know that $\phi(c\delta)/(c\delta) \leq \phi(\delta)/\delta$, so $\phi(c\delta) \leq c\phi(\delta)$. Then for $\delta \geq \underline{\delta}_n$:

$$E^* \sup_{d(\theta,\theta_0)\geq\delta} \frac{(\mathbb{G}_n m_\theta)_+}{d^2(\theta,\theta_0)+\delta^2}$$

Model selection: General Result

Proof Lemma 3.5.9 1/4.

Assume $m_{\theta_0} \equiv 0$ by increasing the constant B to $2B$. For $c \geq 1$ we know that $\phi(c\delta)/(c\delta) \leq \phi(\delta)/\delta$, so $\phi(c\delta) \leq c\phi(\delta)$. Then for $\delta \geq \underline{\delta}_n$:

$$E^* \sup_{d(\theta,\theta_0)\geq\delta} \frac{(\mathbb{G}_n m_\theta)_+}{d^2(\theta,\theta_0)+\delta^2} \leq \sum_{j=1}^{\infty} E^* \sup_{2^{j-1}\delta\leq d(\theta,\theta_0)<2^j\delta} \frac{(\mathbb{G}_n m_\theta)_+}{d^2(\theta,\theta_0)+\delta^2}$$

Proof Lemma 3.5.9 1/4.

Assume $m_{\theta_0} \equiv 0$ by increasing the constant B to $2B$. For $c \geq 1$ we know that $\phi(c\delta)/(c\delta) \leq \phi(\delta)/\delta$, so $\phi(c\delta) \leq c\phi(\delta)$. Then for $\delta \geq \underline{\delta}_n$:

$$\begin{aligned} E^* \sup_{d(\theta,\theta_0)\geq\delta} \frac{(\mathbb{G}_n m_\theta)_+}{d^2(\theta,\theta_0)+\delta^2} &\leq \sum_{j=1}^{\infty} E^* \sup_{2^{j-1}\delta\leq d(\theta,\theta_0)<2^j\delta} \frac{(\mathbb{G}_n m_\theta)_+}{d^2(\theta,\theta_0)+\delta^2} \\ &\leq \sum_{j=1}^{\infty} \frac{\phi(2^j\delta)}{2^{2j-2}\delta^2+\delta^2} \end{aligned}$$

Proof Lemma 3.5.9 1/4.

Assume $m_{\theta_0} \equiv 0$ by increasing the constant B to $2B$. For $c \geq 1$ we know that $\phi(c\delta)/(c\delta) \leq \phi(\delta)/\delta$, so $\phi(c\delta) \leq c\phi(\delta)$. Then for $\delta \geq \underline{\delta}_n$:

$$
\begin{aligned}
E^* \sup_{d(\theta,\theta_0)\geq\delta} \frac{(\mathbb{G}_n m_\theta)_+}{d^2(\theta,\theta_0)+\delta^2} &\leq \sum_{j=1}^{\infty} E^* \sup_{2^{j-1}\delta\leq d(\theta,\theta_0)<2^j\delta} \frac{(\mathbb{G}_n m_\theta)_+}{d^2(\theta,\theta_0)+\delta^2} \\
&\leq \sum_{j=1}^{\infty} \frac{\phi(2^j\delta)}{2^{2j-2}\delta^2+\delta^2} \\
&\leq \sum_{j=1}^{\infty} \frac{2^j}{2^{2j-2}} \frac{\phi(\delta)}{\delta^2}
\end{aligned}
$$

Model selection: General Result

Proof Lemma 3.5.9 1/4.

Assume $m_{\theta_0} \equiv 0$ by increasing the constant B to $2B$. For $c \geq 1$ we know that $\phi(c\delta)/(c\delta) \leq \phi(\delta)/\delta$, so $\phi(c\delta) \leq c\phi(\delta)$. Then for $\delta \geq \underline{\delta}_n$:

$$
\begin{aligned}
E^* \sup_{d(\theta,\theta_0)\geq\delta} \frac{(\mathbb{G}_n m_\theta)_+}{d^2(\theta,\theta_0)+\delta^2} &\leq \sum_{j=1}^{\infty} E^* \sup_{2^{j-1}\delta \leq d(\theta,\theta_0) < 2^j\delta} \frac{(\mathbb{G}_n m_\theta)_+}{d^2(\theta,\theta_0)+\delta^2} \\
&\leq \sum_{j=1}^{\infty} \frac{\phi(2^j\delta)}{2^{2j-2}\delta^2+\delta^2} \\
&\leq \sum_{j=1}^{\infty} \frac{2^j}{2^{2j-2}} \frac{\phi(\delta)}{\delta^2} \\
&= 4\frac{\phi(\delta)}{\delta^2}.
\end{aligned}
$$

Model selection: General Result

Proof Lemma 3.5.9 2/4.

We have

$$E^* \sup_{d(\theta,\theta_0)<\delta} \frac{(\mathbb{G}_n m_\theta)_+}{d^2(\theta,\theta_0)+\delta^2}$$

Proof Lemma 3.5.9 2/4.

We have

$$E^* \sup_{d(\theta,\theta_0)<\delta} \frac{(\mathbb{G}_n m_\theta)_+}{d^2(\theta,\theta_0)+\delta^2} \leq E^* \sup_{d(\theta,\theta_0)<\delta} \frac{(\mathbb{G}_n m_\theta)_+}{\delta^2}$$

Model selection: General Result

Proof Lemma 3.5.9 2/4.

We have

$$E^* \sup_{d(\theta,\theta_0)<\delta} \frac{(\mathbb{G}_n m_\theta)_+}{d^2(\theta,\theta_0)+\delta^2} \leq E^* \sup_{d(\theta,\theta_0)<\delta} \frac{(\mathbb{G}_n m_\theta)_+}{\delta^2}$$

$$\leq \phi(\delta)/\delta^2.$$

Model selection: General Result

Proof Lemma 3.5.9 2/4.

We have

$$E^* \sup_{d(\theta,\theta_0)<\delta} \frac{(\mathbb{G}_n m_\theta)_+}{d^2(\theta,\theta_0)+\delta^2} \leq E^* \sup_{d(\theta,\theta_0)<\delta} \frac{(\mathbb{G}_n m_\theta)_+}{\delta^2}$$
$$\leq \phi(\delta)/\delta^2.$$

Thus

$$E^* \sup_{\theta\in\Theta} \frac{(\mathbb{G}_n m_\theta)_+}{d^2(\theta,\theta_0)+\delta^2} \leq 5\phi(\delta)/\delta^2.$$

Model selection: General Result

Proof Lemma 3.5.9 3/4.

Note that $m_\theta/(d^2(\theta,\theta_0)+\delta^2) \leq 2B/\delta^2$

Model selection: General Result

Proof Lemma 3.5.9 3/4.

Note that $m_\theta / (d^2(\theta, \theta_0) + \delta^2) \leq 2B/\delta^2$ and

$$P(m_\theta / (d^2(\theta, \theta_0) + \delta^2))^2 \leq \frac{d^2(\theta, \theta_0)}{d^4(\theta, \theta_0) + 2d^2(\theta, \theta_0)\delta^2 + \delta^4} \leq \frac{1}{\delta^2}.$$

Model selection: General Result

Proof Lemma 3.5.9 3/4.

Note that $m_\theta / (d^2(\theta, \theta_0) + \delta^2) \leq 2B/\delta^2$ and

$$P(m_\theta / (d^2(\theta, \theta_0) + \delta^2))^2 \leq \frac{d^2(\theta, \theta_0)}{d^4(\theta, \theta_0) + 2d^2(\theta, \theta_0)\delta^2 + \delta^4} \leq \frac{1}{\delta^2}.$$

Lemma (2.15.9)

Let $\mathcal{F}$ be a countable class of measurable functions $f : \mathcal{X} \to \mathbb{R}$ such that $\|f\|_\infty \leq M$ and $Pf^2 \leq \frac{1}{\delta^2}$ for every $f \in \mathcal{F}$. Then, for every $x > 0$, with probability at least $1 - e^{-x}$,

$$\forall f \in \mathcal{F} : \frac{1}{2}\mathbb{G}_n f \leq E \sup_{f \in \mathcal{F}} \mathbb{G}_n f + \frac{Mx}{\sqrt{n}} + \frac{1}{\delta}\sqrt{x}.$$

Model selection: General Result

Proof Lemma 3.5.9 3/4.

Note that $m_\theta/(d^2(\theta,\theta_0)+\delta^2) \le 2B/\delta^2$ and

$$P(m_\theta/(d^2(\theta,\theta_0)+\delta^2))^2 \le \frac{d^2(\theta,\theta_0)}{d^4(\theta,\theta_0)+2d^2(\theta,\theta_0)\delta^2+\delta^4} \le \frac{1}{\delta^2}.$$

Lemma (2.15.9)

Let $\mathcal{F}$ be a countable class of measurable functions $f:\mathcal{X}\to\mathbb{R}$ such that $\|f\|_\infty \le M$ and $Pf^2 \le \frac{1}{\delta^2}$ for every $f\in\mathcal{F}$. Then, for every $x>0$, with probability at least $1-e^{-x}$,

$$\forall f\in\mathcal{F}: \frac{1}{2}\mathbb{G}_n f \le E\sup_{f\in\mathcal{F}}\mathbb{G}_n f + \frac{Mx}{\sqrt{n}} + \frac{1}{\delta}\sqrt{x}.$$

There is a separability argument missing in the book.

Model selection: General Result

Proof Lemma 3.5.9 4/4.

Thus with probability at least $1 - e^{-x}$,

$$\forall \theta \in \Theta : \mathbb{G}_n m_\theta \leq 15 \left(\frac{\phi(\delta)}{\delta^2} + \frac{Bx}{\delta^2 \sqrt{n}} + \frac{\sqrt{x}}{\delta} \right) \left(d^2(\theta, \theta_0) + \delta^2 \right).$$

Proof Lemma 3.5.9 4/4.

Thus with probability at least $1 - e^{-x}$,

$$\forall \theta \in \Theta : \mathbb{G}_n m_\theta \leq 15 \left(\frac{\phi(\delta)}{\delta^2} + \frac{Bx}{\delta^2 \sqrt{n}} + \frac{\sqrt{x}}{\delta} \right) (d^2(\theta, \theta_0) + \delta^2).$$

For $\delta > \delta_n D/\eta > \delta_n$, we have

$$\phi(\delta)/\delta^2 \leq \phi(\delta_n D/\eta)/(\delta_n D/\eta)^2$$

□

Model selection: General Result

Proof Lemma 3.5.9 4/4.

Thus with probability at least $1 - e^{-x}$,

$$\forall \theta \in \Theta : \mathbb{G}_n m_\theta \leq 15 \left(\frac{\phi(\delta)}{\delta^2} + \frac{Bx}{\delta^2 \sqrt{n}} + \frac{\sqrt{x}}{\delta} \right) (d^2(\theta, \theta_0) + \delta^2).$$

For $\delta > \delta_n D/\eta > \delta_n$, we have

$$\begin{aligned} \phi(\delta)/\delta^2 &\leq \phi(\delta_n D/\eta)/(\delta_n D/\eta)^2 \\ &\leq \eta\sqrt{n} \end{aligned}$$

□

Model selection: General Result

Proof Lemma 3.5.9 4/4.

Thus with probability at least $1 - e^{-x}$,

$$\forall \theta \in \Theta : \mathbb{G}_n m_\theta \leq 15 \left(\frac{\phi(\delta)}{\delta^2} + \frac{Bx}{\delta^2 \sqrt{n}} + \frac{\sqrt{x}}{\delta} \right) (d^2(\theta, \theta_0) + \delta^2).$$

For $\delta > \delta_n D/\eta > \delta_n$, we have

$$\begin{aligned} \phi(\delta)/\delta^2 &\leq \phi(\delta_n D/\eta)/(\delta_n D/\eta)^2 \\ &\leq \eta\sqrt{n} \end{aligned}$$

Now choose $\delta = \delta_n D/\eta + \sqrt{Bx/(n\eta)} + \sqrt{x/n}/\eta$ so the three terms are smaller than $\sqrt{n}\eta$. $\square$

Proof Theorem 3.5.1 1/8.

By definition of $\hat{\theta}$ and $\hat{k}$ we have $\mathbb{P}_n m_{\hat{\theta}} + J(\hat{k}) \leq \mathbb{P}_n m_\theta + J(k)$ for all $\theta \in \Theta_k$.

Model selection: General Result

Proof Theorem 3.5.1 1/8.

By definition of $\hat{\theta}$ and $\hat{k}$ we have $\mathbb{P}_n m_{\hat{\theta}} + J(\hat{k}) \leq \mathbb{P}_n m_\theta + J(k)$ for all $\theta \in \Theta_k$.So for $\theta = \theta_k^*$ we have

$$\begin{aligned} P(m_{\hat{\theta}} - m_{\theta^*}) \leq P(m_{\theta_k^*} - m_{\theta^*}) + J(k) - J(\hat{k}) \\ + \frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\theta_k^*} - m_{\bar{\theta}_k}) + \frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\bar{\theta}_k} - m_{\hat{\theta}}). \end{aligned}$$

Proof Theorem 3.5.1 1/8.

By definition of $\hat{\theta}$ and $\hat{k}$ we have $\mathbb{P}_n m_{\hat{\theta}} + J(\hat{k}) \leq \mathbb{P}_n m_\theta + J(k)$ for all $\theta \in \Theta_k$.So for $\theta = \theta_k^*$ we have

$$P(m_{\hat{\theta}} - m_{\theta^*}) \leq P(m_{\theta_k^*} - m_{\theta^*}) + J(k) - J(\hat{k}) + \frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\theta_k^*} - m_{\bar{\theta}_k}) + \frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\bar{\theta}_k} - m_{\hat{\theta}}).$$

We bound the first empirical process. Since $d(\bar{\theta}_k, \theta^*) \leq d(\theta, \theta^*)$, for any $k, k' \in \mathcal{K}$ we have

$$P(m_{\theta_k^*} - m_{\bar{\theta}_{k'}})^2 \leq d^2(\theta_k^*, \bar{\theta}_{k'}) \lesssim d^2(\theta_k^*, \theta^*) + d^2(\hat{\theta}_{k'}, \theta^*).$$

Model selection: General Result

Proof Theorem 3.5.1 2/8.

By Lemma 2.15.9 applied to the class $\{m_{\theta_k^*} - m_{\bar{\theta}_{k'}}\}$ we have with probability at least $1 - e^{-x_k - x_{k'} - \xi}$,

$$\frac{1}{\sqrt{n}} \mathbb{G}_n(m_{\theta_k^*} - m_{\bar{\theta}_{k'}}) \lesssim (B_k \vee B_{k'}) \frac{x_k + x_{k'} + \xi}{n} + (d(\theta_k^*, \theta^*) \vee d(\hat{\theta}_{k'}, \theta^*)) \sqrt{\frac{x_k + x_{k'} + \xi}{n}}.$$

Model selection: General Result

Proof Theorem 3.5.1 2/8.

By Lemma 2.15.9 applied to the class $\{m_{\theta_k^*} - m_{\bar\theta_{k'}}\}$ we have with probability at least $1 - e^{-x_k - x_{k'} - \xi}$,

$$\frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\theta_k^*} - m_{\bar\theta_{k'}}) \lesssim (B_k \vee B_{k'})\frac{x_k + x_{k'} + \xi}{n} + (d(\theta_k^*, \theta^*) \vee d(\hat\theta_{k'}, \theta^*))\sqrt{\frac{x_k + x_{k'} + \xi}{n}}.$$

By the inequality $2\sqrt{ab} \leq a^2/c + cb^2$ we bound the second term by

Model selection: General Result

Proof Theorem 3.5.1 2/8.

By Lemma 2.15.9 applied to the class $\{m_{\theta_k^*} - m_{\bar{\theta}_{k'}}\}$ we have with probability at least $1 - e^{-x_k - x_{k'} - \xi}$,

$$\frac{1}{\sqrt{n}} \mathbb{G}_n(m_{\theta_k^*} - m_{\bar{\theta}_{k'}}) \lesssim (B_k \vee B_{k'}) \frac{x_k + x_{k'} + \xi}{n} + (d(\theta_k^*, \theta^*) \vee d(\hat{\theta}_{k'}, \theta^*)) \sqrt{\frac{x_k + x_{k'} + \xi}{n}}.$$

By the inequality $\cancel{2\sqrt{ab} \leq a^2/c + cb^2}$ $ab \leq a^2/c + cb^2$ we bound the second term by

Model selection: General Result

Proof Theorem 3.5.1 2/8.

By Lemma 2.15.9 applied to the class $\{m_{\theta_k^*} - m_{\bar{\theta}_{k'}}\}$ we have with probability at least $1 - e^{-x_k - x_{k'} - \xi}$,

$$\frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\theta_k^*} - m_{\bar{\theta}_{k'}}) \lesssim (B_k \vee B_{k'})\frac{x_k + x_{k'} + \xi}{n} + (d(\theta_k^*, \theta^*) \vee d(\hat{\theta}_{k'}, \theta^*))\sqrt{\frac{x_k + x_{k'} + \xi}{n}}.$$

By the inequality $\cancel{2\sqrt{ab} \leq a^2/c + cb^2}$ $ab \leq a^2/c + cb^2$ we bound the second term by

$$\eta\frac{d^2(\theta_k^*, \theta^*) \vee d^2(\hat{\theta}_{k'}, \theta^*)}{C_k \vee C_{k'}} + \frac{1}{\eta}(C_k \vee C_{k'})\frac{x_k + x_{k'} + \xi}{n}.$$

with $c = \frac{\eta}{C_k \vee C_{k'}}$.

Proof Theorem 3.5.1 3/8.

Comonotivity of (x_k, B_k, C_k) show that

$$\frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\theta_k^*} - m_{\bar{\theta}_{k'}}) \lesssim \sum_{\kappa=k,k'} \left(B_\kappa + \frac{C_\kappa}{\eta}\right) \frac{x_\kappa + \xi}{n} + \frac{\eta d^2(\theta_k^*, \theta^*)}{C_k} + \frac{\eta d^2(\hat{\theta}_{k'}, \theta^*)}{C_{k'}}.$$

Model selection: General Result

Proof Theorem 3.5.1 3/8.

Comonotivity of (x_k, B_k, C_k) show that

$$\frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\theta_k^*} - m_{\bar{\theta}_{k'}}) \lesssim \sum_{\kappa=k,k'} \left(B_\kappa + \frac{C_\kappa}{\eta}\right)\frac{x_\kappa + \xi}{n} + \frac{\eta d^2(\theta_k^*, \theta^*)}{C_k} + \frac{\eta d^2(\hat{\theta}_{k'}, \theta^*)}{C_{k'}}.$$

This is false for k, k' with probability at most $e^{-x_k - x_{k'} - \xi}$, hence this does not hold for at least one pair k, k' with probability at most

$$\sum_k \sum_{k'} e^{-x_k - x_{k'} - \xi} = e^{-\xi}\left(\sum_k e^{-x_k}\right)\left(\sum_{k'} e^{-x_{k'}}\right) \leq e^{-\xi}.$$

Proof Theorem 3.5.1 3/8.

Comonotivity of (x_k, B_k, C_k) show that

$$\frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\theta_k^*} - m_{\bar{\theta}_{k'}}) \lesssim \sum_{\kappa=k,k'}\left(B_\kappa + \frac{C_\kappa}{\eta}\right)\frac{x_\kappa + \xi}{n} + \frac{\eta d^2(\theta_k^*, \theta^*)}{C_k} + \frac{\eta d^2(\hat{\theta}_{k'}, \theta^*)}{C_{k'}}.$$

This is false for k, k' with probability at most $e^{-x_k - x_{k'} - \xi}$, hence this does not hold for at least one pair k, k' with probability at most

$$\sum_k \sum_{k'} e^{-x_k - x_{k'} - \xi} = e^{-\xi}\left(\sum_k e^{-x_k}\right)\left(\sum_{k'} e^{-x_{k'}}\right) \leq e^{-\xi}.$$

We take $k' = \hat{k}$.

Model selection: General Result

Proof Theorem 3.5.1 4/8.

We now consider bounding $\frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\bar{\theta}_{\hat{k}}} - m_{\hat{\theta}})$. We use Lemma 3.5.9 to the functions $-m_\theta$.

Model selection: General Result

Proof Theorem 3.5.1 4/8.

We now consider bounding $\frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\bar{\theta}_{\hat{k}}} - m_{\hat{\theta}})$. We use Lemma 3.5.9 to the functions $-m_\theta$.

Lemma (Lemma 3.5.9)

Let $\{m_\theta : \theta \in \Theta\}$ satisfies certain conditions, then for any $\delta_n \geq \underline{\delta}_n$ with $\phi(\delta_n) \leq D^2\sqrt{n}\delta_n^2$, every $0 < \eta < D$ and every $x > 0$, with probability at least $1 - e^{-x}$,

$$\forall\theta \in \Theta : \frac{1}{135\sqrt{n}}\mathbb{G}_n(m_\theta - m_{\theta_0}) \leq \eta d^2(\theta, \theta_0) + \frac{\delta_n^2 D}{\eta} + \left(B + \frac{1}{\eta}\right)\frac{x}{n}.$$

Model selection: General Result

Proof Theorem 3.5.1 4/8.

We now consider bounding $\frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\bar\theta_{\hat k}} - m_{\hat\theta})$. We use Lemma 3.5.9 to the functions $-m_\theta$.

Lemma (Lemma 3.5.9)

Let $\{m_\theta : \theta \in \Theta\}$ satisfies certain conditions, then for any $\delta_n \geq \underline{\delta}_n$ with $\phi(\delta_n) \leq D^2\sqrt{n}\delta_n^2$, every $0 < \eta < D$ and every $x > 0$, with probability at least $1 - e^{-x}$,

$$\forall \theta \in \Theta : \frac{1}{135\sqrt{n}}\mathbb{G}_n(m_\theta - m_{\theta_0}) \leq \eta d^2(\theta, \theta_0) + \frac{\delta_n^2 D}{\eta} + \left(B + \frac{1}{\eta}\right)\frac{x}{n}.$$

Thus for all $\theta \in \Theta_k$, with probability at least $1 - e^{-x_k - \xi}$,

$$\frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\bar\theta_k} - m_\theta) \lesssim \frac{\eta}{C_k} d^2(\theta, \bar\theta_k) + \frac{\delta_{n,k}^2 C_k D_k^2}{\eta} + \left(B_k + \frac{C_k}{\eta}\right)\frac{x_k + \xi}{n}.$$

Model selection: General Result

Proof Theorem 3.5.1 5/8.

$$\frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\bar{\theta}_k} - m_\theta) \lesssim \frac{\eta}{C_k} d^2(\theta, \bar{\theta}_k) + \frac{\delta_{n,k}^2 C_k D_k^2}{\eta} + \left(B_k + \frac{C_k}{\eta}\right)\frac{x_k + \xi}{n}.$$

Model selection: General Result

Proof Theorem 3.5.1 5/8.

$$\frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\bar{\theta}_k} - m_\theta) \lesssim \frac{\eta}{C_k} d^2(\theta, \bar{\theta}_k) + \frac{\delta_{n,k}^2 C_k D_k^2}{\eta} + \left(B_k + \frac{C_k}{\eta}\right)\frac{x_k + \xi}{n}.$$

We choose $\theta = \hat{\theta}_k$. Note that $d(\hat{\theta}_k, \bar{\theta}_k) \leq 2d(\hat{\theta}_k, \theta^*)$.

Model selection: General Result

Proof Theorem 3.5.1 5/8.

$$\frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\bar{\theta}_k} - m_\theta) \lesssim \frac{\eta}{C_k} d^2(\theta, \bar{\theta}_k) + \frac{\delta_{n,k}^2 C_k D_k^2}{\eta} + \left(B_k + \frac{C_k}{\eta}\right)\frac{x_k + \xi}{n}.$$

We choose $\theta = \hat{\theta}_k$. Note that $d(\hat{\theta}_k, \bar{\theta}_k) \leq 2d(\hat{\theta}_k, \theta^*)$. This is true simultaneously for all $k \in \mathcal{K}$ with probability at least $1 - e^{-\xi}$, thus we choose $k = \hat{k}$ to get

$$\frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\bar{\theta}_{\hat{k}}} - m_\theta) \lesssim \frac{\eta}{C_{\hat{k}}} d^2(\theta, \bar{\theta}_{\hat{k}}) + \frac{\delta_{n,\hat{k}}^2 C_{\hat{k}} D_{\hat{k}}^2}{\eta} + \left(B_{\hat{k}} + \frac{C_{\hat{k}}}{\eta}\right)\frac{x_{\hat{k}} + \xi}{n}$$

with probability at least $1 - e^{-\xi}$.

Proof Theorem 3.5.1 6/8.

Combining everything, for some $c > 0$, with probability at least $1 - 2e^{\xi}$, we have

$$\begin{aligned}
& cP(m_{\hat{\theta}} - m_{\theta^*}) \\
& \leq cP(m_{\theta_k^*} - m_{\theta^*}) + cJ(k) - cJ(\hat{k}) \\
& + \sum_{\kappa = k, k'} \left(B_\kappa + \frac{C_\kappa}{\eta} \right) \frac{x_\kappa + \xi}{n} + \frac{\eta d^2(\theta_k^*, \theta^*)}{C_k} + \frac{\eta d^2(\hat{\theta}, \theta^*)}{C_{\hat{k}}} \\
& + \frac{\delta_{n,\hat{k}}^2 C_{\hat{k}} D_{\hat{k}}}{\eta}.
\end{aligned}$$

Model selection: General Result

Proof Theorem 3.5.1 6/8.

Combining everything, for some $c > 0$, with probability at least $1 - 2e^{\xi}$, we have

$$\begin{aligned}
&cP(m_{\hat{\theta}} - m_{\theta^*}) \\
&\leq cP(m_{\theta_k^*} - m_{\theta^*}) + cJ(k) - cJ(\hat{k}) \\
&+ \sum_{\kappa = k, k'} \left(B_\kappa + \frac{C_\kappa}{\eta} \right) \frac{x_\kappa + \xi}{n} + \frac{\eta d^2(\theta_k^*, \theta^*)}{C_k} + \frac{\eta d^2(\hat{\theta}, \theta^*)}{C_{\hat{k}}} \\
&+ \frac{\delta_{n,\hat{k}}^2 C_{\hat{k}} D_{\hat{k}}}{\eta}.
\end{aligned}$$

By assumption, $d(\theta, \theta^*) \leq C_k P(m_\theta - m_{\theta^*})$.

Model selection: General Result

Proof Theorem 3.5.1 6/8.

Combining everything, for some $c > 0$, with probability at least $1 - 2e^{\xi}$, we have

$$
\begin{aligned}
&cP(m_{\hat{\theta}} - m_{\theta^*}) \\
&\leq cP(m_{\theta_k^*} - m_{\theta^*}) + cJ(k) - cJ(\hat{k}) \\
&+ \sum_{\kappa = k, k'} \left(B_\kappa + \frac{C_\kappa}{\eta} \right) \frac{x_\kappa + \xi}{n} + \frac{\eta d^2(\theta_k^*, \theta^*)}{C_k} + \frac{\eta d^2(\hat{\theta}, \theta^*)}{C_{\hat{k}}} \\
&+ \frac{\delta_{n,\hat{k}}^2 C_{\hat{k}} D_{\hat{k}}}{\eta}.
\end{aligned}
$$

By assumption, $d(\theta, \theta^*) \leq C_k P(m_\theta - m_{\theta^*})$. Young's inequality says $B\xi \leq 2B \log B + e^{\xi/2}$.

Model selection: General Result

Proof Theorem 3.5.1 7/8.

We get

$$
\begin{aligned}
&(c-\eta)P(m_{\hat{\theta}} - m_{\theta^*}) \\
&\leq (c+\eta)P(m_{\theta_k^*} - m_{\theta^*}) + cJ(k) - cJ(\hat{k}) + \frac{\delta_{n,\hat{k}}^2 C_{\hat{k}} D_{\hat{k}}^2}{\eta} \\
&+ \sum_{\kappa=k,k'} \frac{1}{n}\left(B_\kappa + \frac{C_\kappa}{\eta}\right)\left[x_\kappa + 2\log\left(B_\kappa + \frac{C_\kappa}{\eta}\right)\right] + 2\frac{e^{\xi/2}}{n}.
\end{aligned}
$$

Model selection: General Result

Proof Theorem 3.5.1 7/8.

We get

$$
\begin{aligned}
&(c-\eta)P(m_{\hat{\theta}} - m_{\theta^*}) \\
&\leq (c+\eta)P(m_{\theta_k^*} - m_{\theta^*}) + cJ(k) - cJ(\hat{k}) + \frac{\delta_{n,\hat{k}}^2 C_{\hat{k}} D_{\hat{k}}^2}{\eta} \\
&+ \sum_{\kappa=k,k'} \frac{1}{n}\left(B_\kappa + \frac{C_\kappa}{\eta}\right)\left[x_\kappa + 2\log\left(B_\kappa + \frac{C_\kappa}{\eta}\right)\right] + 2\frac{e^{\xi/2}}{n}.
\end{aligned}
$$

Thus for a fixed $\eta < c$ we see that for all $k \in \mathcal{K}$:

$$
P(m_{\hat{\theta}} - m_{\theta^*}) \leq P(m_{\theta_k^*} - m_{\theta^*}) + J(k) + \frac{2e^{\xi/2}}{n},
$$

Proof Theorem 3.5.1 8/8.

We now write, with $a := \inf_{k\in\mathcal{K}} \left(P(m_{\theta_k^*} - m_{\theta^*}) + J(k) \right)$,

Model selection: General Result

Proof Theorem 3.5.1 8/8.

We now write, with $a := \inf_{k \in \mathcal{K}} \Big(P(m_{\theta_k^*} - m_{\theta^*}) + J(k) \Big)$,

$$\mathbb{E}^* P(m_{\hat{\theta}} - m_{\theta^*})$$
$$= \int_0^\infty P^*(P(m_{\hat{\theta}} - m_{\theta^*}) > t)dt$$

Model selection: General Result

Proof Theorem 3.5.1 8/8.

We now write, with $a := \inf_{k \in \mathcal{K}} \left(P(m_{\theta_k^*} - m_{\theta^*}) + J(k) \right)$,

$$\begin{aligned}
&\mathbb{E}^* P(m_{\hat{\theta}} - m_{\theta^*}) \\
&= \int_0^\infty P^*(P(m_{\hat{\theta}} - m_{\theta^*}) > t) dt \\
&\leq \int_0^a P^*(P(m_{\hat{\theta}} - m_{\theta^*}) > t) dt + \int_0^\infty P^*(P(m_{\hat{\theta}} - m_{\theta^*}) > a + t) dt
\end{aligned}$$

Proof Theorem 3.5.1 8/8.

We now write, with $a := \inf_{k \in \mathcal{K}} \Big(P(m_{\theta_k^*} - m_{\theta^*}) + J(k) \Big)$,

$$\mathbb{E}^* P(m_{\hat\theta} - m_{\theta^*})$$

$$= \int_0^\infty P^*(P(m_{\hat\theta} - m_{\theta^*}) > t) dt$$

$$\leq \int_0^a P^*(P(m_{\hat\theta} - m_{\theta^*}) > t) dt + \int_0^\infty P^*(P(m_{\hat\theta} - m_{\theta^*}) > a + t) dt$$

$$\leq a + \int_{-\infty}^\infty P^*(P(m_{\hat\theta} - m_{\theta^*}) > a + \frac{1}{n} e^{\xi/2}) \frac{1}{2n} e^{\xi/2} d\xi$$

Proof Theorem 3.5.1 8/8.

We now write, with $a := \inf_{k\in\mathcal{K}}\left(P(m_{\theta_k^*} - m_{\theta^*}) + J(k)\right)$,

$$
\begin{aligned}
&\mathbb{E}^* P(m_{\hat\theta} - m_{\theta^*}) \\
&= \int_0^\infty P^*(P(m_{\hat\theta} - m_{\theta^*}) > t)dt \\
&\leq \int_0^a P^*(P(m_{\hat\theta} - m_{\theta^*}) > t)dt + \int_0^\infty P^*(P(m_{\hat\theta} - m_{\theta^*}) > a + t)dt \\
&\leq a + \int_{-\infty}^\infty P^*(P(m_{\hat\theta} - m_{\theta^*}) > a + \frac{1}{n}e^{\xi/2})\frac{1}{2n}e^{\xi/2}d\xi \\
&\leq a + \frac{1}{2n}\int_{-\infty}^0 e^{\xi/2}dt + \frac{1}{2n}\int_0^\infty e^{-\xi}e^{\xi/2}d\xi
\end{aligned}
$$

Proof Theorem 3.5.1 8/8.

We now write, with $a := \inf_{k\in\mathcal{K}} \Big(P(m_{\theta_k^*} - m_{\theta^*}) + J(k) \Big)$,

$$\begin{aligned}
&\mathbb{E}^* P(m_{\hat\theta} - m_{\theta^*}) \\
&= \int_0^\infty P^*(P(m_{\hat\theta} - m_{\theta^*}) > t)dt \\
&\leq \int_0^a P^*(P(m_{\hat\theta} - m_{\theta^*}) > t)dt + \int_0^\infty P^*(P(m_{\hat\theta} - m_{\theta^*}) > a + t)dt \\
&\leq a + \int_{-\infty}^\infty P^*(P(m_{\hat\theta} - m_{\theta^*}) > a + \frac{1}{n}e^{\xi/2})\frac{1}{2n}e^{\xi/2}d\xi \\
&\leq a + \frac{1}{2n}\int_{-\infty}^0 e^{\xi/2}dt + \frac{1}{2n}\int_0^\infty e^{-\xi}e^{\xi/2}d\xi \\
&\lesssim \inf_{k\in\mathcal{K}} \Big(P(m_{\theta_k^*} - m_{\theta^*}) + J(k) \Big) + \frac{1}{n}.
\end{aligned}$$

Theorem (3.5.11)

Let $\{m_\theta : \theta \in \Theta\}$ *be a class of measurable functions* $m_\theta : \mathcal{X} \to \mathbb{R}$ *indexed by a metric space* (Θ, d) *such that* $\|m_\theta\|_\infty \leq B$ *for every* $\theta \in \Theta_k$.

Model selection: Statistical Learning

Theorem (3.5.11)

Let $\{m_\theta : \theta \in \Theta\}$ *be a class of measurable functions* $m_\theta : \mathcal{X} \to \mathbb{R}$ *indexed by a metric space* (Θ, d) *such that* $\|m_\theta\|_\infty \leq B$ *for every* $\theta \in \Theta_k$. *Let* $J : \mathcal{K} \to [0, \infty)$ *be (possible random) functions satisfying, for every* $k \in \mathcal{K}$ *and some* $C > 0$*, with probability at least* $1 - e^{-x_k - \xi}$,

$$\sqrt{n} J(k) \geq E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) + B\sqrt{2x_k} - C\sqrt{2\xi},$$

for given numbers x_k *such that* $\sum_k e^{-x_k} \leq 1$.

Model selection: Statistical Learning

Theorem (3.5.11)

Let $\{m_\theta : \theta \in \Theta\}$ be a class of measurable functions $m_\theta : \mathcal{X} \to \mathbb{R}$ indexed by a metric space (Θ, d) such that $\|m_\theta\|_\infty \leq B$ for every $\theta \in \Theta_k$. Let $J : \mathcal{K} \to [0, \infty)$ be (possible random) functions satisfying, for every $k \in \mathcal{K}$ and some $C > 0$, with probability at least $1 - e^{-x_k - \xi}$,

$$\sqrt{n} J(k) \geq E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) + B\sqrt{2x_k} - C\sqrt{2\xi},$$

for given numbers x_k such that $\sum_k e^{-x_k} \leq 1$. Then

$$E^* P(m_{\hat{\theta}} - m_{\theta^*}) \leq \inf_{k \in \mathcal{K}} \Big(P(m_{\theta_k^*} - m_{\theta^*}) + EJ(k) \Big) + \frac{\sqrt{2\pi}(B + C)}{\sqrt{n}}.$$

Proof Theorem 3.5.11 1/3.

We use Theorem 2.15.1:

Theorem (2.15.1)

If $\mathcal{F}$ is a class of measurable functions $f : \mathcal{X} \to \mathbb{R}$ such that $|f(x) - f(y)| \leq 1$ for every $f \in \mathcal{F}$ and every $x, y \in \mathcal{X}$, then, for all $t \geq 0$,

$$P^* \left(\left| \sup_f \mathbb{G}_n f - E^* \sup_f \mathbb{G}_n f \right| \geq t \right) \leq 2 \exp(-2t^2),$$

Proof Theorem 3.5.11 1/3.

We use Theorem 2.15.1:

Theorem (2.15.1)

If $\mathcal{F}$ is a class of measurable functions $f : \mathcal{X} \to \mathbb{R}$ such that $|f(x) - f(y)| \leq 1$ for every $f \in \mathcal{F}$ and every $x, y \in \mathcal{X}$, then, for all $t \geq 0$,

$$P^* \left(\left| \sup_f \mathbb{G}_n f - E^* \sup_f \mathbb{G}_n f \right| \geq t \right) \leq 2 \exp(-2t^2),$$

Take $\mathcal{F} = \{-\frac{m_\theta}{2B} : \theta \in \Theta_k\}$ and $t = \frac{1}{2}\sqrt{2(x_k + \xi)}$:

Proof Theorem 3.5.11 1/3.

We use Theorem 2.15.1:

Theorem (2.15.1)

If $\mathcal{F}$ is a class of measurable functions $f : \mathcal{X} \to \mathbb{R}$ such that $|f(x) - f(y)| \leq 1$ for every $f \in \mathcal{F}$ and every $x, y \in \mathcal{X}$, then, for all $t \geq 0$,

$$P^* \left(\left| \sup_f \mathbb{G}_n f - E^* \sup_f \mathbb{G}_n f \right| \geq t \right) \leq 2 \exp(-2t^2),$$

Take $\mathcal{F} = \{-\frac{m_\theta}{2B} : \theta \in \Theta_k\}$ and $t = \frac{1}{2}\sqrt{2(x_k + \xi)}$:

$$\sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) \leq E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) + B\sqrt{2x_k + 2\xi}$$

with probability at least $1 - e^{-x_k - \xi}$.

Proof Theorem 3.5.11 2/3.

With probability at least $1 - e^{-x_k - \xi}$,

$$\sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) \leq E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) + B\sqrt{2x_k + 2\xi},$$

Model selection: Statistical Learning

Proof Theorem 3.5.11 2/3.

With probability at least $1 - e^{-x_k - \xi}$,

$$\sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) \leq E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) + B\sqrt{2x_k + 2\xi},$$

and by assumption, with probability at least $1 - e^{-x_k - \xi}$,

$$E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) \leq \sqrt{n} J(k) + C\sqrt{2\xi} - B\sqrt{2x_k}$$

Model selection: Statistical Learning

Proof Theorem 3.5.11 2/3.

With probability at least $1 - e^{-x_k - \xi}$,

$$\sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) \leq E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) + B\sqrt{2x_k + 2\xi},$$

and by assumption, with probability at least $1 - e^{-x_k - \xi}$,

$$E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) \leq \sqrt{n} J(k) + C\sqrt{2\xi} - B\sqrt{2x_k}$$

So with probability at least $1 - 2e^{-x_k - \xi}$:

$$\sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) \leq \sqrt{n} J(k) + (B + C)\sqrt{2\xi}.$$

Proof Theorem 3.5.11 2/3.

With probability at least $1 - e^{-x_k - \xi}$,

$$\sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) \le E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) + B\sqrt{2x_k + 2\xi},$$

and by assumption, with probability at least $1 - e^{-x_k - \xi}$,

$$E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) \le \sqrt{n} J(k) + C\sqrt{2\xi} - B\sqrt{2x_k}$$

So with probability at least $1 - 2e^{-x_k - \xi}$:

$$\sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) \le \sqrt{n} J(k) + (B + C)\sqrt{2\xi}.$$

This is true for all k simultaneously with probability $1 - 2e^{-\xi}$.

Proof Theorem 3.5.11 3/3.

Choose $\theta = \hat{\theta}_k$ and $k = \hat{k}$ to get, with probability at least $1 - 2e^{-\xi}$

$$-\frac{1}{\sqrt{n}}\mathbb{G}_n m_{\hat{\theta}} \leq J(\hat{k}) + (B + C)\sqrt{\frac{2\xi}{n}}.$$

Proof Theorem 3.5.11 3/3.

Choose $\theta = \hat{\theta}_k$ and $k = \hat{k}$ to get, with probability at least $1 - 2e^{-\xi}$

$$-\frac{1}{\sqrt{n}}\mathbb{G}_n m_{\hat{\theta}} \leq J(\hat{k}) + (B + C)\sqrt{\frac{2\xi}{n}}.$$

Note from earlier that

$$\begin{aligned} P(m_{\hat{\theta}} - m_{\theta^*}) \leq\ & P(m_{\theta_k^*} - m_{\theta^*}) + J(k) - J(\hat{k}) \\ & + \frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\theta_k^*} - m_{\bar{\theta}_k}) + \frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\bar{\theta}_k} - m_{\hat{\theta}}), \end{aligned}$$

Proof Theorem 3.5.11 3/3.

Choose $\theta = \hat{\theta}_k$ and $k = \hat{k}$ to get, with probability at least $1 - 2e^{-\xi}$

$$-\frac{1}{\sqrt{n}}\mathbb{G}_n m_{\hat{\theta}} \leq J(\hat{k}) + (B + C)\sqrt{\frac{2\xi}{n}}.$$

Note from earlier that

$$\begin{aligned} P(m_{\hat{\theta}} - m_{\theta^*}) \leq\ & P(m_{\theta_k^*} - m_{\theta^*}) + J(k) - J(\hat{k}) \\ & + \frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\theta_k^*} - m_{\bar{\theta}_k}) + \frac{1}{\sqrt{n}}\mathbb{G}_n(m_{\bar{\theta}_k} - m_{\hat{\theta}}), \end{aligned}$$

hence with probability at least $1 - 2e^{-\xi}$

$$P(m_{\hat{\theta}} - m_{\theta^*}) \leq P(m_{\theta_k^*} - m_{\theta^*}) + J(k) + \frac{1}{\sqrt{n}}\mathbb{G}_n m_{\theta_k^*} + (B + C)\sqrt{\frac{2\xi}{n}}$$

Model selection: Statistical Learning

Theorem (3.5.11 result)

$$E^* P(m_{\hat{\theta}} - m_{\theta^*}) \leq \inf_{k \in \mathcal{K}} \Big(P(m_{\theta_k^*} - m_{\theta^*}) + EJ(k) \Big) + \frac{\sqrt{2\pi}(B + C)}{\sqrt{n}}.$$

- The right side is never smaller than $O(n^{-1/2})$, so the rate is never better than $n^{-1/4}$.

Model selection: Statistical Learning

Theorem (3.5.11 result)

$$E^* P(m_{\hat{\theta}} - m_{\theta^*}) \leq \inf_{k \in \mathcal{K}} \Big(P(m_{\theta_k^*} - m_{\theta^*}) + EJ(k) \Big) + \frac{\sqrt{2\pi}(B+C)}{\sqrt{n}}.$$

- The right side is never smaller than $O(n^{-1/2})$, so the rate is never better than $n^{-1/4}$.

Theorem (3.5.11 condition)

$$\sqrt{n} J(k) \geq E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) + B\sqrt{2x_k} - C\sqrt{2\xi}$$

- Hence the rate is never smaller than $n^{-1/2} E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta)$. This is significantly worse than the lower bound on the rate with $E^* \sup_{\theta \in \Theta_k : d(\theta, \bar{\theta}_k) < \delta} \mathbb{G}_n(m_{\bar{\theta}_k} - m_\theta)$ in Theorem 3.5.4.

Model selection: Statistical Learning

Theorem (3.5.11 condition)

$$\sqrt{n}J(k) \geq E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) + B\sqrt{2x_k} - C\sqrt{2\xi}$$

$J(k)$ is allowed to be data-dependent. If not, then it should hold with probability one (typically $C = \xi = 0$).

Theorem (3.5.11 condition)

$$\sqrt{n}J(k) \geq E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) + B\sqrt{2x_k} - C\sqrt{2\xi}$$

$J(k)$ is allowed to be data-dependent. If not, then it should hold with probability one (typically $C = \xi = 0$). This yields

$$J(k) \geq \frac{1}{\sqrt{n}} E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) + B\sqrt{\frac{2x_k}{n}}.$$

We need an upper bound on the first term that does not depend on the distribution.

Example (3.5.12)

We bound

$$E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) \leq 2E^* \|\mathbb{G}_n^o\|_{\mathcal{M}_k}.$$

Model selection: Statistical Learning

Example (3.5.12)

We bound

$$E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) \leq 2E^* \|\mathbb{G}_n^o\|_{\mathcal{M}_k}.$$

By Proposition 2.15.3, with probability at least $1 - e^{-x_k - \xi}$

$$2E^* \|\mathbb{G}_n^o\|_{\mathcal{M}_k} \leq 2\|\mathbb{G}_n^o\|_{\mathcal{M}} + 2B\sqrt{2x_k + 2\xi}.$$

Model selection: Statistical Learning

Example (3.5.12)

We bound

$$E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) \leq 2E^* \|\mathbb{G}_n^o\|_{\mathcal{M}_k}.$$

By Proposition 2.15.3, with probability at least $1 - e^{-x_k - \xi}$

$$2E^* \|\mathbb{G}_n^o\|_{\mathcal{M}_k} \leq 2\|\mathbb{G}_n^o\|_{\mathcal{M}} + 2B\sqrt{2x_k + 2\xi}.$$

Thus with probability at least $1 - e^{-x_k - \xi}$

$$2E^* \|\mathbb{G}_n^0\|_{\mathcal{M}_k} + B\sqrt{2x_k} - 2B\sqrt{2\xi}$$

Model selection: Statistical Learning

Example (3.5.12)

We bound

$$E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) \leq 2E^* \|\mathbb{G}_n^o\|_{\mathcal{M}_k}.$$

By Proposition 2.15.3, with probability at least $1 - e^{-x_k - \xi}$

$$2E^* \|\mathbb{G}_n^o\|_{\mathcal{M}_k} \leq 2\|\mathbb{G}_n^o\|_{\mathcal{M}} + 2B\sqrt{2x_k + 2\xi}.$$

Thus with probability at least $1 - e^{-x_k - \xi}$

$$\begin{aligned} &2E^* \|\mathbb{G}_n^0\|_{\mathcal{M}_k} + B\sqrt{2x_k} - 2B\sqrt{2\xi} \\ &\leq 2\|\mathbb{G}_n^o\|_{\mathcal{M}_k} + 2B\sqrt{2x_k} + 2B\sqrt{2\xi} + B\sqrt{2x_k} - 2B\sqrt{2\xi} \end{aligned}$$

Model selection: Statistical Learning

Example (3.5.12)

We bound

$$E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) \leq 2E^* \|\mathbb{G}_n^o\|_{\mathcal{M}_k}.$$

By Proposition 2.15.3, with probability at least $1 - e^{-x_k - \xi}$

$$2E^* \|\mathbb{G}_n^o\|_{\mathcal{M}_k} \leq 2\|\mathbb{G}_n^o\|_{\mathcal{M}} + 2B\sqrt{2x_k + 2\xi}.$$

Thus with probability at least $1 - e^{-x_k - \xi}$

$$\begin{aligned} &2E^* \|\mathbb{G}_n^0\|_{\mathcal{M}_k} + B\sqrt{2x_k} - 2B\sqrt{2\xi} \\ &\leq 2\|\mathbb{G}_n^o\|_{\mathcal{M}_k} + 2B\sqrt{2x_k} + 2B\sqrt{2\xi} + B\sqrt{2x_k} - 2B\sqrt{2\xi} \\ &= 2\|\mathbb{G}_n^o\|_{\mathcal{M}_k} + 3B\sqrt{2x_k}. \end{aligned}$$

Model selection: Statistical Learning

Example (3.5.12)

Therefore we set

$$J(k) = 2 \sup_{\theta \in \mathcal{M}_k} |\mathbb{P}_n^o m_\theta| + 3B\sqrt{\frac{2x_k}{n}},$$

which satisfies the condition with probability at least $1 - e^{-x_k - \xi}$.

Theorem (3.5.11 condition)

$$\sqrt{n}J(k) \geq E^* \sup_{\theta \in \Theta_k} \mathbb{G}_n(-m_\theta) + B\sqrt{2x_k} - C\sqrt{2\xi}$$