Problem Set Solutions

1. **Optimization of Marginal Likelihood**

   The optimization of the marginal likelihood w.r.t. the hyperparameters is generally not possible in closed form. Consider, however, the situation where one hyperparameter, θ_0 gives the overall scale of the covariance

   $$k_y(\mathbf{x}, \mathbf{x}') = \theta_0 \tilde{k}_y(\mathbf{x}, \mathbf{x}'),$$

   where k_y is the covariance function for the noisy targets (i.e. including noise contributions) and $\tilde{k}_y(\mathbf{x}, \mathbf{x}')$ may depend on further hyperparameters, $\theta_1, \theta_2, \ldots$. Show that the marginal likelihood can be optimized w.r.t. θ_0 in closed form.

   Solution: The log marginal likelihood for a Gaussian process regression model with zero mean is given by:

   $$L = -\frac{1}{2}\mathbf{y}^\top K_y^{-1}\mathbf{y} - \frac{1}{2}\log|K_y| - \frac{n}{2}\log(2\pi)$$

   where K_y is the $n \times n$ covariance matrix of the targets, evaluated at the training points. Substituting $K_y = \theta_0 \tilde{K}_y$, we have $K_y^{-1} = \frac{1}{\theta_0}\tilde{K}_y^{-1}$ and $\log|K_y| = \log|\theta_0 \tilde{K}_y| = n\log\theta_0 + \log|\tilde{K}_y|$.

   The log marginal likelihood becomes:

   $$L(\theta_0) = -\frac{1}{2\theta_0}\mathbf{y}^\top \tilde{K}_y^{-1}\mathbf{y} - \frac{n}{2}\log\theta_0 - \frac{1}{2}\log|\tilde{K}_y| - \frac{n}{2}\log(2\pi)$$

   To find the optimal θ_0, we take the derivative of L with respect to θ_0 and set it to zero:

   $$\frac{\partial L}{\partial \theta_0} = \frac{1}{2\theta_0^2}\mathbf{y}^\top \tilde{K}_y^{-1}\mathbf{y} - \frac{n}{2\theta_0} = 0$$

   Multiplying by $2\theta_0^2$ (assuming $\theta_0 > 0$), we get:

   $$\mathbf{y}^\top \tilde{K}_y^{-1}\mathbf{y} - n\theta_0 = 0 \implies \hat{\theta}_0 = \frac{1}{n}\mathbf{y}^\top \tilde{K}_y^{-1}\mathbf{y}$$

   This provides a closed-form solution for optimizing the scale hyperparameter θ_0 given the others.

2. **Brownian motion and Brownian Bridge**

   The Brownian motion is a Gaussian process defined for $x \geq 0$ and has $f(0) = 0$. It has mean zero and a non-stationary covariance function $k(x, x') = \min(x, x')$. If we condition it on passing

through $f(1) = 0$ we obtain a process known as the Brownian bridge (or *tied-down* Wiener process). Show that this process has covariance $k(x, x') = \min(x, x') - xx'$ for $0 \leq x, x' \leq 1$ and mean 0.

Solution: The joint distribution of the Wiener process evaluated at points x, x', and 1 is a multivariate Gaussian:

$$\begin{pmatrix} f(x) \\ f(x') \\ f(1) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} x & \min(x, x') & x \\ \min(x, x') & x' & x' \\ x & x' & 1 \end{pmatrix} \right)$$

where we used the fact that for $x, x' \leq 1$, $\min(x, 1) = x$ and $\min(x', 1) = x'$.

We want to find the conditional distribution of $(f(x), f(x'))$ given $f(1) = 0$. Using the standard formula for conditional Gaussian distributions ($\mu_{A|B} = \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B)$ and $\Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$), we have:

Conditional Mean:

$$\mu(x, x')_{|f(1)=0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} x \\ x' \end{pmatrix} 1^{-1}(0 - 0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Conditional Covariance:

$$\Sigma(x, x')_{|f(1)=0} = \begin{pmatrix} x & \min(x, x') \\ \min(x, x') & x' \end{pmatrix} - \begin{pmatrix} x \\ x' \end{pmatrix} 1^{-1} \begin{pmatrix} x & x' \end{pmatrix}$$

$$= \begin{pmatrix} x & \min(x, x') \\ \min(x, x') & x' \end{pmatrix} - \begin{pmatrix} x^2 & xx' \\ xx' & (x')^2 \end{pmatrix}$$

Thus, the cross-covariance is $k(x, x') = \min(x, x') - xx'$.

3. **Monotonicity of Predictive Variance**

Let $\mathrm{Var}_n(f(\mathbf{x}_*))$ be the predictive variance of a Gaussian process regression model at $\mathbf{x}_*$ given a dataset of size n.

$$\mathrm{Var}_n(f(\mathbf{x}_*)) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{k}_*, \qquad \mathbf{k}_* = \begin{bmatrix} k(\mathbf{x}_*, \mathbf{x}_1) \\ k(\mathbf{x}_*, \mathbf{x}_2) \\ \vdots \\ k(\mathbf{x}_*, \mathbf{x}_n) \end{bmatrix},$$

and K is the kernel covariance matrix of the observed indices. The corresponding predictive variance using a dataset of only the first $n-1$ training points is denoted $\mathrm{Var}_{n-1}(f(\mathbf{x}_*))$. Show that $\mathrm{Var}_n(f(\mathbf{x}_*)) \leq \mathrm{Var}_{n-1}(f(\mathbf{x}_*))$, i.e. that the predictive variance at $\mathbf{x}_*$ cannot increase as more training data is obtained.

Solution: The predictive variance represents the conditional variance of $f(\mathbf{x}_*)$ given the observations. Let $Y_{n-1} = \{y_1, \ldots, y_{n-1}\}$ be the first $n-1$ observations, and y_n be the n-th observation. We are interested in comparing $\text{Var}(f_* \mid Y_{n-1})$ and $\text{Var}(f_* \mid Y_{n-1}, y_n)$.

Because the joint distribution of f_*, Y_{n-1}, and y_n is Gaussian, we can use the law of conditional variance (or the properties of partitioned Gaussian covariance matrices). Treating conditioning on Y_{n-1} as our base measure, the joint distribution of f_* and y_n given Y_{n-1} is Gaussian.

The conditional variance formula gives:

$$\text{Var}(f_* \mid Y_{n-1}, y_n) = \text{Var}(f_* \mid Y_{n-1}) - \frac{\text{Cov}(f_*, y_n \mid Y_{n-1})^2}{\text{Var}(y_n \mid Y_{n-1})}$$

Since variance is always non-negative, the term $\frac{\text{Cov}(f_*, y_n|Y_{n-1})^2}{\text{Var}(y_n|Y_{n-1})}$ is greater than or equal to zero. Therefore:

$$\text{Var}(f_* \mid Y_{n-1}, y_n) \leq \text{Var}(f_* \mid Y_{n-1})$$

which directly translates to $\text{Var}_n(f(\mathbf{x}_*)) \leq \text{Var}_{n-1}(f(\mathbf{x}_*))$.

4. **From Basis Functions to GP Marginal Likelihood**

We have seen that a linear-in-the-parameters model with a Gaussian prior on the weights is mathematically equivalent to a Gaussian Process. Consider a model for periodic, smooth functions defined by the following finite basis:

$$f(x) = w_1 \cos(x) + w_2 \sin(x)$$

where the weights are assigned independent Gaussian priors: $w_1 \sim \mathcal{N}(0, \alpha)$ and $w_2 \sim \mathcal{N}(0, \alpha)$.

(a) Prove that this model induces a Gaussian Process. Determine its exact mean function $m(x)$ and covariance function $k(x, x')$.

Solution: Because $f(x)$ is a linear combination of Gaussian random variables w_1 and w_2, any finite vector of function evaluations $[f(x_1), \ldots, f(x_N)]^\top$ will follow a joint multivariate Gaussian distribution. By definition, this implies $f(x)$ is a Gaussian Process.

The mean function is:

$$\begin{aligned} m(x) &= \mathbb{E}[f(x)] \\ &= \mathbb{E}[w_1 \cos(x) + w_2 \sin(x)] \\ &= \mathbb{E}[w_1] \cos(x) + \mathbb{E}[w_2] \sin(x) \\ &= 0 \end{aligned}$$

The covariance function is:

$$\begin{aligned} k(x, x') &= \mathbb{E}[(f(x) - 0)(f(x') - 0)] \\ &= \mathbb{E}[(w_1 \cos(x) + w_2 \sin(x))(w_1 \cos(x') + w_2 \sin(x'))] \\ &= \mathbb{E}[w_1^2] \cos(x) \cos(x') + \mathbb{E}[w_2^2] \sin(x) \sin(x') + \mathbb{E}[w_1 w_2](\dots) \end{aligned}$$

Since w_1 and w_2 are independent, $\mathbb{E}[w_1 w_2] = 0$. Using $\mathbb{E}[w_1^2] = \mathbb{E}[w_2^2] = \alpha$, we get:

$$k(x, x') = \alpha(\cos(x) \cos(x') + \sin(x) \sin(x')) = \alpha \cos(x - x')$$

(b) Suppose we record a single noisy observation y_1 at an arbitrary input location x_1, where $y_1 = f(x_1) + \epsilon_1$ and $\epsilon_1 \sim \mathcal{N}(0, \sigma_n^2)$. Using the Gaussian conditioning formulas, derive the explicit posterior predictive mean $\mu_{*|y}$ and predictive variance $\Sigma_{*|y}$ for an arbitrary test point x_*.

Solution: First, we construct the necessary covariance matrices evaluated at the training point $X = [x_1]$ and test point $X_* = [x_*]$. Using our kernel $k(x, x') = \alpha \cos(x - x')$:

- $K(X, X) = k(x_1, x_1) = \alpha \cos(x_1 - x_1) = \alpha \cos(0) = \alpha$
- $K(X_*, X) = k(x_*, x_1) = \alpha \cos(x_* - x_1)$
- $K(X_*, X_*) = k(x_*, x_*) = \alpha \cos(x_* - x_*) = \alpha$

Using the predictive distribution formulas for GP regression:

$$\mu_{*|y} = K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} y_1$$

$$\mu_{*|y} = (\alpha \cos(x_* - x_1))(\alpha + \sigma_n^2)^{-1} y_1 = \frac{\alpha y_1}{\alpha + \sigma_n^2} \cos(x_* - x_1)$$

For the predictive variance:

$$\Sigma_{*|y} = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*)$$

$$\Sigma_{*|y} = \alpha - (\alpha \cos(x_* - x_1))(\alpha + \sigma_n^2)^{-1}(\alpha \cos(x_* - x_1)) = \alpha - \frac{\alpha^2 \cos^2(x_* - x_1)}{\alpha + \sigma_n^2}$$

(c) The marginal likelihood provides a principled way to optimize hyperparameters by balancing the data fit term and the complexity penalty. Formulate the log marginal likelihood for the single observation y_1 from Part (b) as a function of the hyperparameter α. Find the value of $\hat{\alpha} \geq 0$ that maximizes this likelihood (assuming σ_n^2 is fixed). What happens if $y_1^2 \leq \sigma_n^2$?

Solution: The general log marginal likelihood is:

$$\log p(y \mid X, \theta) = -\frac{1}{2} y^\top [K + \sigma_n^2 I]^{-1} y - \frac{1}{2} \log |K + \sigma_n^2 I| - \frac{n}{2} \log(2\pi)$$

For our single data point y_1 at x_1, we have $K = k(x_1, x_1) = \alpha$ and $n = 1$. Substituting these in:

$$\log p(y_1 \mid \alpha) = -\frac{y_1^2}{2(\alpha + \sigma_n^2)} - \frac{1}{2} \log(\alpha + \sigma_n^2) - \frac{1}{2} \log(2\pi)$$

Notice that the marginal likelihood does not depend on the input location x_1 due to the stationarity of the prior variance. To find the optimal hyperparameter α, we take the derivative with respect to α and set it to zero:

$$\frac{\partial}{\partial \alpha} \log p(y_1 \mid \alpha) = \frac{y_1^2}{2(\alpha + \sigma_n^2)^2} - \frac{1}{2(\alpha + \sigma_n^2)} = 0$$

Multiplying by $2(\alpha + \sigma_n^2)^2$, we obtain:

$$y_1^2 - (\alpha + \sigma_n^2) = 0 \implies \hat{\alpha} = y_1^2 - \sigma_n^2$$

Because α represents a variance, it must be non-negative. If $y_1^2 \leq \sigma_n^2$, the unconstrained optimal value would be negative. Applying the constraint $\alpha \geq 0$, the maximum marginal likelihood is achieved at $\hat{\alpha} = 0$.

Conceptually, if the magnitude of the observation y_1 is entirely explainable by the assumed noise variance σ_n^2, the model heavily penalizes complexity and collapses the GP variance to zero, predicting a flat function $f(x) = 0$.

5. **Minimax lower bound for sparse sequences**

Let $\Theta := \ell_0[s_n]$ as in PS8, we propose to demonstrate, as $n \to \infty$,

$$\inf_T \sup_{\theta \in \ell_0[s_n]} E_\theta \|T(X) - \theta\|^2 \geq 2 s_n \log(n/s_n)(1 + o(1)),$$

where the infimum is taken over all possible estimators of θ and $s_n = o(n)$, $\log(n)/s_n = o(1)$.

(a) Show that for Π any prior distribution on the whole $\mathbb{R}^n$,

$$R_M := \inf_T \sup_{\theta \in \Theta} E_\theta \|T(X) - \theta\|^2 \geq \inf_T \int_\Theta E_\theta \|T(X) - \theta\|^2 d\Pi(\theta).$$

Solution: For any estimator T, the supremum of the risk over the parameter space Θ is always greater than or equal to the average risk computed with respect to any probability measure Π restricted to Θ. That is, for a fixed T:

$$\sup_{\theta \in \Theta} E_\theta \|T(X) - \theta\|^2 \geq \int_\Theta E_\theta \|T(X) - \theta\|^2 d\Pi(\theta)$$

Taking the infimum over all possible estimators T on both sides preserves the inequality, directly yielding the stated result.

We consider the following prior distribution on $\mathbb{R}^n$, where $\alpha_n \in (0,1)$ and $M_n > 0$ are arbitrary for now,

$$\Pi \sim \bigotimes_{i=1}^{n} (1-\alpha_n)\delta_0 + \alpha_n \delta_{M_n}.$$

(b) Show that this is a distribution on the set $\Theta_1 := \{0, M_n\}^n$. Do we have $\Pi(\Theta) = 1$?

Solution: The prior Π is constructed as an independent product over the n coordinates. For each coordinate i, the marginal prior is a mixture of Dirac delta functions at 0 and M_n, meaning coordinate i only takes values in $\{0, M_n\}$. Thus, the joint distribution is fully supported on the Cartesian product $\Theta_1 = \{0, M_n\}^n$.

We do **not** have $\Pi(\Theta) = 1$ in general. Recall $\Theta = \ell_0[s_n]$, the set of vectors with at most s_n non-zero coordinates. Under Π, the number of non-zero elements follows a Binomial distribution $\mathcal{B}(n, \alpha_n)$. Thus, $\Pi(\Theta) = P(\mathcal{B}(n, \alpha_n) \leq s_n)$, which is not strictly 1 unless $\alpha_n = 0$ or $s_n = n$.

(c) Show that if θ is drawn according to Π, the preceding infimum can be restricted to the class $\mathcal{S}$ of estimators taking values in $[-2M_n, 2M_n]^n$ only. One could show that the quadratic risk of any estimator is at least as large as that of its 'projection' onto $[-2M_n, 2M_n]$.

Solution: Let $T(X)$ be an arbitrary estimator. We can define a projected estimator $T^{proj}(X)$ coordinate-wise such that $T_i^{proj}(X)$ is the closest point to $T_i(X)$ in the interval $[-2M_n, 2M_n]$. Since the true parameter θ_i drawn from Π is always in $\{0, M_n\} \subset [-2M_n, 2M_n]$, projecting $T_i(X)$ onto $[-2M_n, 2M_n]$ can only decrease its distance to θ_i. Therefore, $|T_i^{proj}(X) - \theta_i| \leq |T_i(X) - \theta_i|$ almost surely. Squaring and summing over i implies $\|T^{proj}(X) - \theta\|^2 \leq \|T(X) - \theta\|^2$. Thus, the infimum over all estimators is achieved (or arbitrarily well-approximated) by estimators in $\mathcal{S}$.

(d) Show that, for $\mathcal{S}$ defined in the previous question,

$$\begin{aligned} R_M &\geq \inf_{T \in \mathcal{S}} \int_{\mathbb{R}^n} E_\theta \|T(X) - \theta\|^2 d\Pi(\theta) - \sup_{T \in \mathcal{S}} \int_{\Theta^c} E_\theta \|T(X) - \theta\|^2 d\Pi(\theta) \\ &\geq \inf_{T} \int_{\mathbb{R}^n} E_\theta \|T(X) - \theta\|^2 d\Pi(\theta) - 10 n M_n^2 \Pi(\Theta^c). \end{aligned}$$

Solution: From Part 1, we can split the integral over $\mathbb{R}^n$: $\int_\Theta = \int_{\mathbb{R}^n} - \int_{\Theta^c}$. Taking the infimum over $T \in \mathcal{S}$, we get:

$$\inf_{T \in \mathcal{S}} \left(\int_{\mathbb{R}^n} \cdots - \int_{\Theta^c} \cdots \right) \geq \inf_{T \in \mathcal{S}} \int_{\mathbb{R}^n} \cdots - \sup_{T \in \mathcal{S}} \int_{\Theta^c} \cdots$$

For the second term, since $T \in \mathcal{S}$ ($T_i \in [-2M_n, 2M_n]$) and $\theta_i \in \{0, M_n\}$, the maximum distance per coordinate is $|2M_n - (-M_n)| = 3M_n$. Thus, the maximum squared distance is $(3M_n)^2 = 9M_n^2 \leq 10M_n^2$. Summing over n dimensions gives a maximum squared norm of $10nM_n^2$. Hence, the integral over Θ^c is trivially bounded by $10nM_n^2\Pi(\Theta^c)$.

Let $n\alpha_n = s_n - D_n\sqrt{s_n}$, and $M_n := \sqrt{2\log(1/\alpha_n) - C_n}$, with C_n and D_n two sequences that tend to infinity slowly.

(e) Using the deviation inequality $P(|\text{Bin}(n,\alpha_n) - n\alpha_n| > x) \le 2e^{-x^2/(2a_n(x))}$, with $a_n(x) = n\alpha_n(1-\alpha_n) + x/3$, show that for $D_n = 4(\log n)^{1/2}$, as $n \to \infty$, $nM_n^2\Pi(\Theta^c) = o(1)$.

Solution: The event Θ^c corresponds to the number of non-zero coefficients exceeding s_n. Under Π, this number is $K \sim \text{Bin}(n, \alpha_n)$. We want to bound $\Pi(\Theta^c) = P(K > s_n) \le P(|K - n\alpha_n| > s_n - n\alpha_n)$. Let $x = s_n - n\alpha_n = D_n\sqrt{s_n} = 4\sqrt{s_n \log n}$. For large n, $a_n(x) \approx n\alpha_n + x/3 \sim s_n$. Applying the deviation inequality:

$$\Pi(\Theta^c) \le 2\exp\left(-\frac{16 s_n \log n}{2s_n(1+o(1))}\right) = 2\exp(-8\log n) = 2n^{-8}.$$

We know $M_n^2 \approx 2\log(1/\alpha_n) \approx 2\log(n/s_n) \le 2\log n$. So $nM_n^2\Pi(\Theta^c) \le n(2\log n)(2n^{-8}) = \frac{4\log n}{n^7} = o(1)$.

(f) Express $\inf_T \int_{\mathbb{R}^n} E_\theta\|T(X) - \theta\|^2 d\Pi(\theta)$ as a function of the posterior mean $\overline{\theta} = \int \theta d\Pi(\theta \mid X)$ by interpreting it as a Bayesian risk, then calculate $\overline{\theta}$ explicitly.

Solution: The term $\inf_T \int_{\mathbb{R}^n} E_\theta\|T(X) - \theta\|^2 d\Pi(\theta)$ represents the Bayes risk under the prior Π and squared error loss. The estimator that minimizes the Bayes risk is the posterior mean $\overline{\theta}(X) = E[\theta \mid X]$. Thus, the infimum equals $\int E_\theta\|\overline{\theta} - \theta\|^2 d\Pi(\theta) = \mathbb{E}_{X,\theta}[\|\overline{\theta} - \theta\|^2]$. Because the prior and likelihood factorize, the posterior mean is computed coordinate-wise:

$$\overline{\theta}_i(X_i) = P(\theta_i = M_n \mid X_i)M_n + P(\theta_i = 0 \mid X_i) \cdot 0$$

Using Bayes' rule:

$$\overline{\theta}_i(X_i) = \frac{\alpha_n \phi(X_i - M_n)}{(1-\alpha_n)\phi(X_i) + \alpha_n\phi(X_i - M_n)} M_n$$

where ϕ is the density of the standard normal $\mathcal{N}(0,1)$.

(g) Show that for all i,

$$\begin{aligned} m_i &:= \int E_\theta(\overline{\theta}_i - \theta_i)^2 d\Pi(\theta_i) \ge \Pi(\theta_i = M_n) E_{M_n}(\overline{\theta}_i - M_n)^2 \\ &\ge \Pi(\theta_i = M_n) E_{M_n}[(\overline{\theta}_i - M_n)^2 \mathbf{1}_{|\varepsilon_i| \le K_n}], \end{aligned}$$

for any sequence $K_n \to \infty$ and deduce that for $C_n \to \infty$ sufficiently slowly, $m_i \ge \alpha_n M_n^2(1 + o(1))$, uniformly in i.

Solution: The first inequality trivially drops the positive contribution when $\theta_i = 0$. The second inequality drops the contribution when $|\varepsilon_i| > K_n$ (where $X_i = M_n + \varepsilon_i$). Under $\theta_i = M_n$, $X_i = M_n + \varepsilon_i$, so we rewrite $\overline{\theta}_i$:

$$\overline{\theta}_i = \frac{M_n}{1 + \frac{1-\alpha_n}{\alpha_n}\frac{\phi(M_n + \varepsilon_i)}{\phi(\varepsilon_i)}}$$

The likelihood ratio is $\frac{\phi(M_n+\varepsilon_i)}{\phi(\varepsilon_i)} = \exp\left(-\frac{M_n^2}{2} - M_n\varepsilon_i\right)$. Since $M_n^2 = 2\log(1/\alpha_n) - C_n$, we have:

$$\frac{1-\alpha_n}{\alpha_n} e^{-M_n^2/2 - M_n\varepsilon_i} \approx \frac{1}{\alpha_n} e^{-\log(1/\alpha_n) + C_n/2 - M_n\varepsilon_i} = e^{C_n/2 - M_n\varepsilon_i}$$

On the set $|\varepsilon_i| \le K_n$, if $C_n \to \infty$ slower than $M_n K_n$, the exponent goes to $+\infty$, meaning the denominator goes to $+\infty$. Consequently, $\overline{\theta}_i \to 0$. Thus, $(\overline{\theta}_i - M_n)^2 \to M_n^2$. Taking the expectation on this high-probability set gives $E_{M_n}[\dots] \sim M_n^2$. Since $\Pi(\theta_i = M_n) = \alpha_n$, we get $m_i \ge \alpha_n M_n^2(1 + o(1))$.

(h) Conclude.

Solution: From parts 4 and 6, the minimax risk R_M is bounded below by the Bayes risk minus a $o(1)$ term. The Bayes risk is $\sum_{i=1}^n m_i$. Using the uniform lower bound from part 7:

$$R_M \ge \sum_{i=1}^n m_i - o(1) \ge n\alpha_n M_n^2(1 + o(1)).$$

We have $n\alpha_n = s_n - D_n\sqrt{s_n} \sim s_n$. Also, $M_n^2 = 2\log(1/\alpha_n) - C_n$. Since $\alpha_n \sim s_n/n$, we have $\log(1/\alpha_n) \sim \log(n/s_n)$. Therefore, $M_n^2 \sim 2\log(n/s_n)$. Substituting these approximations yields:

$$R_M \ge s_n \left(2\log\frac{n}{s_n}\right)(1 + o(1)) = 2s_n \log\left(\frac{n}{s_n}\right)(1 + o(1)),$$

which is the desired minimax lower bound.