Appendix

In this appendix we provide proofs of all the technical derivations. For the reader convenience the corresponding lemmas and theorems are restated.

Appendix A: Normal Projections

The following standard lemma on normal projections will be needed later for computing expectations of projection estimators and linking them to the dimension of the corresponding subspace.

Lemma 1. *Let ϵ be a random $\mathcal{N}(0, I_N)$ vector in $\mathbb{R}^N$, and let $\mathcal{S}$ be a linear subspace in $\mathbb{R}^N$. Then the random vector $\mathbb{P}_{\mathcal{S}}\,\epsilon = U_{\mathcal{S}}\epsilon$ has an $\mathcal{N}(0, U_{\mathcal{S}})$ distribution with $U_{\mathcal{S}}$ being the projection matrix onto $\mathcal{S}$.*
The squared norm $\|\,\mathbb{P}_{\mathcal{S}}\,\epsilon\|^2$ is χ^2_d-distributed with $d = \dim(\mathcal{S})$. In particular $\mathbb{E}[\|\,\mathbb{P}_{\mathcal{S}}\,\epsilon\|^2] = d$.

Proof. By definition of Gaussian vectors, for every $a \in \mathbb{R}^N$ we have that

$$\mathbb{E}[\exp(\langle a, U_{\mathcal{S}}\epsilon\rangle)] = \mathbb{E}[\exp(\langle U_{\mathcal{S}}^T a, \epsilon\rangle)] = \exp(\frac{1}{2}\langle a, U_{\mathcal{S}}U_{\mathcal{S}}^T a\rangle),$$

and since $U_{\mathcal{S}}$ is a projection matrix, $U_{\mathcal{S}}U_{\mathcal{S}}^T = U_{\mathcal{S}}$, so $U_{\mathcal{S}}\epsilon \sim \mathcal{N}(0, U_{\mathcal{S}})$. Let $\{v_1, \ldots, v_d\}$ be an orthonormal basis for $\mathcal{S}$ and $V = [v_1, \ldots, v_d]$; then $V^TV = I_d$ and $V\epsilon \sim \mathcal{N}(0, I_d)$. Also,

$$\|\,\mathbb{P}_{\mathcal{S}}\,\epsilon\|^2 = \sum_{i=1}^{d}(v_i^T\epsilon)^2 = \|V\epsilon\|^2,$$

thus $\|\,\mathbb{P}_{\mathcal{S}}\,\epsilon\|^2$ is χ^2-distributed with d degrees of freedom, so

$$\mathbb{E}[\|\,\mathbb{P}_{\mathcal{S}}\,\epsilon\|^2] = d.$$

□

Appendix B: Two-Pass Dynamic Programming

Proof of (8). Since we defined $\pi_{\bar{m}}$ to be the partition having as elements all the segments of π_m, in particular we have $m \subset \bar{m}$ and

$$\mathbb{P}_{\mathcal{F}_m} = \mathbb{P}_{\mathcal{F}_{\bar{m}}}\,\mathbb{P}_{\mathcal{F}_m} = \mathbb{P}_{\mathcal{F}_m}\,\mathbb{P}_{\mathcal{F}_{\bar{m}}}\,.$$

Moreover, since $\mathbb{P}_{\mathcal{F}_{\bar{m}}}\,Y - \mathbb{P}_{\mathcal{F}_m}\,Y \in \mathcal{F}_{\bar{m}}$, then by the projection theorem we have that

$$(Y - \mathbb{P}_{\mathcal{F}_{\bar{m}}}\,Y)\perp(\mathbb{P}_{\mathcal{F}_{\bar{m}}}\,Y - \mathbb{P}_{\mathcal{F}_m}\,Y),$$

so the Pythagorean theorem implies that

$$\|Y - \mathbb{P}_{\mathcal{F}_m}\,Y\|^2 = \|Y - \mathbb{P}_{\mathcal{F}_{\bar{m}}}\,Y\|^2 + \|\,\mathbb{P}_{\mathcal{F}_{\bar{m}}}\,Y - \mathbb{P}_{\mathcal{F}_m}\,Y\|^2.$$

Thus, the minimization of criterion (4) simplifies to

$$\min_{m\in\mathcal{M}}\mathrm{Crit}(m) = \min_{0\leqslant d'\leqslant d''\leqslant D}\left\{\min_{|\bar{m}|=d''}\left\{\|Y - \mathbb{P}_{\mathcal{F}_{\bar{m}}}\,Y\|^2 + \min_{\substack{m\subset\bar{m}\\|m|=d'}}\|\,\mathbb{P}_{\mathcal{F}_{\bar{m}}}\,Y - \mathbb{P}_{\mathcal{F}_m}\,\mathbb{P}_{\mathcal{F}_{\bar{m}}}\,Y\|^2\right\} + \sigma^2 K\,\mathrm{pen}(d', d'')\right\}.$$

Instead of computing this minimum exactly we will take a greedy step in the second minimum by defining

$$\tilde{m} := \arg\min_{|\bar{m}|=d''}\|Y - \mathbb{P}_{\mathcal{F}_{\bar{m}}}\,Y\|^2,$$

and plugging it into the third minimum to obtain the following relaxation:

$$\begin{aligned}\min_{m\in\mathcal{M}}\mathrm{Crit}(m) &\leqslant \min_{0\leqslant d'\leqslant d''\leqslant D}\left\{\|Y - \mathbb{P}_{\mathcal{F}_{\tilde{m}}}\,Y\|^2 + \min_{\substack{m\subset\tilde{m}\\|m|=d'}}\|\,\mathbb{P}_{\mathcal{F}_{\tilde{m}}}\,Y - \mathbb{P}_{\mathcal{F}_m}\,\mathbb{P}_{\mathcal{F}_{\tilde{m}}}\,Y\|^2 + \sigma^2 K\,\mathrm{pen}(d', d'')\right\}\\ &= \min_{0\leqslant d'\leqslant d''\leqslant D}\left\{\min_{|m|=d''}\|Y - \mathbb{P}_{\mathcal{F}_m}\,Y\|^2 + \min_{\substack{m\subset\tilde{m}\\|m|=d'}}\|\,\mathbb{P}_{\mathcal{F}_{\tilde{m}}}\,Y - \mathbb{P}_{\mathcal{F}_m}\,\mathbb{P}_{\mathcal{F}_{\tilde{m}}}\,Y\|^2 + \sigma^2 K\,\mathrm{pen}(d', d'')\right\}.\end{aligned}$$

The second inner minimization of the last equation can then be relaxed by restricting it to partitions satisfying the clustering property, since

$$\min_{m\in\mathcal{M}_{\tilde{m},d'}} \|\,\mathbb{P}_{\mathcal{F}_{\tilde{m}}}Y-\mathbb{P}_{\mathcal{F}_m}\mathbb{P}_{\mathcal{F}_{\tilde{m}}}Y\|^2 \leqslant \min_{m\in\mathcal{M}_{\bar{y}_{\tilde{m}},d'}} \|\,\mathbb{P}_{\mathcal{F}_{\tilde{m}}}Y-\mathbb{P}_{\mathcal{F}_m}\mathbb{P}_{\mathcal{F}_{\tilde{m}}}Y\|^2.$$

This leads to the following upper bound:

$$\min_{m\in\mathcal{M}}\mathrm{Crit}(m) \leqslant \min_{0\leqslant d''\leqslant D}\left\{\min_{|m|=d''}\|Y-\mathbb{P}_{\mathcal{F}_m}Y\|^2+\min_{\substack{0\leqslant d'\leqslant d''\\ m\in\mathcal{M}_{\bar{y}_{\tilde{m}},d''}}}\|\,\mathbb{P}_{\mathcal{F}_{\tilde{m}}}Y-\mathbb{P}_{\mathcal{F}_m}\mathbb{P}_{\mathcal{F}_{\tilde{m}}}Y\|^2+\sigma^2K\,\mathrm{pen}(d',d'')\right\}.$$

Therefore, we can define the following relaxation for the minimization of the criterion in (4):

$$\mathrm{Crit}_r(d'') := \min_{|m|=d''}\|Y-\mathbb{P}_{\mathcal{F}_m}Y\|^2+\min_{\substack{0\leqslant d'\leqslant d''\\ m\in\mathcal{M}_{\bar{y}_{\tilde{m}},d''}}}\left\{\|\,\mathbb{P}_{\mathcal{F}_{\tilde{m}}}Y-\mathbb{P}_{\mathcal{F}_m}\mathbb{P}_{\mathcal{F}_{\tilde{m}}}Y\|^2+\sigma^2K\,\mathrm{pen}(d'_m,d''_m)\right\},$$

which corresponds to (8). □

Theorem 4.1. *Let $(y_i)_{i=1}^N\subset\mathbb{R}$, $D\in\mathbb{N}$ and $K>0$. Then, recalling the dynamic programming recursions in (9) and (10),*

- *for all $1\leqslant d\leqslant D$,*

$$\tilde{m}_d\in\arg\min_{|\tilde{m}|=d}\|Y-\mathbb{P}_{\mathcal{F}_{\tilde{m}}}Y\|^2,$$

- *for all $1\leqslant\delta\leqslant d\leqslant D$,*

$$\tilde{\tilde{m}}_{(d,\delta)}\in\arg\min_{m\in\mathcal{M}_{\bar{y}_{\tilde{m}}\delta}}\|\,\mathbb{P}_{\mathcal{F}_{\tilde{m}}}Y-\mathbb{P}_{\mathcal{F}_m}\mathbb{P}_{\mathcal{F}_{\tilde{m}}}Y\|^2.$$

Furthermore, Algorithm 1 correctly solves the minimization problem in (8), with time and space complexity $\mathcal{O}(N^3+D^4)$ and $\mathcal{O}(N^2+D^3)$, respectively.

Proof. Here we abuse the notation of m and $\mathcal{M}$ so that if Y^n is the sub-vector of the first n component of Y then m and $\mathcal{M}$ are still defined by mere restriction to the first n component and $\mathbb{P}_{\mathcal{F}_m}Y^n$ still makes sense for $m\in\mathcal{M}$.

To prove the 1^{st} point we need to show that $C_d(n)$, defined inductively as

$$\begin{aligned}C_1(n)&:=R_{[1,n]}, && n\in[\![1,N]\!],\\ C_d(n)&:=\min_{i\in[\![d,n]\!]}\{C_{d-1}(i-1)+R_{[i,n]}\}, && 2\leqslant d\leqslant D,\quad d\leqslant n\leqslant N,\end{aligned}$$

is equal to $\min_{|\tilde{m}|=d}\|Y^n-\mathbb{P}_{\mathcal{F}_{\tilde{m}}}Y^n\|^2$, with $Y^n=(Y_1,\dots,Y_n)$. This implies that for $n=N$ we obtain the result for all d. This is straightforward since if, for Y^n, $\pi_m=\{0=i_0<i_1<\cdots<i_d<i_{d+1}=n\}$ is a partition, then

$$\|Y^n-\mathbb{P}_{\mathcal{F}_{\tilde{m}}}Y^n\|^2=\sum_{k=0}^{d}R_{[i_k+1,i_{k+1}]}.$$

Taking the minimum over $|\bar{m}|=d$ or, equivalently, over the values of $i_1,i_2,\dots,i_d$, we obtain

$$\begin{aligned}\min_{|\tilde{m}|=d}\|Y^n-\mathbb{P}_{\mathcal{F}_{\tilde{m}}}Y^n\|^2&=\min_{0=i_0<i_1<\cdots<i_d<i_{d+1}=n}\sum_{k=0}^{d}R_{[i_k+1,i_{k+1}]}\\ &=\min_{d\leqslant i_d+1\leqslant n}\left\{\min_{0<i_1<\cdots<i_{d-1}<i_d}\left\{\sum_{k=0}^{d-1}R_{[i_k+1,i_{k+1}]}\right\}+R_{[i_d+1,n]}\right\}\\ &=\min_{i\in[\![d,n]\!]}\{C_{d-1}(i-1)+R_{[i,n]}\}.\end{aligned}$$

This yields our 1^{st} point:

$$\tilde{m}_d \in \arg\min_{|\bar{m}|=d} \|Y - \mathbb{P}_{\mathcal{F}_{\bar{m}}} Y\|^2.$$

To prove the second point, we first define, in the same notations of Algorithm 1, Y_r as a rearrangement of Y according to the permutation ϕ_d, and $\tilde{m}_r$ as a rearrangement of $\tilde{m}$. Also $Y_r^{(t)}$ denotes the truncation of Y_r to the t^{th}-segment. By the clustering property in (7) we have that $m \in \mathcal{M}_{\bar{y}_{\tilde{m}_r},\delta}$ if and only if $\bar{m} \in \mathcal{M}_{\bar{y}_{\tilde{m}_r},\delta}$, hence

$$\begin{aligned}\min_{m\in\mathcal{M}_{\bar{y}_{\tilde{m}},\delta}} \| \mathbb{P}_{\mathcal{F}_{\tilde{m}}} Y - \mathbb{P}_{\mathcal{F}_m} \mathbb{P}_{\mathcal{F}_{\tilde{m}}} Y\|^2 &= \min_{m\in\mathcal{M}_{\bar{y}_{\tilde{m}_r},\delta}} \| \mathbb{P}_{\mathcal{F}_{\tilde{m}_r}} Y_r - \mathbb{P}_{\mathcal{F}_m} \mathbb{P}_{\mathcal{F}_{\tilde{m}_r}} Y_r\|^2 \\ &= \min_{m\in\mathcal{M}_{\bar{y}_{\tilde{m}_r},\delta}} \| \mathbb{P}_{\mathcal{F}_{\tilde{m}_r}} Y_r - \mathbb{P}_{\mathcal{F}_{\tilde{m}}} \mathbb{P}_{\mathcal{F}_{\tilde{m}_r}} Y_r\|^2.\end{aligned}$$

Here we are back to the same setup of the 1^{st} point, so we need to show that $G_{(t,\delta)}$ defined inductively as

$$\begin{aligned} G_{(t,1)} &:= \bar{R}_{[1,t]}, && t \in [\![1, d]\!], \\ G_{(t,\delta)} &:= \min_{i\in[\![\delta,t]\!]} \{G_{(i-1,\delta-1)} + \bar{R}_{[i,t]}\}, && 2 \leqslant \delta \leqslant t \leqslant d, \end{aligned}$$

is equal to $\min_{\bar{m}\in\mathcal{M}_{\bar{y}_{\tilde{m}_r},\delta}} \| \mathbb{P}_{\mathcal{F}_{\bar{m}_r}} Y_r^t - \mathbb{P}_{\mathcal{F}_{\bar{m}}} \mathbb{P}_{\mathcal{F}_{\bar{m}_r}} Y_r^t\|^2$ and we obtain the desired result at $t = d$. The same derivation as for the 1^{st} point carries over by using $\bar{\bar{y}}_{(k,l)} := \frac{\sum_{i=k}^l \alpha_{(i)}\bar{y}_{(i)}}{\sum_{i=k}^l \alpha_{(i)}}$ and $\bar{R}_{[k,l]} := \sum_{i=k}^l \alpha_{(i)}(\bar{y}_{(i)} - \bar{\bar{y}}_{(k,l)})^2$ for all $1 \leqslant k \leqslant l \leqslant d$ as defined in Algorithm 1, and we obtain the result since Y_r^t is constant over every segment.

Since both points hold, then by the definition of $m_{(\hat{d},\hat{\delta})}$ in Algorithm 1 we obtain a solution to the minimization criterion in (8), thus establishing the correctness of Algorithm 1.

Regarding the complexity of the algorithm, the first step consists in making the $\bar{y}_{[k,l]}$ and $R_{[k,l]}$ matrices. This can be done efficiently by making a cumulative sum matrix $(\sum_{i=k}^l y_i)_{k,l}$, whose rows can be formed in $\mathcal{O}(N)$ time and the whole matrix in $\mathcal{O}(N^2)$ time. We compute the $\bar{y}_{[k,l]}$ matrix in $\mathcal{O}(N^2)$ time and $R_{[k,l]}$ in $\mathcal{O}(N^2)$ time, hence this first step has time complexity $\mathcal{O}(N^3)$ and space complexity $\mathcal{O}(N^2)$.

The 1^{st} dynamic programming recurrence has time complexity $\mathcal{O}(DN^2)$ and space complexity $\mathcal{O}(DN)$, since there are $\mathcal{O}(DN)$ comparisons to perform in order to find the minimum of $\mathcal{O}(N)$ elements.

D backtracking operations are needed for the 1^{st} dynamic programming recurrence. They run in $\mathcal{O}(D)$ time to obtain the optimal models $\tilde{m}_d := \{0 = i_0 \leqslant i_1 < i_2 < \cdots < i_d \leqslant i_{d+1} = N\}$ from 1 to D. This backtracking procedure has time complexity $\mathcal{O}(D^2)$ and space complexity $\mathcal{O}(D^2)$.

Each of the D sorting operations that return ϕ_d for all d can be done with $\mathcal{O}(D \ln D)$ space and time complexity; more efficient sorting algorithms can be used but since this is not the bottleneck operation in Algorithm 1 we do not require more efficiency. On the other hand $\mathcal{O}(D \ln D)$ space for sorting- D sorting stages can be done using the same memory space- is overcome by the D^2 storage cost of D storages. More efficient sorting algorithms are described in (Cormen et al., 2009) and (Goodrich & Tamassia, 2001) Overall, these steps have time complexity $\mathcal{O}(D^2 \ln D)$ and space complexity $\mathcal{O}(D^2)$.

For the second preprocessing steps we need to compute $(\alpha_{(k)})_{k=0}^d$, $\bar{\bar{y}}_{(k,l)}$ and $\bar{R}_{[k,l]}$. As before, we do this via a cumulative sum matrix $(\alpha_{(k)})_{k=0}^d$ which is built in $\mathcal{O}(D^2)$ time, and a weighed cumulative sum matrix $(\sum_{i=k}^l \alpha_{(i)}\bar{y}_i)_{k,l}$ built in $\mathcal{O}(D^2)$ time. We can then compute $(\bar{\bar{y}})_{[k,l]}$ in $\mathcal{O}(D^2)$ and $R_{[k,l]}$ in $\mathcal{O}(D^2)$ time, and doing this for D models requires a time complexity of $\mathcal{O}(D^4)$ and a space complexity of $\mathcal{O}(D^2)$.

The 2^{nd} dynamic programming step with backtracking now requires a time complexity of $\mathcal{O}(D^4)$ and a space complexity of $\mathcal{O}(D^3)$ to store $\tilde{\tilde{m}}_{(d,\delta)}$ for all d and δ.

Computing $B_{(d,\delta)}$ requires obtaining $\text{pen}((d, \delta))$ (see (20) and (19)), which can be done recursively using the s (2) and (3) in $\mathcal{O}(DN + D^2)$ in time and $\mathcal{O}(DN + D^2)$ in space; the minimization to compute $\text{Crit}(m_{(\hat{d},\hat{\delta})})$, which requires $\mathcal{O}(D^2)$ time and $\mathcal{O}(1)$ space; and backtracking to obtain $m_{(\hat{d},\hat{\delta})}$, which requires $\mathcal{O}(D)$ time and $\mathcal{O}(D)$ space, since everything is already stored so we just need to look up $\tilde{m}_{\hat{d}}$ and rearrange back $\tilde{\tilde{m}}_{(\hat{d},\hat{\delta})}$ using $\phi_{\hat{d}}$.

The overall complexity of the algorithm is dominated by $\mathcal{O}(N^3 + D^4)$ in time and $\mathcal{O}(N^2 + D^3)$ in space. □

The time and space complexity can be improved upon using the efficient implementation in (Celisse et al., 2017) for the dynamic programming of Steps 3 and 11, this implementation change the order of the for loops of d' and d'' and computes recursively the values of $R_{[i,n]}$ and $\bar{R}_{[i,t]}$ as needed, so no preprocessing of steps 1 and 9 is needed. With this, we get a time and space complexity $\mathcal{O}(N^2D + D^4)$ and $\mathcal{O}(DN + D^3)$, respectively. The useful regime in which the result of this algorithm are significant is $d''_{m^*} = o(\frac{N}{\log N})$ according to corollary 6.1 so we only to choose D within those constraint. To balance computational performance and statistical performance we can choose $D = \mathcal{O}(N^{1/2})$ giving us a time and space complexity $\mathcal{O}(N^{5/2}$ and $\mathcal{O}(N^{3/2})$, respectively.

Appendix C: Model Selection criterion for change points and clustering

Proof of (12). We start by computing the probability distribution of F; the law of total probability yields

$$d\mu_F = \sum_{m \in \mathcal{M}} p_m d\mu_{F/m},$$

and Bayes' theorem then provides the posterior distribution of Y/m,

$$\begin{aligned}\frac{d\mu_{m/Y}}{d\mu_m} &= \frac{d\mu_{Y/m}}{d\mu_Y} \\ &= \frac{d\mu_{Y/m}}{\int d\mu_{Y/m'} d\mu_{m'}} \\ &= \frac{\int d\mu_{Y/F} d\mu_{F/m}}{\sum_{m' \in \mathcal{M}} p_{m'} \int d\mu_{Y/F} d\mu_{F/m'}}.\end{aligned}$$

Both $\mu_{m/Y}$ and $d\mu_m$ are absolutely continuous with respect to the counting measure, hence we have

$$\frac{d\mu_{m/Y}}{d\mu_m} = \frac{p_{m/Y}}{p_m}.$$

We denote by ϕ_N the density of the multivariate $\mathcal{N}(0, I_N)$ distribution. The law of total probability again gives

$$d\mu_{Y/m} = \int_{f \in \mathcal{F}_m} \phi_N \left(\frac{Y - f}{\sigma} \right) l_{f/m}(f) df.$$

Putting these conditional probabilities together, we obtain the following a-posteriori distribution for the random variable m given the observation Y:

$$p_{m/Y} = \frac{p_m \int_{f \in \mathcal{F}_m} \phi_N \left(\frac{Y - f}{\sigma} \right) l_{f/m}(f) df}{\sum_{m' \in \mathcal{M}} p_{m'} \int_{f' \in \mathcal{F}_m} \phi_N \left(\frac{Y - f'}{\sigma} \right) l_{f/m'}(f') df'}.$$

This complete the proof. □

Lemma 2. *Let $l \colon \mathbb{R} \to \mathbb{R}^+$ be a four times differentiable probability density function. Define, for $f \in \mathbb{R}$, $y^n \in \mathbb{R}^n$:*

$$\begin{aligned}L_n(f, y^n) &:= -\frac{\|y^n - f\mathbb{1}^n\|_2^2}{2n\sigma^2} + \frac{1}{n} \ln l(f), \\ \sigma^2_{L_n}(f) &:= \frac{1}{|L''_n(f, y^n)|}.\end{aligned}$$

Let $\mathcal{A} \subset \mathbb{R}^\infty$ such that for all $y \in \mathcal{A}$ the following holds:

- *the integrals $\int_{\mathbb{R}} \exp(L_n(f, y^n)) df$ are bounded uniformly (in n and y^n) by some constant β.*
- *the sequence $\hat{f}_n := \frac{\sum_{k=1}^n y_k}{n}$ converges and has as limit $\hat{f} = \lim_{n \to \infty} \frac{\sum_{k=1}^n y_k}{n}$.*

- *$L_n(\cdot, y^n)$ has a sequence of maximizers $(\bar{f}_n(y^n))_{n=1}^{\infty}$ at which $L_n(\cdot, y^n)$ has a negative second derivative.*
- *there is a $\delta_0,\ N_0$ and $m, M > 0$ such that for all $n \geqslant N_0$ we have $|L_n^{(i)}(f, y^n)| < M$ for all $|f - \bar{f}_n| < \delta_0$ for $i \in [\![2, 4]\!]$ and $m < |L_n^{(2)}(f, y^n)|$, where $L_n^{(i)}$ is the i^{th}-derivative of $L_n(f, y^n)$ with respect to f.*

then the sequence $(\bar{f}_n(y^n))_{n\in\mathbb{N}}$ converges to $\bar{f} = \lim\limits_{n\to\infty} \frac{\sum_{k=1}^n y_k}{n}$ and we have:

$$\int_{\mathbb{R}} \exp\left(-\frac{\|y^n - f\mathbb{1}^n\|_2^2}{2\sigma^2} + \ln l(f)\right) df = \frac{\sqrt{2\pi}\sigma_{L_n}(\bar{f}_n)}{\sqrt{n}} \exp(nL_n(\bar{f}_n))(1 + \mathcal{O}(n^{-3/2})).$$

Proof. We begin by observing that $\hat{f}_n = \frac{\sum_{i=1}^n y_i}{n} = \arg\max_{f\in\mathbb{R}} -\frac{\|y^n - f\mathbb{1}^n\|_2^2}{2n\sigma^2}$. We will show that $(\bar{f}_n)_{n=1}^{\infty}$ should have the same limit as $(\hat{f}_n)_{n=1}^{\infty}$.
Since $\frac{d^2}{df^2}\sum_{k=1}^n (y_k - f)^2 = 2n > 0$, we have, by integration, that

$$-\sum_{k=1}^n (y_k - f)^2 + \sum_{k=1}^n (y_k - \hat{f}_n)^2 = -n(f - \hat{f}_n)^2.$$

Thus, we obtain

$$\limsup_{n\to\infty} \frac{1}{2n\sigma^2} \sup_{|f-\hat{f}_n|>\delta} \left\{-\sum_{k=1}^n (y_k - f)^2 + \sum_{k=1}^n (y_k - \hat{f}_n)^2\right\} = \limsup_{n\to\infty} \frac{1}{2\sigma^2} \sup_{|f-\hat{f}_n|>\delta} -(f - \hat{f}_n)^2 = -\frac{\delta^2}{2\sigma^2}.$$

If there is $\delta_0 > 0$ and a subsequence $(n_l)_{l=1}^{\infty}$ such that $|\hat{f}_{n_l} - \bar{f}_{n_l}| > \delta_0$ for all l, then

$$-\frac{1}{2n_l\sigma^2}\sum_{k=1}^{n_l} (y_k - \bar{f}_{n_l})^2 \leqslant -\frac{\delta_0^2}{2\sigma^2} - \frac{1}{2n_l\sigma^2}\sum_{k=1}^{n_l} (y_k - \hat{f}_{n_l})^2.$$

By assumption $\frac{\sum_{k=1}^n y_k}{n} = \hat{f}_n \to \hat{f}$ then $\ln l(\hat{f}_n) \to \ln l(\hat{f})$. Thus, there is an N_1 such for all $n \geqslant N_1$ we obtain

$$-\frac{1}{n} + \frac{l(\hat{f})}{n} \leqslant \frac{l(\hat{f}_n)}{n}.$$

On the other hand, l is bounded by some C, hence

$$\frac{1}{n}\ln l(\bar{f}_n) \leqslant \frac{1}{n}\ln(C).$$

This gives

$$\frac{1}{n}\ln l(\bar{f}_n) - \frac{1}{n}\ln l(\hat{f}_n) \leqslant \frac{1}{n}\ln(C) + \frac{1}{n} - \frac{1}{n}\ln l(\hat{f}) = \mathcal{O}\left(\frac{1}{n}\right).$$

Therefore, for all $\hat{f}_{n_l}$ and $\bar{f}_{n_l}$ such that $|\hat{f}_{n_l} - \bar{f}_{n_l}| > \delta_0$ we have that

$$\begin{aligned} L_{n_l}(\bar{f}_{n_l}, y^{n_l}) =& -\frac{1}{2n_l\sigma^2}\sum_{k=1}^{n_l} (y_k - \bar{f}_{n_l})^2 + \frac{1}{n_l}\ln l(\bar{f}_{n_l}) \\ \leqslant& -\frac{1}{2n_l\sigma^2}\sum_{k=1}^{n_l} (y_k - \hat{f}_{n_l})^2 + \frac{1}{n_l}\ln l(\hat{f}_{n_l}) + \mathcal{O}(n_l^{-1}) - \frac{\delta_0^2}{2\sigma^2} \\ =& L_{n_l}(\hat{f}_{n_l}, y^{n_l}) - \frac{\delta_0^2}{2\sigma^2} + \mathcal{O}(n_l^{-1}). \end{aligned} \tag{18}$$

This implies that there is an l_0 such that $L_{n_{l_0}}(\hat{f}_{n_{l_0}}, y^{n_{l_0}}) < L_{n_{l_0}}(\bar{f}_{n_{l_0}}, y^{n_{l_0}}) - \frac{\delta_0^2}{4\sigma^2}$, which in turn contradicts the fact that $\bar{f}_n$ is a maximizer of L_n for all n.

From here we conclude that for all δ there is $N_2 \in \mathbb{N}$ for all $n \geqslant N_2$ we have $|\bar{f}_n - \hat{f}_n| < \delta/2$, and since the sequence $(\hat{f}_n)_{n=1}^{\infty}$ converges, there is an N_3 for all $n \geqslant N_3$ we have that $|\hat{f}_n - \hat{f}| < \delta/2$, which in turn implies that for all $n \geqslant N_4 = \max\{N_2, N_3\}$ we have $|\bar{f}_n - \hat{f}| < \delta$. Thus our original claim and first part of the lemma is proved, namely, that the sequence $(\bar{f}_n)_{n=1}^{\infty}$ is convergent and $\hat{f} = \lim_{n\to\infty} \hat{f}_n = \lim_{n\to\infty} \bar{f}_n = \bar{f}$.
For a fixed $\delta > 0$ we will approximate both terms of the following decomposition:

$$\int_{\mathbb{R}} \exp(nL_n(f, y^n))df = \int_{\mathbb{R}-(\hat{f}-\delta,\hat{f}+\delta)} \exp(nL_n(f, y^n))df + \int_{(\hat{f}-\delta,\hat{f}+\delta)} \exp(nL_n(f, y^n))df.$$

Just as before, since $\bar{f}_n \to \bar{f}$ there is an N_5 such for all $n \geqslant N_5$, for all $f \in \mathbb{R} - (\hat{f}_n - \delta, \hat{f}_n + \delta)$,

$$-\frac{1}{2n\sigma^2}\sum_{k=1}^{n}(y_k - f)^2 \leqslant -\frac{1}{2n\sigma^2}\sum_{k=1}^{n}(y_k - \hat{f}_n)^2 - \frac{\delta^2}{2\sigma^2}.$$

Since both $\hat{f}_n$ and $\bar{f}_n$ have the same limit, then by continuity there is N_6 starting from which we get:

$$-\frac{1}{2n\sigma^2}\sum_{k=1}^{n}(y_k - f)^2 \leqslant -\frac{1}{2n\sigma^2}\sum_{k=1}^{n}(y_k - \bar{f}_n)^2 - \frac{\delta^2}{4\sigma^2}.$$

Following the same steps leading to (18), we obtain:

$$\frac{1}{n}\ln l(f) \leqslant \frac{1}{n}\ln l(\bar{f}_n) + \mathcal{O}\left(\frac{1}{n}\right) \leqslant \frac{1}{n}\ln(C) + \mathcal{O}\left(\frac{1}{n}\right).$$

These two results together imply for that there is an N_7 such that for all $n \geqslant N_7$ we have

$$L_n(f, y^n) \leqslant -\frac{\delta^2}{8\sigma^2}.$$

Hence for $n \geqslant N_6$ we can bound the first integral as follows:

$$\begin{aligned}\int_{\mathbb{R}-(\bar{f}_n-\delta,\bar{f}_n+\delta)} \exp(nL_n(f, y^n))df &= \int_{\mathbb{R}-(\bar{f}_n-\delta,\bar{f}_n+\delta)} \exp((n-1)L_n(f, y^n))\exp(L_n(f, y^n))df \\ &\leqslant \exp\left(-(n-1)\frac{\delta^2}{8\sigma^2}\right)\int_{\mathbb{R}} \exp(L_n(f, y^n))df \\ &\leqslant \exp\left(-(n-1)\frac{\delta^2}{8\sigma^2}\right)\beta.\end{aligned}$$

Now we turn to the integral over $(\bar{f}_n - \delta, \bar{f}_n + \delta)$. Taking the Taylor series of L_n around $\bar{f}_n$ (and omitting the second argument, y^n, for simplicity) we obtain

$$L_n(f) = L_n(\hat{f}) + (f - \bar{f}_n)L_n^{(1)}(\bar{f}_n) + \frac{1}{2}(f - \bar{f}_n)^2 L_n^{(2)}(\bar{f}_n) + \frac{1}{6}(f - \bar{f}_n)^3 L_n^{(3)}(\bar{f}_n) + \mathcal{O}\left((f - \bar{f}_n)^4\right),$$

where $L_n^{(1)}(\bar{f}_n) = 0$, and using the Taylor expansion of $\exp(y)$ around 0 for the higher order terms we obtain

$$\exp(nL_n(f)) = \exp(nL_n(\bar{f}_n))\exp\left(\frac{n}{2}(f - \bar{f}_n)^2 L_n^{(2)}(\bar{f}_n)\right) \times \left(1 + \frac{n}{6}(f - \bar{f}_n)^3 L_n^{(3)}(\bar{f}_n) + n\mathcal{O}\left((f - \bar{f}_n)^4\right)\right).$$

For the term with odd derivative it is easy to see that the integral is zero

$$\int_{(\bar{f}_n-\delta,\bar{f}_n+\delta)} \frac{n}{6}(f - \bar{f}_n)^3 L_n^{(3)}(\bar{f}_n)\exp\left(\frac{n}{2}(f - \bar{f}_n)^2 L_n^{(2)}(\bar{f}_n)\right)df = 0$$

The big-$\mathcal{O}$ term coming from the residual of the expansion can be neglected since by definition there is a $C \geqslant 0$ such that

$$\begin{aligned}
\int_{(\bar{f}_n-\delta,\bar{f}_n+\delta)} n\mathcal{O}\left((f-\bar{f}_n)^4\right)\exp\left(\frac{n}{2}(f-\bar{f}_n)^2L_n^{(2)}(\bar{f}_n)\right)df &\leqslant 2Cn\int_{\bar{f}_n}^{\infty}(f-\bar{f}_n)^4\exp\left(-\frac{n}{2}(f-\bar{f}_n)^2|L_n^{(2)}(\bar{f}_n)|\right)df \\
&= Cn\int_0^{\infty}u^3\exp\left(-\frac{n}{2}u^2|L_n^{(2)}(\bar{f}_n)|\right)d(u^2) \\
&= Cn\int_0^{\infty}u^{5/2-1}\exp\left(-\frac{n}{2}u|L_n^{(2)}(\bar{f}_n)|\right)du \\
&= Cn^{-3/2}\frac{\Gamma(5/2)}{|L_n^{(2)}(\bar{f}_n)|} \\
&\leqslant Cn^{-3/2}\frac{\Gamma(5/2)}{m} \\
&= \mathcal{O}(n^{-3/2}).
\end{aligned}$$

The second derivative term can be approximated as follows:

$$\begin{aligned}
\int_{(\bar{f}_n-\delta,\bar{f}_n+\delta)}\exp\left(\frac{n}{2}(f-\bar{f}_n)^2L_n^{(2)}(\bar{f}_n)\right)df &= \int_{-\infty}^{\infty}\exp\left(\frac{n}{2}(f-\bar{f}_n)^2L_n^{(2)}(\bar{f}_n)\right)df + \mathcal{O}\left(e^{-(n-1)\frac{\delta^2}{8\sigma^2}}\right) \\
&= \frac{\sqrt{2\pi}\sigma_{L_n}(\bar{f}_n)}{\sqrt{n}} + \mathcal{O}\left(e^{-(n-1)\frac{\delta^2}{16\sigma^2}}\right).
\end{aligned}$$

Since the last term is exponentially small on n for fixed δ putting everything together we get the final claim:

$$\int_{\mathbb{R}}\exp(nL_n(f))df = \frac{\sqrt{2\pi}\sigma_{L_n}(\bar{f}_n)}{\sqrt{n}}\exp(nL_n(\bar{f}_n))(1+\mathcal{O}(n^{-3/2}))$$

□

Proof of (14). Observe first that the conditions of Lemma 2 hold for a large class of sufficiently smooth and bounded (possibly improper) priors $l(f)$ and for the data generated according to the sampling scheme in (11), since under these assumptions the number of samples in each cluster $[k]$ tends to infinity as $N \to \infty$ a.s. and the sample mean converges almost surely (a.s.), *i.e.*, $\hat{f}_k = \lim_{n\to\infty}\hat{f}_{kn}$ a.s. From Lemma 2, the sequence $(\bar{f}_{kn})_{n=1}^{\infty}$ is convergent for every cluster $[k]$, and $\hat{f}_k = \lim_{n\to\infty}\hat{f}_{kn} = \lim_{n\to\infty}\bar{f}_{kn} = \bar{f}_k$ a.s. Thus, by continuity,

$$\begin{aligned}
\ln\left(\sigma_{L_n}(\bar{f}_{kn})\exp(nL_n(\bar{f}_{kn}))(1+\mathcal{O}(n^{-3/2}))\right) &= \ln\left(\sigma_{L_n}(\hat{f}_{kn})\exp(nL_n(\hat{f}_{kn}))(1+\mathcal{O}(n^{-3/2}))+\mathcal{O}(1))\right) \\
&= \ln(\sigma_{L_n}(\hat{f}_{kn})) + nL_n(\hat{f}_{kn}) + \ln(1+\mathcal{O}(n^{-3/2}))+\mathcal{O}(1)),
\end{aligned}$$

with $\sigma^2_{L_n}(f) = \frac{1}{|L_n''(f,y^n)|} = \frac{\sigma^2}{1+\frac{\sigma^2}{n}(\ln(l(f_n)))^{(2)}}$. From Lemma 2 we can rewrite the logarithm of the posterior distribution

in (13) as follows:

$$\begin{aligned}
\ln p_{m/Y} &= \ln p_m - \sum_{k=1}^{d'+1} \frac{|[k]|}{2} \ln(2\pi\sigma^2) - \sum_{k=1}^{d'+1} \frac{1}{2} \ln |[k]| + \frac{1}{2} \sum_{k=1}^{d'+1} \ln \sigma_{L_n}^2 + (d'+1) \ln \sqrt{2\pi} \\
&\quad - \sum_{k=1}^{d'+1} \frac{\|y_{[k]} - \hat{f}_k e^k\|_2^2}{2\sigma^2} + \sum_{k=1}^{d'+1} \ln \left(1 + o\left(\frac{1}{|[k]|}\right)\right) + \mathcal{O}(d'_m) \\
&= \ln p_m - \frac{N}{2} \ln(2\pi\sigma^2) - \sum_{k=1}^{d'+1} \frac{1}{2} \ln |[k]| - \frac{1}{2} \sum_{k=1}^{d'+1} \ln \left(1 + \frac{\sigma^2}{|[k]|} \ln l(f_k)\right)^{(2)} + (d'+1) \ln(\sqrt{2\pi}\sigma) \\
&\quad - \frac{\|y - \hat{f}_m\|_2^2}{2\sigma^2} + \sum_{k=1}^{d'+1} \ln \left(1 + o\left(\frac{1}{|[k]|}\right)\right) + \mathcal{O}(d'_m) \\
&= \ln C_N + \ln p_m - \sum_{k=1}^{d'+1} \frac{1}{2} \ln |[k]| - \frac{1}{2} \sum_{k=1}^{d'+1} \ln \left(1 + \mathcal{O}\left(\frac{1}{|[k]|}\right)\right) + (d'+1) \ln(\sqrt{2\pi}\sigma) \\
&\quad - \frac{\|y - \mathbb{P}_{\mathcal{F}_m} y\|_2^2}{2\sigma^2} + \sum_{k=1}^{d'+1} \ln \left(1 + o\left(\frac{1}{|[k]|}\right)\right) + \mathcal{O}(d'_m),
\end{aligned}$$

where C_N is constant depending only on N. If we want to maximize the likelihood we would be interested in finding $m \in \mathcal{M}$ such that for all $m' \in \mathcal{M}$ we have $\ln \frac{p_m}{p_{m'}} \geqslant 0$. Since $\sum_{k=1}^{d'+1} \frac{1}{|[k]|} \leqslant d'+1$ and $\ln(1+x) = x + o(x)$, this leads to

$$\begin{aligned}
\ln \frac{p_{m/Y}}{p_{m'/Y}} &= \ln \frac{p_m}{p_{m'}} + (d'_m - d'_{m'}) \ln(\sqrt{2\pi}\sigma) - \frac{1}{2} \left(\sum_{k=1}^{d'_m+1} \ln |[k_m]| - \sum_{k=1}^{d'_{m'}+1} \ln |[k_{m'}]| \right) \\
&\quad - \frac{\|y - \mathbb{P}_{\mathcal{F}_m} y\|_2^2 - \|y - \mathbb{P}_{\mathcal{F}_{m'}} y\|_2^2}{2\sigma^2} + \mathcal{O}(d'_m + d'_{m'}),
\end{aligned}$$

hence maximizing the likelihood is equivalent to minimizing:

$$\mathrm{Crit}_{\mathrm{MAP}}(m) = \ln \frac{1}{p_m} + \frac{1}{2} \sum_{k=1}^{d'_m+1} \ln |[k_m]| + \frac{\|y - \mathbb{P}_{\mathcal{F}_m} y\|_2^2}{2\sigma^2} + \mathcal{O}(d'_m).$$

To avoid the dependency of the criterion on the number of elements in each cluster we observe that

$$\sum_{k=1}^{d'+1} \ln |[k]| = (d'+1) \ln(\prod_{k=1}^{d'+1} |[k]|^{\frac{1}{d'+1}}) \leqslant (d'+1) \ln \frac{N}{d'},$$

from the arithmetic-geometric mean inequality, so we have that

$$0 \leqslant (d'+1) \ln \frac{N}{d'} - (d'+1) \ln(\prod_{k=1}^{d'+1} |[k]|^{\frac{1}{d'+1}}) \leqslant (d'+1) \ln \frac{N}{d'},$$

Thus for all $K \geqslant 1$ we obtain the following upper bound, which corresponds to (14):

$$\mathrm{Crit}_{\mathrm{MAP}}(m) \leqslant \frac{\|y - \mathbb{P}_{\mathcal{F}_m} y\|_2^2}{2\sigma^2} + K \left(\ln \frac{1}{p_m} + \frac{1}{2}(d'_m + 1) \ln \frac{N}{d'_m} \right) + \mathcal{O}(d'_m).$$

□

Appendix D: Oracle Inequality and Upper Bound on the Risk

Lemma 3. *There exists a sequence $(B_N)_{N \in \mathbb{N}}$ for which for all $m \in \mathcal{M}$*

$$p_m = \frac{\exp(-d'_m - d''_m)}{B_N S^2(d''_m + 1, d'_m + 1) C_{d''_m}^N}, \tag{19}$$

is a valid probability mass function on $\mathcal{M}$, and this sequence satisfies the following bounds:

$$\frac{e^3}{(e-1)^2(e+1)}(1-3e^{-N-1}) \leqslant B_N \leqslant \frac{e^3}{(e-1)^2(e+1)}.$$

Proof. Fixing d' and d'', we can make a model π_m by first choosing a subset of cardinality d'' from $\{1,2,\dots,N\}$; there are $C^N_{d''}$ ways to make this choice. This leaves us with $d''+1$ segments. We then partition the segments into exactly $d'+1$ parts, which corresponds to taking a partition into $d'+1$ parts of $\{1,2,\dots,d''+1\}$, where the distance between any two elements of the same part is at least two. This can be done in $S^2(d''+1,d'+1)$ ways. Thus if we let $\mathcal{A}(d',d'') = \{\pi_m \in \Pi_N \colon |\pi_m| = d' \text{ and } |\pi_m|_0 = d''\}$, we have that

$$|\mathcal{A}(d',d'')| = S^2(d''+1,d'+1)C^N_{d''}.$$

Since $S^2(\cdot,\cdot)$, $C^N_{d''_m}$ and the exponential are all non-negative, for $p=(p_m)_{m\in\mathcal{M}}$ to be a valid probability mass function we only need to find a positive sequence $(B_N)_{N\in\mathbb{N}}$ such that p_m sum to 1. Now,

$$\begin{aligned}
\sum_{m\in\mathcal{M}} p_m &= (B_N)^{-1}\sum_{d'=0}^{N-1}\sum_{d''=d'}^{N}\sum_{m\in\mathcal{A}(d',d'')}\left(S^2(d''+1,d'+1)C^N_{d''}\right)^{-1}\exp(-d'-d'') \\
&= (B_N)^{-1}\sum_{d'=0}^{N-1}\sum_{d''=d'}^{N}|\mathcal{A}(d',d'')|\left(S^2(d''+1,d'+1)C^N_{d''}\right)^{-1}\exp(-d'-d'') \\
&= (B_N)^{-1}\sum_{d'=0}^{N-1}\exp(-d')\sum_{d''=d'}^{N}\exp(-d'') \\
&= (B_N)^{-1}\sum_{d'=0}^{N-1}\exp(-d')\frac{e^{-d'}-e^{-N-1}}{1-e^{-1}} \\
&= (B_N)^{-1}\frac{1}{1-e^{-1}}\left(\frac{1-e^{-2N}}{1-e^{-2}} - \frac{e^{-N-1}-e^{-2N-1}}{1-e^{-1}}\right) \\
&= (B_N)^{-1}\frac{e^3}{(e-1)^2(e+1)}\left(1-e^{-2N}-e^{-N-1}+e^{-2N-1}-e^{-N-2}+e^{-2N-2}\right) \\
&= 1,
\end{aligned}$$

which holds for the choice $B_N = \frac{e^3}{(e-1)^2(e+1)}\left(1-e^{-2N}-e^{-N-1}+e^{-2N-1}-e^{-N-2}+e^{-2N-2}\right)$. The bounds on B_N are easy to verify. □

The choice of probability mass function (p_m) in Lemma 3 distributes the mass evenly among models of the same dimensions. On the other hand, to balance the exponential increase in the number of models the exponential factor makes p_m decrease exponentially with the dimensions. Together with the fact that the prior $l_{Y/m}$ was absorbed in the error term of the approximation in Lemma 2, we obtain a set of what can be considered as least favorable priors for our Bayesian setting, and this will have the effect of reducing the upper bound on the risk.

Lemma 4. *For $1 \leqslant k \leqslant N-1$ and $N \geqslant 4$, with the convention $0\ln 0 = 0$, the following bounds hold for the binomial coefficients:*

$$\left(\frac{N}{k}\right)^k \leqslant C^N_k \leqslant \left(\frac{Ne}{k}\right)^k,$$

and the following bounds hold for the Stirling numbers of the second kind:

$$k^{N-k} \leqslant S(N,k) \leqslant \frac{1}{2}(Ne)^k k^{N-2k}.$$

In particular, we have that

$$d''_m\ln[d''_m e] - d'_m\ln\frac{d''_m}{e} + d''_m\ln\frac{N}{d''_m} \leqslant \ln\frac{1}{p_m} \leqslant d'_m\ln[d''_m e^2] + d''_m\ln[d'_m e^2] + d''_m\ln\frac{N}{d''_m}.$$

Proof. For $1 \leqslant l < k \leqslant N$ we have

$$\frac{N}{k} \leqslant \frac{N-l}{k-l} \leqslant \frac{N}{k-l}.$$

Hence,

$$\left(\frac{N}{k}\right)^k \leqslant \frac{N(N-1)\dots(N-l+1)}{k(k-1)\dots 1} = C_k^N \leqslant \frac{N^k}{k!}.$$

On the other hand we have that

$$\begin{aligned} k\ln k - k &= \int_1^k \ln x dx - 1 \\ &= \sum_{l=1}^{k-1} \int_l^{l+1} \ln x dx - 1 \\ &\leqslant \sum_{l=1}^{k-1} \ln(l+1) - 1 \\ &\leqslant \ln(k!), \end{aligned}$$

thus we obtain $k\ln N - \ln(k!) \leqslant k(\ln N - \ln k + 1)$, and taking the exponential yields

$$C_k^N \leqslant \frac{N^k}{k!} \leqslant \left(\frac{Ne}{k}\right)^k.$$

The following bound for Stirling numbers of the second kind can be found in (Rennie & Dobson, 1969):

$$\frac{1}{2}(k^2+k+2)k^{N-k-1} - 1 \leqslant S(N,k) \leqslant \frac{1}{2}C_k^N k^{N-k}, \qquad 1 \leqslant k \leqslant N-1.$$

Using the upper bound on the binomial coefficients we can easily derive bounds for the Stirling numbers. In particular, since

$$\begin{aligned} \frac{1}{2}(k^2+k+2)k^{N-k-1} - k^{N-k} &= \frac{1}{2}(k^2+k+2)k^{N-k-1} - \frac{2k}{2}k^{N-k-1} \\ &= \frac{1}{2}(k^2-k+2)k^{N-k-1} \\ &\geqslant 1, \end{aligned}$$

we obtain the lower bound

$$S(N,k) \geqslant \frac{1}{2}(k^2+k+2)k^{N-k-1} - 1 \geqslant k^{N-k}.$$

The upper bound is derived as follows:

$$S(N,k) \leqslant \frac{1}{2}C_k^N k^{N-k} \leqslant \frac{1}{2}(Ne)^k k^{N-2k}.$$

Finally, since $S^2(N+1,k+1) = S(N,k)$, we have from Lemma 3 that

$$\begin{aligned} \ln\frac{1}{p_m} &= \ln\left(B_N S^2(d_m''+1, d_m'+1) C_{d_m''}^N \exp(d_m'+d_m'')\right) \\ &= \ln B_N + \ln S(d_m'', d_m') + \ln C_{d_m''}^N + d_m' + d_m'' \\ &\leqslant d_m' \ln[d_m'' e^2] + d_m'' \ln[d_m' e^2] + d_m'' \ln\frac{N}{d_m''} \end{aligned}$$

and

$$\ln \frac{1}{p_m} \geqslant \ln \frac{e^3}{(e-1)^2(e+1)} + \ln(1-3e^{-N-1}) + (d''_m - d'_m)\ln d''_m + d'_m \ln \frac{N}{d'_m} + d'_m + d''_m$$
$$\geqslant d''_m \ln[d''_m e] - d'_m \ln \frac{d''_m}{e} + d''_m \ln \frac{N}{d''_m},$$

since $\ln \frac{e^3}{(e-1)^2(e+1)} \approx 0.604$, and for $N \geqslant 4$ we have $\ln(1-3e^{-N-1}) \geqslant \ln(1-3e^{-5}) \approx -0.0204$. This leads to the desired result. □

From this lemma, the penalty term in (15) behaves like $d''_m \ln \frac{N}{d''_m}$, which would correspond to the behavior of the first part when the dimensions are close but would penalize more those models with $d'_m \ll d''_m$.

Theorem 6.1 (Oracle inequality for $\hat{f}_{\hat{m}}$)**.** *With $\mathcal{M}$ restricted to models such that $ed'_m \leqslant N$ and for the choice of $K = 3a$, p_m as in 3, $\mathrm{pen}(m)$ as in 15 and $\hat{m} \in \mathcal{M}$ corresponding to*

$$\hat{m} \in \arg\min_{m\in\mathcal{M}} \|y - \hat{f}_m\|_2^2 + \sigma^2 K\,\mathrm{pen}(m), \tag{20}$$

We obtain for all $a > 1$,

$$E_{f^*}[\|\mathbb{P}_{\mathcal{F}_{\hat{m}}} Y - f^*\|^2] \leqslant \arg\min_{m\in\mathcal{M}} \left\{ \frac{a}{a-1}\mathbb{E}_{f^*}[\|\mathbb{P}_{\mathcal{F}_m} Y - f^*\|^2] + \frac{a^2\sigma^2}{a-1}\left(7 + 3(d'_m+1)\ln\frac{N}{d'_m} + 6\ln\frac{1}{p_m}\right)\right\}. \tag{21}$$

The proof of this theorem follows the line of reasoning described in (Massart, 2003). To establish this result we rely on the Gaussian concentration inequality stated below and whose proof can be found in (Cirel'son et al., 1976; Boucheron et al., 2013):

Lemma 5 (Tsirelson-Ibragimov-Sudakov Inequality)**.** *Assume that $F\colon \mathbb{R}^d \to \mathbb{R}$ is 1-Lipschitz and X is a Gaussian random vector following $\mathcal{N}(0, \sigma^2 I)$. Then, there exists a variable $\xi \sim exp(1)$ following an exponential distribution of parameter 1, such that*

$$F(X) \leqslant \mathbb{E}[F(X)] + \sigma\sqrt{2\xi}.$$

Proof of Theorem 6.1. By definition of $\hat{m}$ we have that, for all $m \in \mathcal{M}$,

$$\|y - \hat{f}_{\hat{m}}\|_2^2 + \sigma^2 K\,\mathrm{pen}(\hat{m}) \leqslant \|y - \hat{f_m}\|_2^2 + \sigma^2 K\,\mathrm{pen}(m).$$

By expanding the squares and using $Y = f^* + \epsilon$ we obtain

$$\|y - \hat{f}_m\|_2^2 = \|f^* - \hat{f}_m\|_2^2 + 2\langle \epsilon, f^* - \hat{f}_m\rangle + \|\epsilon\|_2^2.$$

On the other hand, we have by expanding the squares,

$$\|f^* - \hat{f}_{\hat{m}}\|_2^2 \leqslant \|f^* - \hat{f}_m\|_2^2 + 2\langle \epsilon, f^* - \hat{f}_m\rangle - 2\langle \epsilon, f^* - \hat{f}_{\hat{m}}\rangle + \sigma^2 K\,\mathrm{pen}(m) - \sigma^2 K\,\mathrm{pen}(\hat{m}). \tag{22}$$

The rest of the proof will consist in upper bounding the expected value of the terms of the right hand side of (22).

Again, since $Y = f^* + \epsilon$ we also have, for all $m \in \mathcal{M}$, that

$$\hat{f}_m = \mathbb{P}_{\mathcal{F}_m} Y = \mathbb{P}_{\mathcal{F}_m} f^* + \mathbb{P}_{\mathcal{F}_m} \epsilon. \tag{23}$$

We can use Lemma 1 to derive a simple bound on $\mathbb{E}[\langle \epsilon, f^* - \hat{f}_m\rangle]$ as follows:

$$\begin{aligned}
\mathbb{E}[\langle \epsilon, f^* - \hat{f}_m\rangle] &= -\mathbb{E}[\langle \epsilon, \hat{f}_m\rangle] \\
&= -\mathbb{E}[\langle \epsilon, \mathbb{P}_{\mathcal{F}_m} f^* + \mathbb{P}_{\mathcal{F}_m} \epsilon\rangle] \\
&= -\mathbb{E}[\langle \epsilon, \mathbb{P}_{\mathcal{F}_m} \epsilon\rangle] \\
&= -\mathbb{E}[\|\mathbb{P}_{\mathcal{F}_m} \epsilon\|_2^2] \\
&= -\sigma^2 d'_m \leqslant 0,
\end{aligned}$$

so we can discard this term since it has a negative contribution of small order on the bound.

To bound $2\langle \epsilon, \hat{f}_{\hat{m}} - f^* \rangle$ we use Young's inequality $2\langle u, v\rangle \leqslant a\|u\|_2^2 + \frac{1}{a}\|v\|_2^2$ for all $a > 0$ as follows:

$$\begin{aligned}
2\langle \epsilon, \hat{f}_{\hat{m}} - f^* \rangle &= 2\langle \mathbb{P}_{\mathcal{F}_{\hat{m}} \oplus \langle f^* \rangle} \epsilon, \hat{f}_{\hat{m}} - f^* \rangle \\
&= 2\langle \mathbb{P}_{\mathcal{F}_{\hat{m}} \ominus \langle f^* \rangle} \epsilon + \mathbb{P}_{\langle f^* \rangle} \epsilon, \hat{f}_{\hat{m}} - f^* \rangle \\
&\leqslant a\| \mathbb{P}_{\mathcal{F}_{\hat{m}} \ominus \langle f^* \rangle} \epsilon + \mathbb{P}_{\langle f^* \rangle} \epsilon\|_2^2 + \frac{1}{a}\|\hat{f}_{\hat{m}} - f^*\|_2^2 \\
&= a(\| \mathbb{P}_{\mathcal{F}_{\hat{m}} \ominus \langle f^* \rangle} \epsilon\|_2^2 + \| \mathbb{P}_{\langle f^* \rangle} \epsilon\|_2^2) + \frac{1}{a}\|\hat{f}_{\hat{m}} - f^*\|_2^2, \qquad a > 0.
\end{aligned}$$

This gives

$$2\langle \epsilon, \hat{f}_{\hat{m}} - f^* \rangle - \frac{1}{a}\|\hat{f}_{\hat{m}} - f^*\|_2^2 \leqslant a\sigma^2(\| \mathbb{P}_{\mathcal{F}_{\hat{m}} \ominus \langle f^* \rangle} \epsilon\|_2^2/\sigma^2 + \| \mathbb{P}_{\langle f^* \rangle} \epsilon\|_2^2/\sigma^2). \tag{24}$$

Since $\| \mathbb{P}_{\langle f^* \rangle} \epsilon\|_2^2/\sigma^2$ follows a χ_1^2 distribution,

$$\mathbb{E}[\| \mathbb{P}_{\langle f^* \rangle} \epsilon\|_2^2/\sigma^2] = 1. \tag{25}$$

Similarly, for all $m \in \mathcal{M}$, $\| \mathbb{P}_{\mathcal{F}_m \ominus \langle f^* \rangle} \epsilon\|_2^2/\sigma^2$ follows a $\chi^2_{\bar{d}_m}$ distribution, where

$$\bar{d}_m := \dim(\mathcal{F}_m \ominus \langle f^* \rangle) = \begin{cases} d'_m, & \text{if } f^* \in \mathcal{F}_m, \\ d'_m + 1, & \text{otherwise.} \end{cases}$$

Thus,

$$\mathbb{E}[\| \mathbb{P}_{\mathcal{F}_m \ominus \langle f^* \rangle} \epsilon\|_2^2/\sigma^2] = \bar{d}_m \leqslant d'_m + 1. \tag{26}$$

We now use a maximal inequality to bound $\mathbb{E}(a\| \mathbb{P}_{\mathcal{F}_{\hat{m}} \ominus \langle f^* \rangle} \epsilon\|_2^2 + \sigma^2 K \operatorname{pen}(\hat{m}))$:

$$\begin{aligned}
\mathbb{E}\left[\frac{\| \mathbb{P}_{\mathcal{F}_{\hat{m}} \ominus \langle f^* \rangle} \epsilon\|_2^2}{\sigma^2} - \frac{K}{a}\operatorname{pen}(\hat{m})\right] &\leqslant \mathbb{E}\left[\max_{m \in \mathcal{M}} \frac{\| \mathbb{P}_{\mathcal{F}_{\hat{m}} \ominus \langle f^* \rangle} \epsilon\|_2^2}{\sigma^2} - \frac{K}{a}\operatorname{pen}(\hat{m})\right] \\
&\leqslant \sum_{m \in \mathcal{M}} \mathbb{E}\left[\max\left\{0, \frac{\| \mathbb{P}_{\mathcal{F}_{\hat{m}} \ominus \langle f^* \rangle} \epsilon\|_2^2}{\sigma^2} - \frac{K}{a}\operatorname{pen}(\hat{m})\right\}\right].
\end{aligned} \tag{27}$$

On the other hand, since the norm is 1-Lipschitz, the Gaussian concentration inequality from Lemma 5 implies that there is an exponential random variable $\xi \sim \exp(1)$ such that

$$\begin{aligned}
\| \mathbb{P}_{\mathcal{F}_m \ominus \langle f^* \rangle} \epsilon\|_2/\sigma &\leqslant \mathbb{E}[\| \mathbb{P}_{\mathcal{F}_m \ominus \langle f^* \rangle} \epsilon\|_2/\sigma] + \sqrt{2\xi} \\
&\leqslant (\mathbb{E}[(\| \mathbb{P}_{\mathcal{F}_m \ominus \langle f^* \rangle} \epsilon\|_2/\sigma)^2])^{1/2} + \sqrt{2\xi} \\
&\leqslant \sqrt{d'_m + 1} + \sqrt{2\xi},
\end{aligned}$$

where we used (26) in the last step. Taking the square we obtain

$$\begin{aligned}
\frac{\| \mathbb{P}_{\mathcal{F}_m \ominus \langle f^* \rangle} \epsilon\|_2^2}{\sigma^2} &\leqslant \left(\sqrt{d'_m + 1} + \sqrt{2\xi}\right)^2 \\
&\leqslant \left(\sqrt{d'_m + 1} + \sqrt{2\ln\frac{1}{p_m}} + \sqrt{2\max\left\{0, \xi - \ln\frac{1}{p_m}\right\}}\right)^2 \\
&\leqslant \left(\sqrt{(d'_m + 1)\ln\frac{N}{d'_m}} + \sqrt{2\ln\frac{1}{p_m}} + \sqrt{2\max\left\{0, \xi - \ln\frac{1}{p_m}\right\}}\right)^2 \\
&\leqslant 3(d'_m + 1)\ln\frac{N}{d'_m} + 6\ln\frac{1}{p_m} + 6\max\left\{0, \xi - \ln\frac{1}{p_m}\right\} \\
&= 3\operatorname{pen}(m) + 6\max\left\{0, \xi - \ln\frac{1}{p_m}\right\},
\end{aligned}$$

where we used the inequalities $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ in the second step, the assumption $N \geqslant e d'_m$ in the third step and the inequality $(a+b+c)^2 \leqslant 3a^2 + 3b^2 + 3c^2$ in the fourth step. Since the second term in the last line is nonnegative, this implies also that

$$\max\left\{0, \frac{\| \mathbb{P}_{\mathcal{F}_m \ominus \langle f^* \rangle} \epsilon\|_2^2}{\sigma^2} - 3\,\mathrm{pen}(m)\right\} \leqslant 6 \max\left\{0, \xi - \ln\frac{1}{p_m}\right\}.$$

On the other hand, we have that

$$\begin{aligned}
\mathbb{E}\left[\max\left\{0, \xi - \ln\frac{1}{p_m}\right\}\right] &= \int_0^\infty \max\left\{0, t - \ln\frac{1}{p_m}\right\} e^{-t} dt \\
&= \int_{\ln\frac{1}{p_m}}^\infty \left(t - \ln\frac{1}{p_m}\right) e^{-t} dt \\
&= -\left[\left(t + 1 - \ln\frac{1}{p_m}\right) e^{-t}\right]_{t=\ln\frac{1}{p_m}}^\infty \\
&= p_m,
\end{aligned}$$

hence taking $K = 3a$ in (27) yields

$$\begin{aligned}
\mathbb{E}\left[\frac{\| \mathbb{P}_{\mathcal{F}_{\hat{m}} \ominus \langle f^* \rangle} \epsilon\|_2^2}{\sigma^2} - \frac{K}{a}\,\mathrm{pen}(\hat{m})\right] &\leqslant \sum_{m \in \mathcal{M}} \mathbb{E}\left[\max\left\{0, \frac{\| \mathbb{P}_{\mathcal{F}_{\hat{m}} \ominus \langle f^* \rangle} \epsilon\|_2^2}{\sigma^2} - 3\,\mathrm{pen}(\hat{m})\right\}\right] \\
&\leqslant 6 \sum_{m \in \mathcal{M}} \mathbb{E}\left[\max\left\{0, \xi - \ln\frac{1}{p_m}\right\}\right] \\
&\leqslant 6 \sum_{m \in \mathcal{M}} p_m \\
&\leqslant 6.
\end{aligned}$$

Combining the last result and equations (24), (25) and (27) brings us to

$$\begin{aligned}
\mathbb{E}\left[2\langle \epsilon, \hat{f}_{\hat{m}} - f^* \rangle - \frac{1}{a}\|\hat{f}_{\hat{m}} - f^*\|_2^2 - 3a\sigma^2\,\mathrm{pen}(\hat{m})\right] &\leqslant a\sigma^2 \mathbb{E}\left[\frac{\| \mathbb{P}_{\mathcal{F}_{\hat{m}} \ominus \langle f^* \rangle} \epsilon\|_2^2}{\sigma^2} + \frac{\| \mathbb{P}_{\langle f^* \rangle} \epsilon\|_2^2}{\sigma^2} - 3\,\mathrm{pen}(\hat{m})\right] \\
&\leqslant a\sigma^2 \sum_{m \in \mathcal{M}} \mathbb{E}\left[\max\left\{0, \frac{\| \mathbb{P}_{\mathcal{F}_m \ominus \langle f^* \rangle} \epsilon\|_2^2}{\sigma^2} - 3\,\mathrm{pen}(m)\right\}\right] + a\sigma^2 \\
&\leqslant 7a\sigma^2.
\end{aligned}$$

Going back to (22), substituting K by its value and subtracting $\frac{1}{a}\|\hat{f}_{\hat{m}} - f^*\|_2^2$ from both sides we obtain

$$\begin{aligned}
&\frac{a-1}{a}\| \mathbb{P}_{\mathcal{F}_{\hat{m}}} Y - f^*\|_2^2 \leqslant \\
&\quad \|f^* - \mathbb{P}_{\mathcal{F}_m} Y\|_2^2 - 2\langle \epsilon, \mathbb{P}_{\mathcal{F}_m} Y - f^* \rangle + 2\langle \epsilon, \mathbb{P}_{\mathcal{F}_{\hat{m}}} Y - f^* \rangle - \frac{1}{a}\| \mathbb{P}_{\mathcal{F}_{\hat{m}}} Y - f^*\|_2^2 - 3a\sigma^2\,\mathrm{pen}(\hat{m}) + 3a\sigma^2\,\mathrm{pen}(m).
\end{aligned}$$

Taking the minimum of this expression over $m \in \mathcal{M}$ and the expectation, and omitting negative terms we obtain the desired result for all $a > 1$:

$$E_{f^*}[\| \mathbb{P}_{\mathcal{F}_{\hat{m}}} Y - f^*\|^2] \leqslant \arg\min_{m \in \mathcal{M}} \left\{\frac{a}{a-1}\mathbb{E}_{f^*}[\| \mathbb{P}_{\mathcal{F}_m} Y - f^*\|^2] + \frac{a^2\sigma^2}{a-1}\left(7 + 3(d'_m + 1)\ln\frac{N}{d'_m} + 6\ln\frac{1}{p_m}\right)\right\}.$$

□

Corollary 6.1. *For the set of models described in (1) with $f^* \in \mathcal{F}_{m^*}$ the following properties hold:*

- *Adaptation and Risk Upper bound: The following adaptive upper bound in terms of d'_{m^*} and d''_{m^*} holds for $a = 2$:*

$$E_{f^*}[\| \mathbb{P}_{\mathcal{F}_{\hat{m}}} Y - f^*\|^2] \leqslant 4\sigma^2\left(7 + 3(d'_{m^*} + 1)\ln\frac{N}{d'_{m^*}} + 6\left(d'_{m^*}\ln[d''_{m^*} e^{\frac{13}{6}}] + d''_{m^*}\ln[d'_{m^*} e^2] + d''_{m^*}\ln\frac{N}{d''_{m^*}}\right)\right).$$

- *Consistency: If* $d''_{m^*} = o(N/\ln N)$*, then* $\lim_{N\to\infty} N^{-1}\mathbb{E}_{f^*}[\|\hat{f}_{\hat{m}} - f^*\|^2] = 0$.

Proof. Equation (21) implies in particular that for m^* such that $f^* \in \mathcal{F}_{m^*}$ we obtain

$$E_{f^*}[\|\mathbb{P}_{\mathcal{F}_{\hat{m}}} Y - f^*\|^2] \leqslant \frac{a}{a-1}\mathbb{E}_{f^*}[\|\mathbb{P}_{\mathcal{F}_{m^*}} Y - f^*\|^2] + \frac{a^2\sigma^2}{a-1}\left(7 + 3(d'_{m^*}+1)\ln\frac{N}{d'_{m^*}} + 6\ln\frac{1}{p_{m^*}}\right).$$

To simplify the first expectation of the right hand side, we can use (23) and Lemma **??** to obtain

$$\mathbb{E}_{f^*}[\|\mathbb{P}_{\mathcal{F}_{m^*}} Y - f^*\|^2] = \mathbb{E}_{f^*}[\|\mathbb{P}_{\mathcal{F}_{m^*}} \epsilon\|^2] = \sigma^2 d'_{m^*}.$$

Then, using Lemma 4 we can upper bound the rest of the right hand side, and choosing $a = \frac{C}{C-1}$ yieds, for all $C > 1$,

$$\begin{aligned}
E_{f^*}[\|\mathbb{P}_{\mathcal{F}_{\hat{m}}} Y - f^*\|^2] &\leqslant C\sigma^2 d'_{m^*} + \frac{C^2\sigma^2}{C-1}\left(7 + 3(d'_{m^*}+1)\ln\frac{N}{d'_{m^*}} + 6\left(d'_{m^*}\ln[d''_{m^*}e^2] + d''_{m^*}\ln[d'_{m^*}e^2] + d''_{m^*}\ln\frac{N}{d''_{m^*}}\right)\right) \\
&= \frac{C^2\sigma^2}{C-1}\left(\left(1-\frac{1}{C}\right)d'_{m^*} + 7 + 3(d'_{m^*}+1)\ln\frac{N}{d'_{m^*}} + 6\left(d'_{m^*}\ln[d''_{m^*}e^2] + d''_{m^*}\ln[d'_{m^*}e^2] + d''_{m^*}\ln\frac{N}{d''_{m^*}}\right)\right) \\
&\leqslant \frac{C^2\sigma^2}{C-1}\left(7 + 3(d'_{m^*}+1)\ln\frac{N}{d'_{m^*}} + 6\left(d'_{m^*}\ln[d''_{m^*}e^{\frac{13}{6}}] + d''_{m^*}\ln[d'_{m^*}e^2] + d''_{m^*}\ln\frac{N}{d''_{m^*}}\right)\right).
\end{aligned}$$

This upper bound achieves a minimum for $C = 2$, yielding the Adaptation and Risk Upper bound result:

$$E_{f^*}[\|\mathbb{P}_{\mathcal{F}_{\hat{m}}} Y - f^*\|^2] \leqslant 4\sigma^2\left(7 + 3(d'_{m^*}+1)\ln\frac{N}{d'_{m^*}} + 6\left(d'_{m^*}\ln[d''_{m^*}e^{\frac{13}{6}}] + d''_{m^*}\ln[d'_{m^*}e^2] + d''_{m^*}\ln\frac{N}{d''_{m^*}}\right)\right).$$

Given that $d'_{m^*} \leqslant d''_{m^*} \leqslant N$, this last equation also implies that

$$E_{f^*}[\|\mathbb{P}_{\mathcal{F}_{\hat{m}}} Y - f^*\|^2] = \mathcal{O}(d'_{m^*}\ln N + d''_{m^*}\ln N) = \mathcal{O}(d''_{m^*}\ln N).$$

Hence, as long as $d''_{m^*} = o(N/\ln N)$, we obtain $\lim_{N\to\infty} N^{-1}\mathbb{E}_{f^*}[\|\hat{f}_{\hat{m}} - f^*\|^2] = 0$, which establishes the Consistency result. □