The Mondrian Kernel Supplementary material

A Proofs

Definition 1. The *linear dimension* of an axis-aligned box $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_D \subseteq \mathbb{R}^D$ is $|\mathcal{X}| := |\mathcal{X}_1| + \cdots + |\mathcal{X}_D|$.

Our first result is a tail bound on the number of partition cells generated by a Mondrian process. We will use it as a Lemma in Proposition 4, but it also confirms that with probability 1, the Mondrian process does not explode (does not generate infinitely many partition cells in finite time).

Proposition 3. *Let $\mathcal{M}$ be a Mondrian process on an axis-aligned box $\mathcal{X}$. For $t \geq 0$, let N_t be the number of partition cells generated by $\mathcal{M}$ until time t. Then*

$$\forall n \in \mathbb{R}_+ \qquad \mathbb{P}[N_t > n] \leq \frac{e^{|\mathcal{X}|t}}{n}.$$

In particular, the Mondrian process does not explode.

Proof. At any time s, by lack of memory of the exponential distribution, the residual time until a partition cell c splits into two has $\mathrm{Exp}(|c|)$ distribution and is independent of all other cells by construction of the Mondrian process. As $|c| \leq |\mathcal{X}|$, this cell splitting process is dominated by a Yule process with birth rate $|\mathcal{X}|$. The number $\tilde{N}_t$ of individuals at time t of a Yule process with birth rate $|\mathcal{X}|$ has geometric distribution with mean $e^{|\mathcal{X}|t}$ and Markov's inequality yields

$$\mathbb{P}[N_t > n] \leq \mathbb{P}[\tilde{N}_t > n] \leq \frac{e^{|\mathcal{X}|t}}{n}.$$

as claimed. Hence $\mathbb{P}[N_t = \infty] = \lim_{n\to\infty} \mathbb{P}[N_t > n] = 0$ for any t. □

We define an ε-grid covering a (closed) interval as a set of points at most ε distance apart, including the boundary points, and with minimal possible cardinality:

Definition 2. Let $\mathcal{X}_1 = [a_1, b_1]$ be an interval of length $|\mathcal{X}_1| = b_1 - a_1$ and let $0 < \varepsilon < |\mathcal{X}_1|$. Define $K := \lceil \frac{|\mathcal{X}_1|}{\varepsilon} \rceil$. An *$\varepsilon$-grid* covering $\mathcal{X}_1$ is a set $\mathcal{U}_1$ of $K+1$ points $u_0 < u_1 < \cdots < u_K$ in $\mathcal{X}_1$ such that $u_0 = a_1$, $u_K = b_1$ and $|u_i - u_{i-1}| \leq \varepsilon$ for all $1 \leq i \leq K$.

Note that such an ε-grid exists by our choice of K, as we can take, e.g., $u_i = i\varepsilon$ for $1 \leq i < K$. The next lemma bounds the probability that two arrivals of a Poisson process running on a bounded interval occur between two consecutive points of an ε-grid covering that interval.

Lemma 1. *Consider a Poisson process with rate λ running on a bounded interval $[0, L]$. Let $\mathcal{U}$ be an ε-grid covering of $[0, L]$. Then the probability that two or more arrivals of the process occur between any two consecutive points of $\mathcal{U}$ is at most $2\lambda^2 L\varepsilon$.*

Proof. As the distance between any two consecutive points of the ε-grid is at most ε by definition, the number of arrivals in a line segment between such two points is dominated by a Poisson random variable with mean $\lambda\varepsilon$. As there are $\lceil \frac{L}{\varepsilon} \rceil$ such segments, the sought probability p can be upper bounded using a union bound as

$$p \leq \left\lceil \frac{L}{\varepsilon} \right\rceil \left(1 - e^{-\lambda\varepsilon} - e^{-\lambda\varepsilon}\lambda\varepsilon\right)$$

and using $1 - e^{-x} \leq x$ twice, we obtain as claimed

$$p \leq \left\lceil \frac{L}{\varepsilon} \right\rceil \left(\lambda\varepsilon - e^{-\lambda\varepsilon}\lambda\varepsilon\right) \leq \left\lceil \frac{L}{\varepsilon} \right\rceil (\lambda\varepsilon)^2 \leq 2L\lambda^2\varepsilon. \quad \square$$

Definition 2 also set us up for defining the concept of an ε-grid on higher-dimensional axis-aligned boxes:

Definition 3. Let $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_D \subseteq \mathbb{R}^D$ be an axis-aligned box and let $\varepsilon > 0$. An *ε-grid* covering $\mathcal{X}$ is a cartesian product $\mathcal{U} = \mathcal{U}_1 \times \cdots \times \mathcal{U}_D$, where each $\mathcal{U}_d$ is an ε-grid covering of $\mathcal{X}_d$ in the sense of Definition 2.

Proposition 4. *For any bounded input domain $\mathcal{X} \subseteq \mathbb{R}^D$ and $\delta > 0$, as $M \to \infty$,*

$$\mathbb{P}\left[\sup_{\mathbf{x},\mathbf{x}' \in \mathcal{X}} |k_M(\mathbf{x}, \mathbf{x}') - k_\infty(\mathbf{x}, \mathbf{x}')| > \delta\right] = \mathcal{O}\left(M^{2/3} e^{-M\delta^2/(12D+2)}\right).$$

Proof. By extending $\mathcal{X}$ if necessary, we may assume without loss of generality that $\mathcal{X}$ is an axis-aligned box with linear dimension $|\mathcal{X}|$.

Recall that a Mondrian kernel of order M corresponds to a random features obtained from M independent Mondrians with lifetime λ. Let $\mathcal{U}$ be an ε-grid covering $\mathcal{X}$, where $\varepsilon > 0$ will be specified later. The proof will upper bound the probability of the following three "bad" events:

A_1 := { any of the M Mondrian samples contains more than n partition cells }
A_2 := { the common refinement of the M Mondrian partitions, disregarding any potential cuts after generating n cells in one Mondrian, has a partition cell that does not contain an element of $\mathcal{U}$ }
A_3 := { $\frac{\delta}{2}$-approximation fails on $\mathcal{U}$, i.e., for some $\mathbf{u}_1$, $\mathbf{u}_2 \in \mathcal{U}$, $|k_M(\mathbf{u}_1, \mathbf{u}_2) - k_\infty(\mathbf{u}_1, \mathbf{u}_2)| > \frac{\delta}{2}$ }

The constant $n \in \mathbb{R}+$ will be specified (optimized) later. Note that $A_1 \cap A_2$ implies that all partition cells in the common refinement of all M Mondrian partitions contain a grid point from $\mathcal{U}$. Since k_M is constant in each such cell, making ε small enough, smoothness of the Laplace kernel k_∞ will ensure that if A_3^c holds then δ-approximation holds throughout $\mathcal{X}$.

Proposition 3 and a union bound over the M Mondrian samples give immediately that

$$\mathbb{P}(A_1) \leq M \frac{e^{|\mathcal{X}|\lambda}}{n}.$$

Note that the ε-grid $\mathcal{U}$ contains at most $(2|\mathcal{X}|/\varepsilon)^D$ grid points. Hoeffding's inequality and a union bound over all pairs of grid points gives for any $\varepsilon > 0$:

$$\mathbb{P}(A_3) \leq \left[\left(2\frac{|\mathcal{X}|}{\varepsilon} \right)^D \right]^2 \left[2\exp\left(-M\delta^2/2 \right) \right].$$

To upper bound the probability of A_2, note that at any time $t < \lambda$, in each partition cell generated so far by any of the M Mondrian processes, an exponential clock is associated to each dimension d of the cell, and if that clock rings, the cell is split at a random location a by a hyperplane lying in dimension d. Consider the point process obtained by projecting the cut points from all partition cells onto their respective coordinate axes. If each Mondrian process generates no more than than n partition cells until its lifetime λ is exhausted, the cut points on the d-th coordinate axis come from at most Mn partition cells, each having width at most $|\mathcal{X}_d|$ in dimension d. Therefore this point process on the d-th coordinate axis can be thought of as taking a suitable subset of points generated by a Poisson point process with intensity $Mn|\mathcal{X}_d|\lambda$. Thus by Lemma 1, the probability that two cut points in dimension d fall between two adjacent coordinates of the ε-grid $\mathcal{U}$ is upper bounded by $2(Mn\lambda)^2|\mathcal{X}_d|\varepsilon$. Observe that if this does not happen in any of the D dimensions then all partition cells in the common refinement must contain a grid point from $\mathcal{U}$. Hence, taking the union bound over all D dimensions,

$$\mathbb{P}(A_2) \leq \sum_{d=1}^{D} 2(Mn\lambda)^2|\mathcal{X}_d|\varepsilon = 2(Mn\lambda)^2|\mathcal{X}|\varepsilon.$$

Thus the probability of a "bad" event occuring is at most

$$\begin{aligned}
&\mathbb{P}(A_1 \cup A_2 \cup A_3) \\
&\leq \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) \\
&\leq M\frac{e^{|\mathcal{X}|\lambda}}{n} + 2(Mn\lambda)^2|\mathcal{X}|\varepsilon + 2\left(2\frac{|\mathcal{X}|}{\varepsilon} \right)^{2D} e^{-M\delta^2/2}.
\end{aligned}$$

and minimizing over $n \in \mathbb{R}_+$ gives

$$\begin{aligned}
&\mathbb{P}(A_1 \cup A_2 \cup A_3) \\
&\leq \left(4\lambda^2 M^2 |\mathcal{X}|\varepsilon e^{2\lambda|\mathcal{X}|} \right)^{1/3} + 2\left(\frac{|\mathcal{X}|}{\varepsilon} \right)^{2D} e^{-M\delta^2/2}.
\end{aligned}$$

If $A_1 \cap A_2$ holds then each cell in the common refinement of the M Mondrian partitions contains an element of the ε-grid $\mathcal{U}$, and the Laplace kernel of lifetime λ changes by at most $1 - e^{-D\lambda\varepsilon}$ when moving from any point in $\mathcal{X}$ to the nearest grid point in its partition cell (in the common refinement). Therefore, as long as $2(1 - e^{-D\lambda\varepsilon}) < \frac{\delta}{2}$ (i.e., $\varepsilon \leq \frac{1}{\lambda D}\ln(1 - \frac{\delta}{4})$), the event $(A_1 \cup A_2 \cup A_3)^c$ implies that δ-approximation holds throughout $\mathcal{X}$. The upper bound on $\mathbb{P}(A_1 \cup A_2 \cup A_3)$ above is minimized for

$$\varepsilon_0 = \left(\frac{12D|\mathcal{X}|^{2D} e^{-M\delta^2/2}}{(4\lambda|\mathcal{X}|)^{1/3} e^{2\lambda|\mathcal{X}|/3}} \right)$$

which tends to 0 as $M \to \infty$ and so for large enough M, we do have $\varepsilon_0 \leq \frac{1}{\lambda D}\ln(1 - \frac{\delta}{4})$. For these large enough M it then holds that

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\mathbf{x},\mathbf{x}' \in \mathcal{X}} \left| \phi(\mathbf{x})^T \phi(\mathbf{x}') - k(\mathbf{x},\mathbf{x}') \right| > \delta \right] \\
&\leq \mathbb{P}(A_1 \cup A_2 \cup A_3) \\
&\leq \left(4\lambda^2 M^2 |\mathcal{X}|\varepsilon_0 e^{2\lambda|\mathcal{X}|} \right)^{1/3} + 2\left(\frac{|\mathcal{X}|}{\varepsilon_0} \right)^{2D} e^{-M\delta^2/2} \\
&= \left(2^{1/(2D)} 4\lambda^2 M^2 |\mathcal{X}|^2 e^{2\lambda L}/D \right)^{1/(3+1/2D)} e^{-\frac{M\delta^2}{12D+2}} \\
&\in \mathcal{O}\left(M^{2/3} e^{-\frac{M\delta^2}{12D+2}} \right). \qquad \square
\end{aligned}$$

Proposition 5. *In a Mondrian regression forest with a factorizing Gaussian prior over leaf predictions, the learning objective function can be stated as*

$$\min_{\mathbf{w} \in \mathbb{R}^C} \sum_{n=1}^{N} \frac{1}{M} \sum_{m=1}^{M} \mathit{loss}(y_n, \hat{y}_n^{(m)}) + \gamma^2 \|\mathbf{w}\|_2^2.$$

Proof. The predictive mean parameters $\mathbf{w}^{(m)}$ in the leaves of the m-th tree are fitted by solving

$$\min_{\mathbf{w}^{(m)} \in \mathbb{R}^{C^{(m)}}} \sum_{n=1}^{N} (y_n - \mathbf{w}^{(m)T} \boldsymbol{\phi}_n^{(m)})^2 + \gamma^2 \|\mathbf{w}^{(m)}\|_2^2$$

where γ^2 is the ratio of noise and prior variance in the predictive model. The parameters $\mathbf{w}^{(m)}$ are disjoint for different trees, so these M independent optimization problems are equivalent to minimizing the average

$$\min_{\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(M)}} \frac{1}{M} \sum_{m=1}^{M} \left(\sum_{n=1}^{N} (y_n - \hat{y}_n^{(m)})^2 + \gamma^2 \|\mathbf{w}^{(m)}\|_2^2 \right)$$

where $\hat{y}_n^{(m)} := \mathbf{w}^{(m)T} \boldsymbol{\phi}_n^{(m)}$ is the m-th tree's prediction at data point n. Rewriting in terms of the squared loss $\mathrm{loss}(y, \hat{y}) := (y - \hat{y})^2$ and the normalized concatenated weights $\mathbf{w} := M^{-1/2}[\mathbf{w}^{(1)T} \cdots \mathbf{w}^{(M)T}]^T$, the learning objective function becomes

$$\min_{\mathbf{w} \in \mathbb{R}^C} \sum_{n=1}^{N} \frac{1}{M} \sum_{m=1}^{M} \mathrm{loss}(y_n, \hat{y}_n^{(m)}) + \gamma^2 \|\mathbf{w}\|_2^2. \qquad \square$$

B Bayesian kernel width learning

Section 4.2.1 described how in a ridge regression setting, the marginal likelihood $\mathcal{L}(\lambda) = p(\mathbf{y}|\mathbf{X},\lambda)$ can be efficiently computed for all $\lambda \in [0,\Lambda]$. With a prior $p(\lambda)$ over the lifetime (inverse kernel width) λ whose support is included in $[0,\Lambda]$, the posterior distribution over λ is

$$p(\lambda|\mathbf{y},\mathbf{X}) \propto p(\lambda)p(\mathbf{y}|\mathbf{X},\lambda)$$

with normalizing constant

$$p(\mathbf{y}|\mathbf{X}) = \sum_{c=0}^{C-M} p(\mathbf{y}|\mathbf{X},\lambda=\tau_c)\int_{\tau_c}^{\tau_{c+1}} p(\lambda)\,\mathrm{d}\lambda$$

where $0 = \tau_0 < \tau_1 < \cdots < \tau_{C-M}$ is the sequence of times when new cuts appeared in any of the M Mondrian samples. The predictive distribution at a new test point $\mathbf{x}_*$ is obtained by marginalizing out λ:

$$\begin{aligned}
&p(y_*|\mathbf{x}_*,\mathbf{X},\mathbf{y})\\
&= \int p(y_*|\mathbf{x}_*,\mathbf{X},\mathbf{y},\lambda)p(\lambda|\mathbf{y},\mathbf{x})\,\mathrm{d}\lambda\\
&= \sum_{c=0}^{C-M} p(y_*|\mathbf{x}_*,\lambda=\tau_c)p(\tau_c \le \lambda < \tau_{c+1}|\mathbf{y},\mathbf{X})\\
&= \sum_{c=0}^{C-M} p(y_*|\mathbf{x}_*,\lambda=\tau_c)\int_{\tau_c}^{\tau_{c+1}} p(\lambda|\mathbf{y},\mathbf{X})\,\mathrm{d}\lambda\\
&= \sum_{c=0}^{C-M} p(y_*|\mathbf{x}_*,\lambda=\tau_c)\int_{\tau_c}^{\tau_{c+1}} \frac{p(\lambda)p(\mathbf{y}|\mathbf{X},\lambda)}{p(\mathbf{y}|\mathbf{X})}\,\mathrm{d}\lambda\\
&= \sum_{c=0}^{C-M} p(y_*|\mathbf{x}_*,\lambda=\tau_c)p(\mathbf{y}|\mathbf{X},\lambda=\tau_c)\frac{\int_{\tau_c}^{\tau_{c+1}} p(\lambda)\,\mathrm{d}\lambda}{p(\mathbf{y}|\mathbf{X})}\\
&= \frac{\sum_{c=0}^{C-M} p(y_*|\mathbf{x}_*,\lambda=\tau_c)p(\mathbf{y}|\mathbf{X},\lambda=\tau_c)\int_{\tau_c}^{\tau_{c+1}} p(\lambda)\,\mathrm{d}\lambda}{\sum_{c=0}^{C-M} p(\mathbf{y}|\mathbf{X},\lambda=\tau_c)\int_{\tau_c}^{\tau_{c+1}} p(\lambda)\,\mathrm{d}\lambda}\\
&= \sum_{c=0}^{C-M} k_c p(y_*|\mathbf{x}_*,\lambda=\tau_c)
\end{aligned}$$

where the mixing coefficients

$$k_c := \frac{p(\mathbf{y}|\mathbf{X},\lambda=\tau_c)\int_{\tau_c}^{\tau_{c+1}} p(\lambda)\,\mathrm{d}\lambda}{\sum_{c=0}^{C-M} p(\mathbf{y}|\mathbf{X},\lambda=\tau_c)\int_{\tau_c}^{\tau_{c+1}} p(\lambda)\,\mathrm{d}\lambda}$$

can be precomputed and cached for faster predictions. The integrals $\int_{\tau_c}^{\tau_{c+1}} p(\lambda)\,\mathrm{d}\lambda$ can be readily evaluated if we have access to the cumulative distribution function of our prior $p(\lambda)$, which we assume.

C Online learning

Mirroring Section 4.2.1, we discuss the example of ridge regression where exact online updates can be carried out. Assume we have access to the regularized feature covariance matrix $\mathbf{A} = \mathbf{\Phi}^T\mathbf{\Phi} + \delta^2\mathbf{I}_C$ and its inverse $\mathbf{A}^{-1}$ or Cholesky decomposition $\mathrm{chol}(\mathbf{A})$ before a new data point $\mathbf{x} \in \mathbb{R}^D$ arrives, and we wish to update these efficiently.

If the dimensionality of ϕ increases by k due to $\mathbf{x}$ creating k new non-empty partition cells, we first append k rows and columns to $\mathbf{A}$, containing $0s$ only, except on the main diagonal we put δ^2. Correspondingly, $\mathbf{A}^{-1}$ or $\mathrm{chol}(A)$ are updated by appending k rows and columns, with non-zero entries only on the main diagonal. (These entries would equal δ^{-2} in $\mathbf{A}^{-1}$ and δ in $\mathrm{chol}(\mathbf{A})$). This ensures the feature map ϕ now incorporates all necessary features.

Noting that the (i,j)-entry of $\mathbf{A} - \delta^2\mathbf{I}_C$ counts data points belonging to partition cells i and j at the same time (this can be non-zero only if i, j correspond to different Mondrian samples), normalized by $1/M$, and that the (i,j)-entry of the outer product $\phi(\mathbf{x})\phi(\mathbf{x})^T$ is $1/M$ if the new data point $\mathbf{x}$ falls into both cells i and j, and 0 otherwise, we see that

$$\mathbf{A}_{\text{new}} \leftarrow \mathbf{A}_{\text{old}} + \phi(\mathbf{x})\phi(\mathbf{x})^T$$

is a rank-1 update. Therefore both $\mathbf{A}^{-1}$ and $\mathrm{chol}(\mathbf{A})$ can be updated efficiently in $\mathcal{O}(C^2)$ time and the new MAP weights $\hat{\mathbf{w}}_{\text{new}} = \mathbf{A}_{\text{new}}^{-1}(\mathbf{\Phi}^T\mathbf{y})$ in $\mathcal{O}(MC)$ by exploiting sparsity of $\phi(\mathbf{x})$. The determinant of the rank-1 updated matrix $\mathbf{A}_{\text{new}}$ can also be updated in $\mathcal{O}(C^2)$ time using the Matrix determinant lemma, or obtained directly from the Cholesky decomposition (as the squared product of its diagonal entries) in $\mathcal{O}(C)$ time, allowing the training marginal likelihood to be updated in $\mathcal{O}(NM + C^2)$.