Supplementary Materials

A Unified and Fast Explainable Model for Predictive Analytics

1 Proofs of Lemmas

Lemma 1. FXAM's normal equations satisfy stationarity conditions of $\mathcal{L}$.

Proof. According to equation (20) in [29], we equivalently need to prove that normal equations should satisfy the bellowing criteria (here M^{-} denote generalized inverse satisfying $MM^{-}M = M$):

$$\mathcal{L} = \left\| y - \sum_{j=1}^{p} f_j - f_Z - f_T - f_S \right\|^2 + \sum_{j=1}^{p} f_j^T (M_j^- - I) f_j + f_z^T (M_z^- - I) f_z + f_T^T (M_T^- - I) f_T + f_s^T (M_s^- - I) f_s$$

Recall that $M_j = \left(I + \lambda K_j\right)^{-1}$ implies M_j is inversible, so $M_j^- = I + \lambda K_j \Rightarrow f_j^T (M_j^- - I) f_j = \lambda f_j^T K_j f_j$. Same deduction for M_T.

Re-write $M_z = ZAZ^T$ where $A = (Z^T Z + \lambda_z I)^{-1}$which is a $Q \times Q$ symmetric and invertible matrix. and because $M_z M_z^- M_z = M_z$, expanding it we get $ZAXA^T Z^T = ZAZ^T$ where $X = Z^T S_z^- Z$. Then we get to know $X = A^{-1} + H$, where H is some matrix satisfying $ZAHA^T Z^T = 0$. Recall that $f_Z = \mathrm{Z}\beta$, thus $f_z^T (M_z^- - I) f_z = \beta^T X \beta - \beta^T Z^T Z \beta = \beta^T (A^{-1} + H - Z^T Z)\beta = \beta^T (\lambda_z I + H)\beta = \lambda_z \beta^T \beta$. The last equality holds because $\beta = AZ^T \tilde{y}$ according to normal equations, thus $\beta^T H \beta = \tilde{y}^T ZAHA^T Z^T \tilde{y} = 0$.

Because permutation matrix P is orthogonal matrix, thus $M_S^- = P^T \begin{bmatrix} I + \lambda_S K'_{S_0} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & I + \lambda_S K'_{S_{d-1}} \end{bmatrix} P = I + \lambda_S P^T \begin{bmatrix} K'_{S_0} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & K'_{S_{d-1}} \end{bmatrix} P$, and considering K'_{S_φ} is a $\left|\mathcal{T}_\varphi\right| \times \left|\mathcal{T}_\varphi\right|$ matrix only applies to phase-φ data points, thus it is easy to see $f_s^T (M_s^- - I) f_s = \lambda_S \sum_{\varphi=0}^{d-1} f_{S_\varphi}^T K_{S_\varphi} f_{S_\varphi}$. ■

Theorem 1. Solutions of FXAM's normal equations exists and are global optimal.

Proof. According to theorem 2 of [29], the normal equations exists and are global optimal if each smoothing matrix M_j, M_z, M_T, or M_S is symmetric and shrinking (i.e., with eigenvalues in [0,1]). Thus we check M_j, M_z, M_T, and M_S one by one:

M_j, M_T are indeed symmetric and shrinking according to standard analysis of cubic spline smoothing matrix.

Re-write $M_z = ZAZ^T$ where $A = (Z^T Z + \lambda_z I)^{-1}$. It is easy to see that A is a symmetric matrix thus $M_z^T = (Z^T)^T A^T Z^T = ZAZ^T$=$M_z$.

Denote singular value decomposition of Z is $Z = UDV^T$, where U and V are orthogonal matrices, D is a $Q \times Q$ diagonal matrix, with diagonal entries $d_{11} \geq \cdots \geq d_{QQ} \geq 0$. Thus we have

$M_z y = Z(Z^T Z + \lambda I)^{-1} Z^T y = UDV^T(VD^2V^T + \lambda I)^{-1} VDU^T y = UDV^T(VD^2V^T + \lambda VIV^T)^{-1} VDU^T y = UDV^T V(D^2 + \lambda I)^{-1} V^{-1} VDU^T y = UD(D^2 + \lambda I)^{-1} DU^T y = \sum_{j=1}^{Q} u_j \frac{d_j^2}{d_j^2+\lambda} u_j^T y$, thus the eigenvalues $\frac{d_j^2}{d_j^2+\lambda}$ are in [0,1] considering $\lambda > 0$.

Re-write $M_s = P^T \Theta P$. Since each $K'_{S_\varphi}$is a symmetric matrix, thus $\left(I + \lambda_S K'_{S_\varphi}\right)^{-1}$ is symmetric and Θ is symmetric, thus M_s is symmetric. Due to the shrinking property of $(I + \lambda_S K'_{S_\varphi})^{-1}$, and considering Θ is a block-diagonal matrix with $(I + \lambda_S K'_{S_\varphi})^{-1}$as its blocks, thus Θ is also shrinking: $\|\Theta y\|^2 \le \|y\|^2 \; \forall\, y$. So $\|M_S y\|^2 = y^T M_S^T M_S y = y^T P^T \Theta^T \Theta P y = \|\Theta P y\|^2 \le \|Py\|^2 = \|y\|^2$, thus M_S is shrinking. ■

Theorem 2. TSI algorithm converges to a solution of FXAM's normal equations.
Proof. Denote the index set $I \coloneqq \{1,2, \dots, p, Z, T, S\}$. We only need to prove that TSI converges to a solution of FXAM's homogenous equations (i.e., FXAM's normal equations with $y = 0$), because a general solution is a solution of homogenous equations plus an arbitrary solution of FXAM's normal equations. Denote the loss function of homogenous equations as $\mathcal{L}_0(f) \coloneqq \left\|\sum_{j\in I} f_j\right\|^2 + \sum_{j\in I} f_j^T (M_j^- - I) f_j$.

We define a linear map T_j to describe the updating of jth component in TSI when $y = 0$:

$$T_j \colon \begin{bmatrix} f_1 \\ \vdots \\ f_p \\ f_z \\ f_T \\ f_S \end{bmatrix} \equiv f \to \begin{bmatrix} f_1 \\ \vdots \\ -M_j \sum_{i\in I, i\neq j} f_i \\ f_z \\ f_T \\ f_S \end{bmatrix}, \forall j \in I$$

A full cycle of backfitting over numerical features is then described by $K = T_p T_{p-1} \dots T_1$. Denote the m full cycles as K^m. It is obvious that K^m converges to a limit K^∞ (we can view this as a standard task of backfitting over pure numerical features) therefore with property $KK^\infty = K^\infty$. Note that K^∞ describes the procedure of stage 1, thus the full cycle of entire TSI is $\mathcal{K} = T_S T_T T_Z K^\infty$. Since each component of $\mathcal{K}$ is minimizer of $\mathcal{L}_0(f)$ and since $\mathcal{L}_0$ is a quadratic form, hence $\mathcal{L}_0(\mathcal{K}f) \le \mathcal{L}_0(f)$. When $\mathcal{L}_0(\mathcal{K}f) = \mathcal{L}_0(f)$, no strict descent is possible on any component, thus $T_S f = f, T_T f = f, T_Z f = f, K^\infty f = f$. Considering $KK^\infty = K^\infty$, thus $KK^\infty f = K^\infty f \Leftrightarrow Kf = f$ when descent vanishes. Since each component T_j of K only updates separate f_j, thus $Kf = f \Leftrightarrow T_j f = f \; \forall j \in \{1, \dots, p\}$. So descent vanishes on any f equivalent to $T_j f = f \; \forall j \in I$. Meanwhile, such f satisfies homogenous equations, which indicates $\mathcal{L}_0(f) = 0$ according to theorem 5 in [29]. In summary, we have a linear mapping $\mathcal{K}$ satisfying $\mathcal{L}_0(\mathcal{K}f) < \mathcal{L}_0(f)$ when $\mathcal{L}_0(f) > 0$ and $\mathcal{K}f = f$ when $\mathcal{L}_0(f) = 0$. According to theorem 8 of [29], $\mathcal{K}^m$ converges to $\mathcal{K}^\infty$. ■

Lemma 2. $LOSS = \sum_{i=1}^{n} (f(x_i) - y_i)^2 \le n((2LBh)^2 + 2\sigma^2)$ where B is the bounded support of kernel K_h.

Proof. $\sum_{i=1}^{n}(f(x_i)-y_i)^2=\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{n}(y_j-y_i)K_h(x_i-x_j)}{\sum_{j=1}^{n}K_h(x_i-x_j)}\right)^2\le\sum_{i=1}^{n}\frac{\sum_{j=1}^{n}(y_j-y_i)^2K_h(x_i-x_j)}{\sum_{j=1}^{n}K_h(x_i-x_j)}$ (Jensen's inequality). Considering item $(y_j-y_i)^2K_h(x_i-x_j)$: Since K_h is bounded with B, $(y_j-y_i)^2K_h(x_i-x_j)$ only accounts for x_i such that $x_i\in B(x_j)$. Denote the indexes of all the data points within $B(x_j)$ are $\{q_1,\dots,q_k\}$, denote $u\in\{1\sim k\}, v\in\{1\sim k\}$ such that $y_{q_u}\le y_{q_i}\ \forall i\ne u$; $y_{q_v}\ge y_{q_i}\ \forall i\ne v$, then $y_{q_u}\le f(x)=\frac{\sum_{j=1}^{n}y_jK_h(x-x_j)}{\sum_{j=1}^{n}K_h(x-x_j)}=\frac{\sum_{j=1}^{k}y_{q_j}K_h(x-x_{q_j})}{\sum_{j=1}^{k}K_h(x-x_{q_j})}\le y_{q_v}$, hence $(y_j-y_i)^2K_h(x_i-x_j)\le(y_{Q_v}-y_{Q_u})^2K_h(x_i-x_j)$. Considering y_{Q_v} is a sample drawn from $Y_{Q_v}=F(x_{Q_v})+\epsilon_{Q_v}$, and y_{Q_u} is a sample drawn from $Y_{Q_u}=F(x_{Q_u})+\epsilon_{Q_u}$, we get to know that $E\big(Y_{Q_v}-Y_{Q_u}\big)^2=E\big(F(x_{Q_v})-F(x_{Q_u})+\epsilon_{Q_v}-\epsilon_{Q_u}\big)^2=E\left(F(x_{Q_v})-F(x_{Q_u})\right)^2+E\big(\epsilon_{Q_v}-\epsilon_{Q_u}\big)^2\le(2LBh)^2+2\sigma^2$. Then approximately, we write $(y_{Q_v}-y_{Q_u})^2\lesssim(2LBh)^2+2\sigma^2$ which leads to $LOSS=\sum_{i=1}^{n}(f(x_i)-y_i)^2\le n((2LBh)^2+2\sigma^2)$. ■

Lemma 3. $E\|f_n-f_s\|^2\le 4c\left(\frac{(\sigma^2+\sup|F|^2)L}{s}\right)^{\frac{2}{3}}$

Proof. $E\|f_n-f_s\|^2=\int\big(f_n(x)-f_s(x)\big)^2\mu(dx)=\int\big(f_n(x)-F(x)+F(x)-f_s(x)\big)^2\mu(dx)\le\int\big(f_n(x)-F(x)+F(x)-f_s(x)\big)^2\mu(dx)+\int\big(f_n(x)-F(x)-F(x)+f_s(x)\big)^2\mu(dx)=2(E\|f_n-F\|^2+E\|f_s-F\|^2)\le 4E\|f_s-F\|^2=4c\left(\frac{(\sigma^2+\sup|F|^2)L}{s}\right)^{\frac{2}{3}}$. ■