A. Modeling Multi-Way Interactions

In certain applications, we might expect that there are interactions of order greater than two. For example, suppose we are trying to predict college admissions. Then, we might expect a three-way interaction between a candidate's SAT score, GPA, and extracurricular involvement. Individually, these variables might only exhibit moderate association but together they could have a multiplicative effect. For example, we might expect that candidates who have high SAT scores, high GPAs, and excellent extracurricular activities will be accepted with near certainty, while candidates who only possess one/two of these qualities are borderline applicants.

We now show how to extend our results to handle such three-way, or more generally, *r-way interactions*.

Definition A.1. (r-way interactions) The *r-way interactions* of a covariate vector $x \in \mathbb{R}^p$ are generated from the feature map

$$\Phi_r(x) \coloneqq \bigoplus_{d=1}^{r} \bigoplus_{k:k_1+\cdots+k_p=d} \prod_{j=1}^{p} x_j^{k_j}, \quad k \in \mathbb{N}^p,$$

where $\bigoplus_{j=1}^{m} a_j \coloneqq (a_{11}, \cdots, a_{1k_1}, \cdots, a_{m1}, \cdots, a_{mk_m})$ denotes the concatenation of vectors $a_j \in \mathbb{R}^{k_j}$.

To model r-way interactions, we must use degree r polynomial kernels to generate all the necessary interactions. Hence, we recommend using the following generalized two-way interaction kernel, which we call the r-way interaction kernel.

Definition A.2. (r-way interaction kernel) A kernel k is called an *r-way interaction kernel* if for some choice of $M_1, M_2, M_3 \in \mathbb{N}$, $\alpha, \psi, \lambda^{(m)} \in \mathbb{R}_+^p$ $(m = 1, \ldots, M_1)$, $\nu^{(m)} \in \mathbb{R}_+$ $(m = 1, \ldots, M_2)$, and $\nabla^{(m)} \in \mathbb{R}_+^p$ $(k = 1, \ldots, M_3)$ it can be re-expressed as

$$\sum_{m=1}^{M_1} k_{poly,r}^1(\lambda^{(m)} \odot x, \lambda^{(m)} \odot y) + \sum_{m=1}^{M_2} \nu^{(m)} \left[\prod_{s=1}^{r} x_{i_{sm}} \prod_{s=1}^{r} y_{i_{sm}} \right] + \sum_{m=1}^{M_3} k_{r-1}(\nabla^{(m)} \odot x, \nabla^{(m)} \odot y),$$

where $\odot$ is the Hadamard product and k_{r-1} is an $r-1$ degree interaction kernel. The base case kernel (i.e., when $r = 2$) is provided in Definition 4.2.

To select the weights for an r-way interaction kernel, we must solve a system of equations similar to Eq. (9), except for a target prior covariance matrix $\Sigma_\tau \in \mathbb{R}^{\dim(\Phi_r) \times \dim(\Phi_r)}$.

B. Proofs

B.1. Proof of Proposition 4.1

Let $g(\cdot) = \theta^T \Phi_2(\cdot)$ and $\theta \mid \tau \sim \mathcal{N}(0, \Sigma_\tau)$. Then, $y^{(n)} = g(x^{(n)}) + \epsilon^{(n)}$. The first claim follows by taking $\phi = \Phi_2$ and $f = g$ in Rasmussen & Williams (2006, Equation 2.12).

The second claim follows directly from the duality between the weight-space and function-space view of a GP (Rasmussen & Williams, 2006, Chapter 2).

B.2. Proof of Theorem 4.3

The proof of Theorem 4.3 depends critically on Lemma B.1 below, which characterizes the relation between adding two kernels and the resulting induced prior covariance matrix.

Lemma B.1. *Let k_1 and k_2 be two kernels such that there exists vectors $a^{(1)}, a^{(2)} \in \mathbb{R}^{dim(\Phi_2)}$ for which $k_i(x, y) = \langle a^{(i)} \odot \Phi_2(x), a^{(i)} \odot \Phi_2(y) \rangle$. Let $k_3(x, y) = k_1(x, y) + k_2(x, y)$. Then,*

$$k_3(x, y) = \langle \Sigma_3^{\frac{1}{2}} \Phi_2(x), \Sigma_3^{\frac{1}{2}} \Phi_2(y) \rangle \quad \textit{s.t.} \quad \Sigma_3 = \textit{diag}(a^{(1)} \odot a^{(1)} + a^{(2)} \odot a^{(2)}). \tag{12}$$

Proof. By the sum property of kernels,

$$\begin{aligned} k_1(x,y) + k_2(x,y) &= \langle [a_1\ a_2] \odot [\Phi_2(x)\ \Phi_2(x)], [a_1\ a_2] \odot [\Phi_2(y)\ \Phi_2(y)] \rangle \\ &= \langle a^{(1)} \odot \Phi_2(x), a^{(1)} \odot \Phi_2(y) \rangle + \langle a^{(2)} \odot \Phi_2(x), a^{(2)} \odot \Phi_2(y) \rangle \\ &= \langle a^{(1)} \odot a^{(1)} \odot \Phi_2(x), \Phi_2(y) \rangle + \langle a^{(2)} \odot a^{(2)} \odot \Phi_2(x), \Phi_2(y) \rangle \\ &= \langle a^{(1)} \odot a^{(1)} \odot \Phi_2(x) + a^{(2)} \odot a^{(2)} \odot \Phi_2(x), \Phi_2(y) \rangle \\ &= \langle (a^{(1)} \odot a^{(1)} + a^{(2)} \odot a^{(2)}) \odot \Phi_2(x), \Phi_2(y) \rangle \\ &= \Phi_2^T(x)\ \mathrm{diag}((a^{(1)} \odot a^{(1)} + a^{(2)} \odot a^{(2)})\ \Phi_2(y) \\ &= k_3(x,y). \end{aligned} \tag{13}$$

□

By Lemma B.1, it suffices to write out the feature map of each kernel in Definition 4.2. The induced feature maps of each respective kernel term in Definition 4.2 are given by $a_i \odot \Phi_2(x), 1 \le i \le 4$ for

$$\begin{aligned} a_1 &\coloneqq ((\lambda_1^{(m)})^2, \cdots, (\lambda_p^{(m)})^2, \sqrt{2}\lambda_1^{(m)}\lambda_2^{(m)}, \cdots, \sqrt{2}\lambda_{p-1}^{(m)}\lambda_p^{(m)}, \sqrt{2}\lambda_1^{(m)}, \cdots, \sqrt{2}\lambda_p^{(m)}, 1) \\ a_2 &\coloneqq (0, \cdots, 0, 0, \cdots, 0, \alpha_1, \cdots, \alpha_p, \sqrt{A}) \\ a_3 &\coloneqq (\psi_1, \cdots, \psi_p, 0, \cdots, 0, 0, \cdots, 0, 0) \\ a_4 &\coloneqq (0, \cdots, 0, 0, \cdots, 0, \sqrt{\nu^{(m)}}, 0, \cdots, 0, 0, \cdots, 0, 0) \end{aligned} \tag{14}$$

The first claim follows from Eq. (14) and Lemma B.1.

To prove the second claim, take an arbitrary diagonal prior covariance matrix $S \in \mathbb{R}^{\dim(\Phi_2)\times\dim(\Phi_2)}$. It suffices to show that there exists a solution of,

$$\begin{aligned} \mathrm{diag}(S)_{(i)} &= \alpha_i^2 + 2\sum_{m=1}^{M_1}\left[\lambda_i^{(m)}\right]^2 \\ \mathrm{diag}(S)_{(ij)} &= 2\sum_{m=1}^{M_1}\left[\lambda_i^{(m)}\lambda_j^{(m)}\right]^2 + \sum_{m:i_m=i,j_m=j}^{K_2}\nu^{(m)} \\ \mathrm{diag}(S)_{(ii)} &= \psi_i^2 + \sum_{m=1}^{M_1}\left[\lambda_i^{(m)}\right]^4 \\ \mathrm{diag}(S)_{(0)} &= M_2 + A. \end{aligned}$$

for some choice of $M_1, M_2 \in \mathbb{N}$, $\alpha, \psi, \lambda^{(m)} \in \mathbb{R}_+^p$ $(m = 1, \ldots, M_1)$, $\nu^{(m)} \in \mathbb{R}_+$ $(m = 1, \ldots, M_2)$, and $A \in \mathbb{R}$. Take $\alpha_i^2 = \mathrm{diag}(S)_{(i)}$ and $\psi_i^2 = \mathrm{diag}(S)_{(ii)}$, for $i = 1, \cdots, p$. Take $\lambda^{(m)} = 0$. Let $M_2 = \frac{p(p-1)}{2}$ and $\nu^{(1)} = \mathrm{diag}(S)_{(12)}, \cdots, \nu^{(M_2)} = \mathrm{diag}(S)_{((p-1)p)}$. Finally, letting $A = \mathrm{diag}(S)_{(0)} - M_2$ solves the system.

Remark. While we have shown one of the *many* ways to solve the above system for an arbitrary S, the strategy taken above is not practically useful; computing the kernel in this fashion will take $\Theta(p^2)$ time because $M_2 = \Theta(p^2)$. In practice, we must leverage the polynomial kernels (i.e., those in the M_1 sum) to avoid making M_2 large. We show how such a strategy works in Appendix C.

B.3. Proof of Theorem 5.1

Define $g(A^{ij}) \coloneqq (g(e_i), g(-e_i), g(e_j), g(e_{ij}))$. Then,

$$\begin{aligned} g(A^{ij}) \mid D, \tau \sim \mathcal{N}(\mu_{g_{ij}}, \Sigma_{ij}) \quad \text{s.t.} \quad \mu_{g_{ij}} &\coloneqq K_\tau(A^{ij}, X) H_\tau Y, \\ \Sigma_{ij} &\coloneqq \left[K_\tau(A^{ij}, A^{ij}) - K_\tau(A^{ij}, X) H_\tau K_\tau(X, A^{ij})\right], \end{aligned} \tag{15}$$

which follows directly from Rasmussen & Williams (2006, Equation 2.21). Notice that,

$$\theta_{x_i} = \frac{g(e_1)}{2} - \frac{g(-e_1)}{2} = a_i^T g(A^{ij}) \quad \text{and} \quad \theta_{x_i x_j} = \frac{g(e_1)}{2} - \frac{g(-e_1)}{2} - g(e_j) + g(e_{ij}) = a_{ij}^T g(A^{ij}), \tag{16}$$

where $a_i = (1/2, -1/2, 0, 0)$ and $a_{ij} = (-1/2, 1/2, -1, 1)$. The proof follows from Eq. (15), Eq. (16), and recalling that an affine transformation $h : x \mapsto Ax$ of a multivariate Gaussian distribution $Z \sim \mathcal{N}(\mu, \Sigma)$ is given by $h(Z) \sim \mathcal{N}(A\mu, A\Sigma A^T)$.

B.4. Proof of Corollary 5.2

Corollary 5.2 follows immediately once we can show that $K_\tau(A_{ij}, X)$ takes $O(1)$ time. It suffices to show $k_\tau(x^{(n)}, e_i)$ and $k_\tau(x^{(n)}, e_i + e_j)$ take $O(1)$ time. Since k_τ is a sum of polynomial kernels, $k_\tau(x, y)$ only depends on $x, y \in \mathbb{R}^p$ through the inner product $x^T y$. Hence, for vectors $\tilde{x}, \tilde{y} \in \mathbb{R}^M$, $k_\tau(\tilde{x}, \tilde{y})$ is well-defined and just depends on $\tilde{x}^T \tilde{y}$. Now, $k_\tau(x^{(n)}, e_i) = k_\tau(x_i^{(n)}, 1)$ and $k_\tau(x^{(n)}, e_i + e_j) = k_\tau((x_i^{(n)}, x_j^{(n)}), (1, 1))$. Since $k_\tau(x_i^{(n)}, 1)$ and $k_\tau((x_i^{(n)}, x_j^{(n)}), (1, 1))$ do not depend on p, these terms each take $O(1)$ time to compute.

B.5. The General Kernel Interaction Trick

In this section, we generalize the kernel interaction trick, namely show how to access the distribution of arbitrary components of θ. First, we require some new notation. For $E \subseteq \{1, \cdots, p\}$, $|E| = M$, define

$$\theta_E \coloneqq (\theta_{x_{i_1}}, \cdots, \theta_{x_{i_M}}, \theta_{x_{i_1} x_{i_2}}, \cdots, \theta_{x_{i_{M-1}} x_{i_M}}), \quad i_j \in E. \tag{17}$$

We show how to compute $\theta_E \mid \tau, D$ from the GP posterior predictive distribution. Without any lost of generality, we may assume $E = \{1, \cdots, M\}$ by relabeling the covariates.

Theorem B.2. *(General kernel interaction trick) Let* $H_\tau \coloneqq (K_\tau + \sigma^2 I_N)^{-1}$ *and*

$$A_M \coloneqq [e_1, -e_1, \cdots e_M, -e_M, e_1 + e_2, \cdots, e_{M-1} + e_M]^T.$$

Let $K_\tau(A_M, X) = K_\tau(X, A_M)^T$ *be the matrix formed by taking the kernel between each row of* A_M *with each row of* X. *Let*

$$\begin{aligned} a_i &\coloneqq (0, 0, \cdots, 1/2, -1/2, \cdots, 0, 0, \cdots, 0) \in \mathbb{R}^{2M + \frac{M(M-1)}{2}} \\ a_{ij} &\coloneqq (0, 0, \cdots, 1/2, -1/2, \cdots, -1, \cdots, 0, 0, \cdots, 1, \cdots, 0) \in \mathbb{R}^{2M + \frac{M(M-1)}{2}} \end{aligned} \tag{18}$$

for $i < j$. *That is,* a_i *has non-zero entries at* e_i *and* $-e_i$ *and* a_{ij} *has non-zero entries at* e_i, $-e_i$, $-e_j$, *and* $e_i + e_j$. *Let*

$$R_M \coloneqq [a_1 \cdots a_M \;\; a_{12} \cdots a_{(M-1)M}]^T. \tag{19}$$

Then, $\theta_E \mid \tau, D$ *is a multivariate Gaussian distribution with mean* $R_M K_\tau(A_M, X) H_\tau Y$ *and covariance matrix*

$$R_M \left[K_\tau(A_{ij}, A_{ij}) - K_\tau(A_{ij}, X) H_\tau K_\tau(X, A_{ij}) \right] R_M^T.$$

Proof. Following the proof of Theorem 5.1,

$$\begin{aligned} g(A^M) \mid D, \tau \sim \mathcal{N}(\mu_{g_M}, \Sigma_M) \quad \text{s.t.} \quad \mu_{g_M} &\coloneqq K_\tau(A^M, X) H_\tau Y, \\ \Sigma_M &\coloneqq \left[K_\tau(A^M, A^M) - K_\tau(A^M, X) H_\tau K_\tau(X, A^M) \right]. \end{aligned} \tag{20}$$

Similar to Eq. (16),

$$\theta_{x_i} = \frac{g(e_1)}{2} - \frac{g(-e_1)}{2} = a_i^T g(A^M) \quad \text{and} \quad \theta_{x_i x_j} = \frac{g(e_1)}{2} - \frac{g(-e_1)}{2} - g(e_j) + g(e_{ij}) = a_{ij}^T g(A^M). \tag{21}$$

The proof follows from Eq. (20), Eq. (21), and recalling that an affine transformation $h : x \mapsto R_M^T x$ of a multivariate Gaussian distribution $Z \sim \mathcal{N}(\mu, \Sigma)$ is given by $h(Z) \sim \mathcal{N}(R_M \mu, R_M \Sigma R_M^T)$.

□

Corollary B.3. *Given* K_τ, *the distribution* $\theta_E \mid \tau, D$ *takes* $O(M^2)$ *time and memory to compute.*

Proof. The proof is identical to the one provided in Appendix B.4. □

B.6. Proof of Proposition 6.1

See Appendix C.2.

C. Example Bayesian Interaction Models

In the following subsections, we show how to solve Eq. (9) for several classes of models.

C.1. Block-Degree Priors

Suppose we would like to set the prior variance of all terms with the same degree equal. That is, we would like to use a prior of the form,

$$\begin{aligned} \eta \in \mathbb{R}^3 &\sim p(\eta) \\ \theta_{x_i} \mid \eta &\sim \mathcal{N}(0, \eta_1^2) \\ \theta_{x_i x_j} \mid \eta &\sim \mathcal{N}(0, \eta_2^2) \\ \theta_{x_i^2} \mid \eta &\sim \mathcal{N}(0, \eta_3^2) \\ \theta_0 \mid \eta &\sim \mathcal{N}(0, c^2) \end{aligned} \tag{22}$$

To find the corresponding kernel, let $\lambda = (\frac{1}{\sqrt[4]{2}}\sqrt{\eta_2}, \cdots, \frac{1}{\sqrt[4]{2}}\sqrt{\eta_2})$, $M_1 = 1$ and $M_2 = 0$. Then, $\text{diag}(S)_{(ij)} = \eta_2^2$. Setting $\psi_i^2 = \eta_3^2 - \frac{1}{2}\eta_2^2$, implies that $\text{diag}(S)_{(ii)} = \eta_3^2$. Finally, letting $\alpha_i^2 = \tau_1^2 - \frac{2\eta_2}{\sqrt{2}}$ and $A = c^2 - 1$ implies that $\text{diag}(S)_{(i)} = \eta_1^2$ and $\text{diag}(S)_{(0)} = c^2$ as desired. We may equivalently re-write the induced kernel as,

$$k_{block,\eta}(x, y) = \frac{\eta_2^2}{2} k^1_{\text{poly},2}(x, y) + (\eta_3^2 - \frac{\eta_2^2}{2}) k^0_{\text{poly},1}(x \odot x, y \odot y) + \left(\eta_1^2 - \eta_2^2\right) k^{c^2 - \frac{\eta_2}{2}}_{\text{poly},1}(x, y). \tag{23}$$

Hence, Eq. (22) admits a kernel that only takes $O(p)$ time to compute.

C.2. Sparsity Priors

By Lemma B.1, the sparsity prior model provided in Eq. (11) equals $k_{block,\eta}(\kappa \odot x, \kappa \odot y)$.

D. SKIM Model Details

The full hierarchal form of SKIM is provided below, which is based closely on the *regularized horseshoe prior* (Piironen & Vehtari, 2017) and the model proposed in Griffin & Brown (2017):

$$\begin{aligned} m^2 \sim \text{InvGamma}(\alpha_1, \beta_1) \qquad & \xi^2 \sim \text{InvGamma}(\alpha_2, \beta_2) \\ \phi := \frac{s}{p - s}\frac{\sigma}{\sqrt{N}} \qquad & \sigma \sim N^+(0, \alpha_3) \\ \kappa_i = \frac{m\lambda_i}{\sqrt{m^2 + \eta_1^2 \lambda_i^2}} \qquad & \lambda_i \sim C^+(0, 1) \\ \eta_1 \sim C^+(0, \phi) \qquad & \eta_2 \sim \frac{\eta_1^2}{m^2}\xi \\ \theta_{x_i} \mid \eta, \kappa &\sim \mathcal{N}(0, \eta_1^2 \kappa_i^2) \\ \theta_{x_j} \mid \eta, \kappa &\sim \mathcal{N}(0, \eta_1^2 \kappa_j^2) \\ \theta_{x_i x_j} \mid \eta, \kappa &\sim \mathcal{N}(0, \eta_2^2 \kappa_i^2 \kappa_j^2) \\ \theta_0 \mid \eta &\sim \mathcal{N}(0, c^2) \end{aligned}$$

where s, α_i, and β_i are user-specified hyperparameters, $C^+(0, 1)$ is a half-Cauchy distribution, and N^+ is a half-normal distribution. More details, such as selecting the hyperparameters, desirable properties, and interpretations of SKIM, are provided below.

D.1. SKIM Details

Recall that we are primarily interested in the case when θ is sparse and satisfies strong-hierarchy. In order to promote sparsity in the main effects, we require two ingredients: (1) a prior on the *global shrinkage* parameter η_1 and (2) a prior on the *local shrinkage* parameters $\kappa \in \mathbb{R}^p$, which together express the prior variance of θ_{x_i} as (Carvalho et al., 2009; Piironen & Vehtari, 2017):

$$\theta_{x_i} \mid \kappa, \eta_1 \sim \mathcal{N}(0, \eta_1^2 \kappa_i^2), \quad i = 1, \cdots, p. \tag{24}$$

η_1 controls the overall sparsity level of the model; in particular, the model becomes more sparse as η_1 decreases. If we expect s non-zero main effects, then setting $\eta_1 = \frac{s}{p-s}\frac{\sigma}{\sqrt{N}}$ will yield an expected prior sparsity level of s by Piironen & Vehtari (2017, Equation 3.12). However, we often do not know exactly how to select s. Hence, Piironen & Vehtari (2017) instead suggest drawing,

$$\phi := \frac{s}{p-s}\frac{\sigma}{\sqrt{N}} \qquad \eta_1 \sim C^+(0, \phi), \tag{25}$$

to express our uncertainty of not knowing the true main effect sparsity level.

The prior variance of θ_{x_i} is non-negligible only when κ_i is large enough to escape the global shrinkage of η_1. Hence, we want to draw κ_i from a heavy-tailed distribution so that certain main effects can escape global shrinkage. Carvalho et al. (2009) suggest drawing κ_i from a half-Cauchy distribution since this distribution has fat tails. However, such a prior often leads to undesirable numerical stability issues when using NUTS (Piironen & Vehtari, 2017). As a result, Piironen & Vehtari (2017) instead propose using a *regularized horseshoe* prior which truncates the half-Cauchy distribution to have support only on $[0, m)$ instead of $[0, \infty)$. This truncation (empirically) turns out to lead to better mixing properties, and is equivalent to the following generative mechanism:

$$\kappa_i = \frac{m\lambda_i}{\sqrt{m^2 + \eta_1^2 \lambda_i^2}} \qquad \lambda_i \sim C^+(0, 1) \tag{26}$$

As $\lambda_i \to \infty$, $\kappa_i \to \frac{m}{\eta_1}$. Hence, as $\lambda_i \to \infty$, the prior variance of θ_{x_i} equals m. Since we might not know the scale m of the non-zero main effects, we place a prior on m, namely,

$$m^2 \sim \text{InvGamma}(\alpha_1, \beta_1) \tag{27}$$

for hyperparameters α_1 and α_2.

Next, we model the interactions. If strong-hierarchy holds, sparsity comes for free; if there are only $s \ll p$ non-zero main effects, then there are at most $\frac{s(s-1)}{2} \ll p^2$ possible pairwise interactions. We must be careful, however, because strong-hierarchy trivially holds; our main effect estimates will, with probability one, never equal zero because the prior variances of the main effects are greater than 0 with probability one by Eq. (24) and our choice of priors. Instead, we aim for a relaxed version of strong-hierarchy. Namely, that the prior variance of an interaction $\theta_{x_i x_j}$ is large only if θ_{x_i} and θ_{x_j} are both large. θ_{x_i} and θ_{x_j} are large only when κ_i and κ_j are large. Hence, it suffices to make the prior variance of $\theta_{x_i x_j}$ large only when κ_i and κ_j are both large. Let $\tilde{\kappa}_i^2 = \frac{\eta_1^2}{m^2}\kappa_i^2$. Then, $0 \leq \tilde{\kappa}_i^2 \leq 1$ and $\tilde{\kappa}_i$ approaches 1 as $\lambda_i \to \infty$. Since, $\tilde{\kappa}_i^2$ and $\tilde{\kappa}_i^2$ are bounded by 1, $\tilde{\kappa}_i^2\tilde{\kappa}_i^2$ will only be close to 1 when each term is close to one. That is, when both λ_i *and* λ_j are large, or equivalently when κ_i and κ_j are both large. Hence, it suffices to let

$$\begin{aligned} \theta_{x_i x_j} \mid \eta_1, \kappa &\sim \mathcal{N}(0, \xi^2 \tilde{\kappa}_j^2 \tilde{\kappa}_j^2) \\ &= \mathcal{N}(0, \eta_2^2 \kappa_i^2 \kappa_j^2) \quad \text{for} \quad \eta_2 := \frac{\eta_1^2}{m^2}\xi, \end{aligned} \tag{28}$$

to promote strong-hierarchy, where ξ has the interpretation of the scale of the non-zero interaction effects; as λ_i and λ_j tend to infinity, the prior variance of $\theta_{x_i x_j}$ approaches ξ^2. Again, since we might not now this scale, we draw

$$\xi^2 \sim \text{InvGamma}(\alpha_2, \beta_2), \tag{29}$$

for some choice of hyperparameters α_2 and β_2.

E. Woodbury Identity and the Matrix Determinant Lemma

The Woodbury matrix identity implies that,

$$(A^{-1} + UU^T)^{-1} = A - AU(I_K + U^T AU)^{-1}U^T A, \tag{30}$$

where $A \in \mathbb{R}^{M\times M}$, $U \in \mathbb{R}^{M\times K}$, and I_K is the $K \times K$ identity matrix. The matrix determinant lemma implies that,

$$\det(A^{-1} + UU^T) = \det(I + U^T AU)\det(A^{-1}) \tag{31}$$

Then, by the Woodbury identity,

$$\Sigma_{\tau,N} = (\Sigma_\tau^{-1} + \frac{1}{\sigma^2}\Phi_2(X)^T\Phi_2(X))^{-1} = \Sigma_\tau - \Sigma_\tau\Phi_2(X)^T(I_N + \Phi_2(X)\Sigma_\tau\Phi_2(X)^T)^{-1}\Phi_2(X)\Sigma_\tau. \tag{32}$$

Computing $p(D \mid \tau)$ requires computing $\det(\Sigma_{\tau,N})$. By the matrix determinant lemma,

$$\det(\Sigma_{\tau,N}) = (\det(I_N + \Phi_2(X)\Sigma_\tau\Phi_2(X)^T)\det(\Sigma_\tau^{-1}))^{-1}. \tag{33}$$

When Σ_τ is diagonal, the determinant equals the product of the diagonal, and its inverse equals one over the diagonal. Both of these quantities can be computed in $O(p^2)$ time. Hence, the time complexity for computing $\det(\Sigma_{\tau,N})$ is dominated by computing $\det(I_N + \Phi_2(X)\Sigma_\tau\Phi_2(X)^T)$, which takes $O(N^2p^2 + N^3)$ time and $O(Np^2)$ memory to store $\Phi_2(X)$.

F. Standard Polynomial Kernel

The feature map induced by the standard degree two polynomial kernel is given by,

$$\begin{aligned}\Phi^c_{\text{poly},2}(x) &\coloneqq (x_1^2, \cdots, x_p^2, \sqrt{2}x_1x_2, \cdots, \sqrt{2}x_{p-1}x_p, \sqrt{2c}x_1, \cdots, \sqrt{2c}x_p, c)\\ &= a_{poly,2} \odot \Phi_2(x),\ a_{poly,2} \coloneqq (1, \cdots, 1, \sqrt{2}, \cdots, \sqrt{2}, \sqrt{2c}, \cdots, \sqrt{2c}, c).\end{aligned} \tag{34}$$

Hence, Eq. (34) implies that

$$\text{diag}(\Sigma_{poly,2}) = a_{poly,2} \odot a_{poly,2}. \tag{35}$$

Eq. (35) shows that the prior covariance of the interaction terms are given higher prior variance than the main effects when $c \leq 1$, which is often undesirable. Furthermore, this prior does not promote sparsity, which is typically expected in high-dimensional problems.