On learning functions over biological sequence space: relating Gaussian process priors, regularization, and gauge fixing

Samantha Petti[1*], Carlos Martí-Gómez[2], Justin B. Kinney[2], Juannan Zhou[3], and David M. McCandlish[2]

[1]Department of Mathematics, Tufts University, Medford, MA, 02155.
[2]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724
[3]Department of Biology, University of Florida, Gainesville, FL, 32611

Abstract

Mappings from biological sequences (DNA, RNA, protein) to quantitative measures of sequence functionality play an important role in contemporary biology. We are interested in the related tasks of (i) inferring predictive sequence-to-function maps and (ii) decomposing sequence-function maps to elucidate the contributions of individual subsequences. Because each sequence-function map can be written as a weighted sum over subsequences in multiple ways, meaningfully interpreting these weights requires "gauge-fixing," i.e., defining a unique representation for each map. Recent work has established that most existing gauge-fixed representations arise as the unique solutions to L_2-regularized regression in an overparameterized "weight space" where the choice of regularizer defines the gauge. Here, we establish the relationship between regularized regression in overparameterized weight space and Gaussian process approaches that operate in "function space," i.e. the space of all real-valued functions on a finite set of sequences. We disentangle how weight space regularizers both impose an implicit prior on the learned function and restrict the optimal weights to a particular gauge. We also show how to construct regularizers that correspond to arbitrary explicit Gaussian process priors combined with a wide variety of gauges. Next, we derive the distribution of gauge-fixed weights implied by the Gaussian process posterior and demonstrate that even for long sequences this distribution can be efficiently computed for product-kernel priors using a kernel trick. Finally, we characterize the implicit function space priors associated with the most common weight space regularizers. Overall, our framework unifies and extends our ability to infer and interpret sequence-function relationships.

1 Introduction

A fundamental goal of biology is to understand how sequence-level differences in DNA, RNA or protein result in different observable outcomes. This mapping from DNA, RNA or protein sequences to some quantitative measure of sequence functionality, e.g. the growth rate (fitness) of a microbe or binding affinity of a protein, can be difficult to predict and interpret because combinations of mutations interact in complex ways [34, 42, 47, 54, 10, 59, 21]. Recent technological advances have produced datasets that can help us characterize such relationships on unprecedented scales. It is now possible to construct libraries of millions of sequences and simultaneously measure an associated

[*] E-mail: samantha.petti@tufts.edu

function for each sequence [15, 23, 2]. This has given rise to interest in method development for characterizing and interpreting sequence-function relationships [4, 5, 19, 35, 43, 30, 58, 36, 49, 59, 6, 13, 14, 32, 8, 25]. When successful, such methods can provide biological insight into the mechanisms that determine the function of the sequence and can be applied to practical problems such as protein engineering [55, 16, 27].

Here we focus on two related tasks: (i) inferring sequence-function maps, and (ii) decomposing such maps to elucidate the contribution of individual subsequences (including gapped subsequences whose positions are not necessarily contiguous). An accurate approximation of the sequence-function map is useful in that it can be applied to make predictions about the function of sequences that have not been measured as well as to mitigate the influence of measurement noise. Decomposing a sequence-function map into contributions from individual subsequences, which correspond to particular chemical subunits that are present in the cell, is one important strategy for interpreting sequence-function mappings and is particularly useful for suggesting mechanistic hypotheses that explain the observed sequence-function data. However, interpreting a decomposition into contributions of individual subsequences is made complicated by a lack of identifiability.

To see the simplest way in which this lack of identifiability arises, note that in principle we can define one feature per subsequence and hence learn a weight for each. The predicted function of a sequence can then be obtained by summing the weights corresponding to all of its subsequences. For sequences of length ℓ on an α character alphabet, we can view the weights as an $(\alpha+1)^\ell$-dimensional vector in *weight space* where each dimension corresponds to a possible subsequence. This model is non-identifiable: a particular sequence-function map can be expressed by many different weight space vectors because there are only α^ℓ possible sequences for which to measure values, leading to $(\alpha+1)^\ell - \alpha^\ell$ extra degrees of freedom. Thus, meaningfully interpreting the weights requires "fixing the gauge," [38] that is, imposing additional constraints to ensure that the learned vector of weights lies in a particular gauge, where a gauge is defined as a subset of weight space in which each sequence-function map can be expressed uniquely. Weights have different interpretations in different gauges, and indeed, strategic choices of gauge can help guide the exploration and interpretation of complex functional landscapes [38]. Importantly, this need to fix the gauge arises even in the simplest models of sequence-function relationships such as pairwise interaction models [52, 12, 48], and essentially occurs for any model whose form respects certain symmetries of the space of possible sequences [37], meaning that the issue of how to appropriately fix the gauge arises quite generally.

Recent work established that gauges that take the form of linear subspaces arise as the set of unique solutions to L_2-regularized regression in weight space [38], where the choice of positive-definite regularizer specifies the gauge. However, such regularizers result in two very different types of shrinkage. First, a regularizer designed for a particular gauge imposes a penalty on the component of each weight vector that is not in the gauge, which ultimately shrinks that component to zero and hence enforces that the optimal solution lies in the corresponding gauge space. Second, any positive-definite regularizer on the weights also enforces a pattern of shrinkage on the α^ℓ-dimensional vector of estimates produced by the regression procedure, where each dimension corresponds to a sequence and the value of a sequence represents our estimate of some measured function of the sequence. That is, by shrinking the estimated values of the weights, the regularizer also produces shrinkage in "function space," but the geometric features and biological interpretation of this shrinkage remains unclear.

A different approach to modeling sequence-function mappings is given by Gaussian process regression [39], which works by directly specifying a Bayesian prior over function space. To formulate Gaussian process regression in *function space*, we consider each sequence-function map as an α^ℓ-dimensional vector where each dimension corresponds to a sequence and the value of a sequence represents some measured function of the sequence. Then we assume the function space vector of measured values is drawn from a multivariate Gaussian distribution. The covariance of this distribution, which is specified by a kernel matrix, expresses a prior belief about which sequences should have similar measured function values. Gaussian process regression yields a posterior distribution over sequence-function maps. A variety of kernels for Gaussian process regression have been proposed and applied to successfully predict sequence-function maps and quantify the uncertainty of predictions [45, 50, 41, 56, 59, 1, 57].

Here, we provide a treatment of the relationship between gauge fixing, L_2-regularized linear regression, and Gaussian process regression for learning sequence-function mappings. These connections arise naturally due to the well-known connection between L_2-regularized linear regression and Bayesian linear regression in the case where the prior on the weights is multivariate Gaussian. From this point of view, besides fixing the gauge, L_2-regularization corresponds to implicitly imposing a Bayesian prior on weight space. We will show how these implicit priors on weight space can be translated into priors on function space, thus clarifying the relationship between regularized regression in $(\alpha+1)^\ell$-dimensional weight space and Gaussian process regression in α^ℓ-dimensional function space.

In addition to relating function space Gaussian processes to the implicit priors induced by weight space regularized regression, we consider the gauge-fixed weights of draws from posterior distributions of function space Gaussian processes. These distributions are fundamentally different than posterior distributions of weights achieved via Bayesian regression: while it is possible to design Gaussian priors on weights such that the maximum a posteriori (MAP) estimate is guaranteed to be in a certain gauge, the support of this posterior is always the entire weight space. This is a problem because the lack of gauge fixing greatly decreases the biological interpretability of these posterior draws. Analyzing the gauge-fixed weights corresponding to draws from the posterior distribution of a Gaussian process in function space allows us to express uncertainty in the weights while maintaining the interpretability gained by fixing the gauge. Naively, to compute the posterior distribution of gauge-fixed weights or even simply the gauge-fixed weights of the MAP estimate of a function space Gaussian process, one would need to compute an entire sequence-function map and project it into the gauge space of interest. However, for even moderately sized α and ℓ, this approach is intractable as function space is α^ℓ-dimensional. We introduce a kernel trick that allows us to efficiently compute any subset of the gauge-fixed weights corresponding to the MAP estimate or posterior distribution for function space Gaussian processes. These results allow us to explicitly specify the function space prior via an appropriate Gaussian process, analyze the results via interpretable gauge-fixed contributions of sub-sequences, and provide uncertainty estimates on these gauge-fixed parameters, all in a computationally tractable manner. Moreover, our kernel trick is more general; it can be applied to compute the posterior distribution of most other previously proposed representations of real-valued functions over sequence space including background-averaged epistasis coefficients [14] and coefficients of the function in the Fourier basis [6].

1.1 Summary of our contributions

Our contributions can be summarized as follows:

1. The choice of regularizer for weight space regression both induces a prior on function space and determines the gauge of the optimal solution. We show that for any Gaussian prior on function space and any linear gauge space, there exists a regularizer that induces the prior and whose optimum is in the linear gauge space (Theorem 1 of Section 3). Moreover, we establish that such regularizers can be constructed by taking the sum of two matrices: one that determines the prior on function space and the other that determines the gauge. We construct such matrices for priors and gauges of interest in Section 4.

2. One way to analyze a complex sequence-function map is to express it in a particular gauge and interpret the weights of each subsequence. Given measurements for a fraction of the possible sequences, how do we infer corresponding gauge-fixed weights and quantify the uncertainty? In Theorem 5 of Section 5, we establish a kernel trick that allows us to efficiently compute the posterior distribution of a large class of linear transformations of functional values without having to explicitly compute an α^ℓ-dimensional sequence-function map. In Theorem 7 of Section 6, we show how to apply this result to gauge-fixed weights. Given a training set of size t, any set of j subsequences, and a function space prior with a product form, we can compute the distribution over j gauge-fixed weights corresponding to the posterior distribution of the function space Gaussian process. Doing so only requires matrix and vector operations with dimensions at most j and t.

3. Diagonal matrices are a natural choice for regularizers in weight space. However, it is not clear a priori what form of shrinkage these regularizers induce in function space. Treating the diagonal regularizers as corresponding to independent zero-mean Gaussian priors on the values of the individual weights, we derive analytic formulas for the function space priors implicitly induced by these diagonal regularizers, and in doing so demonstrate how the choice of diagonal regularizer controls the rate of correlation decay of these induced priors (Section 7).

1.2 Related work

The flexibility of Gaussian processes make them an attractive method for modeling complex sequence-relationships so that now many families of kernels have been considered [45, 50, 41, 56, 59, 1]. Here we focus on isotropic kernels and non-isotropic product kernels in which each feature corresponds to a sequence position, see [59, 57]. Due to the mathematical structure of such kernels and recent advances in GPU acceleration [17, 51], these kernels are tractable for inference with hundreds of thousands of sequences and have been shown to exhibit state-of-the-art predictive performance [59, 57]. Our work is a conceptual bridge between these families of function space priors and the theory of gauge-fixing for parameter interpretation explored in [38].

Our work also provides a bridge between the inference and analysis of empirical sequence-function relationships and the theoretical literature on “fitness landscapes”[24, 3]. In particular, special cases of the function space and weight space priors we consider here are equivalent to several notable models in the fitness landscape literature including the connectedness model [40] (see Section 4.2) and the NK and GNK models [22, 7, 28, 20, 6] (see Section 7).

1.3 Applicability beyond biological sequences

While our results were developed with biological sequences in mind, they apply to regression and Gaussian processes over arbitrary finite discrete product spaces. We treat each position in a biological sequence as a categorical variable whose value indicates the character at the position; our models do not take into account sequence order. Therefore, we can apply our results to learn functions of any finite number of categorical features, including the common case where all features are binary. Our results apply directly if the number of categories for each feature is constant and can be extended if not. Example applications of prediction over spaces of categorical features include predicting patient outcomes based on the presence or absence of risk factors and treatments, predicting the productivity of microbial communities based on species composition, and making predictions based on responses to multiple choice surveys [29, 44, 9].

1.4 Outline

We begin with a preliminaries section to introduce notation (Section 2.1), define gauge-fixing and highlight how the choice of gauge can guide the interpretation of the function (Section 2.2), and introduce Gaussian processes on sequence space (Section 2.3). In Section 3 we establish the relationship between regularized regression in weight space and Gaussian process regression in function space. In doing so, we derive a general formula for weight space regularizers in terms of the induced function space prior and the gauge of the optimizer. In Section 4, we describe how to design regularizers for gauges of interest and various function space kernels that have been shown to perform well in practice. Then, in Section 5, we describe a kernel trick for computing the distribution of a class of linear transformations of functional values corresponding to draws from the posterior distributions of Gaussian processes, and in Section 6 we demonstrate how to apply this kernel trick to compute the posterior distribution of gauge-fixed weights. Finally, in Section 7 we describe the function space priors implicitly induced by diagonal weight space regularizers.

2 Preliminaries

2.1 Notation.

Let $\mathcal{A}$ be the alphabet of characters with $\alpha = |\mathcal{A}|$, and let ℓ be the length of the sequences. Our goal is to learn mappings of the form $f : \mathcal{A}^\ell \to \mathbb{R}$. We can equivalently consider each such map as a vector in $\mathbb{R}^{\alpha^\ell}$ indexed by sequences $x \in \mathcal{A}^\ell$. We refer to $\mathbb{R}^{\alpha^\ell}$ as *function space.*

Let $\mathcal{S}$ denote the set of possible subsequences,

$$\mathcal{S} = \{(S, s) : S \subseteq [\ell], s \in \mathcal{A}^{|S|}\},$$

where $[\ell] = \{1, 2, \ldots \ell\}$, S denotes the set of positions, and s denotes the sequence of length $|S|$ corresponding to the characters present at those positions. For a sequence x, let $x[S]$ denote the characters that appear at the positions in S. For example, if $x = abcde$, $x[\{2, 5\}] = be$. Let $w_{S,s}$ be denote the weight for subsequence (S, s). For ease of notation, when $S = \emptyset$ and s is the empty string, we write $w_\emptyset$ to refer to the corresponding coefficient. Note $|\mathcal{S}| = (\alpha + 1)^\ell$. A real-valued function on sequence space can be written as the weighted sum of indicator functions on these subsequences,

$$f(x) = \sum_{(S,s)\in\mathcal{S}} w_{(S,s)}\delta_{x[S]=s}.$$

Alternately, we can write this expression as

$$f = \Phi w$$

where f is an α^ℓ-dimensional vector indexed by sequences with $f_x = f(x)$, w is a $(\alpha+1)^\ell$-dimensional vector of weights indexed by the subsequence features, and Φ is an $\alpha^\ell \times (\alpha + 1)^\ell$ matrix indexed by sequences and subsequence features with $\Phi_{x,(S,s)} = \delta_{x[S]=s}$. We refer to $\mathbb{R}^{(\alpha+1)^\ell}$ as *weight space.* The non-identifiablity of representing a function as a vector in weight space is clear from the dimensionality; $w \in \mathbb{R}^{(\alpha+1)^\ell}$ and $f \in \mathbb{R}^{\alpha^\ell}$, so there are many w such that $f = \Phi w$.

All vectors and matrices we consider are indexed by sequences or subsequences and the order does not matter. We therefore use the notation M_U to restrict a matrix to the rows corresponding to the elements in the set U. Let X be a set of training sequences, and let y be corresponding observed measurements. We use Φ_X to denote the matrix Φ restricted to the rows corresponding to sequences in the training set and f_X to denote the function vector restricted to the sequences in the training set. When two sets appear as subscripts, e.g. $M_{X,Y}$ we restrict the rows of M to the elements of X and the columns to the elements in Y.

Throughout we let W and Λ be $(\alpha+1)^\ell \times (\alpha+1)^\ell$ dimensional matrices indexed by subsequences, i.e. indexed by the elements of $\mathcal{S}$. We use W when the matrix is a covariance matrix and Λ when the matrix is used as a regularizer. Similarly, we let K and Δ be $\alpha^\ell \times \alpha^\ell$-dimensional matrices indexed by sequences, and use K when the matrix is a covariance matrix and Δ when it is used as a regularizer. We let $N(\mu, K)$ denote the multivariate Gaussian distribution with mean μ and covariance K.

2.2 Gauges

Here we review the theory of gauge-fixing for biological sequences established in [38]. There are many vectors in weight space that give rise to the same real-valued function on sequence space. For example, given a vector w in weight space, we can add any real value a to the weight for the empty set, $w'_\emptyset = w_\emptyset + a$, and subtract a from the weights for the subsequences on the first position, $w'_{(\{1\},c)} = w_{(\{1\},c)} - a$ for each character $c \in \mathcal{A}$. The resultant weight space vector w' produces the same function on sequence space as w, $\Phi w = \Phi w'$. The space of gauge freedoms describes all such directions in weight space along which moving does not change the corresponding function on sequence space, i.e. vectors g such that $\Phi(w+g) = \Phi w$. The following definition gives an equivalent characterization.

Definition 1. *The space of gauge freedoms G is the subspace of weight space defined by*

$$G = \{w \in \mathbb{R}^{(\alpha+1)^\ell} : \Phi w = 0\}.$$

Definition 2. *A subspace of weight space Θ is a linear gauge space if it is complementary to the space of gauge freedoms G.*

Note that for a fixed linear gauge space Θ, every function on sequence space can be represented as precisely one vector in Θ.

2.2.1 The λ-π family of gauges

We focus on the λ-π family of gauges introduced in [38]. Every gauge $\Theta^{\lambda,\pi}$ in this family is defined by two parameters: $\lambda \in [0,\infty]$ and a product distribution defined by assigning probabilities to the characters at each position $\pi(x) = \prod_{p\in[\ell]} \pi^p_{x_p}$. The parameter λ controls the relative magnitudes of the weights between longer and shorter subsequences. When $\lambda = 0$ we have the "trivial gauge" in which the weights for all subsequences of length less than ℓ are zero; as λ increases more weight is pushed to the shorter subsequences. The distribution π controls the relative magnitude of weights for different subsequences as a function of how likely they are under π.

Formally, λ-π gauges are tensor products of single-position gauges Θ^{λ,π^p} of the following form

$$V_\lambda = span\left\{\begin{pmatrix}\lambda\\1\\\vdots\\1\end{pmatrix}\right\}, \quad V_\perp^{\pi^p} = \left\{\begin{pmatrix}0\\v_{c_1}\\\vdots\\v_{c_\alpha}\end{pmatrix} : \sum_{c\in\mathcal{A}} v_c \pi^p_c = 0\right\}, \quad \Theta^{\lambda,\pi^p} = V_\lambda \oplus V_\perp^{\pi^p},$$

where π^p is a distribution over characters at position p, $\sum_{c\in\mathcal{A}} \pi^p_c = 1$ and $\pi^p_c \geq 0$ for all characters $c \in \mathcal{A}$. The gauge space $\Theta^{\lambda,\pi}$ is the tensor product

$$\Theta^{\lambda,\pi} = \bigotimes_{p=1}^{\ell} \Theta^{\lambda,\pi^p}.$$

The projection matrix into gauge $\Theta^{\lambda,\pi}$ is given by

$$P^{\lambda,\pi}_{(S,s),(T,t)} = \prod_{p\in S\cap T} \left(\delta_{s_p=t_p} - \pi^p_{t_p}\eta\right) \prod_{p\in S\setminus T} (1-\eta) \prod_{p\in T\setminus S} \pi^p_{t_p}\eta \prod_{p\notin S\cup T} \eta \tag{1}$$

where $\eta = \lambda/(1+\lambda)$ (see [38]). Note that projection matrix for $\lambda = \infty$ is well-defined as this simply corresponds to the case $\eta = 1$.

We now highlight two gauges of interest in the λ-π family; for further discussion of these and other specific gauges in the family see [38]. Additionally, in Section 4.1.1 we establish a new marginalization property for λ-π gauges and use this to further interpret the role of λ and π.

1. Hierarchical gauge. Hierarchical gauges are obtained by taking $\lambda = \infty$ or equivalently $\eta = 1$ in Equation (1). The zero-sum gauge is the hierarchical gauge with π as the uniform distribution. In this gauge, the mean function value for sequences with a particular subsequence (S, s) can be expressed simply in terms of the weights. The mean function value over all sequences is $w_\emptyset$, the mean function value over all sequences with c at position p is $w_\emptyset + w_{(\{p\},c)}$, the mean function value over all sequences with c at position p and c' at position p' is $w_\emptyset + w_{(\{p\},c)} + w_{(\{p'\},c')} + w_{(\{p,p'\},cc')}$, and so on. The function obtained by summing the weights on subsequences of length up to k is the least-squares approximation of the function with up to k^{th} order terms. General hierarchical gauges can be interpreted similarly with π as the distribution used to compute the mean and least-squares approximation.

2. Wild-type gauge. The wild-type gauge is obtained by taking the limit as π approaches the probability distribution that has support only at a fixed "wild-type" sequence (see [38] for details). In the wild-type gauge, weights of subsequences that agree with the wild-type sequence at any position are zero, except for $w_\emptyset$ which gives the function value of the wild-type sequence. The weight $w_{(\{p\},c)}$ quantifies the effect of the mutation c at position p on the wild-type background, while higher order weights $w_{(S,s)}$ quantify the epistatic (i.e. interaction) effect due to the group of mutations (S, s).

2.2.2 Gauge fixing and regularization

In regularized regression over weight space, the choice of the positive definite regularizer Λ determines the gauge of the optimal solution, e.g. for every positive-definite Λ there is a gauge such that the optimizer always lies in this gauge.

Definition 3. *Let Λ be a positive definite matrix, and define*

$$w^{OPT}(\Lambda,\beta) = \underset{w\in\mathbb{R}^{(\alpha+1)^\ell}}{\arg\min}\ \|y - \Phi_X w\|_2^2 + \beta w^T \Lambda w.$$

We say Λ is a Θ-regularizer if Λ is positive definite and $w^{OPT} \in \Theta$ for all $\beta > 0$, sequences X, and measurements y.

In Section 4, we will discuss how to construct $\Theta^{\lambda,\pi}$ regularizers.

2.3 Gaussian process regression in sequence space.

We now outline Gaussian process regression as applied to sequence space. A Gaussian process is defined by a Gaussian prior distribution on functions f given by a covariance matrix K called a *kernel*. We always assume K defines a proper prior, i.e. K is positive-definite.

Definition 4. *Suppose $y = f_X + \varepsilon$ where $f \sim N(0, K)$ and $\varepsilon \sim N(0, \sigma^2 I)$. Let $f^{MAP}(K, \sigma^2)$ be the Maximum a Posteriori (MAP) estimate for f under this Gaussian process.*

It is well known that the posterior distribution of f is given by

$$f \sim N(K_{*,X}\left(K_{X,X} + \sigma^2 I\right)^{-1} y, K_{*,*} - K_{*,X}\left(K_{X,X} + \sigma^2 I\right)^{-1} K_{X,*})$$

where the subscript X restricts K to the rows and/or columns corresponding to sequences in the training set X and the subscript $*$ indicates all rows and/or columns, and that moreover the MAP estimate is given by the mean of this posterior distribution (see [39]).

In regularized regression over weight space, the choice of the positive-definite regularizer Λ implicitly induces a prior distribution on function space in addition to determining the gauge of the optimal solution. We say that a regularizer Λ induces the prior defined by K if the optimal weights under the regularizer Λ yield the function consistent with the MAP estimate under K.

Definition 5. *We say a regularizer Λ induces the prior K if $f^{MAP}(K, \sigma^2) = \Phi w^{OPT}(\Lambda, \sigma^2)$ for all $\sigma^2 > 0$.*

In Section 4.2 we define two broad families of kernels that we will consider: variance component kernels [59] and product kernels [57].

3 The relationship between regularization, Gaussian process priors, and gauge fixing

In this section we affirmatively resolve the question: for any linear gauge Θ and prior K on function space, does there exists a Θ-regularizer Λ that induces the prior K?

Theorem 1. *For any linear gauge Θ and prior K on function space, there exists a Θ-regularizer that induces the prior K. The matrix $\Lambda = \Phi^T K^{-1} \Phi + B^T B$ where B is a matrix with nullspace Θ is one such regularizer.*

The theorem provides a recipe for building a Θ-regularizer that induces a function space prior K. Such a regularizer can be written as the sum of two matrices: one that determines the induced prior and one that determines the gauge. In Section 4.1, we derive a simple matrix of the form $B^T B$ for the λ-π gauges, then in Section 4.3 we compute $\Phi^T K^{-1} \Phi$ for several useful classes of prior. We can also phrase Theorem 1 in terms of a Bayesian regression in weight space rather than Gaussian process regression in function space.

Corollary 1. *For any linear gauge Θ and prior on function space K, there exists a Gaussian prior W over weight space that both induces the function space prior K, meaning $w \sim N(0, W)$ implies $\Phi w \sim N(0, K)$, and enforces the specified gauge, meaning $w \sim N(0, W)$ implies $w^{MAP} \in \Theta$.*

To prove Theorem 1, we first establish a new general condition for Θ-regularizers in Section 3.1. Then in Section 3.2 we establish the equivalence of MAP estimates and optimal solutions of regularized regression across function and weight space and prove Theorem 1.

3.1 Matrices that act as Θ-regularizers

Posfai et al. establish a sufficient condition for a positive-definite matrix to act as a Θ-regularizer [38].

Definition 6. *Let V_1 and V_2 be complementary subspaces of an m-dimensional vector space V. An $m \times m$ positive-definite matrix Λ orthgonalizes V_1 and V_2 if for all $v_1 \in V_1$ and $v_2 \in V_2$,*

$$v_1^T \Lambda v_2 = 0.$$

If Λ orthgonalizes V_1 and V_2, we say that V_1 and V_2 are Λ-orthogonal.

The following lemma is a special case of Claim 6 of Posfai et al. [38].

Lemma 1. *Let Θ be a gauge space, and let G be the space of gauge freedoms. If Λ orthogonalizes Θ and G, then Λ is a Θ-regularizer.*

Moreover, Posfai et al. establish two conditions that are equivalent to Λ orthogonalizing a pair of complementary subspaces V_1 and V_2, see Lemma 7. In the following lemma, we establish another equivalent condition that we will use in the proof of Theorem 1, see Appendix A.1 for the proof.

Lemma 2. *Let V_1 and V_2 be complementary subspaces. A matrix Λ orthogonalizes V_1 and V_2 if and only if $\Lambda = A^T A + B^T B$ for matrices A and B such that* $\text{nullspace}(A) = V_2$ *and* $\text{nullspace}(B) = V_1$.

3.2 Equivalence of MAP estimates and optimizers of regularized regression

In this section we will establish the equivalence of the following prediction methods:

1. **Penalized regression in weight space** (Definition 3).

$$w^{OPT}(\Lambda, \beta) = \underset{w \in \mathbb{R}^{(\alpha+1)^\ell}}{\arg\min} \; \|y - \Phi_X w\|_2^2 + \beta w^T \Lambda w$$

2. **Gaussian process regression in function space** (Definition 4). Suppose $y = f_X + \varepsilon$ where $f \sim N(0, K)$ and $\varepsilon \sim N(0, \sigma^2 I)$. Let $f^{MAP}(K, \sigma^2)$ be the MAP estimate for f under this Gaussian process.

To do so, we introduce two additional prediction methods as intermediaries and establish the equivalences depicted in Figure 1.

3. **Penalized regression in function space.**

$$f^{OPT}(\Delta, \beta) = \underset{f \in \mathbb{R}^{\alpha^\ell}}{\arg\min} \|y - f_X\|_2^2 + \beta f^T \Delta f$$

4. **Bayesian regression in weight space.** Suppose $y = \Phi_X w + \varepsilon$ where $w \sim N(0, W)$ and $\varepsilon \sim N(0, \sigma^2 I)$. Let $w^{MAP}(W, \sigma^2)$ be the MAP estimate for w under this Gaussian prior.

With respect to the last of these methods, it is well-known that the posterior distribution of w is

$$N(w^{MAP}, A^{-1}) \quad \text{where} \quad w^{MAP} = \sigma^{-1} A^{-1} \Phi_X^T y \quad \text{and} \quad A = \sigma^2 \Phi_X^T \Phi_X + W^{-1}.$$

We begin with two well-known lemmas. Lemma 3 relates the optimizer of penalized regression and the MAP estimate of a Gaussian process (see Section 6.2 of [39]) and Lemma 4 establishes equivalence of the MAP estimates for Bayesian regression in weight space and Gaussian process regression in function space (see Section 2.1.2 of [39]).

$$\begin{tikzcd}
f^{OPT}(\Delta, \beta) \arrow[r, leftrightarrow, "\begin{array}{c}\text{Equal when } \Delta K = K\Delta = I \text{ and } \sigma^2 = \beta \\ \text{(Lemma 3)}\end{array}"] & f^{MAP}(K, \sigma^2) \\
w^{OPT}(\Lambda, \beta) \arrow[u, "\begin{array}{c}\Lambda = \Phi^T \Delta \Phi \\ +B^T B \implies \\ f^{OPT} = \Phi w^{OPT} \\ \text{(Lemma 5)}\end{array}"] \arrow[r, leftrightarrow, "\begin{array}{c}\text{Equal when } \Lambda W = W\Lambda = I \text{ and } \sigma^2 = \beta \\ \text{(Lemma 3)}\end{array}"'] \arrow[d, "\begin{array}{c}\Lambda = A^T A + B^T B \\ \text{orthogonalizes} \\ \text{gauge } \Theta \\ \text{(Lemmas 1 and 2)}\end{array}"'] & w^{MAP}(W, \sigma^2) \arrow[u, "\begin{array}{c}K = \Phi W \Phi^T \implies \\ f^{MAP} = \Phi w^{MAP} \\ \text{(Lemma 4)}\end{array}"'] \\
w^{OPT} \in \Theta &
\end{tikzcd}$$

Figure 1: An illustration of the equivalences established. Let A and B be matrices such that $nullspace(A) = G$ and $nullspace(B) = \Theta$.

Lemma 3. *Suppose $y = \Phi_X b + \varepsilon$ where $b \sim N(0, W)$ and $\varepsilon \sim N(0, \sigma^2 I)$. Then the MAP estimate for b is equal to*

$$\arg\min_b \|y - \Phi_X b\|_2^2 + \sigma^2 b^T W^{-1} b.$$

Lemma 4. *Suppose $w \sim N(0, W)$. Then $\Phi w \sim N(0, K)$ where $K = \Phi W \Phi^T$. Moreover, $f^{MAP}(K, \sigma^2) = \Phi w^{MAP}(W, \sigma^2)$.*

The following lemma establishes equivalence of the optimizers for regularized regression in weight space and function space.

Lemma 5. *Let B be a matrix with nullspace Θ, and let Δ be positive definite. If $\Lambda = \Phi^T \Delta \Phi + B^T B$ then $f^{OPT}(\Delta, \beta) = \Phi w^{OPT}(\Lambda, \beta)$ for all $\beta > 0$.*

Proof. We will use two facts. First, suppose $w \in \Theta$. Then

$$w^T \Lambda w = w^T (\Phi^T \Delta \Phi + B^T B) w = (\Phi w)^T \Delta (\Phi w).$$

Second, $w^{OPT} \in \Theta$. This follows directly from Claim 1 and Lemmas 1 and 2.

Let w_0 be the representation of f^{OPT} in gauge Θ, $\Phi w_0 = f^{OPT}$. By the first fact, $\left(f^{OPT}\right)^T \Delta f^{OPT} = (\Phi w_0)^T \Delta (\Phi w_0) = w_0^T \Lambda w_0$. It follows that

$$\psi = \|y - \Phi_X w_0\|_2^2 + \beta w_0^T \Lambda w_0 = \|y - f_X^{OPT}\|_2^2 + \beta \left(f^{OPT}\right)^T \Delta f^{OPT}.$$

The optimality of w^{OPT} and f^{OPT} imply that

$$\|y - \Phi_X w^{OPT}\|_2^2 + \beta \left(w^{OPT}\right)^T \Lambda w^{OPT} \leq \psi \leq \|y - \left(\Phi w^{OPT}\right)_X \|_2^2 + \beta \left(\Phi w^{OPT}\right)^T \Delta \left(\Phi w^{OPT}\right).$$

Note that $\|y - \Phi_X w^{OPT}\|_2^2 = \|y - \left(\Phi w^{OPT}\right)_X \|$. The fact that $w^{OPT} \in \Theta$ implies $\left(\Phi w^{OPT}\right)^T \Delta \left(\Phi w^{OPT}\right) = \left(w^{OPT}\right)^T \Lambda w^{OPT}$. Thus the upper and lower bounds are equal meaning that inequalities are satisfied at equality. Thus we have

$$\psi = \|y - \Phi_X w_0\|_2^2 + \beta w_0^T \Lambda w_0 = \|y - \Phi_X w^{OPT}\|_2^2 + \beta \left(w^{OPT}\right)^T \Lambda w^{OPT},$$

and the uniqueness the optimizer implies that $w^{OPT} = w_0$ meaning $f^{OPT} = \Phi w^{OPT}$, as desired. □

Claim 1. *Let Δ be positive-definite. Then $\Phi^T \Delta \Phi = A^T A$ for some matrix A with nullspace equal to the space of gauge freedoms G.*

Proof. Since Δ is positive-definite, $\Delta = Z^T Z$ for some invertible matrix Z, and we can write $\Phi^T \Delta \Phi = (Z\Phi)^T Z\Phi$. Note since Z is invertible, $Z\Phi v = 0$ implies $\Phi v = 0$, so the nullspace of $Z\Phi$ is equal to the nullspace of Φ, which is G. □

With this result established, we can now prove Theorem 1, which describes how to build a Θ-regularizer that induces a function space prior K.

Proof. (of Theorem 1) Take $\Lambda = \Phi^T K^{-1} \Phi + B^T B$ where B is a matrix with nullspace Θ. Claim 1 and Lemmas 1 and 2 directly imply that Λ is a Θ-regularizer. The fact that Λ induces the prior K follows directly from Lemmas 3 and 5. □

4 Building gauge-specific regularizers that induce variance component and product kernel priors

In this section we demonstrate how to build Θ-regularizers that induce function space priors K for a useful class of gauges Θ and two useful classes of function space priors K. Recall that for a matrix B with nullspace Θ,

$$\Lambda = \Phi^T K^{-1} \Phi + B^T B$$

is a Θ-regularizer that induces prior K. This additive form gives us a recipe for building Θ-regularizers that induce a specific function space prior K, where the regularizer takes the form of a sum of two matrices, one that determines the induced prior and the other that determines the gauge. In Section 4.1, we derive a simple formula for a matrix of the form $B^T B$ where the nullspace of B is $\Theta^{\lambda,\pi}$. Then in Section 4.3 we compute $\Phi^T K^{-1} \Phi$ for VC and product kernels. Together these results serve as a recipe for constructing regularizers that induce a particular VC or product prior in a particular λ-π gauge.

4.1 Building Θ-regularizers for λ-π gauges

We first establish a marginalization property for λ-π gauges (Section 4.1.1) that helps to interpret the gauge parameters. Then in Section 4.1.2 we employ this marginalization property to build simple matrices of the form $B^T B$ where the nullspace of B is a λ-π gauge $\Theta^{\lambda,\pi}$.

4.1.1 A marginalization property for λ-π gauges

Definition 7. *We say that a vector $w \in \mathbb{R}^{(\alpha+1)^\ell}$ satisfies the λ-π marginalization property if for all (U, u, p) where U is a subset of positions that does not contain p and u is a subsequence on U,*

$$\sum_{c\in\mathcal{A}} \pi_c^p w_{(U\cup\{p\},u^{+c})} = \frac{w_{(U,u)}}{\lambda} = \left(\frac{1-\eta}{\eta}\right) w_{(U,u)}, \tag{2}$$

where $\eta = \lambda/(1+\lambda)$, and u^{+c} denotes the sequence on $U \cup \{p\}$ that agrees with u on U and has character c at position p.

Lemma 6 establishes that w satisfies the λ-π marginalization property if and only if w is in the λ-π gauge. This marginalization property was established for the hierarchical gauge ($\lambda = \infty, \eta = 1$) in Claim 21 of [38]. In the zero-sum gauge (hierarchical gauge with uniform π), the weights for any set of α subsequences on the same set of positions that differ only at one particular position average to zero. Direct consequences of this are the interpretation of the weights as mean effects and the fact that the truncated model (the function obtained by summing weights on subsequences of length up to k)is the least-squares approximation of the function with up to k^{th} order terms, as described in Section 2.2.1. For a non-uniform product distribution π, the results are analogous with a weighted average and weighted least squares with respect to π.

Our marginalization property is novel for finite λ, and its interpretation provides insight into how λ affects magnitude of the weights as a function of the size of the corresponding subsequence. The hierarchical ($\lambda = \infty$) case can be thought of as pushing as much information into the lower order weights as possible; the best k^{th} order approximation can be obtained by truncating up to weights corresponding to subsequences of length k. Consider the other extreme when λ is very small. The weights for any set of α subsequences on the same set of k positions that differ only at one particular position average to $1/\lambda$ times the weight for the length $k-1$ subsequence in common, yielding relatively larger weights for longer subsequences. As $\lambda \to 0$, all except the highest order weights (corresponding to length ℓ subsequences) vanish, yielding the trivial gauge.

Lemma 6. *A vector of weights w satisfies the λ-π marginalization property if and only if $w \in \Theta^{\lambda,\pi}$.*

Proof. First we show that if $w \in \Theta^{\lambda,\pi}$, then w satisfies the λ-π marginalization property. It suffices to fix a basis for $\Theta^{\lambda,\pi}$ and show that each basis vector satisfies the λ-π marginalization property. We can choose a basis for $\Theta^{\lambda,\pi}$ where each basis vector has the form $w = \bigotimes_{p=1}^{\ell} \theta^p$ where $\theta^p \in \Theta^{\lambda,\pi^p}$. Note θ^p is an $\alpha + 1$ dimensional vector. We index the last α positions with the corresponding character c and index the first position with 0. Since only the V_λ component of w contributes to a nonzero value to θ_0,

$$\theta^p - \frac{\theta_0^p}{\lambda}\begin{pmatrix}\lambda\\1\\\vdots\\1\end{pmatrix} \in V_\perp^{\pi^p}, \qquad \text{so} \sum_{c\in\mathcal{A}} \pi_c^p\left(\theta_c^p - \frac{\theta_0^p}{\lambda}\right) = 0 \quad \text{and thus} \quad \sum_{c\in\mathcal{A}} \pi_c^p\theta_c^p = \frac{\theta_0^p}{\lambda}.$$

Let U be a subset of positions that does not contain p and u be a subsequence on U. We now show that the basis element w satisfies Equation (2). Note

$$w_{(U\cup\{p\},u^{+c})} = \zeta\theta_c^p \quad \text{and} \quad w_{(U,u)} = \zeta\theta_0^p \quad \text{where} \quad \zeta = \prod_{q\in U}\theta_{u_q}^q \prod_{q\notin U\cup\{p\}}\theta_0^q.$$

It follows that

$$\sum_{c\in\mathcal{A}} \pi_c^p w_{(U\cup\{p\},u^{+c})} = \zeta \sum_{c\in\mathcal{A}} \pi_c^p \theta_c^p = \frac{\zeta\theta_0^p}{\lambda} = \frac{w_{(U,u)}}{\lambda},$$

as desired.

Next we show that any w that satisfies the λ-π marginalization property is in $\Theta^{\lambda,\pi}$ by showing that $P^{\lambda,\pi}w = w$. Let (S,s) be an arbitrary subsequence. We will show that $(P^{\lambda,\pi}w)_{(S,s)} = w_{(S,s)}$, or equivalently

$$\sum_{(T,t)} b(T,t) w_{(T,t)} = w_{(S,s)},$$

where for ease of notation we let

$$b(T,t) = P^{\lambda,\pi}_{(S,s),(T,t)} = \prod_{p\in S\cap T}\left(\delta_{s_p=t_p} - \pi^p_{t_p}\eta\right) \prod_{p\in S\setminus T}(1-\eta) \prod_{p\in T\setminus S} \pi^p_{t_p}\eta \prod_{p\notin S\cup T}\eta.$$

Let F be a subset of positions, and let $\mathcal{T}_F$ be the set of subsequences that agree with s on positions in $S\cap F$ and do not include positions in $S^c\cap F$,

$$\mathcal{T}_F = \{(T,t) : t_p = s_p \text{ for all } p\in S\cap F\cap T, \text{ and } p\notin T \text{ for all } p\in S^c\cap F\}.$$

We iterative apply Claims 4 and 5 (proved in Appendix A.2) for each position $p\in S$ and each position $p\notin S$ respectively to obtain the desired result

$$\begin{aligned}\sum_{(T,t)\in\mathcal{T}_\emptyset} b(T,t)w_{(T,t)} &= \left(\prod_{p\in S}\frac{1}{1-\pi^p_{s_p}\eta}\prod_{p\notin S}\frac{1}{\eta}\right)\sum_{(T,t)\in\mathcal{T}_{\{1,2,\dots L\}}} b(T,t)w_{(T,t)} \\ &= \left(\prod_{p\in S}\frac{1}{1-\pi^p_{s_p}\eta}\prod_{p\notin S}\frac{1}{\eta}\right) b(S,s)w_{(S,s)} = w_{(S,s)}.\end{aligned}$$

□

4.1.2 A simple formula for B^TB for $\Theta^{\lambda,\pi}$ regularizers

We derive a simple formula for B^TB for $\Theta^{\lambda,\pi}$ regularizers that is as sparse as the Laplacian of the Hamming graph for words of length ℓ with alphabet size $\alpha+1$.

Theorem 2. *Let*

$$Z_{(S,s),(T,t)} = \begin{cases}(1-\eta)\,\eta^2 + \sum_{p\in S}\left(\pi^p_{s_p}\right)^2 & S=T,\ s=t\\ \pi^p_{s_p}\pi^p_{t_p} & S=T,\ d(s,t)=1 \textit{ with } s_p\neq t_p\ .\\ 0 & \textit{otherwise}\end{cases}$$

Then $Z = B^TB$ for a matrix B with null space equal to the gauge space $\Theta^{\lambda,\pi}$.

Proof. Let B be an $\ell(\alpha+1)^{\ell-1}\times(\alpha+1)^\ell$ matrix with columns indexed by subsequences (T,t) and rows indexed by triples (U,u,p) where U is a subset of positions that does not contain p and u is a subsequence on U. Recall that u^{+c} denotes the sequence on $U\cup\{p\}$ that agrees with u on U and has character c at position p. Define

$$B_{(U,u,p),(T,t)} = \begin{cases} \pi_c^p & T = U \cup \{p\}, t = u^{+c} \\ \frac{-(1-\eta)}{\eta} & T = U, t = u \\ 0 & otherwise, \end{cases}.$$

where $\eta = \lambda/(1+\lambda)$. Note that $w \in \text{nullspace}(B)$ is equivalent to w satisfying the λ-π property. Lemma 6 implies that $\Theta^{\lambda,\pi} = \text{nullspace}(B)$. It is straightforward to verify that $Z = B^T B$. □

4.2 Variance component and product kernels

Here we introduce two classes of kernels that have shown to perform well in practice: variance component kernels [59] and product kernels [57].

Variance component (VC) kernels are the class of isotropic kernels, i.e. kernels whose values depend only on the Hamming distance between the pair of sequences. These kernels are parameterized in terms of how much variance is due to different orders of interaction. We can decompose function space into orthogonal subspaces $V_0, V_1, \dots V_\ell$, where V_0 is the constant subspace and each V_k includes function vectors that express interactions between exactly k sites; for further details see [59, 26, 18]. Consequently we can decompose any vector uniquely as the sum of orthogonal vectors $f = \sum_{k=0}^{\ell} f_k$ where $f_k \in V_k$, and decompose the variance $f^T f = \sum_{k=0}^{\ell} f_k^T f_k$. For f drawn from a distribution, we define the *variance of order* k of this distribution as $\mathsf{E}\left(f_k^T f_k\right)$. Krawtchouk polynomials are used to build kernels that are parameterized by these variances.

Definition 8. *The Krawtchouk polynomial is*

$$\mathcal{K}_k(d) = \mathcal{K}_k(d; \ell, \alpha) = \sum_{i=0}^{k} (-1)^i (\alpha - 1)^{k-i} \binom{d}{i} \binom{\ell - d}{k - i}.$$

Definition 9. *Let $\lambda_0, \dots, \lambda_\ell > 0$. The variance component (VC) kernel is given by*

$$K_{x,y} = \sum_{k=0}^{\ell} \lambda_k \mathcal{K}_k(d(x,y))$$

where $d(x,y)$ denotes the Hamming distance between sequences x and y.

For f drawn from a VC kernel, $\mathsf{E}\left(f_k^T f_k\right) = \lambda_k \binom{\ell}{k} (\alpha - 1)^k$. The dimension of V_k is $\binom{\ell}{k} (\alpha - 1)^k$, so we call λ_k the *dimension-normalized variance of order* k. Note that λ_k can be interpreted as the mean squared interaction coefficients of order k, see [59].

We will also consider product kernels, which are not necessarily isotropic priors that can express the fact that different position-character combinations can play different roles. We recently introduced two subclasses of product kernels, Connectedness and Jenga kernels, that achieve state-of-the-art predictive performance on several sequence-function datasets and yield hyperparameters that can be interpreted to provide insight into the mechanisms by which the sequence determines the function [57]. We formally define these kernels in Appendix A.3.1.

Definition 10. *A product kernel on sequences of length ℓ has the form*

$$K = \bigotimes_{p=1}^{\ell} K^p,$$

where for $p \in [\ell]$, K^p is a symmetric $\alpha \times \alpha$ positive definite matrix.

4.3 Building regularizers that induce VC and product kernels.

First we compute $\Phi^T K^{-1}\Phi$ when K is a VC kernel. While this matrix is dense, it contains only order ℓ^2 distinct entries which can be precomputed.

Theorem 3. *Let K be a VC kernel, $K_{x,y} = \sum_{k=0}^{\ell} \lambda_k \mathcal{K}_k(d(x,y))$, then*

$$(\Phi^T K^{-1}\Phi)_{(S,s),(T,t)} = \alpha^{\ell-|S\cup T|} \sum_{d=j}^{\ell-(|S\cap T|-j)} \binom{\ell-|S\cap T|}{d-j}(\alpha-1)^{d-j}\left(\sum_{k=0}^{\ell}\lambda_k^{-1}\mathcal{K}_k(d)\right),$$

where j is the Hamming distance between s and t among their common positions $S\cap T$.

Proof. Note

$$\left(\Phi^T K^{-1}\Phi\right)_{(S,s),(T,t)} = \sum_x \Phi_{x,(S,s)}\left(\sum_y K^{-1}_{x,y}\Phi_{y,(T,t)}\right) = \sum_{\substack{x,y\\ x[S]=s,y[T]=t}} K^{-1}_{x,y}.$$

To compute this sum, we count how many pairs of sequences x,y are at distance d and satisfy $x[S]=s, y[T]=t$. In building such pairs, the subsequence of x on S is fixed and the subsequence of y on T is fixed. Let j be the Hamming distance between s and t among their common positions $S\cap T$. All valid pairs of sequences will have Hamming distance at least j. There are $\alpha^{\ell-|S\cup T|}$ choices for the subsequence of x on $(S\cup T)^c$. Having fixed one such subsequence on these positions in x, each position in $(S\cap T)^c$ is fixed in either x or y (in $S\setminus T$, x is fixed; in $T\setminus S$, y is fixed; in $(S\cup T)^c$, x is fixed). To build a pair of sequences at distance d, we need to choose $d-j$ positions from $(S\cap T)^c$ and pick characters so that x and y disagree at those positions. There are $\binom{\ell-|S\cap T|}{d-j}(\alpha-1)^{d-j}$ ways to do so. We apply Lemma 8 to compute K^{-1} and obtain

$$\left(\Phi^T K^{-1}\Phi\right)_{(S,s),(T,t)} = \sum_{d=j}^{\ell-(|S\cap T|-j)} \alpha^{\ell-|S\cup T|}\binom{\ell-|S\cap T|}{d-j}(\alpha-1)^{d-j}\left(\sum_{k=0}^{\ell}\lambda_k^{-1}\mathcal{K}_k(d)\right).$$

□

Next we compute $\Phi^T K^{-1}\Phi$ for product kernels; this matrix is again dense, but it can easily be computed as the tensor product of ℓ matrices with simple forms. From Theorem 4, it is straightforward to compute $\Phi^T K^{-1}\Phi$ for Jenga, Connectedness, and Geometric kernels. For completeness we include these computations in Appendix A.3.2.

Theorem 4. *Let K be a product kernel with $(K^p)^{-1}_{c,c'} = b^p_{c,c'}$. Then*

$$\left(\Phi^T K^{-1}\Phi\right)_{(S,s),(T,t)} = \prod_{p\in S\cap T} b^p_{x_p,y_p}\prod_{p\in S\setminus T}\left(\sum_{c\in\mathcal{A}} b^p_{s_p,c}\right)\prod_{p\in T\setminus S}\left(\sum_{c\in\mathcal{A}} b^p_{t_p,c}\right)\prod_{p\notin S\cup T}\left(\sum_{c,c'} b^p_{c,c'}\right)$$

Proof. Since $K^{-1} = \bigotimes_p (K^p)^{-1}$, $K^{-1}_{x,y} = \prod_p b^p_{x_p,y_p}$. Recall $\Phi_{X,(S,s)} = \delta_{x[S]=s}$. It follows that

$$\left(\Phi^T K^{-1}\Phi\right)_{(S,s),(T,t)} = \sum_{\substack{x,y\\ x[S]=s,y[T]=t}} K^{-1}_{x,y} = \sum_{\substack{x,y\\ x[S]=s,y[T]=t}} \prod_p b^p_{x_p,y_p}.$$

We construct pairs of sequences x and y for which $x[S]=s, y[T]=t$ and compute the factor that each position contributes to the summand.

- If $p \in S \cap T$, there is only one option: $x_p = s_p$ and $y_p = t_p$, and this position contributes a factor of $b^p_{x_p,y_p}$ to the summand.
- Consider $p \in S \setminus T$. We must have $x_p = s_p$ and y_p can take any value. If $y_p = c$, this position contributes a factor of $b^p_{x_p,c}$ to the summand. The case $p \in T \setminus S$ is analogous.
- Consider $p \notin S \cup T$. Then x_p and y_p can take any value. If $x_p = c$ and $y_p = c'$, this position contributes a factor of $b^p_{c,c'}$ to the summand.

Note that each term in the expansion of the following expression corresponds to the $K^{-1}_{x,y}$ value for a pair of sequences with the property that $x[S] = s, y[T] = t$. The result follows.

$$\prod_{p \in S \cap T} b^p_{x_p,y_p} \prod_{p \in S \setminus T} \left(\sum_{c \in \mathcal{A}} b^p_{x_p,c} \right) \prod_{p \in T \setminus S} \left(\sum_{c \in \mathcal{A}} b^p_{y_p,c} \right) \prod_{p \notin S \cup T} \left(\sum_{c,c'} b^p_{c,c'} \right).$$

□

5 A kernel trick for inferring the posterior distribution of transformations of functions

For many applications of interest, α and L are such that writing down the α^L-dimensional function vector is impractical. Instead, much can be learned about the function by studying interpretable representations of it. Here we highlight three classes of representations and establish a general kernel trick that can be applied to efficiently compute the posterior distribution of arbitrary subsets of coefficients from these representations under a function space Gaussian process prior specified by a product kernel.

We will consider the following representations of real-valued functions over sequence space, each of which can be obtained by a linear transformation M of the function vector f.

- **Gauge-fixed weights** with respect to a λ-π gauge. The function value of a sequence is given by the sum of weights of its subsequences:

  $$f_x = \sum_{(S,s) \in \mathcal{S}} w_{(S,s)} \delta_{x[S]=s}.$$

  As discussed in 2.2.1, the choice of λ and π guides the interpretation of the weights $w_{(S,s)}$. Restricting the projection matrix (Equation (1)) to columns corresponding to the whole sequence, $(T,t) : |T| = L$, gives the linear transformation M. We can compute $w_{(S,s)} = M_{(S,s),*} f$ where $M_{(S,s),*}$ is a row vector whose formula is given in Table 1. Taking $\lambda = \infty$ and π uniform yields the zero-sum gauge, which is utilized in the "reference-free analysis" (RFA) approach described in [32]. Taking $\lambda = \infty$ and letting π be the point mass distribution on a wild-type sequence ($\pi^p_{WT_p} = 1$) yields the wild-type gauge. As discussed in [38], in the wild-type gauge $w_{(S,s)} = 0$ for all subsequences s that agree with the wild-type in at least one position; thus, we only define coefficients $w_{(S,u)}$ where u does not agree with the wild-type.

- **Background-averaged epistastic coefficients.** These coefficients, defined in [35] for binary alphabets and extended to arbitrary alphabet size in [14], are again defined with respect to a wild-type sequence. For each set of positions S and substring u on S that does not contain any wild-type characters, the corresponding background-averaged epistastic coefficient $\varepsilon_{(S,u)}$

represents the epistastic effect of the combination of mutations u on S, averaged over all backgrounds outside the subsequence. Following Equation 9 of [14], we have $\varepsilon_{(S,u)} = M_{(S,u),*}f$ where $M_{(S,u),*}$ is a row vector whose formula is given in Table 1. In the bi-allelic ($\alpha = 2$) case, the background-averaged epistastic coefficients are a rescaling of the Walsh-Hadamard coefficients (see Equation 6 of [53]).

- **Fourier coefficients.** In the Fourier basis [6], for each subset of positions S, there are $(\alpha-1)^{|S|}$ coefficients that together describe the $|S|$-way interaction at those positions. The Fourier basis vectors are the eigenvectors of the Hamming graph Laplacian and are natural way to express the GNK model [6]. The Fourier coefficients are also defined with respect to a wild-type sequence, and we index the $(\alpha-1)^{|S|}$ coefficients for the subset of positions S by substrings that do not contain a particular reference (i.e. wild-type) allele, which without loss of generality we denote as allele zero. Following Equation 10 of [6], we have each coefficient $\beta_{(S,u)} = M_{(S,u),*}f$ where $M_{(S,u),*}$ is a row vector whose formula is given in Table 1. In the bi-allelic ($\alpha = 2$) case, the Fourier coefficients are the Walsh-Hadamard coefficients (see Equation 6 of [53]).

For each of these transformations, naively computing each entry of Mf requires computing an α^L dimensional dot product. We leverage the factorizable form of the rows of M to establish a kernel trick that allows us to efficiently compute the posterior distribution of Mf.

Representation	**Transformation matrix**
λ-π gauge-fixed weights [38]	$M_{(S,s),x} = \prod_{p\in S}\left(\delta_{s_p=x_p} - \pi^p_{x_p}\eta\right)\prod_{p\notin S}\pi^p_{x_p}\eta$
Hierarchical ($\lambda = \infty$)	$M_{(S,s),x} = \prod_{p\in S}\left(\delta_{s_p=x_p} - \pi^p_{x_p}\right)\prod_{p\notin S}\pi^p_{x_p}$
Zero-sum ($\lambda = \infty, \pi$ uniform), same as RFA in [32]	$M_{(S,s),x} = \prod_{p\in S}\left(\delta_{s_p=x_p} - \frac{1}{\alpha}\right)\prod_{p\notin S}\frac{1}{\alpha}$
Wild-type ($\lambda = \infty, \pi$ WT)	$M_{(S,u),x} = \prod_{p\in S}\left(\delta_{x_p=u_p} - \delta_{x_p=WT_p}\right)\prod_{p\notin S}\delta_{x_p=WT_p}$
Background averaged epistastic coefficients [14]	$M_{(S,u),x} = \prod_{p\in S}\left(\delta_{x_p=u_p} - \delta_{x_p=WT_p}\right)\prod_{p\notin S}\frac{1}{\alpha}$
Bi-allelic ($\alpha = 2$), same as rescaled Walsh-Hadamard [53]	$M_{S,x} = \frac{1}{2}^{L-\lvert S\rvert}\prod_{p\in S}\left(\delta_{x_p\neq WT_p} - \delta_{x_p=WT_p}\right)$
Fourier coefficients [6]	$M_{(S,u),x} = \frac{1}{\sqrt{\alpha^L}}\prod_{p\in S}\left(\delta_{x_p=0} - \frac{1}{\sqrt{\alpha}-1}\delta_{x_p\neq 0} + \sqrt{\alpha}\delta_{x_p=u_p}\right)$
Bi-allelic ($\alpha = 2$), same as Walsh-Hadamard [53]	$M_{S,x} = \frac{1}{\sqrt{2^L}}\prod_{p\in S}\left(\delta_{x_p=0} - \delta_{x_p=1}\right)$

Table 1: The linear transformations M that map the function vector f to a specific representation. Throughout S is used to denote a subset of positions, s is used to denote any subsequence on those positions, and u is used to denote any subsequence on those positions that differs from a wild-type or reference allele at each position. We use u instead of s for the background averaged epistatic coefficients because these coefficients are only defined for subsequences that differ from a wild-type at all positions. We again use u instead of s for the wild-type gauge because all coefficients corresponding to sequences that agree with the wild-type are zero. For bi-allelic alphabets, we suppress the u since there is only one option for such a subsequence. For the λ-π gauges, $\eta = \lambda/(1+\lambda)$.

For f drawn from a Gaussian process posterior, the distribution of Mf is normal with mean and covariance given in Theorem 5a. While deriving analytic formulas for the mean and covariance of Mf is straightforward, computing the mean and covariance from these formulas requires computing entries of MK and MKM^T, which when done naively involves taking α^L dimensional dot products. We show that we can compute MK and MKM^T much more efficiently when the Gaussian process prior K is a product kernel and each row of M has a factorizable form (in the sense of Equation (3) in Theorem 5b below). In particular, Theorem 5b gives an expression for each entry of MK and MKM^T as the product of L factors, each of which is a sum of α or α^2 values. Given the expressions for MK and MKM^T, we can compute each element of the representation Mf with only matrix and vector operations with dimensions at most the number of training sequences.

Since our focus here is on gauge spaces, we dedicate the following section to describing the posterior distribution of gauge-fixed weights and applying Theorem 5 to compute an explicit formula for the posterior (Theorem 7). It is easy to verify that the matrices M given in Table 1 for the background averaged epistastic coefficients and the Fourier coefficents satisfy the row-factorizable condition (Equation (3)), and thus Theorem 5b can be applied to efficiently compute the posterior distributions of these representations as well.

Theorem 5. *Let f be drawn from the posterior of the a function space Gaussian process with covariance K and variance σ^2 (see Definition 4 or 1 of Section 3.2), and let M be a linear transformation of f.*

(a) The distribution of the random variable Mf is $N(\bar{\theta}, R)$ where $\bar{\theta} = MK_{*,X}Qy$, $R = MKM^T - (MK_{*,X})\left(Q + (Qy)(Qy)^T\right)(MK_{*,X})^T$, and $Q = \left(K_{X,X} + \sigma^2 I\right)^{-1}$.

(b) Let K be a product kernel

$$K_{x,y} = \prod_{p=1}^{L} a^p_{x_p,y_p}$$

where $a^p_{c,c'} = a^p_{c',c}$ for all pairs of characters c and c'. Suppose each row of M can be factorized by position,

$$M_{i,x} = \prod_{p=1}^{L} m^{i,p}_{x_p}. \tag{3}$$

Then

$$(MK)_{i,y} = \prod_{p=1}^{L} \left(\sum_{c \in \mathcal{A}} m^{i,p}_c a^p_{c,y_p} \right)$$

and

$$(MKM^T)_{i,j} = \prod_{p=1}^{L} \left(\sum_{c,c' \in \mathcal{A}} m^{i,p}_c m^{j,p}_{c'} a^p_{c,c'} \right).$$

Proof. (a) Recall the posterior distribution of the function space Gaussian process, $f \sim N(f^{MAP}, C_f)$, where

$$f^{MAP} = K_{*,X}Qy, \quad C_f = K_{*,*} - K_{*,X}QK_{X,*}, \quad \text{and } Q = \left(K_{X,X} + \sigma^2 I\right)^{-1}$$

Claim 3 implies $\bar{\theta} = MK_{*,X}Qy$ and

$$R = MC_fM^T - Mf^{MAP}\left(f^{MAP}\right)^T M^T = MKM^T - (MK_{*,X})\left(Q + (Qy)(Qy)^T\right)(MK_{*,X})^T. \tag{4}$$

(b) Let $\mathcal{S}$ denote the set of sequences. Note that

$$(MK)_{i,y} = \sum_{x \in \mathcal{S}} \left(\prod_{p=1}^{L} m^{i,p}_{x_p} a^{p}_{x_p,y_p} \right) = \prod_{p=1}^{L} \left(\sum_{c \in \mathcal{A}} m^{i,p}_{c} a^{p}_{c,y_p} \right),$$

as each term in the expansion of the rightmost expression corresponds to a sequence in $\mathcal{S}$. Next observe

$$\begin{aligned}(MKM^T)_{i,j} &= \sum_{y \in \mathcal{S}} \left(\prod_{p=1}^{L} \left(\sum_{c \in \mathcal{A}} m^{i,p}_{c} a^{p}_{c,y_p} \right) m^{j,p}_{y_p} \right) \\ &= \prod_{p=1}^{L} \sum_{c' \in \mathcal{A}} m^{j,p}_{c'} \left(\sum_{c \in \mathcal{A}} m^{i,p}_{c} a^{p}_{c,c'} \right) \\ &= \prod_{p=1}^{L} \left(\sum_{c,c' \in \mathcal{A}} m^{i,p}_{c} m^{j,p}_{c'} a^{p}_{c,c'} \right).\end{aligned}$$

□

6 Kernel trick for inferring the posterior distribution over weights in a λ-π gauge

The results in Section 4 describe how to perform regularized regression over weight space in a way that induces a chosen function space prior and yields an optimum in a particular gauge. For applications where α and ℓ are sufficiently large, a more practical approach is to infer a subset of the $(\alpha+1)^\ell$ gauge-fixed weights corresponding to a curated set of subsequences. In this section we demonstrate how to do so within our framework of function space Gaussian processes using a kernel trick described in Theorem 6. First we explain how to gauge-fix posterior distributions from weight space Bayesian regression and function space Gaussian processes (Definition 11), compute these gauge-fixed distributions, and describe when they are equal (Theorem 6). Then we show that in the case of λ-π gauges and product kernels it is incredibly efficient to compute the distribution of gauge-fixed weights for any fixed set of subsequences (Theorem 7).

For function space Gaussian processes, we can map draws from the posterior distribution f to a linear gauge space Θ by $\bar{P}f$, where $\bar{P}$ denotes the projection matrix into gauge Θ (see e.g. Equation (1)) restricted to columns corresponding to the whole sequence, $(T,t) : |T| = L$. Note $\bar{P}$ is $(\alpha+1)^\ell \times \alpha^\ell$ matrix and the coefficients of f in the Θ gauge are given by $\theta = \bar{P}f$. Note that the distribution of θ is fundamentally different than the posterior distribution of the *Bayesian regression in weight space* described in Section 3.2. The former is a distribution over the linear subspace Θ, whereas the latter has support over all of $\mathbb{R}^{(\alpha+1)^\ell}$ and only the MAP estimate w^{MAP} is guaranteed to be in Θ. We can likewise map draws from the posterior distribution of Bayesian regression in weight space to Θ by applying the associated projection matrix P.

Definition 11. *Posterior distributions of gauge-fixed weights:*

- *(Function space Gaussian processes). Let f be drawn from the posterior of the a function space Gaussian process (see Definition 4 or 1 of Section 3.2). The posterior distribution of gauge-fixed weights is the distribution of the random variable $\bar{P}f$.*

- *(Weight-space Bayesian regression.) Let w be drawn from the weight-space Bayesian regression posterior (See 3 of Section 3.2). The posterior distribution of gauge-fixed weights is the distribution of the random variable Pw.*

The posterior distributions of gauge-fixed weights are normal and Theorem 6a and b gives their means and covariances. The posterior covariance matrices are singular, ensuring that all draws lie in the gauge-fixed space. Theorem 6c establishes that when $K = \Phi W \Phi^T$, i.e. when the function space Gaussian process and the weight space Bayesian regression have corresponding MAP estimates $f^{MAP} = \Phi w^{MAP}$, then it is also the case that their posterior distributions over gauge-fixed weights are the same.

Theorem 6. *Posterior distributions of gauge-fixed weights can be written as follows:*

(a) The posterior distribution of gauge-fixed weights corresponding to a function space Gaussian process with covariance K and variance σ^2 is given by $N(\bar\theta, R)$ where

$$\bar\theta = \bar P K_{*,X} Q y, \quad R = \bar P K \bar P^T - (\bar P K_{*,X}) \left(Q + (Qy)(Qy)^T\right) (\bar P K_{*,X})^T, \text{ and } A = \left(K_{X,X} + \sigma^2 I\right)^{-1}.$$

(b) The posterior distribution of gauge-fixed weights corresponding to a Bayesian weight space prior with covariance W and variance σ^2 is given by $N(\bar\theta, R)$ where

$$\bar\theta = P\sigma^{-1} C_w \Phi_X^T y, \quad R = P C_w P^T - P\left(\sigma^{-1} C_w \Phi_X^T y\right)\left(\sigma^{-1} C_w \Phi_X^T y\right)^T P^T$$

$$\text{and} \quad C_w = \left(\sigma^2 \Phi_X^T \Phi_X + W^{-1}\right)^{-1}.$$

(c) If $K = \Phi W \Phi^T$, then the posterior distributions of gauge-fixed weights corresponding to the function space Gaussian process with covariance K and variance σ^2 and the Bayesian weight space prior with covariance W and variance σ^2 are identical.

Next, note that since the posterior distribution over the gauge-fixed weights θ is normal, it is straightforward to compute the distribution for any subset of the weights. To do so for function space Gaussian processes, one needs to compute $\bar P_{T,*} K$ where T denotes the rows of $\bar P$ corresponding to the subset of weights of interest. Naively, this requires dot product operations on vectors of length α^ℓ. However, when K is a product kernel and Θ is a λ-π gauge, we can efficiently compute the entries of $\bar P K$ and arrive at the following simple formula for the posterior distribution of gauge-fixed weights.

Theorem 7. *Let K be a product kernel*

$$K_{x,y} = \prod_{p \in P} a^p_{x_p, y_p}$$

where $a^p_{c,c'} = a^p_{c',c}$ for all pairs of characters c and c', and let Θ be a λ-π gauge. Then the posterior distribution over gauge-fixed weights is given by $\theta \sim N(\bar\theta, R)$ where

$$\bar\theta_{(S,s)} = \left(z^{(S,s)}\right)^T \left(K_{X,X} + \sigma^2 I\right)^{-1} y$$

and

$$R_{(S,s),(T,t)} = \prod_{p\in S\cap T}\left(\bar{\zeta}^p - \zeta^p_{s_p} - \zeta^p_{t_p} + a^p_{s_p,t_p}\right)\prod_{p\in S\setminus T}\left(\zeta^p_{s_p} - \bar{\zeta}^p\right)\prod_{p\in T\setminus S}\left(\zeta^p_{t_p} - \bar{\zeta}^p\right)\prod_{p\notin S\cup T}\bar{\zeta}^p$$
$$+\left(z^{(S,s)}\right)^T\left(\left(K_{X,X}+\sigma^2 I\right)^{-1} + \left(K_{X,X}+\sigma^2 I\right)^{-1} yy^T\left(K_{X,X}+\sigma^2 I\right)^{-1}\right)z^{(T,t)},$$

where

$$\zeta^p_c = \eta\sum_{c'}\pi^p_{c'}a^p_{c,c'}, \qquad \bar{\zeta}^p = \eta^2\sum_{c,c'}\pi^p_c\pi^p_{c'}a^p_{c,c'}$$

with $\eta = \lambda/(1+\lambda)$, *and* $z^{(S,s)}$ *is a* $|X|$ *dimensional vector given by*

$$z^{(S,s)}_x = \prod_{p\in S}\left(a^p_{x_p,s_p} - \zeta^p_{x_p}\right)\prod_{p\notin S}\zeta^p_{x_p}.$$

In Appendix A.4 we prove Theorems 6 and 7.

7 Function space priors induced by diagonal regularizers

A natural choice of a weight space regularizer is a diagonal matrix Λ. The optimizer of weight space regression with a diagonal regularizer equals the MAP estimate under a Gaussian prior for weight space in which the weights are assumed to be drawn independently. There is a large literature on random fitness landscapes constructed as the weighted sum of subsequence indicator features where the weights are drawn independently, see [22, 33, 7, 20]. One such model is the GNK (generalized NK model [7, 28, 20, 6]). In the GNK model, each position p defines a subset of positions N_p with $p\in N_p$ called a neighborhood. Weights for subsequences on the neighborhoods, (N_p, s), are drawn independently from normal distributions whose variance is the inverse of the size of N_p and zero weight is assigned to all other subsequences. Therefore, the GNK model is induced by a diagonal regularizer of the form $\Lambda_{(S,s),(S,s)} = |S|$ when S is a neighborhood $S = N_p$ and ∞ when S is not a neighborhood.

Here we compute the function space priors induced by diagonal regularizers in which the regularization strength is finite for all weights. In Section 7.1, we show that the class of $\Theta^{\lambda,\pi}$ diagonal regularizers (for $\lambda < \infty$) introduced in [38] induce product kernel priors on function space. Then in Section 7.2, we consider the class of diagonal regularizers where the regularization strength depends only on the order of interaction. We show that such regularizers induce VC priors on function space, but not all VC kernels can be induced by these diagonal regularizers.

7.1 Function space priors induced by diagonal $\Theta^{\lambda,\pi}$-regularizers

Diagonal weight space regularizers of the form

$$\Lambda_{(S,s),(S,s)} = \lambda^{|S|}\prod_{p\in S}\pi^p_{s_p}$$

are $\Theta^{\lambda,\pi}$-regularizers [38]. This form emphasizes the relationship between λ and the distribution of weights across different subsequence lengths. With the uniform distribution π, $\Lambda_{(S,s),(S,s)} = (\lambda/\alpha)^{|S|}$. When $\lambda = \alpha$ all weights incur equal penalization. Larger λ disproportionately penalizes the weights for higher-order subsequences, whereas smaller λ disproportionately penalizes the weights for lower-order subsequences. Here we compute the function space prior induced by such regularizers.

Theorem 8. *Let $\lambda > 0$ be finite and π a product distribution with full support. The diagonal regularizer $\Lambda_{(S,s),(S,s)} = \lambda^{|S|}\prod_{p\in S}\pi^p_{s_p}$ induces the prior*

$$K_{x,y} = \prod_{p:x_p=y_p}\left(1+\frac{1}{\pi^p_{x_p}\lambda}\right).$$

Proof. By Lemmas 3 and 4, it suffices to compute $\Phi\Lambda^{-1}\Phi^T$,

$$(\Phi\Lambda^{-1}\Phi^T)_{x,y} = \sum_{\substack{S,s:\\ x[S]=s,\, y[S]=s}} \Lambda^{-1}_{S,s,T,t} = \sum_{\substack{S,s:\\ x[S]=s,\, y[S]=s}} \frac{1}{\lambda^{|S|}\prod_{p\in S}\pi^p_{s_p}} = \prod_{p:x_p=y_p}\left(1+\frac{1}{\pi^p_{x_p}\lambda}\right).$$

□

Note that for non-uniform π, this is a heteroskedastic prior; the variances for sequences more likely under π are comparatively smaller. For uniform π, the induced kernels are scaled geometric decay kernels with $\beta = \left(1+\frac{\alpha}{\lambda}\right)^{-1}$ and scale factor $\left(1+\frac{\alpha}{\lambda}\right)^{\ell}$,

$$K_{x,y} = \left(1+\frac{\alpha}{\lambda}\right)^{\ell-d(x,y)}$$

When λ is smaller, we observe a sharper decay in correlation. This aligns with the observation that when λ is smaller the diagonal regularizer penalizes the weights for higher-order subsequences less strongly; when weights of higher-order subsequences are free to vary widely, the resultant function will have less correlation across similar sequences.

7.2 Function space priors induced by order-dependent diagonal regularizers

Next we consider diagonal regularizers with values depending only on the length of the subsequences.

Theorem 9. *Let Λ be a diagonal regularization matrix whose values depends only on the length of the subsequence,*

$$\Lambda_{(S,s),(S,s)} = a_{|S|},$$

where $a_j > 0$ for $j \in [\ell]$. Then Λ induces the VC prior

$$K_{x,y} = \sum_{k=0}^{\ell}\left(\sum_{j=k}^{\ell}\frac{1}{\alpha^j a_j}\binom{\ell-k}{j-k}\right)\mathcal{K}_k(d(x,y)).$$

While any sequence of positive λ_k's defines a valid VC kernel, we show in Appendix A.5 that not all VC kernels are induced by some order-dependent diagonal regularizer. The following corollary explains how to compute a order-dependent regularizer that induces a VC kernel, if one exists.

Corollary 2. *Let K be the kernel defined by dimension-normalized variance components $\lambda \geq 0$. Let T be an $(\ell+1)\times(\ell+1)$ zero-indexed upper triangular matrix with $T_{ij} = \binom{\ell-i}{j}$ for $i \leq j$. Let W be an $(\ell+1)\times(\ell+1)$ zero-indexed matrix with $W_{ij} = w_j(i)$. If $a = T^{-1}W\lambda > 0$ entry-wise, then the order-dependent diagonal regularizer $\Lambda_{(S,s),(S,s)} = 1/a_{|S|}$ induces the prior K.*

We now prove Theorem 9 using a combinatorial identity given and proven in Lemma 9.

Proof. (of Theorem 9). By Lemmas 3 and 4, it suffices to show that $\Phi\Lambda^{-1}\Phi^T = K$. We apply Lemma 9 and compute

$$(\Phi\Lambda^{-1}\Phi^T)_{x,y} = \sum_{\substack{S,s:\\ x[S]=s,\, y[S]=s}} \Lambda^{-1}_{(S,s),(S,s)} = \sum_{j=0}^{\ell} \sum_{\substack{S,s:\\ x[S]=s,\, y[S]=s\\ |S|=j}} \frac{1}{a_j} = \sum_{j=0}^{\ell-d} \binom{\ell - d(x,y)}{j} \frac{1}{a_j}$$

$$= \sum_{j=0}^{\ell-d} \left(\sum_{k=0}^{j} \binom{\ell-k}{j-k} \mathcal{K}_k(d(x,y)) \right) \frac{1}{\alpha^j a_j}$$

$$= \sum_{k=0}^{\ell} \left(\sum_{j=k}^{\ell} \frac{1}{\alpha^j a_j} \binom{\ell-k}{j-k} \right) \mathcal{K}_k(d(x,y)).$$

□

8 Discussion

The methods used to represent and infer functions over sequence space can have a substantive impact on interpretation and prediction accuracy [32, 11, 31]. Our framework can be used as a guide for computing the implicit function space prior induced by different choices of representation of the function (e.g. gauge-fixed weights [38] or a basis such as [46, 6, 14]) combined with a choice of an L_2 regularization matrix. Because each combination of representation type and regularizer imposes implicit assumptions, computing the induced function space prior can help practitioners quantify these assumptions, guide their choice of representation and regularizer, and inform their downstream interpretation. Our work further clarifies that although historically L_2 regularization on parameter space has been used to both fix the gauge and provide regularization for the estimated function [38], in fact the choice of gauge is a matter of how we choose to represent the learned function and is in principle independent of the form of regularization or Gaussian process prior we impose on function space.

Our main results linking the regularization matrix to the induced function space prior are tailored to our notion of weight space defined by an overcomplete basis of indicator functions. To illustrate the importance of the form of the representation, in Appendix A.6 we demonstrate that the same regularization matrices behave differently when different bases define the weight space. In particular, we apply our framework to two alternate bases for the bi-allelic case ($\alpha = 2$) and show that whereas diagonal regularizers applied when using the Walsh-Hadamard basis induce homoskedastic function space priors, diagonal regularizers applied when using the wild-type basis induce heteroskedastic function space priors. Further investigation is necessary to characterize the assumptions imposed by forms of regularization other than L_2, e.g. L_1 regularization as is also commonly employed in practice [36, 6, 32].

Our work also provides a theoretically-grounded method for estimating representations of real-valued functions over sequence space and quantifying uncertainty for these estimates. Given a function space prior and a set of training points, our kernel trick can be applied to efficiently compute the posterior distribution for a general class of linear transformations of functions over sequence space. This class includes gauge-fixed weights [32, 38], coefficients of the the function when written in the Fourier basis [6], and background-averaged epistastic coefficients [14], providing efficient estimates and uncertainty bounds for these quantities. Importantly, our kernel trick allows the computation of these estimates without requiring us to explicitly reconstruct the full α^ℓ-dimensional function, which opens the possibility of computing these quantities on-the-fly in order to quantify higher-order genetic interactions in much longer sequences than have been investigated to date. Moreover, for the case of gauge-fixed weights that is the focus of our contribution here, these individually computed weights maintain the interpretability properties of the chosen gauge (e.g. the relationship to averages over regions of sequence space [38]) and result in a posterior whose support is limited to the α^ℓ-dimensional gauge space, without ever calculating the full $(\alpha + 1)^\ell$ vector of weights.

Acknowledgements

This work was supported by a Burroughs-Wellcome Careers at the Scientific Interface (CASI) award (SP), NIH grant R35 GM133613 (CMG, DMM), NIH grant R35 GM133777 (JBK), NIH grant R01 HG011787 (JBK, DMM), and NIH grant R35 GM154908 (JZ), as well as additional funding from the Simons Center for Quantitative Biology at CSHL (DMM, JBK) and the College of Liberal Arts and Sciences at the University of Florida (JZ).

References

[1] Alan Nawzad Amin, Eli Nathan Weinstein, and Debora Susan Marks. Biological sequence kernels with guaranteed flexibility, April 2023. arXiv:2304.03775 [cs, q-bio, stat].

[2] Alex N Nguyen Ba, Katherine R Lawrence, Artur Rego-Costa, Shreyas Gopalakrishnan, Daniel Temko, Franziska Michor, and Michael M Desai. Barcoded bulk QTL mapping reveals highly polygenic and epistatic architecture of complex traits in yeast. *Elife*, 11:e73983, 2022.

[3] Claudia Bank. Epistasis and adaptation on fitness landscapes. *Annual Review of Ecology, Evolution, and Systematics*, 53(1):457–479, 2022.

[4] Niko Beerenwinkel, Lior Pachter, and Bernd Sturmfels. Epistasis and shapes of fitness landscapes. *Statistica Sinica*, pages 1317–1342, 2007.

[5] Niko Beerenwinkel, Lior Pachter, Bernd Sturmfels, Santiago F Elena, and Richard E Lenski. Analysis of epistatic interactions and fitness landscapes using a new geometric approach. *BMC Evolutionary Biology*, 7:1–12, 2007.

[6] David H Brookes, Amirali Aghazadeh, and Jennifer Listgarten. On the sparsity of fitness functions and implications for learning. *Proceedings of the National Academy of Sciences*, 119(1):e2109649118, 2022.

[7] Jeffrey Buzas and Jeffrey Dinitz. An analysis of NK landscapes: Interaction structure, statistical properties, and expected number of local optima. *IEEE Transactions on Evolutionary Computation*, 18(6):807–818, 2013.

[8] Kristina Crona and Devin Greene. Walsh coefficients and circuits for several alleles. *arXiv preprint arXiv:2401.00743*, 2024.

[9] Juan Diaz-Colunga, Abigail Skwara, Jean CC Vila, Djordje Bajic, and Alvaro Sanchez. Global epistasis and the emergence of function in microbial consortia. *Cell*, 187(12):3108–3119, 2024.

[10] Júlia Domingo, Pablo Baeza-Centurion, and Ben Lehner. The causes and consequences of genetic interactions (epistasis). *Annual Review of Genomics and Human Genetics*, 20:433–460, 2019.

[11] Thomas Dupic, Angela M Phillips, and Michael M Desai. Protein sequence landscapes are not so simple: on reference-free versus reference-based inference. *bioRxiv*, 2024.

[12] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 87(1):012707, 2013.

[13] Andre J Faure and Ben Lehner. Mochi: neural networks to fit interpretable models and quantify energies, energetic couplings, epistasis, and allostery from deep mutational scanning data. *Genome Biology*, 25(1):303, 2024.

[14] Andre J Faure, Ben Lehner, Verónica Miró Pina, Claudia Serrano Colome, and Donate Weghorn. An extension of the walsh-hadamard transform to calculate and model epistasis in genetic landscapes of arbitrary shape and complexity. *PLoS Computational Biology*, 20(5):e1012132, 2024.

[15] Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8):801–807, 2014.

[16] Chase R Freschlin, Sarah A Fahlberg, and Philip A Romero. Machine learning to navigate fitness landscapes for protein engineering. *Current Opinion in Biotechnology*, 75:102713, 2022.

[17] Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix Gaussian process inference with gpu acceleration. *Advances in Neural Information Processing Systems*, 31, 2018.

[18] Robert Happel and Peter F Stadler. Canonical approximation of fitness landscapes. *Complexity*, 2(1):53–58, 1996.

[19] Trevor Hinkley, João Martins, Colombe Chappey, Mojgan Haddad, Eric Stawiski, Jeannette M Whitcomb, Christos J Petropoulos, and Sebastian Bonhoeffer. A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nature Genetics*, 43(5):487–489, 2011.

[20] Sungmin Hwang, Benjamin Schmiegelt, Luca Ferretti, and Joachim Krug. Universality classes of interaction structures for nk fitness landscapes. *Journal of Statistical Physics*, 172:226–278, 2018.

[21] Milo S Johnson, Gautam Reddy, and Michael M Desai. Epistasis and evolution: recent advances and an outlook for prediction. *BMC Biology*, 21(1):120, 2023.

[22] Stuart A Kauffman and Edward D Weinberger. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, 141(2):211–245, 1989.

[23] Justin B Kinney and David M McCandlish. Massively parallel assays and quantitative sequence–function relationships. *Annual Review of Genomics and Human Genetics*, 20:99–127, 2019.

[24] Susanna Manrubia, José A Cuesta, Jacobo Aguirre, Sebastian E Ahnert, Lee Altenberg, Alejandro V Cano, Pablo Catalán, Ramon Diaz-Uriarte, Santiago F Elena, Juan Antonio García-Martín, et al. From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics. *Physics of Life Reviews*, 38:55–106, 2021.

[25] Carlos Martí-Gómez, Juannan Zhou, Wei-Chia Chen, Justin B Kinney, and David M McCandlish. Inference and visualization of complex genotype-phenotype maps with gpmap-tools. *bioRxiv*, pages 2025–03, 2025.

[26] Johannes Neidhart, Ivan G Szendro, and Joachim Krug. Exact results for amplitude spectra of fitness landscapes. *Journal of Theoretical Biology*, 332:218–227, 2013.

[27] Pascal Notin, Nathan Rollins, Yarin Gal, Chris Sander, and Debora Marks. Machine learning for functional protein design. *Nature Biotechnology*, 42(2):216–228, 2024.

[28] Stefan Nowak and Joachim Krug. Analysis of adaptive walks on nk fitness landscapes with different interaction schemes. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(6):P06014, 2015.

[29] John G Orme and Terri Combs-Orme. *Multiple Regression with Discrete Dependent Variables*. Oxford University Press, 2009.

[30] Jakub Otwinowski, David M McCandlish, and Joshua B Plotkin. Inferring the shape of global epistasis. *Proceedings of the National Academy of Sciences*, 115(32):E7550–E7558, 2018.

[31] Yeonwoo Park, Brian PH Metzger, and Joseph W Thornton. On the analysis of protein genetic architecture: Response to "Protein sequence landscapes are not so simple". *bioRxiv*, pages 2024–09, 2024.

[32] Yeonwoo Park, Brian PH Metzger, and Joseph W Thornton. The simplicity of protein sequence-function relationships. *Nature Communications*, 15(1):7953, 2024.

[33] Alan S Perelson and Catherine A Macken. Protein evolution on partially correlated landscapes. *Proceedings of the National Academy of Sciences*, 92(21):9657–9661, 1995.

[34] Patrick C Phillips. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855, 2008.

[35] Frank J Poelwijk, Vinod Krishna, and Rama Ranganathan. The context-dependence of mutations: a linkage of formalisms. *PLoS Computational Biology*, 12(6):e1004771, 2016.

[36] Frank J Poelwijk, Michael Socolich, and Rama Ranganathan. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature Communications*, 10(1):4213, 2019.

[37] Anna Posfai, David M McCandlish, and Justin B Kinney. Symmetry, gauge freedoms, and the interpretability of sequence-function relationships. *Physical Review Research*, 7(2):023005, 2025.

[38] Anna Posfai, Juannan Zhou, David M McCandlish, and Justin B Kinney. Gauge fixing for sequence-function relationships. *PLoS Computational Biology*, 21(3):e1012818, 2025.

[39] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian Processes for Machine Learning*. MIT press Cambridge, 2006.

[40] Gautam Reddy and Michael M Desai. Global epistasis emerges from a generic model of a complex trait. *Elife*, 10:e64740, 2021.

[41] Philip A. Romero, Andreas Krause, and Frances H. Arnold. Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences*, 110(3):E193–E201, January 2013. Publisher: Proceedings of the National Academy of Sciences.

[42] Timothy B Sackton and Daniel L Hartl. Genotypic context and epistasis in individuals and populations. *Cell*, 166(2):279–287, 2016.

[43] Zachary R Sailer and Michael J Harms. Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics*, 205(3):1079–1088, 2017.

[44] Thomas J Santner and Diane E Duffy. *The Statistical Analysis of Discrete Data*. Springer Science & Business Media, 2012.

[45] Gabriele Schweikert, Gunnar Rätsch, Christian Widmer, and Bernhard Schölkopf. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

[46] Peter F Stadler, Rudi Seitz, and Günter P Wagner. Population dependent Fourier decomposition of fitness landscapes over recombination spaces: Evolvability of complex characters. *Bulletin of Mathematical Biology*, 62:399–428, 2000.

[47] Tyler N Starr and Joseph W Thornton. Epistasis in protein evolution. *Protein Science*, 25(7):1204–1218, 2016.

[48] Richard R Stein, Debora S Marks, and Chris Sander. Inferring pairwise interactions from biological data using maximum-entropy probability models. *PLoS Computational Biology*, 11(7):e1004182, 2015.

[49] Ammar Tareen, Mahdi Kooshkbaghi, Anna Posfai, William T Ireland, David M McCandlish, and Justin B Kinney. MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. *Genome Biology*, 23(1):98, 2022.

[50] Nora C. Toussaint, Christian Widmer, Oliver Kohlbacher, and Gunnar Rätsch. Exploiting physico-chemical properties in string kernels. *BMC Bioinformatics*, 11(8):S7, October 2010.

[51] Ke Wang, Geoff Pleiss, Jacob Gardner, Stephen Tyree, Kilian Q Weinberger, and Andrew Gordon Wilson. Exact Gaussian processes on a million data points. *Advances in Neural Information Processing Systems*, 32, 2019.

[52] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[53] Edward D Weinberger. Fourier and Taylor series on fitness landscapes. *Biological Cybernetics*, 65(5):321–330, 1991.

[54] Daniel M. Weinreich, Yinghong Lan, Jacob Jaffe, and Robert B. Heckendorn. The influence of higher-order epistasis on biological fitness landscape topography. *Journal of Statistical Physics*, 172(1):208–225, 2018.

[55] Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, 16(8):687–694, 2019.

[56] Kevin K Yang, Zachary Wu, Claire N Bedbrook, and Frances H Arnold. Learned protein embeddings for machine learning. *Bioinformatics*, 34(15):2642–2648, August 2018.

[57] Juannan Zhou, Carlos Martí-Gómez, Justin Kinney, Samantha Petti, and David M McCandlish. Manuscript in preparation. 2025.

[58] Juannan Zhou and David M McCandlish. Minimum epistasis interpolation for sequence-function relationships. *Nature Communications*, 11(1):1782, 2020.

[59] Juannan Zhou, Mandy S Wong, Wei-Chia Chen, Adrian R Krainer, Justin B Kinney, and David M McCandlish. Higher-order epistasis and phenotypic prediction. *Proceedings of the National Academy of Sciences*, 119(39):e2204233119, 2022.

A Appendix

A.1 Useful lemmas

Claim 2. *(Woodbury, Sherman, and Morrison matrix inversion lemma) Assuming all relevant inverses exists*

$$\left(Z + UWV^T\right)^{-1} = Z^{-1} - Z^{-1}U\left(W^{-1} + V^T Z^{-1}U\right)^{-1} V^T Z^{-1}.$$

The following is a straightforward computation, included here for completeness.

Claim 3. *Let $x \sim N(\mu, C)$. Then $Px \sim N(P\mu, PCP^T - P\mu\mu^T P^T)$.*

Proof. Since Px is a linear transformation of a Gaussian random variable, it is also a Gaussian random variable. Linearity implies that the mean of Px is equal to $P\mu$. Observe

$$\begin{aligned} Cov(Px) &= \mathsf{E}\big((Px - P\mu)(Px - P\mu)^T\big) \\ &= \mathsf{E}\big(Pxx^T P^T\big) - \mathsf{E}\big(P\mu x^T P^T\big) - \mathsf{E}\big(Px\mu^T P^T\big) + \mathsf{E}\big(P\mu\mu^T P^T\big) \\ &= PCP^T - P\mu\mu^T P^T. \end{aligned}$$

□

Lemma 7. *(Claim 25 of [38].) Let V_1 and V_2 be complementary subspaces of a vector space V. Let P_1 be the projection into V_1 along V_2, and P_2 be the projection into V_2 along V_1. Let Λ be a symmetric positive definite matrix acting on V. Then the following are equivalent:*

1. *V_1 and V_2 are Λ-orthogonal, i.e. $v_1^T \Lambda v_2 = 0$ for all $v_1 \in V_1$ and $v_2 \in V_2$.*
2. *For any fixed $v_1 \in V_1$, $\arg\min_{v_2 \in V_2}(v_1 + v_2)^T \Lambda (v_1 + v_2) = 0$.*
3. *$\Lambda = P_1^T \Lambda P_1 + P_2^T \Lambda P_2$.*

Next we prove Lemma 2, which gives another equivalent condition for Λ orthogonality.

Proof. (of Lemma 2). Assume Λ orthogonalizes V_1 and V_2. Since Λ is positive-definite, we can write $\Lambda = Z^T Z$ where Z is invertible. Moreover, since Λ is orthogonalizing Lemma 7 implies that

$$\Lambda = P_1^T \Lambda P_1 + P_2^T \Lambda P_2 = (ZP_1)^T ZP_1 + (ZP_2)^T ZP_2,$$

where P_1 and P_2 are the projection matrices into V_1 and V_2 along V_2 and V_1 respectively. It remains to show that the null space of ZP_1 is V_2 and the null space of ZP_2 is V_1. Indeed, note that since Z is invertible, if $ZP_1 x = 0$, then $P_1 x = 0$, meaning x is in the null space of P_1, which is equal to V_2. The argument that V_1 is the null space of ZP_2 is analogous.

Next assume that $\Lambda = A^T A + B^T B$ where nullspace$(A) = V_2$ and nullspace$(B) = V_1$. First we show that Λ is positive-definite. Let $v = v_1 + v_2$ where $v_1 \in V_1$ and $v_2 \in V_2$. Note

$$v^T \Lambda v = v_1^T A^T A v_1 + v_2^T B^T B v_2 = \|Av_1\|_2^2 + \|Bv_2\|_2^2$$

is zero if and only if $v_1 = 0$ and $v_2 = 0$. Next we apply Condition 1 of Lemma 7 to establish that Λ orthogonalizes V_1 and V_2. Let $v_1 \in V_1$ and $v_2 \in V_2$. Observe

$$v_1^T (A^T A + B^T B) v_2 = v_1^T A^T (Av_2) + (v_1^T B^T) B v_2 = 0.$$

□

Lemma 8. *If K has the form $K_{x,y} = \sum_{k=0}^{\ell} \lambda_k \mathcal{K}_k(d(x,y))$, then*

$$K^{-1}_{x,y} = \sum_{k=0}^{\ell} \lambda_k^{-1} \mathcal{K}_k(d(x,y)).$$

Proof. This results simply follows from the fact that $K_{x,y} = \sum_{k=0}^{\ell} \lambda_k \mathcal{K}_k(d(x,y))$ provides an eigendecomposition of the matrix K, see [59]. □

The following identity is a generalization of equation (22) in [26].

Lemma 9. *Let $j \in [\ell - d]$. Then*

$$\alpha^{-j} \sum_{k=0}^{j} \binom{\ell-k}{j-k} \mathcal{K}_k(d) = \binom{\ell-d}{j}.$$

Proof. We will show the equivalent statement

$$\sum_{0 \le i \le k \le j} \binom{\ell-k}{j-k} \binom{d}{i} \binom{\ell-d}{k-i} \alpha^j (\alpha-1)^{k-i} (-1)^i = \binom{\ell-d}{j} \alpha^{2j} \tag{5}$$

by showing that each side of the Equation (5) is equal to the coefficient of z^j in the polynomial $(\alpha^2 z + 1)^{\ell-d}$. To see the righthand side, imagine expanding $(\alpha^2 z + 1)^{\ell-d}$ into $2^{\ell-d}$ terms. Each degree j term has coefficient α^{2j} and there are $\binom{\ell-d}{j}$ such terms. To see the lefthand side, note that

$$(\alpha^2 z + 1)^{\ell-d} = (\alpha z + \alpha(\alpha-1)z + 1)^{\ell-d} (\alpha z - \alpha z + 1)^d.$$

Imagine expanding this polynomial into 3^ℓ terms. We consider the possible coefficients for a degree j term in this expansion and compute how many such terms yield this coefficient. Imagine a degree j term that is the product of $j-k$ copies of αz, $k-i$ copies of $\alpha(\alpha-1)z$ and i copies of $(-\alpha z)$. There are

$$\binom{\ell-k}{j-k} \binom{\ell-d}{k-i} \binom{d}{i}$$

such terms, each with a coefficient of

$$(\alpha)^{j-k} (\alpha(\alpha-1))^{k-i} (-\alpha)^i = \alpha^j (\alpha-1)^{k-i} (-1)^i.$$

Thus, the lefthand side of Equation (5) also counts the coefficient of the degree j term in the polynomial $(\alpha^2 z + 1)^{\ell-d}$. □

A.2 Computations that help establish the marginalization property

Claim 4. *Let b and $\mathcal{T}_F$ be defined with respect to a fixed subsequence (S, s) as described in the proof of Lemma 6. Let w satisfy the λ-π marginalization property. Then for $p \in S^c \cap F^c$,*

$$\sum_{(T,t) \in \mathcal{T}_F} b(T,t) w_{(T,t)} = \sum_{(T,t) \in \mathcal{T}_{F \cup \{p\}}} \frac{b(T,t) w_{(T,t)}}{\eta}.$$

Proof. Note that

$$b(U \cup \{p\}, u^{+c}) = \pi_c^p b(U,u).$$

We compute

$$\begin{aligned}
\sum_{(T,t)\in\mathcal{T}_F} b(T,t)w_{(T,t)} &= \sum_{(U,u)\in\mathcal{T}_F : p\notin U} \left(b(U,u)w_{(U,u)} + \sum_{c\in\mathcal{A}} b(U\cup\{p\}, u^{+c})w_{(U\cup\{p\},u^{+c})} \right) \\
&= \sum_{(U,u)\in\mathcal{T}_F : p\notin U} \left(b(U,u)w_{(U,u)} + \sum_{c\in\mathcal{A}} \pi_c^p b(U,u)w_{(U\cup\{p\},u^{+c})} \right) \\
&= \sum_{(U,u)\in\mathcal{T}_F : p\notin U} \left(b(U,u)w_{(U,u)} + \left(\frac{1-\eta}{\eta}\right) b(U,u)w_{(U,u)} \right) \\
&= \sum_{(T,t)\in\mathcal{T}_{F\cup\{p\}}} \frac{b(T,t)w_{(T,t)}}{\eta},
\end{aligned}$$

where we used the assumption that w satisfies Equation (2) to establish the third equality. □

Claim 5. *Let b and $\mathcal{T}_F$ be defined with respect to a fixed subsequence (S,s) as described in the proof of Lemma 6. Let w satisfy the λ-π marginalization property. Then for $p \in S \cap F^c$,*

$$\sum_{(T,t)\in\mathcal{T}_F} b(T,t)w_{(T,t)} = \sum_{(T,t)\in\mathcal{T}_{F\cup\{p\}}} \frac{b(T,t)w_{(T,t)}}{1-\pi_{s_p}^p\eta}.$$

Proof. Note that

$$b(U\cup\{p\}, u^{+s_p}) = \frac{\left(1-\pi_{s_p}^p\right)\eta b(U,u)}{1-\eta} = \frac{-\pi_{s_p}^p \eta b(U,u)}{1-\eta} + \frac{b(U\cup\{p\},u^{+s_p})}{1-\pi_{s_p}^p\eta},$$

and for $c \neq s_p$

$$b(U\cup\{p\}, u^{+c}) = \frac{-\pi_c^p\eta b(U,u)}{1-\eta}.$$

We compute

$$\begin{aligned}
\sum_{(T,t)\in\mathcal{T}_F} b(T,t)w_{(T,t)} &= \sum_{(U,u)\in\mathcal{T}_F : p\notin U} \left(b(U,u)w_{(U,u)} + \sum_{c\in\mathcal{A}} b(U\cup\{p\}, u^{+c})w_{(U\cup\{p\},u^{+c})} \right) \\
&= \sum_{(U,u)\in\mathcal{T}_F : p\notin U} \left(b(U,u)w_{(U,u)} + \frac{b(U\cup\{p\},u^{+s_p})}{1-\pi_{s_p}^p\eta} w_{(U\cup\{p\},u^{+s_p})} + \sum_{c\in\mathcal{A}} \frac{-\pi_c^p\eta b(U,u)}{1-\eta} w_{(U\cup\{p\},u^{+c})} \right) \\
&= \sum_{(U,u)\in\mathcal{T}_F : p\notin U} \frac{b(U\cup\{p\},u^{+s_p})}{1-\pi_{s_p}^p\eta} w_{(U\cup\{p\},u^{+s_p})} \\
&= \sum_{(T,t)\in\mathcal{T}_{F\cup\{p\}}} \frac{b(T,t)w_{(T,t)}}{1-\pi_{s_p}^p\eta},
\end{aligned}$$

where we used the assumption that w satisfies Equation (2) to establish the third equality. □

A.3 Building regularizers for geometric decay, Connectedness, and Jenga kernels

In this section, we discuss three subclasses of product kernels and compute $\Phi^T K^{-1}\Phi$ for each class.

A.3.1 Definitions of geometric decay, Connectedness, and Jenga kernels

We will consider the following three subclasses of product kernels. The most simple subclass are geometric decay kernels, where the covariance decays exponentially with Hamming distance. These isotropic kernels are also a special case of VC kernels corresponding to hyperparameters λ_k that decay exponentially with the order k [59, 26].

Definition 12. *A geometric decay kernel has the following form:*

$$K_{x,y} = \beta^{d(x,y)}$$

where $\beta \in (0,1)$*.*

Next we consider Connectedness kernels, which are a generalization of the geometric decay kernel that can express that making changes at different positions results in different effects on predictability. The covariation between a pair is the product of site specific factors z^p for all positions p where they differ. In the bi-allelic case ($\alpha = 2$) when $z^p \in (0,1)$ this prior on function space is equivalent to the Connectedness model proposed by [40].

Definition 13. *A Connectedness kernel has the form*

$$K_{x,y} = \prod_{p:x_p\neq y_p} z^p$$

where each factor satisfies $\frac{-1}{\alpha-1} < z^p < 1$*.*

The constraints ensure that K is positive-definite, see [57].

Finally, we consider Jenga kernels, a generalization of Connectedness kernels that allows different allele-position (character-position) combinations to affect predictability differently. We assign each character-position a factor z^p_c; the covariance between a pair of sequences is the product of the factors over the positions where the sequences differ.

Definition 14. *A Jenga kernel has the form*

$$K_{x,y} = \prod_{p:x_p\neq y_p} s_p z^p_{x_p} z^p_{y_p}$$

where at each position p *either*

1. $s_p = 1$ *and* $z^p_c \in (0,1)$ *for each* $c \in \mathcal{A}$*, or*

2. $s_p = -1$ *and* $\sum_{c\in\mathcal{A}} \frac{\left(z^p_c\right)^2}{1+\left(z^p_c\right)^2} \leq 1$*.*

The two conditions above correspond to $z^p > 0$ and $z^p \leq 0$ in the connectedness kernel, respectively, and the constraints ensure that K is positive-definite, see [57].

A.3.2 Theorem 4 applied to geometric decay, Connectedness, and Jenga kernels

The following corollaries of Theorem 4 give the form of $\Phi^T K^{-1}\Phi$ for geometric decay, Connectedness, and Jenga kernels.

Corollary 3. *Let K be a Jenga kernel, $K_{x,y} = \prod_{p:x_p\neq y_p} s^p z^p_{x_p} z^p_{y_p}$, then $(\Phi^T K^{-1}\Phi)_{(S,s),(T,t)}$ is as given in Theorem 4 with*

$$b^p_{c,c'} = \frac{\delta_{c=c'}}{1 - s^p\left(z^p_c\right)^2} + \frac{s^p\gamma^p z^p_c z^p_{c'}}{\left(1 - s^p\left(z^p_c\right)^2\right)\left(1 - s^p\left(z^p_{c'}\right)^2\right)}, \qquad \gamma^p = \frac{-1}{1 + s^p\sum_{c\in\mathcal{A}}\frac{\left(z^p_c\right)^2}{1-s^p\left(z^p_c\right)^2}}$$

Corollary 4. *Let K be a connectedness kernel $K_{x,y} = \prod_{p:x_p\neq y_p} z^p$. Then*

$$(\Phi^T K^{-1}\Phi)_{(S,s),(T,t)} = \prod_{p\in[\ell]}\frac{1}{1+(\alpha-1)z^p}\prod_{\substack{p\in S\cap T\\ x_p=y_p}}\frac{1+(\alpha-2)z^p}{1-z^p}\prod_{\substack{p\in S\cap T\\ x_p\neq y_p}}\frac{-z^p}{1-z^p}\prod_{p\notin S\cup T}\alpha.$$

Corollary 5. *Suppose K is a geometric decay kernel, $K_{x,y} = \beta^{d(x,y)}$ with $\beta\in(0,1)$. Then*

$$(\Phi^T K^{-1}\Phi)_{(S,s),(T,t)} = \frac{\alpha^{L-|S\cup T|}}{(1+(\alpha-1)\beta)^{\ell}}\left(\frac{1+(\alpha-2)\beta}{1-\beta}\right)^{|\{p\in S\cap T: s_p=t_p\}|}\left(\frac{-\beta}{1-\beta}\right)^{|\{p\in S\cap T: s_p\neq t_p\}|}.$$

Corollary 3 follows directly from Theorem 4 and Lemma 10, which gives the form of $(K^p)^{-1}$ for a Jenga kernel.

Lemma 10. *Let K be a matrix of the form*

$$K_{ij} = \begin{cases}1 & i=j\\ sa_ia_j & i\neq j\end{cases}$$

where $s\in\{-1,1\}$. Then

$$K^{-1}_{ij} = \frac{\delta_{i=j}}{(1-sa_i^2)} + \frac{s\zeta a_i a_j}{(1-sa_i^2)(1-sa_j^2)} \quad \textit{for} \quad \zeta = \frac{-1}{1+s\sum_i\frac{a_i^2}{(1-sa_i^2)}}.$$

Proof. The result follows directly from applying Sherman-Morrison formula (Lemma 11) with A the diagonal matrix with $A_{ii} = 1 - sa_i^2$ and $u = sv$ with $u_i = a_i$. □

Lemma 11. *(Sherman-Morrison formula) Let A be an invertible matrix. Then*

$$(A+uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}.$$

To arrive at the expression in Corollary 4, we begin with Corollary 3 and plug in $s^p = sgn(z^p)$ and $z^p_c = \sqrt{|z^p|}$ for all c. We obtain

$$b^p_{x_p,y_p} = \begin{cases}\frac{1+(\alpha-2)z^p}{(1-z^p)(1+(\alpha-1)z^p)} & x_p = y_p\\ \frac{-z^p}{(1-z^p)(1+(\alpha-1)z^p)} & x_p\neq y_p\end{cases}$$

$$\sum_{c\in\mathcal{A}} b^p_{c,c'} = \frac{1+(\alpha-2)z^p-(\alpha-1)z^p}{(1-z^p)(1+(\alpha-1)z^p)} = \frac{1}{1+(\alpha-1)z^p}$$

$$\sum_{c,c'} b^p_{c,c'} = \frac{\alpha(1+(\alpha-2)z^p)-\alpha(\alpha-1)z^p}{(1-z^p)(1+(\alpha-1)z^p)} = \frac{\alpha}{1+(\alpha-1)z^p}$$

To arrive at the expression in Corollary 5, we plug $z^p = \beta$ for all p in to Corollary 4.

A.4 Proofs of Theorem 6 and Theorem 7

We now prove Theorem 6.

Proof. (Theorem 6) (a) The results follows by taking $M = \bar{P}$ and applying Theorem 5. (b) Recall the posterior distribution under the weight space Gaussian prior, $w \sim N(w^{MAP}, C_w)$, where

$$w^{MAP} = \sigma^{-1}\left(\sigma^2\Phi_X^T\Phi_X + W^{-1}\right)^{-1}\Phi_X^T y \quad \text{and} \quad C_w = \left(\sigma^2\Phi_X^T\Phi_X + W^{-1}\right)^{-1}.$$

Claim 3 implies $\bar{\theta} = P\sigma^{-1}C_w\Phi_X^T y$ and

$$R = PC_wP^T - Pw^{MAP}\left(w^{MAP}\right)^T P^T = PC_wP^T - P\left(\sigma^{-1}C_w\Phi_X^T y\right)\left(\sigma^{-1}C_w\Phi_X^T y\right)^T P^T \tag{6}$$

(c) By Claim 3, it suffices to show that if $K = \Phi W \Phi^T$, then $\bar{P}f^{MAP} = Pw^{MAP}$ and $\bar{P}C_f\bar{P}^T = PC_wP^T$. Note $P = \bar{P}\Phi$ and $f^{MAP} = \Phi w^{MAP}$ (Lemma 4). It follows that $\bar{P}f^{MAP} = \bar{P}\Phi w^{MAP} = Pw^{MAP}$.

Next we apply the matrix inversion formula (Claim 2) with $W = \sigma^2 I$, $U = \Phi_X^T$, $V^T = \Phi_X$ and $Z = W^{-1}$ to show that $\bar{P}C_f\bar{P}^T = PC_wP^T$. Observe

$$C_w = \left(\sigma^2\Phi_X^T\Phi_X + W^{-1}\right)^{-1} = W - W\Phi_X^T\left(\sigma^{-2}I + \Phi_X W\Phi_X^T\right)^{-1}\Phi_X W = W - W\Phi_X^T Q\Phi_X W.$$

Also note that

$$PWP^T = \bar{P}\Phi W\Phi^T\bar{P}^T = \bar{P}K\bar{P}^T, \quad PW\Phi_X^T = \bar{P}\Phi W\Phi_X^T = \bar{P}K_{*,X}, \quad \text{and} \quad \Phi_X WP^T = K_{*,X}\bar{P}^T.$$

We obtain

$$PC_wP^T = PWP^T - PW\Phi_X^T Q\Phi_X WP^T = \bar{P}K\bar{P}^T - \bar{P}K_{*,X}QK_{*,X}\bar{P}^T = \bar{P}C_f\bar{P}^T,$$

as desired. □

Proof. (of Theorem 7.) Recall

$$P^{\lambda,\pi}_{(S,s),([\ell],t)} = \prod_{p\in S}\left(\delta_{s_p=t_p} - \pi^p_{t_p}\eta\right)\prod_{p\notin S}\pi^p_{t_p}\eta.$$

We apply Theorem 5b with $M = \bar{P}$ and

$$m_c^{(S,s),p} = \begin{cases} \delta_{s_p=c} - \pi^p_c\eta & p \in S \\ \pi^p_c\eta & p \notin S. \end{cases}$$

The result follows directly from the following computations:

$$\begin{aligned}
\left(\bar{P}K\right)_{(S,s),y} &= \prod_{p\in S}\left(\left(1-\pi^p_{s_p}\eta\right)a^p_{y_p,s_p} - \sum_{c\neq s_p}\pi^p_c\eta a^p_{y_p,c}\right)\prod_{p\notin S}\left(\sum_{c\in\mathcal{A}}\pi^p_c\eta a^p_{y_p,c}\right)\\
&= \prod_{p\in S}\left(a^p_{y_p,s_p}-\zeta^p_{y_p}\right)\prod_{p\notin S}\zeta^p_{y_p}
\end{aligned}$$

and

$$\begin{aligned}
\left(\bar{P}K\bar{P}^T\right)_{(S,s),(T,t)} = &\left(\prod_{p\in S\cap T}\sum_{c,c'\in\mathcal{A}}\left(\delta_{s_p=c}-\pi^p_c\eta\right)\left(\delta_{t_p=c'}-\pi^p_{c'}\eta\right)\right)\left(\prod_{p\notin S\cap T}\sum_{c,c'\in\mathcal{A}}\pi^p_c\eta\pi^p_{c'}\eta\right)\\
&\left(\prod_{p\in S\setminus T}\sum_{c,c'\in\mathcal{A}}\left(\delta_{s_p=c}-\pi^p_c\eta\right)\pi^p_{c'}\eta\right)\left(\prod_{p\in T\setminus S}\sum_{c,c'\in\mathcal{A}}\pi^p_c\eta\left(\delta_{t_p=c'}-\pi^p_{c'}\eta\right)\right)\\
= &\prod_{p\in S\cap T}\left(\bar{\zeta}^p-\zeta^p_{s_p}-\zeta^p_{t_p}+a^p_{s_p,t_p}\right)\prod_{p\in S\setminus T}\left(\zeta^p_{s_p}-\bar{\zeta}^p\right)\prod_{p\in T\setminus S}\left(\zeta^p_{t_p}-\bar{\zeta}^p\right)\prod_{p\notin S\cup T}\bar{\zeta}^p.
\end{aligned}$$

□

A.5 Variance component kernels that cannot be induced by order-dependent diagonal regularizers

Every sequence of positive λ_k defines a valid variance component (VC) kernel. Here we show that not all VC kernels can be induced by some order-dependent digaongal regularizer. Theorem 9 establishes that priors induced by order-dependent diagonal regularizers have dimension-normalized k^{th} order variance given by

$$\lambda_k = \sum_{j=k}^{\ell}\frac{1}{\alpha^j a_j}\binom{\ell-k}{j-k}.$$

Note that such λ_k's decrease with k, meaning that any sequence of non-decreasing λ_k's cannot be induced by an order-dependent diagonal regularizer. Moreover, it is not the case that any VC prior defined by decreasing λ_k is induced by an order-dependent diagonal regularizer. Indeed note that

$$\lambda_{\ell-1} = \frac{1}{\alpha^{\ell-1}a_{\ell-1}}+\frac{1}{\alpha^{\ell}a_{\ell}}$$

and for $k\leq\ell-1$,

$$\lambda_k \geq \frac{\ell-k}{\alpha^{\ell-1}a_{\ell-1}}+\frac{1}{\alpha^{\ell}a_{\ell}} \geq \lambda_{\ell-1}+\frac{\ell-k-1}{\alpha^{\ell-1}a_{\ell-1}}.$$

This restriction that λ_k cannot be arbitrarily close to $\lambda_{\ell-1}$ for order-dependent diagonal regularizers implies that not all VC priors with decreasing λ_k can be induced by an order-dependent diagonal regularizer.

A.6 Function space priors induced by diagonal regularizers for alternative bi-allelic weight spaces

Finally, we describe the function space priors induced by diagonal regularizers for different weight spaces in the bi-allelic case ($\alpha = 2$). When $\alpha = 2$, there are natural bases to use for regularized regression that are interpretable and not overparameterized [53, 35]: the Walsh-Hadamard basis (WH) and the wild-type basis (WT). Both have one basis element associated with each subset of positions $S \subseteq [\ell]$, but the meaning of the weight of each basis element is interpreted differently.

Walsh-Hadamard basis. We encode the alleles as $\{-1, +1\}$ and write a function f in terms of 2^ℓ weights w_S as

$$f = \sum_S w_S \prod_{p \in S} x_p = Hw \quad \text{where} \quad H_{x,S} = \prod_{p \in S} x_p.$$

Wild-type basis. We encode the wild-type allele as 0 and the mutant allele as 1. We write a function f in terms of 2^ℓ weights w_S as

$$f = \sum_S w_S \prod_{p \in S} \delta_{x_p=1} = Tw \quad \text{where} \quad T_{x,S} = \prod_{p \in S} \delta_{x_p=1}.$$

We apply the framework established in Section 3.2 to describe how regularized regression in these $\alpha^\ell = 2^\ell$-dimensional weight spaces induce function space priors. Here H or T will play the role of Φ. We show that any diagonal regularizers with the WT basis induces a heterosketdastic prior, whereas any diagonal regularizer with the WH basis induces a homoskedastic prior. Moreover, we describe how a subset of connectedness kernel can be induced with diagonal regularizers with the WH basis.

Theorem 10. *Let Λ be diagonal regularizer indexed by subsets of positions,*

$$\Lambda_{S,S} = \prod_{p \in S} \rho_p$$

where $\rho_p > 0$. Let $f^{MAP}(K, \sigma^2)$ be the MAP estimate for the Gaussian process $y = f_X + \varepsilon$ where $f \sim N(0, K)$ and $\varepsilon \sim N(0, \sigma^2 I)$.

1. *When used with the WH basis, the regularizer Λ induces the function space prior*

$$K^H_{x,y} = \left(\prod_p (1 + \rho_p^{-1}) \right) \prod_{p: x_p \neq y_p} \frac{1 - \rho_p^{-1}}{1 + \rho_p^{-1}},$$

meaning

$$Hw^{OPT}(\Lambda, \sigma^2) = f^{MAP}(K^H, \sigma^2)$$

where

$$w^{OPT}(\Lambda, \sigma^2) = \operatorname*{arg\,min}_{w \in \mathbb{R}^{2^\ell}} \|y - H_X w\|_2^2 + \sigma^2 w^T \Lambda w.$$

2. When used with the WT basis, the regularizer Λ induces the function space prior

$$K^{(T)}_{x,y} = \prod_{\substack{p \in S: \\ x_p = y_p = 1}} (1 + \rho_p^{-1}),$$

meaning

$$Tw^{OPT}(\Lambda, \sigma^2) = f^{MAP}(K^{(T)}, \sigma^2)$$

where

$$w^{OPT}(\Lambda, \sigma^2) = \arg\min_{w \in \mathbb{R}^{2^\ell}} \|y - T_X w\|_2^2 + \sigma^2 w^T \Lambda w.$$

Whereas K^H is a homoskedastic prior, $K^{(T)}$ is heteroskedastic; the variance of the wild-type sequence is one and the variance of the other sequences depends on the extent to which they differ from the wild-type (sequences that are more different from the wild-type tend to have higher variances). We can induce a connectedness prior with $0 < z^p < 1$ through regularized regression in the WH basis by choosing $\rho_p = (1 + z^p)/(1 - z_p)$ and rescaling by $1/\prod_p (1 + \rho_p^{-1})$.

Proof. (of Theorem 10). Taking Φ to be H or T, we apply Lemmas 3 and 4 to conclude that the regularizer Λ induces the kernels $H\Lambda^{-1}H^T$ and $T\Lambda^{-1}T^T$ on function space for the WH and WT bases respectively. Observe

$$\begin{aligned}
(H\Lambda^{-1}H^T)_{x,y} &= \sum_S \Lambda_S^{-1} \prod_{p \in S} x_p y_p \\
&= \sum_S \left(\prod_{p \in S} \rho_p^{-1} \right) \left(\prod_{\substack{p \in S \\ x_p \neq y_p}} (-1) \right) \\
&= \prod_{p : x_p = y_p} (1 + \rho_p^{-1}) \prod_{p : x_p \neq y_p} (1 - \rho_p^{-1}) \\
&= \left(\prod_p (1 + \rho_p^{-1}) \right) \prod_{p : x_p \neq y_p} \frac{1 - \rho_p^{-1}}{1 + \rho_p^{-1}}.
\end{aligned}$$

Similarly

$$(T\Lambda^{-1}T^T)_{x,y} = \sum_S \Lambda_S^{-1} \prod_{p \in S} \delta_{x_p = 1} \delta_{y_p = 1} = \sum_S \left(\prod_{\substack{p \in S: \\ x_p = y_p = 1}} \rho_p^{-1} \right) = \prod_{\substack{p \in S: \\ x_p = y_p = 1}} (1 + \rho_p^{-1}).$$

□