Kernelized Stein Discrepancies for Biological Sequences

Alan N. Amin
Program in Systems, Synthetic, and Quantitative Biology
Harvard University
Cambridge, MA 02138
alanamin@g.harvard.edu

Eli N. Weinstein
Data Science Institute
Columbia University
New York, NY 10027
ew2760@columbia.edu

Debora S. Marks
Department of Systems Biology
Harvard Medical School
Boston, MA 02115
debbie@hms.harvard.edu

Abstract

Generative models of biological sequences are a powerful tool for learning from complex sequence data, predicting the effects of mutations, and designing novel biomolecules with desired properties. The problem of measuring differences between high-dimensional distributions is central to the successful construction and use of generative probabilistic models. In this paper we propose the KSD-B, a novel divergence measure for distributions over biological sequences that is based on the kernelized Stein discrepancy (KSD). As for all KSDs, the KSD-B between a model and dataset can be evaluated even when the normalizing constant of the model is unknown; unlike any previous KSD, the KSD-B can be applied to arbitrary distributions over variable-length discrete sequences, and can take into account biological notions of mutational distance. Our theoretical results rigorously establish that the KSD-B is not only a valid divergence measure, but also that it detects convergence and non-convergence in distribution. We outline the wide variety of possible applications of the KSD-B, including (a) goodness-of-fit tests, which enable generative sequence models to be evaluated on an absolute instead of relative scale; (b) measurement of posterior sample quality, which enables accurate semi-supervised sequence design; and (c) selection of a set of representative points, which enables the design of libraries of sequences that are representative of a given generative model for efficient experimental testing.

1 Introduction

Generative models of biological sequences have wide and growing application, including in phylogenetic analysis, variant effect prediction, and protein design among many other areas [Hopf et al., 2017, Riesselman et al., 2018, Russ et al., 2020, Shin et al., 2021, Frazer et al., 2020, Davidsen et al., 2019]. A central challenge in constructing and using generative biological sequence models, as for all generative models, is evaluating divergences between distributions. Divergences can enable, for instance, careful measurement of mismatch between the model and the data, or mismatch between the model and samples that have been drawn from the model using some approximate sampling procedure. Constructing divergences between distributions over the space of biological sequences – taking into account the fact that sequences can have different lengths – presents unique challenges [Weinstein, 2022]. In particular, although many useful divergences have been constructed over Euclidean space (i.e. $\mathbb{R}^d$), biological sequence space differs in that it is both discrete (there are a finite number of amino acids/nucleotides) and infinite (sequences can be arbitrarily long). Moreover, notions of distance in biological sequence space differ substantially from standard Euclidean distance metrics: two sequences that differ by a single insertion/deletion would be considered close in biological sequence space, whereas "insertions" and "deletions" are not even well-defined concepts in standard

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Euclidean space. These issues present a major barrier to the application of a wide variety of valuable divergence-based methods to generative biological sequence models.

In this paper we construct the KSD-B, a divergence between distributions of sequences based on the kernelized Stein discrepancy (KSD) [Gorham and Mackey, 2017, Liu et al., 2016]. The KSD-B can be tractably computed for two distributions p and q given only unnormalized probabilities from p and samples from q. Moreover, the KSD-B can account for biologically relevant notions of sequence distance, through the choice of kernel [Ben-Hur et al., 2008]. Finally, the KSD-B comes with strong theoretical guarantees: it is faithful – it is zero if and only if q and p are equal – and it detects convergence and non-convergence – converges to zero if and only $q_1, q_2, \ldots$ converge to p.

These properties of the KSD-B make it uniquely able to address a number of challenging practical problems in evaluating and using generative biological sequence models. First, the KSD-B enables construction of nonparameteric goodness-of-fit tests; here the faithfulness of the KSD-B is crucial. Goodness-of-fit tests allow generative biological sequence models to be evaluated on an absolute scale, determining whether they match the data rather than just whether one model is better than another (as is the case, for instance, with standard held-out log likelihood evaluation). Second, the KSD-B enables measurement of the quality of a sequence of approximate samples from a posterior; here, the facts that the KSD-B can be applied to unnormalized probabilities, and can detect non-convergence, are crucial. Sampling from a posterior over sequences is central to the problem of semi-supervised sequence design and ancestral sequence reconstruction among other applications, but standard Markov chain convergence metrics cannot be used to check whether the samples in fact reflect the complete posterior distribution. Third, the KSD-B allows a set of representative sequences to be chosen from a distribution; here, the ability of the KSD-B to detect non-convergence is again crucial. When designing libraries of sequences to synthesize and test experimentally using generative models, choosing a set of representative points provides an efficient way of exploring the full range of model predictions in the laboratory. All of these applications and more make the KSD-B a valuable tool for working with generative biological sequence models.

Stein discrepancies have been previously developed for Euclidean space [Gorham et al., 2016, Gorham and Mackey, 2017, Gorham et al., 2020, Liu et al., 2016] and some finite discrete spaces with certain structures [Shi et al., 2022, Yang et al., 2018, Han et al., 2020]. We develop our method to give guarantees for distributions on the space of all sequences, which in particular is both discrete and infinite. We start by defining a Stein operator, replacing the gradient – which comes from the Langevin diffusion infinitesimal generator Gorham et al. [2016]– with locally balanced sampling Zanella [2020]. The domain of Stein discrepancies in Euclidean space can be interpreted as vector fields, so we define the domain of our Stein operator to also be vector fields, rather than functions, of sequences. Defining an integral probability metric with our Stein operator then gives us a divergence that is computationally tractable for a wide range of distributions, the KSD-B. Finally, we delineate assumptions that hold for many biologically relevant kernels and distributions, and show that they lead to strong theoretical guarantees for the KSD-B.

2 A novel discrepancy for biological sequence distributions

In this section we will define the KSD-B, a novel discrepancy for biological sequences, and show how it can be tractably calculated. The KSD-B builds on and extends the existing notion of a Stein discrepancy, a particular type of integral probability metric.

Integral probability metrics Let S be the infinite space of sequences, i.e. the set of all finite length strings drawn from a fixed alphabet (such as the 20 amino acids or the 4 nucleotides). We will start by considering a probability distribution p on S and data $X_1, \ldots, X_N \in S$. We can represent the data as an empirical distribution $q = \frac{1}{N}\sum_{n=1}^{N} \delta_{X_n}$ where δ_{X_n} is the distribution that has all its mass on $\{X_n\}$, and then compare the distributions p and q. One general method for measuring a discrepancy between distributions p and q is an Integral Probability Metric (IPM), defined as $\sup_{f\in\mathcal{F}} |E_q f - E_p f|$ for a chosen family $\mathcal{F}$ of functions on S, where E_p is the expectation under p.

If $\mathcal{F}$ is large enough, an IPM can detect if $p \neq q$ for any p and q, making it useful for building a consistent goodness-of-fit test. IPMs are a particularly useful choice of discrepancy measure for biological sequence models, where the ultimate goal is often to synthesize and test samples from a distribution p in the laboratory: so long as the (unknown) laboratory genotype-to-phenotype map f^* falls in the class $\mathcal{F}$, a small IPM guarantees that samples from p will have similar phenotypes to samples from q, as $|E_q f^* - E_p f^*| \leq \sup_{f\in\mathcal{F}} |E_q f - E_p f|$ [Weinstein et al., 2021, 2022].

Stein discrepancies. Unfortunately, depending on the family $\mathcal{F}$, evaluating $E_p f$ for $f \in \mathcal{F}$ may require samples or normalized probabilities from p, which are not always available (for instance, if p is an energy-based model, or the posterior of a complex Bayesian model). The Stein discrepancy solves this problem using a transformation $\mathcal{T}_p$ on functions of S, known as the Stein operator, such that $E_p \mathcal{T}_p f = 0$ for all $f \in \mathcal{F}$. Then, replacing $\mathcal{F}$ with $\mathcal{T}_p(\mathcal{F})$, the IPM is simply $\sup_{f \in \mathcal{F}} E_q \mathcal{T}_p f$ which is potentially much easier to compute.

A Stein discrepancy for biological sequences. Existing approaches to constructing Stein discrepancies typically employ Stein operators that rely on gradients of $p(x)$ and $f(x)$ with respect to x Liu et al. [2016]. As the space S is neither continuous nor finite, such approaches cannot be applied directly and are nontrivial to generalize. In order to construct a Stein operator for biological sequences we build on the generator method of Barbour [1990], which constructs a Stein operator $\mathcal{T}_p$ using a continuous-time Markov process with stationary distribution p Gorham et al. [2016], Shi et al. [2022]. The basic intuition behind the generator method is that if we evolve the data distribution q according to an infinitesimal step of the Markov process, the only way for the expectation of all functions $f \in \mathcal{F}$ to be constant is if the data distribution q matches the stationary distribution p exactly. Whereas the gradients in standard Stein discrepancies arise from the use of overdamped Langevin diffusion as the Markov process, we rely on Markov processes appropriate for biological sequence space, with infinitesimal transitions corresponding to substitutions, insertions and deletions.

Our first step is to expand the standard definition of Stein discrepancies: instead of letting each $f \in \mathcal{F}$ take as input a single sequence X, we let each $f \in \mathcal{F}$ take as input two sequences. This extension will allow us to endow $\mathcal{F}$ with enough additional structure to construct tractable Stein discrepancies, while remaining flexible enough to detect differences between any two distributions p and q. Define a relation M on S such that X and Y are related if Y can be reached from X via a single mutation - either a single substitution, a single insertion of a single letter, or a single deletion of a single letter. We will write this as $(X, Y) \in M$ or XMY. Following Chow et al. [2017] we will define vector fields on S to be functions $f : M \to \mathbb{R}$ such that $f(X, Y) = -f(Y, X)$ for all $(X, Y) \in M$, i.e. f must satisfy an anticommutativity property. We will work with families $\mathcal{F}$ consisting of vector fields f. For any $g : S \to \mathbb{R}$, we also define the vector field $\nabla g(X, Y) = g(Y) - g(X)$ for $(X, Y) \in M$. This provides our generalized notion of a gradient in biological sequence space.

Now we will define our Stein operator and use it construct an IPM. To construct the Markov process over sequences, we build on locally balanced sampling procedures [Zanella, 2020, Shi et al., 2022]. Consider a continuous non-negative function χ with the property that $\chi(t) = t\chi(1/t)$ for all $t > 0$ and $\chi(0) = 0$; examples include $\chi(t) = \sqrt{t}$ and $\chi(t) = \min\{t, 1\}$ the latter of which is used in Metropolis Hastings correction steps. Let p be a distribution on S. For $(X, Y) \in M$ with $p(X) > 0$, define the infinitesimal transition probability

$$T_{p,X \to Y} = \#\{\text{single mutations taking } X \text{ to } Y\} \chi\left(\frac{p(Y)}{p(X)}\right).$$

Let $T_{p,X \to Y} = \infty$ on the rest of M. Thus, by our choice of χ, the Markov process satisfies detailed balance, i.e. $T_{p,X \to Y} p(X) = T_{p,Y \to X} p(Y)$ where we define $\infty \times 0 = 0$ throughout. Define the Stein operator $\mathcal{T}_p$ taking vector fields to functions on the support of p, $\text{supp}(p) = \{X \mid p(X) > 0\}$, such that for a vector field f on S,

$$(\mathcal{T}_p f)(X) = \sum_{Y \in S \mid YMX} T_{p,X \to Y} f(X, Y).$$

Comparing pairs (X, Y) and (Y, X) in M, we have informally, applying the antisymmetry property,

$$E_p \mathcal{T}_p f = \frac{1}{2} \sum_{(X,Y) \in M} p(X) T_{p,X \to Y} f(X, Y) + p(Y) T_{p,Y \to X} f(Y, X) = 0.$$

Thus, if we select a family of vector fields $\mathcal{F}$, we can define the IPM on $\mathcal{T}_p(\mathcal{F})$, $\sup_{\tilde{f} \in \mathcal{T}_p(\mathcal{F})} |E_q \tilde{f} - E_p \tilde{f}| = \sup_{f \in \mathcal{F}} |E_q \mathcal{T}_p f - E_p \mathcal{T}_p f| = \sup_{f \in \mathcal{F}} E_q \mathcal{T}_p f$. To compute $E_q \mathcal{T}_p f$ for a given f, one only needs samples from q to take the expectation and unnormalized probabilities from p to calculate $\mathcal{T}_p f$.

The KSD-B: A kernelized Stein discrepancy for biological sequences. Next, we need to choose a specific family of vector fields $\mathcal{F}$ to apply our Stein operator to; this family should be sufficiently large to guarantee that the Stein discrepancy can detect differences between any two distributions, but also provide sufficient structure such that the Stein discrepancy is computationally tractable. A standard existing approach is to use a reproducing kernel Hilbert space (RKHS), $\mathcal{H}_k$, where k is a symmetric positive definite kernel defined over the data space. One can then take $\mathcal{F}$ to be the unit ball

in the RKHS, $\{f \mid \|f\|_k \leq 1\}$, where $\|\cdot\|_k$ is the norm on the RKHS [Gorham and Mackey, 2017, Liu et al., 2016]. In our case, however, we need the RKHS to consist of vector fields. Thus, we define a *vector field kernel* as a kernel k on M such that all $f \in \mathcal{H}_k$ are vector fields. We will discuss in appendix A.4 how to build vector field kernels. Given a vector field kernel k, we define the kernelized Stein discrepancy for biological sequences (KSD-B) as $\text{KSD-B}_{p,k}(q) = \sup_{\|f\|_k \leq 1} E_q \mathcal{T}_p f$.

Previous works on Stein discrepancies for finite discrete spaces did not use vector fields, instead working with scalar fields, defining kernels on the space of fixed-length sequences and using Stein operators of the form $\mathcal{T}_p \nabla$ [Shi et al., 2022, Yang et al., 2018]. This approach is a special case of our KSD-B, using a particular choice of vector field kernel k^∇ for a kernel k on S, as shown in Proposition A.2. We will see in Section 3 that in the infinite discrete setting relevant for biological sequences, the scalar field approach cannot provide strong theoretical guarantees except with pathological kernels.

Finally we show that the KSD-B is computationally tractable and formalize our previous argument that $E_p \mathcal{T}_p f = 0$. We say a distribution q on S is p, k-integrable if $E_{X \sim q} \sum_{Y \in S \mid YMX} T_{p,Y \to X} \sqrt{k((X,Y),(X,Y))} < \infty$. Note this implies $\text{supp}(q) \subseteq \text{supp}(p)$.

Proposition 2.1. *Say k is a vector field kernel and q is a p, k-integrable distribution on S.*

$$\text{KSD-B}_{p,k}(q) = E_{X,X' \sim q} \sum_{YMX, Y'MX'} T_{p,X \to Y} T_{p,X' \to Y'} k((X,Y),(X',Y')). \tag{1}$$

If p is p, k-integrable, then for all $f \in \mathcal{H}_k$, $E_p \mathcal{T}_p f = 0$.

Equation 1 can be computed if one can sample from q and has unnormalized probabilities from p.

3 Detecting convergence and non-convergence of distributions

In this section we will demonstrate the theoretical properties of KSD-Bs that make them useful for goodness-of-fit tests, evaluating sample quality, and choosing representative points. Much of our results are inspired by techniques developed in Gorham et al. [2016], Gorham and Mackey [2017].

KSD-B is faithful. For the KSD-B to be useful as a nonparametric goodness-of-fit test, it must be able to detect if a model distribution p matches a data distribution q. In particular, the divergence must be faithful, that is, $\text{KSD-B}_{p,k}(q) \to 0 \iff p = q$. Given the KSD-B is an IPM, faithfulness will hold provided $\mathcal{H}_k$ is large enough. For KSDs on continuous spaces, faithfulness is usually guaranteed via a universality assumption on the kernel k, namely that $\mathcal{H}_k$ is dense in some function space. In discrete space, however, kernels may satisfy a more powerful condition: their RKHS may include all delta functions, in which case we say the kernel is "deltable". Deltability is formally defined in Definition A.5. In the following proposition, we show deltability ensures faithfulness, and thus so long as we use a deltable kernel, our KSD-B provides a consistent goodness-of-fit test.

Proposition 3.1. *Say supp(p) is connected. If k is a deltable vector field kernel or k is a deltable scalar field kernel on S and $\sup_n E_q \sum_{YMX} T_{p,Y \to X} < \infty$ then $\text{KSD-B}_{p,k}(q) = 0$ only if $p = q$.*

KSD-B detects convergence and non-convergence. For the KSD-B to be useful in evaluating sample quality, it must be able to determine whether or not a sequence of empirical distributions $q_1, q_2, \ldots$ (corresponding to the samples) converges to a distribution p (corresponding to the model). Formally, we hope that the KSD-B converges to zero, i.e. $\text{KSD-B}_{p,k}(q_n) \to 0$, if and only if q_n converges to p by some natural metric of convergence, such as convergence in distribution. The same concern holds if we are choosing a set of representative points: as we optimize $\text{KSD-B}_{p,k}(q_n)$ with respect to the empirical distribution q_n, we hope that $\text{KSD-B}_{p,k}(q_n) \to 0$ implies q_n converges to p and vice versa, i.e. our chosen points reflect p more and more accurately.

We start by showing that the KSD-B detects non-convergence, i.e. $\text{KSD-B}_{p,k}(q_n) \to 0$ implies q_n converges to p in distribution. In Proposition A.9 we give an example that demonstrates that if p has "non-uniformly" decreasing tails, the KSD-B may not detect non-convergence. We will thus need to assume that p has "uniformly" decreasing tails (Assumption A.10): after a certain length, longer and longer sequences are sufficiently less likely under p. This assumption holds for some models and not others (Section A.3). In Propositions A.11 and A.12 we show that the KSD-B may also fail to detect non-convergence if we allow k to have thin tails. In particular, for scalar field KSDs, k cannot be allowed to be bounded; thus no non-pathological choice of k will give us scalar field KSD-Bs that can detect non-convergence. We thus further assume that k has thick (possibly unbounded) tails in Assumption A.13. We provide examples of kernels that satisfy all our required assumptions in Section A.4.3. With these assumptions, we can guarantee that the KSD-B detects non-convergence.

Theorem 3.2. *Say p is a distribution on S obeying assumption A.10 and k is a deltable vector field kernel obeying Assumption A.13 A or a deltable kernel on S obeying Assumption A.13 B. Say $(q_n)_n$ is a sequence of distributions on S. If $\text{KSD-B}_{p,k}(q_n) \to 0$ then q_n converges to p in distribution.*

Finally, we show that the KSD-B detects convergence, i.e. if q_n converges to p in some (weighted) total variation metric, then KSD-B$_{p,k}(q_n) \to 0$.

Proposition 3.3. *Say k is a vector field kernel and $p, q_1, q_2, \ldots$ are p, k-integrable distributions on S. Call $A(X) = \sum_{Y \in MX} T_{p,Y \to X} \sqrt{k((X,Y),(X,Y))}$. If $\sum_X |p(X) - q_n(X)| A(X) \to 0$ then* KSD-B$_{p,k}(q_n) \to 0$.

Note if one is working with a scalar field KSD-B then k must be unbounded, and thus the weight A larger, making more difficult to detect convergence.

4 Conclusion

In this paper we've defined a novel, computationally tractable discrepancy on the space of biological sequences, the KSD-B, and established theoretical results showing it can be used for goodness-of-fit testing, evaluating the quality of approximate samples from a posterior, and choosing a set of representative points from a distribution. In future work we aim to illustrate these applications on simulated and real data. We believe that the KSD-B can serve as a valuable tool for generative biological sequence modeling broadly, helping to ensure that generative models are accurate, reliable and trustworthy as they see growing use across biology, biotechnology and biomedicine.

5 Aknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

A D Barbour. Stein's method for diffusion approximations. *Probab. Theory Related Fields*, 84(3): 297–322, 1990.

Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. Support vector machines and kernels for computational biology. *PLoS Comput. Biol.*, 4(10):e1000173, October 2008.

Shui-Nee Chow, Wuchen Li, and Haomin Zhou. Entropy dissipation of Fokker-Planck equations on graphs. January 2017.

Kristian Davidsen, Branden J Olson, William S DeWitt, 3rd, Jean Feng, Elias Harkins, Philip Bradley, and Frederick A Matsen, 4th. Deep generative models for T cell receptor protein sequences. *Elife*, 8, September 2019.

Randal Douc, Gersende Fort, and Arnaud Guillin. Subgeometric rates of convergence of f-ergodic strong markov processes. *Stochastic Process. Appl.*, 119(3):897–923, March 2009.

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Kelly Brock, Yarin Gal, and Debora S Marks. Large-scale clinical interpretation of genetic variants using evolutionary data and deep learning. *bioRxiv*, page 2020.12.21.423785, 2020.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. March 2017.

Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, and Lester Mackey. Measuring sample quality with diffusions. November 2016.

Jackson Gorham, Anant Raj, and Lester Mackey. Stochastic stein discrepancies. July 2020.

Will Grathwohl, Kevin Swersky, Milad Hashemi, David Duvenaud, and Chris J Maddison. Oops I took a gradient: Scalable sampling for discrete distributions. 2021.

Martin Hairer. Convergence of markov processes, 2021.

Jun Han, Fan Ding, Xianglong Liu, Lorenzo Torresani, Jian Peng, and Qiang Liu. Stein variational inference for discrete distributions. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4563–4572. PMLR, 2020.

Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta P I Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.*, 35(2):128–135, 2017.

Thomas M Liggett. *Continuous time Markov processes: An introduction*. Graduate studies in mathematics. American Mathematical Society, Providence, RI, March 2010.

Qiang Liu, Jason D Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. *33rd International Conference on Machine Learning, ICML 2016*, 1(1):448–461, 2016.

Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods*, 15(10):816–822, 2018.

Philippe Rigollet and Jan-Christian Hütter. High dimensional statistics. lecture notes, November 2019.

William P Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, and Rama Ranganathan. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, 2020.

Jiaxin Shi, Yuhao Zhou, Jessica Hwang, Michalis K Titsias, and Lester Mackey. Gradient estimation with discrete stein operators. February 2022.

Jung-Eun Shin, Adam J Riesselman, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.*, 12(1):2403, April 2021.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert R G Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.*, 12(70):2389–2410, 2011.

Eli N Weinstein. *Generative Statistical Methods for Biological Sequences*. PhD thesis, Harvard University, Ann Arbor, United States, 2022.

Eli N Weinstein, Alan N Amin, Will Grathwohl, Daniel Kassler, Jean Disset, and Debora S Marks. Optimal design of stochastic DNA synthesis protocols based on generative sequence models. October 2021.

Eli N Weinstein, Alan N Amin, Jonathan Frazer, and Debora S Marks. Non-identifiability and the blessings of misspecification in models of molecular fitness and phylogeny. January 2022.

Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-Fit testing for discrete distributions via stein discrepancy. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5561–5570. PMLR, 2018.

Giacomo Zanella. Informed proposals for local MCMC in discrete spaces. *J. Am. Stat. Assoc.*, 115 (530):852–865, April 2020.

A Proofs

In this appendix we prove our assertions made in the main text. First in section A.1 we lay out our notation. Next, in section A.2 we prove results about the ability of the KSD-B to detect convergence and non-convergence. In section A.3 we describe whether or not certain distributions over sequences satisfy the assumption that they have "uniformly thin tails" made in section A.2. Finally, in section A.4, we lay out examples of kernels that may be used to create KSD-Bs that are able to detect convergence and non-convergence.

A.1 Notation

Let our alphabet, $\mathcal{B}$, be a finite set at the set of all sequences be defined as $S = \cup_{i=0}^{\infty} \mathcal{B}^i$ where $\mathcal{B}^0$ is defined to only contain the empty string $\emptyset$. If p is a distribution on S let $\text{supp}(p) = \{X \mid p(X) > 0\}$ and $M_{p,p} = \{(X, Y) \in M \mid X, Y \in \text{supp}(p)\}$. We will say p has connected support if $\text{supp}(p)$ is a connected set in the graph with vertices S and edges M. Finally, for $X \in S$, define $\text{flux}_p(X) = \sum_{YMX} T_{p,X\to Y}$.

We define the set of bounded functions on S $C_b(S)$, the set of functions on S vanishing at infinity C_0, and the set of functions on S that are non-zero at only finitely many points and $C_C(S)$. We also define the set of all vector fields that are non-zero on only finitely many points in M as $C_{C,vf}(M)$. We define $\|\cdot\|_\infty$ as the infinity norm on $C_b(S)$. For two distributions μ, ν on S, call $\|\nu - \mu\|_{\text{TV}}$ their distance in total variation.

For two sequences of real numbers $(a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}}$, both possibly undefined for small n, we write $a_n \lesssim b_n$ to mean that there is a positive constant C such that eventually $a_n \leq Cb_n$. We write $a_n \sim b_n$ when $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. We write $a_n = O(b_n)$ if $a_n \lesssim b_n$ and $a_n = o(b_n)$ if $\frac{|a_n|}{|b_n|} \to 0$. We define $a \wedge b$ as the minimum of a and b, and $a \vee b$ as the maximum.

A kernel on a set H is a symmetric function $k : H \times H \to \mathbb{R}$ that is "non-negative definite", i.e. for all $X_1, \ldots, X_N \in H$, $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$, $\sum_{n=1}^N \sum_{n'=1}^N \alpha_n \alpha_{n'} k(X_n, X_{n'}) \geq 0$. We also require that $k(X, X) > 0$ for all $X \in S$. For every $X \in S$ define the function $k_X = k(X, \cdot)$. Define the dot product $(\cdot|\cdot)_k$ on linear combinations of these functions with $(k_X|k_Y) = k(X, Y)$ and call the associated norm $\|\cdot\|_k$. Let $\mathcal{H}_k$ be the Hilbert space completion of the span of $\{k_X\}_{X\in H}$ under $(\cdot|\cdot)_k$ and call this the reproducing kernel Hilbert space (RKHS) of k. Elements of the RKHS can be understood as functions on H by $(f|k_X) = f(X)$.

Say k is a kernel on a space H and $A : H \to (0, \infty)$. We call $k^A(X, Y) = A(X)k(X, Y)A(Y)$ the kernel k "tilted" by A. k^A is a kernel on H and the transformation that takes $g \in \mathcal{H}_k$ to $X \mapsto g(X)A(X)$ is a unitary isomorphism to $\mathcal{H}_{k^A}$.

A.2 Proofs for properties of KSD-Bs

In this section we prove the results described in the main text for KSD-Bs. We will first look at how to compute KSD-Bs in section A.2.1. In this section we will also show that scalar field KSD-B may be written as an instance of vector field KSD-Bs. Next, in section A.2.2 we will look at a stochastic process related to the KSD-B. We will show that this process exists and list some of its properties that will be useful in later proofs. In section A.2.3 we look at conditions under which the KSD-B is faithful. In section A.2.4 we look at when the KSD-B can detect convergence and non-convergence. We will look at several examples in which the KSD-B cannot detect non-convergence, motivating further assumptions on the distribution and kernel we consider. We state these assumptions and show that, with them, the KSD-B can detect non-convergence. Finally we show that the KSD-B can also detect convergence. In section A.2.5 we will construct and prove the properties of the examples described in section A.2.4. Finally, in section A.2.6 we look at approximating the KSD-B by quantities that are more cheaply computable such that we can bound our error.

A.2.1 Scalar and vector field KSD-Bs and their computability

In this section we demonstrate that KSD-Bs are computable and that scalar field KSD-Bs can be understood as an instance of vector field KSD-Bs.

We first write the KSD-B in two other forms, one of which is easily computable, and the other of which is of theoretical use.

Proposition A.1. *Say k is a vector field kernel and q is a p, k-integrable distribution on S. Then for all $f \in \mathcal{H}_k$,*

$$E_q \mathcal{T}_p f = \frac{1}{2} \sum_{(X,Y)\in M_{p,p}} p(Y) T_{p,Y\to X} \left(\frac{q(X)}{p(X)} - \frac{q(Y)}{p(Y)} \right) f(X, Y) \tag{2}$$

and

$$\text{KSD-B}_{p,k}(q)^2 = E_{X,X'\sim q} \sum_{YMX,Y'MX'} T_{p,X\to Y} T_{p,X'\to Y'} k((X, Y), (X', Y')). \tag{3}$$

If p is p, k-integrable, then for all $f \in \mathcal{H}_k$, $E_p \mathcal{T}_p f = 0$.

Proof. Say q is p, k-integrable. Define $\phi_q : \mathcal{H}_k \to \mathbb{R} \mid f \mapsto E_q \mathcal{T}_p f$ For $f \in \mathcal{H}_k$,

$$\begin{aligned} \phi_q(f) =& E_{X\sim q} \sum_{YMX} T_{p,X\to Y}(f|k_{(X,Y)})_k \\ \leq& \|f\|_k E_{X\sim q} \sum_{YMX} T_{p,X\to Y} \sqrt{k((X,Y),(X,Y))}. \end{aligned} \tag{4}$$

Thus ϕ_q is a bounded linear operator on $\mathcal{H}_k$ and is thus a member of $\mathcal{H}_k$. As well, $\text{KSD-B}_{p,k}(q) = \|\phi_k\|_k$.

$$\begin{aligned} (\phi_q|\phi_q)_k =& \phi_q(\phi_q) \\ =& E_{X\sim q} \sum_{YMX} T_{p,X\to Y} \phi_q(k_{(X,Y)}) \\ =& E_{X\sim q} \sum_{YMX} T_{p,X\to Y} E_{X'\sim q} \sum_{Y'MX'} T_{p,X'\to Y'} k_{(X,Y)}(X',Y') \\ =& E_{X,X'\sim q} \sum_{YMX} \sum_{Y'MX'} T_{p,X\to Y} T_{p,X'\to Y'} k((X,Y),(X',Y')). \end{aligned}$$

Note that since all quantities in the expectation and sum are positive, equation 5 shows the absolute integrability of the expectation and sum. Thus we can rearrange terms to get

$$\begin{aligned} \phi_q(f) =& E_{X\sim q} \sum_{YMX} T_{p,X\to Y} f(X,Y) \\ =& \sum_{(X,Y)\in M_{p,p}} q(X) T_{p,X\to Y} f(X,Y) \\ =& \frac{1}{2} \sum_{(X,Y)\in M_{p,p}} (q(X) T_{p,X\to Y} f(X,Y) + q(Y) T_{p,Y\to X} f(Y,X)) \\ =& \frac{1}{2} \sum_{(X,Y)\in M_{p,p}} p(Y) T_{p,Y\to X} \left(\frac{q(X)}{p(X)} - \frac{q(Y)}{p(Y)} \right) f(X,Y). \end{aligned}$$

The statement about p follows from the above equation with $p = q$ noting $\frac{q(X)}{p(X)} = \frac{q(Y)}{p(Y)}$ for all $(X,Y) \in M_{p,p}$. □

Equation 2 gives some intuition on the effect of the kernel on the value of the KSD-B: note that, for $(X,Y) \in M_{p,p}$, $T_{p,Y\to X} \geq 0$, so the KSD-B uses vector fields $f \in \mathcal{H}_k$ to detect non-zero "differences in slopes" $p(Y) \left(\frac{q(X)}{p(X)} - \frac{q(Y)}{p(Y)} \right) = q(X) \left(\frac{p(Y)}{p(X)} - \frac{q(Y)}{q(X)} \right)$. A kernel with a large enough $\mathcal{H}_k$ can thus detect more subtle differences in these slopes.

We end the section by demonstrating that scalar field KSD-Bs can be written as an instance of vector field KSD-Bs.

Proposition A.2. *For a kernel k on S, define the kernel*

$$k^\nabla((X,Y),(X',Y')) = k(Y,Y') - k(X,Y') - k(Y,X') + k(X,X')$$

for $(X,Y),(X',Y') \in M$. k^∇ is a vector field kernel and if q is a p, k^∇-integrable distribution on S then

$$\sup_{\|f\|_{k^\nabla}\leq 1} E_q \mathcal{T}_p f = \sup_{\|f\|_k \leq 1} E_q \mathcal{T}_p \nabla f.$$

Proof. k^∇ is non-negative definite as if $(X_1,Y_1).\ldots,(X_N,Y_N) \in M$ and $\alpha_1,\ldots,\alpha_N \in \mathbb{R}$ then, calling $f = \sum_n \alpha_n k_{X_n}$ and $g = \sum_n \alpha_n k_{Y_n}$,

$$\sum_{n,m} \alpha_n \alpha_m k^\nabla((X_n,Y_n),(X_m,Y_m)) = (g|g)_k - (f|g)_k - (g|f)_k + (f|f)_k = \|f-g\|_k \geq 0.$$

One can also verify that $k^{\nabla}_{(X,Y)} = -k^{\nabla}_{(Y,X)}$ for all $(X,Y) \in M$, so for every $f \in \mathcal{H}_{k^\nabla}$,

$$f(X,Y) = (f|k^{\nabla}_{(X,Y)})_{k^\nabla} = -(f|k^{\nabla}_{(Y,X)})_{k^\nabla} = -f(Y,X).$$

Define, similar to Proposition A.1, $\tilde{\phi}_q : \mathcal{H}_k \to \mathbb{R} \mid f \mapsto E_q \mathcal{T}_p \nabla f$. Note $k^\nabla((X,Y),(X,Y)) = k(X,X) - 2k(X,Y) + k(Y,Y) = \|k_X - k_Y\|_k^2$. For $f \in \mathcal{H}_k$,

$$\begin{aligned}
\tilde{\phi}_q(f) =& E_{X\sim q} \sum_{YMX} T_{p,X\to Y}(f|k_Y - k_X)_k \\
\leq& \|f\|_k E_{X\sim q} \sum_{YMX} T_{p,X\to Y} \|k_Y - k_X\|_k \\
\leq& \|f\|_k E_{X\sim q} \sum_{YMX} T_{p,X\to Y} \sqrt{k^\nabla((X,Y),(X,Y))}.
\end{aligned} \tag{5}$$

Thus $\tilde{\phi}_q$ is a bounded linear operator on $\mathcal{H}_k$ and is thus a member of $\mathcal{H}_k$. As well, $\left(\sup_{\|f\|_k \leq 1} E_q \mathcal{T}_p \nabla f\right)^2 = \|\phi_k\|_k^2$.

$$\begin{aligned}
(\tilde{\phi}_q|\tilde{\phi}_q)_k =& \tilde{\phi}_q(\tilde{\phi}_q) \\
=& E_{X\sim q} \sum_{YMX} T_{p,X\to Y} \left(\tilde{\phi}_q(k_Y - k_X)\right) \\
=& E_{X\sim q} \sum_{YMX} T_{p,X\to Y} \\
& \times \left(E_{X'\sim q} \sum_{Y'MX'} T_{p,X'\to Y'} \left((k_Y(Y') - k_X(X')) - (k_Y(X') - k_X(X'))\right) \right) \\
=& E_{X,X'\sim q} \sum_{YMX,Y'MX'} T_{p,X\to Y} T_{p,X'\to Y'} k^\nabla((X,Y),(X',Y')) \\
=& \text{KSD-B}_{p,k}(q)^2.
\end{aligned}$$

□

A.2.2 Stochastic processes on sequences

As described above, the KSD-B is connected with a particular stochastic process defined by transition rates $T_{p,X\to Y}$ which depend on p and g. In this section we prove that this process is well defined given an integrability condition and prove some properties of this process that will be of use in proving result about the KSD-B.

First we rigorously define this process. Let p be a distribution on S. Define $\mathcal{L}_p = \mathcal{T}_p \nabla$; this is a matrix indexed by $\text{supp}(p)$ defined by $\mathcal{L}_{p,X,Y} = \mathcal{L}_p(\delta_X)(Y)$. We have that $\mathcal{L}_{p,X,Y} = T_{p,X\to Y} \geq 0$ if $X \neq Y$ and $\mathcal{L}_{p,X,X} = -\sum_{Y\neq X} \mathcal{L}_{p,X\to Y} = -\text{flux}_p(X)$. Such a matrix is called a Q-matrix and can define a Markov process as follows. First let $K_{X\to Y} = \mathcal{L}_{p,X,Y}/\text{flux}_p(X)$ if $X \neq Y$ and 0 if $X = Y$. The entries of K are positive and its rows sum to 1 so it defines a discrete-time stochastic process $(Z_0, Z_1, \dots)$ known as the "underlying stochastic process". Now define the transition times $\tau_n \sim \text{Exp}(\text{flux}_p(Z_n))$ and the process $(X_t)_t$ where $X_t = Z_n$ if $\tau_{n-1} \leq t < \tau_n$. This defines a family of transition probabilities $(P_t)_t$ where $P_t(X)$ is a positive measure on S describing the distribution of X_t given $X_0 = X$ for any t, X. Now if $f \in C_C(S)$, we can define the function $P_t f(X) = E_{P_t(X)} f$. These functions are continuously differentiable in t and the backwards Kolmogorov equation holds, i.e. $\frac{d}{dt} P_t f(X) = \mathcal{L}_p P_t f(X)$ (see section 2.5 of Liggett [2010]).

Unfortunately, the above results do not rule out some possible pathologies when $\text{supp}(p)$ is infinite. This is because $(X_t)_t$ may "explode", i.e. transition infinitely many times in finite time. This can manifest in $\sum_Y P_t(X)(Y) < 1$ or the forward Kolmogorov equation - $\frac{d}{dt} P_t f(X) = P_t \mathcal{L}_p f(X)$ - failing to hold. To avoid these pathologies, we add an integrability condition on p, namely that $E_p \text{flux}_p < \infty$. The below lemma shows that in this case P_t are valid probability distributions and

the forward Kolmogorov equation holds. We also list some consequences of these results that will help prove future results.

Let us also introduce a definition: for a Markov Matrix P, we call a measure p on S invariant if $E_p Pf = E_p f$ for all $f \in C_C(S)$. The following lemma proves that the above described stochastic process exists and lists some of its properties.

Lemma A.3. *Say p has connected support and $E_p \text{flux}_p < \infty$.*

(A) There is a Markov process $(X_t)_t$ on $\text{supp}(p)$ such that for all $f \in C_C(S)$, defining $P_t(f)(X) = E[f(X_t)|X_0 = X]$, $P_t f(X)$ is continuously differentiable in t and $\frac{d}{dt} P_t f(X) = \mathcal{L} P_t f(X) = P_t \mathcal{L} f(X)$.

(B) P_t are stationary for p and if q is another distribution with $E_q \text{flux}_p < \infty$ then $q = p$ if and only if $E_q \mathcal{L}_p f = 0$ for all $f \in C_C(S)$.

(C) If $f \in C_C(S)$, $f(X_t) - \int_0^t \mathcal{L}_p f(X_s) ds$ is a Martingale in t conditional on $X_0 = X$ for every $X \in \text{supp}(p)$.

Proof. Take K, $(Z_n)_n$, $(\tau_n)_n$, $(X_t)_t$ and P_t defined as above. $(Z_n)_n$ is an irreducible Markov chain by definition as $\text{supp}(p)$ is connected. To show that the P_t indeed define probability distributions, note that $\nu = \text{flux}_p p$ is a finite measure on S that is stationary with respect to K since $\text{flux}_p(X) p(X) K_{X \to Y} = \text{flux}_p(Y) p(Y) K_{Y \to X}$. This implies that $(Z_n)_n$ will visit each $X \in \text{supp}(p)$ infinitely many times almost surely. To see this, assume $(Z_n)_n$, starting at some point, visits an $X \in \text{supp}(p)$ only finitely many times with positive probability. Since $(Z_n)_n$ is irreducible, every time Z_n hits X there is a fixed chance that it never hits X again, so, almost surely, Z_n hits X only finitely many times. Let $\hat{\nu} = \nu(X)/\nu(S)$ so, since $\hat{\nu}$ is stationary for K,

$$\hat{\nu}(X) = \int d\hat{\nu}(Y)(K^m)_{Y \to X} = E_{Z_0 \sim \hat{\nu}} \left[\mathbb{1}(Z_m = X) \right] \to 0$$

as $m \to \infty$ by dominated convergence, a contradiction. Thus, by Corollary 2.34 (b) of Liggett [2010], P_t are distributions on S and $\sum_n \tau_n = \infty$ almost surely, that is, $(X_t)_t$ is a well defined Markov process. We also have that $P_t f(X) = E[f(X_t)|X_0 = X]$ for all X, t.

For the second claim, first note $E_q \text{flux}_p < \infty$ implies $\text{supp}(q) \subseteq \text{supp}(p)$ since if $X \notin \text{supp}(X)$, $T_{p, X \to Y}$ is defined to be ∞. By equation 2.40 of Liggett [2010], if q is a distribution on S such that $E_q \text{flux}_p(X) < \infty$, q is stationary for all P_t if and only if $E_q \mathcal{L}_p \delta_X = 0$ for all $X \in \text{supp}(p)$. In particular, p is stationary for all P_t. On the other hand, by our construction of $(X_t)_t$, since $\text{supp}(p)$ is connected, by Proposition 2.6 of Hairer [2021], each P_t has at most one stationary distribution for $t > 0$. Thus, $p = q$ if and only if $E_q \mathcal{L}_p f = 0$ for all $f \in C_C(S)$.

To show that we also have the forward Kolmogorov equation it suffices by Theorem 2.39 of Liggett [2010] to show that $P_t \text{flux}_p(X) < \infty$ for all t, X. To see this, note by the fact that p is stationary for P_t,

$$E_p \text{flux}_p \geq E_p \left(\text{flux}_p \vee N \right) = E_p P_t \left(\text{flux}_p \vee N \right) \to E_p P_t \text{flux}_p$$

as $N \to \infty$ by monotone convergence so that $P_t \text{flux}_p(X) < \infty$ for all $t > 0, X \in \text{supp}(p)$.

The statement about Martingales holds by the backwards and forwards Kolmogorov equations and theorem 3.32 of Liggett [2010].

□

We will also note the following theorem from Hairer [2021] which will help us determine the convergence rates of the stochastic processes.

Theorem A.4. *(theorem 4.1 of Hairer [2021]) Say p has connected support and $E_p \text{flux}_p < \infty$. Say $V : S \to [1, \infty)$ is a function such that $V(X) \to \infty$ as $|X| \to \infty$. $\mathcal{L}_p V \leq K - \varphi \circ V$ on $\text{supp}(p)$ for some strictly concave $\varphi : [0, \infty) \to [0, \infty)$ with $\phi(0) = 0$ and increasing to infinity. Now define $H(u) = \int_1^u \varphi(s)^{-1} ds$. Then there is a $C > 0$ such that for all $X \in \text{supp}(p)$,*

$$\|P_t(X) - p\|_{\text{TV}} \leq \frac{CV(X)}{H^{-1}(t)} + \frac{C}{\varphi \circ H^{-1}(t)}.$$

Proof. All conditions of the theorem are obviously satisfied except for the fact that $V(X(t)) - \int_0^t ds\,(K - \varphi \circ V(X(s)))$ is a local super-martingale conditioned on $X_0 = X$ for some $X \in \text{supp}(p)$. This follows from Theorem 3.4 of Douc et al. [2009] if $M_t = V(X_t) - \int_0^t \mathcal{L}V(X_s)ds$ defines a local Martingale when $X_0 = X$ for all $X \in \text{supp}(p)$.

To show this, for every number N and $X \in S$, call $V^N(X) = V(X)$ if $V(X) \leq N$ and $V^N(X) = 0$ otherwise so that $V_N \in C_C(S)$. Also define $T_N = \inf\{t \mid \exists Y s.t.\ Y M X_t \text{ and } V(Y) \geq N\}$. T_N is a stopping time and $T_N \to \infty$ almost surely. By Lemma A.3, $M_t^N = V^N(X_t) - \int_0^t \mathcal{L}V^N(X_s)ds$ is a Martingale conditioned on $X_0 = X$ for any $X \in \text{supp}(p)$ and, by the definition of T_N, $M_t = M_t^N$ for all $t \leq T_N$. Thus, $(M_t)_t$ is a local Martingale.

□

A.2.3 Faithfulness and deltability

In this section we will begin to look at when KSD-Bs can detect non-convergence. We will show that KSD-B can detect tight non-convergence, that is $\text{KSD-B}_{p,k}(q_n) \not\to 0$ if $q_n \not\to p$ and $(q_n)_n$ is uniformly tight, whenever it is faithful. To show that KSD-B is faithful, we will need an assumption that essentially asks that $\mathcal{H}_k$ is large enough.

We now describe this assumption.

Definition A.5. *If k is a vector field kernel, then we say k is deltable if $\tilde{\delta}_{(X,Y)} \in \mathcal{H}_k$ for all $(X, Y) \in M$, where we define $\delta_{(X,Y)}$ to be the vector field on M that is 1 on (X, Y), -1 on (Y, X) and 0 elsewhere. If k is a kernel on S, then we say k is deltable if $\tilde{\delta}_X \in \mathcal{H}_k$ for all $X \in S$, we define δ_X to be the function that is 1 on X and 0 elsewhere.*

We will describe in section A.4 how to build deltable scalar and vector field kernels. To see that deltability implies that $\mathcal{H}_k$ is large, note that a vector field kernel k is deltable if and only if $C_{C,vf}(M) \subset \mathcal{H}_k$ and a scalar field kernel is deltable if and only if $C_C(S) \subset \mathcal{H}_k$. We can also connect deltability with another notion of the size of $\mathcal{H}_k$: if k is deltable, $\mathcal{H}_k$ is dense in any space for which $C_{C,vf}(M)$ or $C_C(S)$ is dense and is in particular C_0 and L^p - universal Sriperumbudur et al. [2011].

This assumption is also where our study of vector field and scalar field KSD-Bs diverge as if k is a kernel on S, k^∇, definined in Proposition A.2 is not deltable as the next proposition shows.

Proposition A.6. *Say k is a kernel on S. Then k^∇ is not deltable.*

Proof. Let X_1, X_2, X_3 three distinct sequences in S such that $X_1 M X_2 M X_3$. For any $(X, Y) \in M$, calling $f = k^\nabla_{(X,Y)}$

$$\begin{aligned} f(X_1, x_2) + f(X_2, X_3) + f(X_3, X_1) =& (k_y|(k_{X_2} - k_{X_1}) + (k_{X_3} - k_{X_2}) + (k_{X_1} - k_{X_3}))_k \\ & - (k_y|(k_{X_2} - k_{X_1}) + (k_{X_3} - k_{X_2}) + (k_{X_1} - k_{X_3}))_k \\ =& 0. \end{aligned}$$

Thus, for all $f \in \mathcal{H}_{k^\nabla}$, $f(X_1, x_2) + f(X_2, X_3) + f(X_3, X_1) = 0$. However,

$$\delta_{(X_1,X_2)}(X_1, x_2) + \delta_{(X_1,X_2)}(X_2, X_3) + \delta_{(X_1,X_2)}(X_3, X_1) = 1.$$

□

Now we will look at proving the faithfulness and detection of tight non-convergence in proposition 3.1. Our assumption of deltability will allow us to see that if $\text{KSD-B}_{p,k}(q) = 0$ then $E_q f = 0$ for all f in $\mathcal{T}_p(C_{C,vf}(M))$ or $\mathcal{T}_p \nabla(C_C(S))$. The next lemma demonstrates that this implies $q = p$. Considering $f \in \mathcal{T}_p(C_{C,vf}(M))$ the proof will simply follow from the logic of equation 2. However, the same logic cannot be used for $f \in \mathcal{T}_p(\nabla C_C(S))$ as $\mathcal{H}_{k^\nabla}$ cannot be deltable. Instead we will make use of Lemma A.3 (B).

Lemma A.7. *Say p has connected support and q is a distribution on S. If $E_q \mathcal{T}_p f \neq \infty$ for all $f \in C_{C,vf}(M)$ or $E_q \mathcal{T}_p \nabla f \neq \infty$ for all $f \in C_C(S)$ then $\text{supp}(q) \subseteq \text{supp}(p)$. If $E_q \mathcal{T}_p f = 0$ for all $f \in C_{C,vf}(M)$ or $E_q \text{flux}_p < \infty$, $E_p \text{flux}_p < \infty$, and $E_q \mathcal{T}_p \nabla f = 0$ for all $f \in C_C(S)$ then $q = p$.*

Proof. Assume $E_q \mathcal{T}_p f$ is well defined and finite for all $f \in C_{C,vf}(M)$. If $\text{supp}(q) \not\subseteq \text{supp}(p)$ then there is a $X \in \text{supp}(q) \setminus \text{supp}(p)$ such that there is a YMX such that $q(Y) = 0$ or $Y \in \text{supp}(p)$; in either case, by equation 2,

$$E_q \mathcal{T}_p \delta_{(X,Y)} = q(X) T_{p,X \to Y} - q(Y) T_{p,Y \to X} = \infty,$$

since the later term is 0 and, recall, $T_{p,X \to Y}$ is defined to be ∞ when $X \notin \text{supp}(p)$, a contradiction. Thus $\text{supp}(q) \subseteq \text{supp}(p)$. Next assume $E_q \mathcal{T}_p \nabla f \neq \infty$ for all $f \in C_C(S)$. Again pick $X \in \text{supp}(q) \setminus \text{supp}(p)$ such that there is a YMX such that $q(Y) = 0$ or $Y \in \text{supp}(p)$.

$$E_q \mathcal{T}_p \nabla \delta_Y = -q(Y) \text{flux}_p(Y) + \sum_{ZMY} q(Z) T_{p,Y \to Z},$$

and in either case the first term is 0 and the second is ∞, a contradiction.

Now say Say $E_q \mathcal{T}_p f = 0$ for all $f \in C_{C,vf}(M)$. If $X \in \text{supp}(q), Y \in \text{supp}(p), YMX$,

$$0 = E_q \mathcal{T}_p \delta_{(X,Y)} = q(X) T_{p,Y \to X} \left(\frac{p(Y)}{p(X)} - \frac{q(Y)}{q(X)} \right)$$

so we have $q(Y)/q(X) = p(Y)/p(X)$. Thus $\text{supp}(q) = \text{supp}(p)$ and $q(Y)/q(X) = p(Y)/p(X)$ for all $(X, Y) \in M_{p,p}$. Since the support of p in connected this implies that $q = p$.

Now if $E_p \text{flux}_p < \infty$ and $E_q \mathcal{T}_p \nabla f = 0$ for all $f \in C_C(S)$ then $q = p$ by Lemma A.3 (B).

□

Now we use this lemma to show detection of tight non-convergence and faithfulness of the KSD-B, in particular proving Proposition 3.1.

Proposition A.8. *Say supp(p) is connected and $(q_n)_n$ is a tight sequence of distributions on S satisfying* KSD-B$_{p,k}(q) \to 0$. *If k is a deltable vector field kernel or k is a deltable kernel on S, $E_p \text{flux}_p < \infty$, and $\sup_n E_{q_n} \text{flux}_p < \infty$ then $q_n \to p$ in distribution. In particular, if k is a deltable vector field kernel or k is a deltable kernel on S, $E_p \text{flux}_p < \infty$, and $E_q \text{flux}_p < \infty$, then* KSD-B$_{p,k}(q) = 0$ *only if $p = q$.*

Proof. Assume k is a deltable vector field kernel. Say KSD-B$_{p,k}(q) \to 0$ but $(q_n)_n$ does not converge in distribution to p for a sequence of distributions on S $(q_n)_n$. Since $(q_n)_n$ is tight, we can pass to a sub-sequence $(q_{n_k})_k$ that converges in distribution to a distribution q on S. Since for all $f \in C_{C,vf}(M)$, $\mathcal{T}_p f$ is non-zero on only finitely many points, $E_q \mathcal{T}_p f = \lim_k E_{q_{n_k}} \mathcal{T}_p f = 0$ since $f \in \mathcal{H}_k$ by assumption. By lemma A.7, $q = p$, a contradiction.

The situation is similar if k is a deltable kernel on S after using Fatou's lemma to conclude

$$E_q \text{flux}_p \leq \liminf_k E_{q_{n_k}} \text{flux}_p < \infty.$$

□

Note in particular that if $\text{supp}(p)$ is finite since any sequence $(q_n)_n$ with KSD-B$_{p,k}(q) \to 0$ must have $\text{supp}(q_n) \subset \text{supp}(p)$ eventually by Lemma A.7, we automatically have $\sup_n E_{q_n} \text{flux}_p < \infty$ and $(q_n)_n$ uniformly tight. Thus if $\text{supp}(p)$ is finite, the KSD-B can detect non-convergence.

A.2.4 Detection of non-convergence

We will now prove results that describe conditions under which the KSD-B detect convergence and non-convergence. We will see through some counter examples that for arbitrary p, k, the KSD-B actually cannot detect non-convergence. These counter examples will motivate extra assumptions that must be placed on p – namely that it have "uniformly decreasing" tails – and k – namely that it have "thick tails".

In our first counter example, proven in section A.2.5 we will show an example of a distribution p for which the KSD-B does not detect non-convergence. The p in this example has tails that do not "decrease uniformly" with the length of the sequence as there are $(X, Y) \in M$ with $|Y| > |X|$ and $p(Y) \not< p(X)$.

Proposition A.9. *Let $p(X) \propto |\mathcal{B}|^{-L} e^{-\mu L}$ for some $\mu > 0$ if $|X| = L$ or $|X| = L+1$ for even L, and say k is a bounded vector field kernel. Then there is a sequence $(q_n)_n$ such that* $\text{KSD-B}_{p,k}(q_n) \to 0$ *and q_n does not converge to p in distribution.*

Thus we place an assumption on p that will essentially ask that it has tails decrease uniformly with the length of the sequence. First we define the quantities

$$\begin{aligned}
\text{del}_p(X) &= \sum_{|Y|=L-1, XMY} T_{p, X \to Y} \\
\text{ins}_p(X) &= \sum_{|Y|=L+1, XMY} T_{p, X \to Y} \\
\text{gap}_p(L) &= \inf_{X \in S \mid |X|=L} \text{del}_p(X) - \text{ins}_p(X).
\end{aligned}$$

Define $\text{del}_p(L) = \infty$, and $\text{ins}_p(X) = 0$ if $X \notin \text{supp}(p)$. $\text{del}_p(X)$ describes the propensity to gain a deletion, $\text{ins}_p(L)$ the propensity to gain an insertion. We now assume that $\text{gap}_p(L)$ is "big enough" as L grows, We measure how big $\text{gap}_p(L)$ using a "Foster-Lyapunov" function V_p that we will show later controls the convergence of the stochastic process described in section A.2.2.

Assumption A.10. *We assume p has connected support, $E_p \text{flux}_p < \infty$, and there is some concave function $V_p : [0, \infty) \to [0, \infty)$ such that $\lim_{L \to \infty} V_p(L) = \infty$ and*

$$\text{gap}_p(L) \gtrsim \frac{V_p(L)^{\frac{1+\epsilon_V}{2+\epsilon_V}}}{V_p(L) - V_p(L-1)}$$

for some $\epsilon_V > 0$.

We now rewrite the asymptotic inequality in more interpretable forms for several examples of V_p. Note that the since V_p is concave and goes to ∞, the right hand side is eventually less than

$$\frac{V_p(L)^{\frac{1+\epsilon_V}{2+\epsilon_V}}}{V_p'(L)} = \frac{V_p(L)}{V_p'(L)} V_p(L)^{-\frac{1}{2+\epsilon_V}} = \left((\log V_p)'(L) V_p^{\frac{1}{2+\epsilon_V}}(L) \right)^{-1}. \tag{$*$}$$

Note that this quantity is larger when V_p grows more slowly. As well, we will see later that a slower growing V_p results in faster convergence of our process described in section A.2.2. Let's now focus on three example choices of V_p to see how fast gap_p must increase. First consider $V_p = L^\alpha$ for some $0 < \alpha \leq 1$ for which $(*)$ is $L^{1 - \frac{\alpha}{2+\epsilon_V}}$. Thus if $\text{gap}_p(L) \gtrsim L^\beta$ for some $\beta > 1/2$ then assumption A.10 is satisfied for $V_p = L^\alpha$ for some $0 < \alpha \leq 1$. Another option is $V_p(L) = (\log(L))^\beta$ for some $\beta > 0$, in which case $(*)$ is $L \log(L)^{1 - \frac{\beta}{2+\epsilon_V}}$. We can thus pick such a V_p for a $\beta > 2$ for example if $\text{gap}_p(L) \gtrsim L$. Finally, if we define $\log^{(N)}$ as log composed with itself N times, $V_p(L) = (\log^{(N)}(L))^\beta$ corresponds to $\text{gap}_p \gtrsim \left(\prod_{n=0}^{N-1} \log^{(n)}(L) \right) (\log^{(N)}(L))^{1 - \frac{\beta}{2+\epsilon_V}}$. In particular, if $\text{gap}_p(L) \gtrsim L^\alpha$ for some $\alpha > 1$, then we can pick a V_p that grows as slowly as desired. Also note that this is satisfied by $\text{supp}(p)$ being finite. Thus, in general, the faster gap_p increases, the slower we can make V_p increase.

Now we turn to requirements on our kernel k, first considering two motivating counter examples, both proved in section A.2.5. For our first counter example, note that if k is a kernel on S bounded by a number N and $f \in \mathcal{H}_k$ with $\|f\|_k \leq 1$, then for all $X \in S$, $f(X) = (f | k_X)_k \leq \|f\|_k \sqrt{k(X, X)} \leq N$. Thus $\|f\|_\infty \leq N$. Our first counter example now shows that any bounded scalar field kernel cannot result in a KSD-B that detect non-convergence.

Proposition A.11. *There is a distribution p on S such that $\text{gap}_p(L) \sim L$ but $\sup_{\|f\|_\infty \leq 1} E_{q_n} \mathcal{T}_p \nabla f \to 0$ for a sequence of distributions q_n that does not converge in distribution to p.*

Now we turn to vector field kernels. Our next example is an illustration of a case where vector field kernels with thin tails cannot detect non-convergence. The construction is similar in idea to the example of theorem 6 of Gorham and Mackey [2017].

Proposition A.12. *Let $p(X) \propto e^{-\mu|X|}|\mathcal{B}|^{-|X|}$ for some $\mu > 0$ and k be a vector field kernel such that, for $(X, Y), (X', Y') \in M$ with $|X| = |X'|$,*

$$|k((X, Y), (X', Y'))| \leq C(d_H(X, X') + 1)^{-4-\epsilon}$$

for some $C, \epsilon > 0$ where d_H is the hamming distance. Then there is a sequence of distributions $(q_n)_n$ in S such that $\text{KSD-B}_{p,k}(q_n) \to 0$ but q_n doesn't converge to p.

Motivated by these last two examples, we now make an assumption on the kernel k that asks that there exists a function in its RKHS that has thick tails. We will call such a kernel *coercive*. In particular, we ask that there is a $\tilde{f}$ such that $\mathcal{T}_p\tilde{f}$ is increasing sufficiently quickly with respect to the tails of p.

Assumption A.13. *Say p is a distribution on S that satisfies assumption A.10 with V_p. We say a kernel k is coercive if (A) k is a vector field kernel such that there is a $\tilde{f} \in \mathcal{H}_k$ with $\lim_{|X|\to\infty} \mathcal{T}_p\tilde{f}(X) = \infty$ and*

$$\sum_L \frac{\inf_{|X|=L} \mathcal{T}_p\tilde{f}(X)}{\left(\sup_{|X|=L} \text{ins}_p(X)\right) V_p(L+1)} = \infty. \tag{6}$$

(B) k is a kernel on S such that $\text{supp}(p)$ is finite or there is a $\tilde{f} \in \mathcal{H}_k$ with $\lim_{|X|\to\infty} \mathcal{T}_p\nabla\tilde{f}(X) = \infty$ and

$$\sum_L C_L \wedge C_{L+1} = \infty \text{ where } C_L = \frac{\inf_{|X|=L} \mathcal{T}_p\nabla\tilde{f}(X)}{\left(\sup_{|X|=L} \text{flux}_p(X)\right) V_p(L+1)}. \tag{7}$$

To understand this condition intuitively, first consider assumption (A). The denominator in the sum is the maximum propensity for insertions $\left(\sup_{|X|=L} \text{ins}_p(X)\right)$, multiplied by our Foster-Lyapunov function $V_p(L+1)$ so that if p has thinner tails we expect this quantity to be smaller and assumption (A) to be easier to satisfy. In section A.4.3, we construct $\tilde{f}$ such that $\inf_{|X|=L} \mathcal{T}_p\tilde{f}(X) \gtrsim \text{gap}_p(|X|)\tilde{f}(X)$. If we assume $\text{gap}_p(|X|) \gtrsim \text{ins}_p(X)$ then the assumption is satisfied if $\sum_L \frac{\inf_{|X|=L} \tilde{f}(X)}{V_p(L+1)} = \infty$, that is, f itself have thick tails.

Assumption (B) is very similar to (A) with the exceptions of 1) the use of the operator $\mathcal{T}_p\nabla$ instead of $\mathcal{T}_p$, 2) ins_p terms having been replaced by a possibly much larger flux_p term, and 3) the sum being of minima of sequential terms. The last difference (3) simply says that the sequence $C_1, C_2, \ldots$ cannot alternate between large and small.

With these assumptions we can now prove the the KSD-B detects non-convergence. Our approach is inspired by the proof of theorem 8 of Gorham and Mackey [2017]. First we will use our assumption on p to prove the following lemma which is similar to theorem 5 of Gorham et al. [2016].

Lemma A.14. *Say p is a distribution on S obeying assumption A.10. If $g \in C_b(S)$ and $g(X) = 0$ for $X \notin \text{supp}(p)$, then there is a $f_g : S \to \mathbb{R}$ such that $f_g(X) = 0$ for $X \notin \text{supp}(p)$, $\mathcal{T}_p\nabla f_g = g - E_p g$, and $f_g(X) \leq CV_p(X)\|g\|_\infty$ for a universal constant C.*

Proof. $\mathcal{L}_p = \mathcal{T}_p\nabla$ is the infinitesimal generator for a semi-group $(P_t)_t$. Also define $\Delta V_{p,L} = V_p(L) - V_p(L-1)$ and $V_p(X)$ as $V_p(|X|)$. If $X \in \text{supp}(p)$ with $|X| = L$,

$$\begin{aligned}
\mathcal{L}V_p(X) =& \sum_{YMX, |Y|=|X|+1} T_{p,X\to Y}\Delta V_{p,L+1} - \sum_{YMX,|Y|=|X|-1} T_{p,X\to Y}\Delta V_{p,L} \\
=& \text{ins}_p(X)\Delta V_{p,L+1} - \text{del}_p(X)\Delta V_{p,L} \\
\leq& \text{ins}_p(X)\left(\Delta V_{p,L+1} - \Delta V_{p,L}\right) - \text{gap}_p(L)\Delta V_{p,L}
\end{aligned}$$

Since V_p in concave, the first term is negative. As well, by assumption, $\text{gap}_p(L)\Delta V_{p,L} \gtrsim \varphi(V_p(L-1))$ where $\varphi(x) = x^{(1+\epsilon)/(2+\epsilon)}$. Thus there are constants C_1, C_2 such that for all $X \in \text{supp}(p)$,

$$\mathcal{L}V_p(X) \leq C_1 - C_2\varphi \circ V_p(X).$$

By theorem A.4, with $H = \int_1^u ds\varphi^{-1}(s) = C_3(u^{\frac{1}{2+\epsilon}} - 1)$, we have

$$\|P_t(X) - p\|_{\text{TV}} \lesssim V_p(X)t^{-(2+\epsilon)} + t^{-(1+\epsilon)}.$$

Now assume $g \in C_b(S)$. We have that

$$|P_t g(X) - E_p g| \leq \|g\|_\infty \|P_t(X) - p\|_{\mathrm{TV}}$$

so $\int_0^\infty dt |P_t g(X) - E_p g| \leq C' \|g\|_\infty V_p(X)$ for some $C' > 0$ for large enough X. Thus we can define

$$f_g(X) = \int_0^\infty dt \, (E_p g - P_t g(X))$$

with $|f_g|(X) \leq C' \|g\|_\infty V_p(X)$. Because we have absolute integrability, and by Lemma A.3 (A) we can also write

$$\mathcal{L} f_g(X) = \int_0^\infty dt \, (-\mathcal{L} P_t g(X)) = \int_0^\infty dt \left(-\frac{d}{dt} P_t g(X) \right) = g(X) - E_p g.$$

□

Finally we show that the KSD-B can detect non-convergence.

Theorem A.15. *Say p is a distribution on S obeying assumption A.10 and k is a deltable vector field kernel obeying assumption A.13 (A) or k a deltable kernel on S obeying assumption A.13 (B). Say $(q_n)_n$ is a sequence of distributions on S. If* $\mathrm{KSD\text{-}B}_{p,k}(q_n) \to 0$ *then q_n converges to p in distribution.*

Proof. First note that by Lemma A.7, $\mathrm{supp}(q_n) \subset \mathrm{supp}(p)$ for all n eventually. Let $g \in C_b(S)$ with $g(X) = 0$ for $X \notin \mathrm{supp}(p)$ and $\|g\|_\infty \leq 1$, so by lemma A.14, there is an $f_g : S \to \mathbb{R}$ such that $f_g \leq \tilde{C} V_p$ for some $\tilde{C} > 0$ and $\mathcal{T}_p \nabla f_g = g - E_p g$. We will show that $E_{q_n} g - E_p g = E_{q_n} \mathcal{T}_p \nabla f_g \to 0$, which will be enough to prove the theorem. We will do so by picking a sequence of $h_m \in \mathcal{H}_k$ such that $\sup_n E_{q_n} |\mathcal{T}_p h_m - \mathcal{T}_p \nabla f_g| \to 0$ as $m \to \infty$. This will show that

$$|E_{q_n} \mathcal{T}_p \nabla f_g| \leq |E_{q_n} \mathcal{T}_p h_m| + E_{q_n} |\mathcal{T}_p h_m - \mathcal{T}_p \nabla f_g| \leq \|h_m\|_k \mathrm{KSD\text{-}B}_{p,k}(q_n) + E_{q_n} |\mathcal{T}_p h_m - \mathcal{T}_p \nabla f_g|$$

which goes to zero as $n \to \infty$ and $m \to \infty$ slow enough.

First assume k is a deltable kernel obeying assumption A.13 (A). Let $\tilde{f} \in \mathcal{H}_k$ satisfy equation 6 and have $\mathcal{T}_p \tilde{f}(X) \to \infty$ as $|X| \to \infty$. There is thus a $\zeta \in \mathbb{R}$ such that $\mathcal{T}_p \tilde{f}(X) + \zeta > 0$ for all $X \in S$. For a sequence $v = (v_1, v_2, \dots)$ of numbers $0 \leq v_n \leq 1$ such that v_n is eventually equal to 0, define the vector field on M $h_v(X, Y) = v_{|X| \wedge |Y|} \nabla f_g(X, Y)$. Since v is eventually 0, by the deltability of k, $h_v \in \mathcal{H}_k$. Then

$$\mathcal{T}_p h_v(X) = v_{|X|} \mathcal{T}_p \nabla f_g(X) + (v_{|X|+1} - v_{|X|}) \sum_{YMX, |Y| = |X|+1} T_{p, X \to Y} \nabla f_g(X, Y).$$

The first term is a better and better approximation of $\mathcal{T}_p \nabla f_g$ as $v \to 1$. We now use our assumption A.13 (A) to bound the second term by $E_{q_n} \mathcal{T}_p \tilde{f} \leq \|\tilde{f}\|_k \mathrm{KSD\text{-}B}_{p,k}(q_n)$. Note

$$\left| \sum_{YMX, |Y| = |X|+1} T_{p, X \to Y} \nabla f_g(X, Y) \right| \leq 2 \tilde{C} V_p(|X| + 1) \mathrm{ins}_p(X).$$

Now call $\Delta v_L = |v_{L+1} - v_L|$ and $R_L := \frac{V_p(L+1)\sup_{|X|=L}\mathrm{ins}_p(X)}{\inf_{|X|=L}\mathcal{T}_p\tilde{f}(X)+\zeta}$, so,

$$\begin{aligned}
E_{q_n}|\mathcal{T}_p h_v - \mathcal{T}_p\nabla f_g| \leq& E_{q_n}[(1-v_{|X|})|\mathcal{T}_p\nabla f_g|] + E_{q_n}\left[\Delta v_{|X|} 2\tilde{C}V_p(|X|+1)\mathrm{ins}_p(X)\right] \\
\leq& 2\|g\|_\infty E_{q_n}[1-v_{|X|}] + 2\tilde{C}E_{q_n}\left[\left(\mathcal{T}_p\tilde{f}+\zeta\right)\Delta v_{|X|}\frac{V_p(|X|+1)\mathrm{ins}_p(X)}{\mathcal{T}_p\tilde{f}+\zeta}\right] \\
\leq& 2E_{q_n}[1-v_{|X|}] + 2\tilde{C}E_{q_n}\left[\mathcal{T}_p\tilde{f}+\zeta\right]\sup_L\left(\Delta v_L R_L\right) \\
\leq& 2E_{q_n}\left[\mathcal{T}_p\tilde{f}+\zeta\right]\sup_L\frac{1-v_{|X|}}{\mathcal{T}_p\tilde{f}+\zeta} + 2\tilde{C}E_{q_n}\left[\mathcal{T}_p\tilde{f}+\zeta\right]\sup_L\left(\Delta v_L R_L\right) \\
=& E_{q_n}\left[\mathcal{T}_p\tilde{f}+\zeta\right]\left(2\sup_L\frac{1-v_{|X|}}{\mathcal{T}_p\tilde{f}+\zeta} + 2\tilde{C}\sup_L\left(\Delta v_L R_L\right)\right) \\
\leq& \left(\|\tilde{f}\|_k\text{KSD-B}_{p,k}(q_n)+\zeta\right)\left(2\sup_L\frac{1-v_{|X|}}{\mathcal{T}_p\tilde{f}+\zeta} + 2\tilde{C}\sup_L\left(\Delta v_L R_L\right)\right) \\
\lesssim& \sup_L\frac{1-v_{|X|}}{\mathcal{T}_p\tilde{f}+\zeta} + \sup_L\left(\Delta v_L R_L\right).
\end{aligned}$$

By assumption $\mathcal{T}_p\tilde{f}+\zeta \to \infty$ and $\sum_L R_L^{-1} = \infty$. For $\epsilon, L' > 0$ define $v_L^{\epsilon,L'} = 1$ for $L \leq L'$ and $\Delta v_L = \epsilon R_L^{-1} \wedge (v_L)$ for $l \geq L$. By assumption $\sum_L R_L^{-1} = \infty$ so $v^{\epsilon,L'}$ is eventually 0. We thus have $\sup_L\left(\Delta v_L^{\epsilon,L'} R_L\right) = \epsilon$ and $\sup_L \frac{1-v_{|X|}^{\epsilon,L'}}{\mathcal{T}_p\tilde{f}+\zeta} \leq \frac{1}{\inf_{|X|\geq L}\mathcal{T}_p\tilde{f}+\zeta}$. By our assumption that $\mathcal{T}_p\tilde{f} \to \infty$, both of these quantities go to 0 as $L' \to \infty$ and $\epsilon \to 0$.

Now assume k is a deltable kernel obeying assumption A.13 (B). The case that $\mathrm{supp}(p)$ is finite was shown in Proposition 3.1 so assume $\mathrm{supp}(p)$ is infinite. The proof is very similar. This time, for a sequence $v = (v_1, v_2, \dots)$ of decreasing numbers $0 \leq v_n \leq 1$ such that v_n is eventually equal to 0, define the function on S, $h_v(X) = v_{|X|}f_g(X)$. Since v is eventually 0, by the deltability of k, $h_v \in \mathcal{H}_k$. Then, by similar reasoning to the previous case,

$$\begin{aligned}
\mathcal{T}_p\nabla h_v(X) =& v_{|X|}\mathcal{T}_p\nabla f_g(X) + (v_{|X|+1} - v_{|X|})\sum_{YMX,|Y|=|X|+1} T_{p,X\to Y}\nabla f_g(X,Y) \\
&+ (v_{|X|-1} - v_{|X|})\sum_{YMX,|Y|=|X|-1} T_{p,X\to Y}\nabla f_g(X,Y).
\end{aligned}$$

Note that since V_p is increasing, the sum of the later two terms is upper bounded by

$$2\tilde{C}\tilde{\Delta}v_L V_p(|X|+1)\mathrm{flux}_p(X)$$

defining $\tilde{\Delta}v_L = |v_{L+1} - v_L| \vee |v_L - v_{L-1}|$. Now call $\tilde{R}_L := \frac{V_p(L+1)\sup_{|X|=L}\mathrm{flux}_p(X)}{\inf_{|X|=L}\mathcal{T}_p\nabla\tilde{f}(X)+\zeta}$, so,

$$\begin{aligned}
E_{q_n}|\mathcal{T}_p\nabla h_v - \mathcal{T}_p\nabla f| \leq& E_{q_n}[(1-v_{|X|})|\mathcal{T}_p\nabla f_g|] + E_{q_n}\left[\tilde{\Delta}v_{|X|} 2\tilde{C}V_p(|X|+1)\mathrm{flux}_p(X)\right] \\
\leq& 2E_{q_n}\left[\mathcal{T}_p\nabla\tilde{f}+\zeta\right]\left(\sup_L\frac{1-v_{|X|}}{\mathcal{T}_p\nabla\tilde{f}+\zeta} + 2\tilde{C}\sup_L\left(\tilde{\Delta}v_L\tilde{R}_L\right)\right) \\
\leq& \left(\|\tilde{f}\|_k\text{KSD-B}_{p,k}(q_n)+\zeta\right)\left(2\sup_L\frac{1-v_{|X|}}{\mathcal{T}_p\nabla\tilde{f}+\zeta} + 2\tilde{C}\sup_L\left(\tilde{\Delta}v_L\tilde{R}_L\right)\right) \\
\lesssim& \sup_L\frac{1-v_{|X|}}{\mathcal{T}_p\nabla\tilde{f}+\zeta} + \sup_L\left(\tilde{\Delta}v_L\tilde{R}_L\right).
\end{aligned}$$

By assumption $\mathcal{T}_p\tilde{f}+\zeta \to \infty$ and $\sum_L \tilde{R}_L^{-1} \wedge \tilde{R}_{L+1}^{-1} = \infty$. For $\epsilon, L' > 0$ define $v_L^{\epsilon,L'} = 1$ for $L \leq L'$ and $v_L = v_{L-1} - \epsilon\tilde{R}_{L-1}^{-1} \wedge \tilde{R}_L^{-1} \wedge (v_{L-1})$ for $l \geq L$. Thus $\tilde{\Delta}v_L \leq \epsilon\tilde{R}_L^{-1}$. By assumption $\sum_L \tilde{R}_L^{-1} \wedge \tilde{R}_{L+1}^{-1} = \infty$ so $v^{\epsilon,L'}$ is eventually 0. We thus have $\sup_L\left(\tilde{\Delta}v_L^{\epsilon,L'}\tilde{R}_L\right) = \epsilon$

and $\sup_L \frac{1-v^{\epsilon,L'}_{|X|}}{\mathcal{T}_p\tilde{f}+\zeta} \leq \frac{1}{\inf_{|X|\geq L}\mathcal{T}_p\tilde{f}+\zeta}$. By our assumption that $\mathcal{T}_p\tilde{f} \to \infty$, both of these quantities go to 0 as $L' \to \infty$ and $\epsilon \to 0$. □

Finally we prove that the KSD-B can detect convergence as in proposition 3.3.

Proposition A.16. *Say k is a vector field kernel and $p, q_1, q_2, \dots$ are p, k-integrable distributions on S. Call $A(X) = \sum_{YMX} T_{p,Y\to X}\sqrt{k((X,Y),(X,Y))}$.*

$$\sum_X |p(X) - q_n(X)|A(X) \to 0 \implies \text{KSD-B}_{p,k}(q_n) \to 0.$$

Proof. Say $f \in \mathcal{H}_k$.

$$|E_p\mathcal{T}_pf - E_q\mathcal{T}_pf| \leq \|f\|_k \sum_X |p(X) - q_n(X)| \sum_{YMX} T_{p,Y\to X}\sqrt{k((X,Y),(X,Y))}$$

which proves the result. □

A.2.5 Proofs of examples

Proposition A.17. *(Proposition A.9) Let $p(X) \propto |\mathcal{B}|^{-L}e^{-\mu L}$ for some $\mu > 0$ if $|X| = L$ or $|X| = L+1$ for even L, and say k is a bounded vector field kernel. Then there is a sequence $(q_n)_n$ such that $\text{KSD-B}_{p,k}(q_n) \to 0$ and q_n does not converge to p in distribution.*

Proof. Define, for even L, $\tilde{q}_L = p\mathbb{1}_{|X|\leq L}$ and $q_L = \tilde{q}_L / \sum_X \tilde{q}_L(X)$. Call $q_L(L') = q_L(X)$ for any $|X| = L'$. Call $N_L = \{(X,Y) \in M \mid |X| = L, |Y| = L+1\}$. The terms of the sum in Equation 2 are non-zero only for $(X,Y) \in N_L$. Thus,

$$\begin{aligned}
\text{KSD-B}_{p,k}(q_n)^2 &= \left(\sup_{\|f\|_k\leq 1} \sum_{(X,Y)\in N_L} q_L(X)T_{p,X\to Y}f(X,Y)\right)^2 \\
&= q_L(L)^2\left(\sup_{\|f\|_k\leq 1}\left(f\middle|\sum_{(X,Y)\in N_L} T_{p,X\to Y}k_{(X,Y)}\right)_k\right)^2 \\
&= q_L(L)^2\left\|\sum_{(X,Y)\in N_L} T_{p,X\to Y}k_{(X,Y)}\right\|_k^2 \\
&= q_L(L)^2 \sum_{(X,Y)\in N_L}\sum_{(X',Y')\in N_L} T_{p,X\to Y}T_{p,X'\to Y'}k((X,Y),(X',Y')).
\end{aligned}$$

If $(X,Y) \in N_L$, then $T_{p,X\to Y} \leq L+1$. Thus, if k is bounded by a number $C > 0$,

$$\text{KSD-B}_{p,k}(q_n)^2 \leq q_L(L)^2(L+1)^2|\mathcal{B}|^{2L}C = \left(\frac{q_L(L)}{e^{-\mu L}|\mathcal{B}|^{-L}}\right)^2 e^{-2\mu L}(L+1)^2C \to 0$$

as $\frac{q_L(L)}{e^{-\mu L}|\mathcal{B}|^{-L}} = \left(\sum_{|X|\leq L}\tilde{p}(X)\right)^{-1} \to 1$. □

Proposition A.18. *(Proposition A.11) There is a distribution p on S such that $\text{gap}_p(L) \sim L$ but $\sup_{\|f\|_\infty\leq 1} E_{q_n}\mathcal{T}_p\nabla f \to 0$ for a sequence of distributions q_n that does not converge in distribution to p.*

Proof. Let p be the distribution supported on $\{\emptyset, A, AA, AAA, \dots\}$ for $A \in \mathcal{B}$ with $p(L) = p(L \times A) = 2^{-(L+1)}$ for any number L. Define $r = \left(\frac{p(L)}{p(L-1)}\right) = (1/2)$ for any L and $\tilde{r} = \left(\frac{p(L-1)}{p(L)}\right) = (2)$

for any L (with $\tilde{r}_0 = 0$). Thus, $r < 1 \leq \tilde{r}$ for all L. Say q is a distribution supported on finitely many $\{\emptyset, A, AA, AAA, \dots\}$, and f is a function on S with $\|f\|_\infty \leq 1$,

$$\begin{aligned}
E_q \mathcal{T}_p \nabla f &= \sum_{L=0}^{\infty} q(L) \left((L+1) r (f(L+1) - f(L)) + L\tilde{r}(f(L-1) - f(L)) \right) \\
&= \sum_{L=0}^{\infty} f(L) \Big(q(L+1)(L+1)\tilde{r} + q(L-1)Lr \\
&\qquad - q(L) \left(L\tilde{r}_L + (L+1)r \right) \Big) \\
&= \sum_{L=0}^{\infty} f(L) \Big(q(L+1)(L+1)\tilde{r} - q(L)L\tilde{r} \\
&\qquad + q(L-1)Lr - q(L)(L+1)r \Big).
\end{aligned}$$

Let $\tilde{q}_{m,n}(L) = L^{-1}$ for $m \leq L \leq n$ and $\tilde{q}_{m,n}(L) = 0$ for $L > n$ and $L < m$. Now let $q_{m,n} = \tilde{q}_{m,n}/Z_{m,n}$ where $Z_{m,n} = \sum_{L=m}^{n} L^{-1}$ which goes to ∞ as $n \to \infty$. Thus,

$$\begin{aligned}
E_{q_{m,n}} \mathcal{T}_p \nabla f =& f(m-1) Z_{m,n}^{-1} \tilde{r} - f(n) Z_{m,n}^{-1} \tilde{r} \\
&+ \sum_{L=m+1}^{n} f(L) \left(q_{m,n}(L-1)Lr - q_{m,n}(L)(L+1)r \right) \\
&- f(m) q_{m,n}(m+1) r + f(n+1) q_{m,n}(n)(n+1) r \\
=& {Z_{m,n}}^{-1} \left(\tilde{r} f(m-1) - \tilde{r} f(n) - f(m) \frac{m+1}{m} + f(n+1) \frac{n+1}{n} \right) \\
&+ \sum_{L=m+1}^{n} f(L) r \left(1 - \frac{(L+1)(L-1)}{L^2} \right) \\
\leq& 6\tilde{r} {Z_{m,n}}^{-1} + \sup_{L > m} Lr \left| 1 - \left(1 - \frac{1}{L^2} \right) \right| \\
=& 6\tilde{r} {Z_{m,n}}^{-1} + r/m.
\end{aligned} \tag{8}$$

This expression goes to 0 as $n, m \to \infty$.

□

Proposition A.19. *(Proposition A.12) Let $p(X) \propto e^{-\mu|X|} |\mathcal{B}|^{-|X|}$ for some $\mu > 0$ and k be a vector field kernel such that, for $(X, Y), (X', Y') \in M$ with $|X| = |X'|$,*

$$|k((X,Y),(X',Y'))| \leq C(d_H(X,X') + 1)^{-4-\epsilon}$$

for some $C, \epsilon > 0$ where d_H is the hamming distance. Then there is a sequence of distributions $(q_n)_n$ in S such that KSD-$\mathrm{B}_{p,k}(q_n) \to 0$ *but q_n doesn't converge to p.*

Proof. First note that for $(X, Y) \in M$, calling $(e^{\mu}|\mathcal{B}|) = c$, $T_{p, X \to Y} \leq c(|X| + 1)$. For distinct points $X_1, \dots, X_N \in \mathcal{B}^L$ let $q = \frac{1}{N} \sum_{n=1}^{N} \delta_{X_n}$. Call $R = \min_{n \neq m} d_H(X_n, X_m) > 0$ and say k is

bounded by a number C. Then by equation 1,

$$\begin{aligned}
\text{KSD-B}_{p,k}(q)^2 \leq & \frac{c^2(L+1)^2}{N^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \sum_{Y \in MX_n} \sum_{Y' \in MX_m} |k((X_n, Y), (X_m, Y'))| \\
= & \frac{c^2(L+1)^2}{N^2} \Bigg(\sum_{n=1}^{N} \sum_{Y \in MX_n} \sum_{Y' \in MX_n} |k((X_n, Y), (X_n, Y'))| \\
& \qquad + \sum_{n \neq m} \sum_{Y \in MX_n} \sum_{Y' \in MX_m} |k((X_n, Y), (X_m, Y'))| \Bigg) \\
\lesssim & \frac{(L+1)^2}{N^2} \left(NL^2 + N^2 L^2 R^{-(4+\epsilon)} \right) \\
= & O\left(L^4 \left(N^{-1} + R^{-(4+\epsilon)} \right) \right).
\end{aligned}$$

We now pick, for each L, $X_1, \ldots, X_{N_L} \in \mathcal{B}^L$ to be the largest set of sequence such that $\min_{n \neq m} d_H(X_n, X_m) > R_L = \frac{L}{20|\mathcal{B}|}$. We will show $L^4 N_L^{-1} \to 0$, so that we will have $L^4 \left(N_L^{-1} + R_L^{-(4+\epsilon)} \right) \to 0$ and the proof will be complete. For $X \in \mathcal{B}^L$, $r > 0$, define the Hamming ball $B(X, r) = \{Y \in \mathcal{B}^L \mid d_H(X, Y) \leq r\}$. Thus $\mathcal{B}^L \subset \cup_n B(X_n, R_L)$, otherwise we could add another sequence to $(X_n)_n$. Thus $|\mathcal{B}|^L \leq \sum_n |B(X_n, R_L)| = N_L |B(X_1, R_L)|$. Let Z be a Binomial random variable with parameters L and $|\mathcal{B}|^{-1}$. Then $|B(X_1, R_L)|/|\mathcal{B}|^L = P(Z \leq R_L)$. On the other hand, calling $t = -\log\left(\frac{R_L |\mathcal{B}|}{L}\right) = \log 20$,

$$\begin{aligned}
P(Z \leq R_L) = & P(e^{-tZ} \geq e^{-tR_L}) \\
\leq & e^{tR_L} E e^{-tZ} \\
= & e^{tR_L} \left(|\mathcal{B}|^{-1} e^{-t} + (1 - |\mathcal{B}|^{-1}) \right)^L \\
= & e^{tR_L} \left(1 + |\mathcal{B}|^{-1} (e^{-t} - 1) \right)^L \\
\leq & \exp\left(tR_L + L|\mathcal{B}|^{-1} (e^{-t} - 1) \right) \\
= & \exp\left(R_L (1 + t) - L|\mathcal{B}|^{-1} \right) \\
= & \exp\left(-L|\mathcal{B}|^{-1} \left(1 - \frac{1}{20} (1 + \log 20) \right) \right) \\
\leq & \exp\left(-\frac{1}{2} L|\mathcal{B}| \right).
\end{aligned}$$

Thus, $N_L \geq e^{\frac{1}{2} L|\mathcal{B}|}$ so that $L^4 N_L^{-1} \to 0$ as $L \to \infty$. □

A.2.6 Efficient approximate kernelized Stein discrepancies

We have shown that the KSD-B is computable by equation 1. However this expression may be expensive to evaluate; namely, the terms inside the expectation,

$$\sum_{X \in MY} \sum_{X' \in MY'} T_{p, X \to Y} T_{p, X' \to Y'} k((X, Y), (X', Y')) \tag{†}$$

cost $O(|X||X'|)$ to evaluate as each sequence X has up to $|\mathcal{B}|(|X|+1)+(|\mathcal{B}|-1)|X|+|X|$ neighbours in M. Instead we can approximate the sum as follows. As in section A.2.2, we define the Markov matrix $K_{X \to Y} = T_{p, X \to Y} / \text{flux}_p(X)$ if $X \neq Y$ and 0 if $X = Y$ and let $(Z_0, Z_1, \ldots), (Z_0', Z_1', \ldots)$ be two independent Markov processes following K. Then we can rewrite (†) as

$$\begin{aligned}
& \text{flux}_p(X) \text{flux}_p(X') \sum_{X \in MY} \sum_{X' \in MY'} K_{X \to Y} K_{X' \to Y'} k((X, Y), (X', Y')) \\
& \qquad = \text{flux}_p(X) \text{flux}_p(X') E_{Z_0 = X} E_{Z_0' = X'} k((X, Z_1), (X', Z_1')).
\end{aligned} \tag{9}$$

Now we can approximate the expectations in a variety of ways. If $T_{p,X\to Y}$ are cheap to evaluate, we may sampling $Y_{X,m} \sim Z_1|Z_0 = X$ for $m = 1, \ldots, M_X$ and use the approximation

$$\text{flux}_p(X)\text{flux}_p(X')\frac{1}{M_X M_{X'}}\sum_{m,m'} k((X, Y_{X,m}), (X', Y_{X',m'}))$$

which only involves M, M' kernel evaluations, potentially much smaller than $O(|X||X'|)$. If $T_{p,X\to Y}$ are expensive, we may approximate the expectation by sampling from an approximate distribution, say using Gibbs with Gradients Grathwohl et al. [2021] or uniform sampling; however we will not explore this later approach further.

Now we ask how large M_X must be to achieve a good approximation. We consider two scenarios. In the first we consider arbitrary q and show that if M_X is $O(|X|)$, although potentially much smaller than the number of neighbours of $|X|$, we achieve a bound on the error of the approximation. In the second, we consider q to be a set of point masses $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. In this case, we show that M_X can be $O\left(\log\left(\max_i |X_i|\right)\log(n)\right)$, likely substantially lower than $O(|X|)$. The proofs of these results in either scenario are roughly based on the use of a sub-Gaussian concentration inequality as used in theorem 4 of Gorham et al. [2020].

Now we consider the first of these scenarios.

Proposition A.20. *Let p be a distribution on S and k is a bounded vector field kernel. For each X, n, let $(Y_{X,n,m})_{m=1}^{M_{X,n}}$ be iid samples distributed as $Z_1|Z_0 = X$. Define the approximate KSD*

$$\text{KSD-B}\hat{}_{p,k}^{n}(q)^2 = E_{X,X'\sim q}\text{flux}_p(X)\text{flux}_p(X')\frac{1}{M_{X,n}M_{X',n}}\sum_{m,m'} k((X, Y_{X,n,m}), (X', Y_{X',n,m'})).$$

Say $(q_n)_n$ is a sequence of distributions on S with $\sup_n E_{q_n}\text{flux}_p < \infty$.

$$\textit{If } M_{X,n} \geq C|X|n(\log n)^2 \textit{ for some } C > 0 \textit{ then } \left|\text{KSD-B}_{p,k}(q_n) - \text{KSD-B}\hat{}_{p,k}^{n}(q_n)\right| \to 0.$$

Proof. Let $M_{X,n}$ be a family of numbers such that $M_{X,n} \geq C|X|n(\log n)^2$ for some $C > 0$. Call $p(Y|X) = K_{X\to Y}$. Sample $(Y_{X,n,m})_{m=1}^{M_{X,n}}$ iid from $p(Y|X)$ for all $X \in S$. Call $\hat{p}_n(Y|X) = \frac{1}{M_{X,n}}\sum_{m=1}^{M_{X,n}} \delta_{Y_{X,n,m}}$. Since k is bounded, say by some number M,

$$E_{X\sim q}\text{flux}_p(X)E_{\hat{p}_n(Y|X)}\sqrt{k((X,Y),(X,Y))} \leq ME_{X\sim q}\text{flux}_p(X) < \infty.$$

Then the functional $\phi_n : \mathcal{H}_k \to \mathbb{R} \mid f \mapsto E_{X\sim q}\text{flux}_p(X)E_{\hat{p}_n(Y|X)}f(X,Y)$ is bounded, and is thus in $\mathcal{H}_k$. As in the proof of proposition 2.1,

$$\text{KSD-B}\hat{}_{p,k}^{n}(q) = \|\phi_n\|_k = \sup_{f\in\mathcal{F}} E_{X\sim q}T_p(X)E_{\hat{p}_n(Y|X)}f(X,Y).$$

We will show that $\sup_X \|p(Y|X) - \hat{p}_n(Y|X)\|_{\text{TV}} \to 0$ as $n \to \infty$ almost surely later; for now, assume this is the case. Thus, since $\|f\|_k \leq 1$ implies that $\|f\|_\infty \leq M$,

$$\begin{aligned}
&\left|\text{KSD-B}_{p,k}(q_n) - \text{KSD-B}\hat{}_{p,k}^{n}(q_n)\right| \\
&\qquad \leq \sup_{\|f\|_k\leq 1} E_{X\sim q_n}\text{flux}_p(X)\left|E_{p(Y|X)}f(X,Y) - E_{\hat{p}_n(Y|X)}f(X,Y)\right| \\
&\qquad \leq ME_{X\sim q_n}\text{flux}_p(X)\left\|p(Y|X) - \hat{p}_n(Y|X)\right\|_{\text{TV}} \\
&\qquad \leq M\left(E_{X\sim q_n}\text{flux}_p(X)\right)\sup_X \left\|p(Y|X) - \hat{p}_n(Y|X)\right\|_{\text{TV}} \\
&\qquad \to 0.
\end{aligned}$$

Now we will show that $\sup_X \|p(Y|X) - \hat{p}_n(Y|X)\|_{\text{TV}} \to 0$ almost surely. Pick a sequence of positive numbers $\epsilon_1, \epsilon_2, \ldots$. Let $X \in \text{supp}(p)$ with $|X| = L$ and call $N(X) = \{Y \in S \mid YMX\}$. Call $\mathcal{F}_X = \{f : N(X) \to \{-1, 1\}\}$ so

$$\|p(Y|X) - \hat{p}_n(Y|X)\|_{\text{TV}} = \max_{f\in\mathcal{F}_X} E_{p(Y|X)}f(Y) - E_{\hat{p}_n(Y|X)}f(Y).$$

Note for each $f \in \mathcal{F}_X$, $E_{\hat{p}_n(Y|X)} f(Y)$ is an average of $M_{X,n}$ iid random variables that take values $\{-1, 1\}$ and is thus a sub-Gaussian random variable with variance-proxy $C'/\sqrt{M_{X,n}}$ for some C'. Then by a maximal concentration inequality (theorem 1.14 of Rigollet and Hütter [2019]), since $|\mathcal{F}| = 2^{CL}$ for some $C > 0$,

$$P(\|p(Y|X) - \hat{p}_n(Y|X)\|_{\mathrm{TV}} > \epsilon_n) \leq C_1 L \exp\left(-C_2 M_{X,n} \epsilon_n^2\right)$$

for some constants $C_1, C_2 > 0$. Thus for some constant $C_3 > 0$, if ϵ_n decreases slowly enough, for large enough n,

$$\begin{aligned}
&\sum_{X,n} P(\|p(Y|X) - \hat{p}_n(Y|X)\|_{\mathrm{TV}} > \epsilon_n) \\
&\qquad \leq C_1 \sum_{L,n} L|\mathcal{B}|^L \exp\left(-C_2 C L n (\log n)^2 \epsilon_n^2\right) \\
&\qquad \lesssim \sum_{L,n} \exp\left(-C_3 L n (\log n)^2 \epsilon_n^2\right) \\
&\qquad \lesssim \sum_{n} \int_0^\infty dL \exp\left(-C_3 L n (\log n)^2 \epsilon_n^2\right) \\
&\qquad \lesssim \sum_{n} \frac{1}{n(\log n)^2 \epsilon_n^2} < \infty.
\end{aligned}$$

By the Borel-Cantelli lemma, the probability that $\|p(Y|X) - \hat{p}_n(Y|X)\|_{\mathrm{TV}} > \epsilon_n$ for infinitely many L, n is 0. Thus, with probability 1, as $n \to \infty$,

$$\sup_X \|p(Y|X) - \hat{p}_n(Y|X)\|_{\mathrm{TV}} \to 0.$$

□

We finish the section by considering the second situation where q is a sum of point masses.

Proposition A.21. *Let p be a distribution on S, $(q_n)_n$ a sequence of distributions on S with $q_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i^n}$ for $X_i^n \in S$ and k is a bounded vector field kernel. For each X, n, let $(Y_{i,n,m})_{m=1}^{M_{i,n}}$ be iid samples distributed as $Z_1|Z_0 = X$. Define the approximate KSD*

$$\text{KSD-B}\hat{}_{p,k}^n(q)^2 = \frac{1}{n^2}\sum_{i,j} \mathrm{flux}_p(X_i^n)\mathrm{flux}_p(X_j^n)\frac{1}{M_{i,n}M_{j,n}}\sum_{m,m'} k((X_i^n, Y_{i,n,m}), (X_j^n, Y_{j,n,m'})).$$

Defining $L_n = \max_i |X_i^n|$, for some large enough C,

$$\textit{if } M_{i,n} \geq C\log(L_n)(\log n) \textit{ then } \left|\text{KSD-B}_{p,k}(q_n) - \text{KSD-B}\hat{}_{p,k}^n(q_n)\right| \to 0.$$

Proof. Let $M_{i,n}$ be a family of number such that $M_{i,n} \geq C\log(L_n)(\log n)$. Again, call $p(Y|X) = K_{X\to Y}$, sample $(Y_{i,n,m})_{m=1}^{M_{i,n}}$ iid from $p(Y|X_i^n)$, and call $\hat{p}_n(Y|X_i^n) = \frac{1}{M_{i,n}}\sum_{m=1}^{M_{i,n}} \delta_{Y_{i,n,m}}$. As in the proof of Proposition A.20,

$$\begin{aligned}
&\left|\text{KSD-B}_{p,k}(q_n) - \text{KSD-B}\hat{}_{p,k}(q_n)\right| \\
&\qquad \leq \sup_{\|f\|_k \leq 1} E_{X\sim q_n}\mathrm{flux}_p(X)\left|E_{p(Y|X)}f(X,Y) - E_{\hat{p}_n(Y|X)}f(X,Y)\right| \\
&\qquad \leq E_{X\sim q_n}\mathrm{flux}_p(X)\left\|p(Y|X) - \hat{p}_n(Y|X)\right\|_{\mathrm{TV}}.
\end{aligned}$$

We will again show that $\sup_i \|p(Y|X_i^n) - \hat{p}_n(Y|X_i^n)\|_{\mathrm{TV}} \to 0$ as $n \to \infty$ almost surely, which will complete the proof.

Pick a sequence of positive numbers $\epsilon_1, \epsilon_2, \ldots$. Then again by a maximal concentration inequality (theorem 1.14 of Rigollet and Hütter [2019]),

$$P(\|p(Y|X_i^n) - \hat{p}_n(Y|X_i^n)\|_{\mathrm{TV}} > \epsilon_n) \leq C_1|X_i^n|\exp\left(-C_2 M_{i,n}\epsilon_n^2\right)$$

for some constants $C_1, C_2 > 0$. Thus, if C is large enough, for some constant $C_3 > 1$, if ϵ_n decreases slowly enough,

$$\begin{aligned}\sum_{i,n} P(\|p(Y|X_i^n) - \hat{p}_n(Y|X_i^n)\|_{\mathrm{TV}} > \epsilon_n) \\ &\leq C_1 \sum_n nL_n \exp\left(-C_2 C \log L_n \log(n)\epsilon_n^2\right) \\ &\lesssim \sum_n \exp\left(-C_3(\log L_n + 1)\log(n)\epsilon_n^2\right) \\ &\lesssim \sum_n n^{-C_3\epsilon_n^2} < \infty.\end{aligned}$$

By the Borel-Cantelli lemma, the probability that $\|p(Y|X_i^n) - \hat{p}_n(Y|X_i^n)\|_{\mathrm{TV}} > \epsilon_n$ for infinitely many i, n is 0. Thus, with probability 1, as $n \to \infty$,

$$\sup_i \|p(Y|X_i^n) - \hat{p}_n(Y|X_i^n)\|_{\mathrm{TV}} \to 0$$

□

A.3 Distributions and their uniform tails

In this section we will consider some examples of distributions on S to see if they satisfy the assumptions made above. In particular we will be interested in calculating $\mathrm{gap}_p, \mathrm{ins}_p, \mathrm{del}_p$, and flux_p. This will allow us to see if they satisfy assumption A.10 and are p, k integrable for reasonable choices of k. We will start with an example illustrating how finite-lag autoregressive models may not satisfy assumption A.10. We then look at some simple examples that do satisfy assumption A.10. Finally, we demonstrate that profile hidden Markov models (pHMMs), which are ubiquitously used in biological sequence analysis, satisfy assumption assumption A.10 for a choice of g.

First we illustrate how an autoregressive model may fail to satisfy assumption A.10. In the next example, we create a lag 2 autoregressive model such that, for letters $A, B \in \mathcal{B}$, the motif ABA is high probability while AAA is low probability. Thus a sequence such as $X = AAAA$ may increase in probability by gaining an insertion of B and so $\mathrm{del}_p(X) < \mathrm{ins}_p(X)$.

Proposition A.22. *Let $A \neq B \in \mathcal{B}$. Let $X \sim p$ such that $X_{:2} = AA$ and $X_L \sim p(b|X_{L-2:L})$ where $p(A|AA) = 0.1$, $p(B|AA) = 0.8$, $p(\$|AA) = 0.1$, $p(A|AB) = 0.8$, $p(B|AB) = 0.1$, $p(\$|AB) = 0.1$, $p(A|BA) = 0.8$, $p(B|BA) = 0.1$, and $p(\$|BA) = 0.1$ where \$ represents the end of the sequence. Then $\mathrm{gap}_p(L) < 0$ for large enough L and in particular, p does not satisfy assumption A.10.*

Proof. Call $X = L \times A$. $p(X) = 0.1^{L-1}$, so $\mathrm{del}_p(X) = L(0.1)$. However, $p(L_1 \times A + B + L_2 \times A) = 0.1^{L_1+L_2-2-1}0.8^3$, so, if $L_1 + L_2 = L$, $p(L_1 \times A + B + L_2 \times A)/p(L \times A) = 0.8^3 0.1^{-2}$. Thus, $\mathrm{ins}_p(X) \geq (L-2)(0.8^3 0.1^{-2}) \geq \mathrm{del}_p(X)$ for large enough L. □

On the other hand there are obvious examples of distributions that do satisfy assumption A.10, such as if $p(X) \propto |\mathcal{B}|^{-L} e^{-\mu L}$, where $\mathrm{gap}_p(L) \sim L$ and $p(X) \propto |\mathcal{B}|^{-L} L!^{-1}$ where $\mathrm{gap}_p(L) \sim L^2$.

We now consider pHMMs, which we now define and show satisfy assumption A.10. We also show that pHMMs are subexponential, i.e. if p is a pHMM then $E_p e^{t|X|} < \infty$ for any t small enough; this will be useful for determining if we have p, k integrability with certain kernels. To define a pHMM, we start with a Markov model with "letter" states ${}_s = \{s_1, s_2, \ldots, s_{\tilde{L}}\}$, "insertion" states ${}_i = \{i_0, i_1, \ldots, i_{\tilde{L}}\}$, a start state s_0, and killing state Δ. s_l and i_l may only transfer to $s_{l'}$ for $l' > l$ or $i_{l'}$ for $l' \geq l$. Then each of these hidden states, except s_0 and Δ, emits a $b \in \mathcal{B}$ with probability $p(b|Z)$ for a state Z. Thus a probability of a sequence X with $|X| = L$ can be written as

$$p(X) = \sum_{Z \in \mathcal{I}_L} p(Z)p(X|Z) = \sum_{Z \in \mathcal{I}_L} p(Z_{|X|+1}|Z_{|X|}) \prod_{l=1}^{L} p(Z_l|Z_{l-1})p(X_l|Z_l)$$

where the sum is over, where we define $\mathcal{I}_L = \{(Z_0, Z_1, \ldots, Z_{L+1}) \mid Z_i \in {}_s \cup {}_i \text{ for } 1 \leq i \leq L, Z_{L+1} = \Delta, Z_0 = s_0\}$.

We add a few conditions to our pHMM: The first is that infinite length sequences are not allowed, namely that $\sup_l p(i_l|i_l) \leq \mu$ for some $\mu < 1$. We also impose that $p(b|Z) > 0$ for all states Z and $b \in \mathcal{B}$ and call $\eta = \min_{b,Z} p(b|Z)$. Finally, we ask that if $p(Z|i_l) > 0$ for some state Z and $p(i_l|s_{l'}) > 0$, then $p(Z|s_l) > 0$, that is, if a state can be reached by $s_{l'}$ by first adding an insertion, then it can be reached by $s_{l'}$ directly as well. This last condition guarantees that removing an insertion from any sequence of states Z does not make the sequence probability 0.

Proposition A.23. *If p is a pHMM and $\chi(t) = t \wedge 1$ then $\mathrm{gap}_p(L) \gtrsim L$. Also, $\mathrm{ins}_p(X) \lesssim \mathrm{del}_p(X) \sim \mathrm{flux}_p(X) \sim |X|$ and $E_p e^{t|X|} < \infty$ for any $t < -\log \mu^-$.*

Proof. Let X be a sequence with $|X| = L > 3\tilde{L}$. Also call, for every $\hat{L}, l$, $\mathcal{I}_{\hat{L},s}(l) = \{Z \in \mathcal{I}_{\hat{L}} \mid Z_l \in_s\}$ and $\mathcal{I}_{\hat{L},i}(l) = \{Z \in \mathcal{I}_{\hat{L}} \mid Z_l \in_i\}$.

$$\mathrm{ins}_p(X) = \sum_{|Y|=L+1, XMY} T_{p,X\to Y} = \frac{1}{p(X)} \sum_{l=0}^{L} \sum_{b\in\mathcal{B}} p(X_{b,+l})$$

where $X_{b,+l}$ is the sequence X with an inserted letter b at position l. Now we use the sum over $\mathcal{B}$ to marginalize out the emission at position l:

$$\begin{aligned}
\sum_{b\in\mathcal{B}} p(X_{b,+l}) =& \sum_{b\in\mathcal{B}} \sum_{Z\in\mathcal{I}_{L+1}} p(Z)p(X_{b,+l}|Z) \\
=& \sum_{b\in\mathcal{B}} \sum_{Z\in\mathcal{I}_{L+1}} p(Z) \left(\prod_{l'=0}^{l-1} p(X_{b,+l,l'}|Z_{l'}) \prod_{l'=l}^{L} p(X_{b,+l,l'+1}|Z_{l'+1}) \right) p(X_{b,+l,l}|Z_l) \\
=& \sum_{Z\in\mathcal{I}_{L+1}} p(Z) \prod_{l'=0}^{l-1} p(X_{l'}|Z_{l'}) \prod_{l'=l}^{L} p(X_{l'}|Z_{l'+1}) \left(\sum_{b\in\mathcal{B}} p(b|Z_l) \right) \\
=& \sum_{Z\in\mathcal{I}_{L+1}} p(Z)p(X|\tilde{Z})
\end{aligned}$$

where, for $Z \in \mathcal{I}_{L+1}$, $\tilde{Z} \in \mathcal{I}_L$ is defined to be Z but with Z_l removed. The idea of the proof is to show that the leading terms of the last sum are ones in which Z_l is in the middle of a multiple insertion. For these Z, $Z \mapsto \tilde{Z}$ is an injection and we can replace $p(Z)$ with its upper bound $\mu p(\tilde{Z})$. Now summing over $\tilde{Z}$ just gives us $\mu p(X)$ and finally summing over l and dividing by $p(X)$ gives our bound $L\mu$.

First we will consider Z with a letter in position l, i.e. $Z \in \mathcal{I}_{L+1,s}(l)$. For each $Z \in \mathcal{I}_{L+1,s}(l)$ pick a position l_Z such that $Z_{l_Z} = Z_{l_Z+1} = Z_{l_Z-1} \in_i$, i.e. l_Z is in the middle of a multiple insertion. Define $\hat{Z} \in \mathcal{I}_{L,s}(l-1) \cup \mathcal{I}_{L,s}(l)$ to be Z with l_Z removed. First note, by our choice of l_Z, $p(Z)/p(\hat{Z}) < 1$. As well, since $\hat{Z}$ differs from $\tilde{Z}$ in at most $2\tilde{L}$ positions, $p(X|\tilde{Z}) \leq p(X|\hat{Z})\eta^{-2\tilde{L}}$. Finally, note that at most $\tilde{L}+1$ Z map to the same $\hat{Z}$. Now write

$$\begin{aligned}
\sum_{Z\in\mathcal{I}_{L+1,s}(l)} p(Z)p(X|\tilde{Z}) =& \sum_{Z\in\mathcal{I}_{L+1,s}(l)} \frac{p(Z)}{p(\hat{Z})} \frac{p(X|\tilde{Z})}{p(X|\hat{Z})} p(\hat{Z})p(X|\hat{Z}) \\
\leq& \eta^{-2\tilde{L}} \sum_{Z\in\mathcal{I}_{L+1,s}(l)} p(\hat{Z})p(X|\hat{Z}) \\
\leq& \eta^{-2\tilde{L}}(\tilde{L}+1) \sum_{Z\in\mathcal{I}_{L,s}(l-1)\cup\mathcal{I}_{L,s}(l)} p(Z)p(X|Z) \\
=& \eta^{-2\tilde{L}}(\tilde{L}+1)p(X, Z \in \mathcal{I}_{L,s}(l-1) \cup \mathcal{I}_{L,s}(l)) \\
=& \eta^{-2\tilde{L}}(\tilde{L}+1) \left(p(X, Z_l \in_s) + p(X, Z_{l-1} \in_s) \right).
\end{aligned}$$

For the first term in the sum write

$$\begin{aligned}\frac{1}{p(X)}\sum_{l=0}^{L}\eta^{-2\tilde{L}}(\tilde{L}+1)p(X,Z_l\in_s) =&\eta^{-2\tilde{L}}(\tilde{L}+1)\sum_{l=0}^{L}p(Z_l\in_s|X)\\ =&\eta^{-2\tilde{L}}(\tilde{L}+1)\sum_{l=0}^{L}E\left[\mathbb{1}(Z_l\in_s)|X\right]\\ =&\eta^{-2\tilde{L}}(\tilde{L}+1)E\left[\sum_{l=0}^{L}\mathbb{1}(Z_l\in_s)\middle|X\right]\\ \leq&\eta^{-2\tilde{L}}(\tilde{L}+1)(\tilde{L}+1)=O(1)\end{aligned}$$

Where the last inequality follows from the fact that if $p(Z)>0$, then at most $\tilde{L}+1$ states are letters. The second term is similar.

Next we consider $Z\in\mathcal{I}_{L+1,i}(l)$. Note that at most $\tilde{L}+1$ Z in $\mathcal{I}_{L+1,i}(l)$ map to the same $\tilde{Z}$ and that by the fact that $p(Z)=p(Z_1|Z_0)\times\ldots p(Z_{L+2}|Z_{L+1})$ and our assumptions, there is a M' such that $p(Z)/p(\tilde{Z})\leq M'$ for all $Z\in\mathcal{I}_{L+1,s}(l)$. We will split $\mathcal{I}_{L+1,s}(l)$ into two parts: define $A_1=\{Z\in\mathcal{I}_{L+1,i}(l)\mid Z_{l-1}\neq Z_{l+1}\}$ and $A_2=\{Z\in\mathcal{I}_{L+1,i}(l)\mid Z_{l-1}=Z_{l+1}\}$. That is, if $Z\in A_2$ then $Z_{l-1}=Z_l=Z_{l+1}$, so Z_l must by in $_i$ and so position l is in a multiple insertion. Thus, if $Z\in A_2$, then $p(Z)/p(\tilde{Z})\leq\mu$, and, $Z\mapsto\hat{Z}$ is injective on A_2. On the other hand, if $Z\in A_1$ then $\tilde{Z}_{l-1}\neq\tilde{Z}_l$. Thus,

$$\begin{aligned}\sum_{Z\in A_1}p(Z)p(X|\tilde{Z})\leq&M'\sum_{Z\in A_1}p(\tilde{Z})p(X|\tilde{Z})\leq(\tilde{L}+1)M'p(X,Z_{l-1}\neq Z_l)\\ \sum_{Z\in A_2}p(Z)p(X|\tilde{Z})\leq&\mu\sum_{Z\in A_2}p(\tilde{Z})p(X|\tilde{Z})\leq\mu p(X)\end{aligned}$$

and,

$$\begin{aligned}\frac{1}{p(X)}\sum_{l=0}^{L}\sum_{Z\in A_1}p(Z)p(X|\tilde{Z})\leq&(\tilde{L}+1)M'E\left[\sum_{l=0}^{L}\mathbb{1}(Z_{l-1}\neq Z_l)\middle|X\right]\leq(\tilde{L}+1)M'(2\tilde{L}+2)\\ \frac{1}{p(X)}\sum_{l=0}^{L}\sum_{Z\in A_2}p(Z)p(X|\tilde{Z})\leq&\mu L.\end{aligned}$$

Combining the above results we finally have

$$\mathrm{ins}_p(X)\leq L\mu+O(1)$$

Considering deletions now, we have

$$\mathrm{del}_p(X)=\sum_{|Y|=L-1,XMY}T_{p,X\to Y}=\frac{1}{p(X)}\sum_{l=1}^{L}p(X_{-l})$$

with X_{-l} defined to be X with position l deleted. In this case,

$$p(X_{-l})=\sum_{|Z|=L}p(Z)\prod_{l'=0}^{l-1}p(X_{l'}|Z_{l'})\prod_{l'=l+1}^{L+1}p(X_{l'}|Z_{l'-1}).$$

For $Z\in\mathcal{I}_{L-1,i}(l-1)$, let $\tilde{Z}$ be Z but with an extra i_k in position l if $Z_{l-1}=i_k$. For $Z\in\mathcal{I}_{L-1,i}(l-1)$, $p(Z)/p(\tilde{Z})\geq\mu^{-1}\geq\mu^{-1}p(X_l|\tilde{Z}_l)$ and $Z\mapsto\tilde{Z}$ is a bijection to elements Z' of $\mathcal{I}_L$ such that $Z'_{l-1}=Z'_l\in_i$. Thus we have,

$$\frac{1}{p(X)}p(X_{-l})\geq\mu^{-1}\frac{1}{p(X)}\sum_{Z\in\mathcal{I}_{L,i}(l-1)}p(\tilde{Z})p(X|\tilde{Z})=\mu^{-1}p(Z_{l-1}=Z_l\in_i|X).$$

Now let $R = \sum_{l=0}^{L} \mathbb{1}(Z_{l-1} = Z_l \in_i)$, which is lower bounded by $L - 3\tilde{L}$. Thus

$$\mathrm{del}_p(X) \geq \sum_{|Y|=L-1, XMY} T_{p,X\to Y} \geq \mu^{-1} E_p\left[R|X\right] \geq \mu^{-1}\left(L - 3\tilde{L}\right) = L\mu^{-1} - O(1).$$

On the other hand, letting, for a $z \in_s \cup_i$, $Z_{l,z}$ be Z but with an extra z in position l, there is an $M'' > 0$ such that $p(Z)/\left(\sum_{z\in_s\cup_i} p(Z_{l,z})\right) \leq M''$ and for any z, $p(X_l|Z_{l,z})^{-1} \leq \eta^{-1}$. Finally note that taking each $Z \in \mathcal{I}_{L-1}$ to the set $\{Z_{l,z}\}_{z\in_s\cup_i}$, each $Z \in \mathcal{I}_L$ is counted exactly once. Thus,

$$\frac{1}{p(X)} p(X_{-l}) \leq \eta^{-1} M'' \frac{1}{p(X)} \sum_{Z\in\mathcal{I}_{L-1}} \sum_{z\in_s\cup_i} p(Z_{l,z}) p(X|Z_{l,z}) = \eta^{-1} M''.$$

Thus, as above, $\mathrm{del}_p(X) \leq \eta^{-1} M'' L$. Thus we have $\mathrm{ins}_p(X) \lesssim \mathrm{del}_p(L) \sim \mathrm{flux}_p(X) \sim |X|$ and $\mathrm{gap}_p(L) \geq \left(L\mu^{-1} - O(1)\right) - \left(L\mu + O(1)\right) \gtrsim L$.

Also note that

$$(\mu + o(1)) \sum_{|X|=L} p(X) \geq \frac{1}{L+1} \sum_{|X|=L} p(X) \frac{1}{p(X)} \sum_{l=0}^{L} \sum_{b\in\mathcal{B}} p(X_{b,+l}) = \sum_{|X|=L+1} p(X)$$

so $p(|X| = L) \lesssim e^{-tL}$ if $e^{-t} > \mu$. In particular, if $t < -\log\mu$, then $E_p e^{t|X|} = \sum_L p(|X| = L) e^{tL} < \infty$. □

Thus by our discussion of assumption A.10, we gain the following corollary.

Corollary A.24. *If p is a pHMM and $\chi(t) = t \wedge 1$ then assumption A.10 is satisfied for $V_p(L) = (\log L)^{2+\epsilon}$ for any $\epsilon > 0$. Also, $E_p \mathrm{flux}_p(X) < \infty$.*

A.4 Building kernels for the KSD-B

We've shown that kernels that are deltable and coercive result in KSD-B's with theoretical guarantees outlined above. In this section we describe how to build deltable and coercive scalar and vector field kernels. First in section A.4.1 we will describe some examples of deltable scalar field kernels. One example will be the alignment kernel for which we will present new results describing the thickness of its tails. Then in section A.4.2 we describe how to build vector field kernels from scalar field kernels. Finally, in section A.4.3 we describe five examples of deltable kernels that are coercive for pHMMs described in section A.3. In section A.4.4 we prove the results we described in section A.4.1.

A.4.1 Deltable scalar field kernels

Before describing specific scalar field kernels, we review some basic results about deltability in the next lemma.

Lemma A.25. *(Propositions ... in ...) If k is a deltable kernel and $A : S \to (0, \infty)$, then the tilted kernel k^A is deltable. If k, k' are deltable kernels, the tensorized kernel $k \otimes k'((X, Y), (X', Y')) = k(X, X')k'(X', Y')$ is deltable. If k is a deltable kernel and k' is a kernel then $k + k'$ is deltable. If k is a deltable kernel and $S' \subseteq S$ then k restricted to S' is deltable.*

We now describe three scalar field kernels we will use as build deltable and coercive kernels. We first define the inverse multiquadratic Hamming kernel (IMQ-H) as

$$k_H(X, Y) = (1 + d_H(X, Y))^{1/2}$$

where d_H is the Hamming distance considering all sequences as ending with infinitely many stop symbols \$. We also define the embedding kernel as

$$k_F(X, Y) = \exp\left(-\frac{1}{2\sigma^2} \|F(X) - F(Y)\|_2^2\right)$$

for some $F : S \to \mathbb{R}^D$ for some D, σ. ... showed that k_H is deltable and under general conditions k_F is deltable as well.

We spend the rest of this section considering the alignment kernel. We will first define two kernels that we will use to define alignments: one for comparing individual bases and the other for penalizing insertions. To compare bases let $k_s(X,Y) = |\mathcal{B}|^{-1}\delta_X(Y) \times \mathbb{1}(|X|=1)$ be the identity kernel on $\mathcal{B}$, that is, it is only positive if X and Y are the same length one sequence. To penalize insertions, let $k_I(X,Y) = \exp\left(-\mu(|X|+|Y|) - \Delta\mu\left(\mathbb{1}(|X|\geq 1) + \mathbb{1}(|Y|\geq 1)\right)\right)$ for $0 < \mu < \infty$ and $0 \leq \Delta\mu \leq \infty$. In this case, sequences compared under k_I are interpreted as insertions, penalized with insertion start penalty $\Delta\mu$ and insertion extension penalty μ. Also define the insertion kernel without start penalty $\tilde{k}_I(X,Y) = \exp\left(-\mu(|X|+|Y|)\right)$.

We define the alignment kernel, for a "decay parameter" $\gamma > 0$, as

$$\tilde{k}(X,Y) = \sum_{l, X_1+\cdots+X_{2l+1}=X, Y_1+\cdots+Y_{2l+1}=Y} \gamma^l k_I(X_1,Y_1)\prod_{i=1}^{l} k_s(X_{2i},Y_{2i})k_I(X_{2i+1},Y_{2i+1}) \tag{10}$$

where the sum is over all numbers l and partitions of X and Y into l aligned letters $X_2, X_4, \ldots, Y_2, Y_4, \ldots$ and $l+1$ insertions $X_1, X_3, \ldots, Y_1, Y_3, \ldots$. We also define the alignment kernel without insertion starts at the ends of the sequence

$$\tilde{k}_{ne}(X,Y) = \sum \gamma^l \tilde{k}_I(X_1,Y_1)\left(\prod_{i=1}^{l-1} k_s(X_{2i},Y_{2i})k_I(X_{2i+1},Y_{2i+1})\right)k_s(X_{2l},Y_{2l})\tilde{k}_I(X_{2l+1},Y_{2l+1}).$$

Let $\zeta = 2\mu + \log\gamma - 2\log|\mathcal{B}|$. Deltability of k and k_{ne} depends on $\Delta\mu$ and ζ.

Proposition A.26. *(Theorems 22 and 25 of ...) $\tilde{k}$ and $\tilde{k}_{ne}$ are deltable if and only if $\Delta\mu = \infty$, or $\Delta\mu > 0$ and $\zeta \geq 0$, or $\Delta\mu = 0$ and $\zeta > 0$.*

Unlike the IMQ-H and embedding kernels, alignment kernels are not normalized, i.e. $\tilde{k}(X,X)$ depends on X. This is usually dealt with by working with the tilted kernel $k(X,Y) = \tilde{k}(X,X)^{-1/2}\tilde{k}(Y,Y)^{-1/2}\tilde{k}(X,Y)$. However, it is difficult to discern if this kernel is coercive. Instead we first tilt $\tilde{k}, \tilde{k}_{ne}$ into more convenient forms: define $k = \tilde{k}^{\tilde{A}}$, $k = \tilde{k}_{ne}^{\tilde{A}}$ for $\tilde{A}(X) = \exp(\mu|X|)$. This form will allow us to find an upper bound for $\sqrt{k(X,X)}$ and find a $h \in \mathcal{H}_k$ with thick tails. Finally we will decide an appropriate tilting function A such that $\sqrt{k^A(X,X)} = A(X)\sqrt{k(X,X)}$ is not too large to preclude a distribution p from being p,k integrable, and hA still has thick enough tails to be coercive for reasonable p.

We now must find a bound for $\sqrt{k(X,X)}$ and a $g \in \mathcal{H}_k$ with thick tails. These results are different when $\Delta\mu < \infty$ and $\Delta\mu = \infty$. We first consider the case when $\Delta\mu < \infty$. To state these results, we will need to define $\xi = 1 - e^{-\Delta\mu} < 1$ and the function

$$r_1(x,\xi) = \frac{1}{2}\left(1 + x + \sqrt{(1+x)^2 - 4\xi x}\right).$$

To get a deltable kernel, we must have $\zeta \geq 0$, in which case $r_1(e^{\zeta/2}|\mathcal{B}|,\xi) \geq r_1(e^{\zeta/2},\xi) > 1$. Now we state our results on the tails of the alignment kernel, which will be proven in section A.4.4.

Proposition A.27. *Say $\Delta\mu < \infty$, then*

$$L^{1/2} r_1(e^{\zeta/2}|\mathcal{B}|,\xi)^{|X|} \leq \sqrt{k(X,X)} \leq r_1(e^{\zeta/2}|\mathcal{B}|,\xi)^{|X|}$$

and for any $\pi < 1$, there is a $h \in \mathcal{H}_k$ such that $h(X)$ depends only on $|X|$ and

$$h(X) = r_1(\pi e^{\zeta/2},\xi)^{|X|} + O(|X|).$$

As well, for any $X \in S$, $k_x \lesssim h$.

To achieve a bounded kernel, we may pick $A = r_1(e^{\zeta/2}|\mathcal{B}|,\xi)^{-|X|}$, in which case $k^A(X,X) \leq 1$. However, in this case $A(X)g(X) \sim \left(\frac{r_1(\pi e^{\zeta/2},\xi)}{r_1(e^{\zeta/2}|\mathcal{B}|,\xi)}\right)^{|X|}$ which decays exponentially and is thus unlikely to be coercive. We will further discuss how to tilt this kernel to get a coercive kernel in section A.4.3.

We now finish the section considering the case when $\Delta\mu = \infty$ and pick $\zeta = -\log|\mathcal{B}|$. In this case $k(X,Y) = 0$ if $X \neq Y$ so k is not interesting. However this is not the case for k_{ne}. In fact, k_{ne} can be interpreted as an infinite k-mer spectrum kernel, i.e. it uses the counts of k-mers in each sequence as its features.

Proposition A.28. *Say $\Delta\mu = \infty$ and $\zeta = -\log|\mathcal{B}|$. For a $Y, X \in S$ call $\phi_Y(X)$ the number of times Y appears in X. Then $k_{ne}(X, X') = \sum_{Y\in S}\phi_Y(X)\phi_Y(X')$. Now let $A(X) = |X|^{-3/2}$. k^A_{ne} is a bounded C_0 kernel, i.e. for all $f \in \mathcal{H}_k$, $f \in C_0(S)$. k^A_{ne} is also non-vanishing, i.e. $\sqrt{k_{ne}(X,X)} \not\to 0$ as $|X| \to \infty$. As well, letting $h = \sum_{X\in\mathcal{B}} k^A_{ne,X}$ then $h(X)$ depends only on $|X|$ and $h(X) = |X|^{-1/2} + C|X|^{-3/2} + o(|X|^{-3/2})$ for some constant C.*

A.4.2 Vector field kernels for KSDs

We now describe how to build vector field kernels from scalar filed kernels. Our main tool will be a correspondence between kernels on some space and vector field kernels. To state this correspondence, we will need the following definition.

Definition A.29. *A sign on M is a $\sigma : M \to \{-1, 1\}$ such that $\sigma(X, Y) = -\sigma(Y, X)$ for all $(X, Y) \in M$. Define $M^\sigma = \{(X, Y) \in M \mid \sigma(X, Y) = 1\}$. For a $(X, Y) \in M$, define $(X, Y)^\sigma = (X, Y)$ if $\sigma(X, Y) = 1$ and (Y, X) otherwise. Say σ is proper if $\sigma(X, Y) = 1$ if $|Y| = |X| - 1$ for $(X, Y) \in M$.*

Proposition A.30. *Let σ be a sign on M. There is a correspondence between kernels on M^σ and vector field kernels such that a kernel on M^σ, k, corresponds to the vector field kernel*

$$((X, Y), (X', Y')) \mapsto \sigma(X, Y)\sigma(X', Y')k((X, Y)^\sigma, (X', Y')^\sigma)$$

and a vector field kernel corresponds to its restriction to M^σ. Deltable kernels k on M^σ, i.e. kernels such that $\delta_{(X,Y)} \in \mathcal{H}_k$ for all $(X, Y) \in M^\sigma$, correspond to deltable vector field kernels.

Proof. The first statement, including the bijectivity of the correspondence, is clear except that the mapping from a kernel M^σ to a vector field kernel defines a non-negative definite vector field kernel. We will now show this. Let k be a kernel on M^σ, $(Z_n)_{n=1}^N \subset M$ be distinct, and $(\alpha_n)_{n=1}^N \subset \mathbb{R}$. For $Z \in M$, call $\alpha_Z = \alpha_n$ if $Z = Z_n$ and 0 if $Z \neq Z_n$ for any n. For $(X, Y) \in M$, call $(X, Y)^{-\sigma} = (Y, X)$ if $\sigma = 1$ and (X, Y) otherwise.

$$\begin{aligned}\sum_n\sum_m \sigma(Z_n)\sigma(Z_m)\alpha_n\alpha_m k(Z_n^\sigma, Z_m^\sigma)& \\ &= \sum_{Z\in M}\sum_{Z'\in M}\sigma(Z)\sigma(Z')\alpha_Z\alpha_{Z'}k(Z^\sigma, Z'^\sigma) \\ &= \sum_{Z\in M^\sigma}\sum_{Z'\in M^\sigma}\left(\alpha_Z - \alpha_{Z^{-\sigma}}\right)\left(\alpha_{Z'} - \alpha_{Z'^{-\sigma}}\right)k\left(Z, Z'\right) \geq 0.\end{aligned}$$

To check that this defines a vector field kernel, call $\tilde{k}$ the extension of the kernel k to M. Then if $f \in \mathcal{H}_{\tilde{k}}$,

$$f(X, Y) = \left(f\middle|\tilde{k}\left((X, Y), \cdot\right)\right)_{\tilde{k}} = -\left(f\middle|\tilde{k}\left((Y, X), \cdot\right)\right)_{\tilde{k}} = -f(Y, X).$$

The second statement follows from the fact that if k is a kernel on M^σ and $\tilde{k}$ is its corresponding vector field kernel, then if $f \in \mathcal{H}_k$ then there is a $\tilde{f} \in \mathcal{H}_{\tilde{k}}$ such that $\tilde{f}(X, Y) = \sigma(X, Y)f((X, Y)^\sigma)$. To see this note that $k_{(X,Y)} \mapsto \tilde{k}_{(X,Y)}$ can define a unitary linear transformation on finite linear combinations of $\{k_{(X,Y)}\}_{(X,Y)\in M^\sigma}$. This transformation takes f that are finite linear combinations of $\{k_{(X,Y)}\}_{(X,Y)\in M^\sigma}$ to the above defined $\tilde{f}$ and can be extended to all of $\mathcal{H}_k$ to obey the same property. □

We can now use this correspondence to build vector field kernels by building kernels on M^σ.

Proposition A.31. *Let k, k' be kernels on S. The following are kernels on M^σ.*

$$\begin{gathered}((X, Y), (X', Y')) \mapsto k(X, X')k'(Y, Y') \\ ((X, Y), (X', Y')) \mapsto \left(k(X, X') + k'(Y, Y')\right)^2 \\ ((X, Y), (X', Y')) \mapsto k(X + Y, X' + Y') \\ ((X, Y), (X', Y')) \mapsto k(X, X') \\ ((X, Y), (X', Y')) \mapsto k(X, X')\mathbb{1}(|X| \neq |Y|, |X'| \neq |Y'|).\end{gathered}$$

If k, k' are deltable then the corresponding vector field kernels of the first two of these kernels are deltable.

Proof. That the first four of these kernels are non-negative definite because they are restrictions of non-negative definite kernels on $S \times S$. The last kernel can be constructed by first defining the kernel $((X,Y),(X',Y')) \mapsto k(X,X')$ on $S \times S$, restricting to $\{(X,Y) \in M^\sigma \mid |X| \neq |Y|\}$ and then extending to the rest of M^σ by setting $k_{(X,Y)} = 0$ if $|X| = |Y|$.

If k, k' are deltable, the first kernel described above is deltable by Lemma A.25 as it is the restriction of $k \otimes k'$ on $S \times S$. The second kernel is also deltable by Lemma A.25 as $(k(X,X') + k'(Y,Y'))^2 = \left(k(X,X')^2 + k(Y,Y')^2\right) + 2k \otimes k'((X,Y),(X',Y'))$ so that is the sum of two kernels, one deltable. □

A.4.3 Coercive vector field kernels with delta functions

We will now use the above tools to describe three scalar field kernels and two vector field kernels that are deltable and coercive for pHMMs. In this section we will assume $\chi(t) = t \wedge 1$ so that the result sof Proposition A.23 and Corollary A.24 hold. To demonstrate that our kernels are coercive we will use the following lemma.

Lemma A.32. *Say p is a pHMM and f is a vector field such that $f(X,Y) = 0$ if $|X| = |Y|$. Call $f(L) = f(X,Y)$ for $|X| = L$ and $|Y| = L-1$ and assume $f(L+1) \lesssim f(L)$, that is, f does not increase super-exponentially. Say g is another vector field with $g(X,Y) = o(f(|X|))$ as $|X| \to \infty$; in particular, $g = 0$ satisfies this condition. Say f is eventually positive and $(\sup_{|X|=L} \mathrm{ins}_p(X))\,(f(L) - f(L+1)) > -(1-\epsilon)\mathrm{gap}_p(L)f(L)$ eventually for some $\epsilon > 0$; in particular the later condition is satisfied if f is non-increasing in L. Then $\mathcal{T}_p(f+g)(X) \geq \frac{1}{2}\epsilon \mathrm{gap}_p(L)f(L)$ eventually.*

Proof. Say $X \in S$ and $|X| = L$. Since $\mathrm{flux}_p(X) \sim \mathrm{gap}_p(|X|)$ by proposition A.23,

$$\begin{aligned}\mathcal{T}_p(f+g)(X) =& -\mathrm{ins}_p(X)f(L+1) + \mathrm{del}_p(X)f(L) + \mathrm{flux}_p(X)o(f(L)+f(L+1)) \\ \geq& \mathrm{gap}_p(L)f(L) + \mathrm{ins}_p(X)\,(f(L) - f(L+1)) + \mathrm{flux}_p(X)o(f(L)) \\ \geq& \mathrm{gap}_p(L)f(L)(\epsilon + o(1)).\end{aligned}$$

□

In this case, with the notation of the lemma, since $\mathrm{ins}_p(X) \lesssim \mathrm{gap}_p(|X|) \sim \mathrm{flux}_p(X)$ and we can set $V_p(X) = (\log|X|)^{2+\epsilon}$ for some $\epsilon > 0$ by Proposition A.23 and Corollary A.24,

$$\sum_L \frac{\inf_{|X|=L} \mathcal{T}_p(f+g)(X)}{\left(\sup_{|X|=L} \mathrm{flux}_p(X)\right) V_p(L)} \gtrsim \sum_L \frac{f(L)}{(\log L)^{2+\epsilon}}.$$

Thus, if we can find f, g such that $f(L) \gtrsim L^{-(1-\delta)}$ for some $\delta > 0$ then we can satisfy coercitivity for pHMMs.

As well, recall from Proposition A.23 that $E_p e^{t|X|} < \infty$ for small enough t and $\mathrm{flux}_p(X) \sim |X|$ so that any kernel with $\sqrt{k(X,X)} \leq e^{t'|X|}$ for small enough t' will be such that p is p,k integrable.

Now we introduce our examples of kernels. For the rest of this section, assume σ is a *proper* ordering. We may for example let σ be the lexicographic ordering for some ordering of the letters in $\mathcal{B}$.

Unboundedly tilted IMQ-H. Our first scalar field kernel will be the unboundedly tilted IMQ-H (UT IMQ-H):

$$k(X,Y) = A(X)k_H(X,Y)A(Y)$$

for $A(X) = (|X|+1)^{3/2}$. k is deltable and $\sqrt{k(X,X)} = (|X|+1)^{3/2}$, so if p is a pHMM, p is p,k-integrable. Let $h = -k_\emptyset$, so $h(X) = -|X| - 1$. $\nabla h(X,Y) = |X| - |Y|$, so, by Lemma A.32 with $f = \nabla h$ and $g = 0$ and the following discussion, k is coercive for pHMMs.

Unboundedly tilted alignment kernel with gaps. Our second scalar field kernel will be the unboundedly tilted alignment kernel with gaps (UT AwG):

$$k(X,Y) = A(X)k(X,Y)A(Y)$$

for $\Delta\mu < \infty, \zeta \geq 0$ and $A(X) = r_1(e^{\zeta/2}, \xi)^{-(1-\epsilon)|X|}$ for some small $\epsilon > 0$. Let h be as defined in Proposition A.27 picking $\pi < 1$ such that $\frac{r_1(\pi e^{\zeta/2},\xi)}{r_1(e^{\zeta/2},\xi)^{1-\epsilon}} = 1 + \delta$ for a small δ. h is a function only of the length of the sequence and $h(X) = (1+\delta)^{|X|} + o(1)$. Call $f = -\nabla\left(X \mapsto (1+\delta)^{|X|}\right)$ and $g = f - \nabla(-h) = o(1)$.

$$f(X,Y) = 0 \text{ if } |X| = |Y|$$

$$f(X,Y) = \delta(\epsilon)(1+\delta)^{|X|-1} \text{ if } |Y| = |X| - 1$$

$$f(X,Y) - f(Z,Y) = -\delta^2(1+\delta)^{|X|-1} = -\delta^2(1+\delta)^{-1}h(X) \text{ if } |Y| < |X| < |Z|.$$

Thus, if

$$\left(\sup_{|X|=L} \text{ins}_p(X)\right)\delta^2(1+\delta)^{-1} \leq (1-\epsilon')\text{gap}_p(L)$$

eventually for some $\epsilon' > 0$, i.e. δ is small enough, then k is coercive for pHMMs by Lemma A.32 with f and g. However, by Proposition A.27,

$$L^{-1/2}\left(\frac{r_1(e^{\zeta/2}|\mathcal{B}|,\xi)}{r_1(e^{\zeta/2},\xi)^{1-\epsilon}}\right)^L \leq \sup_{|X|=L}\sqrt{k(X,X)}.$$

One can check that this ratio is minimized in the limit $\epsilon = 0, \xi = 0, \zeta = 0$ in which case

$$\frac{r_1(e^{\zeta/2}|\mathcal{B}|,\xi)}{r_1(e^{\zeta/2},\xi)^{1-\epsilon}} \geq \frac{r_1(|\mathcal{B}|^{1/2},0)}{r_1(1,0)} = \frac{|\mathcal{B}|^{1/2}+1}{2}.$$

This may be too large for some pHMMs to have p,k integrability given that $\frac{1}{2}\left(|\mathcal{B}|^{1/2}+1\right)$ is 3/2 in the case when $\mathcal{B}$ is the set of nucleotide where $|\mathcal{B}| = 4$ and approximately 2.74 in the case when $\mathcal{B}$ is the set of amino acids where $|\mathcal{B}| = 20$.

Unboundedly tilted alignment kernel without gaps. Our third scalar field kernel will be the unboundedly tilted alignment kernel without gaps (UT AwoG):

$$k(X,Y) = k_{ne}(X,Y)$$

with $\Delta\mu = \infty$ and $\zeta = -|\mathcal{B}|$. In this case, k is deltable and $\sqrt{k(X,X)} \sim L^{3/2}$ so that if p is a pHMM, it has p,k-integrability. Finally, Let h be as defined in Proposition A.28 so that h is a function only of sequence length and $h(X) = |X| + C' + o(1)$ for come $C > 0$. Now by Lemma A.32 with $f(X,Y) = |X| - |Y|$ and $g = \nabla(-h) - f = o(1)$, k is coercive for pHMMs.

IMQ-H plus alignment. We define our first vector field kernel to be the IMQ-H plus alignment kernel (IMQ-H+A). Our strategy will be to add a thick tailed IMQ-H kernel that is not deltable with a thin tailed alignment kernel that is. First define the Hamming to be

$$k_1((X,Y),(X',Y')) = k_H(X,X')\mathbb{1}(|X| \neq |Y|, |X'| \neq |Y'|)$$

on M^σ. Next define the alignment component to be

$$k_2((X,Y),(X',Y')) = (k^A(X,X') + k^A(Y,Y'))^2$$

where $k^A(X,X') = A(X)A(X')k^{\tilde{A}}(X,X')$ where k is the alignment kernel with $\Delta\mu < \infty$ and $\Delta\mu > 0$ or $\zeta > 0$ for $A(X) = \left(r_1(e^{\zeta/2}|\mathcal{B}|^{1/2},\xi)\right)^{-|X|}$ on M^σ. Now define $k = k_1 + k_2$. k is deltable as k_2 is deltable by Lemma A.25. k is also bounded as k_1, k_2 are. Let $h = -k_{(\emptyset,A)}$ for some $A \in \mathcal{B}$. Now let $f = -k_{1,(\emptyset,A)}$ so that $f(X,Y) = (|Y|+1)^{-1/2}$ if $|Y| < |X|$ and $f(X,Y) = 0$ if $|X| = |Y|$. Finally define $g = h - f = k_{2,(\emptyset,A)}$ Let $\tilde{h}$ be as defined in Proposition A.27 for some $0 < \pi < 1$. We have that

$$k^A(X,X') \lesssim h(X)A(X) \sim \left(\frac{r_1(\pi e^{\zeta/2},\xi)}{r_1(e^{\zeta/2}|\mathcal{B}|^{1/2},\xi)}\right)\exp(-c|X'|)$$

for some $c > 0$ when $|X| = 0$ or $|X| = 1$. Thus, $g(X,Y) = O(e^{-c|X|})$. Thus, the kernel is coercive for pHMMs by Lemma A.32 with f and g.

Alignment without gaps plus alignment. We define our final vector field kernel to be the alignment without gaps plus alignment kernel (AwoG+A). First define the coercive alignment-without-gaps component to be

$$k_1((X,Y),(X',Y')) = A(X)A(X')k_{ne}^{\tilde{A}}(X,X')\mathbb{1}(|X| \neq |Y|, |X'| \neq |Y'|)$$

where k_{ne} is the alignment kernel with $\Delta\mu = \infty$ and $A(X) = L^{-3/2}$. Next define the alignment component to be the same as above,

$$k_2((X,Y),(X',Y')) = (k^A(X,X') + k^A(Y,Y'))^2$$

where $k^A(X,X') = A(X)A(X')k^{\tilde{A}}(X,X')$ where k is the alignment kernel with $\Delta\mu < \infty$ and $\Delta\mu > 0$ or $\zeta > 0$ for $A(X) = \left(r_1(e^{\zeta/2}|\mathcal{B}|^{1/2},\xi)\right)^{-|X|}$ on M^σ. Now define $k = k_1 + k_2$. Again, k is deltable as k_2 is deltable. k is also bounded as k_1, k_2 are. Let $h = -\sum_{X\in\mathcal{B}} k_{(X,X+X)}$. Now let $f = -\sum_{X\in\mathcal{B}} k_{1,(X,X+X)}$ so that, by Proposition A.28, $f(X,Y) = (|Y|+1)^{-1/2}$ if $|Y| < |X|$ and $f(X,Y) = 0$ if $|X| = |Y|$. Finally define $g = h - f = -\sum_{X\in\mathcal{B}} k_{2,(X,X+X)}$ As in the previous example, $g(X,Y) = O(e^{-c|X|})$. Thus, the kernel is coercive for pHMMs by Lemma A.32 with f and g.

We will also combine these last two kernels with an embedding kernel.

A.4.4 Proofs for]' the alignment kernel

In this section we will be interested in bounding $\sqrt{k(X,X)}$ and finding a thick tailed $h \in \mathcal{H}_k$ where k is the alignment kernel. We will also do the same for the alignment kernel without an affine gap penalty at the ends of sequences k_{ne}.

Let us review some results for the case when $|\mathcal{B}| = 1$ that will be useful. If $\mathcal{B} = \{A\}$, call $k(L,L') = k(L\times A, L'\times A)$. ... showed that there is an orthogonal basis $(e_L)_L$ such that $\|e_L\|_k = e^{-\zeta/2}$ where $\zeta = 2\mu + \log\gamma$ in this case, $(e_{L'}, k_L)_k \geq 0$ for all L, L' and $(e_{L'}, k_L)_k = 0$ if $L' > L$. Then, defining the infinite upper triangular matrix M such that $M_{L',L} = (e_{L'}, k_L)_k$, we get

$$k(L,L') = (k_L|k_{l'})_k = \left(\sum_{L''} M_{L'',L}e^{\zeta}e_{L''}\middle|\sum_{L''} M_{L'',L'}e^{\zeta}e_{L''}\right)_k = \sum_{L''=0}^{\infty} M_{L'',L}M_{L'',L'}e^{L''\zeta}. \tag{11}$$

The same equation holds for k_{ne} for another matrix M_{ne}.

The exact values of the entries of the matrices M, M_{ne} will be important to achieve bounds on the tails of the alignment kernel. ... showed that if we define $\xi = 1 - e^{-\Delta\mu}$, $f_\xi(y) = \frac{1-\xi y}{1-y}$, and the formal power series

$$F_\xi(x,y) = \frac{f_\xi(y)}{1 - xyf_\xi(y)} = \frac{1-\xi y}{1-(1+x)y+\xi xy^2}$$

$$F_{\xi,ne}(x,y) = xy\frac{\left(\frac{1}{1-y}\right)^2}{1-xyf_\xi(y)} + \frac{1}{1-y} = \frac{xy}{(1-y)^2\left(1-(1+x)y+\xi xy^2\right)} + \frac{1}{1-y}$$

then $M_{L',L} = [x^{L'}y^L]F_\xi(x,y)$ and $M_{ne,L',L} = [x^{L'}y^L]F_{\xi,ne}(x,y)$ where $[x^{L'}y^L]$ denotes the coefficient in front of the term $x^{L'}y^L$ of the following formal power series.

We now show that we can write the size of $k(X,X)$ and $h(X)$ in terms of $F_\xi, F_{\xi,\neq}$.

Proposition A.33. *Calling $C_L = [y^L]F_\xi(e^{\zeta/2}|\mathcal{B}|^{1/2}, y)$, $L^{-1/2}C_L \leq \sup_{|X|=L}\sqrt{k(X,X)} \leq C_L$ and the same inequality is true for k_{ne} and $F_{\xi,ne}$.*

Proof. First of all, by equation 10, if $A \in \mathcal{B}$, we clearly have $k(X,X) \leq k(|X|\times A, |X|\times A)$. $k_s(A,A) = |\mathcal{B}|^{-1}$, so k restricted to $\{\emptyset, A, AA, AAA, \dots\}$ is identical to the string kernel in the

case $|\mathcal{B}| = 1$ and with decay parameter $\gamma|\mathcal{B}|^{-1}$. Thus, by equation 11 with $2\mu + \log\left(\gamma|\mathcal{B}|^{-1}\right) = \zeta + \log|\mathcal{B}|$,

$$\begin{aligned}
k(L,L) &= \sum_{L'=0}^{L} e^{\zeta L'} |\mathcal{B}|^{L'} M_{L',L}^2 \\
&\leq \left(\sum_{L'=0}^{\infty} e^{\zeta L'/2} |\mathcal{B}|^{L'/2} M_{L',L} \right)^2 \\
&= \left(\sum_{L'=0}^{\infty} \left(e^{\zeta/2} |\mathcal{B}|^{1/2} \right)^{L'} [x^{L'} y^L] F_\xi(x,y) \right)^2 \\
&= \left([y^L] F_\xi(e^{\zeta/2}|\mathcal{B}|^{1/2}, y) \right)^2 .
\end{aligned}$$

The result is identical with $F_{\xi,ne}$. On the other hand,

$$\begin{aligned}
k(L,L) =& L \left(\frac{1}{L} \sum_{L'=0}^{L} \left(e^{\zeta L'/2} |\mathcal{B}|^{L'/2} M_{L',L} \right)^2 \right) \\
\geq& L \left(\frac{1}{L} \sum_{L'=0}^{L} e^{\zeta L'/2} |\mathcal{B}|^{L'/2} M_{L',L} \right)^2 \\
=& \frac{1}{L} \left([y^L] F(e^{\zeta/2}|\mathcal{B}|^{1/2}, y) \right)^2 .
\end{aligned}$$

□

Now we build h.

Proposition A.34. *Say $0 < \pi < 1$. There is an $h \in \mathcal{H}_k$ such that $(h|k_X)_k = [y^{|X|}] F_\xi\left(\pi e^{\zeta/2}, y\right)$. There is a $h \in \mathcal{H}_{k_{ne}}$ such that $(h|k_{ne,X})_k = C + [y^{|X|}] F_{\xi,ne}\left(e^{\zeta/2}\pi, y\right)$ for some constant C. As well, for any $X \in S$, $k_X \lesssim h$.*

Proof. Define $k_L = |\mathcal{B}|^L \sum_{|X|=L} k_X$. If $Y, Y' \in S$ and $|Y| = |Y'| = L'$, we have, by Proposition 21 of ..., that $(k_L|k_Y)_k = (k_L|k_{Y'})_k$, thus $(k_L|k_Y)_k = (k_L|k_{L'})_k$. Now note that ... showed in Theorem 23 that k restricted to $\{k_0, k_1, \dots\}$ is is identical to the string kernel in the case $|\mathcal{B}| = 1$ with decay parameter $\gamma|\mathcal{B}|^{-2}$. We will create a h for this kernel with $(h|k_L)_k = [y^L] F_\xi\left(e^{\zeta/2}\pi, y\right)$ and the Proposition will follow from the fact that $(h|k_Y)_k = (h|k_{|Y|})_k$.

We define $h = \sum_L \alpha^L k_L$ for some α. Note that since $k(X,Y) \geq 0$ for all X, Y, $k_X \lesssim h$ for any X. Now write

$$(h|e_{L'})_k = \sum_L \alpha^L [x^{L'} y^L] F_\xi(x,y) = [x^{L'}] F_\xi(x, \alpha).$$

Thus, since $F_\xi(x,y) = f_\xi(y)(1 - xy f_\xi(y))^{-1} = f(y) \sum_{L=0}^{\infty} x^L (y f_\xi(y))^L$,

$$\|h\|_k^2 = \sum_L e^{\zeta L} \left([x^L] F_\xi(x,\alpha) \right)^2 = f_\xi(\alpha)^2 \sum_L e^{\zeta L} \left(\alpha f_\xi(\alpha) \right)^{2L}$$

which is finite as long as $\pi = \alpha f_\xi(\alpha) e^{\zeta/2} < 1$. We can pick α to let π be any positive value < 1. In this case

$$\begin{aligned}
(h|k_{L'})_k =& \sum_L e^{\zeta L} (h|e_{L'})_k M_{L,L'} \\
=& \sum_L e^{\zeta L} \left(f_\xi(\alpha) (\alpha f_\xi(\alpha))^L \right) [x^L y^{L'}] F_\xi(x,y) \\
=& f_\xi(\alpha) \sum_L \left(\pi e^{\zeta/2} \right)^L [x^L y^{L'}] F_\xi(x,y) \\
=& f_\xi(\alpha) [y^{L'}] F_\xi\left(\pi e^{\zeta/2}, y \right).
\end{aligned}$$

We now turn to the very similar case of h_{ne}. The norm of $h = \sum_l \alpha^l k_{ne,l}$ is

$$\sum_L e^{\zeta L} \left([x^L]F_{\xi,ne}(x,\alpha)\right)^2 = \left(\frac{\alpha}{1-\alpha}\right)^2 + \left(\frac{\alpha}{(1-\alpha)^2}\right)^2 \sum_{L=1}^{\infty} e^{\zeta L} \left(\alpha f_\xi(\alpha)\right)^{2(L-1)}$$

which is finite again as long as $\pi = \alpha f_\xi(\alpha) e^{\zeta/2} < 1$.

$$\begin{aligned}(h|k_{ne,L'})_k =& \sum_L e^{\zeta L} M_{ne,L,L'} \left([x^L]F_\xi(x,\alpha)\right) \\ =& \frac{1}{1-\alpha} M_{ne,0,L'} + \frac{\alpha}{(1-\alpha)^2 \alpha f_\xi(\alpha)} \sum_{L=1}^{\infty} e^{\zeta L/2} M_{ne,L,L'} \pi^L \\ =& \frac{1}{1-\alpha} - \frac{\alpha}{(1-\alpha)^2 \alpha f_\xi(\alpha)} + \frac{\alpha}{(1-\alpha)^2 \alpha f_\xi(\alpha)} [y^{L'}] F_{\xi,ne}(e^{\zeta/2}\pi, y).\end{aligned}$$

□

Thus, to analyze the tails of the alignment kernel, we will need to analyze $[y^L]F(x,y)$ and $[y^L]F_{\xi,ne}(x,y)$. The coefficients of $F, F_{\xi,ne}$ will depend on the polynomial $1-(1+x)y+\xi y^2$. We rewrite the polynomial $1-(1+x)y+\xi y^2 = (1-r_1 y)(1-r_2 y)$ for $r_1(\xi,x) \geq r_2(\xi,x)$, which are

$$\frac{1}{2}\left(1+x \pm \sqrt{(1+x)^2 - 4\xi x}\right).$$

These values are decreasing with ξ, positive, and distinct when $\xi < 1$ since $(1+x)^2 - 4\xi x > (1+x)^2 - 4x = (x-1)^2 \geq 0$. When $\xi < 1$, r_1 is also always > 1 since it is $> \frac{1}{2}(1+x+|x-1|) = x \vee 1$. When $\xi = 0$, $r_1 = x+1, r_2 = 0$. We now see that if $\Delta\mu < \infty$ then the coefficients of F_ξ and $F_{\xi,ne}$ grow exponentially. However, if $\Delta\mu = \infty$ the coefficients may grow or shrink exponentially or, in the case of $F_{\xi,ne}$ grow exponentially or polynomially.

Proposition A.35. *If $\xi < 1$ and $x \geq 0$, both $[y^L]F_\xi(x,y)$ and $[y^L]F_{\xi,ne}(x,y)$ are equal to $Cr_1(x,\xi)^L + O(L)$ for some (different) $C > 0$. If $\xi = 1$ then $[y^L]F_1(x,y) = x^L$ and if $x > 1$, $[y^L]F_{1,ne}(x,y) = x^L + O(L)$ otherwise if $x < 1$, $[y^L]F_{1,ne}(x,y) = CL + C' + o(1)$ for some $C > 0, C'$ and if $x = 1$, $[y^L]F_{1,ne}(x,y) = L(L-1)/2 + 1$.*

Proof. First let us consider the case of $\xi = 0$.

$$F_0(x,y) = \frac{1}{1-(1+x)y} = \sum_{L=0}^{\infty} (1+x)^L y^L$$

$$F_{0,ne}(x,y) = \frac{xy}{(1-y)^2(1-(1+x)y)} + \frac{1}{1-y}.$$

By partial fraction decomposition, for some A, B, C with $A, B \neq 0$ and $C \neq 1$, constant c_1, c_2,

$$\begin{aligned}F_{0,ne}(x,y) =& \frac{Axy}{1-(1+x)y} + \frac{xy(By-C)}{(1-y)^2} + \frac{1}{1-y} \\ =& c_0 + c_1 y + \sum_{L=2}^{\infty} \left(Ax(1+x)^{L+1-1} + Bx\binom{L+1-2}{1} - BCx\binom{L-1}{1} + 1\right) y^L.\end{aligned}$$

The leading term in the brackets is $Ax(1+x)^{L-1}$ and, since the coefficients of $F_{0,ne}$ are positive, $A > 0$.

Now we consider the case when $0 < \xi < 1$.

$$F_\xi(x,y) = \frac{(1-\xi y)}{(1-r_1 y)(1-r_2 y)}$$

so by partial fraction decomposition, for $A, B \neq 0$,

$$F_\xi(x,y) = (1-\xi y)\left(\frac{A}{1-r_1 y} + \frac{b}{1-r_2 y}\right) = c_1 \sum_{L=0}^{\infty} \left(Ar_1^L - A\xi r_1^{L-1} + Br_2^L - B\xi r_2^{L-1}\right) y^L.$$

Since $r_1 > 1 > \xi$, the leading term in the brackets is $A(1-\xi/r_1)r_1^L$. Similarly, $[y^L]F_{\xi,ne} = Cr_1^L + O(L)$ for some $C > 0$.

Now we look at when $\xi = 1$. Here $f_\xi(y) = 1$. Thus,

$$F_1(x,y) = \frac{1}{1-xy} = \sum_{L=0}^{\infty} x^L y^L$$

$$F_{1,ne}(x,y) = \frac{xy}{(1-y)^2(1-xy)} + \frac{1}{1-y}.$$

If $x \neq 1$, again, by partial fraction decomposition, By partial fraction decomposition, for some A, B, C with $A, B \neq 0$ and $C \neq 1$, constant c_1, c_2,

$$\begin{aligned} F_{1,ne}(x,y) =& \frac{Axy}{1-xy} + \frac{Bxy(y-C)}{(1-y)^2} + \frac{1}{1-y} \\ =& c_0 + c_1 y + \sum_{L=2}^{\infty} \left(Ax^L + Bx\binom{L+1-2}{1} - BCx\binom{L+1-1}{1} + 1 \right) y^L \end{aligned}$$

So that the leading term is Ax^L if $x > 1$ or $Bx\binom{L-1}{1} - BCx\binom{L}{1}$ if $x < 1$. since $C \neq 1$, the later term is $= CL + C'$ for some $C, C' > 0$. If $x = 1$,

$$F_{1,ne}(x,y) = 1 + \sum_{L=1}^{\infty} \left(\binom{L+1-1}{2} + 1 \right) y^L$$

so that $[y^L]F_{1,ne}(x,y) = L(L-1)/2$. □

Now combining the results of these last three propositions, we have proven Proposition A.27. To begin proving Proposition A.28, we first tighten our estimate of $\sqrt{k(X,X)}$ in the case $\Delta\mu = \infty$.

Proposition A.36. *Say* $\Delta\mu = \infty$. $\sup_{|X|=L} \sqrt{k(X,X)} = \left(e^{\zeta/2}|\mathcal{B}|^{1/2}\right)^L$ *and* $\sup_{|X|=L} \sqrt{k_{ne}(X,X)}$ *is* $\sim \left(e^{\zeta/2}|\mathcal{B}|^{1/2}\right)^L$ *if* $e^{\zeta/2}|\mathcal{B}|^{1/2} > 1$, *is* $\sim L^{3/2}$ *if* $e^{\zeta/2}|\mathcal{B}|^{1/2} = 1$, *and is* $\sim L$ *if* $e^{\zeta/2}|\mathcal{B}|^{1/2} < 1$.

Proof. When $\Delta\mu = \infty$, M is the identity matrix, so that

$$k(L \times A, L \times A) = e^{L\zeta}|\mathcal{B}|^L.$$

On the other hand, $M_{ne,0,L} = 1$ for all L and, since, for $L \geq L' > 0$,

$$[x^{L'}y^L]F_{1,ne}(x,y) = [y^L]\frac{y}{(1-y)^2}y^{L'-1} = [y^{L-L'}](1-y)^{-2} = L - L' + 1.$$

Thus, calling $\lambda = e^\zeta|\mathcal{B}|$,

$$\begin{aligned} k(L \times A, L \times A) =& 1 + \sum_{L'=1}^{L} \lambda^{L'}(L-L')^2 \\ =& 1 + \lambda^{L+1} \sum_{L'=1}^{L} \lambda^{-(L-L'+1)}(L-L'+1)^2 \\ =& 1 + \lambda^{L+1} \sum_{L'=1}^{L} \lambda^{-L'}L'^2. \end{aligned}$$

If $\lambda > 1$, the sum is increasing and bounded, so, $k(L \times A, L \times A) = 1 + C\lambda^L(1+o(1))$ for some $C > 0$. If $\lambda = 1$, we have $k(L \times A, L \times A) = 1 + CL^3(1+o(1))$ for some $C > 0$. Finally, if $\lambda < 1$, since

$$k(L \times A, L \times A) = 1 + L^2 \sum_{L'=1}^{L} \lambda^{L'} \left(1 - \frac{L'}{L}\right)^2,$$

the sum is increasing and bounded with L so that $k(L \times A, L \times A) = 1 + CL^2(1+o(1))$ for some $C > 0$. □

Next we must look at when a tilted alignment kernel is C_0.

Proposition A.37. *Say $\tilde{A} : \mathbb{N} \to (0, \infty)$ and $A(X) = \tilde{A}(|X|)$. If k^A is a bounded kernel, then it is C_0 if and only if $\tilde{A}(L)[y^L]F_\xi(\pi e^{\zeta/2}|\mathcal{B}|^{1/2}, y) \to 0$ for any $\pi < 1$ if and only if the same condition holds for any $\pi > 0$.*

Proof. Let $b \in \mathcal{B}$ and define $g = C\sum_L \alpha^L k_{L\times b}$ for some α, C. By the same logic as Proposition A.34, for any $0 < \pi < 1$, we can pick α, C, such that $g_\pi \in \mathcal{H}_k$ and $g_\pi(L \times b) = [y^L]F(\pi e^{\zeta/2}|\mathcal{B}|, y)$. Thus $g_\pi A \in \mathcal{H}_{k^A}$. Since k^A is bounded, if it is C_0 then $g_\pi A$ must be in $C_0(S)$. On the other hand if $g_\pi A \in C_0(S)$ for some π,

$$\sup_{|X|=L,|Y|=L'} k^A(X, Y) = \tilde{A}(L)\tilde{A}(L')k(L \times b, L' \times b) \leq \tilde{A}(L)\tilde{A}(L')(C\alpha^L)^{-1}g_\pi(L') \to 0$$

as $L' \to \infty$ so $k^A_X \in C_0(S)$ for all X. Finally, $g_\pi A \in C_0(S)$ if and only if $\tilde{A}(L)[y^L]F_\xi(\pi e^{\zeta/2}|\mathcal{B}|^{1/2}, y) \to 0$. □

Finally we prove Proposition A.28

Proposition A.38. *(Proposition A.28) Say $\zeta = -\log|\mathcal{B}|$ and $\Delta\mu = \infty$. For a $Y, X \in S$ call $\phi_Y(X)$ the number of times Y appears in X. Then $k_{ne}(X, X') = \sum_{Y\in S}\phi_Y(X)\phi_Y(X')$. Now let $A(X) = |X|^{-3/2}$. k^A_{ne} is a bounded C_0 kernel, i.e. for all $f \in \mathcal{H}_k$, $f \in C_0(S)$. k^A_{ne} is also non-vanishing, i.e. $\sqrt{k_{ne}(X, X)} \not\to 0$ as $|X| \to \infty$. As well, letting $h = \sum_{X\in\mathcal{B}} k^A_{ne,X}$ then $h(X)$ depends only on $|X|$ and $h(X) = |X|^{-1/2} + C|X|^{-3/2}$ for some constant C.*

Proof. First note that since $\Delta\mu = \infty$ the insertion penalty kernel k_I defined in section A.4.1 has $k_I(X, Y) = 0$ unless $X = Y = \emptyset$. Thus,

$$k_{ne}(X, Y) = e^{\mu(|X|+|Y|)} \sum_{X_2=Y_2} \gamma^l \tilde{k}_I(X_1, Y_1)|\mathcal{B}|^{-l}\tilde{k}_I(X_3, Y_3)$$

where the sum is over numbers l, and partitions $X_1 + X_2 + X_3 = X$ and $Y_1 + Y_2 + Y_3 = Y$ where $X_2 = Y_2$ and $|X_2| = |Y_2| = l$. Note $\tilde{k}_I(X_1, Y_1)\tilde{k}_I(X_3, Y_3) = e^{-\mu(|X|-|X_2|+|Y|-|Y_2|}$, so,

$$k_{ne}(X, Y) = \sum_{X_2=Y_2} \gamma^l|\mathcal{B}|^{-l}e^{\mu l}.$$

Since $\gamma|\mathcal{B}|^{-1}e^\mu = e^\zeta|\mathcal{B}| = 1$, $k_{ne}(X, Y)$ is simply a sum of matching substrings of X and Y, which is equal to $\sum_{Y\in S}\phi_Y(X)\phi_Y(X')$.

First note $e^{\zeta/2}|\mathcal{B}|^{1/2} = 1$, so, by proposition A.36, $\sup_{|X|=L}\sqrt{k_{ne}(X, X)} = CL^{3/2} + C' + o(1)$ for some $C > 0$, C'. Thus, k^A_{ne} is bounded and non-vanishing. On the other hand, if $\pi < 1$, $[y^L]F_\xi(\pi e^{\zeta/2}|\mathcal{B}|^{1/2}, y) \sim L$, so, by Proposition A.37, k^A_{ne} is C_0.

Finally, letting $h = \sum_{X\in\mathcal{B}} k_{ne,X}$ and noting that if $X \in \mathcal{B}$, $k_{ne,X}(Y) = \#(X \text{ in } Y) + 1$ (the plus one for $\phi_\emptyset$), we have that $h(Y) = |Y| + 4$. □